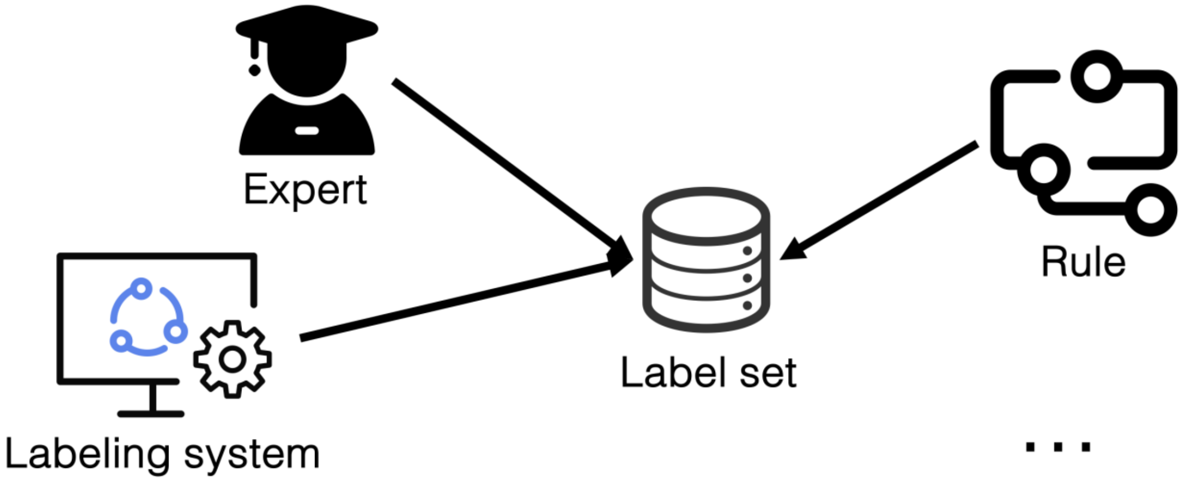
Collaborative Refining for Learning from Inaccurate Labels

Bin Han, Yi-Xuan Sun, Ya-Lin Zhang, Libang Zhang, Haoran Hu
Longfei Li, Jun Zhou [†], Guo Ye, Huimei He

Background



In industry, obtaining accurate labels can often be costly and time-consuming. Instead, **noisy labels from multiple sources** are more convenient to collect.

Overview

The Collaborative Refining for Learning from Inaccurate Labels (CRL) framework operates in two main steps:

- **Partitioning the Dataset.**
- **Refining Labels and Samples**:
  - For samples with disagreement, we propose a method called **Label Refining for Disagreements (LRD) to get reliable labels**
  - For samples where annotators agree, we apply **Robust Union Selection (RUS)** to select the most trustworthy samples based on theoretical bounds.

CRL: Overall Framework

Dataset

x		$\tilde{y}^1$	$\tilde{y}^2$	$\tilde{y}^3$
0.8	0.5	0	1	1
0.3	1.0	0	0	0
0.1	0.3	0	0	1
0.4	0.2	1	0	1
0.9	0.8	1	1	1

Training Data D

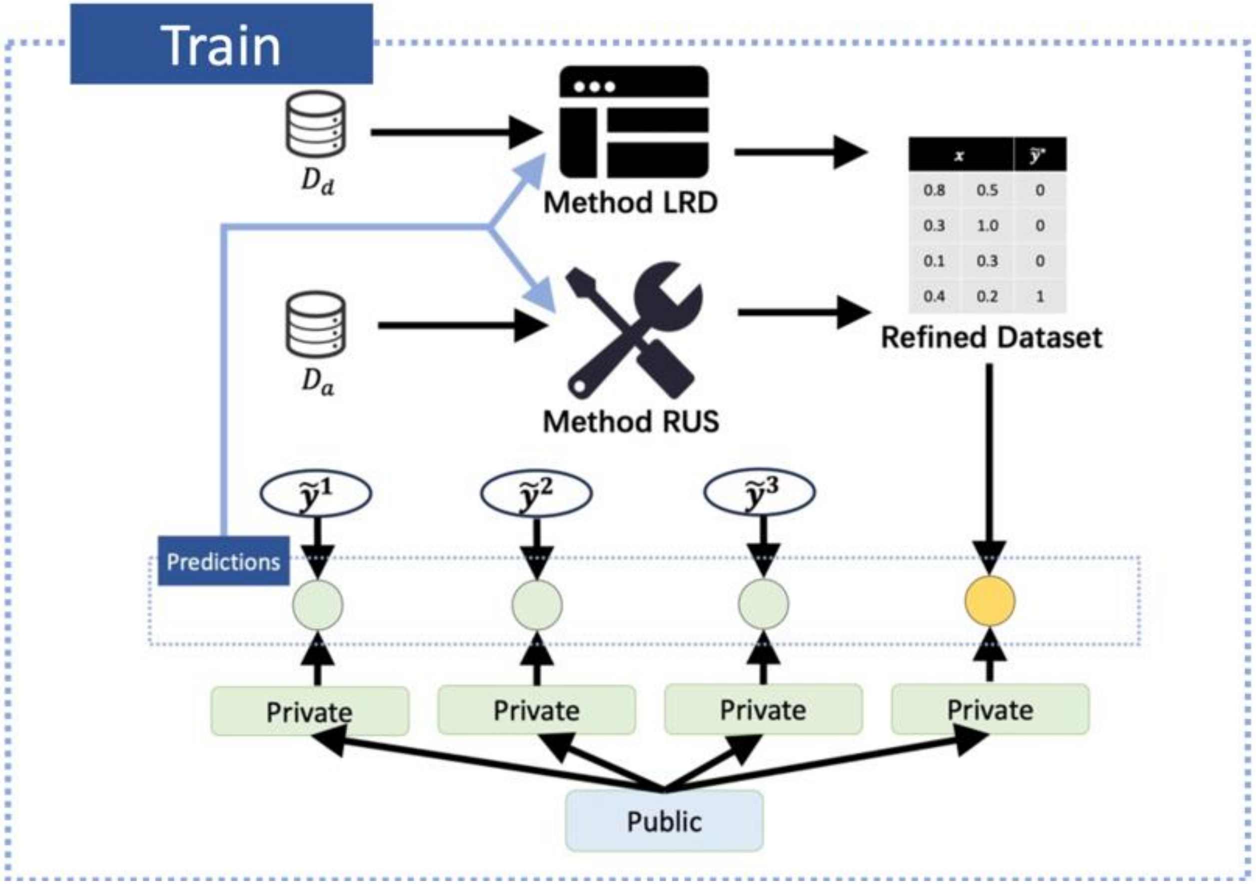
x		$\tilde{y}^1$	$\tilde{y}^2$	$\tilde{y}^3$
0.3	1.0	0	1	1
0.1	0.3	0	0	1
0.4	0.2	1	0	1

Samples with disagreements D_d

x		$\tilde{y}^1$	$\tilde{y}^2$	$\tilde{y}^3$
0.8	0.5	0	0	0
0.9	0.8	1	1	1

Samples with agreements D_a

Train



LRD: Label Refining for Samples with Disagreements

Theorem 1. *Let* $(\boldsymbol{x}, y^*, \tilde{y}^0, \tilde{y}^1)$ *be any sample with ground-truth label* y^* *and two conflicting labels* $\tilde{y}^0$, $\tilde{y}^1$ *from two annotators, i.e.,* $\tilde{y}^0 \neq \tilde{y}^1$. *Assume* T^0 *and* T^1 *satisfy* $T^0_{ii} > 0.5$ *and* $T^1_{ii} > 0.5$, $\forall i \in \{0, 1\}$, $\ell(f_{\Theta_0^*}(\boldsymbol{x}), \tilde{y}^0) < \ell(f_{\Theta_1^*}(\boldsymbol{x}), \tilde{y}^1)$ *if and only if* $y^* = \tilde{y}^0$.

Corollary 1. *Let* $(\boldsymbol{x}, y^*, \{\tilde{y}^r\}_{r=1}^R)$ *be any sample with ground-truth label* y^* *and* R *conflicting labels* $\{\tilde{y}^r\}_{r=1}^R$ *from* R *annotators, i.e.,* $\exists r_0, r_1 \subseteq \{1, ..., R\}$, $\tilde{y}^{r_0} \neq \tilde{y}^{r_1}$. *Assume* $T^r_{ii} > 0.5$, $\forall i \in \{0, 1\}$ *and* $r \in \{1, ..., R\}$, *if* $\ell(f_{\Theta_k^*}(\boldsymbol{x}), \tilde{y}^k) = \min(\{\ell(f_{\Theta_r^*}(\boldsymbol{x}), \tilde{y}^r)\}_{r=1}^R)$, $y^* = \tilde{y}^k$.

- The most reliable index k for sample x_i is acquired through:

$$k = \operatorname*{argmin}_r \{\ell(f_{\Theta_r}(\boldsymbol{x}_i), \tilde{y}_i^r)\}_{r=1}^R$$

RUS: Robust Union Selection

- The average loss

$$\tilde{\mu}_i = \frac{1}{R}\sum_{r=1}^{R} \ell(f_{\Theta_r}(\boldsymbol{x}_i), \tilde{y}_i^r)$$

- Smooth function

$$\phi(z) = \log(1 + z + \frac{z^2}{2})$$

- Selection criterion for sample x_i

$$\tilde{\mu}_i^{\phi} = \frac{1}{R|T|}\sum_{r=1}^{R}\sum_{t \in T} \phi(\ell(f_{\Theta_r^t}(\boldsymbol{x}_i), \tilde{y}_i^r))$$

$$c_i = \tilde{\mu}_i^{\phi} - \frac{\tilde{\sigma}_i^2 (n + \frac{\tilde{\sigma}_i^2 \log(2n)}{n^2})}{n - \tilde{\sigma}_i^2}$$

Experiments

Table 1: Main results with AUC as the evaluation metric. The best results are in bold.

Noise	Dataset	Methods												
		Single	NN-Mjv	HE_A	HE_M	CL	DN	NN-EBCC	NN-IBCC	WeaSEL	SLF	CoNAL	ADMoE	Ours
Class	AgNews	0.660	0.634	0.723	0.724	0.626	0.757	0.703	0.746	0.833	0.791	0.769	0.762	**0.855**
	20News	0.746	0.684	0.755	0.756	0.749	0.729	0.796	0.779	0.824	0.768	0.778	0.765	**0.849**
	IMDb	0.666	0.602	0.667	0.670	0.614	0.689	0.673	0.699	0.709	0.700	0.702	0.707	**0.766**
	Yelp	0.725	0.713	0.783	0.785	0.779	0.779	0.782	0.786	0.805	0.807	0.769	0.799	**0.867**
	Amazon	0.586	0.567	0.681	0.687	0.631	0.661	0.635	0.657	0.718	0.679	0.672	0.652	**0.775**
	Diabetes	0.648	0.610	0.576	0.592	0.633	0.686	0.657	0.663	0.680	0.577	0.696	0.646	**0.728**
	Backdoor	0.530	0.535	0.681	0.687	0.651	0.765	0.668	0.771	0.640	0.816	0.716	0.814	**0.937**
	Campaign	0.561	0.558	0.628	0.636	0.574	0.663	0.619	0.632	0.629	0.694	0.697	0.680	**0.783**
	Waveform	0.772	0.744	0.660	0.663	0.792	0.770	0.788	0.802	**0.840**	0.807	0.818	0.823	0.840
	Celeba	0.738	0.710	0.723	0.725	0.784	0.849	0.758	0.768	0.782	0.851	0.859	0.824	**0.891**
	SVHN	0.637	0.639	0.671	0.671	0.692	0.701	0.676	0.679	0.671	0.730	0.726	0.718	**0.761**
	F-MNIST	0.607	0.590	0.684	0.691	0.682	0.667	0.664	0.631	0.678	0.723	0.705	0.737	**0.776**
	CIFAR-10	0.541	0.573	0.574	0.576	0.596	0.570	0.590	0.590	0.587	0.587	0.634	0.625	**0.655**

Table 2: Real-world results with AUC as the evaluation metric. The best results are in bold.

Dataset	Methods												
	Single	NN-Mjv	HE_A	HE_M	CL	DN	NN-EBCC	NN-IBCC	WeaSEL	SLF	CoNAL	ADMoE	Ours
Sentiment	0.712	0.727	0.744	0.730	0.724	0.732	0.728	0.736	0.730	0.686	0.741	0.722	**0.753**
CIFAR-10N	0.791	0.853	0.788	0.786	0.788	0.807	0.850	0.849	0.851	0.821	0.816	0.761	**0.866**

Experiments

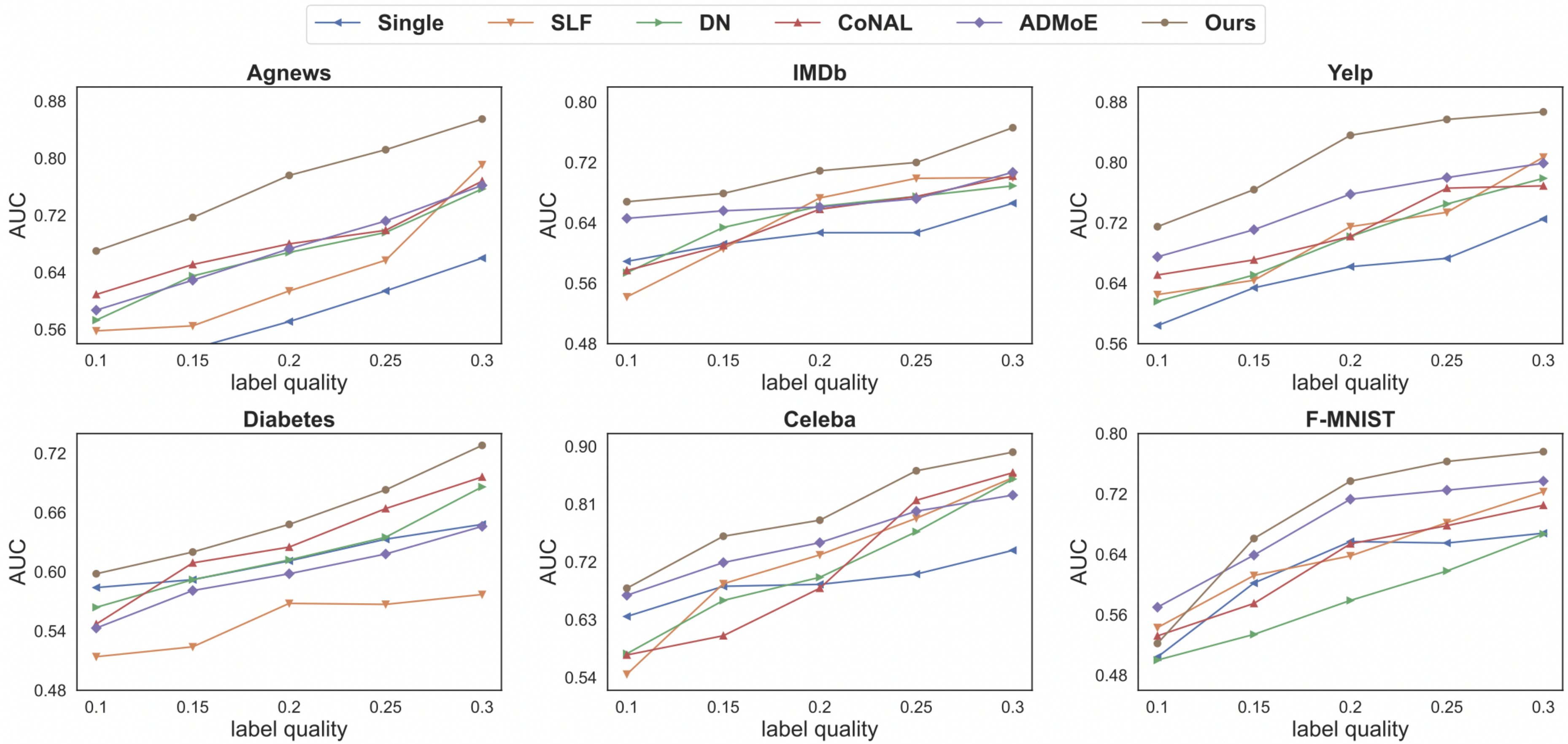


Figure 1: AUC comparison under different label quality k.

Thanks for watching!