Revisiting Pseudo-Label for Single-Positive Multi-Label Learning

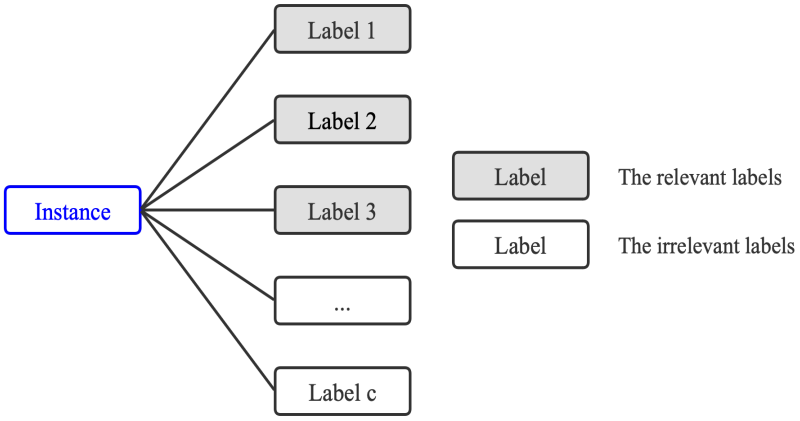
Biao Liu[1], Ning Xu[1*], Jiaqi Lv[2], Xin Geng[1*]

1 School of Computer Science and Engineering, Southeast University

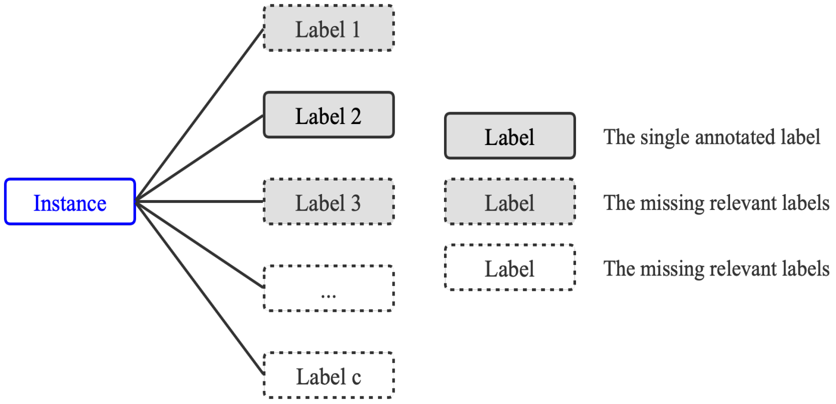
2 RIKEN Center for Advanced Intelligence Project

Single-Positive Multi-Label Learning

Multi-Label Learning



Single-Positive Multi-Label Learning



SPMLL问题中伪标记生成的可学习性

Small Unreliability Degree Condition

Intuitive explanation: From the perspective of pseudo-label generation, for any pseudo-label of an instance, it must be not always be mislabeled.

Theorem 1: Suppose a SPMLL pseudo-label-based method has an unreliability degree of pseudo-label $\xi = \sup_{\substack{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{l})\sim p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{l}) \\ j\in\{1,2,\dots,c\}}} \Pr(l^j \neq y^j)$, $0 \leq \xi \leq 1$. Let $\theta_1 = c\log\frac{2}{1+\xi}$, and suppose the Natarajan dimension of the hypothesis $\mathcal{H}$ is $d_{\mathcal{H}}$, define

$$n_0(\mathcal{H},\epsilon,\delta) = \frac{4}{\theta_1\epsilon}\left(d_{\mathcal{H}}\left(\log(4d_{\mathcal{H}}) + 2c\log c + \log\frac{1}{\theta_1\epsilon}\right) + \log\frac{1}{\delta} + 1\right).$$

Then when $n > n_0(\mathcal{H},\epsilon,\delta)$, $\mathcal{R}_{ham}\left(\mathcal{A}(\overline{\mathcal{D}})\right) < \epsilon$ with probability $1-\delta$.

SPMLL问题中伪标记生成的可学习性

Non-Zero Minimum Positive Label Sampling Probability Condition:

Intuitive explanation: From the perspective of data generation, every relevant label of each instance can possibly be sampled as the single label.

Theorem 1: Suppose a SPMLL pseudo-label-based method has an unreliability degree of pseudo-label $\tau = \inf_{\substack{(\boldsymbol{x},\boldsymbol{y},\gamma)\sim p(\boldsymbol{x},\boldsymbol{y},\gamma) \\ y_j=1, j\in\{1,2,\ldots,c\}}} \Pr(j \neq \gamma)$, $\tau > 0$. Let $\theta_2 = c\log\frac{2}{2-\tau}$, and suppose the Natarajan dimension of the hypothesis $\mathcal{H}$ is $d_{\mathcal{H}}$, define

$$n_0(\mathcal{H},\epsilon,\delta) = \frac{4}{\theta_2\epsilon}\left(d_{\mathcal{H}}\left(\log(4d_{\mathcal{H}}) + 2c\log c + \log\frac{1}{\theta_2\epsilon}\right) + \log\frac{1}{\delta} + 1\right).$$

Then when $n > n_0(\mathcal{H},\epsilon,\delta)$, $\mathcal{R}_{ham}\left(\mathcal{A}(\bar{\mathcal{D}})\right) < \epsilon$ with probability $1-\delta$.

The MIME Approach

The information-based objective function:

$$\mathcal{L}_{IB} = \sum_{j=1}^{c} I(\boldsymbol{z}^j, y^j) - \beta_j I(\boldsymbol{z}^j, \boldsymbol{x})$$
$$\geq \sum_{j=1}^{c} I(\boldsymbol{z}^j, l^j) - \beta_j I(\boldsymbol{z}^j, \boldsymbol{x})$$

Expressive about the label ↑ ↑ Compressive about $\boldsymbol{x}$

Variational Bayes Techniques:

$$\mathcal{L} \approx \frac{1}{nc} \sum_{i=1}^{n} \sum_{j=1}^{c} \mathbb{E}_{\epsilon \sim p(\epsilon)} \left[-\log q_{\boldsymbol{\phi}} \left(l_i^j \mid f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \epsilon) \right) \right] + \beta_j \mathrm{KL}\left[p(\boldsymbol{z}^j \mid \boldsymbol{x}_i), r(\boldsymbol{z}^j) \right]$$

Experimental Results

Table 1: Predictive performance of each comparing methods on four MLIC datasets in terms of *mean average precision (mAP)* (mean ± std). The best performance is highlighted in bold (the larger the better).

	VOC	COCO	NUS	CUB
AN	85.546±0.294	64.326±0.204	42.494±0.338	18.656±0.090
AN-LS	87.548±0.137	67.074±0.196	43.616±0.342	16.446±0.269
WAN	87.138±0.240	65.552±0.171	45.785±0.192	14.622±1.300
EPR	85.228±0.444	63.604±0.249	45.240±0.338	19.842±0.423
ROLE	88.088±0.167	67.022±0.141	41.949±0.205	14.798±0.613
EM	88.674±0.077	70.636±0.094	47.254±0.297	20.692±0.527
EM-APL	88.860±0.080	70.758±0.215	47.778±0.181	21.202±0.792
SMILE	86.311±0.450	63.331±0.112	43.611±0.172	18.611±0.144
LAGC	88.021±0.121	70.422±0.062	46.211±0.155	21.840±0.237
MIME	**89.199±0.157**	**72.920±0.255**	**48.743±0.428**	**21.890±0.347**

THANK YOU

liubiao01@seu.edu.cn