Carnegie Mellon University Heinz College

FAIROD: Fairness-aware Outlier Detection

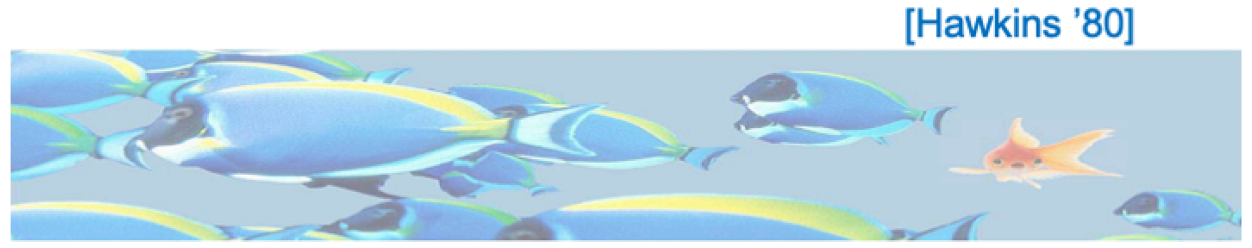
Shubhranshu Shekhar, Neil Shah and Leman Akoglu

Introduction

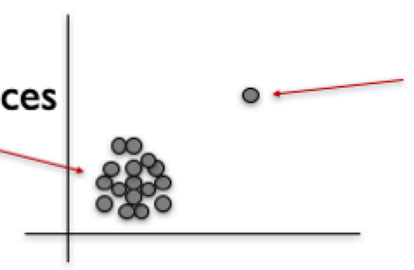
What is an outlier?

Observations that…

- "… **inconsistent** with the remainder…" [Barnett&Lewis'94]
- "… **deviate markedly** from other members of sample in which it occurs" [Grubbs '69]
- "… deviate so much … as to arouse suspicions … they were generated by a **different mechanism**" [Hawkins '80]

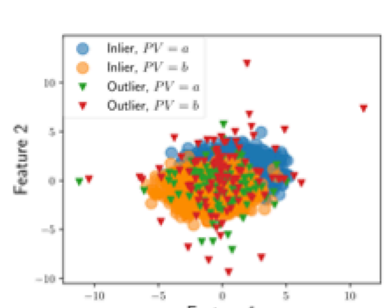


Outlier Detection

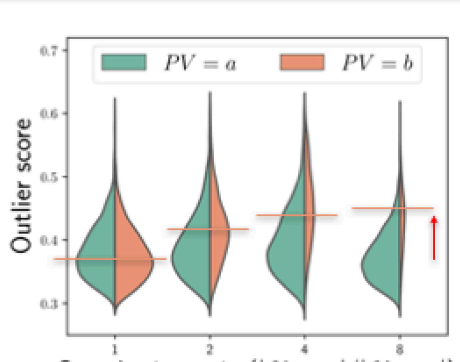


- designed to spot/flag rare, minority samples
  - e.g. suspicious activity, abnormal heart rate etc.
- facilitates auditing ("*policing*") by human experts
  - e.g. stop-and-frisk in automated surveillance flagged instances

Bias in Outlier Detection



- Simulated dataset
  - equal sized groups
  - groups induced by PV ∈ {●, ●}



Higher outlier scores as sample size of $PV = b$ is decreased

Bias in Outlier Detection

- Societal minorities may be statistical minorities
  - defined by protected variable (PV) race/ ethnicity/gender/age etc.
  - societal minority ≠ riskiness
- Disparate Impact
  - unjust flagging leading to over-policing
    - exacerbated by correlated variables with PVs
  - feedback loop results in further skewness

Problem

Fair Outlier Detection

- Given:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^N \subseteq \mathbb{R}^d$
  - $\mathcal{PV} = \{PV_i\}_{i=1}^N,\ PV_i \in \{a, b\}$
    - $PV_i = a$ identifies majority group
- Build a **detector** that estimates outlier scores $\mathcal{S}$ and assigns outlier labels $\mathcal{O}$ s.t.
  - i. assigned labels and scores are "fair" w.r.t. the PV
  - ii. higher scores correspond to higher riskiness encoded by the underlying (unobserved) true labels $\mathcal{Y}$

Proposed Desiderata

D1. Detection effectiveness

$P(Y = 1 \mid O = 1) > P(Y = 1)$

D2. Treatment parity

$P(O{=}1|X) = P(O{=}1|X, PV{=}v), \forall v$

D3. Statistical parity (SP)

$P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$

✓Enforceable

D4. Group fidelity

$P(O{=}1|Y{=}1, PV{=}a) = P(O{=}1|Y{=}1, PV{=}b)$

✓Enforceable via proposed proxy

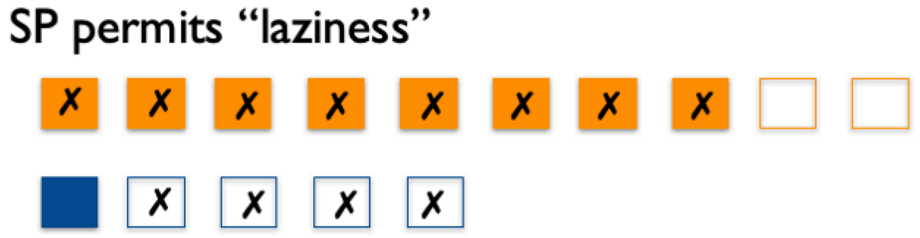
D5. Base rate preservation

$P(Y = 1|O = 1, PV = v) = P(Y = 1|PV = v)$, $\forall v \in \{a, b\}$

✗Can't be enforced

SP and Group Fidelity

- SP permits "laziness"



- Group Fidelity depends on Y
  - proxy enforces group-level rank preservation
  - fidelity to within-group ranking from the base model i.e. $\pi_{PV=v}^{BASE} = \pi_{PV=v}; \forall v \in \{a, b\}$, π denotes ranking
  - addresses laziness

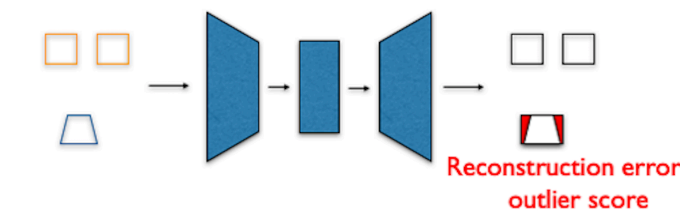
Fairness-aware Outlier Detection

Fairness-aware Outlier detection

- Given:
  - Observations $\mathcal{X} = \{X_i\}_{i=1}^N \subseteq \mathbb{R}^d$
  - $\mathcal{PV} = \{PV_i\}_{i=1}^N,\ PV_i \in \{a, b\}$
    - $PV_i = a$ identifies majority group
- Build a **detector** that estimates outlier scores $\mathcal{S}$ and assigns outlier labels $\mathcal{O}$ to achieve
  - i. $P(Y = 1 \mid O = 1) > P(Y = 1)$ [D1]
  - ii. $P(O{=}1|X) = P(O{=}1|X, PV{=}v), \forall v$ [D2]
  - iii. $P(O{=}1|PV{=}a) = P(O{=}1|PV{=}b)$ [D3]
  - iv. $\pi_{PV=v}^{BASE} = \pi_{PV=v}; \forall v$, [D4]
    BASE is fairness-agnostic detector

FAIROD

- Instantiates deep-autoencoder as BASE detector



- Minimizes the regularized loss

$$\mathcal{L} = \alpha \underbrace{\mathcal{L}_{\text{BASE}}}_{\text{Reconstruction}} + (1-\alpha) \underbrace{\mathcal{L}_{SP}}_{\text{Statistical Parity}} + \gamma \underbrace{\mathcal{L}_{GF}}_{\text{Group Fidelity}}$$

Experiments

Datasets

Dataset	N	d	PV	PV = b	$\lvert\mathcal{X}_{PV=a}\rvert/\lvert\mathcal{X}_{PV=b}\rvert$	% outliers	Labels
Adult	25262	11	gender	*female*	4	5	{income ≤ 50K, income > 50K}
Credit	24593	1549	age	*age ≤ 25*	4	5	{paid, delinquent}
Tweets	3982	10000	racial dialect	*African-American*	4	5	{normal, abusive}
Ads	1682	1558	simulated	1	4	5	{non-ad, ad}
Synth1	2400	2	simulated	1	4	5	{0, 1}
Synth2	2400	2	simulated	1	4	5	{0, 1}

Evaluation Measures

- Fairness = $\min\left(r, \frac{1}{r}\right)$, where $r = \frac{P(O{=}1|PV{=}a)}{P(O{=}1|PV{=}b)}$
- Group Fidelity = $HM(NDCG_{PV=a}, NDCG_{PV=b})$
- Top-k rank agreement = $\frac{|\pi_{[1:k]}^{BASE} \cap \pi_{[1:k]}^{detector}|}{|\pi_{[1:k]}^{BASE} \cup \pi_{[1:k]}^{detector}|}$
- AUC-ratio = $\frac{AUC_{PV=a}}{AUC_{PV=b}}$
- AP-ratio = $\frac{AP_{PV=a}}{AP_{PV=b}}$

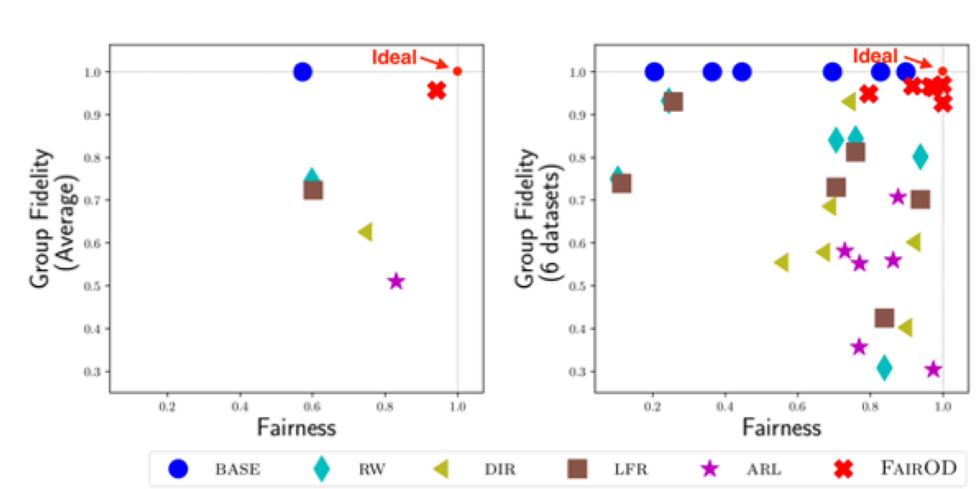
used when ground truth labels are available

Baselines

- BASE – Deep anomaly detector based on autoencoder
- RW – utilizes reweighting to counterbalance under-representation of minority group
- DIR – edits feature values decorrelateing features and PV
- LFR – finds latent representation of the data while obfuscating information about PV
- ARL – finds latent representation by employing an adversarial training process to remove PV information

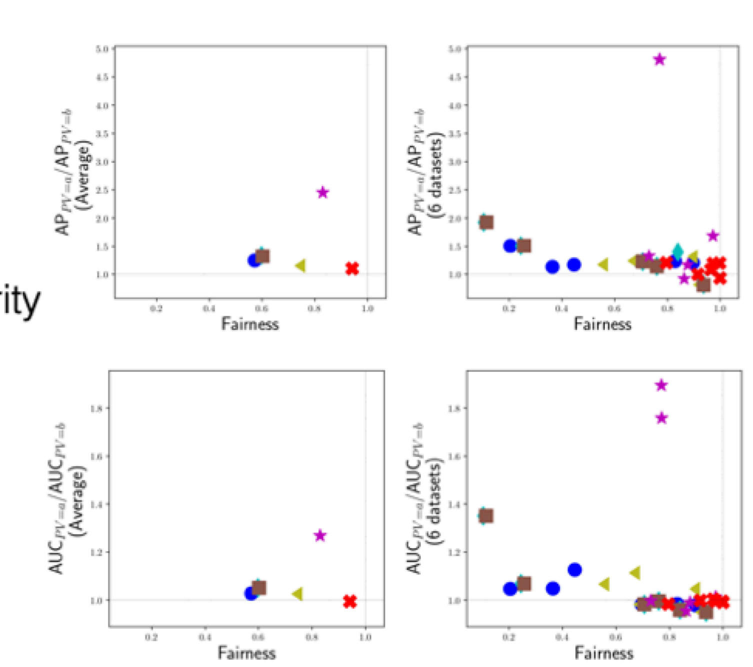
Results

Fairness



Group Fidelity vs Fairness

Label aware parity measures vs Fairness



Fairness-accuracy trade-off