Learning with Multiple Complementary Labels

Lei Feng[1*], Takuo Kaneko[2,3*], Bo Han[3,4], Gang Niu[3], Bo An[1], Masashi Sugiyama[2,3]

[1]Nanyang Technological University, Singapore
[2]The University of Tokyo, Tokyo, Japan
[3]RIKEN Center for Advanced Intelligent Project, Tokyo, Japan
[4]Hong Kong Baptist University, Hong Kong SAR, China
[*]Equal Contribution

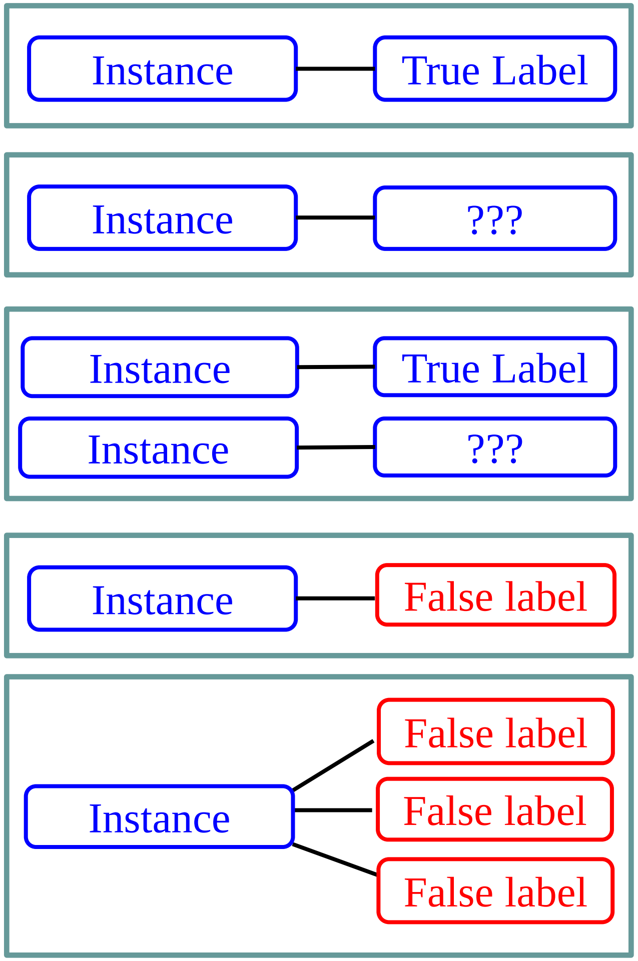
ICML 2020

Outline

- Learning Frameworks
- Problem Formulation
- The Proposed Methods
  - Wrappers
  - Unbiased Risk Estimator
  - Upper-Bound Surrogate Losses
- Experiments
- Conclusion

Learning Frameworks

- Supervised Learning:
- Unsupervised Learning:
- Semi-Supervised Learning [Chapelle et al., 2006]:
- Complementary-Label Learning [Ishida et al., 2017;2019]:
- **Learning with Multiple Complementary Labels (our paper):**



Instance — True Label

Instance — ???

Instance — True Label
Instance — ???

Instance — False label

Instance — False label, False label, False label

Data Distribution

For complementary-label (CL) learning [Ishida et al., 2017; 2019]:

$$\bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{k-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y).$$

For learning with multiple complementary labels (MCLs):

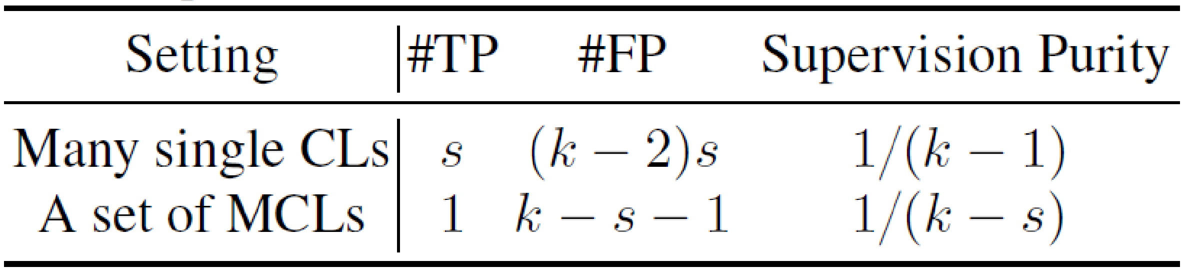
$$\bar{p}(x, \bar{Y}) = \sum_{j=1}^{k-1} p(s = j) \bar{p}(\boldsymbol{x}, \bar{Y} \mid s = j),$$

where

$$\bar{p}(\boldsymbol{x}, \bar{Y} \mid s = j) := \begin{cases} \frac{1}{\binom{k-1}{j}} \sum_{j \notin \bar{Y}} p(\boldsymbol{x}, y) & \text{if } |\bar{Y}| = j, \\ 0 & \text{otherwise.} \end{cases}$$

- k: the number of classes
- $\bar{p}(\boldsymbol{x}, \bar{y})$: joint distribution with a single CL
- $p(\boldsymbol{x}, y)$: joint distribution with a single true label
- $\bar{p}(\boldsymbol{x}, \bar{Y})$: joint distribution with MCLs
- $p(s = j)$: the probability of the size of the set of MCLs being j

Wrappers



Setting	#TP	#FP	Supervision Purity
Many single CLs	s	$(k-2)s$	$1/(k-1)$
A set of MCLs	1	$k-s-1$	$1/(k-s)$

- #TP: how many times the correct label serves as a non-complementary label for each instance
- #FP: how many times the other labels except the correct label serve as a non-complementary label for each instance
- Supervision Purity: #TP/(#TP+#FP)

Decomposing a set of MCLs into many single CLs: Decomposition after Shuffle/Decomposition before Shuffle.

E.g., suppose $\bar{Y} = (\bar{y}_1, \bar{y}_2), (x, \bar{Y})$ is decomposed into $(x, \bar{y}_1)$ and $(x, \bar{y}_2)$.

Using the wrappers, we can apply any existing complementary-label learning methods. However, the supervision purity would be diluted after decomposition, as shown in the above table.

Unbiased Risk Estimator

The classification risk can be equivalently expressed as

$$R(f) = \sum_{j=1}^{k-1} p(s=j)\bar{R}_j(f),$$

where

$$\bar{R}_j(f) \coloneqq \mathbb{E}_{\bar{p}(x,\bar{Y} \mid s=j)}[\bar{\mathcal{L}}_j(f(x),\bar{Y})],$$

and

$$\bar{\mathcal{L}}_j(f(x),\bar{Y}) \coloneqq \sum_{y\notin\bar{Y}} \mathcal{L}(f(x),y) - \frac{k-1-j}{j}\sum_{y'\notin\bar{Y}} \mathcal{L}(f(x),y').$$

- $R(f)$: the classification risk defined as $\mathbb{E}_{p(x,y)}[\mathcal{L}(f(x),y)]$
- $\mathcal{L}(f(x),y)$: multi-class loss function

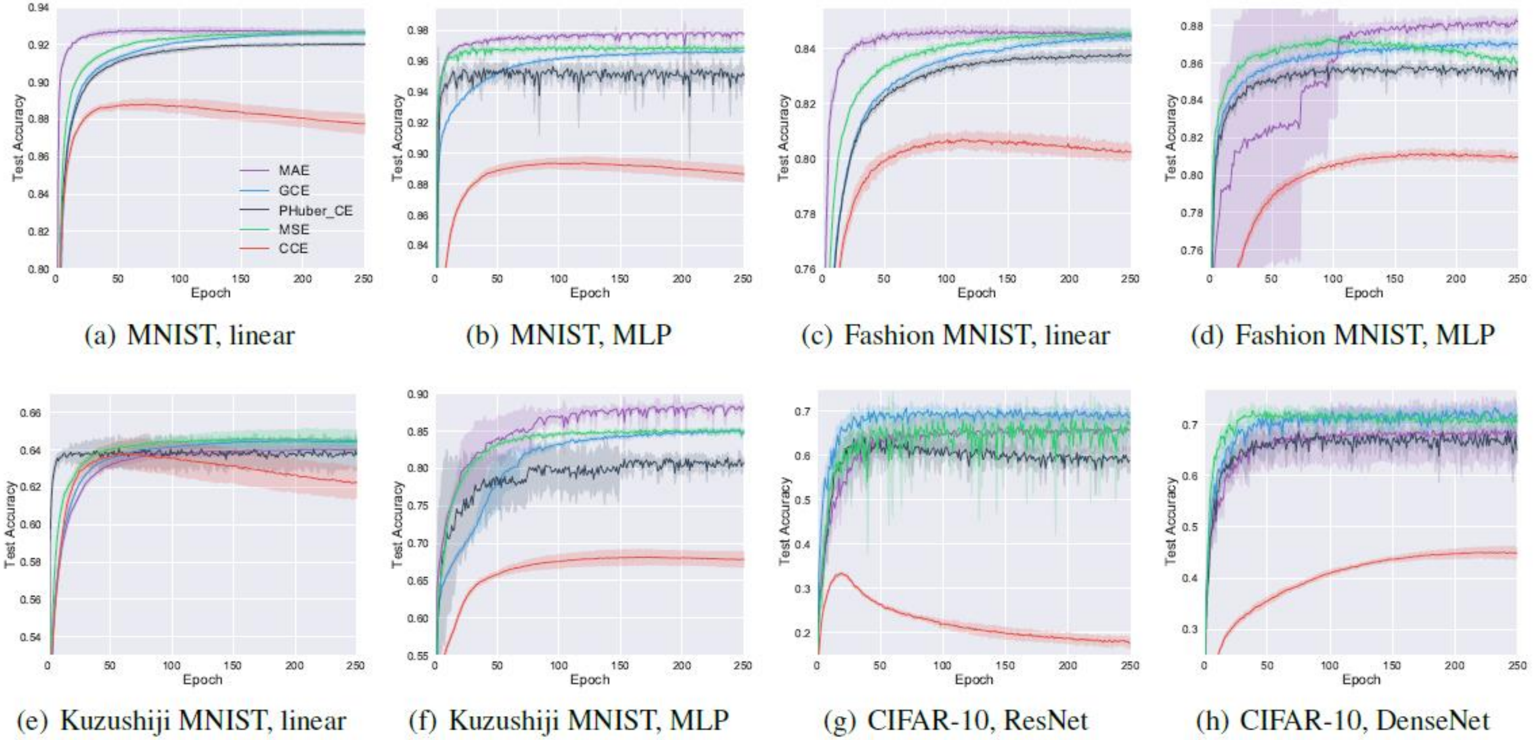
Each set of MCLs is taken as a whole!

Practical Implementation

Observation: The empirical risk estimator may become unbounded below if the used loss function is unbounded, thereby leading to over-fitting.

Conjecture: Bounded loss is better than unbounded loss.

Results: We validate via experiments that MAE, MSE, GCE [Zhang & Sabuncu, 2018], and Phuber-CE [Menon et al., 2020] outperform CCE.



(a) MNIST, linear

(b) MNIST, MLP

(c) Fashion MNIST, linear

(d) Fashion MNIST, MLP

(e) Kuzushiji MNIST, linear

(f) Kuzushiji MNIST, MLP

(g) CIFAR-10, ResNet

(h) CIFAR-10, DenseNet

Is Bounded Loss Good Enough?

Is the performance of the unbiased risk estimator with bounded loss good enough?

We take MAE for example, and insert MAE into the empirical risk estimator, and obtain an equivalent formulation as

$$\mathcal{L}'_{\mathrm{MAE}}(f(\boldsymbol{x}_i), \bar{Y}_i) = 1 - \sum_{j \notin \bar{Y}_i} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i),$$

Its gradient is expressed as

$$\frac{\partial \mathcal{L}'_{\mathrm{MAE}}}{\partial \boldsymbol{\theta}} = \begin{cases} -\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i) \cdot 1 & \text{if } j \notin \bar{Y}_i, \\ 0 & \text{otherwise.} \end{cases}$$

Each example is treated equally important for optimization.

Upper-Bound Surrogate Losses

We propose the following upper-bound surrogate losses:

$$\mathcal{L}_{\mathrm{EXP}}(f(\boldsymbol{x}_i), \bar{Y}_i) = \exp(-\textstyle\sum_{j\notin\bar{Y}_i} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i)),$$

$$\mathcal{L}_{\mathrm{LOG}}(f(\boldsymbol{x}_i), \bar{Y}_i) = -\log(\textstyle\sum_{j\notin\bar{Y}_i} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i)).$$

Their gradient can be expressed as

$$\frac{\partial \mathcal{L}_{\mathrm{EXP}}}{\partial \boldsymbol{\theta}} = \begin{cases} -\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i) \cdot w_{\mathrm{EXP}} & \text{if } j \notin \bar{Y}_i, \\ 0 & \text{otherwise,} \end{cases}$$

$$\frac{\partial \mathcal{L}_{\mathrm{LOG}}}{\partial \boldsymbol{\theta}} = \begin{cases} -\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i) \cdot w_{\mathrm{LOG}} & \text{if } j \notin \bar{Y}_i, \\ 0 & \text{otherwise,} \end{cases}$$

where $w_{\mathrm{EXP}} = \exp(-\sum_{j\notin\bar{Y}_i} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i))$ and $w_{\mathrm{LOG}} = (\sum_{j\notin\bar{Y}_i} p_{\boldsymbol{\theta}}(j \mid \boldsymbol{x}_i))^{-\mathbf{1}}$.

Higher weights will be given to hard examples!

Experiments

- Benchmark datasets: MNIST, Kuzushiji-MNIST, Fashion-MNIST, CIFAR-10.
- UCI datasets: Yeast, Texture, Dermatology, Synthetic Control, 20Newsgroups.
- Compared methods: GA, NN, and Free [Ishida et al., 2019], PC [Ishida et al., 2017], Forward [Yu et al., 2018], CLPL [Cour et al., 2011], unbiased risk estimator with bounded losses MAE, MSE, GCE [Zhang & Sabuncu, 2018], and PHuber-CE (Menon et al., 2020) and unbounded loss CCE, and the two upper-bound surrogate losses EXP and LOG.

Extensive experimental results clearly demonstrate the effectiveness of our proposed methods.

Conclusion

❑ A novel problem setting that generalizes learning with a single CL to learning with MCLs.

❑ Solutions including the wrappers and an unbiased risk estimator.

❑ Upper-bound surrogate losses.

Thank you!