Is Complementary-Label Learning Realistic?

Hsuan-Tien Lin

林軒田

Professor, National Taiwan University

September 25, 2024

Department of Mathematics, National University of Singapore

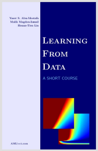
About Me

Hsuan-Tien Lin

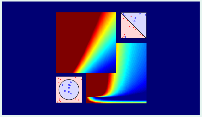
Professor
National Taiwan University

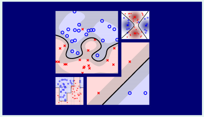
Chief Data Science Consultant
(former Chief Data Scientist)
Appier Inc.

Co-author
Learning from Data

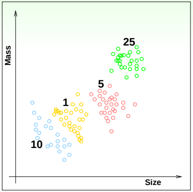


Instructor
NTU-Coursera Mandarin MOOCs
ML Foundations/Techniques





Supervised Learning
(Slide Modified from My ML Foundations MOOC)



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

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

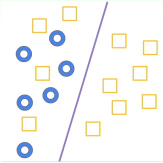
learning algorithm
$\mathcal{A}$

final hypothesis
$g \approx f$

hypothesis set
$\mathcal{H}$

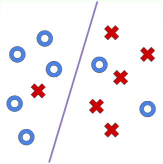
supervised learning:
every input vector $\mathbf{x}_n$ with
its (possibly expensive) label y_n,

Weakly-supervised: Learning without True y_n



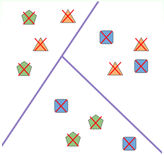
(a) Positive-unlabeled Learning [CE2008]

incomplete



(b) Learning with Noisy Labels [NN2013]

inaccurate



(c) Complementary-label Learning [TI2017]

inexact

- positive-unlabeled: some of true $y_n = +1$ revealed
- noisy: possibly incorrect label y'_n instead of true y_n
- complementary: false label $\overline{y}_n$ instead of true y_n

weakly-supervised: claimed to be a realistic route for reducing labeling burden

Complementary-Label Learning

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

Figure 1 of [XY2018]

potential to reducing labeling burden [TI2017]

- 1 ordinary label per instance
- $(K-1)$ complementary labels per instance, just need one of them

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

Example: Fruit Labeling Task

 → 

(left: from 2020 AICup in Taiwan; right: `publicdomainvectors.org`)

hard: true label

- orange ?
- mango ?
- cherry
- banana

easy: complementary label

- orange
- mango
- cherry
- banana ✗

can also help improve other ML tasks,
like semi-supervised learning [QD2024]

Comparison to Ordinary Learning

Ordinary (Supervised) Learning

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

Complementary-Label Learning

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

goal during testing: $\operatorname*{argmax}_{k \in [K]} \Big(g \Big($  $\Big)\Big)_k \rightarrow$ cherry

ordinary versus complementary:
same goal via different training data

Formal Setup of Complementary-Label Learning

input	complementary label
	banana

Given

size-N data $\mathcal{D} = \{(\text{input } \mathbf{x}_n \in \mathcal{X}, \text{complementary label } \overline{y}_n \in [K])\}_{n=1}^{N}$
such that $\overline{y}_n \neq y_n$ for some hidden ordinary label $y_n \in [K]$

Goal

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

todo: two CLL models, and more!

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. ICML 2020.

Review: Risk Minimization in Ordinary Learning

- goal: minimize the 0/1 loss

$$\ell_{01}(y, g(\mathbf{x})) = [\![y \neq \operatorname*{argmax}_{k \in [K]} (g(\mathbf{x}))_k]\!]$$

  with risk (average loss) $R_{01} = \mathbb{E}_{(\mathbf{x},y)} \{\ell_{01}(y, g(\mathbf{x}))\}$

- consider a surrogate loss ℓ that replaces ℓ_{01}

$$\ell \colon [K] \times \mathbb{R}^K \to \mathbb{R}_+$$

  with risk $R_\ell = \mathbb{E}_{(\mathbf{x},y)} \{\ell(y, g(\mathbf{x}))\}$

Empirical Risk Minimization (ERM):
estimate R_ℓ by training data and minimize it

Unbiased Risk Estimation for CLL

Ordinary Learning

- ERM: minimizes

$$\hat{R}_\ell = \mathop{\mathbb{E}}_{(\mathbf{x}_n, y_n) \in \mathcal{D}} \{\ell(y_n, g(\mathbf{x}_n))\},$$

the empirical version of the surrogate risk $R_\ell = \mathbb{E}_{(\mathbf{x}, y)} \{\ell(y, g(\mathbf{x}))\}$

Unbiased Risk Estimator for CLL [TI2019]

- [under assumption on $P(\overline{y}|y)$] rewrite ℓ to some $\overline{\ell}$ such that

$$\overline{R}_{\overline{\ell}} = \mathbb{E}_{(\mathbf{x}, \overline{y})} \overline{\ell}(\overline{y}, g(\mathbf{x})) = \mathbb{E}_{(\mathbf{x}, y)} \ell(y, g(\mathbf{x})) = R_\ell$$

- $\overline{R}_{\overline{\ell}}$ called unbiased risk estimator (URE)
- URE-CLL: minimize empirical version $\hat{\overline{R}}_{\overline{\ell}}$ of URE

URE-CLL: pioneer model for CLL, with theoretical guarantees like consistency

Example of URE-CLL

cross-entropy loss

for $g(\mathbf{x}) = \boldsymbol{p}(k \mid \mathbf{x})$,

- ℓ_{CE}: surrogate of ℓ_{01} derived by maximum likelihood, with risk

$$R_{CE} = \mathbb{E}_{(\mathbf{x},y)}\Big\{\underbrace{-\log \boldsymbol{p}(y \mid \mathbf{x})}_{\ell_{CE}}\Big\}$$

URE for cross-entropy loss [TI2019]

$$\overline{R}_{CE} = \mathbb{E}_{(\mathbf{x},\overline{y})}\Big\{\overbrace{(K-1)\log \boldsymbol{p}(\overline{y} \mid \mathbf{x}) - \sum_{k=1}^{K} \log \boldsymbol{p}(k \mid \mathbf{x})}^{\overline{\ell}}\Big\}$$

under uniform $\overline{y}$ (that $\neq y$) assumption

URE-CLL: $\min_{\boldsymbol{p}} \hat{\overline{R}}_{CE}$

Issue: URE-CLL Overfits Easily

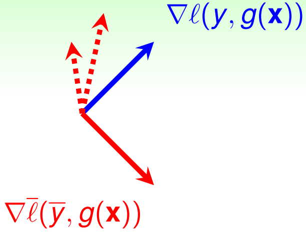
$$
\begin{aligned}
\ell_{CE} &= -\log \boldsymbol{p}(y \mid \mathbf{x}) \\
\bar{\ell}_{CE} &= \underbrace{(K-1)\log \boldsymbol{p}(\bar{y} \mid \mathbf{x})}_{\text{negative}} - \sum_{k=1}^{K} \log \boldsymbol{p}(k \mid \mathbf{x})
\end{aligned}
$$

ordinary risk and URE are very different

- $\ell > 0$: ordinary risk R non-negative
- often small $\boldsymbol{p}(\bar{y} \mid \mathbf{x})$: $\bar{\ell}$ often very negative
- empirically, negative $\hat{\bar{R}}_{\bar{\ell}}$
  —since only some $\bar{y}_n$ is observed
- observation: negative empirical URE $\rightarrow$ overfitting (but why?)

practical remedy NN-URE [TI2019]:
constrain empirical URE to be non-negative

Our Contributions



(to be discussed)

an analytical and algorithmic study of URE-CLL, which . . .

- constructs a novel loss-design framework
- results in promising empirical performance
- leads to novel insights light on why negative empirical URE causes overfitting

will first describe key idea
behind our proposed framework

Key Idea: URE on 0/1 instead of ℓ

Minimize Complementary 0/1

- goal: minimize R_{01}, not surrogate R_ℓ
- URE of R_{01}: need

$$\overline{R}_{\overline{01}} = \mathbb{E}_{(\mathbf{x},\overline{y})}\overline{\ell}_{01}(\overline{y}, g(\mathbf{x})) = \mathbb{E}_{(\mathbf{x},y)} \underbrace{\ell_{01}(y, g(\mathbf{x}))}_{[\![y \neq \mathrm{argmax}_k (g(\mathbf{x}))_k]\!]}$$

- simple solution:

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

- intuition: all we need is to discourage $g(\mathbf{x})$ from predicting $\overline{y}$ —minimum likelihood "principle"

Surrogate Complementary Loss (SCL): minimize (empirical) surrogate risk of $\overline{\ell}_{01}$

Illustrative Difference between URE and SCL

$$\begin{tikzcd}
\text{URE} & & \boxed{R_\ell} \arrow[r, "E"] & \boxed{\overline{R}_{\overline{\ell}}} \arrow[r] & \boxed{\hat{\overline{R}}_{\overline{\ell}}} \\
& \boxed{R_{01}} \arrow[ur, "A"] \arrow[dr, "E"'] & & & \\
\text{SCL} & & \boxed{\overline{R}_{01}} \arrow[r, "A"'] & \boxed{\overline{R}_\phi} \arrow[r] & \boxed{\hat{\overline{R}}_\phi}
\end{tikzcd}$$

URE: ripple effect of error

- theoretical motivation [TI2017]
- estimation step (E) amplifies approximation error (A) in $\overline{\ell}$

SCL: “directly” minimize complementary likelihood

- non-negative surrogate loss ϕ for $\overline{\ell}_{01}$ to be minimized
- potentially preventing ripple effect
- unify previous studies as different ϕ [XY2018, YK2019]

SCL: swapping (E) and (A) for loss design

Example of Avoiding Negative Risk

Unbiased Risk Estimator (URE)

URE loss $\bar{\ell}_{CE}$ [TI2019] from ℓ_{CE},

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

Surrogate Complementary Loss (SCL)

[YK2019]

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

—a non-negative surrogate of $\bar{\ell}_{01}$

SCL opens new possibilities
on studying different ϕ

Experimental Results

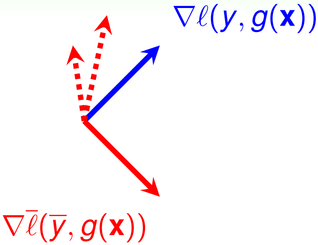
Models

1. Unbiased Risk Estimator (URE) with $\bar{\ell}_{CE}$ [TI2017]
2. Non-Negative Correction of URE (NN-URE) with $\bar{\ell}_{CE}$ [TI2019]
3. Surrogate Complementary Loss (SCL) with exponential ϕ (ours)

Dataset + Model	URE	NN-URE	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

SCL is significantly better
than URE and NN-URE

Analysis Using Gradients



Gradient Direction of URE

- very diverse directions on each $\bar{y}$ to maintain unbiasedness
- low correlation to the target ℓ_{01}

Gradient Direction of SCL

- targets towards minimum likelihood objective
- higher correlation to the target $\bar{\ell}_{01}$

will quantify this
with bias-variance decomposition

Gradient Estimation Error

Gradient Estimation

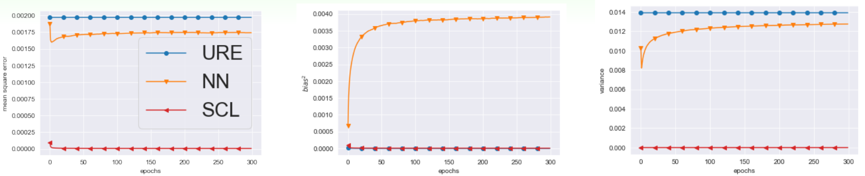
1. ordinary gradient $\boldsymbol{f} = \nabla \ell(y, g(\mathbf{x}))$
2. complementary gradient $\boldsymbol{c} = \nabla \bar{\ell}(\bar{y}, g(\mathbf{x}))$
3. expected complementary gradient $\boldsymbol{h}$ = average of $\boldsymbol{c}$ over $\bar{y}$

Bias-Variance Decomposition

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

is unbiased risk/gradient estimator good?

Bias-Variance Tradeoff on Gradient



(a) MSE (b) Bias^2 (c) Variance

	Bias	Variance	MSE
URE	0	Big	Big
NN-URE	Big	Smaller	Big
SCL	Small	Smallest	Small

SCL reduces variance of URE while introducing small bias

Some Issues for Mathematicians

minimize $\bar{\ell}_{01}$—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 mostly based on uniform complementary labels:

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

some (but not all) answered in the next paper

Mini-Summary

Explain Overfitting of URE

- URE only expected to do well
- fixed CLs cause high variance (hence overfitting)

Surrogate Complementary Loss (SCL)

- avoids negative risk issue by design
- minimum likelihood principle

Experiment Results

- SCL significantly outperforms others
- trade small gradient bias for lower variance

"traditional" statistics tools
can be useful for modern problem

Wei-I Lin and Hsuan-Tien Lin.
Reduction from complementary-label learning to probability estimates. PAKDD 2023
Best Paper Runner-up Award.

Reflection on CLL Model Design

$\boxed{\text{reduction to ordinary learning}}$

$$\text{CLL data} \xrightarrow[\text{special loss}]{\text{training}} \boxed{\text{ordinary classifier}} \xrightarrow{\text{inference}} \text{prediction}$$

Inference: Easy

simply $\text{argmax}_k (g(\mathbf{x}))_k$

Training: Challenging

- indirect estimation from CLs
- prone to overfitting
- mostly only tested on deep models

can we make training easier?

Our Contributions

$$R_{01}(\text{dec}(\overline{g}, L_1)) \le \frac{4\sqrt{2}}{\gamma}\sqrt{R(\overline{g}, \ell_{KL})}$$

(to be discussed)

a principled study of CLL Model Design, which ...

- promotes a novel reduction framework
- leads to sound explanations on several existing models
- results in promising empirical performance in some scenarios

again, will first describe key idea
behind our proposed framework

Key Idea: Complementary Probability Estimation

reduction to complementary probability estimation (CPE)

CLL data $\xrightarrow[\text{existing loss}]{\text{training}}$ complementary probability estimator $\xrightarrow{\text{inference}}$ prediction

Training: Easy

learn complementary probability estimates $\overline{g}(\mathbf{x})$ with CLs

- direct learning from CLs
- many existing deep/non-deep models
- easy to validate too

inference: how (under what assumption)?

Assumption: How are CLs Generated?

uniform assumption

$$P(\overline{y} \mid y) = \frac{1}{K-1}[\![\overline{y} \neq y]\!]$$

conditional generation assumption

$$P(\overline{y} \mid \mathbf{x}, y) = P(\overline{y} \mid y) = T_{y,\overline{y}}$$

e.g. transition matrix

$$T = \begin{bmatrix} 0 & 0.3 & 0.3 & 0.4 \\ 0.4 & 0 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0.4 & 0.3 & 0.3 & 0 \end{bmatrix}$$

how to do inference with known T after CPE?

Nearest Transition Vector Decoder

CLL data $\xrightarrow[\text{existing loss}]{\text{training}}$ complementary probability estimator $\xrightarrow{\text{inference}}$ prediction

$$T = \begin{bmatrix} 0 & 0.3 & 0.3 & 0.4 \\ 0.4 & 0 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0.4 & 0.3 & 0.3 & 0 \end{bmatrix}$$

looks like $y = 1$ if $\overline{g}(\mathbf{x}) = [0.03, 0.27, 0.25, 0.45]$

proposed nearest-transition-vector decoder for inference:

$$\text{dec}(\overline{g}, d) \colon \mathbf{x} \to \operatorname*{argmin}_{y \in [K]} d(\overline{g}(\mathbf{x}), T_y)$$

Theoretical Guarantee of CPE

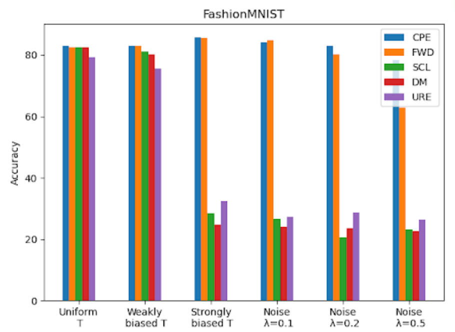
When using $d = L_1$ distance,

$$R_{01}(\text{dec}(\overline{g}, L_1)) \leq \frac{4\sqrt{2}}{\gamma}\sqrt{R_{KL}(\overline{g})}$$

- γ: minimum L_1 distance between rows of transition vectors
- smaller CPE error (KL divergence) $\rightarrow$ smaller R_{01}
- explains SCL as special case of L1 decoding under uniform assumption
- can be used to validate with CLs only

other distance measures possible
(but we did not study much)

Experimental Results



Models

1. Unbiased Risk Estimator (URE) [TI2017]
2. Discriminative model (DM*) [YG2021]
3. Surrogate Complementary Loss (SCL*, our previous work)
4. Forward (FWD*) [XY2018]
5. Complementary Probability Estimator (CPE, ours)

CPE better than others and special cases (*), especially with noisy T

Some Issues for Mathematicians Revisited

minimize $\bar{\ell}_{01}$—hypothesis that least matches complementary data:

is minimum likelihood principle well-justified? Yes, special case of CPE.

bias-variance decomposition of gradient based on empirical findings:

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

current results mostly based on uniform complementary labels:

the assumptions to make CLL 'learnable'? any known T with $\gamma > 0$.

some answered in the this paper

Mini-Summary

Explain SCL (and Others)

- via a different reduction route

Complementary Probability Estimation (CPE)

- estimate complementary probabilities during training (easy)
- nearest transition vector decoding (theoretical guarantees)

Experiment Results

- CPE outperforms (?) others
- potential for noisy CLL and CL-only validation

now, is CLL realistic?

Hsiu-Hsuan Wang, Tan-Ha Mai,
Nai-Xuan Ye, Wei-I Lin, Hsuan-Tien Lin.
CLImage: Human-Annotated Datasets for
Complementary-Label Learning. arXiv:2305.08295

Recall: Assumptions in CLL Model Design

noise-free assumption

$P(\bar{y} = y \mid y) = 0$

uniform assumption

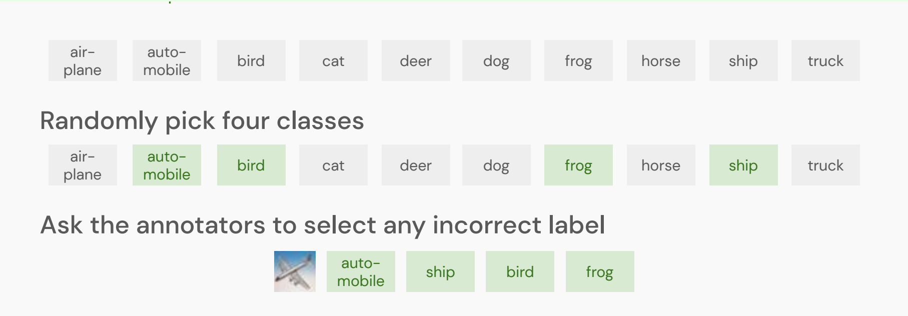
$$P(\bar{y} \mid y) = \frac{1}{K-1}[\![\bar{y} \neq y]\!]$$

conditional generation assumption

$$P(\bar{y} \mid \mathbf{x}, y) = P(\bar{y} \mid y) = T_{y,\bar{y}}$$

do they hold in reality?

CLImage: Protocol for Collecting CL from Annotators



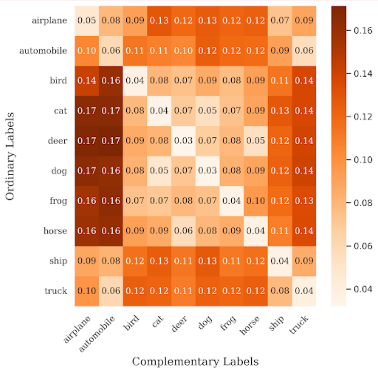
(courtesy of Wei-I Lin)

play here: `https://github.com/ntucllab/CLImage_Dataset/`

Analysis of Collected Data

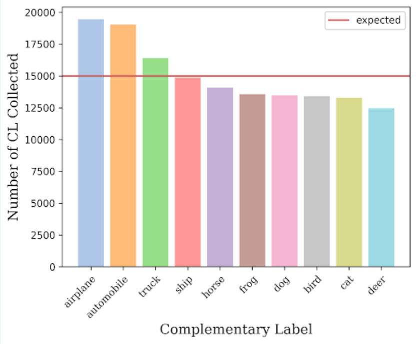
is it noise-free?

no (not surprisingly), and it affects performance significantly



is it uniform?

no (not surprisingly), and it affects performance a bit



more studies on noisy CLL is needed

An Insider Secret

CLImage

- CLCIFAR10
- CLCIFAR20 (20 meta-classes)
- CLMicroImageNet10 (10 random classes)
- CLMicroImageNet20 (20 random classes)

–why only data of 10 or 20 classes?

Truth

tried CIFAR100 but failed

- higher accuracy than random guess
- much lower than ordinary classification, even after noise cleaning

pure CLL overly weak and may not be realistic

Summary (Finally)

Surrogate Complementary Loss

run URE before doing surrogate instead

Complementary Probability Estimation

consider probability estimation on CLs instead

CLImage

attempt to benchmark how realistic CLL is, with a dataset collection and a library in its beta version

`https://github.com/ntucllab/libcll`

Thank you and all my students/collaborators!

References

- [CE2008] Learning classifiers from only positive and unlabeled data, KDD 2008
- [NN2013] Learning with noisy labels, NeurIPS 2013
- [TI2017] Learning from complementary labels, NeurIPS 2017
- [XY2018] Learning with biased complementary labels, ECCV 2018
- [TI2019] Complementary-Label Learning for Arbitrary Losses and Models, ICML 2019
- [YK2019] NLNL: Negative learning for noisy labels, ICCV 2019
- [YG2021] Discriminative complementary-label learning with weighted loss, ICML 2021
- [QD2024] Boosting Semi-Supervised Learning with Contrastive Complementary Labeling, Neural Networks 2024