FairOD: Fairness-aware Outlier Detection

Appendix

Shubhranshu Shekhar
shubhras@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Neil Shah
nshah@snap.com
Snap Inc.
Seattle, WA, USA

Leman Akoglu
lakoglu@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

This document supplements [3] by giving proofs of the presented claims, detailed dataset description, and details on hyperparameters and additional experiments.

A PROOFS

A.1 Proof of Claim 1

Proof. We want OD to exhibit detection effectiveness i.e. $P(Y = 1|O = 1) > P(Y = 1)$.

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a|O=1)\cdot\\ & P(Y=1|PV=a,O=1)+\\ & P(PV=b|O=1)\cdot\\ & P(Y=1|PV=b,O=1)\end{aligned}$$

Given SP, we have

$$\begin{aligned} & P(O=1|PV=a) = P(O=1|PV=b)\\ \implies & P(PV=a|O=1) = P(PV=a),\ \text{and}\\ & P(PV=b|O=1) = P(PV=b)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a)\cdot\\ & P(Y=1|PV=a,O=1)+\\ & P(PV=b)\cdot\\ & P(Y=1|PV=b,O=1)\end{aligned}\tag{1}$$

Now,

$$\begin{aligned}P(Y=1) =& P(PV=a)\cdot P(Y=1|PV=a)+\\ & P(PV=b)\cdot P(Y=1|PV=b)\end{aligned}$$

Therefore, if we want $P(Y = 1|O = 1) > P(Y = 1)$, then

$$\begin{aligned} & P(PV=a)\cdot P(Y=1|PV=a,O=1)+\\ & P(PV=b)\cdot P(Y=1|PV=b,O=1)\\ & \qquad >\\ & P(PV=a)\cdot P(Y=1|PV=a)+\\ & P(PV=b)\cdot P(Y=1|PV=b)\end{aligned}\tag{2}$$

$$\begin{aligned}\implies \exists v\in\{a,b\}\quad s.t.\ & P(Y=1|PV=v,O=1)\\ & >\\ & P(Y=1|PV=v)\end{aligned}$$

□

A.2 Proof of Claim 2

Proof. Without loss of generality, assume that $P(Y = 1|PV = a, O = 1) > P(Y = 1|PV = a)$ i.e. (i.e. $P(Y = 1|PV = a, O = 1) = K \cdot P(Y = 1|PV = a); K > 1$), and let $\frac{P(Y=1|PV=a)}{P(Y=1|PV=b)} = \frac{P(Y=1|PV=a,O=1)}{P(Y=1|PV=b,O=1)} = \frac{1}{r}$ then
Case 1: When $P(Y = 1|PV = b, O = 1) < P(Y = 1|PV = b)$

$$\begin{aligned} & P(Y=1|PV=b,O=1) < P(Y=1|PV=b)\\ \implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a)\\ \implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a,O=1),\\ & \qquad [\because P(Y=1|PV=a,O=1) > P(Y=1|PV=a)]\end{aligned}$$

This contradicts our assumption that $P(Y = 1|PV = b, O = 1) = r \cdot P(Y = 1|PV = a, O = 1)$, therefore it must be that $P(Y = 1|PV = b, O = 1) \boldsymbol{\geq} P(Y = 1|PV = b)$.

Case 2: When $P(Y = 1|PV = b, O = 1) = P(Y = 1|PV = b)$

$$\begin{aligned} & P(Y=1|PV=b,O=1) = P(Y=1|PV=b)\\ \implies & P(Y=1|PV=b,O=1) = r\cdot P(Y=1|PV=a)\\ \implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a,O=1),\\ & \qquad [\because P(Y=1|PV=a,O=1) > P(Y=1|PV=a)]\end{aligned}$$

This contradicts our assumption that $P(Y = 1|PV = b, O = 1) = r \cdot P(Y = 1|PV = a, O = 1)$, therefore it must be that $P(Y = 1|PV = b, O = 1) \boldsymbol{>} P(Y = 1|PV = b)$.

Case 3: When $P(Y = 1|PV = b, O = 1) > P(Y = 1|PV = b)$ i.e. $(P(Y = 1|PV = b, O = 1) = L \cdot P(Y = 1|PV = b); L > 1)$
Now, we know that,

$$\begin{aligned} & P(Y=1|PV=a)\cdot P(Y=1|PV=b,O=1)\\ & \quad = P(Y=1|PV=b)\cdot P(Y=1|PV=a,O=1)\\ \implies & P(Y=1|PV=a)\cdot P(Y=1|PV=b,O=1)\\ & \quad = P(Y=1|PV=b)\cdot K\cdot P(Y=1|PV=a)\\ \implies & P(Y=1|PV=b,O=1) = K\cdot P(Y=1|PV=b)\\ \implies & P(Y=1|PV=b,O=1) > P(Y=1|PV=b)\end{aligned}$$

And, for ratio to be preserved, it must be that $L = K$.

Hence, enforcing preservation of ratios implies base-rates in flagged observations are larger than their counterparts in the population. □

B DATA DESCRIPTION

B.1 Synthetic data

We illustrate the effectiveness of FairOD on two synthetic datasets, namely Synth1 and Synth2 (as illustrated in Fig. 3 in [3]). These datasets are constructed to present scenarios that mimic real-world settings, where we may have features which are uncorrelated with respect to outcome labels but partially correlated with PV, or features which are correlated both to outcome labels and PV.

- Synth1: In Synth1, we simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1 is correlated with the protected variable PV, but does not offer any predictive value with respect to ground-truth outlier labels $\mathcal{Y}$, while x_2 is correlated with these labels $\mathcal{Y}$ (see Fig. 3a in [3]). We draw 2400 samples, of which $PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. x_1 differs in terms of shifted means, but equal variances, for both majority and minority groups. x_2 is distributed similarly for both majority and minority groups, drawn from a normal distribution for outliers, and an exponential for inliers. The detailed generative process for the data is below, and Fig. 3a shows a visual.

Synth1

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$
$$Y \sim \text{Bernoulli}(1/20)$$
$$x_1 \sim \begin{cases} \text{Normal}(-1, 1.44) & \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1.44) & \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$
$$x_2 \sim \begin{cases} \text{Normal}(-1, 1) & \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1) & \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

- Synth2: In Synth2, we again simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1, x_2 are partially correlated with both the protected variable PV as well as ground-truth outlier labels $\mathcal{Y}$ (see Fig. 3b in [3]). We draw 2400 samples, of which $PV = a$ (majority) for 2000 points, and $PV = b$ (minority) for 400 points. 120 (5%) of these points are outliers. For inliers, both x_1, x_2 are normally distributed, and differ across majority and minority groups only in terms of shifted means, but equal variances. Outliers are drawn from a product distribution of an exponential and linearly transformed Bernoulli distribution (product taken for symmetry). The detailed generative process for the data is below (right), and Fig. 3b shows a visual.

Synth2

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$
$$Y \sim \text{Bernoulli}(1/20)$$
$$x_1 \sim \begin{cases} \text{Normal}(180, 10) & \text{if} \quad PV = 1 \quad \text{[a, majority]} \\ \text{Normal}(150, 10) & \text{if} \quad PV = 0 \quad \text{[b, minority]} \end{cases}$$
$$x_2 \sim \begin{cases} \text{Normal}(10, 3) & \text{if} \quad Y = 1 \quad \text{[outlier]} \\ \text{Exponential}(1) & \text{if} \quad Y = 0 \quad \text{[inlier]} \end{cases}$$

B.1.1 *Real-world data.*

We conduct experiments on 4 real-world datasets and select them from diverse domains that have different types of (binary) protected variables, specifically gender, age, and race. Detailed descriptions are as follows.

- **Adult** [2] (Adult). The dataset is extracted from the 1994 Census database where each data point represents a person. The dataset records income level of an individual along with features encoding personal information on education, profession, investment and family. In our experiments, *gender* ∈ {*male, female*} is used as the protected variable where *female* represents minority group and high earning individuals who exceed an annual income of 50,000 i.e. annual $income > 50,000$ are assigned as outliers ($Y = 1$). We further downsample *female* to achieve a *male* to *female* sample size ratio of 4:1 and ensure that percentage of outliers remains the same (at 5%) across groups induced by the protected variable.

- **Credit-defaults** [2] (Credit). This is a risk management dataset from the financial domain that is based on Taiwan's credit card clients' default cases. The data records information of credit card customers including their payment status, demographic factors, credit data, historical bill and payments. Customer *age* is used as the protected variable where $age > 25$ indicates the majority group and $age \leq 25$ indicates the minority group. We assign individuals with delinquent *payment status* as outliers ($Y = 1$). The $age > 25$ to $age \leq 25$ imbalance ratio is 4:1 and contains 5% outliers across groups induced by the protected variable.

- **Abusive Tweets** [1] (Tweets). The dataset is a collection of Tweets along with annotations indicating whether a tweet is abusive or not. The data are not annotated with any protected variable by default; therefore, to assign protected variable to each Tweet, we employ the following process: We predict the racial dialect — *African-American* or *Mainstream* — of the tweets in the corpus using the language model proposed by [1]. The dialect is assigned to a Tweet only when the prediction probability is greater than 0.7, and then the predicted *racial dialect* is used as protected variable where *African-American dialect* represents the minority group. In this setting, abusive tweets are labeled as outliers ($Y = 1$) for the task of flagging abusive content on Twitter. The group sample size ratio of *racial dialect* = *African-American* to *racial dialect* = *Mainstream* is set to 4:1. We further sample data points to ensure equal percentage (5%) of outliers across dialect groups.

- **Internet ads** [2] (Ads). This is a collection of possible advertisements on web-pages. The features characterize each ad by encoding phrases occurring in the ad URL, anchor text, alt text, and encoding geometry of the ad image. We assign observations with class label *ad* as outliers ($Y = 1$) and downsample the data to get an outlier rate of 5%. There exists no demographic information available, therefore we simulate a binary protected variable by randomly assigning each observation to one of two values (i.e. groups) ∈ {0, 1} such that the group sample size ratio is 4:1.

C HYPERPARAMETERS

We choose the hyperparameters of FairOD from $\alpha \in \{0.01, 0.5, 0.9\} \times \gamma \in \{0.01, 0.1, 1.0\}$ by evaluating the Pareto curve for fairness and group fidelity criteria. The base and FairOD methods both

Table 1: Evaluation measures are reported for the competing methods on the datasets presented in Appendix B.1.

(a) Synth1

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0262	0.1282	1.0	1.0	0.9594	0.9168	0.8819	0.5849
RW	0.033	0.135	0.9299	0.9309	0.9794	0.9168	0.8819	0.5849
DIR	0.0445	0.0775	0.3953	0.9281	0.9742	0.9138	0.8814	0.7529
LFR	0.0330	0.1350	0.9299	0.9309	0.9794	0.9168	0.8819	0.5849
ARL	0.0520	0.0400	0.9136	0.3955	0.9786	0.5565	0.886	0.1842
FairOD	0.0500	0.0500	0.9639	0.9671	0.9666	0.9634	0.8166	0.7557
FairOD-L	0.0495	0.0525	0.9149	0.9295	0.9017	0.8714	0.599	0.5214
FairOD-C	0.0480	0.0600	0.8929	0.9082	0.9499	0.9284	0.7542	0.6501

(b) Synth2

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0361	0.0811	1.0	1.0	0.6153	0.5464	0.273	0.2335
RW	0.0205	0.1975	0.9242	0.6313	0.7544	0.5586	0.3973	0.2064
DIR	0.0465	0.0675	0.4224	0.9164	0.7892	0.7089	0.3921	0.317
LFR	0.0205	0.1975	0.9242	0.6313	0.7544	0.5586	0.3973	0.2064
ARL	0.0520	0.0400	0.1801	0.1386	0.9786	0.5165	0.886	0.1842
FairOD	0.0500	0.0500	0.9339	0.9201	0.6357	0.6419	0.2726	0.2918
FairOD-L	0.0500	0.0500	0.8984	0.8843	0.6385	0.6472	0.2742	0.2838
FairOD-C	0.0450	0.0750	0.8997	0.9095	0.5957	0.5419	0.2665	0.2339

(c) Adult

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0358	0.0433	1.0	1.0	0.6344	0.6449	0.1105	0.0898
RW	0.0515	0.0391	0.8399	0.8479	0.6323	0.6351	0.1303	0.1141
DIR	0.0515	0.0391	0.9299	0.9309	0.6323	0.6351	0.1303	0.1141
LFR	0.0515	0.0391	0.8099	0.8099	0.6323	0.6351	0.1303	0.1141
ARL	0.0507	0.0444	0.9147	0.5765	0.5951	0.6009	0.0987	0.0848
FairOD	0.0497	0.0511	0.9646	0.9616	0.6374	0.6404	0.1085	0.0912
FairOD-L	0.0513	0.0403	0.9178	0.9005	0.6425	0.6312	0.1213	0.1048
FairOD-C	0.0527	0.0302	0.8119	0.7877	0.6533	0.6229	0.1872	0.1435

(d) Credit

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0445	0.064	1.0	1.0	0.7376	0.7512	0.1938	0.1582
RW	0.0467	0.06627	0.8399	0.8409	0.7376	0.7512	0.1938	0.1582
DIR	0.0467	0.06627	0.6899	0.6809	0.7376	0.7512	0.1938	0.1582
LFR	0.0467	0.06627	0.7299	0.7309	0.7376	0.7512	0.1938	0.1582
ARL	0.0471	0.0645	0.5533	0.6118	0.7242	0.7263	0.1396	0.1054
FairOD	0.0468	0.066	0.9235	0.9421	0.7368	0.7494	0.2134	0.1725
FairOD-L	0.0475	0.062	0.7147	0.6564	0.7276	0.7394	0.1246	0.1025
FairOD-C	0.0467	0.0662	0.7871	0.8029	0.7327	0.7484	0.1333	0.1091

(e) Tweets

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0369	0.1015	1.0	1.0	0.5739	0.5476	0.061	0.0539
RW	0.0479	0.0571	0.2882	0.3312	0.5583	0.582	0.0466	0.0334
DIR	0.0494	0.0507	0.388	0.4178	0.5552	0.5307	0.0454	0.0345
LFR	0.0479	0.0571	0.4082	0.4422	0.5583	0.582	0.0466	0.0334
ARL	0.0482	0.0558	0.5432	0.5762	0.4912	0.5146	0.0504	0.0442
FairOD	0.0488	0.0532	0.9668	0.9671	0.569	0.5699	0.0617	0.0617
FairOD-L	0.0331	0.1167	0.9137	0.8986	0.5091	0.4237	0.0574	0.0425
FairOD-C	0.0501	0.0488	0.6753	0.6903	0.5592	0.5891	0.0627	0.1002

(f) Ads

	Flag-rate		GroupFidelity		AUC		AP	
Method	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$	$PV = a$	$PV = b$
BASE	0.0286	0.0318	1.0	1.0	0.7077	0.7234	0.2555	0.2124
RW	0.0491	0.0523	0.8236	0.7813	0.7286	0.7672	0.4227	0.5183
DIR	0.0491	0.0523	0.6236	0.5813	0.7286	0.7672	0.4296	0.5253
LFR	0.0491	0.0523	0.7236	0.6813	0.7286	0.7672	0.4257	0.5253
ARL	0.0499	0.0500	0.5028	0.2181	0.6572	0.6487	0.0885	0.0525
FairOD	0.0499	0.0500	0.9698	0.9699	0.7179	0.7216	0.2592	0.2163
FairOD-L	0.0683	0.0588	0.5551	0.8684	0.7179	0.7345	0.0005	0.0005
FairOD-C	0.0499	0.0500	0.6611	0.6966	0.7007	0.7251	0.2636	0.2455

use an auto-encoder with two hidden layers. We fix the number of hidden nodes in each layer to 2 if $d \leq 100$, and 8 otherwise. The representation learning methods LFR and ARL use the model configurations as proposed by their authors. The hyperparameter grid for the preprocessing baselines are set as follows: $repair\_level \in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ for DIR, $A_z \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ and $A_x = 1 - A_z$ for LFR, and $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ for ARL. We pick the best model for the preprocessing baselines using Fairness as they only optimize for statistical parity. The best BASE model is selected based on reconstruction error through cross validation upon multiple runs with different random seeds.

D SUPPLEMENTAL RESULTS

In this section, we report Flag-rate, GroupFidelity, AUC and AP (see Table 1) for the competing methods on a set of datasets (see Appendix B.1.1) w.r.t groups induced by $PV = v;\ v \in \{a, b\}$ to supplement the experimental results presented in Sec. 4. Notice that in most cases (see Table 1a through Table 1f), FairOD outperforms the BASE model on label-aware parity metrics (AUC-ratio, AP-ratio) and, furthermore, outperforms BASE on at least one of the performance metrics (e.g. AUC, AP); fairness need not imply worse OD performance.

REFERENCES

[1] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[2] Moshe Lichman et al. 2013. UCI machine learning repository.

[3] Shubhranshu Shekhar, Neil Shah, and Leman Akoglu. 2021. FairOD: Fairness-aware Outlier Detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*.