On Non-Random Missing Labels In SSL [ICLR 2022]

▷ SSL is a missing label problem.

label { Missing Not At Random (MNAR) ★ → labeled & unlabeled have different distribution
Missing Completely At Random (MCAR)

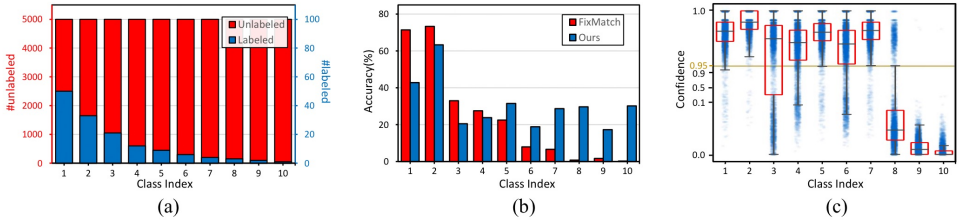


Figure 1: Visualization of an MNAR example and its experimental results on CIFAR-10. (a) Class distribution of the labeled and unlabeled training data. (b) Test accuracy of the supervised model using FixMatch and our CAP (Section 4.1). (c) The distribution of FixMatch's confidence scores on unlabeled data. Samples with confidence larger than a fixed threshold (0.95, yellow line) are imputed. The corresponding box-plots display the (minimum, first quartile, median, third quartile, maximum) summary. The confidence-axis is not equidistant scaling for better visualization.

FixMatch tends to impute the labels of more samples from the popular classes.

▷ Missing Labels in SSL

$$D \begin{cases} D_L = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_L} \\ D_U = \{(x^{(i)})\}_{i=N_L+1}^{N} \end{cases}$$

label missing indicator $M = \{m^{(i)}\}_{i=1}^{N}$, $m^{(i)}=1 \rightarrow$ unlabeled

$$D = (X, Y, M) = \{(x^{(i)}, y^{(i)}, m^{(i)})\}_{i=1}^{N}$$

① Traditional SSL is MCAR, that is, M is independent with Y.

$$P(Y|X=x, M=0) = P(Y|X=x)$$

$$\mathbb{E}[\hat{y}] = \mathbb{E}[y|\hat{\theta}] = \sum_{(x,y)\in D_L} y \cdot P(y|x) = \sum_{(x,y)\in D} y \cdot P(y|x, M=0)$$

$$= \sum_{(x,y)\in D} y \cdot P(y|x) = \mathbb{E}[y], \quad \text{unbiased}$$

(2) M is dependent with Y. $P(Y|X=x, M=0) \neq P(Y|X=x)$

$$\mathbb{E}[\hat{y}] = \sum_{(x,y)\in D} y \cdot P(y|x, M=0) = \sum_{(x,y)\in D} y \cdot P(y|x) \cdot \frac{P(M=0|x,y)}{P(M=0|x)}$$
$$\neq \sum_{(x,y)\in D} y \cdot P(y|x) = \mathbb{E}[y].$$

biased

We overlook the role of "class" Y in causing the non-randomness of M.

⊢ Class-Aware Doubly Robust Framework

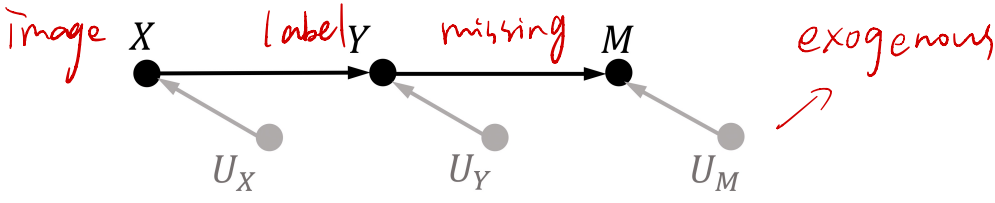
(1) Class-Aware Propensity for Labeled Data

Traditional SSL methods estimate the model parameters via maximum likelihood estimation on the labeled data:

$$\hat{\theta} = \arg\max_{\theta} \log P(Y|X, M=0; \theta) = \arg\max_{\theta} \sum_{(x,y)\in D_L} \log P(y|x;\theta) \tag{4}$$

$\hat{\theta}$ is estimated over D_L rather than the entire D (difference)

Inspired by causal inference, we have:



$P(X|Y, M=0) = P(X|Y)$

In this way, we can obtain the unbiased $\hat{\theta}$ on the labeled subset of data by maximizing $P(X|Y)$:

$$\hat{\theta} = \arg\max_{\theta} \log P(X|Y;\theta) = \arg\max_{\theta} \sum\nolimits_{(x,y)\in D_L} \log P(x|y;\theta) \tag{5}$$

$$= \arg\max_{\theta} \sum\nolimits_{(x,y)\in D_L} \log \frac{P(y|x;\theta)P(x;\theta)}{P(y;\theta)} \tag{6}$$

$$= \arg\max_{\theta} \sum\nolimits_{(x,y)\in D_L} \log \frac{P(y|x;\theta)}{P(y;\theta)} \tag{7}$$

$$= \arg\max_{\theta} \sum\nolimits_{(x,y)\in D_L} \log P(y|x;\theta) \cdot \frac{\log P(y|x;\theta) - \log P(y;\theta)}{\log P(y|x;\theta)} \tag{8}$$

$$\triangleq \arg\max_{\theta} \sum\nolimits_{(x,y)\in D_L} \log P(y|x;\theta) \cdot \frac{1}{s(x,y)}. \tag{9}$$

Estimate the propensity score $s(x,y)$ for image x and adjust $P(y|x;\theta)$ by a class-wise prior $P(y;\theta)$ for each labeled data (x,y).

To estimate $P(y;\theta)$, we estimate the propensity within a mini-batch and use a moving averaging strategy.

$$\hat{P}(Y) \leftarrow \mu\hat{P}(Y) + (1-\mu)P(Y; B_t, \theta_t)$$

(2) Class-Aware Imputation For Unlabeled data

Fixed threshold is too coarse the impute missing labels.

To tackle this challenge, we propose a Class-Aware Imputation (CAI) strategy that dynamically adjusts the pseudo-label assignment threshold for different classes. Let C_x denote the potential imputed label for image x, *i.e.*, $C_x = \arg\max_y P(y|x;\theta)$. We use a class-aware threshold $\tau(x)$ for image x as:

$$\tau(x) = \tau_o \cdot \left(\frac{\hat{P}(C_x)}{\max\limits_{y\in\{1,\cdots,C\}} \hat{P}(y)}\right)^{\beta}, \tag{11}$$

(3) Class-Aware Doubly Robust Estimation

Following the formulation of DR estimator, we first rewrite the training objective of CAP and CAI in semi-supervised learning as:

$$\mathcal{L}_{\text{CAP}} = \frac{1}{N}\sum_{i=1,\cdots,N} \frac{(1-m^{(i)})\mathcal{L}_s(x^{(i)}, y^{(i)})}{p^{(i)}} \tag{12}$$

$$\mathcal{L}_{\text{CAI}} = \frac{1}{N}\sum_{i=1,\cdots,N} (m^{(i)}\mathcal{L}_u(x^{(i)}, q^{(i)})\,\mathbb{I}(\text{con}(q^{(i)}) > \tau(x^{(i)})) + (1-m^{(i)})\mathcal{L}_s(x^{(i)}, y^{(i)})), \tag{13}$$

where $m^{(i)}$ is the missing state, $p^{(i)}$ is the propensity score, $q^{(i)}$ is the imputed label with confidence $\text{con}(q^{(i)})$, and $\mathbb{I}(\cdot)$ is the indicator function. As introduced in Section 4.1, we estimate the propensity score $p^{(i)}$ as $s(x^{(i)}, y^{(i)})$ in CAP. Then the optimization of CADR estimator is implemented as

$$\hat{\theta}_{\text{CADR}} = \arg\min_{\theta} \mathcal{L}_{\text{CADR}} = \arg\min_{\theta} \mathcal{L}_{\text{CAP}} + \mathcal{L}_{\text{CAI}} + \mathcal{L}_{\text{supp}}, \tag{14}$$

$$\text{where} \quad \mathcal{L}_{\text{supp}} = \frac{1}{N}\sum_{i=1,\cdots,N} (1 - m^{(i)} - \frac{1-m^{(i)}}{p^{(i)}})\mathcal{L}_u(x^{(i)}, q^{(i)})\,\mathbb{I}(\text{con}(q^{(i)}) > \tau) \tag{15}$$

$$- \frac{1}{N}\sum_{i=1,\cdots,N} (1-m^{(i)})\mathcal{L}_s(x^{(i)}, y^{(i)}),$$

which is a supplementary loss to guarantee the unbiasedness. In this design, $\mathcal{L}_{\text{CAI}} + \mathcal{L}_{\text{supp}}$ is expected to be 0 given correct CAP, and $\mathcal{L}_{\text{CAP}} + \mathcal{L}_{\text{supp}}$ is expected to be 0 given correct CAI. These results guarantee the double robustness in case that either the propensity or imputation is inaccurate.