On Improving Data Utility of DECAF

Hieu M. Vu
vmhieu17@gmail.com

Abstract

DECAF: Generating FairSynthetic Data Using Causally-Aware Generative Networks by Van Breugel et. al. is a novel GAN-based approach for synthesizing fair tabular data. It aims to synthesize an equivalent unbiased data set given a biased dataset with minimum loss of data utility. While, by assuming an underlying DAG for causality representation, the approach is capable of producing fairness with respect to multiple algorithmic fairness definitions, it leaves room for improvement in the data utility aspect. This work proposes a simple extension to DECAF that improves its data utility while maintaining a similar level of fairness. Furthermore, it provides discussions and suggestions on the choice of data utility metrics and carries out experiments on the Adult data set.[1]

1 Introduction

DECAF [1] is a method that aims to generate fair synthetic tabular data using a GAN-based model with an assumed causal DAG. The method consists of 2 stages:

- First, given a DAG describing the underlying causal relations of the training data, a GAN-based model is trained to generate realistic synthetic data. To model the causal relations, the topology order of the attributes is computed, and the generator then sequentially generates each attribute following said order. The discriminators task is, as usual, to learn to differentiate between real and synthesized samples.
- During inference, a different modified DAG is used with the generator to synthesize data. The modified DAG is a relaxed version of the original DAG, in which some edges are removed to eliminate the causal relation between some protected attributes and the target attribute, thus increasing fairness.

Fairness is defined algorithmically using 3 standards with various degrees of strictness, each corresponding to a modified DAG:

1. Fairness Through Unawareness (FTU) requires that the protected attributes A are not explicitly used to predict the target attribute. This is the most tolerant one as it does not take into account indirect causal relations.
2. Demographic Parity (DP) requires that different classes of protected attributes should receive positive outcomes at equal rates. This is the most strict one as it implied strong constraints upon the DAG and eliminated both direct and indirect discrimination.
3. Conditional Fairness (CF) lies in between FTU and DP in terms of strictness, which allow causality flows from the protected attributes A to the target attribute only through some explanatory attributes R.

[1] Code for the experiment is available at https://github.com/lone17/DECAF

Preprint. Work in progress.

Data Utility is another goal of the original paper [1]. However, compared to fairness, it was given much less focus regarding the proposed formulation and methodology. The original paper[1] also did not give any formal definition of data utility as well as what is considered minimal data utility loss. Moreover, the metrics used to measure data utility appears to be inefficient as it suggests inconsistent interpretation. Further details on this are discussed in the next section.

Contribution This work takes a deeper look at the data utility aspect of DECAF[1] and attempts to make an improvement in that regard. Additionally, new evaluation metrics for data utility and a discussion on the choice of metrics are also included in this work.

Novelty and Related Works To the best of my knowledge, the idea of using a GAN-based model embedded with a causal DAG to generate fair synthetic data is still the only of its kind amongst related literature. Thus, by being a direct extension of DECAF[1], this work further extends its novelty along the same basis.

2 Problem Statement

Table 1: (Table 3 in the original paper) Full table of bias removal experiment on Adult dataset [2] including protected removal (PR) metrics.

	Data Quality			**Fairness**	
Method	Precision↑	Recall↑	AUROC↑	FTU↓	DP↓
Original data $\mathcal{D}$	0.920 ± 0.006	0.936 ± 0.008	0.807 ± 0.004	0.116 ± 0.028	0.180 ± 0.010
GAN	0.607 ± 0.080	0.439 ± 0.037	0.567 ± 0.132	0.023 ± 0.010	0.089 ± 0.008
WGAN-GP	0.683 ± 0.015	0.914 ± 0.005	0.798 ± 0.009	0.120 ± 0.014	0.189 ± 0.024
FairGAN	0.681 ± 0.023	0.814 ± 0.079	0.766 ± 0.029	0.009 ± 0.002	0.097 ± 0.018
GAN-PR	0.632 ± 0.077	0.509 ± 0.110	0.612 ± 0.106	$\ddagger 0.0 \pm 0.0$	0.120 ± 0.012
WGAN-GP-PR	0.640 ± 0.019	0.848 ± 0.028	0.739 ± 0.034	$\ddagger 0.0 \pm 0.0$	0.078 ± 0.014
DECAF-PR	0.717 ± 0.021	0.839 ± 0.033	0.769 ± 0.020	$\ddagger 0.0 \pm 0.0$	0.044 ± 0.013
DECAF-ND	0.780 ± 0.023	0.920 ± 0.045	0.781 ± 0.007	0.152 ± 0.013	0.198 ± 0.013
DECAF-FTU	0.763 ± 0.033	0.925 ± 0.040	0.765 ± 0.010	0.004 ± 0.004	0.054 ± 0.005
DECAF-CF	0.743 ± 0.022	0.875 ± 0.038	0.769 ± 0.004	0.003 ± 0.006	0.039 ± 0.011
DECAF-DP	0.781 ± 0.018	0.881 ± 0.050	0.672 ± 0.014	0.001 ± 0.002	0.001 ± 0.001

2.1 Interpretation of Metrics

Table 1 and Table 2 respectively show results of bias removal on the Adult [2] and Credit Approval data sets. Looking at which, there are a few inconsistency to point out:

- The original paper claims minimal loss of data utility as one of its goal, yet did not providing further explanation and definition about what is considered to be minimal data utility loss. And while DECAF achieves the same level of AUROC compared to FairGAN - which is another state-of-the-art method, there is still a significant gap in terms of AUROC score between DECAF and DECAF-ND (in Table 1). Here DECAF-ND can be seen as the upper bound of utility score of DECAF and its variations. This issue is also pointed out by [4].
- In Table 1, while both WGAN-GP and DECAF has higher AUROC scores compared to their *-PR counterparts, GAN does not. It is intuitive to assume that the removal of protected attributes should not lead to increase in data utility, assuming there is causal relations between the protected attributes and the target attribute in the underlying DAG and the provided DAG is correct. Furthermore, the opposite effect appears in Table 2 where GAN has better AUROC score than GAN-PR but WGAN-PR and ADSGAN-PR have better AUROC scores than their non-PR counterparts.
- The original paper mentioned the trade-off between data utility and fairness, but did not include any metric to account for this. It would be easier to compare the quality between debiased synthetic data sets should there be a metric that evaluates the trade-off between the two aspects.

Table 2: (Table 6 form the original paper) Bias removal experiment on Credit Approval dataset.

	Data Quality			Fairness	
Method	Precision↑	Recall↑	AUROC↑	DP↓	FTU↓
GAN	0.921 ± 0.036	0.335 ± 0.029	0.743 ± 0.047	0.405 ± 0.077	0.194 ± 0.058
WGAN	0.970 ± 0.007	0.804 ± 0.057	0.698 ± 0.009	0.520 ± 0.036	0.461 ± 0.029
ADSGAN	0.963 ± 0.009	0.841 ± 0.052	0.708 ± 0.009	0.506 ± 0.013	0.429 ± 0.059
GAN-PR	0.794 ± 0.117	0.368 ± 0.080	0.727 ± 0.047	0.203 ± 0.196	0.0 ± 0.0
WGAN-PR	0.941 ± 0.004	0.880 ± 0.017	0.814 ± 0.019	0.406 ± 0.022	0.0 ± 0.0
ADSGAN-PR	0.945 ± 0.008	0.880 ± 0.019	0.827 ± 0.008	0.413 ± 0.029	0.0 ± 0.0
FairGAN	0.951 ± 0.012	0.663 ± 0.046	0.680 ± 0.008	0.510 ± 0.075	0.474 ± 0.054
DECAF	0.954 ± 0.012	0.601 ± 0.015	0.713 ± 0.045	0.511 ± 0.130	0.432 ± 0.127
DECAF-FTU	0.936 ± 0.017	0.901 ± 0.034	0.877 ± 0.009	0.099 ± 0.065	0.014 ± 0.012
DECAF-DP	0.940 ± 0.007	0.922 ± 0.024	0.875 ± 0.010	0.011 ± 0.029	0.015 ± 0.017

2.2 Choice of Metrics

It is usually the case for biased data sets to have unbalanced distributions skewed towards the negative class. For example, in the Adult data set [2], the '<=50k' class is 76.07% and the '>50k' class is 23.93%. Futhermore, the AUROC metric is known to be unsuitable for unbalanced data sets [3]. As it measures both True Positive Rate and False Positive Rate, a high value of AUROC might due to large False Positive instead of large True Positive.

Additionally, a possible effect of debiasing negative-skewed dataset is that its debiased version would contains many more positive samples, which in turns makes any classifer that trained on it to produce more positive predictions. This prediction when compared with the original label will result in a lot of False Positive, thus decreasing AUROC score. Moreover, as the debiased dataset is assumed to be fair, it is unlikely that it will be skewed towards the opposite side, the AUROC will then tend to come closer to 0.5 in stead of 0 as the dataset is becoming fairer and fairer.

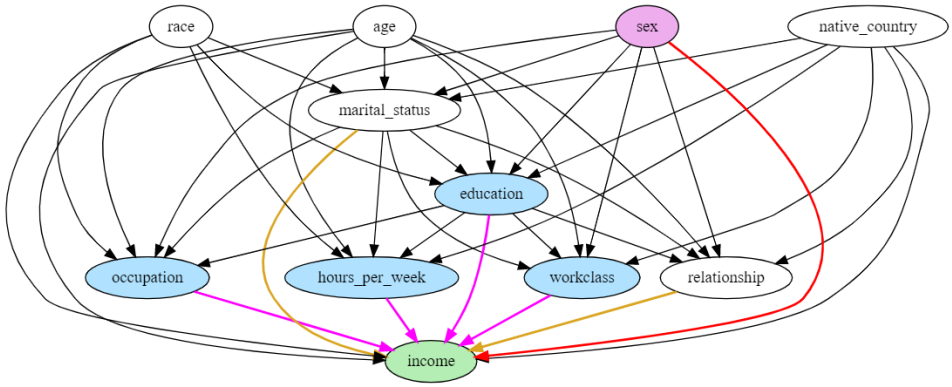

Figure 1: (Figure 6 from the original paper) Adult dataset DAG. The target variable is in green, the protected attribute in purple, and the allowed CF variables in blue. *FTU is achieved by removing:* ✗ *; DP:* ✗ ✗ ✗ *; CF:* ✗ ✗ .

2.3 Data Utility of the Target attribute

As the removal of edges is to block the flow of information from the attribute attributes A to the target attribute Y, this might lead to low data utility in generating the target value. By removing edges to the target attribute from its parents that are on the path from A to Y, the information flow from A to Y is block, but not only that, it will also block the information flow from the meditors along the paths to Y. For example, in Figure 1, edge from *education* to *income* is removed because *education* is a mediator on the path from *sex* to *income*. While the information from *sex* through *education* to *income* is biased, there could exists unbiased influence from *education* to *income*. Thus

such removal of edges has an unnecessary effect on the data utility of *income*. The next section will introduce a new extension to the existing DECAF that overcomes this issue.

3 Method

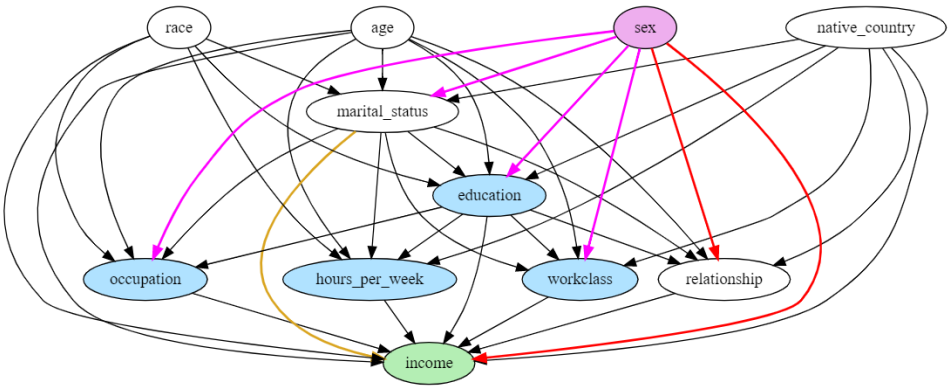

Figure 2: (Figure 6 from the original paper) Adult dataset DAG for generate *income*. The target variable is in green, the protected attribute in purple, and the allowed CF variables in blue. *DP is achieved by removing:* ✗ ✗ ✗ *; CF:* ✗ ✗ .

3.1 Fairness-Utility Scores

As discussed in the previous section, AUROC is not suitable for data sets with skewed amount of negative values examples, which is usually the case for real biased data sets. Instead the F1-score metric should be used as it is more attenuated by skewed distributions, this is also the recommended metric widely used in practice for said type of data [**jen2013facing**]. As such, this work includes F1-score to be another metric for data utility, along side with the AUROC score.

To evaluate teh Fairness-Utility trade-off, this work proposes to use a harmonic mean between the utility metric and the fairness metric, i.e. between each of *F1-score* and *AUROC* and each of *1 - FTU* and *1 - DP*.

3.2 Alternating Graph Generation

To improve the data utility while generating the target label Y, one would want to maximize the information flow to Y while not violate any d-separatedness constraints of the fairness definitions (FTU, CF, DP) as proposed in the DECAF paper [1]. This can be achieved by changing the set of edge to be removed, the idea is to prioritize edges closer to the protected attributes A rather than closer to Y and prioritize information flow into node with more children.

- FTU: As FTU focuses only on direct discrimination, no other set of edges can be removed to increase data utility.
- DP: All edge A -> B for every B who is an ancestor of Y (A is the protected attribute)
- CF: All edge A -> B if no node in the explanatory factors R is a mediator between A and Y, all edge B -> Y if otherwise.

This however will lead to leaking bias since parents of Y is generated before Y. Hence this work uses 2 separated DAGs to perform generation. All attributes except Y is generated as before using the set of node removal described in the DECAF paper. Then alternate to another graph as described above to generate Y. This help increase the information flow into Y with a simple extension and can be done during inference time. An implementation of the Alternating DAG Generation (ADG) for the target attribute of the Adult data set is shown in Figure 2.

Table 3: Experimental results on the Adult dataset. *-y indicates method using alternating DAG generation for the target attribute.

Method	Data Quality				Fairness		Fairness-Utility			
	precision	recall	f1-score	auroc	ftu	dp	ftu-f1	ftu-auroc	dp-f1	dp-auroc
original	0.877+-0.005	0.930+-0.007	0.903+-0.001	0.764+-0.008	0.030+-0.009	0.175+-0.012	0.935+-0.004	0.855+-0.006	0.862+-0.007	0.793+-0.002
decaf_nd	0.887+-0.021	0.758+-0.038	0.816+-0.015	0.729+-0.023	0.089+-0.043	0.347+-0.061	0.861+-0.023	0.809+-0.017	0.724+-0.041	0.686+-0.024
decaf_ftu	0.888+-0.016	0.759+-0.031	0.818+-0.013	0.732+-0.018	0.028+-0.020	0.297+-0.039	0.888+-0.011	0.835+-0.013	0.756+-0.027	0.716+-0.016
decaf_cf	0.777+-0.013	0.879+-0.042	0.824+-0.015	0.551+-0.028	0.036+-0.022	0.041+-0.029	0.889+-0.013	0.701+-0.026	0.886+-0.018	0.699+-0.018
decaf_dp	0.762+-0.009	0.914+-0.034	0.831+-0.015	0.518+-0.021	0.021+-0.021	0.016+-0.012	0.899+-0.010	0.677+-0.014	0.901+-0.010	0.679+-0.017
decaf_cf-y	0.758+-0.009	0.970+-0.025	0.851+-0.006	0.509+-0.021	0.012+-0.011	0.021+-0.021	0.914+-0.007	0.672+-0.016	0.910+-0.011	0.669+-0.014
decaf_dp-y	0.756+-0.005	0.976+-0.020	0.852+-0.007	0.504+-0.012	0.014+-0.010	0.020+-0.019	0.914+-0.008	0.667+-0.010	0.911+-0.011	0.665+-0.013

3.3 Experiment

An experiment on the the Adult data set is conducted and results are reported in Table **??**. The methods with ADG is shown to achieve better f1-score while still having similar level of fairness. When considering the Fariness-Utility score, methods with ADG consistently outperform their non-ADG counterparts on ftu-f1 and dp-f1. For AUROC, the score goes closer to 0.5 as the less biased the methode is. This agrees with results from the orginal paper and the discussion on Section 2.2. [2]

Note on implementation As the official code was incomplete and insufficient to use, this work instead adapts the code from [4] which is an re-implementation of DECAF. While this code make several changes to the original implementation, I found it sufficient as I agree with their interpretation of the original paper, as well as they was able to reproduce the experiments on the Adult dataset to some degree, which is also the experiment focused in this work.

4 Conclusion and Future works

This work proposed a novel extension to the DECAF method called Alternating DAG Generation which decreases the loss of utility significantly while maintaining the same level of fairness. It also suggested several metrics to more effectively measure data utility and the trade-off between data utility and fairness.

For future work, it would be desirable to more formally prove the correctness of the proposed methods, as well as perform more experiments such as on synthetically injected bias data set. The idea of using alternating DAG can also be extended to solving longitudinal data with underlying dynamic DAG, where the DAg is changing over time and the DAG from one time step is used to calculate root attributes of the DAG at the next time step. This can also be applied to modal Cyclic Graph and Undirected Graph by alternatingly considering edges inside loops.

References

[1] Boris van Breugel et al. "DECAF: Generating fair synthetic data using causally-aware generative networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22221–22233.

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[3] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. 2013, pp. 245–251. DOI: 10.1109/ACII.2013.47.

[4] Shuai Wang et al. "Replication Study of DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks". In: *ML Reproducibility Challenge 2021 (Fall Edition)*. 2022. URL: https://openreview.net/forum?id=SVx46hzmhRK.

[2]The code is available at https://github.com/lone17/DECAF