ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABELS

ICLR 2024

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

- Problem / objective
  - Noisy label 들이 DNN 학습 방해
- Contribution / Key idea
  - Shifted gaussian label noise model 을 통해 noisy label 의 방해 최소화

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABELS

ICLR 2024

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Ours 등장 배경

[문제]

Noisy label 들이 DNN 학습 방해

[해결]

1. Loss reweighting -> noisy label 샘플들이 loss 에 주는 기여도 줄임
2. Label correction -> noisy label 을 모델이 예측하는 true label 로 수정

[Ours]

기존에 이 문제를 다룬 논문들은 각 해결법들을 독립적으로 적용했는데 우리는 unified approach를 제안함

[Ours 등장 배경]

Q. 모델이 예측한 레이블이 true label 이라고 확신할 수 있겠어?

학습 초반 : 학습 부족으로 모델의 예측 불안정하여 믿을 수 없음.

학습 후반 : 이미 모델이 noisy label 들에 오버핏 되어있어서 믿을 수 없음.

A. 학습 도중 모델이 예측한 레이블이 true label 이라고 믿을 수 있는 시간은 굉장히 짧다.

Ours. Loss reweighting 을 통해 noisy label 들에 오버핏 되는 것을 최대한 지연시킴으로서 label correction 하기 위한 충분한 시간을 확보하겠다.

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABELS

ICLR 2024

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Preliminary - Compositional data

1. Compositional data 란? 각 데이터의 상대적 정보에 초점

2. Compositional data 의 특징 : 각 compositions 들이 제약 변수임.

- D-1 차원 probability simplex

$$\Delta^{D-1} = \left\{ \mathbf{p} = (p_1, p_2, \ldots, p_D) \in \mathbb{R}^D \;\middle|\; p_i \geq 0 \, \forall i, \; \sum_{i=1}^{D} p_i = 1 \right\}$$

(vs D 차원 Euclidean space)

$$\mathbb{R}^D = \{ \mathbf{x} = (x_1, x_2, \ldots, x_D) \mid x_i \in \mathbb{R}, \forall i = 1, 2, \ldots, D \}$$

3. 일반적 통계 기법들은 비제약 변수에 적합하도록 설계되어 있음.

4. 결론 : Compositional data 에 일반적 통계 기법들을 바로 적용할 수 없다.

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABELS

ICLR 2024

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Preliminary - Log-Ratio Transform

1. Compositional data 에 통계 기법 적용하기 위해 Log-ratio transform 사용

1) Log-ratio transform 을 통해 비제약 공간으로 나와서

2) 통계 기법 적용하고

3) Inverse log-ratio transform 을 통해 원래 있던 제약 공간으로 돌아가는 방식.

2. Log-ratio transform 종류 3가지.

1) additive log-ratio (alr) transform $alr(p) = \log(\left[p_1, p_2, \ldots, p_{D-1}\right]/p_D)$: 비대칭성 문제

2) centered log-ratio (clr) transform $clr(p) = \log(\left[p_1, p_2, \ldots, p_D\right]/g(p))$: $\mathbb{R}^D$ 의 D-1 차원 hyperplane (합이 0) 문제

3) isometric log-ratio (ilr) transform $ilr(p) = V^T clr(p)$: $\mathbb{R}^{D-1}$

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABELS

ICLR 2024

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Overview

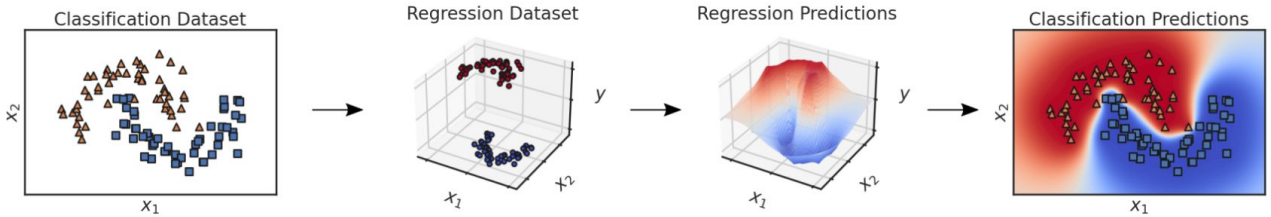
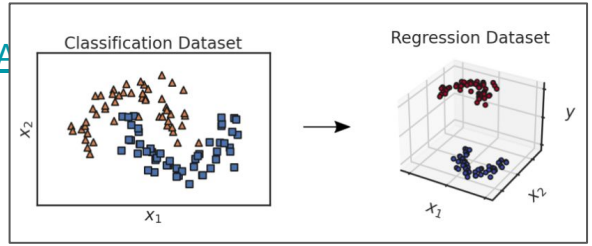


Figure 1: **Method Overview.** Our method is a three-step process: i) classification labels are transformed to regression labels, ii) the regression task is robustly solved with our loss reweighting and label correction method, iii) regression predictions are transformed to classification predictions.

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LA

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology



Step1. Classification dataset 을 Regression dataset 으로 변환

$$\mathcal{D}_\eta^{class} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \rightarrow \mathcal{D}_\eta^{comp} = \{(\boldsymbol{x}_i, LS(y_i))\}_{i=1}^N \rightarrow \mathcal{D}_\eta^{reg} = \{(\boldsymbol{x}_i, ilr(LS(y_i)))\}_{i=1}^N \quad (1)$$

Class ID to Simplex Δ^{K-1}

Class ID to Interior Simplex Δ^{K-1}

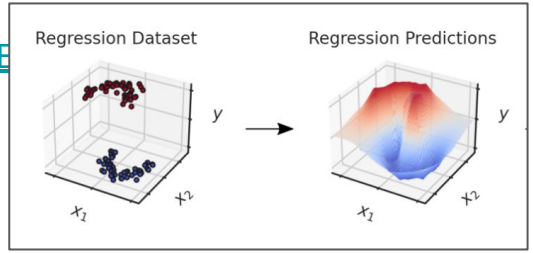
Class ID to Euclidean space $\mathbb{R}^{K-1}$

$$LS(y) = (1-\gamma)\boldsymbol{\delta}_y + \gamma \boldsymbol{u} \quad (4)$$

$$\boldsymbol{\delta}_y \in \Delta^{K-1}$$

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABE

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology



Step2. Regression 알고리즘 진행 (Loss reweighting & Label correction)

$$\underbrace{ilr(LS(y_i))}_{\mathbb{R}^{K-1}} = \boldsymbol{t}(\boldsymbol{x}_i) = \boldsymbol{\mu}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}(\boldsymbol{x}_i), \quad \boldsymbol{\epsilon}(\boldsymbol{x}_i) \sim \mathcal{N}(\boldsymbol{\Delta}(\boldsymbol{x}_i), \boldsymbol{\Sigma}(\boldsymbol{x}_i)) \tag{2}$$

1. Gaussian Noise Model 을 통해 loss reweighting (Maximum Log likelihood Estimation)

$$\arg\min_{\boldsymbol{\theta}} \underbrace{-\sum_{i=1}^{N} \log \mathcal{N}(t_i; \mu_{\boldsymbol{\theta}}(x_i), \sigma^2_{\boldsymbol{\theta}}(x_i))}_{\text{log likelihood}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \frac{(t_i - \mu_{\boldsymbol{\theta}}(x_i))^2}{2\sigma^2_{\boldsymbol{\theta}}(x_i)} + \frac{1}{2} \log \sigma^2_{\boldsymbol{\theta}}(x_i) \tag{5}$$

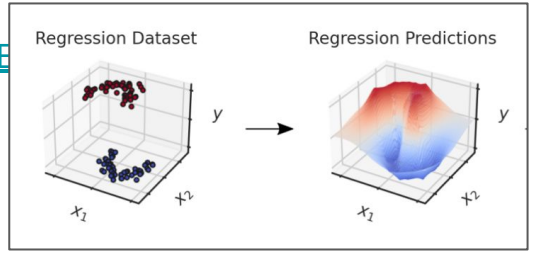
2. Shifted Gaussian Noise Model 을 통해 label correction

$$\begin{aligned} -\log \mathcal{N}(t; \mu_{\boldsymbol{\theta}} + \Delta, \sigma^2_{\boldsymbol{\theta}}) &= \frac{(t - (\mu_{\boldsymbol{\theta}} + \Delta))^2}{2\sigma^2_{\boldsymbol{\theta}}} + \frac{1}{2} \log \sigma^2_{\boldsymbol{\theta}} \\ &= \frac{((t - \Delta) - \mu_{\boldsymbol{\theta}})^2}{2\sigma^2_{\boldsymbol{\theta}}} + \frac{1}{2} \log \sigma^2_{\boldsymbol{\theta}} \\ &= -\log \mathcal{N}(t - \Delta; \mu_{\boldsymbol{\theta}}, \sigma^2_{\boldsymbol{\theta}}) \end{aligned} \tag{6}$$

$$\Delta = \underbrace{t}_{\text{noisy target}} - \underbrace{\mu}_{\text{gt label}}$$

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LABE

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology



Step2. Regression 알고리즘 진행 (Loss reweighting & Label correction)

$$\underbrace{ilr(LS(y_i))}_{\mathbb{R}^{K-1}} = \boldsymbol{t}(\boldsymbol{x}_i) = \boldsymbol{\mu}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}(\boldsymbol{x}_i), \quad \boldsymbol{\epsilon}(\boldsymbol{x}_i) \sim \mathcal{N}(\boldsymbol{\Delta}(\boldsymbol{x}_i), \boldsymbol{\Sigma}(\boldsymbol{x}_i)) \qquad (2)$$

- Label correction with Δ

$$\Delta = t - \mu$$

gt label (알수없음) (간주)

$$\Delta_{\bar{\boldsymbol{\theta}}}(x_i) = t_i - \mu_{\bar{\boldsymbol{\theta}}}(x_i)$$

EMA model's estimated gt label

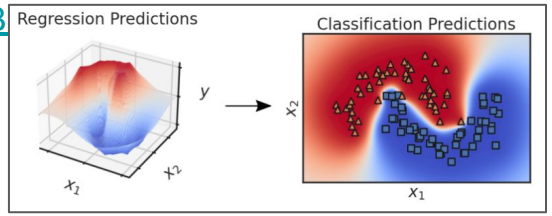
$$t = \mu + (1 - \alpha^e)\Delta$$

결론. 레이블이 observed target t 에서 estimated ground-truth $\mu_{\bar{\boldsymbol{\theta}}}$ 로 부드럽게 변화됨

$$\log \mathcal{N}(t_i; \mu_{\boldsymbol{\theta}}(x_i), \sigma^2_{\boldsymbol{\theta}}(x_i)) \longrightarrow \log \mathcal{N}(t; \mu_{\boldsymbol{\theta}} + \Delta_{\bar{\boldsymbol{\theta}}}, \sigma^2_{\boldsymbol{\theta}}) = \log \mathcal{N}(\mu_{\bar{\boldsymbol{\theta}}}; \mu_{\boldsymbol{\theta}}, \sigma^2_{\boldsymbol{\theta}}) \qquad (7)$$

ROBUST CLASSIFICATION VIA REGRESSION FOR LEARNING WITH NOISY LAB

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology



Step3. Regression prediction 을 Classification prediction 으로 변환

$$\underbrace{\hat{\boldsymbol{\mu}}}_{\text{Regression prediction}} \in \mathbb{R}^{K-1} \quad \rightarrow \quad \underbrace{\hat{\boldsymbol{\pi}} = ilr^{-1}(\hat{\boldsymbol{\mu}})}_{\Delta^{K-1}} \quad \rightarrow \quad \underbrace{\hat{y} = \arg\max_k \hat{\boldsymbol{\pi}}_k}_{\text{Classification prediction}} \tag{3}$$

모델의 아웃풋 = $\hat{\mu}, \hat{\sigma}^2$

$\hat{\mu}$: estimated true mean

$\hat{\sigma}^2$: estimated noise variance

Erik Englesson, Ho

Experiments

Table 1: **Synthetic Noise: CIFAR-10 and CIFAR-100.** All methods are implemented in a common code base, and hyperparameters are searched for. We report the mean and standard deviation of five different runs, where results in bold has no statistically significant difference compared to the method with the highest mean accuracy. Our method consistently demonstrates strong robustness compared to the baselines across various noise rates, noise types and datasets.

	Method	No Noise	Symmetric Noise Rate			Asymmetric Noise Rate		
		0%	20%	40%	60%	20%	30%	40%
CIFAR-10	CE	90.67 ± 0.80	73.54 ± 1.01	56.56 ± 1.44	39.44 ± 1.87	81.35 ± 1.26	76.01 ± 2.67	71.89 ± 1.67
	GCE	90.83 ± 0.44	87.55 ± 0.41	84.72 ± 0.82	79.25 ± 0.93	85.68 ± 0.69	83.97 ± 0.52	72.90 ± 1.61
	LS	89.78 ± 0.39	79.09 ± 0.96	64.27 ± 1.50	43.57 ± 3.13	81.99 ± 1.22	76.49 ± 1.17	71.66 ± 1.78
	HET	90.82 ± 0.42	77.16 ± 0.94	62.85 ± 1.88	44.20 ± 3.05	81.55 ± 0.80	77.05 ± 0.33	72.69 ± 0.89
	NAN	89.61 ± 0.93	83.86 ± 1.03	79.80 ± 0.59	73.58 ± 0.41	84.32 ± 1.05	76.79 ± 2.28	72.90 ± 1.92
	LN	90.17 ± 0.55	86.13 ± 1.03	81.37 ± 1.97	76.08 ± 0.63	87.64 ± 0.78	86.91 ± 1.03	82.18 ± 1.30
	ELR	91.78 ± 0.26	90.15 ± 0.54	88.19 ± 0.68	81.87 ± 2.42	90.59 ± 0.36	89.72 ± 0.22	87.37 ± 0.55
	SOP	91.57 ± 0.38	89.86 ± 0.45	88.45 ± 0.51	**85.56 ± 0.93**	89.84 ± 0.55	87.60 ± 0.65	83.90 ± 1.04
	NAL	92.80 ± 0.23	89.79 ± 0.47	86.25 ± 0.28	78.82 ± 0.43	90.80 ± 0.76	89.51 ± 0.76	84.36 ± 1.19
	SGN (Ours)	**94.12 ± 0.22**	**93.02 ± 0.17**	**91.29 ± 0.25**	**86.03 ± 1.19**	**93.35 ± 0.21**	**92.71 ± 0.11**	**91.26 ± 0.27**
CIFAR-100	CE	64.87 ± 0.88	47.39 ± 0.43	33.62 ± 0.79	20.04 ± 0.58	50.98 ± 0.88	44.04 ± 0.73	36.95 ± 0.58
	GCE	64.33 ± 0.83	61.67 ± 0.67	53.96 ± 1.40	42.85 ± 0.79	59.63 ± 1.28	49.21 ± 0.53	36.78 ± 0.50
	LS	65.39 ± 0.40	57.08 ± 0.70	44.03 ± 1.20	26.13 ± 1.45	55.47 ± 0.76	44.70 ± 0.73	38.56 ± 0.66
	HET	65.18 ± 0.90	54.83 ± 0.46	41.49 ± 1.53	22.42 ± 0.95	61.29 ± 0.46	56.44 ± 0.53	45.75 ± 1.02
	NAN	64.25 ± 0.64	56.93 ± 0.77	50.03 ± 0.62	40.45 ± 0.41	56.40 ± 1.07	52.78 ± 0.85	40.59 ± 0.84
	LN	64.88 ± 0.98	60.58 ± 1.07	55.55 ± 1.30	46.43 ± 1.15	64.31 ± 0.98	64.07 ± 0.77	61.20 ± 1.22
	ELR	67.74 ± 0.61	64.70 ± 0.85	59.92 ± 0.95	48.85 ± 0.85	66.32 ± 0.88	65.99 ± 1.16	63.80 ± 0.35
	SOP	62.50 ± 0.76	61.40 ± 1.18	60.92 ± 1.34	50.80 ± 0.74	54.19 ± 0.48	47.22 ± 1.27	39.20 ± 0.60
	NAL	69.59 ± 0.37	64.27 ± 0.18	57.09 ± 0.51	46.23 ± 0.45	66.59 ± 0.48	64.46 ± 0.62	58.01 ± 0.79
	SGN (Ours)	**73.88 ± 0.34**	**71.79 ± 0.26**	**66.86 ± 0.35**	**56.83 ± 0.57**	**72.83 ± 0.31**	**72.16 ± 0.86**	**71.01 ± 0.71**

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Experiments

Table 2: **Natural Noise: CIFAR-N.** All methods are implemented in a common code base, and hyperparameters are searched for. We report the mean and standard deviation of five different runs, where results in bold has no statistically significant difference compared to the method with the highest mean accuracy. Our method consistently demonstrates strong robustness across all settings.

Method	CIFAR-10N					CIFAR-100N
	Aggregate	Random 1	Random 2	Random 3	Worst	
CE	83.59 ± 0.98	77.75 ± 0.74	75.52 ± 1.08	76.25 ± 1.26	59.01 ± 0.98	42.75 ± 0.93
GCE	86.66 ± 0.68	85.66 ± 0.73	85.58 ± 0.65	84.78 ± 0.62	77.48 ± 1.22	48.81 ± 0.46
LS	85.08 ± 0.54	80.07 ± 0.82	79.81 ± 0.62	79.41 ± 0.68	63.07 ± 1.93	45.98 ± 1.44
HET	84.45 ± 0.57	78.87 ± 0.47	76.24 ± 0.96	77.68 ± 1.93	63.27 ± 2.62	45.58 ± 0.80
NAN	85.53 ± 0.83	81.85 ± 1.13	83.40 ± 0.84	82.77 ± 0.78	75.47 ± 0.76	50.00 ± 0.72
LN	85.35 ± 1.33	83.70 ± 0.80	83.65 ± 0.82	84.07 ± 0.71	74.31 ± 1.08	50.37 ± 0.50
ELR	89.61 ± 0.12	89.05 ± 0.89	88.79 ± 0.72	88.88 ± 0.61	82.59 ± 0.54	54.91 ± 1.11
SOP	89.54 ± 0.57	89.65 ± 0.79	89.43 ± 0.56	89.54 ± 0.39	82.17 ± 0.90	50.20 ± 0.84
NAL	91.30 ± 0.16	89.22 ± 0.60	89.09 ± 0.24	89.22 ± 0.39	81.39 ± 1.22	56.33 ± 1.21
SGN (Ours)	**92.06 ± 0.12**	**91.94 ± 0.19**	**91.69 ± 0.22**	**91.91 ± 0.10**	**86.67 ± 0.42**	**60.36 ± 0.71**

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Experiments

Table 3: **Natural Noise: Clothing1M & (mini) WebVision.** All baseline results (except NAL) are from the work of Liu et al. (2022). We follow their evaluation setup of doing early stopping, and also report mean and standard deviation at the end of training (†). NAL results are from Lu et al. (2022), but they follow a different evaluation setup on WebVision, thus marked with *.

Method	Clothing1M	WebVision Validation		WebVision Test (ILSVRC12)	
	Top 1	Top 1	Top 5	Top 1	Top 5
Forward	69.8	61.1	-	57.3	-
Co-teaching	69.2	63.6	-	61.5	-
ELR	72.9	76.3	91.3	68.7	87.8
SOP	73.5	76.6	-	69.1	-
NAL	73.6	77.4*	92.3*	74.1*	92.1*
SGN (Ours)	73.9	77.2	91.2	72.6	90.5
SGN† (Ours)	73.6 ± 0.27	76.12 ± 0.36	90.74 ± 0.29	72.72 ± 0.17	90.35 ± 0.28

Erik Englesson, Hossein Azizpour KTH Royal Institute of Technology

Experiments

Table 4: **Ablation Study.** We report results on synthetic and natural noise on the CIFAR datasets when we systematically deactivate the three components of the method: loss reweighting (LR), label correction (LC), and EMA prediction (EP). Synthetic noise rates are all at 40%.

LR	LC	EP	CIFAR-10		CIFAR-10N		CIFAR-100		CIFAR-100N
			Symmetric	Asymmetric	Random 1	Worst	Symmetric	Asymmetric	
✗	✗	✗	65.51 ± 2.43	71.73 ± 1.35	80.47 ± 1.52	63.05 ± 1.38	47.87 ± 1.28	37.75 ± 0.79	47.29 ± 0.67
✓	✗	✗	77.62 ± 0.59	78.15 ± 0.58	83.47 ± 0.81	72.58 ± 1.08	53.54 ± 0.83	59.03 ± 1.12	48.61 ± 1.24
✓	✓	✗	87.66 ± 0.68	86.49 ± 0.83	88.67 ± 1.09	84.24 ± 0.28	61.10 ± 0.75	63.80 ± 1.16	55.72 ± 0.65
✗	✗	✓	73.12 ± 0.64	76.15 ± 0.51	87.04 ± 0.32	71.64 ± 0.42	57.06 ± 0.44	44.22 ± 0.35	55.18 ± 0.28
✓	✗	✓	84.86 ± 0.60	83.97 ± 0.25	89.04 ± 0.13	80.30 ± 0.33	64.04 ± 0.47	67.76 ± 1.30	57.25 ± 0.48
✓	✓	✓	**91.29 ± 0.25**	**91.26 ± 0.27**	**91.94 ± 0.19**	**86.67 ± 0.42**	**66.86 ± 0.35**	**71.01 ± 0.71**	**60.36 ± 0.71**