ロバストな半教師あり異常検知のための Positive-Unlabeled Autoencoder

高橋大志, 岩田具治, 熊谷充敏, 山中友貴
NTT

1. 問題設定

- 半教師あり異常検知は、ラベルなしデータ $\mathcal{U} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ に加えて少量の異常データ $\mathcal{A} = \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_M\}$ を用いて、検出性能の向上を目指している
- 既存手法では、**ラベルなしデータを全て正常データと仮定**し、ラベルなしデータに対する異常スコアを最小化、異常データに対する異常スコアを最大化するようにモデルを学習する
- しかし、**ラベルなしデータは異常を含んでいる**ことが多く、モデルの性能が低下してしまう問題がある

MNISTによる例 (1が正常・他が異常)

ラベルなし(学習)　異常 (学習)　正常 (テスト)　異常 (テスト)

2. 既存手法: Autoencoder

- Autoencoder (AE)[1] はデータ $\mathbf{x}$ を低次元の表現 $\mathbf{z}$ に変換するエンコーダ $E_\theta(\mathbf{x})$ と、 $\mathbf{z}$ から $\mathbf{x}$ を再構成するデコーダ $D_\theta(\mathbf{z})$ の二つのニューラルネットワークから構成される
- AEは、各データに対して下記の再構成誤差を最小化するようにパラメータ θ を学習する:

$$\ell(\mathbf{x};\theta) = \|D_\theta(E_\theta(\mathbf{x})) - \mathbf{x}\|$$

- AEは、学習データの再構成には成功するが、学習していないデータの再構成は失敗するため、再構成誤差は異常スコアとして用いることができる

3. 既存手法: Autoencoding Binary Classifier

- Autoencoding Binary Classifier (ABC)[2] はデータ $\mathbf{x}$ が正常 $(y=0)$ もしくは異常 $(y=1)$ である条件付き確率 $p_\theta(y|\mathbf{x})$ を、再構成誤差を用いて下記で定義する:

$$p_\theta(y|\mathbf{x}) = \begin{cases} \exp(-\ell(\mathbf{x};\theta)) & (y=0) \\ 1-\exp(-\ell(\mathbf{x};\theta)) & (y=1) \end{cases}$$

- 再構成誤差が小さい時 $p_\theta(y=0|\mathbf{x})$ が大きくなり、再構成誤差が大きい時 $p_\theta(y=1|\mathbf{x})$ が大きくなる
- この条件付き確率を用いて、下記の損失を導入する:

$$\begin{aligned}\ell_{\mathrm{BCE}}(\mathbf{x}, y;\theta) &= -\log p_\theta(y|\mathbf{x}) \\ &= \underbrace{(1-y)\ell(\mathbf{x};\theta)}_{\text{再構成誤差の最小化}} - \underbrace{y\log(1-\exp(-\ell(\mathbf{x};\theta)))}_{\text{再構成誤差の最大化}}\end{aligned}$$

- ABCでは、下記の目的関数を最小化するように学習する:

$$\mathcal{L}_{\mathrm{ABC}}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \ell_{\mathrm{BCE}}(\underbrace{\mathbf{x}_n}_{\text{ラベルなしデータ}}, 0;\theta) + \frac{1}{M}\sum_{m=1}^{M} \ell_{\mathrm{BCE}}(\underbrace{\tilde{\mathbf{x}}_m}_{\text{異常データ}}, 1;\theta)$$

- しかし、ラベルなしデータは異常を含んでいることが多く、モデルの性能が低下してしまう

4. 提案手法

- この問題を解決するため、PU Learning[3] とABCに基づく Positive-Unlabeled Autoencoder (PUAE) を提案する
- まず最初に、$p_{\mathcal{N}}(\mathbf{x})$, $p_{\mathcal{A}}(\mathbf{x})$, $p_{\mathcal{U}}(\mathbf{x})$ を、それぞれ正常、異常、ラベルなしデータの確率分布として導入し、データセット $\mathcal{U}$ と $\mathcal{A}$ はそれぞれ $p_{\mathcal{U}}(\mathbf{x})$ と $p_{\mathcal{A}}(\mathbf{x})$ に従うと仮定する
- $p_{\mathcal{U}}(\mathbf{x})$ は、異常の発生率を表すハイパーパラメータ $0 \le \alpha \le 1$ を用いて、下記で表現できると仮定する:

$$p_{\mathcal{U}}(\mathbf{x}) = \alpha p_{\mathcal{A}}(\mathbf{x}) + (1-\alpha)p_{\mathcal{N}}(\mathbf{x})$$

- したがって、 $p_{\mathcal{N}}(\mathbf{x})$ は下記で表現できる:

$$(1-\alpha)p_{\mathcal{N}}(\mathbf{x}) = p_{\mathcal{U}}(\mathbf{x}) - \alpha p_{\mathcal{A}}(\mathbf{x})$$

- もし $p_{\mathcal{N}}(\mathbf{x})$ にアクセスできる場合、下記の理想的な目的関数を最小化するうことで、モデルを学習できる

$$\mathcal{L}_{\mathrm{PN}}(\theta) = \underbrace{\alpha\mathbb{E}_{p_{\mathcal{A}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 1;\theta)]}_{\text{再構成誤差の最大化}} + \underbrace{(1-\alpha)\mathbb{E}_{p_{\mathcal{N}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)]}_{\text{再構成誤差の最小化}}$$

- 実際には $p_{\mathcal{N}}(\mathbf{x})$ が未知なため第二項を直接計算できないが、$p_{\mathcal{U}}(\mathbf{x})$ と $p_{\mathcal{A}}(\mathbf{x})$ を用いることで下記のように計算できる:

$$\underbrace{(1-\alpha)\mathbb{E}_{p_{\mathcal{N}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)]}_{p_{\mathcal{U}}(\mathbf{x}) - \alpha p_{\mathcal{A}}(\mathbf{x})} = \mathbb{E}_{p_{\mathcal{U}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)] - \alpha\mathbb{E}_{p_{\mathcal{A}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)]$$

- したがって、$\mathcal{L}_{\mathrm{PN}}(\theta)$ は $\mathcal{U}$ と $\mathcal{A}$ を用いて下記で近似できる:

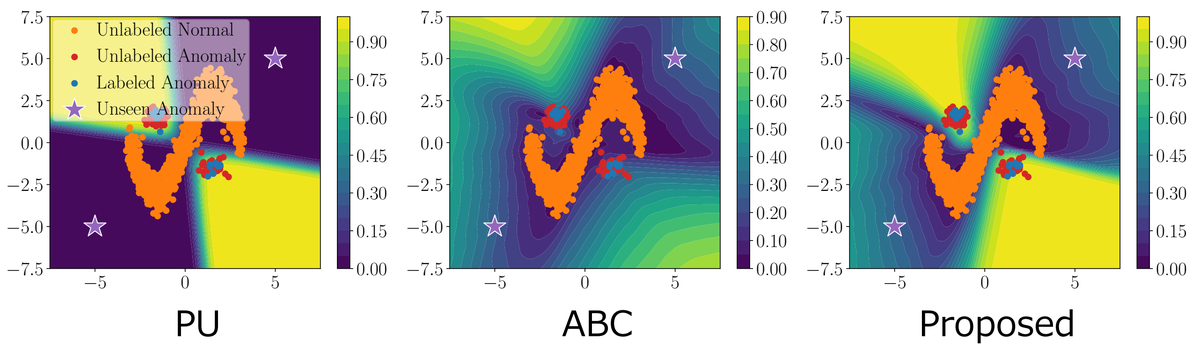
$$\begin{aligned}\mathcal{L}_{\mathrm{PN}}(\theta) &= \alpha\mathbb{E}_{p_{\mathcal{A}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 1;\theta)] + \mathbb{E}_{p_{\mathcal{U}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)] - \alpha\mathbb{E}_{p_{\mathcal{A}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)] \\ &\simeq \alpha\underbrace{\frac{1}{M}\sum_{m=1}^{M}\ell_{\mathrm{BCE}}(\tilde{\mathbf{x}}_m, 1;\theta)}_{\mathcal{L}^+_{\mathcal{A}}(\theta)} + \underbrace{\frac{1}{N}\sum_{n=1}^{N}\ell_{\mathrm{BCE}}(\mathbf{x}_n, 0;\theta)}_{\mathcal{L}^-_{\mathcal{U}}(\theta)} - \alpha\underbrace{\frac{1}{M}\sum_{m=1}^{M}\ell_{\mathrm{BCE}}(\tilde{\mathbf{x}}_m, 0;\theta)}_{\mathcal{L}^-_{\mathcal{A}}(\theta)}\end{aligned}$$

- 過学習を防ぐため、下記を最小化するように学習する[3]:

$$\mathcal{L}_{\mathrm{PUAE}}(\theta) = \alpha\mathcal{L}^+_{\mathcal{A}}(\theta) + \max\left\{0, \underbrace{\mathcal{L}^-_{\mathcal{U}}(\theta) - \alpha\mathcal{L}^-_{\mathcal{A}}(\theta)}_{(1-\alpha)\mathbb{E}_{p_{\mathcal{N}}}[\ell_{\mathrm{BCE}}(\mathbf{x}, 0;\theta)] \text{ の近似で、本来は正だが負になり得る}}\right\}$$

5. 実験

トイデータによる比較



PU　ABC　Proposed

異常検知性能の定量評価 (AUROC)

	MNIST	FashionMNIST	SVHN	CIFAR10
IF	0.829 ± 0.089	0.911 ± 0.060	0.514 ± 0.013	$\mathbf{0.561 \pm 0.097}$
AE[1]	0.894 ± 0.056	0.823 ± 0.094	0.598 ± 0.027	0.532 ± 0.127
DeepSVDD	0.763 ± 0.060	0.667 ± 0.070	0.513 ± 0.014	0.556 ± 0.032
LOE	0.905 ± 0.050	0.860 ± 0.101	0.594 ± 0.028	0.533 ± 0.125
ABC[2]	0.897 ± 0.055	0.826 ± 0.093	0.599 ± 0.027	0.533 ± 0.127
DeepSAD	0.795 ± 0.054	0.693 ± 0.072	0.514 ± 0.016	0.560 ± 0.033
SOEL	0.944 ± 0.034	0.896 ± 0.075	0.601 ± 0.032	0.535 ± 0.123
PU[3]	0.943 ± 0.071	0.811 ± 0.180	0.500 ± 0.014	0.488 ± 0.066
PUAE	$\mathbf{0.971 \pm 0.019}$	$\mathbf{0.940 \pm 0.047}$	$\mathbf{0.608 \pm 0.030}$	$\mathbf{0.554 \pm 0.115}$
PUSVDD	0.922 ± 0.029	0.884 ± 0.060	0.538 ± 0.030	$\mathbf{0.575 \pm 0.037}$

1. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." science 313.5786 (2006): 504-507.
2. Yamanaka, Yuki, et al. "Autoencoding binary classifiers for supervised anomaly detection." PRICAI2019.
3. Kiryo, Ryuichi, et al. "Positive-unlabeled learning with non-negative risk estimator." NeurIPS2017.