Trustworthy Machine Learning under Noisy Data

Dr. Bo Han
HKBU TMLR Group / RIKEN AIP Team
Assistant Professor / BAIHO Visiting Scientist

https://bhanml.github.io/
Overview of This Tutorial

• Part I: Why and What Noisy Labels
• Part II: Current Progress and Tutorial Perspectives
• Part III: Training Perspective
• Part IV: Data Perspective
• Part V: Regularization Perspective
• Part VI: Future Directions

Part I: Why Noisy Labels

(Credit to Amazon)  (Credit to Google)

Why Noisy Labels

(Credit to Clothing1M)

(Credit to Outlook)

What are Noisy Labels

\[ R_{\ell,D}(f_0) := E_{(x,y) \sim D}[\ell(f_0(x), y)] \]

\[ \hat{R}_{\ell,D}(f_0) := \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}(f_0(x_i), \tilde{y}_i) \]

(Credit to Dr. Gang Niu)

Part II: Current Progress

B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama.  
(Not orthogonal fully)

Tutorial Perspectives

Training on Selected Samples

Algorithm 1 General procedure on using sample selection to combat noisy labels.

1: for $t = 0, \ldots, T - 1$ do  
2: draw a mini-batch $\tilde{D}$ from $\mathcal{D}$;  
3: select $R(t)$ small-loss samples $\tilde{D}_f$ from $\tilde{D}$ based on network’s predictions;  
4: update network parameter using $\tilde{D}_f$;  
5: end for
Self-teaching (MentorNet, 2018)

M-Net

Error accumulation!


Co-teaching (2018)


Find “bugs” by peers

Divergence Matters

Co-teaching+ (2019)


Test accuracy depends on selecting rules.

\[ R(t) = 1 - \tau \cdot \min((t/t_k)^c, 1) \]
DivideMix (2020)

Co-teaching + Semi-supervised Learning

Co-refinement and Co-guessing

MentorMix (2020)

Weight $\rightarrow$ Sample $\rightarrow$ Mixup $\rightarrow$ Weight


The estimation for the noisy class posterior is unstable

• Uncertainty about small loss: adopting interval estimation instead of point estimation

\[
\bar{\ell} = \frac{1}{t} \sum_t \phi(\ell_i)
\]
reduce the effect of extreme values, e.g., exponential function

• Uncertainty about large loss: large loss data also have the possibility to be selected.

\[
\ell^* = \bar{\ell} - f(n_t)
\]

\(n_t\) is the number of selected times, \(f\) is a decreasing function
UniCon (2022)

Selected clean set suffers from data imbalance

**Uniform Selection**: enforce the class-balance prior by selecting equal number of clean data per class.

**SSL Training**: contrastive learning on un-selected noisy data.

CoDis (2023)

Model divergence should be maintained to prevent two networks from convergence.

\[ \ell(p_1(x_i), \tilde{y}_i) - \alpha \ast \text{JS}(p_1(x_i) \parallel p_2(x_i)) \]

Small-loss data should be selected

High discrepancy data should be selected

Trade-off between small loss and high discrepancy

Summary

• **Memorization effect** in deep learning is new and important.

• MentorNet and Co-teaching series are developed.

• Many **applications** have leveraged Co-teaching series.

Part IV: Data Perspective

Noise Transition Matrix

Adaptation Layer (2017)

Forward Correction (2017)

(Credit to Dr. Tongliang Liu)

Correct the loss function to offset the impact of label noise

Masking (2018)


Fine-tuning (2019)

Parts-dependent (2020)

the weighted combination of the transition matrices for the parts of the instance

Wrong estimation of noise posterior deteriorates transition matrix estimation.

\[
T_{ij} = P(\tilde{Y} = j|Y = i) = \sum_l P(\tilde{Y} = j|Y' = l, Y = i) P(Y' = l|Y = i)
\]

Introduce an **intermediate class** \(Y'\) to avoid directly estimating the noisy class posterior.

VolMinNet (2021)

Without anchor points, the transition matrix is hard to be estimated.

Among all simplexes that enclose $P(\tilde{Y}|X)$, the one with minimum volume is the optimal.

Cluster-dependent Transition: data belong to different clusters have different transition matrix.

Meta Extended Transition: \((c + 1) \times c\) transition matrix \(T^*\), where the extra \(1 \times c\) vector \(T^\circ\) represent the open-set class.
Constrain the transition matrix in the Dirichlet space
A good transition matrix should simultaneously lead to the optimal forward correction loss and the noise robust loss.

$$\min_T L_{rob}(f_{\hat{\theta}(T)}, \bar{D}_v) \text{ s.t.} \hat{\theta}(T) = \arg\min L(T f_\theta, \bar{D}_{tr})$$

Less estimation error than MGEO
Summary

- **Noise transition matrix** is the key in data perspective.

- A potential direction is how to estimate this matrix **easily**.

- Another potential direction is how to leverage this matrix **effectively**.

Part V: Regularization Perspective

(Credit to Analytics Vidhya)

Bootstrapping (2015)


\[
\ell_{\text{soft}}(q, t) = \sum_{k=1}^{L} \left[ \beta t_k + (1 - \beta) q_k \right] \log(q_k)
\]

\[
\ell_{\text{hard}}(q, t) = \sum_{k=1}^{L} \left[ \beta t_k + (1 - \beta) z_k \right] \log(q_k)
\]

Interpolate between noisy targets and model prediction.
Mixup (2018)


MixMatch & FixMatch (2019&20)

SIGUA (2020)


Algorithm 1 SIGUA-prototype (in a mini-batch).

| Require: base learning algorithm $\mathcal{B}$, optimizer $\mathcal{O}$, mini-batch $\mathcal{S}_b = \{(x_i, \tilde{y}_i)\}_{i=1}^{n_b}$ of batch size $n_b$, current model $f_\theta$ where $\theta$ holds the parameters of $f$, good- and bad-data conditions $c_{\text{good}}$ and $c_{\text{bad}}$ for $\mathcal{B}$, underweight parameter $\gamma$ such that $0 \leq \gamma \leq 1$
| for $i = 1, \ldots, n_b$, do
| if $c_{\text{good}}(x_i, \tilde{y}_i)$ then
| $\ell_{b} \leftarrow \ell_{b} + \ell_i$ # accumulate loss positively
| else if $c_{\text{bad}}(x_i, \tilde{y}_i)$ then
| $\ell_{b} \leftarrow \ell_{b} - \gamma\ell_i$ # accumulate loss negatively
| end if
| # ignore any uncertain data
| end for
| $\ell_b \leftarrow \ell_b / n_b$ # average accumulated loss
| $\nabla_\theta \leftarrow \mathcal{B}.\text{backward}(f_\theta, \ell_b)$ # backward pass
| $\mathcal{O}.\text{step}(\nabla_\theta)$ # update model

Gradient Ascent

Setting: Both the class and the feature distributions have biases between labelled and unlabelled datasets.

First detecting data in the shared class set, then conducting domain adaptation via adversarial generation.
The consistency of forward/backward correction can better regularize models in against label noise.
Which one is better, SSL or transition matrix?

(a) $P(x)$ contains information of labelling, thus modeling label noise is better

(b) $P(x)$ contains no information of labelling, thus SSL is better

The causal structure can be detected intuitively

Summary

• Regularization is very popular for **semi-supervised learning**.

• Explicit regularization is in the level of **objective function**.

• Implicit regularization is in the level of **algorithm** and **data**.
B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama.
Instance-dependent LNRL

(a) Class-conditional noise.
(b) Instance-dependent noise (boundary-consistent noise).
(c) Confidence-scored instance-dependent noise.
Confidence Score: \( r_x = P(Y = \bar{y} | \bar{Y} = y, X = x) \)
UPM (2021)


**PGM:**

\[ P(\tilde{y}|y, x) = (1 - \eta)I\{y = \tilde{y}\} + \eta \phi \]

\[ \phi = P(\tilde{y}|x) \text{ and } \eta = P(s = 1|x) \]

Noisy label distribution possibility to make confusion
CausalNL (2021)

Instance modelling helps transition matrix estimate

Instance-dependent confidence threshold:

\[ \tau(x) = T_{k,k}(x)P(y = s|x) + \sum T_{i,k}(x)P(y = i|x) \]

Adversarial LNRL

Noisy Feature


Noisy Domain


Noisy Similarity

(a) Supervised Classification  (b) SU Classification  (c) NSU Classification

S. Wu et al. Learning from Noisy Pairwise Similarity and Unlabeled Data. JMLR, 2022.


Noisy Demonstration


Noisy Machine Translation

<table>
<thead>
<tr>
<th></th>
<th>Der Elektroden Schalter KARI EL22 dient zur Füllstandserfassung und -regelung von elektrisch leitfähigen Flüssigkeiten.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Src:</strong></td>
<td>The KARI EL22 electrode switch is designed for the control of conductive liquids.</td>
</tr>
<tr>
<td><strong>Tgt:</strong></td>
<td>The electrode switch KARI EL22 is used for level detection and control of electrically conductive liquids.</td>
</tr>
<tr>
<td><strong>Human:</strong></td>
<td></td>
</tr>
</tbody>
</table>

Noisy Prompt


Noisy Model

noisy data hurt pre-trained models

Datasets and Benchmark


Conclusions

• Current progress mainly focuses on class-conditional noise.

• The new trend focuses on instance-dependent noise.

• Besides noisy labels, we should pay more efforts on noisy data.

B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama.
Appendix

• Survey:

• Book:

• Tutorial:
  • IJCAI 2021 Tutorial on Learning with Noisy Supervision
  • CIKM 2022 Tutorial on Learning and Mining with Noisy Labels
  • ACML 2023 Tutorial on Trustworthy Learning under Imperfect Data
  • AAAI 2024 Tutorial on Trustworthy Machine Learning under Imperfect Data

• Workshops:
  • IJCAI 2021 Workshop on Weakly Supervised Representation Learning
  • ACML 2022 Workshop on Weakly Supervised Learning
  • RIKEN 2023 Workshop on Weakly Supervised Learning
  • HKBU-RIKEN 2024 Joint Workshop on Artificial Intelligence and Machine Learning