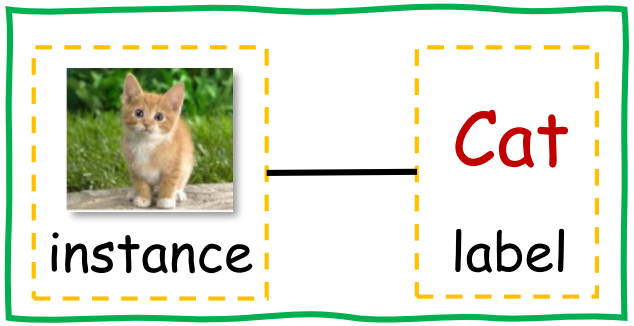
Provably Consistent Partial-Label Learning

Lei Feng[1*] **Jiaqi Lv**[2] **Bo Han**[3] **Miao Xu**[4,5]
Gang Niu[5] **Xin Geng**[2] **Bo An**[1†] **Masashi Sugiyama**[5,6]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]School of Computer Science and Engineering, Southeast University, Nanjing, China
[3]Department of Computer Science, Hong Kong Baptist University, China
[4]The University of Queensland, Australia
[5]Center for Advanced Intelligence Project, RIKEN, Japan
[6]The University of Tokyo, Japan

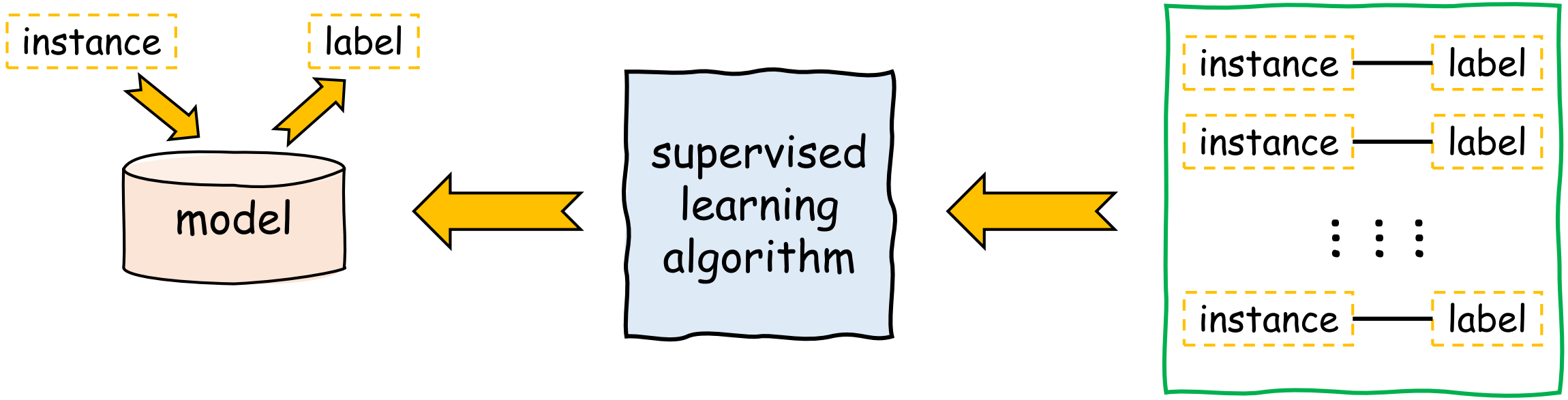
NeurIPS 2020

Ordinary Supervised Learning



Input space:
represented by a single instance (feature vector) characterizing its properties

Target space:
associated with a single label characterizing its semantics



Basic Assumption: Strong Supervision

Successful Learning needs: supervised information + regularities

Strong Supervision Assumption

- Sufficient labeling

  abundant labeled training data are available

- Accurate labeling

  object labeling is accurate

- Explicit labeling

  object labeling is unique and unambiguous

training error

hypothesis space complexity

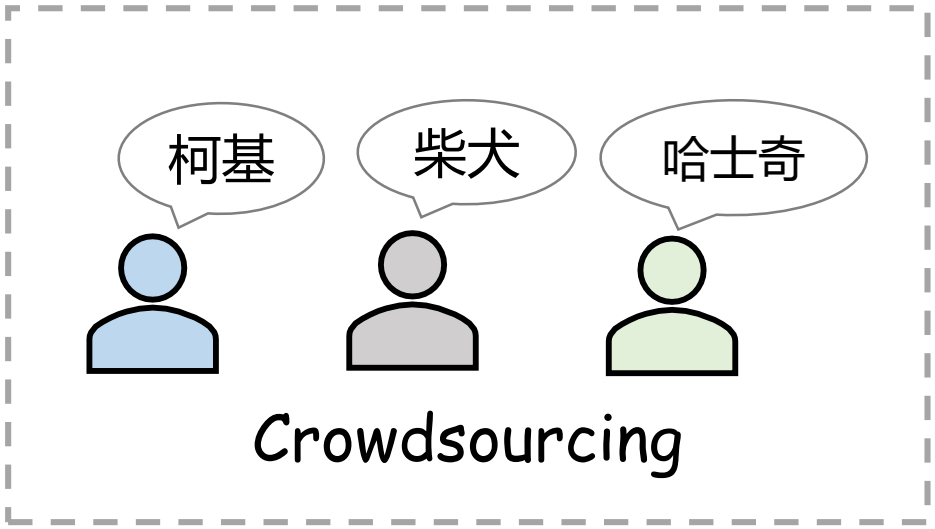
estimation error

approximation error

However, supervision is usually weak in practice

Partial Label Learning (PLL)

- Each object is associated with multiple candidate labels
- Only one candidate label is the unknown ground-truth label



$Y=\{$柯基，柴犬，哈士奇$\}$

Partial Label Learning: Review

PLL Methods

- Identification-based disambiguation: treat the ground-truth label as latent variable identified via iterative refining procedure
- Averaging-based disambiguation: treat all candidate labels in an equal manner and make final prediction by averaging
- Transformation strategy: binary decomposition, dictionary learning, graph matching, regression

EM

However, previous works never consider the generation process of the candidate sets

Notations

Learning with ordinary labels

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}), y)] \qquad \mathcal{Y} = [k]$$

Learning with partial labels

$$\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n} \qquad Y_i \in \mathcal{C} \qquad \mathcal{C} = \{2^{\mathcal{Y}} \setminus \emptyset \setminus \mathcal{Y}\}$$

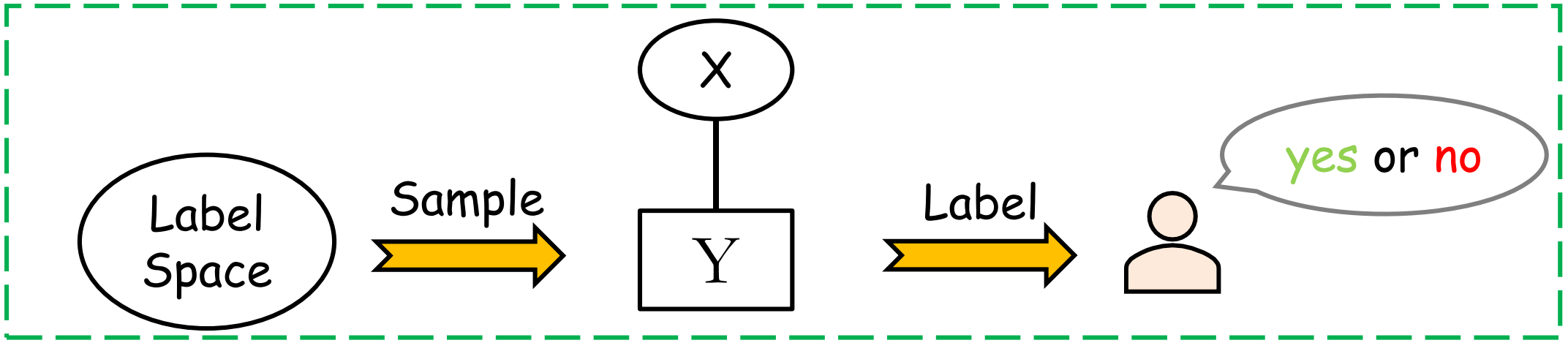
$$|\mathcal{C}| = 2^k - 2$$

$$p(y_i \in Y_i \mid \boldsymbol{x}_i, Y_i) = 1, \ \forall (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, \ \forall Y_i \in \mathcal{C}.$$

PLL: Data Generation Model

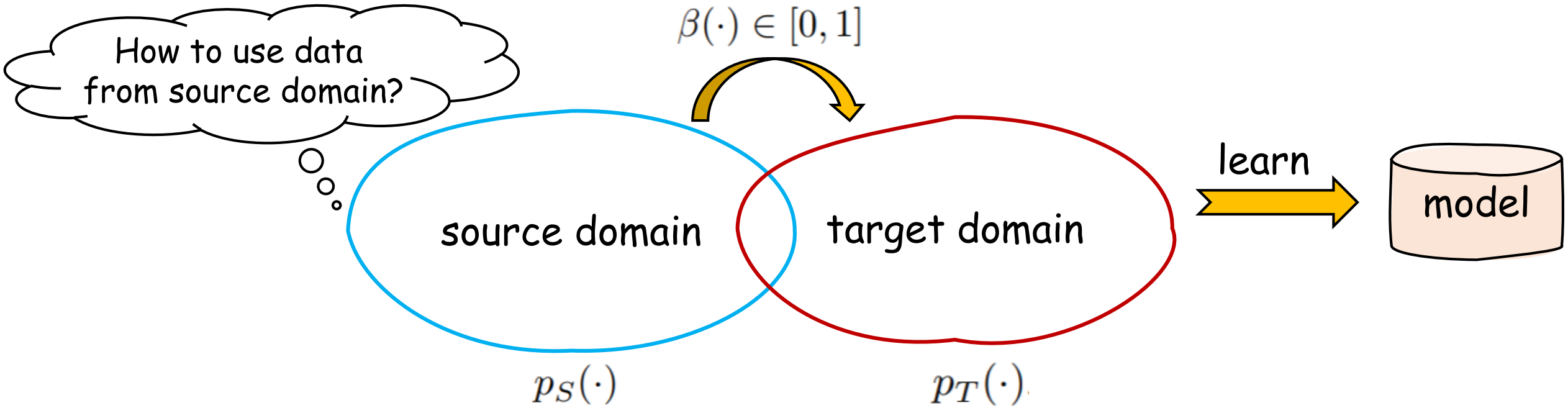
Partially labeled data distribution

$$\widetilde{p}(\boldsymbol{x}, Y) = \sum_{i=1}^{k} p(Y \mid y = i)\, p(\boldsymbol{x}, y = i) \qquad p(Y \mid y = i) = \begin{cases} \frac{1}{2^{k-1}-1} & \text{if } i \in Y, \\ 0 & \text{if } i \notin Y. \end{cases}$$



Lemma 1. *Given any instance $\boldsymbol{x}$ with its correct label y, for any unknown label set Y that is uniformly sampled from $\mathcal{C}$, the equality $p(y \in Y \mid \boldsymbol{x}) = 1/2$ holds.*

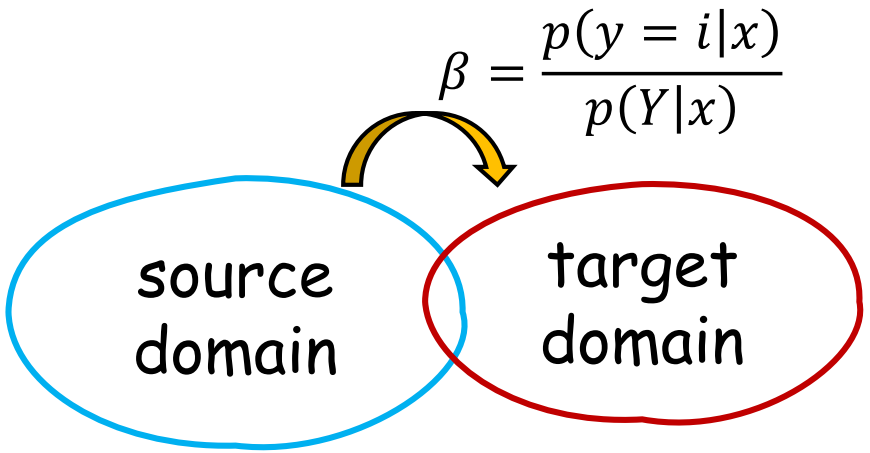
Preliminaries: Importance Reweighting

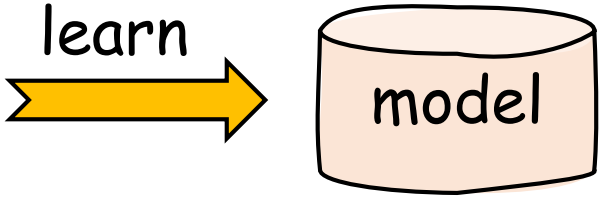


reweighted samples: $p_{\beta(S)}(x) = \frac{p_S(x)\beta(x)}{\int \beta(x)p_S(x)dx}$

minimize $\mathbb{D}\left(p_{\beta(S)}, p_T\right)$ $\Longrightarrow$ $\beta(x) \propto \frac{p_T(x)}{p_S(x)}$

Risk-Consistent Method





$$\frac{p(y=i,x)}{p(Y,x)} = \frac{p(y=i|x)p(x)}{p(Y|x)p(x)}$$

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[\mathcal{L}(f(\boldsymbol{x}),y)] = \int_{\boldsymbol{x}} \sum_{i=1}^{k} p(y=i \mid \boldsymbol{x}) \mathcal{L}(f(\boldsymbol{x}),i) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \int_{\boldsymbol{x}} \sum_{i=1}^{k} \frac{1}{|\mathcal{C}|} \sum_{Y \in \mathcal{C}} p(Y \mid \boldsymbol{x}) \frac{p(y=i|\boldsymbol{x})}{p(Y|\boldsymbol{x})} \mathcal{L}(f(\boldsymbol{x}),i) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$= \frac{1}{2^k-2} \mathbb{E}_{\widetilde{p}(\boldsymbol{x},Y)} \left[\sum_{i=1}^{k} \frac{p(y=i|\boldsymbol{x})}{p(Y|\boldsymbol{x})} \mathcal{L}(f(\boldsymbol{x}),i) \right] = R_{\mathrm{rc}}(f)$$

$$p(Y \mid \boldsymbol{x}) = \sum_{j=1}^{k} p(Y \mid y=j) p(y=j \mid \boldsymbol{x}) = \frac{1}{2^{k-1}-1} \sum_{j \in Y} p(y=j \mid \boldsymbol{x})$$

Risk-Consistent Method

Risk: $R_{\mathrm{rc}}(f) = \frac{1}{2}\mathbb{E}_{\widetilde{p}(\boldsymbol{x},Y)}\left[\sum_{i=1}^{k}\frac{p(y=i|\boldsymbol{x})}{\sum_{j\in Y}p(y=j|\boldsymbol{x})}\mathcal{L}\big(f(\boldsymbol{x}),i\big)\right]$ $\quad \{\boldsymbol{x}_o, Y_o\}_{o=1}^{n}$

Empirical Risk: $\widehat{R}_{\mathrm{rc}}(f) = \frac{1}{2n}\sum_{o=1}^{n}\left(\sum_{i=1}^{k}\frac{p(y_o=i|\boldsymbol{x}_o)}{\sum_{j\in Y_o}p(y_o=j|\boldsymbol{x}_o)}\mathcal{L}\big(f(\boldsymbol{x}_o),i\big)\right)$

Algorithm 1 RC Algorithm

Input: Model f, epoch $T_{\max}$, iteration $I_{\max}$, partially labeled training set $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n}$.

Initialize $p(y_i = j \mid \boldsymbol{x}_i) = 1, \forall j \in Y_i$, otherwise $p(y_i = j \mid \boldsymbol{x}_i) = 0$;
for $t = 1, 2, \ldots, T_{\max}$ **do**
  Shuffle $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n}$;
  for $j = 1, \ldots, I_{\max}$ **do**
    Fetch mini-batch $\widetilde{\mathcal{D}}_j$ from $\widetilde{\mathcal{D}}$;
    Update model f by $\widehat{R}_{\mathrm{rc}}$ in Eq. (9);
    Update $p(y_i \mid \boldsymbol{x}_i)$ by Eq. (10);
  end for
end for
Output: f.

M step: update parameters of the model

E step: estimate the label distribution

$p(y = i \mid \boldsymbol{x}) = g_i(\boldsymbol{x})$ if $i \in Y$, otherwise $p(y = i \mid \boldsymbol{x}) = 0$, $\forall (\boldsymbol{x}, Y) \sim \widetilde{p}(\boldsymbol{x}, Y)$.

Classifier-Consistent Method

$$q_j(\boldsymbol{x}) = p(Y = C_j \mid \boldsymbol{x})$$
$$C_j \in \mathcal{C}\ (j \in [2^k - 2])$$

→ Transition matrix Q

$$Q_{ij} = p(C_j \mid y = i)$$

← We want to know:

$$g_i(\boldsymbol{x}) = p(y = i \mid \boldsymbol{x})$$

$$\rightarrow \quad q(\boldsymbol{x}) = \boldsymbol{Q}^\top g(\boldsymbol{x}) \quad \rightarrow \quad \widehat{R}_{\mathrm{cc}}(f) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{2^{k-1}-1}\sum_{y \in Y_i} \frac{\exp(f_y(\boldsymbol{x}))}{\sum_j \exp(f_j(\boldsymbol{x}))}\right)$$

Algorithm 2 CC Algorithm

Input: Model f, epoch $T_{\max}$, iteration $I_{\max}$, partially labeled training set $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n}$;

1: **for** $t = 1, 2, \ldots, T_{\max}$ **do**
2: **Shuffle** the partially labeled training set $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n}$;
3: **for** $j = 1, \ldots, I_{\max}$ **do**
4: **Fetch** mini-batch $\widetilde{\mathcal{D}}_j$ from $\widetilde{\mathcal{D}}$;
5: **Update** model f by minimizing the empirical risk estimator $\widehat{R}_{\mathrm{cc}}$ in Eq. (12);
6: **end for**
7: **end for**

Output: f.

Experiments

Table 3: Test performance (mean±std) of each method using linear model on UCI datasets.

	Texture	Yeast	Dermatology	Har	20Newsgroups
RC	99.24±0.14%	59.89±1.27%	99.41±1.00%	98.03±0.09%	**75.99±0.53%**
CC	98.02±2.91%•	**59.97±1.57%**	**99.73±0.85%**	**98.10±0.18%**	75.97±0.54%
SURE	95.38±0.28%•	54.39±1.32%•	97.48±0.32%•	97.43±0.24%•	69.82±0.26%•
CLPL	91.93±0.97%•	54.58±2.11%•	99.62±0.85%	97.48±0.18%•	71.44±0.55%•
PLECOC	69.69±4.82%•	37.37±9.73%•	87.84±5.30%•	96.97±0.29%•	15.32±7.86%•
PLSVM	49.38±9.99%•	45.70±8.01%•	80.00±7.53%•	91.64±1.43%•	32.59±8.91%•
PLkNN	96.78±0.31%•	47.79±2.41%•	80.54±5.06%•	94.17±0.59%•	27.18±0.65%•
IPAL	**99.45±0.23%**	48.99±3.84%•	98.65±2.27%•	96.55±0.40%•	48.36±0.85%•

Table 4: Test performance (mean±std) of each method using linear model on real-world datasets.

	Lost	MSRCv2	BirdSong	Soccer Player	Yahoo! News
RC	**79.43±3.26%**	46.56±2.71%	71.94±1.72%	**57.00±0.97%**	**68.23±0.83%**
CC	79.29±3.19%	47.22±3.02%	**72.22±1.71%**	56.32±0.64%	68.14±0.81%
SURE	71.33±3.57%•	46.88±4.67%	58.92±1.28%•	49.41±086%•	45.49±1.15%•
CLPL	74.87±4.30%•	36.53±4.59%•	63.56±1.40%•	36.82±1.04%•	46.21±0.90%•
PLECOC	49.03±8.36%•	41.53±3.25%•	71.58±1.81%	53.70±2.02%•	66.22±1.01%•
PLSVM	75.31±3.81%•	35.85±4.41%•	49.90±2.07%•	46.29±0.96%•	56.85±0.91%•
PLkNN	36.73±2.99%•	41.36±2.89%•	64.94±1.42%•	49.62±0.67%•	41.07±1.02%•
IPAL	72.12±4.48%•	**50.80±4.46%**◦	72.06±1.55%	55.03±0.77%•	66.79±1.22%•

Thanks