Informative Labels in Semi-Supervised Learning

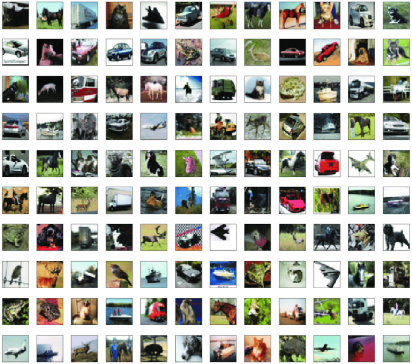
Aude Sportisse
Charles Bouveyron, Pierre-Alexandre Mattei, Hugo Schmutz
Centre Inria d'Université Côte d'Azur, Maasai Team

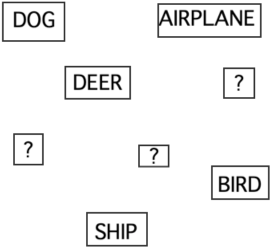
December 13., 2022

Outline

Context

- Huge amount of data is available.
- Labeling the data is costly and time-consuming.





How to leverage from the unlabeled data?

SSL is a missing data problem

- Unlabeled data are seen as observations having a **missing label**.
- $r \in \{0,1\}^n$ indicates where are the missing values in the label y

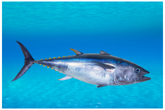
$$\forall i \in \{1, \ldots, n\}, r_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

- Remark: y is partially missing, but r is fully observed.
- r is sometimes **informative**: when some classes are popular



DEER DEER DEER

and other classes are not...


? TUNA ?

Missing-data mechanism

- Not-informative labels (MCAR): the process that causes the lack of data is **totally independent** from the data values.

Not-informative labels: one can ignore the mechanism.

- Informative labels (MNAR): *People are more inclined to label images of some classes which are easy to recognize.*

Informative labels: one should consider the mechanism.

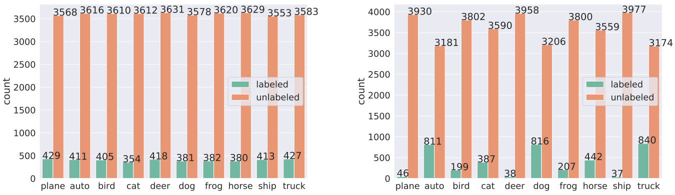


Figure: Artificial missing labels in CIFAR10 datasets.
Left: MCAR labels. Right: MNAR labels.

Issues raised by informative labels

1. **How to consider the mechanism ?**
   - model the conditional distribution $\mathcal{L}(R|X, Y)$ (Bernoulli distribution)
   - take it implicitly into account: [Mohan et al., 2018] (Estimation in linear models) and [Hu et al., 2021] (SSL)

2. **Are the estimators still identifiable ?**

Not always : 2 equal observed distributions can lead to different parameters of the data distribution.

3. **How to adapt the existing methods ?**

4. **How to test the assumption on the mechanism ?**

Discussions with experts are very important. Sometimes, it is possible to do it automatically.

Outline

SSL setting

n i.i.d. samples $D = \{(x_i, y_i)\}_{i=1}^{n}$

- $x_i \in \mathbb{R}^d$ the features (e.g. images)
- $y_i \in \mathcal{C} = \{0, \ldots, K\}$ the labels

We want to estimate θ, parameter of $p(y|x;\theta)$

In practice, $p(y|x;\theta)$ can be a neural network.

What we observe

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

Typically: $n_\ell << n_u$.

How to use all the data to estimate θ?

Reminder in supervised learning

[Supervised learning]

- Objective: learn a predictive model $p(y|x;\theta)$.
- The oracle estimate is the minimizer of the theoretical risk:

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

with L the loss function (measures the error committed by the model to retrieve y).

The theoretical risk is always intractable.

- Minimize the empirical risk:

$$\hat{\theta} = \mathrm{argmin}_{\theta\in\Theta} \quad \hat{\mathcal{R}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} L(\theta; x_i, y_i).$$

The empirical risk is still unobserved in presence of missing labels.

Classical SSL estimator (for MCAR labels)

[Semi-supervised learning]

1) Complete-case: learning with labeled data

Minimize the complete-case empirical risk:

$$\hat{\mathcal{R}}^{\mathrm{CC}}(\theta) := \frac{1}{n_\ell} \underbrace{\sum_{i=1}^{n} r_i L(\theta; x_i, y_i)}_{\text{only the labeled data are used}}$$

2) Incorporating the unlabeled data

$$\hat{\mathcal{R}}^{\mathrm{SSL}}(\theta) := \frac{1}{n_\ell} \underbrace{\sum_{i=1}^{n} r_i L(\theta; x_i, y_i)}_{\text{term on labeled data}} + \frac{\lambda}{n_u} \underbrace{\sum_{i=1}^{n} (1 - r_i) H(\theta; x_i)}_{\text{term on unlabeled data}}$$

$\lambda > 0$: regularization parameter

H: surrogate of L

Choice of the SSL regularization

High-confident imputations for the unlabeled data

- Shannon entropy [Grandvalet and Bengio, 2004]:

$$H(\theta; x) = -\sum_y p(y|x;\theta)\log(p(y|x;\theta)).$$

- Pseudo-labels [Rizve et al., 2021]:
    - choose the class with the maximum predicted probability

    $$c \in \mathrm{argmax}_y\, p(y|x;\theta)$$

    - only the pseudo-labels which have a maximum predicted probability larger than a predefined threshold τ are used as target
    $$H(\theta; x) = -\log p(c|x:\theta)\mathbb{1}_{\max_y p(y|x;\theta)>\tau}$$

Robustness of the model to data augmentation of the features
Recent state-of-the-art method: Fixmatch [Sohn et al.,] and many extensions.

Is SSL a promising approach?

100 labeled images per class for CIFAR10

Error with supervised learning (neural network): 12%
Error using a large unlabeled dataset (FixMatch SSL): **2,5%**

But...

- Popular deep SSL techniques are generally **not safe**, meaning that their theoretical guarantees are not stronger than the complete case baseline [Schmutz et al., 2022].
- **Performances of SSL classical techniques are degraded when the labeled and unlabeled set have different distributions (MNAR)** [Oliver et al.,].

And also...

- Without data augmentation, the gap in performance between using SSL and using only labeled data is smaller.
- Many papers perform not realistic numerical experiments (e.g. too large complete validation set, costly hypertunning parameters) [Oliver et al.,].

Safe MCAR SSL

Get a **debiased estimate** of the theoretical risk for MCAR labels [Schmutz et al., 2022]:

$$\hat{\mathcal{R}}^{\mathrm{SSL}}(\theta) := \frac{1}{n_\ell} \underbrace{\sum_{i=1}^{n} r_i L(\theta; x_i, y_i)}_{\text{term on labeled data}} + \frac{\lambda}{n_u} \underbrace{\sum_{i=1}^{n} (1 - r_i) H(\theta; x_i)}_{\text{term on unlabeled data}} - \frac{\lambda}{n_\ell} \underbrace{\sum_{i=1}^{n} r_i H(\theta; x_i)}_{\text{to get unbiased estimate}}$$

$\lambda > 0$: regularization parameter

H: surrogate of L

Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing

Hugo Schmutz HUGO.SCHMUTZ@INRIA.FR
Université Côte d'Azur
TIRO-MATOS, UMR CEA E4320
Inria, Maasai project-team
Laboratoire J.A. Dieudonné, UMR CNRS 7351
Nice, France

Olivier Humbert OLIVIER.HUMBERT@UNIV-COTEDAZUR.FR
Université Côte d'Azur
TIRO-MATOS, UMR CEA E4320
Centre Antoine Lacassagne
Nice, France

Pierre-Alexandre Mattei PIERRE-ALEXANDRE.MATTEI@INRIA.FR
Université Côte d'Azur
Inria, Maasai project-team
Laboratoire J.A. Dieudonné, UMR CNRS 7351
Nice, France

Towards realistic scenarios

- Class-imbalanced SSL / MCAR: [Kim et al., , Wei et al., , Lee et al.,].
- Different class distribution /MNAR : [Hu et al., 2022]
- Class distribution mismatch / MNAR: **(a)** [Guo et al., , Cao et al.,] or **(b)** [Chen et al.,]
- Class & feature distribution mismatch: **(c)** [Huang et al.,]

	Assumption	Labeled data	Unlabeled data
(a)	$\mathcal{C}^\ell \subset \mathcal{C}^u$	 Pigeon Blackbird	
(b)	$\mathcal{C}^\ell \neq \mathcal{C}^u$	 Pigeon Blackbird Parakeet	
(c)	$\mathcal{C}^\ell \neq \mathcal{C}^u$ $p^\ell(x\|y) \neq p^u(x\|y)$	 Pigeon Blackbird Parakeet	

Outline

Our assumptions

A1. The labels sets are identical: $\mathcal{C}^\ell = \mathcal{C}^u = \mathcal{C} = \{0, \ldots, K\}$.
It implies that we can not have a "new" class in the unlabeled dataset.

A2. The labels are informative (self-masked MNAR): $r \perp\!\!\!\perp x | y$.
Our model can reflect the classes popularity.

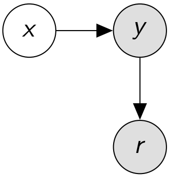


Figure: Structural causal graph of the self-masked mechanism.

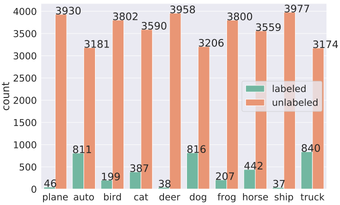


Figure: CIFAR 10 dataset with 10% labeled data (in total).

Our proposal

- **Estimate the mechanism**.
- Prove the **identifiability** of the parameters.

Proposition: identifiability

Under Assumptions **A2.** (**self-masked MNAR**), **identifiability of** θ for the marginal distribution $p(y|x;\theta)$ and **completness** (features has a larger support than the labels), the parameters (θ, ϕ) are identifiable.

[Miao et al., 2015]

- **Debiase the classical estimator** to handle informative labels.

Debiased estimator (for MNAR labels)

1) Complete-case: learning with labeled data

Weight the labeled data by the inverse of the probability (IPW) of being observed.

$$\hat{\mathcal{R}}^{\mathrm{CC,MNAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i L(\theta; x_i, y_i)}{\hat{\phi}_{y_i}},$$

with $\hat{\phi}_{y_i} = \mathbb{P}(r_i = 1 | y_i)$.

2) Incorporating the unlabeled data

$$\hat{\mathcal{R}}^{\mathrm{SSL,MNAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i L(\theta; x_i, y_i)}{\hat{\phi}_{y_i}} + \frac{\lambda}{n}\left(\sum_{i=1}^{n}(1-r_i)H(\theta; x_i) - \sum_{i=1}^{n} r_i \frac{(1-\hat{\phi}_{y_i})}{\hat{\phi}_{y_i}} H(\theta; x_i)\right)$$

Estimation of the mechanism

- **Maximum likelihood estimator** (MLE):

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

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

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

- Implicitly taking into account the MNAR nature of the data [Hu et al., 2021].

Naive estimators in specific cases

- We know that the class are balanced:

$$\hat{\phi}_y = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{r=1, y_i=y\}}}{n} \frac{1}{K},$$

  where K is the number of classes.
- We know the class probabilities $p(y)$:
    - we have data in the general population (e.g. the rate of nodule with such a malignancy level in the general population).

$$\hat{\phi}_y = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{r=1, y_i=y\}}}{n} \frac{1}{p(y)}$$

Numerical experiment

Method	Loss	Acc
MLE	1.312	63.26
MM	**0.3643**	**92.17**
Implicit meca	0.4885	90.54

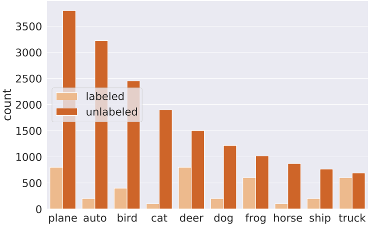


Figure: CIFAR10 with informative missing values and unbalanced classes (18% labeled data in total).

Outline

Conclusion

Surprising facts:

- classical method using MLE for estimation of ϕ fails in many cases.
- "Simple" missing-data setting (MNAR but only one variable is missing!) but complex data (images; need of using CNN).

Continued work:

- Apply the method to a **real medical dataset** (Collaboration with Olivier Humbert, Pr CHU Nice).
- Propose a **likelihood ratio test** to verify the assumption on the mechanism.

Thanks for your attention !

Visit our website !

https://rmisstastic.netlify.app/

Imke Mayer, Julie Josse, Nicholas Tierney and Nathalie Vialaneix and **many other contributors**

Home Workflows Lectures Bibliography Implementations Data People News & links Contact

R-miss-tastic

A resource website on missing values - Methods and references for managing missing data

Welcome!

Mon Apr 19, 2021 by R-miss-tastic

This website provides the main methods, references and implementations (in R and python) for managing missing data, whether to impute, estimate or predict.
Click here for the article introducing this project.
Read more →

FAQ

Sun Apr 18, 2021 by R-miss-tastic

When it comes to analyses with missing values, some questions are raised regularely during classes or seminars. We try to list the most popular questions with some elements of response. If you have another question related to the handling of missing values, feel free to contact us via the Contact form.

Read more →

About

This website is sponsored by R Consortium and maintained by Julie Josse, Imke Mayer, Aude Sportisse, Nicholas Tierney and Nathalie Vialaneix.

Article on arXiv →
Read more →
FAQ →

Follow us!

Events
GitHub
Twitter paper bot
MissCausal

References I

Cao, K., Brbic, M., and Leskovec, J.
Open-world semi-supervised learning, 2021.

Chen, Y., Zhu, X., Li, W., and Gong, S.
Semi-supervised learning under class distribution mismatch, 2020.

Grandvalet, Y. and Bengio, Y. (2004).
Semi-supervised learning by entropy minimization.
Advances in neural information processing systems, 17.

Guo, L.-Z., Zhang, Z.-Y., Jiang, Y., Li, Y.-F., and Zhou, Z.-H.
Safe deep semi-supervised learning for unseen-class unlabeled data, 2020.

Hu, X., Niu, Y., Miao, C., Hua, X.-S., and Zhang, H. (2021).
On non-random missing labels in semi-supervised learning.
In *International Conference on Learning Representations*.

Hu, X., Niu, Y., Miao, C., Hua, X.-S., and Zhang, H. (2022).
On non-random missing labels in semi-supervised learning.
In *International Conference on Learning Representations*.

References II

Huang, Z., Xue, C., Han, B., Yang, J., and Gong, C., .
Universal semi-supervised learning.

Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S., and Shin, J.
Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning, 2020.

Lee, H., Shin, S., and Kim, H.
Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning, 2021.

Miao, W., Liu, L., Tchetgen, E. T., and Geng, Z. (2015).
Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable.
arXiv preprint arXiv:1509.02556.

Mohan, K., Thoemmes, F., and Pearl, J. (2018).
Estimation with incomplete data: The linear case.
In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization.*

References III

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I.
Realistic evaluation of deep semi-supervised learning algorithms, 2018.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. (2021).
In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning.
arXiv preprint arXiv:2101.06329.

Schmutz, H., Humbert, O., and Mattei, P.-A. (2022).
Don't fear the unlabelled: safe deep semi-supervised learning via simple debiaising.
arXiv preprint arXiv:2203.07512.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C.
Fixmatch: Simplifying semi-supervised learning with consistency and confidence, 2020.

References IV

Wei, C., Sohn, K., Mellina, C., Yuille, A., and Yang, F.
Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning, 2021.

Very closed work: [Hu et al., 2022]

- classical: $\hat{\theta} = \mathrm{argmax}_\theta p(y|x;\theta)$.
- For MNAR: $\mathcal{L}(y|x, r = 1) \neq \mathcal{L}(y|x, r = 0)$
- **Proposition:** consider the regression $x|y$ and assume $x \perp\!\!\!\perp r|y$
  - $\hat{\theta} = \mathrm{argmax}_\theta p(x|y;\theta)$
  - $\mathcal{L}(x|y, r = 1) = \mathcal{L}(x|y, r = 0)$
  - $\mathrm{argmax}_\theta p(x|y;\theta) = \mathrm{argmax}_\theta p(y|x;\theta)\frac{1}{s(x,y)}$.
  - $s(x, y)$ depends on the unknown class probabilities $p(y)$.
  - $\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} p(y_i|x_i;\hat{\theta})$ (as $p(y) = \int p(y|x;\theta)p(x)dx$).
- **Double-robustness property** ever θ or $s(x, y)$ can be biased, the theoretical risk will be unbiased.

Some remarks

- The gradient over θ is not propagated though the weight $s(x, y)$ while θ is used to compute it.
- Double-robust: if $s(x, y)$ is biased, the proposition requires perfect imputations for unlabeled data.

How to estimate the mechanism?

Our first idea was to use the maximum likelihood estimate.

Maximum Likelihood Estimate

- Mechanism: $\phi_{y_i} = \mathbb{P}(r_i = 1|y_i)$
- $\ell(\theta, \phi) = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i, y_i, r_i; \theta, \phi)$ untractable.
- Integrate over the missing values: observed log-likelihood.

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

$$\hat{\theta}, \hat{\phi} = \mathrm{argmin}_{\theta\in\Theta,\phi\in\Phi} \quad \ell(\theta, \phi).$$

Advantages:

- Convexity of the observed log-likelihood in $\phi \in \Phi$ for a fixed $\theta \in \Theta$.
- Possible use of a prior on ϕ (with regularization of the log-likelihood)
- Likelihood ratio test is easily derived in practice.

Choice of the SSL regularization

2) **Robustness of the model to data augmentation of the features**
Recent state-of-the-art method: Fixmatch [Sohn et al.,] and many extensions.

- compute a pseudo-labels predicted using a weakly-augmented version of x.
- minimize the likelihood with predictions of the model on a strongly-augmented version of x.

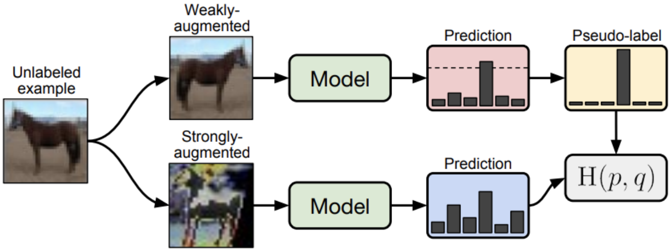


Figure: Credits [Sohn et al.,]