One Positive Label is Sufficient: Single-Positive Multi-Label Learning with Label Enhancement

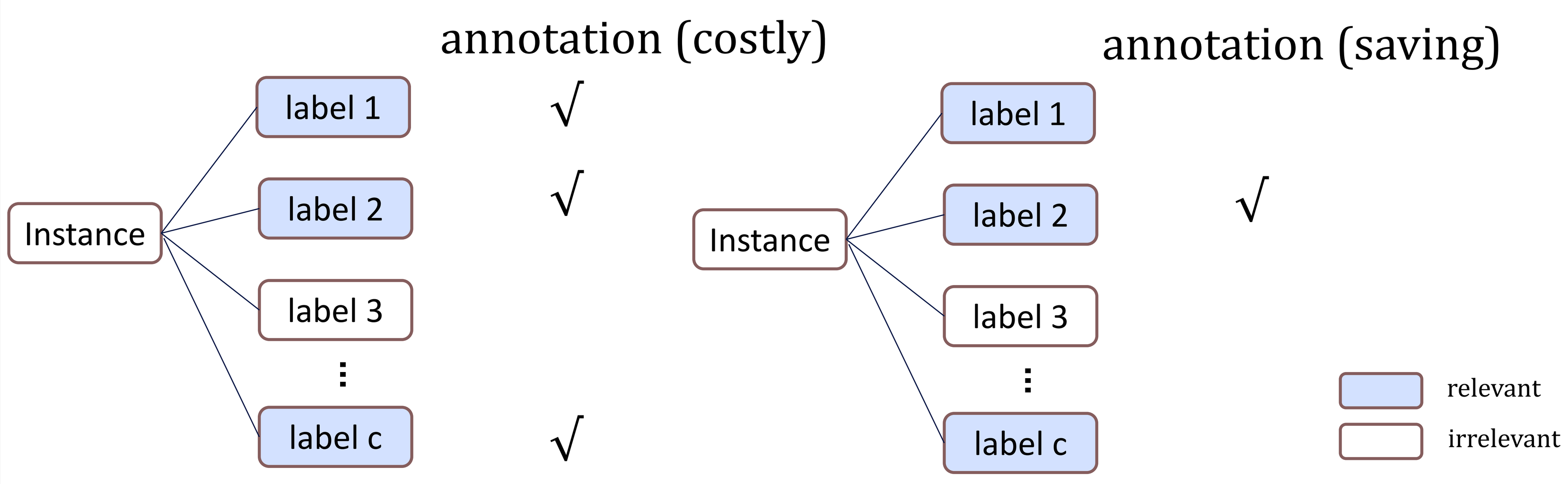
Ning Xu[1], Congyu Qiao[1], Jiaqi Lv[2], Xin Geng[1] and *Min-Ling Zhang[1]*

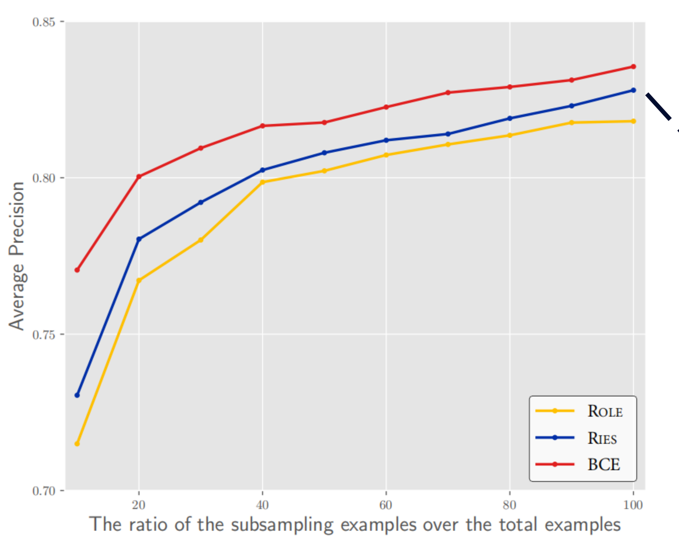
[1] Southeast University; [2] RIKEN Center for Advanced Intelligence Project

{xning, qiaocy, xgeng, zhangml}@seu.edu.cn; is.jiaqi.lv@gmail.com

NeurIPS 2022

Introduction





Comparing with fully labeled case, the SPMLL approaches on single-positive labeled examples only incur a tolerable drop in the performance but drastically reduce the amount of supervision required to train multi-label classifiers.

Our work:

- **Theoretically, we for the first time derive an unbiased risk estimator for SPMLL. Based on this, an estimation error bound is established that guarantees the risk-consistency.**
- **Practically, we propose the method SMILE for SPMLL via adopting the latent soft labels recovered by label enhancement.**

The Proposed Approach

Multi-label Learning → Single Positive Multi-label Learning

Unbiased Esitmator

MLL Expected risk

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

set of relevant labels

SPMLL Expected risk

$$R_{sp}(f)=\mathbb{E}_{p(\boldsymbol{x},\gamma)}\left[\frac{1}{p(y^{\gamma}=1|\boldsymbol{x})c}\sum_{Y\in\mathcal{C}}\mathcal{L}\left(f\left(\boldsymbol{x}\right),Y\right)p(Y|\boldsymbol{x})\right]$$

observed single-positive label

Binary cross-entropy loss

Empirical Risk Estimator

$$\widehat{R}_{sp}(f)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{p(y^{\gamma_i}=1|\boldsymbol{x}_i)c}\sum_{j=1}^{c}d_i^j\ell_i^j+\left(1-d_i^j\right)\bar{\ell}_i^j\right)$$

$\log f_j(\boldsymbol{x})$

soft label, recovered via the label enhancement process

$p(y^j=1|\boldsymbol{x})$

$\log\left(1-f_j(\boldsymbol{x})\right)$

Estimation Error Bound

Theorem 1 *Assume the loss function $\ell\left(f(\boldsymbol{x}),y\right)$ and $\bar{\ell}\left(f(\boldsymbol{x}),y\right)$ are ρ^+-Lipschitz and ρ^--Lipschitz with respect to $f(\boldsymbol{x})$ $(0<\rho^+<\infty$ and $0<\rho^-<\infty)$ for all $y\in\mathcal{Y}$ and the loss function $\mathcal{L}_{sp}$ are bounded by M, i.e., $M=\sup_{\boldsymbol{x}\in\mathcal{X},f\in\mathcal{F},y\in\mathcal{Y}}\mathcal{L}_{sp}\left(f(\boldsymbol{x}),y\right)$, with probability at least $1-\delta$,*

$$R(\widehat{f}_{sp})-R(f^*)\leq 4\sqrt{2}\kappa c(\rho^++\rho^-)\sum_{j=1}^{c}\mathfrak{R}_n(\mathcal{H}_y)+M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Label Enhancement

Aims to infer $q_{\boldsymbol{w}}(\mathbf{D},\mathbf{Z}|\mathbf{L},\mathbf{X},\mathbf{A})$ $\xrightarrow{\text{Approximate}}$ $p(\mathbf{D},\mathbf{Z}|\mathbf{L},\mathbf{X},\mathbf{A})$

decomposed into

$$q_{w_1}(\mathbf{D}\mid\mathbf{L},\mathbf{X},\mathbf{A})=\prod_{i=1}^{n}\prod_{j=1}^{c}\mathrm{Beta}\left(d_i^j|\alpha_i^j,\beta_i^j\right)$$

The parameters $\Delta=[\alpha_1,\alpha_2,\ldots,\alpha_n]$ and $\Phi=[\beta_1,\beta_2,\ldots,\beta_n]$are outputs of the inference model parameterized by ω_1, which is defined as a GCN.

$$q_{w_2}(\mathbf{Z}|\mathbf{D},\mathbf{X})=\prod_{i=1}^{n}Gau(\boldsymbol{z}_i|\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i)$$

The parameters $\Delta=[\mu_1,\mu_2,\ldots,\mu_n,\sigma_1,\sigma_2,\ldots,\sigma_n]$are outputs of the inference model parameterized by ω_2, which is defined as a MLP.

Variational Bayes Techniques

$$\mathcal{L}_{ELBO}=\mathbb{E}_{q_{\boldsymbol{w}}(\mathbf{D},\mathbf{Z}|\mathbf{L},\mathbf{X},\mathbf{A})}[\log p(\mathbf{X}|\mathbf{D},\mathbf{Z})+\log p(\mathbf{L}|\mathbf{D})+\log p(\mathbf{A}|\mathbf{D})]$$
$$-\mathrm{KL}[q_{\boldsymbol{w}_1}(\mathbf{D}|\mathbf{L},\mathbf{X},\mathbf{A})||p(\mathbf{D})]-\mathrm{KL}[q_{\boldsymbol{w}_2}(\mathbf{Z}|\mathbf{D},\mathbf{X})||p(\mathbf{Z})].$$

Compatibility Loss

$$T_C=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}l_i^j\log d_i^j+\left(1-l_i^j\right)\left(1-\log d_i^j\right)$$

promote the label enhancement process via enforcing that the estimated soft label should inherit the labeling-information of observed labels

Optimization problem

$$T_{LE}=-\lambda\mathcal{L}_{ELBO}+T_C$$

Experiments

Datasets

- Twelve widely used MLL datasets and five datasets, where we generate the single positive training data by randomly selecting one positive label to keep for each training example.
- We run the comparing methods with 80%/10%/10% train/validation/test split.

Dataset	$\lvert\mathcal{S}\rvert$	dim($\mathcal{S}$)	$L(\mathcal{S})$	Domain
CAL500	502	68	174	Music
image	2000	294	5	Images
scene	2407	294	6	Images
yeast	2417	103	14	Biology
corel5k	5000	499	374	Images
rcv1-s1	6000	944	101	Text
corel16k-s1	13766	500	153	Images
delicious	16105	500	983	Text
iaprtc12	19627	1000	291	Images
espgame	20770	1000	268	Images
mirflickr	25000	1000	38	Images
tmc2007	28596	981	22	Text

Evaluation Metrics

- *Ranking loss* ↓
- *Hamming loss* ↓
- *One-error* ↓
- *Coverage* ↓
- *Average precision* ↑

Experimental Results

Baselines

SMILE vs. SPMLL approaches

SMILE vs. MLL with missing label approaches

Datasets	SMILE	AN	AN-LS	WAN	ROLE	GLOCAL	MLML	D2ML
CAL500	**0.401±0.011**	0.382±0.044	0.253±0.031	0.393±0.011	0.288±0.008	0.227±0.002	0.233±0.000	0.223±0.001
image	**0.784±0.044**	0.613±0.081	0.621±0.073	0.685±0.058	0.696±0.039	0.771±0.003	0.652±0.001	0.274±0.003
scene	**0.841±0.070**	0.740±0.127	0.741±0.117	0.801±0.020	0.717±0.067	0.825±0.001	0.814±0.000	0.285±0.002
yeast	**0.758±0.003**	0.755±0.003	0.753±0.003	0.757±0.003	0.753±0.003	0.646±0.002	0.456±0.002	0.323±0.001
corel5k	**0.303±0.007**	0.299±0.005	0.272±0.005	0.302±0.004	0.215±0.011	0.218±0.001	0.072±0.001	0.028±0.001
rcv1-s1	**0.616±0.001**	0.604±0.004	0.581±0.002	0.610±0.005	0.570±0.004	0.229±0.000	0.221±0.003	0.053±0.001
corel16k-s1	**0.344±0.003**	0.337±0.003	0.316±0.002	0.344±0.003	0.288±0.004	0.029±0.001	0.081±0.001	0.029±0.004
delicious	0.319±0.001	0.297±0.009	0.193±0.005	**0.320±0.001**	0.199±0.004	0.027±0.001	0.086±0.001	0.028±0.001
iaprtc12	**0.314±0.003**	0.292±0.008	0.244±0.008	0.266±0.006	0.243±0.005	0.035±0.002	0.126±0.001	0.026±0.001
espgame	**0.259±0.003**	0.248±0.002	0.208±0.003	0.259±0.002	0.216±0.004	0.038±0.000	0.086±0.002	0.038±0.001
mirflickr	**0.635±0.004**	0.629±0.003	0.604±0.004	0.611±0.004	0.545±0.019	0.476±0.000	0.253±0.003	0.132±0.002
tmc2007	**0.820±0.002**	0.815±0.003	0.802±0.003	0.815±0.001	0.798±0.005	0.649±0.000	0.415±0.000	0.161±0.001

Table 1: Predictive performance of each comparing approach (mean±std) in terms of *Average precision* ↑. The best performance (the larger the better) is shown in bold face.

SMILE against	AN	AN-LS	WAN	ROLE	GLOCAL	MLML	D2ML
Average precision	**win**[0.0005]	**win**[0.0005]	**win**[0.0092]	**win**[0.0005]	**win**[0.0005]	**win**[0.0005]	**win**[0.0005]
One-error	**win**[0.0122]	**win**[0.0005]	**win**[0.0015]	**win**[0.0005]	**win**[0.0005]	**win**[0.0342]	**win**[0.0005]
Ranking loss	**win**[0.0269]	**win**[0.0005]	**tie**[0.1533]	**win**[0.0005]	**win**[0.0005]	**win**[0.0024]	**win**[0.0005]
Hamming loss	**win**[0.0277]	**win**[0.0178]	**win**[0.0005]	**win**[0.0277]	**win**[0.0277]	**win**[0.0277]	**win**[0.0077]
Coverage	**win**[0.0425]	**win**[0.0005]	**tie**[0.1819]	**win**[0.0005]	**win**[0.0005]	**win**[0.0024]	**win**[0.0015]

Wilcoxon signed-ranks test at 0.05 significance level

SMILE achieves superior performance against all the comparing approaches on all evaluation metrics (except on Ranking loss and Coverage where SMILE achieves comparable performance against WAN), which provides a strong evidence for the effectiveness of risk-consistent estimator for SPMLL.

Conclusion

Conclusion

- We study single-positive multi-label learning and propose a novel approach SMILE.
- We derive an unbiased risk estimator, which suggests that one positive label of each instance is sufficient to train predictive models for multi-label learning.
- We design a benchmark solution via estimating the soft label corresponding to each example in a label enhancement process. The effectiveness of the proposed method is validated on twelve corrupted MLL datasets.

More information

https://github.com/palm-ml/smile
http://palm.seu.edu.cn/