Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism

Aude Sportisse [1] Hugo Schmutz [1,2] Olivier Humbert [2] Charles Bouveyron [1] Pierre-Alexandre Mattei [1]

[1]Université Côte d'Azur, Inria, Maasai, LJAD, CNRS [2]Université Côte d'Azur, TIRO-MATOS, UMR CEA E4320

Semi-supervised learning (SSL)

- **Huge amount** of data is available, but labeling the data can be **costly** or **time-consuming**.
- Goal: learn a predictive model

$$\theta^\star = \text{argmin}_{\theta\in\Theta} \quad \mathcal{R}(\theta) := \mathbb{E}_{(x,y)\sim p(x,y)}[\ell_\ell(\theta; x, y)]$$

by leveraging both labeled and unlabeled data:
- n_ℓ **labeled data**: $D_\ell = \{(x_i, y_i\}_{i=1}^{n_\ell}$
- n_u **unlabeled data**: $D_u = \{(x_i)\}_{i=n_\ell+1}^{n}$

Informative labels in SSL

- $r \in \{0,1\}^n$ indicates **where are the missing values in** y

$$\forall i \in \{1,\dots,n\}, r_i = \begin{cases} 1 \text{ if } y_i \text{ is observed} \\ 0 \text{ otherwise.} \end{cases}$$

- Classical assumption (Missing Completely At Random): $r \perp\!\!\!\perp x, y$.
- **Missing Not At Random (MNAR):** $r \not\perp\!\!\!\perp x, y$.
For example, doctors may prioritize labeling the class of sick patients or leave unlabeled the data with an ambiguous diagnosis.

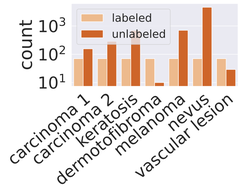


Figure 1:MNAR labels for dermaMNIST (log count of labeled & unlabeled images)

Main issues raised by MNAR labels:

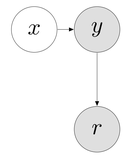
- **consider the mechanism** $\mathbb{P}(r = 1|x, y)$ (otherwise, biased results).
- **prove the identifiability** (two equal observed distribution can lead to different parameters of the data distribution).

Our proposal:

- **Estimate the missing-data mechanism**: (i) to debias any SSL algorithm in presence of MNAR labels; (ii) to provide a heuristic procedure to test whether the labels are indeed MNAR.
- Existing work (Hu et al., 2021): they debias the risk estimator, but they do not directly model the missing-data mechanism.

Identification & estimation of the mechanism

Assumption: the labels are **self-masked MNAR**, i.e. $r \perp\!\!\!\perp x|y$.



✓ it can reflect the classes popularity
✓ it implies $\mathbb{P}(r = 1|x, y) = \mathbb{P}(r = 1|y)$.

Identification of the joint distribution $p(y, x, r)$

Under **self-masked MNAR** labels, the joint distribution is **identified**, i.e. it can be expressed with quantities involving only observed data.

Estimating the missing-data mechanism

- **Method of moments estimator** (ME):

$$(\hat{\phi}_y^M)_\theta = \underbrace{\frac{\sum_{i=1}^n \mathbb{1}_{\{r=1, y_i=y\}}}{n}}_{\text{numbers of labeled data in class y}} \frac{1}{\hat{p}(y;\theta)}, \quad \hat{p}(y;\theta) = \frac{1}{n}\sum_{i=1}^n p(y_i|x_i;\theta)$$

- **Maximum likelihood estimator** (MLE): $\hat{\theta}^L, \hat{\phi}^L = \text{argmin}_{\theta,\phi} \quad \ell(\theta, \phi)$

$$\ell(\theta,\phi) \propto -\frac{1}{n}\sum_{i=1}^{n_\ell} \log p(y_i|x_i;\theta)\phi_{y_i} - \frac{1}{n}\sum_{i=n_\ell+1}^{n} \log \sum_{\tilde{y}\in\mathcal{C}} p(\tilde{y}|x_i;\theta)(1-\phi_{\tilde{y}})$$

Debiasing the classical SSL estimator

Classical SSL estimator of the risk $\mathcal{R}(\theta)$ (for MCAR labels):

$$\hat{\mathcal{R}}^{\text{SSL}(\theta)} := \frac{1}{n}\sum_{i=1}^n r_i \frac{\ell_\ell(\theta; x_i, y_i)}{n_\ell/n} + \frac{\lambda}{n}\sum_{i=1}^n (1-r_i)\frac{\ell_u(\theta; x_i)}{n_u/n}, \ \lambda \geq 0$$

where ℓ_u is a loss function which does not depend on the labels.
Popular approach: select the **pseudo-labels** with predicted proba. $> \tau$.

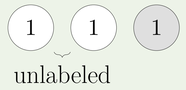
$$\ell_u(\theta; x) = -\log \max p(y|x:\theta)\mathbb{1}_{\max_y p(y|x;\theta)>\tau}$$

Inverse Propensity Weighting (IPW)

$$\hat{\mathcal{R}}_{\hat{\phi}}^{\text{SSL}}(\theta) := \frac{1}{n}\sum_{i=1}^n \frac{r_i \ell_\ell(\theta; x_i, y_i)}{\hat{\phi}_{y_i}} - \frac{\lambda}{n}\sum_{i=1}^n \frac{(r_i - \hat{\phi}_{y_i})}{\hat{\phi}_{y_i}}\ell_u(\theta; x_i)),$$

where $\hat{\phi}_{y_i}$ is an estimator of the mechanism $\phi_{y_i} = \mathbb{P}(r_i = 1|y_i)$.
Idea behind IPW technique: weight the labeled data by the inverse of the probability of being observed.

Probability of observing class 1: $\phi_1 = 1/3$
$\rightarrow$ the observed sample will be counted 3 times

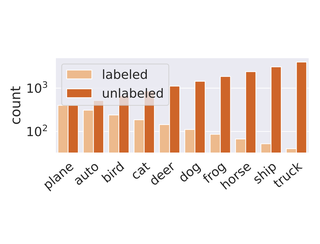


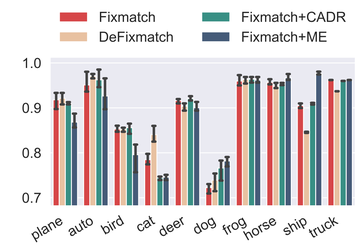
Theoretical results

Consistency

- The moment estimator $(\hat{\phi}_y^M)_\theta$ is consistent for a fixed $\theta \in \Theta$.
- Under mild assumptions on the joint distribution and assuming that ϕ is in the interior of the set Φ, the MLE $\hat{\phi}^L$ is consistent.
- If $\hat{\phi}$ is a consistent estimator of ϕ, the risk $\hat{\mathcal{R}}_{\hat{\phi}}^{\text{SSL}}(\theta)$ is consistent estimator of the theoretical risk.

Unbalanced MNAR setting in CIFAR10





Method	Loss	Accuracy	Mechanism
Complete Case Fixmatch	1.647 ± 0.025	68.26 ± 0.56	MCAR
Fixmatch (Sohn et al., 2020)	0.426 ± 0.017	90.91 ± 0.12	MCAR
DeFixmatch (Schmutz et al., 2023)	0.536 ± 0.020	89.71 ± 0.16	MCAR
Fixmatch + CADR (Hu et al., 2021)	0.452 ± 0.006	91.14 ± 0.30	MNAR
Fixmatch + ME (Ours)	0.321 ± 0.016	91.88 ± 0.24	MNAR

dermaMNIST: identifying nevi

- 10,015 dermatoscopic images (Codella et al., 2019) Unbalanced dataset: 70% of the images are benign nevi (class **nevus**).
- *Pseudo-realistic MNAR scenario:* we assume that a medical doctor would like to classify the conditions equally and select 70 images per class for labeling (7% of observed labels, see Figure 1).

Method	Loss	Accuracy	Accuracy Nevus	MSE ϕ_{nevus}
PseudoLabel	1.34 ± 0.16	57.72 ± 1.95	66.14 ± 5.86	0.80
CADR (Hu et al.)	1.42 ± 0.060	49.36 ± 1.91	50.41 ± 5.38	0.77 ± 0.02
MLE (Ours)	0.993 ± 0.020	66.4 ± 0.81	91.16 ± 2.26	0.34 ± 0.03
MEg (Ours)	1.19 ± 0.148	66.65 ± 1.76	93.54 ± 2.30	0.42 ± 0.08
ME (Ours)	1.24 ± 0.087	65.8 ± 0.78	85.91 ± 3.05	0.38 ± 0.15

✓ Our methods determine if a lesion is a nevus or not with a high accuracy and give the best MSE for the estimation of the mechanism ϕ_{nevus} in the **nevus** class.