Understanding the Effect of Bias in Deep Anomaly Detection

Ziyu Ye  Yuxin Chen  Haitao Zheng
University of Chicago

1 Motivation: Bias from Additional Labeled Anomalies

Existing approaches for Anomaly Detection (AD)

A compact enclosing of the normal.
Unable to use additional labels.
Underfitting bias.

Train with additional labeled anomalies.

A compact enclosing of the normal.
Discriminating on known anomalies.

A Counter-Intuitive Example

Training with additional labeled anomalies can induce disastrous harmful bias.

2 Define Bias: an ERM Framework

Scoring Bias
\[ \text{bias}(x, y) = \arg \max_{(x', y') \in \mathcal{S} \setminus \{x, y\}} \text{TPR}(x', y') - \text{TPR}(x, y) \]

Relative Scoring Bias
\[ \xi(x, y) = \text{bias}(x, y) - \text{bias}(x', y') \]

Empirical Relative Scoring Bias
\[ \hat{\xi}(x, y) = \text{TPR}(x', y') - \text{TPR}(x, y) \]

3 Estimate Bias: a PAC Analysis

Theorem 3. Assume that \( \mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4 \) are Lipschitz continuous with Lipschitz constant \( c_1, c_2, c_3, c_4 \), respectively. Let \( \mathcal{L} \) be the fraction of abnormal data from the nuisance distribution. Then, w.p. at least 1 - \( \delta \), with

\[ N = \mathcal{O}\left(\frac{1}{\mathcal{L} \log \frac{1}{\delta}}\right) \]

the empirical relative scoring bias satisfies

\[ |\hat{\xi}(x, y) - \xi(x, y)| \leq c_3 \sqrt{\frac{2}{N}} + c_4 \frac{1}{\sqrt{N}} \]

The estimation error decreases at the rate of

4 Characterize Bias: Empirical Experiments

Scenario 1: Training w/ hard anomalies.

Scenario 2: Training w/ easy anomalies.

5 Takeaways and Future Directions

Research Question

- Will unseen anomalies suffer from bias due to additional labeled data in training?
- If so, how can we estimate the bias? What is the impact of the bias?

[Clarification] Bias in AD ≠ Bias in Supervised Learning

Table 1: Comparison of anomaly detection tasks with other relevant classification tasks.

Additional labeled data in AD poses a hidden threat for model practitioners.

Data-Based Debiasing Strategy

- Using active learning to obtain representatives anomaly labels.
- Leveraging synthetic examples.

Model-Based Debiasing Strategy

- Using robust model design (e.g., ensembles of semi-supervised and supervised models).