Conformal Credal Self-Supervised Learning

Julian Lienen[1], Caglar Demir[1], Eyke Hüllermeier[2]

[1] Paderborn University, Germany [2] LMU Munich, Germany

September 14th, 2023

Setting

- We assume instances $\boldsymbol{x} \in \mathcal{X}$ with associated ground-truth $p^*(\cdot \mid \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$ for categorical targets (classes) $\mathcal{Y} = \{y_1, ..., y_K\}$

PADERBORN UNIVERSITY

Setting

- We assume instances $\boldsymbol{x} \in \mathcal{X}$ with associated ground-truth $p^*(\cdot \mid \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$ for categorical targets (classes) $\mathcal{Y} = \{y_1, ..., y_K\}$
- In practice, only realization $y \in \mathcal{Y}$ of random variable $Y \sim p^*(\cdot \mid \boldsymbol{x})$ given as training information in $\mathcal{D}_{\text{labeled}}$

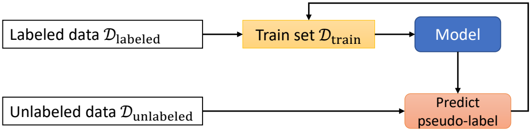
Setting

- We assume instances $\boldsymbol{x} \in \mathcal{X}$ with associated ground-truth $p^*(\cdot \mid \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$ for categorical targets (classes) $\mathcal{Y} = \{y_1, ..., y_K\}$
- In practice, only realization $y \in \mathcal{Y}$ of random variable $Y \sim p^*(\cdot \mid \boldsymbol{x})$ given as training information in $\mathcal{D}_{\text{labeled}}$
- Here, we consider a semi-supervised learning setting with unlabeled data without labels in $\mathcal{D}_{\text{unlabeled}}$

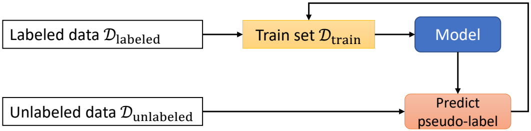
Setting

- We assume instances $\boldsymbol{x} \in \mathcal{X}$ with associated ground-truth $p^*(\cdot \mid \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$ for categorical targets (classes) $\mathcal{Y} = \{y_1, ..., y_K\}$
- In practice, only realization $y \in \mathcal{Y}$ of random variable $Y \sim p^*(\cdot \mid \boldsymbol{x})$ given as training information in $\mathcal{D}_{\text{labeled}}$
- Here, we consider a semi-supervised learning setting with unlabeled data without labels in $\mathcal{D}_{\text{unlabeled}}$
- **Goal**: Learn probabilistic classifier $\widehat{p} : \mathcal{X} \mapsto \mathbb{P}(\mathcal{Y})$

Self-Training



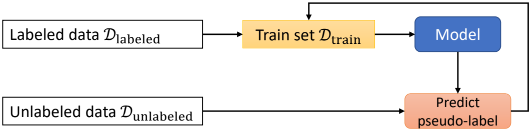
Self-Training



- Typically, pseudo-labels are in the form of single prob. distributions [Lee13]

PADERBORN UNIVERSITY

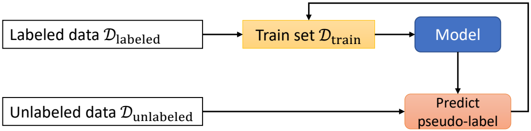
Self-Training



- Typically, pseudo-labels are in the form of single prob. distributions [Lee13]
  - E.g., degenerate distributions p_y with $p_y(y \mid \boldsymbol{x}) = 1$ and $p_y(y' \mid \boldsymbol{x}) = 0$ for $y \neq y'$, where $y := \text{argmax}_{y \in \mathcal{Y}} \widehat{p}(y \mid \boldsymbol{x})$ [SBC+20]

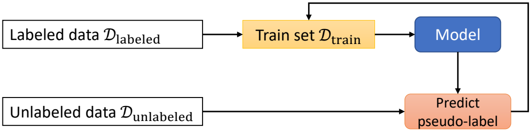
PADERBORN UNIVERSITY

Self-Training



- Typically, pseudo-labels are in the form of single prob. distributions [Lee13]
  - E.g., degenerate distributions p_y with $p_y(y \mid \boldsymbol{x}) = 1$ and $p_y(y' \mid \boldsymbol{x}) = 0$ for $y \neq y'$, where $y := \text{argmax}_{y \in \mathcal{Y}} \widehat{p}(y \mid \boldsymbol{x})$ [SBC$^+$20]
  - Single distribution is incapable in reflecting uncertainty properly, additional uncertainty-awareness means are required [RDRS21]

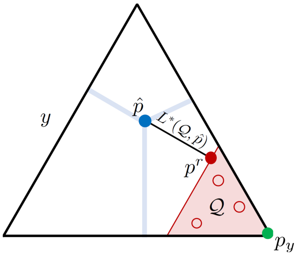
Self-Training



- Typically, pseudo-labels are in the form of single prob. distributions [Lee13]
  - E.g., degenerate distributions p_y with $p_y(y \,|\, \boldsymbol{x}) = 1$ and $p_y(y' \,|\, \boldsymbol{x}) = 0$ for $y \neq y'$, where $y := \operatorname{argmax}_{y \in \mathcal{Y}} \widehat{p}(y \,|\, \boldsymbol{x})$ [SBC+20]
  - Single distribution is incapable in reflecting uncertainty properly, additional uncertainty-awareness means are required [RDRS21]
  - Too extreme distributions p_y may lead to biased and overconfident classifiers $\widehat{p}$ [LH21b]

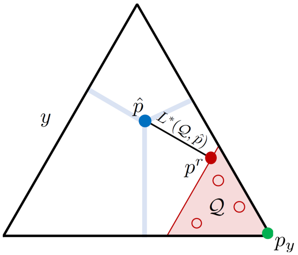
Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision



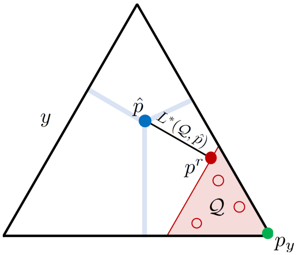
Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision
  - ☐ Supposed to cover p^* with high probability



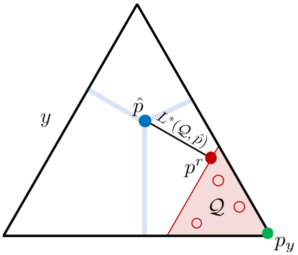
Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision
  - ☐ Supposed to cover p^* with high probability
  - ☐ Relieve from committing to single target distribution



Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision
  - ☐ Supposed to cover p^* with high probability
  - ☐ Relieve from committing to single target distribution

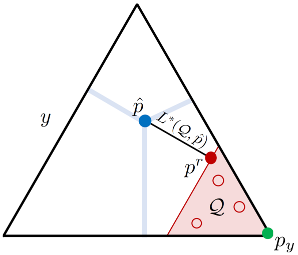


- Learning from set-valued targets by (optimistic) superset learning [HC15]:

$$\mathcal{L}^*(\mathcal{Q}, \hat{p}) := \min_{p \in \mathcal{Q}} \mathcal{L}(p, \hat{p})$$

Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision
  - ☐ Supposed to cover p^* with high probability
  - ☐ Relieve from committing to single target distribution



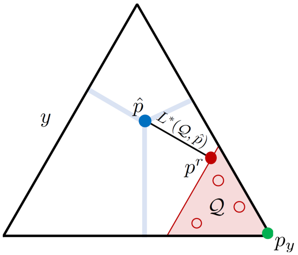
- Learning from set-valued targets by (optimistic) superset learning [HC15]:

$$\mathcal{L}^*(\mathcal{Q}, \hat{p}) := \min_{p \in \mathcal{Q}} \mathcal{L}(p, \hat{p})$$

  - ☐ No prediction $\hat{p} \in \mathcal{Q}$ is penalized (relaxation, data disambiguation)

Credal Label Learning

- Credal labeling [LH21b]: Use *credal sets* (sets of probability distributions) $\mathcal{Q} \subseteq \mathbb{P}(\mathcal{Y})$ as supervision
  - ☐ Supposed to cover p^* with high probability
  - ☐ Relieve from committing to single target distribution



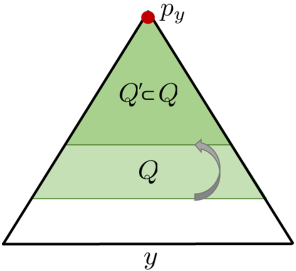
- Learning from set-valued targets by (optimistic) superset learning [HC15]:

$$\mathcal{L}^*(\mathcal{Q}, \hat{p}) := \min_{p \in \mathcal{Q}} \mathcal{L}(p, \hat{p})$$

  - ☐ No prediction $\hat{p} \in \mathcal{Q}$ is penalized (relaxation, data disambiguation)
  - ☐ With $\mathcal{L} = D_{KL}$ and credal sets as depicted, $\mathcal{L}^*$ has convex closed-form expression (efficient optimization)

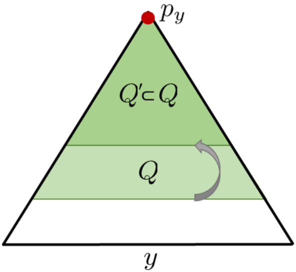
Credal Self-Supervised Learning [LH21a]

- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances



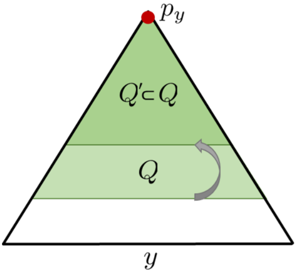
Credal Self-Supervised Learning [LH21a]

- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - Ad-hoc set construction based on the model confidence



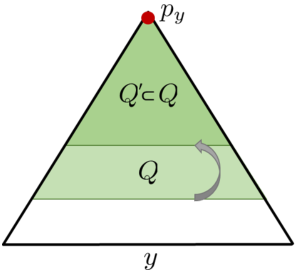
Credal Self-Supervised Learning [LH21a]

- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - ☐ Ad-hoc set construction based on the model confidence
  - ☐ More cautious yet uncertainty-aware learning behavior



Credal Self-Supervised Learning [LH21a]

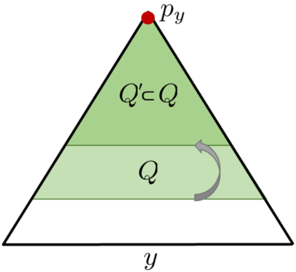
- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - ☐ Ad-hoc set construction based on the model confidence
  - ☐ More cautious yet uncertainty-aware learning behavior



- However, the credal set quality is solely subject to the model confidence

Credal Self-Supervised Learning [LH21a]

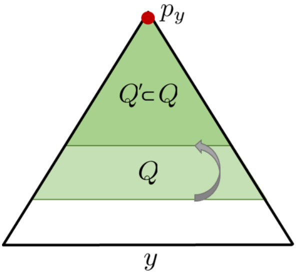
- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - ☐ Ad-hoc set construction based on the model confidence
  - ☐ More cautious yet uncertainty-aware learning behavior



- However, the credal set quality is solely subject to the model confidence
  - ☐ No *objective* validity guarantee

Credal Self-Supervised Learning [LH21a]

- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - ☐ Ad-hoc set construction based on the model confidence
  - ☐ More cautious yet uncertainty-aware learning behavior

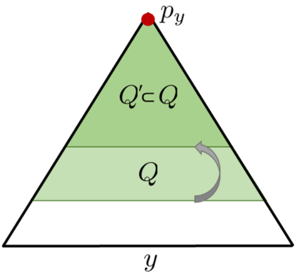


- However, the credal set quality is solely subject to the model confidence
  - ☐ No *objective* validity guarantee

$\Rightarrow$ (Inductive) Conformal prediction (CP) to the rescue [VGS05]: Quantifying the uncertainty of $\widehat{p}(\boldsymbol{x})$ with (marginal) validity guarantees

Credal Self-Supervised Learning [LH21a]

- *Credal self-supervised learning* (CSSL) [LH21a] maintains credal sets as (pseudo-)supervision for all unlabeled instances
  - Ad-hoc set construction based on the model confidence
  - More cautious yet uncertainty-aware learning behavior



- However, the credal set quality is solely subject to the model confidence
  - No *objective* validity guarantee

$\Rightarrow$ (Inductive) Conformal prediction (CP) to the rescue [VGS05]: Quantifying the uncertainty of $\widehat{p}(\boldsymbol{x})$ with (marginal) validity guarantees

- Separated calibration data $\mathcal{D}_{\text{calib}}$ from the original labeled set $\mathcal{D}_{\text{labeled}}$

Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities

Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities
  - Intuitively, $\pi_{\boldsymbol{x}}(y') \in [0, 1]$ upper bounds $p^*(y')$ for all $y' \in \mathcal{Y}$

Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities
  - Intuitively, $\pi_{\boldsymbol{x}}(y') \in [0, 1]$ upper bounds $p^*(y')$ for all $y' \in \mathcal{Y}$
  - Needs to be normalized such that $\max_{y' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(y') = 1$

PADERBORN UNIVERSITY

Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities
  - ☐ Intuitively, $\pi_{\boldsymbol{x}}(y') \in [0,1]$ upper bounds $p^*(y')$ for all $y' \in \mathcal{Y}$
  - ☐ Needs to be normalized such that $\max_{y' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(y') = 1$
  - ☐ Then, it is guaranteed for the true outcome y assoc. with $\boldsymbol{x}$:

$$\Pr(\pi_{\boldsymbol{x}}(y) \leq \delta) \leq \delta$$

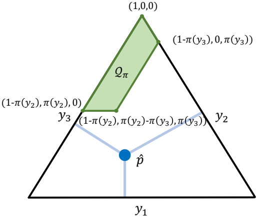
Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities
  - ☐ Intuitively, $\pi_{\boldsymbol{x}}(y') \in [0,1]$ upper bounds $p^*(y')$ for all $y' \in \mathcal{Y}$
  - ☐ Needs to be normalized such that $\max_{y' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(y') = 1$
  - ☐ Then, it is guaranteed for the true outcome y assoc. with $\boldsymbol{x}$:

$$\Pr(\pi_{\boldsymbol{x}}(y) \leq \delta) \leq \delta$$

- *Conformal credal sets* $\mathcal{Q}_{\pi_{\boldsymbol{x}}}$ in accordance with conformal possibilites $\pi_{\boldsymbol{x}}$ can be constructed by

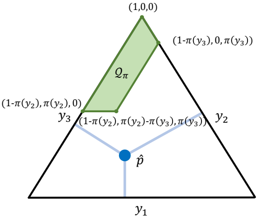
$$\mathcal{Q}_{\pi_{\boldsymbol{x}}} = \left\{ p \in \mathbb{P}(\mathcal{Y}) \,|\, \forall\, Y \subseteq \mathcal{Y} : \sum_{y \in Y} p(y) \leq \max_{y \in Y} \pi_{\boldsymbol{x}}(y) \right\}$$



PADERBORN UNIVERSITY

Conformal Credal Labeling

- [CM22] suggest to interpret p-values $\pi_{\boldsymbol{x}}$ for $\hat{p}(\boldsymbol{x})$ as possibilities
  - Intuitively, $\pi_{\boldsymbol{x}}(y') \in [0,1]$ upper bounds $p^*(y')$ for all $y' \in \mathcal{Y}$
  - Needs to be normalized such that $\max_{y' \in \mathcal{Y}} \pi_{\boldsymbol{x}}(y') = 1$
  - Then, it is guaranteed for the true outcome y assoc. with $\boldsymbol{x}$:
$$\Pr(\pi_{\boldsymbol{x}}(y) \leq \delta) \leq \delta$$
- *Conformal credal sets* $\mathcal{Q}_{\pi_{\boldsymbol{x}}}$ in accordance with conformal possibilites $\pi_{\boldsymbol{x}}$ can be constructed by
$$\mathcal{Q}_{\pi_{\boldsymbol{x}}} = \left\{ p \in \mathbb{P}(\mathcal{Y}) \,|\, \forall\, Y \subseteq \mathcal{Y} : \sum_{y \in Y} p(y) \leq \max_{y \in Y} \pi_{\boldsymbol{x}}(y) \right\}$$
- Learning from conformal credal labels via
$$\mathcal{L}^*(\mathcal{Q}_{\pi_{\boldsymbol{x}}}, \hat{p}) := \min_{p \in \mathcal{Q}_{\pi_{\boldsymbol{x}}}} D_{KL}(p \,||\, \hat{p})$$
  requires more sophisticated optimization algorithm



Generalized Credal Learning

Algorithm Generalized Credal Learning Loss

Require: Predicted distribution $\widehat{p} \in \mathbb{P}(\mathcal{Y})$, (normalized) possibility distribution $\pi : \mathcal{Y} \longrightarrow [0,1]$

if $\widehat{p} \in \mathcal{Q}_\pi$ **then return** $D_{KL}(\widehat{p} \,||\, \widehat{p}) = 0$

Initialize set of unassigned classes $Y = \mathcal{Y}$

while Y is not empty **do**

Determine $y^* \in Y$ with highest $\pi(y^*)$, such that the probabilities

$$\bar{p}(y) = \left(\pi(y^*) - \sum_{y' \notin Y} p^r(y') \right) \cdot \frac{\widehat{p}(y)}{\sum_{y' \in Y'} \widehat{p}(y')}$$

for all $y \in Y' := \{y \in Y \,|\, \pi(y) \leq \pi(y^*)\}$ do not violate any possibility constraints for classes $y' : \pi(y') \leq \pi(y^*)$

Assign $p^r(y) = \bar{p}(y)$ for all $y \in Y'$

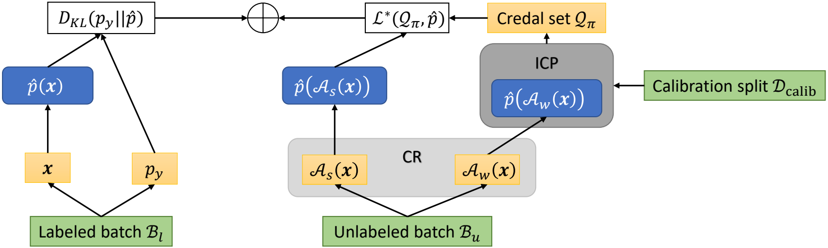
$Y = Y \setminus Y'$

end while

return $D_{KL}(p^r \,||\, \widehat{p})$

Conformal Credal Self-Supervised Learning (C^2S^2L)

Batch-wise loss calculation:

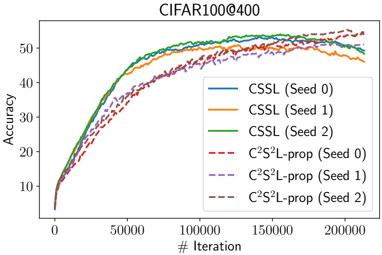


CR = Consistency Regularization

PADERBORN UNIVERSITY

Results

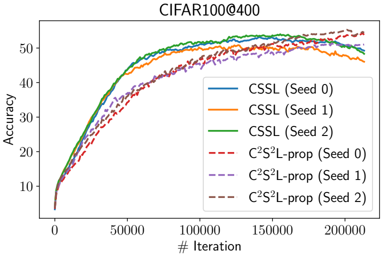
- Superior generalization performance compared to CSSL, especially in later phases of the training



PADERBORN UNIVERSITY

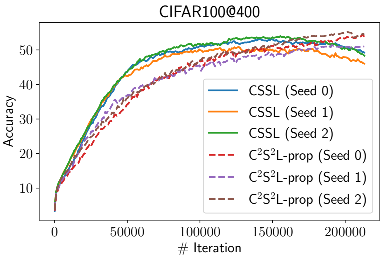
Results

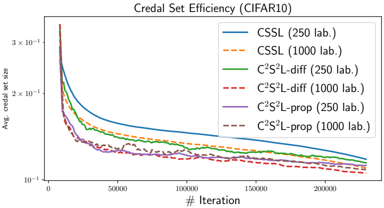
- Superior generalization performance compared to CSSL, especially in later phases of the training
  - ☐ Although calibration data for CP is separated from $\mathcal{D}_{\text{labeled}}$



Results

- Superior generalization performance compared to CSSL, especially in later phases of the training
  - Although calibration data for CP is separated from $\mathcal{D}_{\text{labeled}}$
- Higher quality of pseudo-supervision in C^2S^2L compared to CSSL

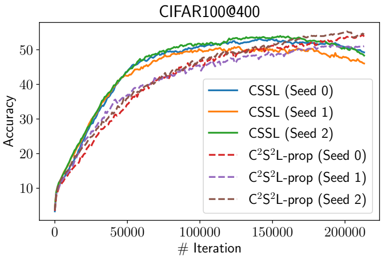


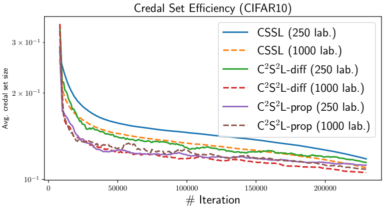


PADERBORN UNIVERSITY

Results

- Superior generalization performance compared to CSSL, especially in later phases of the training
  - ☐ Although calibration data for CP is separated from $\mathcal{D}_{\text{labeled}}$
- Higher quality of pseudo-supervision in C^2S^2L compared to CSSL
  - ☐ Smaller credal set sizes

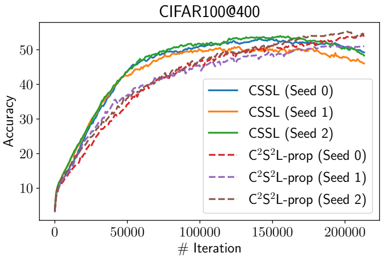


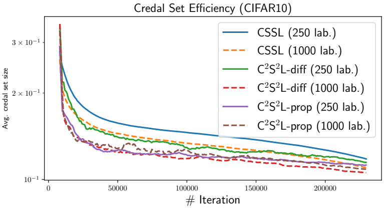


PADERBORN UNIVERSITY

Results

- Superior generalization performance compared to CSSL, especially in later phases of the training
  - ☐ Although calibration data for CP is separated from $\mathcal{D}_{\text{labeled}}$
- Higher quality of pseudo-supervision in C^2S^2L compared to CSSL
  - ☐ Smaller credal set sizes
  - ☐ Improved validity in terms of $\mathbb{1}(\pi(y) \leq \delta)$ for true outcomes y of the unlabeled instances





PADERBORN UNIVERSITY

Conclusion & Outlook

- Conformal prediction can be used to achieve model-agnostic guarantees in self-training

Conclusion & Outlook

- Conformal prediction can be used to achieve model-agnostic guarantees in self-training
  - Combination of superset learning with conformal prediction appears promising in this regard

Conclusion & Outlook

- Conformal prediction can be used to achieve model-agnostic guarantees in self-training
  - Combination of superset learning with conformal prediction appears promising in this regard
  - But how about conditional validity? Can we achieve (at least by an approximation) it?

Conclusion & Outlook

- Conformal prediction can be used to achieve model-agnostic guarantees in self-training
  - Combination of superset learning with conformal prediction appears promising in this regard
  - But how about conditional validity? Can we achieve (at least by an approximation) it?
- Generalized credal label learning allows to learn from arbitrary credal sets

Conclusion & Outlook

- Conformal prediction can be used to achieve model-agnostic guarantees in self-training
  - Combination of superset learning with conformal prediction appears promising in this regard
  - But how about conditional validity? Can we achieve (at least by an approximation) it?
- Generalized credal label learning allows to learn from arbitrary credal sets
  - But how about the scalability?

References

[CM22] Leonardo Cella and Ryan Martin. Validity, consonant plausibility measures, and conformal prediction. *Int. J. Approx. Reason.*, 141:110–130, 2022.

[HC15] Eyke Hüllermeier and Weiwei Cheng. Superset learning based on generalized loss minimization. In *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD, Porto, Portugal, September 7-11, Proc. Part II*, volume 9285 of *LNCS*, pages 260–275. Springer, 2015.

[Hül14] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.*, 55(7):1519–1534, 2014.

[Lee13] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, International Conference on Machine Learning, ICML, Atlanta, GA, USA, June 16-21*, volume 3, 2013.

[LH21a] Julian Lienen and Eyke Hüllermeier. Credal self-supervised learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-14*, pages 14370–14382, 2021.

[LH21b] Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In *Proc. of the 35th AAAI Conference on Artificial Intelligence, virtual, February 2-9*, 2021.

[RDRS21] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *Proc. of the 9th International Conference on Learning Representations, ICLR, virtual, May 3-7*. OpenReview.net, 2021.

[SBC+20] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS, virtual, December 6-12*, 2020.

[VGS05] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.