Learning from Semi-Supervised Weak-Label Data

Hao-Chen Dong, Yu-Feng Li, Zhi-Hua Zhou
donghc@lamda.nju.edu.cn

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

What is the work about?

- In many multi-label learning tasks, it's difficult to get full relevant label set and label incompleteness significantly influences the performance of multi-label learning.

- Previous studies usually consider only one scenario in which labels are incomplete. In fact, there are various scenarios. The combination of these conditions have not been fully studied.

- We propose SSWL method to address semi-supervised weak-label problem which labels can be partially known and completely unknown.

Outline

- Background
- Proposed Method
- Experiments
- Conclusion

Background

- In many real-world tasks, one instance usually has more than one label. Traditional supervised learning based on one label per instance is out of its capability to cope with this problem.

car, *road*, *sky*, *grass*

tower, *sky*

- **Multi-label learning** deals with instances associated with a set of labels and it has attracted much attention.

Background

- But, it's difficult to get full label set.

car, *road*, *sky*, *grass*

full

tower, *sky*

empty

- We can use some labeled instances and many unlabeled instances for training. This kind of multi-label learning problem is **semi-supervised multi-label learning problem**.

Background

- But, it's difficult to get full label set.

car, *road*, *sky*, *grass*

partial

tower, *sky*

partial

- In many cases, we only use labeled instances for training, but we can't make sure if we get full relevant label set, it may be partial. This is **weak-label learning problem**.

Background

- Semi-supervised multi-label learning tries to handle the issue of empty relevant label set.

- Transductive multi-label learning methods (Liu, Jin and Yang 2006; Chen et al. 2008; Guo and Schuurmans 2012; Kong, Ng and Zhou 2013) that assume testing instances are from unlabeled instances.

- Pure semi-supervised multi-label learning methods (Zhao and Guo 2015; Zhan and Zhang 2017) try to make multi-label prediction for any unseen instance.

Background

- Weak-label learning focuses on the issue of partial relevant label set.
- WELL (Sun, Zhang and Zhou 2010) is based on the assumption that instance similarities are determined by a group of low-rank similarity matrixes.
- MLR-GL (Bucak, Jin and Jain 2011) uses group lasso to regularize the training errors.

Background

- More general, we usually have instances with partial or empty label sets for training.

car, *road*, *sky*, *grass*

partial

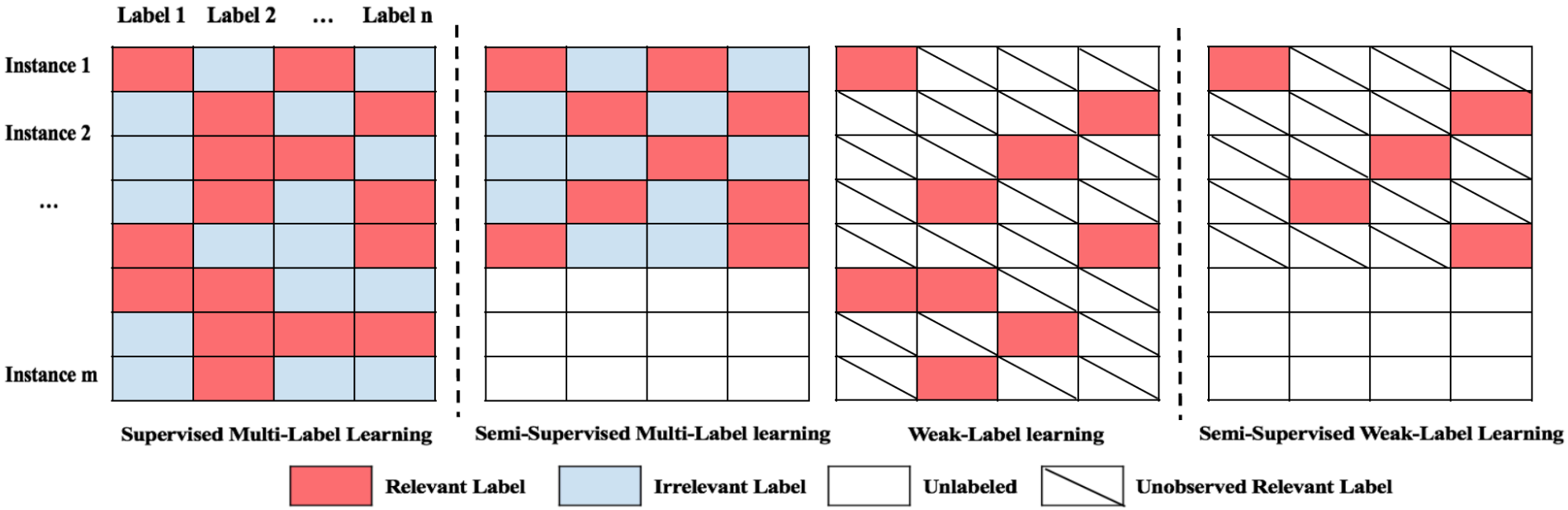
tower, *sky*

empty

- We call this problem as **semi-supervised weak-label learning problem**. This problem considers that labels are partially known and completely unknown.

Proposed Method

- We've summarized four problem settings.



- The number of known labels are not more than the existing problems'. To learn a promising predictive label, we need to complement missing relevant labels at first.

Proposed Method

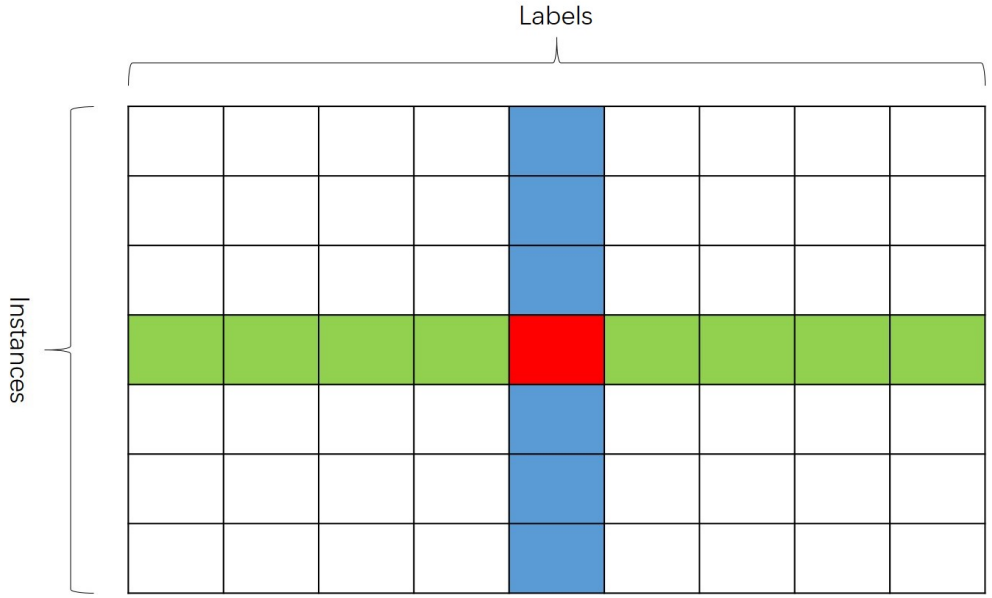
- Traditional label propagation method use instance similarity. The relevant label set of an instance can be derived by that of its nearest neighbors.

- In multi label learning, the assignment of one certain label on training instances can be derived by the assignments of its adjacent labels.

- In order to take label correlation into account, we propose to use both the instance and label similarity for the complementation of missing relevant labels.

Proposed Method

- We have this picture in our mind.



- Then we can get a new regularization term to complement missing relevant labels.

$$\begin{aligned}\Omega(\widehat{\mathbf{Y}},\mathbf{S},\mathbf{L}) &= \sum_{ij}(\hat{Y}_{ij}-\sum_{p\in\mathcal{N}_i}\sum_{q\in\hat{\mathcal{N}}_j}S_{ip}\hat{Y}_{pq}L_{qj})^2 \\ &= \|\widehat{\mathbf{Y}}-\mathbf{S}\widehat{\mathbf{Y}}\mathbf{L}\|_F^2\end{aligned}$$

Proposed Method

- In semi-supervised weak-label learning problem, we don't have the irrelevant labels, we only have some relevant labels.

- Basic ideas
  - We can find the relevant labels in the uncertainty labels, and the rest of them are irrelevant labels.
  - We can employ ensemble learning which is known to be more robust than a single model.

Proposed Method

- We first build two models with the new regularization term for labeled and unlabeled instances respectively. Formally, let $\mathbf{XW}$ and $\mathbf{X}\bar{\mathbf{W}}$ denote two linear multi-label models.

- The first model $\mathbf{XW}$ is initialized to predict the observed relevant labels, whose objective is formulated as $\|(\mathbf{XW}) \circ \mathbf{C} - \mathbf{C}\|_F^2$.

- The second model $\mathbf{X}\bar{\mathbf{W}}$ is initialized to predict the uncertain elements in the label occurrence matrix $\mathbf{C} \in \{0,1\}^{m \times n}$, whose objective is formulated as $\|(\mathbf{X}\bar{\mathbf{W}}) \circ (\mathbf{E} - \mathbf{C}) + (\mathbf{E} - \mathbf{C})\|_F^2$.

- We then leverage them via the promising co-regularization framework to derive a robust predictive result, whose objective is cast as $\|(\mathbf{X}(\mathbf{W} - \bar{\mathbf{W}})) \circ (\mathbf{E} - \mathbf{C})\|_F^2$.

Proposed Method

- So we combine these and get the final objective function.

$$\begin{aligned}\min_{\mathbf{W},\bar{\mathbf{W}},\mathbf{L}} \quad & \|(\mathbf{XW}) \circ \mathbf{C} - \mathbf{C}\|_F^2 + \alpha\Omega(\mathbf{U},\mathbf{S},\mathbf{L}) + \\ & \beta\|(\mathbf{X}(\mathbf{W}-\bar{\mathbf{W}})) \circ (\mathbf{E}-\mathbf{C})\|_F^2 + \\ & \zeta\|(\mathbf{X}\bar{\mathbf{W}}) \circ (\mathbf{E}-\mathbf{C}) + (\mathbf{E}-\mathbf{C})\|_F^2 \\ \text{s.t.} \quad & \mathbf{U} = (\mathbf{XW}) \circ \mathbf{C} + (\mathbf{X}\bar{\mathbf{W}}) \circ (\mathbf{E}-\mathbf{C})\end{aligned}$$

- Where α, β, ζ are the parameters. $\mathbf{U} = (\mathbf{XW}) \circ \mathbf{C} + (\mathbf{X}\bar{\mathbf{W}}) \circ (\mathbf{E}-\mathbf{C})$ is the integrated prediction of two models and we can smooth two models at the same time.

- It is worth noting that classical label propagation techniques can be realized as a special case of our proposal and our method can make prediction for any unseen instance.

Proposed Method

- We observe the objective function is a bi-convex function.

- We extend an efficient block coordinate descend algorithm to optimize the objective function.

Algorithm 1: SSWL Method

Input : $\mathbf{X}$: $m \times d$ instance matrix
$\mathbf{C}$: $m \times n$ label occurrence matrix
$\mathbf{S}$: $m \times m$ similarity matrix of instances
Output: $\mathbf{W}$ and $\bar{\mathbf{W}}$: $d \times n$ coefficient matrixes

Initialize $\mathbf{W}$, $\bar{\mathbf{W}}$, $\mathbf{L}$;
while *not converged* **do**
    Fix $\bar{\mathbf{W}}$ and $\mathbf{L}$, update $\mathbf{W}$ by Eq.5;
    Fix $\mathbf{W}$ and $\mathbf{L}$, update $\bar{\mathbf{W}}$ by Eq.7;
    Fix $\mathbf{W}$ and $\bar{\mathbf{W}}$, update $\mathbf{L}$ by Eq.9;
end

Proposed Method

- If we directly use the derivative of the objective function, we will get the complex matrix equation.

- We can use this theorem to transform the matrix equation to the normal linear equation.

> **Theorem 1.** *(Horn and Johnson 1991) Suppose a matrix* $\hat{\mathbf{X}}$ *satisfies an equation,* $\sum_{i=1}^{b} \mathbf{A}_i \hat{\mathbf{X}} \mathbf{B}_i = \boldsymbol{V}$*, where* $\{\mathbf{A}_i\}_{i=1}^{b}$*,* $\{\mathbf{B}_i\}_{i=1}^{b}$ *and* $\boldsymbol{V}$ *are known. To obtain the solution* $\hat{\mathbf{X}}$*, one could solve the following equivalent problem instead,* $(\sum_{i=1}^{b} \mathbf{B}_i' \otimes \mathbf{A}_i)\boldsymbol{vec}(\hat{\mathbf{X}}) = \boldsymbol{vec}(\boldsymbol{V})$*, which is a normal linear equation.*

- In this way, we can use some efficient algorithms, such as the conjugate gradient algorithm, to solve the equation.

Proposed Method

- We use the optimized solution of the objective function as our final $\mathbf{L}$.

- After getting the coefficient matrix $\mathbf{W}$, we use 0 as the threshold to discretize the predictive result. If the predictive result is greater than 0, we set the label as 1, otherwise we set the label as -1.

Experiments

- Datasets
  - *TMC* dataset
  - *Yeast* dataset
  - *SceneImage* dataset
  - *msrc dataset*
- Comparative methods
  - state-of-the-art weak-label learning methods

  MLR-GL (Bucak, Jin, and Jain 2011) **SSML** (Zhao and Guo 2015)
  - state-of-the-art supervised multi-label learning method

  ML-KNN (Zhang and Zhou 2007)
  - three naive methods that directly decompose the task into multiple binary classification problems via treating labels independently.

  Well-SVM (Li et al. 2013) **S4VM** (Li and Zhou 2015) **BSVM**
  - our method without using unlabeled data

  SSWL-wo

Experiments

- Incomplete label ratio (I.L. Ratio)
  - we randomly drop $\{0\%, 20\%, 40\%, 60\%\}$ of the observed labels on the labeled training data.

- Evaluation criteria
  - **Micro-F1** and **Macro-F1**

  take both precision and recall into account

  the larger the value, the better the performance
  - **Hamming Loss** (**H.L.**)

  evaluates the fraction of misclassified instance-label pairs

  the smaller the value, the better the performance

Experiments

- Results on *TMC* dataset are shown in Table 2

Table 2: Experimental results (mean±std) on *TMC*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.640 ± .001**	.639 ± .001	.612 ± .003	.615 ± .001	.638 ± .001	.501 ± .001	.578 ± .002	.487 ± .002
	20%	**.602 ± .003**	.578 ± .001	.556 ± .002	.596 ± .002	.580 ± .002	.213 ± .002	.506 ± .001	.292 ± .001
	40%	**.582 ± .001**	.455 ± .004	.356 ± .002	.461 ± .003	.423 ± .001	.032 ± .001	.365 ± .003	.023 ± .002
	60%	**.570 ± .002**	.505 ± .001	.113 ± .002	.563 ± .002	.160 ± .022	.012 ± .001	.215 ± .002	.007 ± .003
Macro-F1(↑)	0%	.618 ± .002	**.620 ± .003**	.586 ± .002	.588 ± .002	.613 ± .002	.467 ± .001	.545 ± .002	.464 ± .002
	20%	**.582 ± .001**	.568 ± .001	.519 ± .001	.567 ± .001	.543 ± .002	.175 ± .001	.457 ± .002	.244 ± .001
	40%	**.566 ± .003**	.409 ± .005	.295 ± .003	.413 ± .002	.368 ± .002	.024 ± .001	.309 ± .002	.017 ± .001
	60%	**.553 ± .002**	.494 ± .001	.089 ± .005	.537 ± .001	.125 ± .029	.008 ± .001	.279 ± .002	.004 ± .001
H.L.(↓)	0%	**.065 ± .002**	.069 ± .001	.067 ± .002	.076 ± .001	.067 ± .001	.082 ± .002	.085 ± .002	.080 ± .001
	20%	.075 ± .002	.075 ± .001	.072 ± .001	.078 ± .002	**.071 ± .002**	.092 ± .002	.087 ± .005	.088 ± .001
	40%	**.079 ± .001**	.082 ± .002	.083 ± .002	.086 ± .002	.080 ± .001	.099 ± .001	.092 ± .002	.100 ± .001
	60%	.087 ± .003	.089 ± .002	.097 ± .003	**.084 ± .002**	.095 ± .003	.101 ± .002	.111 ± .003	.101 ± .002

Experiments

- Results on *TMC* dataset are shown in Table 2

Table 2: Experimental results (mean±std) on *TMC*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.640 ± .001**	.639 ± .001	.612 ± .003	.615 ± .001	.638 ± .001	.501 ± .001	.578 ± .002	.487 ± .002
	20%	**.602 ± .003**	.578 ± .001	.556 ± .002	.596 ± .002	.580 ± .002	.213 ± .002	.506 ± .001	.292 ± .001
	40%	**.582 ± .001**	.455 ± .004	.356 ± .002	.461 ± .003	.423 ± .001	.032 ± .001	.365 ± .003	.023 ± .002
	60%	**.570 ± .002**	.505 ± .001	.113 ± .002	.563 ± .002	.160 ± .022	.012 ± .001	.215 ± .002	.007 ± .003
Macro-F1(↑)	0%	.618 ± .002	**.620 ± .003**	.586 ± .002	.588 ± .002	.613 ± .002	.467 ± .001	.545 ± .002	.464 ± .002
	20%	**.582 ± .001**	.568 ± .001	.519 ± .001	.567 ± .001	.543 ± .002	.175 ± .001	.457 ± .002	.244 ± .001
	40%	**.566 ± .003**	.409 ± .005	.295 ± .003	.413 ± .002	.368 ± .002	.024 ± .001	.309 ± .002	.017 ± .001
	60%	**.553 ± .002**	.494 ± .001	.089 ± .005	.537 ± .001	.125 ± .029	.008 ± .001	.279 ± .002	.004 ± .001
H.L.(↓)	0%	**.065 ± .002**	.069 ± .001	.067 ± .002	.076 ± .001	.067 ± .001	.082 ± .002	.085 ± .002	.080 ± .001
	20%	.075 ± .002	.075 ± .001	.072 ± .001	.078 ± .002	**.071 ± .002**	.092 ± .002	.087 ± .005	.088 ± .001
	40%	**.079 ± .001**	.082 ± .002	.083 ± .002	.086 ± .002	.080 ± .001	.099 ± .001	.092 ± .002	.100 ± .001
	60%	.087 ± .003	.089 ± .002	.097 ± .003	**.084 ± .002**	.095 ± .003	.101 ± .002	.111 ± .003	.101 ± .002

- SSWL obtains quite promising performance and consistently performs robustly while the I.L. Ratio getting higher.

Experiments

- Results on *TMC* dataset are shown in Table 2

Table 2: Experimental results (mean±std) on *TMC*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.640 ± .001**	.639 ± .001	.612 ± .003	.615 ± .001	.638 ± .001	.501 ± .001	.578 ± .002	.487 ± .002
	20%	**.602 ± .003**	.578 ± .001	.556 ± .002	.596 ± .002	.580 ± .002	.213 ± .002	.506 ± .001	.292 ± .001
	40%	**.582 ± .001**	.455 ± .004	.356 ± .002	.461 ± .003	.423 ± .001	.032 ± .001	.365 ± .003	.023 ± .002
	60%	**.570 ± .002**	.505 ± .001	.113 ± .002	.563 ± .002	.160 ± .022	.012 ± .001	.215 ± .002	.007 ± .003
Macro-F1(↑)	0%	.618 ± .002	**.620 ± .003**	.586 ± .002	.588 ± .002	.613 ± .002	.467 ± .001	.545 ± .002	.464 ± .002
	20%	**.582 ± .001**	.568 ± .001	.519 ± .001	.567 ± .001	.543 ± .002	.175 ± .001	.457 ± .002	.244 ± .001
	40%	**.566 ± .003**	.409 ± .005	.295 ± .003	.413 ± .002	.368 ± .002	.024 ± .001	.309 ± .002	.017 ± .001
	60%	**.553 ± .002**	.494 ± .001	.089 ± .005	.537 ± .001	.125 ± .029	.008 ± .001	.279 ± .002	.004 ± .001
H.L.(↓)	0%	**.065 ± .002**	.069 ± .001	.067 ± .002	.076 ± .001	.067 ± .001	.082 ± .002	.085 ± .002	.080 ± .001
	20%	.075 ± .002	.075 ± .001	.072 ± .001	.078 ± .002	**.071 ± .002**	.092 ± .002	.087 ± .005	.088 ± .001
	40%	**.079 ± .001**	.082 ± .002	.083 ± .002	.086 ± .002	.080 ± .001	.099 ± .001	.092 ± .002	.100 ± .001
	60%	.087 ± .003	.089 ± .002	.097 ± .003	**.084 ± .002**	.095 ± .003	.101 ± .002	.111 ± .003	.101 ± .002

- SSWL-wo also obtains good performance but is not as good as SSWL.

Experiments

- Results on *TMC* dataset are shown in Table 2

Table 2: Experimental results (mean±std) on *TMC*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.640 ± .001**	.639 ± .001	.612 ± .003	.615 ± .001	.638 ± .001	.501 ± .001	.578 ± .002	.487 ± .002
	20%	**.602 ± .003**	.578 ± .001	.556 ± .002	.596 ± .002	.580 ± .002	.213 ± .002	.506 ± .001	.292 ± .001
	40%	**.582 ± .001**	.455 ± .004	.356 ± .002	.461 ± .003	.423 ± .001	.032 ± .001	.365 ± .003	.023 ± .002
	60%	**.570 ± .002**	.505 ± .001	.113 ± .002	.563 ± .002	.160 ± .022	.012 ± .001	.215 ± .002	.007 ± .003
Macro-F1(↑)	0%	.618 ± .002	**.620 ± .003**	.586 ± .002	.588 ± .002	.613 ± .002	.467 ± .001	.545 ± .002	.464 ± .002
	20%	**.582 ± .001**	.568 ± .001	.519 ± .001	.567 ± .001	.543 ± .002	.175 ± .001	.457 ± .002	.244 ± .001
	40%	**.566 ± .003**	.409 ± .005	.295 ± .003	.413 ± .002	.368 ± .002	.024 ± .001	.309 ± .002	.017 ± .001
	60%	**.553 ± .002**	.494 ± .001	.089 ± .005	.537 ± .001	.125 ± .029	.008 ± .001	.279 ± .002	.004 ± .001
H.L.(↓)	0%	**.065 ± .002**	.069 ± .001	.067 ± .002	.076 ± .001	.067 ± .001	.082 ± .002	.085 ± .002	.080 ± .001
	20%	.075 ± .002	.075 ± .001	.072 ± .001	.078 ± .002	**.071 ± .002**	.092 ± .002	.087 ± .005	.088 ± .001
	40%	**.079 ± .001**	.082 ± .002	.083 ± .002	.086 ± .002	.080 ± .001	.099 ± .001	.092 ± .002	.100 ± .001
	60%	.087 ± .003	.089 ± .002	.097 ± .003	**.084 ± .002**	.095 ± .003	.101 ± .002	.111 ± .003	.101 ± .002

- Our proposal works better than state-of-the-art weak-label learning and semi-supervised multi-label learning algorithms.

Experiments

- Results on *yeast* dataset.

Table 3: Experimental results (mean±std) on *yeast*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	**.647 ± .001**	.619 ± .002	.612 ± .002	.623 ± .003	.584 ± .001	.625 ± .002	.592 ± .001	.623 ± .002
	20%	**.638 ± .001**	.626 ± .002	.534 ± .003	.618 ± .005	.510 ± .002	.506 ± .001	.511 ± .001	.509 ± .001
	40%	**.604 ± .002**	.554 ± .003	.394 ± .003	.379 ± .004	.152 ± .001	.103 ± .002	.432 ± .001	.188 ± .003
	60%	**.616 ± .002**	.568 ± .002	.241 ± .002	.209 ± .002	.046 ± .031	.002 ± .002	.320 ± .002	.019 ± .007
Macro-F1(↑)	0%	**.635 ± .001**	.594 ± .001	.582 ± .003	.600 ± .001	.557 ± .002	.602 ± .001	.578 ± .002	.592 ± .002
	20%	**.618 ± .001**	.613 ± .001	.494 ± .002	.593 ± .004	.476 ± .001	.470 ± .001	.476 ± .002	.478 ± .002
	40%	**.574 ± .001**	.538 ± .005	.359 ± .004	.340 ± .002	.126 ± .001	.083 ± .001	.397 ± .001	.145 ± .002
	60%	**.595 ± .002**	.554 ± .001	.194 ± .003	.177 ± .002	.039 ± .025	.001 ± .001	.280 ± .005	.016 ± .003
H.L.(↓)	0%	**.207 ± .001**	.209 ± .001	.211 ± .001	.213 ± .002	.215 ± .002	.208 ± .001	.214 ± .001	.209 ± .001
	20%	**.210 ± .001**	.216 ± .001	.221 ± .002	.211 ± .002	.224 ± .002	.225 ± .001	.253 ± .002	.224 ± .001
	40%	**.225 ± .001**	.252 ± .004	.246 ± .003	.251 ± .001	.286 ± .002	.294 ± .001	.257 ± .001	.279 ± .002
	60%	**.231 ± .003**	.268 ± .003	.278 ± .002	.275 ± .003	.299 ± .010	.305 ± .001	.286 ± .002	.302 ± .002

Experiments

- Results on *SceneImage* dataset.

Table 4: Experimental results (mean±std) on *SceneImage*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	.589 ± .003	.583 ± .002	**.613 ± .002**	.528 ± .002	.499 ± .002	.456 ± .001	.489 ± .001	.538 ± .002
	20%	**.572 ± .002**	.545 ± .001	.558 ± .003	.419 ± .001	.394 ± .002	.389 ± .002	.458 ± .003	.392 ± .001
	40%	**.540 ± .002**	.534 ± .002	.395 ± .001	.220 ± .001	.174 ± .001	.094 ± .001	.289 ± .001	.205 ± .001
	60%	**.521 ± .003**	.517 ± .002	.251 ± .002	.000 ± .001	.019 ± .010	.000 ± .000	.300 ± .001	.010 ± .002
Macro-F1(↑)	0%	**.576 ± .002**	.553 ± .001	.567 ± .001	.454 ± .003	.407 ± .001	.362 ± .001	.466 ± .001	.437 ± .001
	20%	**.550 ± .003**	.505 ± .001	.494 ± .002	.320 ± .001	.295 ± .002	.295 ± .002	.417 ± .001	.284 ± .001
	40%	**.523 ± .002**	.499 ± .001	.306 ± .001	.140 ± .001	.113 ± .001	.056 ± .001	.207 ± .002	.134 ± .002
	60%	**.510 ± .004**	.495 ± .001	.165 ± .002	.000 ± .001	.011 ± .005	.000 ± .000	.230 ± .001	.005 ± .003
H.L.(↓)	0%	.184 ± .001	.186 ± .001	**.167 ± .002**	.192 ± .001	.193 ± .001	.192 ± .002	.245 ± .002	.167 ± .001
	20%	.199 ± .001	.201 ± .001	**.182 ± .002**	.199 ± .001	.203 ± .001	.204 ± .002	.236 ± .002	.194 ± .001
	40%	**.206 ± .001**	.214 ± .001	**.206 ± .001**	.219 ± .001	.227 ± .001	.240 ± .002	.226 ± .001	.221 ± .001
	60%	**.208 ± .002**	.222 ± .001	.220 ± .002	.249 ± .001	.247 ± .003	.250 ± .002	.229 ± .002	.248 ± .002

Experiments

- Results on *msrc* dataset.

Table 5: Experimental results (mean±std) on *msrc*. ↑ (↓) indicates the larger (smaller) the better. The best performance and its comparable performances are bolded (pairwise t-tests at 95% significance level).

	I.L. Ratio	SSWL	SSWL-wo	Well-SVM	MLR-GL	SSML	ML-kNN	S4VM	BSVM
Micro-F1(↑)	0%	.590 ± .003	.550 ± .003	.505 ± .003	.533 ± .003	**.604 ± .002**	.355 ± .001	.351 ± .001	.487 ± .001
	20%	**.571 ± .001**	.513 ± .003	.444 ± .001	.472 ± .002	.569 ± .003	.307 ± .001	.331 ± .002	.292 ± .002
	40%	.507 ± .001	.449 ± .005	.216 ± .002	.500 ± .001	**.533 ± .002**	.122 ± .002	.316 ± .001	.047 ± .003
	60%	**.465 ± .001**	.367 ± .002	.118 ± .003	.244 ± .001	.462 ± .001	.034 ± .001	.294 ± .001	.007 ± .001
Macro-F1(↑)	0%	.562 ± .003	.505 ± .004	.412 ± .002	.465 ± .004	**.612 ± .001**	.291 ± .002	.378 ± .001	.464 ± .001
	20%	**.522 ± .001**	.419 ± .004	.302 ± .002	.390 ± .002	.518 ± .002	.216 ± .001	.369 ± .002	.244 ± .002
	40%	**.491 ± .002**	.396 ± .004	.151 ± .003	.424 ± .002	.472 ± .001	.072 ± .002	.317 ± .002	.032 ± .002
	60%	**.430 ± .001**	.298 ± .001	.082 ± .002	.166 ± .001	.391 ± .002	.037 ± .001	.282 ± .001	.004 ± .001
H.L.(↓)	0%	**.083 ± .001**	.085 ± .001	**.083 ± .002**	.093 ± .002	.108 ± .002	.094 ± .001	.154 ± .001	.092 ± .001
	20%	**.085 ± .001**	.091 ± .001	.086 ± .001	.101 ± .001	.105 ± .002	.095 ± .001	.163 ± .002	.088 ± .001
	40%	.090 ± .001	.098 ± .002	.098 ± .003	**.083 ± .002**	.100 ± .001	.101 ± .001	.182 ± .002	.100 ± .002
	60%	**.092 ± .001**	.099 ± .001	.103 ± .002	.100 ± .002	.094 ± .001	.106 ± .001	.201 ± .001	.101 ± .001

Conclusion

- We consider semi-supervised weak-label learning problem which is a new kind of multi-label learning problem.

- We propose the SSWL method to address this problem.
  - both instance similarity and label similarity are considered for the complement of missing labels
  - ensemble of multiple models is employed which is more robust than a single model when the label information is insufficient
  - experiments on a number of real tasks validate the effectiveness of SSWL in handling the semi-supervised weak-label learning problem

- Code is on http://lamda.nju.edu.cn/donghc/

Thanks!

Review opinions

Masked Reviewer ID: Assigned_Reviewer_1
Review:

Question	
[Summary] Please summarize the main claims/contributions of the paper in your own words.	This manuscript essentially elaborates the problem for semi-supervised weak-label learning and proposes a bi-convex optimization problem with block coordinate descend algorithm. The authors have given a clear comparison between other works and theirs. In its present form, the organization is clear and easy to follow.
[Relevance] Is this paper relevant to an AI audience?	Likely to be of interest to a large proportion of the community
[Significance] Are the results significant?	Significant
[Novelty] Are the problems or approaches novel?	Novel
[Soundness] Is the paper technically sound?	Technically sound
[Evaluation] Are claims well-supported by theoretical analysis or experimental results?	Very convincing
[Clarity] Is the paper well-organized and clearly written?	Good
[Detailed Comments] Please elaborate on your assessments and provide constructive feedback.	Comments with the order of appearance are as follows: 1. In this paper, the authors propose the semi-supervised problem, but the contribution of unlabeled data is not clear. More details should be provided. 2. In this paper, the authors propose ensemble learning model. While in this way, the relevant labels have less effect on irrelevant labels, which seems unreasonable. 3. In this paper, the authors only provide the optimization of label similarity L, but how to determine the final L, only select top K elements? More details should be provided. 4. For the experiments, the compared algorithms should have more state-of-the-art algorithms.
[QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.	1. The contribution of unlabeled data is not clear. More details should be provided. 2. Explain why the ensemble learning model used. 3. More details should be provided for L. 4. Add more state-of-the-art algorithms, e.g. Semi-supervised Multi-label Learning with Incomplete Labels in IJCAI 2015.
[OVERALL SCORE]	Accept (Top 50% accepted papers (est.))
Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons below.	After reading the author's rebuttal and reviews from fellow reviewers, I would like to maintain my previous decision.

Review opinions

Masked Reviewer ID: Assigned_Reviewer_2
Review:

Question	
[Summary] Please summarize the main claims/contributions of the paper in your own words.	This paper focuses on multi-label learning. Specifically, the authors notice an important observation which is ignored in most previous multi-label works that the labels could be partially and completely missing simultaneously, called the semi-supervised weak-label learning. The authors demonstrate that treating them independently may harm the overall accuracy. To address this issue, this paper proposes a novel ensemble algorithm based on both label similarity and instance similarly and then building SSWL models based on the co-regularization strategy. Experiments on several benchmark datasets show that the SSWL accuracy is significantly improved based on the ensemble algorithm.
[Relevance] Is this paper relevant to an AI audience?	Likely to be of interest to a large proportion of the community
[Significance] Are the results significant?	Highly significant
[Novelty] Are the problems or approaches novel?	Very novel
[Soundness] Is the paper technically sound?	Technically sound
[Evaluation] Are claims well-supported by theoretical analysis or experimental results?	Sufficient
[Clarity] Is the paper well-organized and clearly written?	Good
[Detailed Comments] Please elaborate on your assessments and provide constructive feedback.	Generally, I am satisfied with this paper. It is an interesting paper whose observation is meaningful and solution is simple yet effective. Most multi-label papers focus on supervised learning or one type of missing labels. But this paper pays attention to two kinds of missing labels simultaneously, partially known or completely unknown. The demonstration in Figure 1 is convictive and impressive which clearly points out the difference between previous setting and the proposed new setup, and a novel method SSWL is proposed to ensemble two classifiers via the regularization from both instance and label similarities. This paper notices this problem and improves SSWL accuracy from the ensemble-learning perspective, which is very good. The experiments also show that ensemble learning is helpful in many application data sets for many multi-label models. I believe this work can inspire many other multi-label works focusing on other parts in an algorithm to cope with complicated missing label issue in multi-label learning.
[QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.	I have some questions which I wish the authors could address in the rebuttal. 1. Based on Eq. (2), the authors proposed a new regularization for both instance and label similarities. Since label similarity seems a little bit complicated when combined with subsequent SSWL model training, I am more interested that can the authors provide the results if only instance similarity is used? How much performance gain is achieved with only instance similarity? 2. Ensemble learning is an effective way to help improve the learning performance. To demonstrate the necessity of ensemble learning, a comprehensive evaluation should be given. What is the average performance by simply combining two initial leaners? 3. I notice that macro and micro F1s have more gain than hamming loss for the proposed method. Could the authors give some explanation? 4. The authors use an iterative algorithm to solve Eq. (3). How about the efficiency of the algorithm?
[OVERALL SCORE]	Accept (Top 15%)
Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons below.	Thanks for the feedback.

Review opinions

Masked Reviewer ID: Assigned_Reviewer_3
Review:

Question	
[Summary] Please summarize the main claims/contributions of the paper in your own words.	The paper presents a optimization problem which solves the problem of multi-label classification in the semi-supervised learning setting. Novelty of the proposed approach is that it works with weak-labels.
[Relevance] Is this paper relevant to an AI audience?	Likely to be of interest to a large proportion of the community
[Significance] Are the results significant?	Highly significant
[Novelty] Are the problems or approaches novel?	Novel
[Soundness] Is the paper technically sound?	Technically sound
[Evaluation] Are claims well-supported by theoretical analysis or experimental results?	Very convincing
[Clarity] Is the paper well-organized and clearly written?	Good
[Detailed Comments] Please elaborate on your assessments and provide constructive feedback.	The paper presents a optimization problem which solves the problem of multi-label classification in the semi-supervised learning setting. Novelty of the proposed approach is that it works with weak-labels. This is a quite interesting task and, in some sense, extends the task of learning with incompletely annotated istances. Despite some typos, the paper is well and clearly written. The part on related works is quite condensed and misses some recent references which adopt complete different approaches (not for weak-labels). This is an example: Jurica Levatic, Michelangelo Ceci, Dragi Kocev, Saso Dzeroski: Self-training for multi-target regression with tree ensembles. Knowl.-Based Syst. 123: 41-60 (2017) The method is quite standard and the authors clearly presented it. Theorem 1 is at the basis of the optimization solution the authors present. Probably, more details on this theorem shoud be given. Results are obtained with a correct experimental setting on several datasets coming from different domains. The only negative aspect that can be raised is that the number of used datasets is relatively small. Another limitation is that results should be integrated with statistical tests. However, from the results I see I'm pretty sure that statistical tests will confirm the superiority of SSWL. In my opinion, the paper is relatively solid, even is not extraordinarily novel from the methodological point of view: However, the extension to weak labels is enough to justify the paper. Typos (non-exhaustive list): We further compare with a variant of our proposal that does not usING the unlabeled instances".
[QUESTIONS FOR THE AUTHORS] Please provide questions for authors to address during the author feedback period.	See detailed comments
[OVERALL SCORE]	Accept (Top 15%)
Please acknowledge that you have read the author rebuttal. If your opinion has changed, please summarize the main reasons below.	I acknowledge I have read the author rebuttal. It did not change my opinion about the paper: I'm still positive about acceptance.