Achieving Counterfactual Fairness for Anomaly Detection

Kyungseon Lee, Choeun Kim, Hankyo Jung

March 21, 2024

Seoul National University

Outline

Introduction

Introduction

1. Challenge
   - Existing fair anomaly detection approaches mainly focus on **association-based fairness** notions.
2. Contribution
   - CFAD model: **Counterfactually fair** anomaly detection.

Related Work

Related Work

1. Counterfactual fairness.
   - A distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different in a causal sense.
2. Definition
   - (Counterfactual fairness). Predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$,

$$P\left(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a\right),$$

   for all y and for any value a' attainable by A.

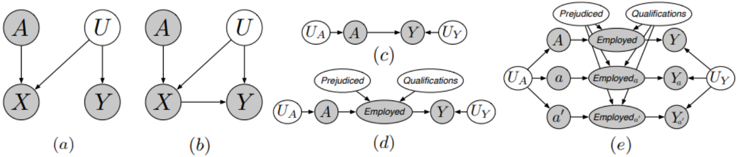
Related Work

1. Definition

$$P\left(\hat{Y}_{A\leftarrow a}(U)=y \mid X=x, A=a\right)=P\left(\hat{Y}_{A\leftarrow a'}(U)=y \mid X=x, A=a\right),$$

2. Intervention on variable V_i
   - substitution of equation $V_i=f_i\left(pa_i, U_{pa_i}\right)$ with the equation $V_i=v$
3. Counterfactual
   - $Y_{A\leftarrow a}(u)$ or Y_a : the value of Y if A had taken value a Causal model is defined by (U, V, F)
     - V : observable variables
     - U : set of latent background variable, which are factors not caused by V
     - F is a set of functions $\{f_1, \ldots, f_n\}$ such that $V_i=f_i\left(pa_i, U_{pa_i}\right)$ where $pa_i \subseteq V\setminus\{V_i\}, U_{pa_i} \subseteq U$
     - pa_i refers to the "parents" of V_i

Related Work



- V : observable variables
- U : set of latent background variable, which are factors not caused by V
- F is a set of functions $\{f_1, \ldots, f_n\}$ such that $V_i = f_i\left(pa_i, U_{pa_i}\right)$ where $pa_i \subseteq V \setminus \{V_i\}, U_{pa_i} \subseteq U$
- pa_i refers to the "parents" of V_i
- A : sensitive attributes

Methodology

Methodology

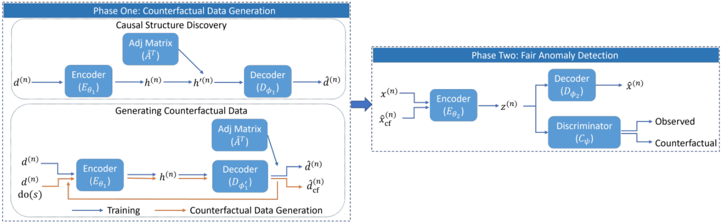


Fig. 1: Framework of CFAD

1. Causal Structure Discovery
2. Generating Counterfactual Data
3. Fair Anomaly Detection

Fair Anomaly Detection

1. θ_2 : Encoder model parameters
2. ϕ_2 : Decoder model parameters
3. D_{ϕ_2} : Decoder
4. E_{θ_2} : Encoder
5. $\left(d^{(n)}\right)_{n=1}^{N} = \left[s^{(n}, x^{(n}\right]$: A training set
6. $s^{(n)}$: A binary sensitive variable
7. $x^{(n)}$: All other variables
8. C_{ψ} : Discriminator
9. $z^{(n)}$: The hidden representations
10. $g(x)$: Anomaly score function
11. $x_{cf}^{(n)}$: Generated counterfactual data

Methodology

Counterfactual fairness is defined as

Definition

An anomaly detection model is counterfactually fair if for each individual n we have $g\left(x^{(n)}\right) = g\left(x_{cf}^{(n)}\right)$.

Fair Anomaly Detection

1. Objective function of AE

$$\mathcal{L}_{\mathrm{AE}}\left(\theta_2, \phi_2\right) = \frac{1}{2N} \sum_{n=1}^{N} \left\| d^{(n)} - D_{\phi_{\mathbf{2}}} \circ E_{\theta_{\mathbf{2}}}\left(x^{(n)}\right) \right\|_2^2$$

$$\min_{\theta_{\mathbf{2}}, \phi_{\mathbf{2}}} \max_{\psi} \mathcal{L}_{\mathrm{AE}}\left(\theta_2, \phi_2\right) + \lambda \mathcal{L}_{\mathrm{C}}\left(\theta_2, \psi\right),$$

2. Objective function of discrimminator

$$\mathcal{L}_{\mathrm{C}}\left(\theta_2, \psi\right) = \frac{1}{N} \sum_{n=1}^{N} \left[\log\left(C_\psi\left(z^{(n)}\right)\right) + \log\left(1 - C_\psi\left(z_{\mathrm{cf}}^{(n)}\right)\right)\right]$$

3. Anomaly score

$$g(x) = \left\| x - D_{\phi_{\mathbf{2}}} \circ E_{\theta_{\mathbf{2}}}(x) \right\|_2^2 .$$

Experiment

Experiment

Datasets & evaluation metric

Table 1: Statistics of datasets.

	Synthetic		Adult		COMPAS	
	Training	Test	Training	Test	Training	Test
Normal (Y=0)	12000	4000	12000	4000	2000	1283
Abnormal (Y=1)	N/A	400	N/A	800	N/A	384

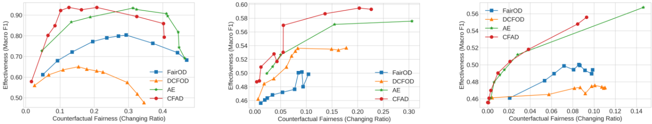
- Datasets: We conduct experiments on a synthetic dataset and two real-world datasets, Adult and COMPAS.
- Synthetic Dataset: We first build a synthetic dataset with 21 variables where we can obtain the ground truth of counterfactuals.
- Evaluation metrics: AUC, PRAUC, Macro F1-score

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i$$

Experiment

Table 2: Anomaly detection on synthetic and real datasets with threshold $\tau = 0.95$. For AUC-PR, AUC-ROC and Macro-F1, the higher the value the better the effectiveness; for Changing Ratio, the lower the value the better the fairness.

Method	Synthetic Dataset				Adult Dataset				COMPAS Dataset			
	AUC-PR	AUC-ROC	Macro-F1	Changing Ratio	AUC-PR	AUC-ROC	Macro-F1	Changing Ratio	AUC-PR	AUC-ROC	Macro-F1	Changing Ratio
PCA	0.992	0.999	0.908	0.478	0.238	0.582	0.476	0.261	0.365	0.642	0.595	0.268
OC-SVM	0.776	0.953	0.477	0.399	0.282	0.638	0.482	0.285	0.337	0.593	0.488	0.376
iForest	0.190	0.693	0.570	0.271	0.312	0.658	0.570	0.279	0.311	0.567	0.564	0.415
AE	0.957	0.996	0.883	0.461	0.349	0.640	0.608	0.590	0.344	0.616	0.581	0.407
DCFOD	0.383	0.832	0.721	0.212	0.249	0.623	0.533	0.071	0.260	0.569	0.466	0.067
FairOD	0.580	0.873	0.689	0.261	0.222	0.621	0.531	0.131	0.265	0.548	0.493	0.068
CFAD	0.947	0.996	0.930	0.199	0.319	0.589	0.576	0.057	0.314	0.596	0.539	0.049



(a) Synthetic (b) Adult (c) COMPAS

Fig. 5: Trade-off between effectiveness and fairness.

- Counterfactual fairness metric

$$\text{changing ratio} = \frac{\sum_{n=1}^{N} 1\left[\hat{y}^{(n)} \neq \hat{y}_{\mathrm{cf}}^{(n)}\right]}{N}$$

Conclusion

Conclusion

- CFAD is able to effectively detect anomalies and also ensure counterfactual fairness.