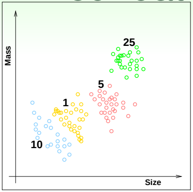
Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, Masashi Sugiyama

ICML 2020 work done during Chou's internship at RIKEN AIP, Japan;
resulting M.S. thesis of Chou won the 2020 thesis award of TAAI

October 30, 2021, SSC, Kaohsiung, Taiwan

Supervised Learning (Slide Modified from My ML Foundations MOOC)



unknown target function
$f : \mathcal{X} \to \mathcal{Y}$

training examples
$\mathcal{D} : (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$

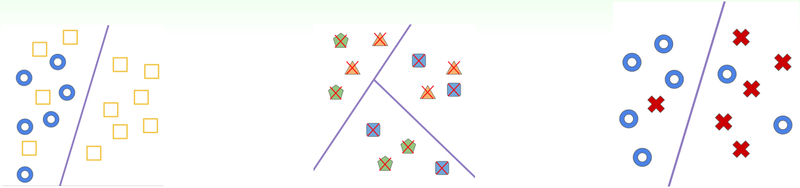
learning algorithm
$\mathcal{A}$

final hypothesis
$g \approx f$

hypothesis set
$\mathcal{H}$

supervised learning:
every input vector $\mathbf{x}_n$ with
its (possibly expensive) label y_n,

Weakly-supervised: Learning without True Labels y_n



(a) positive-unlabeled learning [EN08] (b) learning with complementary labels [Ish+17] (c) learning with noisy labels [Nat+13]

- positive-unlabeled: some of true $y_n = +1$ revealed
- complementary: ‘not label’ $\overline{y}_n$ instead of true y_n
- noisy: noisy label y'_n instead of true y_n

weakly-supervised: a **realistic** and **hot** research direction to reduce labeling burden

[EN08] Learning classifiers from only positive and unlabeled data, KDD’08.

[Ish+17] Learning from complementary labels, NeurIPS’17.

[Nat+13] Learning with noisy labels, NeurIPS’13.

Motivation

popular weakly-supervised models [DNS15; Ish+19; Pat+17]

- derive **Unbiased Risk Estimators (URE)** as new loss
- theoretically, nice properties (unbiased, consistent, etc.) [Ish+17]
- practically, **sometimes bad performance** (overfitting)

our contributions: on Learning w/ Complementary Labels (LCL)

- analysis: **identify weakness** of URE framework
- algorithm: propose an **improved framework**
- experiment: demonstrate **stronger performance**

next: introduction to LCL

[DNS15] Convex formulation for learning from positive and unlabeled data, ICML'15.

[Ish+19] Complementary-Label Learning for Arbitrary Losses and Models, ICML'19.

[Pat+17] Making deep neural networks robust to label noise: A loss correction approach, CVPR'17.

Motivation behind LCL

complementary label $\bar{y}_n$ instead of true y_n

Figure 1 of [Yu+18]

complementary label: **easier/cheaper** to obtain for some applications

Fruit Labeling Task (Image from AICup in 2020)

$\longrightarrow$

hard: true label

- orange **?**
- mango **?**
- cherry
- banana

easy: complementary label

- orange
- mango
- cherry
- banana ✗

complementary: **less labeling cost/expertise** required

Comparison

Ordinary (Supervised) Learning

training: $\{(\mathbf{x}_n =$ $, y_n = \text{mango})\} \rightarrow$ **classifier**

Complementary Learning

training: $\{(\mathbf{x}_n =$ $, \overline{y}_n = \text{banana})\} \rightarrow$ **classifier**

testing goal: **classifier**() $\rightarrow$ cherry

ordinary versus complementary:
same goal via **different training data**

Learning with Complementary Labels Setup

Given

N examples (input $\mathbf{x}_n$, complementary label $\overline{y}_n$) $\in \mathcal{X} \times \{1, 2, \cdots K\}$ in data set $\mathcal{D}$ such that $\overline{y}_n \neq y_n$ for some hidden ordinary label $y_n \in \{1, 2, \cdots K\}$.

Goal

a multi-class classifier $g(\mathbf{x})$ that **closely predicts** (0/1 error) the ordinary label y associated with some **unseen** inputs x

LCL model design: connecting
complementary & ordinary

Unbiased Risk Estimation for LCL

Ordinary Learning

- empirical risk minimization (ERM) on training data

risk: $\mathbb{E}_{(\mathbf{x},y)}[\ell(y, g(\mathbf{x}))]$ **empirical risk:** $\mathbb{E}_{(\mathbf{x}_n,y_n)\in\mathcal{D}}[\ell(y_n, g(\mathbf{x}_n))]$

- loss ℓ: usually **surrogate** of 0/1 error

LCL [Ish+19]

- rewrite the loss ℓ to $\bar{\ell}$, such that

unbiased risk estimator: $\mathbb{E}_{(\mathbf{x},\bar{y})}[\bar{\ell}(\bar{y}, g(\mathbf{x}))] = \mathbb{E}_{(\mathbf{x},y)}[\ell(y, g(\mathbf{x}))]$

- LCL by minimizing **URE**

URE: **pioneer models** for LCL

Example of URE

Cross Entropy Loss

for $g(\mathbf{x}) = \text{argmax}_{k \in \{1,2,\ldots,K\}} \boldsymbol{p}(k \mid \mathbf{x})$,

- ℓ_{CE}: derived by maximum likelihood as surrogate of 0/1

$$\textbf{risk:} \quad R(g; \ell_{CE}) = \mathbb{E}_{(\mathbf{x},y)} \underbrace{\left(-\log(\boldsymbol{p}(y \mid \mathbf{x}))\right)}_{\ell_{CE}}$$

Complementary Learning [Ish+19]

$$\textbf{URE:} \quad \overline{R}(g; \overline{\ell}) = \mathbb{E}_{(\mathbf{x},\overline{y})} \overbrace{\Big[\underbrace{(K-1)\log(\boldsymbol{p}(\overline{y} \mid \mathbf{x}))}_{\text{negative}} - \sum_{k=1}^{K} \log(\boldsymbol{p}(k \mid \mathbf{x})) \Big]}^{\overline{\ell}}$$

under uniform $\overline{y}$ assumption

ERM with URE: $\min_{\boldsymbol{p}} \overline{R}$ with $\mathbb{E}$ taken on $\mathcal{D}$

URE overfits on single label

$$\ell = -\log(\boldsymbol{p}(y \mid \mathbf{x}))$$

$$\bar{\ell} = (K-1)\log(\boldsymbol{p}(\bar{y} \mid \mathbf{x})) - \sum_{k=1}^{K} \log(\boldsymbol{p}(k \mid \mathbf{x}))$$

ordinary risk and URE very different

- $\ell > 0 \rightarrow$ ordinary risk non-negative
- small $\boldsymbol{p}(\bar{y} \mid \mathbf{x})$ (often) $\rightarrow$ possibly very negative $\bar{\ell}$
  empirical URE can be negative: observing **some but not all** $\bar{y}$
- negative empirical URE **drags minimization** towards overfitting

practical remedy: [Ish+19]

NN-URE: constrain emprical URE to be non-negative

how can we avoid negative empirical URE?

Proposed Framework

Minimize Complementary 0/1

- Recall the goal: minimize 0-1 loss, **not** ℓ
- The unbiased estimator of R_{01}

$$\overline{\boldsymbol{R}}_{\overline{\mathbf{01}}}: \quad \mathbb{E}_{\overline{y}}[\overline{\ell}_{01}(\overline{y}, g(\mathbf{x}))] = \ell_{01}(y, g(\mathbf{x}))$$

- We denote $\overline{\ell}_{01}$ as the complementary 0-1 loss:

$$\overline{\ell}_{01}(\overline{y}, g(\mathbf{x})) = [\![\overline{y} = g(\mathbf{x})]\!]$$

Surrogate Complementary Loss (SCL)

- Surrogate loss to optimize $\overline{\ell}_{01}$
- Unify previous work as surrogates of $\overline{\ell}_{01}$ [Yu+18; Kim+19]

[Yu+18] Learning with biased complementary labels, ECCV'18.

[Kim+19] NlNl: Negative learning for noisy labels, ICCV'19.

Negative Risk Avoided

Unbiased Risk Estimator (URE)

URE loss $\bar{\ell}_{CE}$ [Ish+19] from cross-entropy ℓ_{CE},

$$\bar{\ell}_{CE}(\bar{y}, g(\mathbf{x})) = \underbrace{(K-1)\log(\boldsymbol{p}(\bar{y} \mid \mathbf{x}))}_{\text{negative loss term}} - \sum_{j=1}^{K} \log(\boldsymbol{p}(j \mid \mathbf{x}))$$

can go negative.

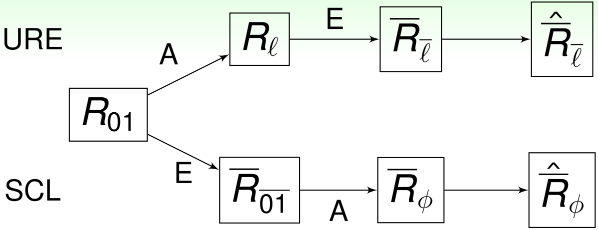
Surrogate Complementary Loss (SCL)

a surrogate of $\bar{\ell}_{01}$ [Kim+19]

$$\phi_{\mathrm{NL}}(\bar{y}, g(\mathbf{x})) = -\log(1 - \boldsymbol{p}(\bar{y} \mid \mathbf{x}))$$

remains non-negative.

Illustrative Difference between URE and SCE



URE: Ripple effect of errors

- Theoretical motivation [Ish+17]
- Estimation step (E) amplifies approximation error (A) in $\overline{\ell}$

SCL: 'Directly' minimize complementary likelihood

- Non-negative loss ϕ
- Practically prevents ripple effect

Classification Accuracy

Methods

1. Unbiased risk estimator (URE) [Ish+19]
2. Non-negative correction methods on URE (NN) [Ish+19]
3. Surrogate complementary loss (SCL)

Table: URE and NN are based on $\bar{\ell}$ rewritten from cross-entropy loss, while SCL is based on exponential loss $\phi_{\mathrm{EXP}}(\bar{y}, g(\mathbf{x})) = \exp(\boldsymbol{p}_{\bar{y}})$.

Data set + Model	URE	NN	SCL
MNIST + Linear	0.850	0.818	**0.902**
MNIST + MLP	0.801	0.867	**0.925**
CIFAR10 + ResNet	0.109	0.308	**0.492**
CIFAR10 + DenseNet	0.291	0.338	**0.544**

Gradient Analysis

Gradient Direction of URE

- Very diverse directions on each $\overline{y}$
- Low correlation to the target ℓ_{01}

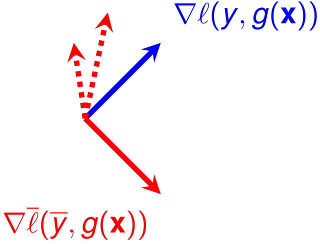


Figure: Illustration of URE

Gradient Direction of SCL

- Targets to minimum likelihood objective
- High correlation to the target $\bar{\ell}_{01}$

Gradient Estimation Error

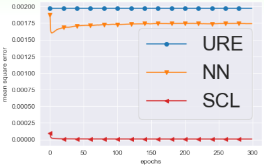
Bias-Variance Decomposition

$$\begin{aligned}\text{MSE} &= \mathbb{E}\left[(\boldsymbol{f} - \boldsymbol{c})^2\right] \\ &= \underbrace{\mathbb{E}\left[(\boldsymbol{f} - \boldsymbol{h})^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\boldsymbol{h} - \boldsymbol{c})^2\right]}_{\text{Variance}}\end{aligned}$$

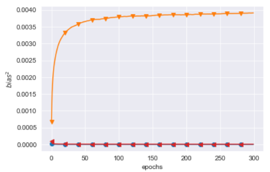
Gradient Estimation

1. Ordinary gradient $\boldsymbol{f} = \nabla \ell(y, g(\mathbf{x}))$
2. Complementary gradient $\boldsymbol{c} = \nabla \bar{\ell}(\bar{y}, g(\mathbf{x}))$
3. Expected complementary gradient $\boldsymbol{h}$

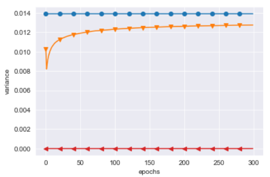
Bias-Variance Tradeoff



(a) MSE



(b) Bias^2



(c) Variance

Findings

- SCL reduces variance by introducing small bias (towards $\overline{y}$)

	Bias	Variance	MSE
URE	0	Big	Big
SCL	Small	Small	Small

Conclusion

Explain Overfitting of URE

- Unbiased methods only do well in expectation
- Single fixed complementary label cause overfitting

Surrogate Complementary Loss (SCL)

- Minimum likelihood principle
- Avoids negative risk issue

Experiment Results

- SCL significantly outperforms other methods
- Introduce small bias for lower gradient variance

Wait: Discussion for Theoreticians

minimize $\bar{\ell}_{0/1}$—hypothesis that **least matches** complementary data:

is this **minimum likelihood** principle well-justified? **Not yet.**

bias-variance decomposition of gradient based on **empirical findings**:

is there a theoretical guarantee to play with the trade-off? **Not yet.**

current results based on **uniform** complementary labels:

do we understand the assumptions to make LCL 'learnable'? **Not yet.**

Thank you!

References

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. “Convex formulation for learning from positive and unlabeled data”. In: *International Conference on Machine Learning*. 2015, pp. 1386–1394.

Charles Elkan and Keith Noto. “Learning classifiers from only positive and unlabeled data”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 213–220.

Takashi Ishida et al. “Learning from complementary labels”. In: *Advances in neural information processing systems*. 2017, pp. 5639–5649.

Takashi Ishida et al. “Complementary-Label Learning for Arbitrary Losses and Models”. In: *International Conference on Machine Learning*. 2019, pp. 2971–2980.

Youngdong Kim et al. “Nlnl: Negative learning for noisy labels”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 101–110.

Nagarajan Natarajan et al. “Learning with noisy labels”. In: *Advances in neural information processing systems*. 2013, pp. 1196–1204.

Vaishnavh Nagarajan and J Zico Kolter. “Uniform convergence may be unable to explain generalization in deep learning”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 11611–11622.

Giorgio Patrini et al. “Making deep neural networks robust to label noise: A loss correction approach”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1944–1952.

Xiyu Yu et al. “Learning with biased complementary labels”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 68–83.