PRELIMINARY VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference

Constructing a Fair Classifier with the Generated Fair Data

Taeuk Jang[1], Feng Zheng[2], Xiaoqian Wang[1]*
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, 47907
[2]Department of Computer Science and Technology, Southern University of Science and Technology, Shenzhen, China, 518055
jang141@purdue.edu, zfeng02@gmail.com, joywang@purdue.edu

Abstract

Fairness in machine learning is getting rising attention as it is directly related to real-world applications and social problems. Recent methods have been explored to alleviate the discrimination between certain demographic groups that are characterized by sensitive attributes (such as race, age, or gender). Some studies have found that the data itself is biased, so training directly on the data causes unfair decision making. Models directly trained on raw data can replicate or even exacerbate bias in the prediction between demographic groups. This leads to vastly different prediction performance in different demographic groups. In order to address this issue, we propose a new approach to improve machine learning fairness by generating fair data. We introduce a generative model to generate cross-domain samples *w.r.t.* multiple sensitive attributes. This ensures that we can generate infinite number of samples that are balanced *w.r.t.* both target label and sensitive attributes to enhance fair prediction. By training the classifier solely with the synthetic data and then transfer the model to real data, we can overcome the under-representation problem which is non-trivial since collecting real data is extremely time and resource consuming. We provide empirical evidence to demonstrate the benefit of our model with respect to both fairness and accuracy.

Introduction

Machine learning has achieved great success in many fields due to its great flexibility and power to represent various types of data. As the model learns from the data to improve the model prediction, it may get influenced by the bias in the data and cause fairness issues. The discrimination in prediction pervasively exists in all kinds of applications including criminal justice (Angwin et al. 2016), banking (Lee and Floridi 2020), and hiring (Raghavan et al. 2020), where state-of-the-art models have been found to make biased predictions towards different demographic groups (*i.e.,* behaving differently when predicting for samples from different gender, race, or age group). For example, the risk assessment tool COMPAS has been found to give more false alert (mistakenly predicting people as future crime) for African-Americans than Caucasians with similar profile (Dressel and Farid 2018; Angwin et al. 2016) - higher false positive rate in *unprivileged group*. In addition, a recent algorithm for advertisement recommendation shows different performance in different gender groups, such that the model tends to promote more job advertisements related with Science, Technology, Engineering, and Math (STEM) fields to men than women (Lambrecht and Tucker 2019).

One major reason behind this inequality in model prediction is data bias. As the collected data are based on human-made decisions, data itself can be biased due to under-representation, sample size disparity, or existence of sensitive relevant features in the data *e.g.,* length of hair, location of residence. As a matter of fact, the predictive bias occurs because the model replicates or even amplifies the bias in the data. The biased prediction towards the unprivileged group of population greatly impairs the trustworthiness in the model and also results in enormous economical and societal issues. As such, fairness in machine learning has becoming a rising concern.

In order to mitigate the *prediction outcome discrimination* and *prediction quality disparity* (Du et al. 2019), recent works have been proposed from different perspectives. One strategy is to minimize the influence of the sensitive attribute - reformulating the data to be independent of the sensitive attribute (Wang and Huang 2019; Zemel et al. 2013; Madras et al. 2018; Adel et al. 2019). In addition to addressing the influence of the sensitive attribute, there are various methods proposed to deal with the bias originated from sample size disparity or under-representation (Kamiran and Calders 2012; Jiang and Nachum 2019). Meanwhile, other works (Frid-Adar et al. 2018) adopt generative adversarial network (GAN) to balance the dataset by augmenting data in the under-represented group.

In spite of the recent efforts, it is inevitable that the processed data still contains bias - such as remaining dependence on the sensitive attribute in the reformulated data, or disparity in data quality. For methods that address the influence of sensitive attribute by cleaning the raw data or learning a fair representation, the models are designed to handle the data bias while sacrificing minimum predictive performance - which will introduce a trade-off between fairness and predictive performance (*e.g.,* accuracy). In the fairness and accuracy trade-off, the processed data may not be "fair" enough when the remaining dependence on the sensitive attribute is needed to ensure accuracy (*i.e.,* the reformulated

*Corresponding author.

data representation is not independent of the sensitive attribute). For methods that address the sample size disparity or under-representation problem by re-sampling or adding synthetic data, since there is disparity in the data quality among different demographic groups and disparity in data distribution when mixing the real and synthetic data, there is disparity in data quality even though the number of samples is the same among demographic groups. The predictive model built on such processed data will get affected by the remaining bias and make discriminatory prediction.

In this paper, we propose to address these challenges from a new perspective: building the model on purely generated data that is balanced (to ensure fairness), and then transfer to real data for fine-tuning (to ensure performance). We use the Variational Autoencoder-Generative Adversarial Network (VAE-GAN) (Larsen et al. 2015) to approximate the data distribution and generate data with desired sensitive attribute and label. The generator in our model facilitates the generation of data with balanced sample from different demographic groups (characterized by sensitive attribute) and classes (characterized by target label). Since the data we use to train the predictive model is entirely from the generator, the sample size is controllable to address the under-representation problem (when different sample size or information is available for different demographic group), and there is no disparity in data quality or distribution. Moreover, we use transfer learning technique to fine tune the model with real data. This makes the model to align well with the real data for the downstream classification task and enhance predictive performance.

We would like to summarize the contributions of our paper as following:

- To the best of our knowledge, this is the first work that uses transfer learning to mitigate the problem of unfair prediction.
- We build a generative model based on VAE-GAN to provide balanced cross-domain data to train a classifier, which addresses the problem of unfairness or imbalance in classification without the burden of collecting more real data.
- We empirically validate our model on benchmark fairness datasets and validate that the transfer learning method improves fairness under different fairness metrics while maintaining comparable predictive performance with state-of-the-art methods.

Related work

In order to achieve a fair outcome from the given unfair data, various works has been proposed to improve prediction output discrimination and prediction quality disparity (Du et al. 2019). To accomplish this goal, some researchers propose to refine the data to be fair (Calmon et al. 2017; Zemel et al. 2013), others revise the classification model to narrow the disparity in the prediction between minority and majority with minimum sacrifice in performance (Woodworth et al. 2017; Zafar et al. 2017; Wang and Huang 2019), or post-process the predictive outcome to be fair given a black-box model in a model-agnostic manner (Hardt, Price, and Srebro 2016).

Data processing methods for fairness are motivated by the fact that data itself is responsible for the inequity because the data collected by humans includes historical discrimination. Also, unbalance in sample size or under-representation of the data in the unprivileged group causes inequality in the prediction. To address this, Chen, Johansson, and Sontag (2018) suggest to consider inadequate or skewed sample size problem to examine discrimination caused by dataset. On the other hand, Escalera et al. (2016) put effort on making fair dataset and introduce the *Faces of the World* dataset which has equally distributed samples with respect to sensitive attributes to earn unbiased results from the balanced data. However, the collection of balanced real data is usually expensive and time-consuming, sometimes even impossible due to security concerns.

To ease the burden of collecting more data, recent research designs adversarial networks to learn a fair representation to filter out the negative impact of sensitive attributes. For example, Madras et al. (2018) propose to learn a representation adversarially with a certain fairness metric as an objective. Subsequent work (Adel et al. 2019) introduces a single network to improve stability of adversarial networks by combining the networks from previous works that had small individual parts.

Our work is inspired by and most related to (Koh and Liang 2017) and (Choi et al. 2018). Koh and Liang (2017) introduce influence function based explanation to examine how particular data point affects the performance of a black-box network. The explanation is obtained by observing the change of loss via re-weighing a certain data point or perturbing the features or labels of the data point. Similarly, in an unbalanced dataset, the model is likely to weigh more in the larger group with more samples and cause imbalance in the prediction. Previous pre-processing works mainly focus on how to effectively remove the impact of the sensitive attribute and learn the representation from which the sensitive attribute cannot be inferred. With such approach, it still cannot overcome the data imbalance and under-representation problem.

In contrast, to train a fair classifier, we generate perfectly balanced synthetic data and transfer to real data to fine-tune the model for classification. Inspired by StarGAN (Choi et al. 2018), we introduce a VAE-GAN model that can translate data in multiple domains so that we can generate infinite number of samples with different combinations to learn a fair classifier. Our generative model is different from StarGAN in that we use VAE-GAN instead of GAN model as the generator, which results in disentangled factors of variation in learning the cross-domain data. Learning fair classifier with balanced synthetic data has significant advantage because acquiring not only good quality but also fair dataset is extremely resource consuming in most learning scenarios.

FairGAN (Xu et al. 2018) also introduce GAN model to ensure fairness, which generates the synthetic data independent to the sensitive attribute that leads to fair classification. Moreover, Sattigeri et al. (2019) extend the FairGAN to multimedia datasets such generating the images from different

gender while imposing equality of opportunity. However, to the extent of our knowledge, this is the first work to build a fair classifier with transfer learning. We generate synthetic data that resembles the original dataset that has different value for sensitive attribute. This allows us to learn fair classifier that can transfer learning pipeline as its realistic and together with the original dataset.

Generating Fair Data for Fair Classification

Motivations

For simplicity, here we consider the binary classification problem as an example, where there is one sensitive attribute in the data and the sensitive attribute can take two possible values. It is notable that the discussion can be easily adapted to other cases, *e.g.,* multiple values of the sensitive attribute or multiple sensitive attributes.

Denote $\mathcal{X} \subset \mathbb{R}^p$ as a compact input space, $\mathcal{Y} = \{0, 1\}$ as the set of labels, and $\mathcal{A} = \{0, 1\}$ as the set of sensitive attribute values. We use $Y = 1$ to denote the preferable label (*e.g.,* getting approved in loan application), and use $A = 0$ to denote the sensitive attribute of the unprivileged group (*e.g.,* race of black in the COMPAS future crime prediction example in Introduction).

Denote $\hat{Y} \in \{0, 1\}$ as the predicted outcome. For the sake of fair performance between different groups, there are several fairness metrics can be considered. For example, the definition of demographic parity (Dwork et al. 2012; Kusner et al. 2017) is that the likelihood of a positive outcome is the same in different demographic groups:

$$\mathrm{P}(\hat{Y} = 1|A = 0) = \mathrm{P}(\hat{Y} = 1|A = 1)\,,$$

yet demographic parity permits random assignment of preferable labels within each group and also conflicts with the perfect classifier ($\hat{Y} = Y$) when the base rate among different groups is different (*i.e.,* when $\mathrm{P}(Y = 1|A = 0) \neq \mathrm{P}(Y = 1|A = 1)$).

The definition of equal true positive rate (TPR), also known as equal opportunity) (Hardt, Price, and Srebro 2016) is

$$\mathrm{P}(\hat{Y} = 1|Y = 1, A = 0) = \mathrm{P}(\hat{Y} = 1|Y = 1, A = 1)\,.$$

Similarly, the definition of equal false positive rate (FPR) or equal true negative rate (TNR) (Chouldechova 2017) is

$$\mathrm{P}(\hat{Y} = 0|Y = 0, A = 0) = \mathrm{P}(\hat{Y} = 0|Y = 0, A = 1)\,,$$

Equal opportunity and equal FPR avoids the random label assignment in demographic parity and also allows the perfect classifier. But it is difficult to directly impose equal TPR and equal FPR as constraints in training the model. In contrast, we focus on the overall accuracy in supervised learning, which can be rewritten as

$$\begin{aligned}
&\mathrm{P}(\hat{Y} = Y) \\
=\ &\mathrm{P}(\hat{Y} = 1, Y = 1) + \mathrm{P}(\hat{Y} = 0, Y = 0) \\
=\ &\mathrm{P}(\hat{Y} = 1|Y = 1, A = 0)\mathrm{P}(Y = 1, A = 0) \\
&+\mathrm{P}(\hat{Y} = 0|Y = 0, A = 0)\mathrm{P}(Y = 0, A = 0) \\
&+\mathrm{P}(\hat{Y} = 1|Y = 1, A = 1)\mathrm{P}(Y = 1, A = 1) \\
&+\mathrm{P}(\hat{Y} = 0|Y = 0, A = 1)\mathrm{P}(Y = 0, A = 1)\,,
\end{aligned} \tag{1}$$

which is a weighted sum of TPR and TNR in different demographic groups, with the group-wise joint probabilities being the weights. The weights are determined by the data distribution itself and cannot be changed by the predictive model.

If the weights are different, *i.e.,* $\mathrm{P}(Y = 1, A = 0) \neq \mathrm{P}(Y = 1, A = 1)$, then there is disparity in TPR between demographic groups when the model is optimizing *w.r.t.* the overall accuracy in Eq. 1 - since larger groups get larger weights thus preference in optimization. The same also holds for the TNR disparity if $\mathrm{P}(Y = 0, A = 0) \neq \mathrm{P}(Y = 0, A = 1)$.

Similarly, we can rewrite the likelihood of a positive outcome as:

$$\begin{aligned}
&\mathrm{P}(\hat{Y} = 1|A = 0) \\
=\ &\mathrm{P}(\hat{Y} = 1|Y = 1, A = 0)\mathrm{P}(Y = 1|A = 0) \\
&+\mathrm{P}(\hat{Y} = 1|Y = 0, A = 0)\mathrm{P}(Y = 0|A = 0)\,,
\end{aligned} \tag{2}$$

and

$$\begin{aligned}
&\mathrm{P}(\hat{Y} = 1|A = 1) \\
=\ &\mathrm{P}(\hat{Y} = 1|Y = 1, A = 1)\mathrm{P}(Y = 1|A = 1) \\
&+\mathrm{P}(\hat{Y} = 1|Y = 0, A = 1)\mathrm{P}(Y = 0|A = 1)\,,
\end{aligned} \tag{3}$$

such that the likelihood of a positive outcome is a weighted sum of TPR and FPR in different demographic groups, with the base rates being the weights. The demographic parity and equal TPR/FPR cannot be obtained at the same time if the base rate of data is different.

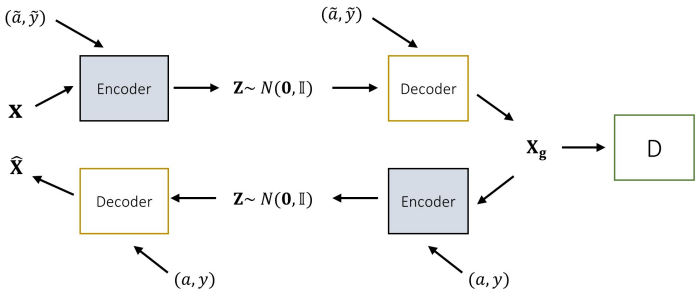
Thus, when the model is built on data with different base rates, there is a natural disparity in TPR, FPR, and likelihood of positive outcome among different demographic groups from the optimization *w.r.t.* the overall accuracy. In this case, fairness will get sacrificed for accuracy or vice versa. To address this, we propose to build data with balanced sample such that the joint probability $P(A, Y)$ is the same among all possible values of A and Y.

In addition, when the data from unprivileged groups is underrepresented and difficult to predict, difference in data quality can be another cause of discrimination in prediction. It cannot be addressed by simply re-sampling or adding generated data to balance the sample size, since the real and generated data comes from different distribution. Disparity in quality of training data among demographic groups still exists in the mixture of data.

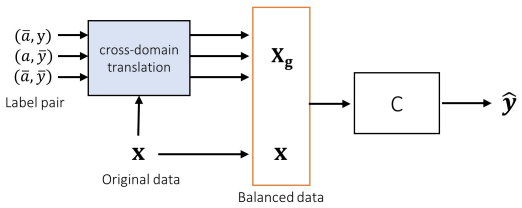
In this paper, we propose a generative model to generate cross-domain data with desired sensitive attribute and target label. We build a fair classifier on the generated data with balanced samples (to ensure fairness) and transfer the model to the real data to fit the prediction task (to ensure predictive performance). Figure 1 illustrates the process of generating synthetic data and training the fair classifier.

Overview of Generative Model

Generative Adversarial Network (GAN) (Goodfellow et al. 2014) proposes a min-max game between the generator G and discriminator D to approximate the data distribution.



(a) Generative model for cross-domain translation



(b) Transfer learning for constructing a fair classifier

Figure 1: The process of constructing a fair classifier includes: 1) learn a generative model for cross domain translation. The encoder E takes the original data $\mathbf{x}$ and the desired sensitive attribute $\tilde{\mathbf{a}}$ and label $\tilde{\mathbf{y}}$ as the input to approximate the prior latent distribution $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$. The decoder F learns the translated data $\mathbf{x}_g$ to fool the discriminator D. The l-1 norm loss is measured between the original data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}$ (after flipping back the sensitive attribute and label); 2) train the classifier C with the generated data that is balanced in sensitive feature $\mathbf{a}$ and label $\mathbf{y}$; 3) fine-tune the classifier C with real data $\mathbf{x}$ along with its generated pair ($\mathbf{x}_g$ in Figure.1b).

The objective function of GAN model is:

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\big[\log D(\mathbf{x})\big] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\big[\log(1 - D(G(\mathbf{z}))\big], \tag{4}$$

where $p(\mathbf{x})$ is the real data distribution and $p(\mathbf{z})$ is a random distribution.

Many variations of GAN model have been proposed to generate synthetic data with good quality and desired properties. StarGAN (Choi et al. 2018) proposes a GAN model with a single generator network for image-to-image translation by changing the image to another domain (*e.g.,* blonde hair, aged face, or happy facial emotion).

In this paper, we generalize the cross-domain translation to non-image data, *i.e.,* translate data to a desired class or demographic group. Different from StarGAN, we build a new cross-domain translator on the basis of VAE-GAN (Larsen et al. 2015) (a mixture of variational autoencoder (VAE) and GAN), which results in disentangled factors of variation in learning the cross-domain data.

Fair Classification with Balanced Data

We propose a generative model based on VAE-GAN to generate *balanced* (same sample size in different groups) and *equal-quality* (all data generated from a unified distribution) for fair classification. Given a data $(\mathbf{x}, \mathbf{a}, \mathbf{y})$, denote $\bar{\mathbf{a}}$ and $\bar{\mathbf{y}}$ as the complement of $\mathbf{a}$ and $\mathbf{y}$ respectively, *e.g.,* $\bar{\mathbf{a}} = 1$ if $\mathbf{a} = 0$.

Our model consists of an encoder E, a decoder F, and a discriminator D. We denote the encoded data from the encoder as $\mathbf{z} \sim E(\mathbf{x}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}})) = q(\mathbf{z}|\mathbf{x}, \tilde{\mathbf{a}}, \tilde{\mathbf{y}})$, and the decoded data from the decoder as $\mathbf{x} \sim F(\mathbf{z}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}})) = p(\mathbf{x}|\mathbf{z}, \tilde{\mathbf{a}}, \tilde{\mathbf{y}})$. The encoder E is optimized with a prior regularization over the latent distribution $p(\mathbf{z})$ where $\mathbf{z} \sim \mathcal{N}(0, I)$. We consider the prior regularization as below:

$$\mathcal{L}_{prior} = \mathbb{E}_{(\mathbf{x},\mathbf{a},\mathbf{y})\sim p(\mathbf{x},\mathbf{a},\mathbf{y})}\big[\mathcal{D}_{KL}\big(E(\mathbf{x}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}}))||p(\mathbf{z})\big) + \mathcal{D}_{KL}\big(E(\mathbf{x}_g, (\mathbf{a}, \mathbf{y}))||p(\mathbf{z})\big)\big], \tag{5}$$

where $\mathcal{D}_{KL}$ denotes the KL-divergence, $(\tilde{\mathbf{a}}, \tilde{\mathbf{y}})$ is randomly sampled from the set of $\{(\bar{\mathbf{a}}, \bar{\mathbf{y}}), (\bar{\mathbf{a}}, \mathbf{y})\}$ (*i.e.,* flipping $\mathbf{a}$ with or without flipping $\mathbf{y}$), and

$$\mathbf{x}_g = F(E(\mathbf{x}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}})), (\tilde{\mathbf{a}}, \tilde{\mathbf{y}})).$$

The decoder F is optimized to maximize the expected log likelihood of the translated data $\mathbf{x}_g$:

$$\mathcal{L}_{like} = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x},\tilde{\mathbf{a}},\tilde{\mathbf{y}})}[-\log\big[F(\mathbf{z}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}}))\big]]. \tag{6}$$

Also, we consider the ℓ_1-norm reconstruction loss, *i.e.,* the reconstructed data is regularized to approximate the original data after flipping back the $\mathbf{a}$ and $\mathbf{y}$ values:

$$\mathcal{L}_{rec} = \mathbb{E}_{(\mathbf{x},\mathbf{a},\mathbf{y})\sim p(\mathbf{x},\mathbf{a},\mathbf{y})}\big[||\mathbf{x} - \hat{\mathbf{x}}||_1\big], \tag{7}$$

where

$$\hat{\mathbf{x}} = F(E(\mathbf{x}_g, (\mathbf{a}, \mathbf{y})), (\mathbf{a}, \mathbf{y})).$$

We optimize the discriminator loss and translation loss for updating the discriminator D:

$$\mathcal{L}_{dis} = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\big[\log D(\mathbf{x})\big] + \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x},\tilde{\mathbf{a}},\tilde{\mathbf{y}})}\big[\log(1 - D(F(\mathbf{z}, \tilde{\mathbf{a}}, \tilde{\mathbf{y}}))\big], \tag{8}$$

$$\mathcal{L}_{trns} = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x},\tilde{\mathbf{a}},\tilde{\mathbf{y}})}[-(\tilde{\mathbf{a}}, \tilde{\mathbf{y}}) \log D_{cls}\big((\tilde{\mathbf{a}}, \tilde{\mathbf{y}})|F(\mathbf{z}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}}))\big)\big], \tag{9}$$

where $D_{cls}\big((\tilde{\mathbf{a}}, \tilde{\mathbf{y}})|F(\mathbf{z}, (\tilde{\mathbf{a}}, \tilde{\mathbf{y}}))\big)$ denotes the predicted probability of $(\tilde{\mathbf{a}}, \tilde{\mathbf{y}})$ by D given the decoded data from F.

After we build the generative model for domain translation, we generate balanced data to train a fair classifier and transfer the model to the real data for fine-tuning. We summarize our model in Algorithm 1.

Experiments

In this section, we examine the proposed method on how the generated balanced data and transfer learning affects the fairness and accuracy of the classifier by comparing with state-of-the-art fairness methods.

Algorithm 1 Optimization Procedure of Our Method

Input dataset $\{(\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where $\mathbf{x}$ is the input feature vector, $\mathbf{a}$ is the sensitive attribute, and $\mathbf{y}$ is the target label.
Output Encoder E, decoder F, and a fair classifier C.
Initialize E, F, C randomly
while not converge **do**
  for t = $1,\ 2,\ \ldots,\ n_b$ **do**
    1. Update encoder E and decoder F to minimize $\mathcal{L}_{vae}$ using Adam optimization
    such that, $\mathcal{L}_{vae} = \mathcal{L}_{like} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{prior}\mathcal{L}_{prior} + \mathcal{L}_{dis}$
    2. Update discriminator D to minimize $\mathcal{L}_D$ with Adam optimization
    such that, $\mathcal{L}_D = \mathcal{L}_{dis} + \mathcal{L}_{trns}$
  end for
end while
STEP 2 Train a preliminary fair classifier C with synthetic balanced data from decoder F with arbitrary random vectors and variations of target domains (**a** and **y**).
STEP 3 Transfer the classifier C to real data **x** with corresponding synthetic pair $\mathbf{x}_g$.

Experimental Setup

We implement the experiments on the four fairness datasets:

- **Adult**: data from the UCI repository (Kohavi 1996): The data contains 48,842 instances described by 14 features (workclass, age, education, sex, race, *etc.*) and the goal is to predict whether the income exceeds 50K USD per year. The feature *sex* is used as the sensitive feature.
- **Compas**[1]: The dataset includes 6,167 samples described by 401 features with the outcome showing if each person gets rearrested within two years. The feature *sex* is used as the sensitive feature in this dataset.
- **German** credit data from the UCI repository (Dua and Graff 2019): The dataset contains 1,000 samples described by 20 features and the goal is to predict the credit risks. The feature *sex* is used as the sensitive feature.
- **MEPS**[2]: The Medical Expenditure Panel Survey (MEPS) dataset contains 15,830 samples with 138 features. The dataset consists of large-scale health assessment surveys from diverse demographics and the outcome shows how much health services was used by each participant. The feature *race* is used as the sensitive feature.

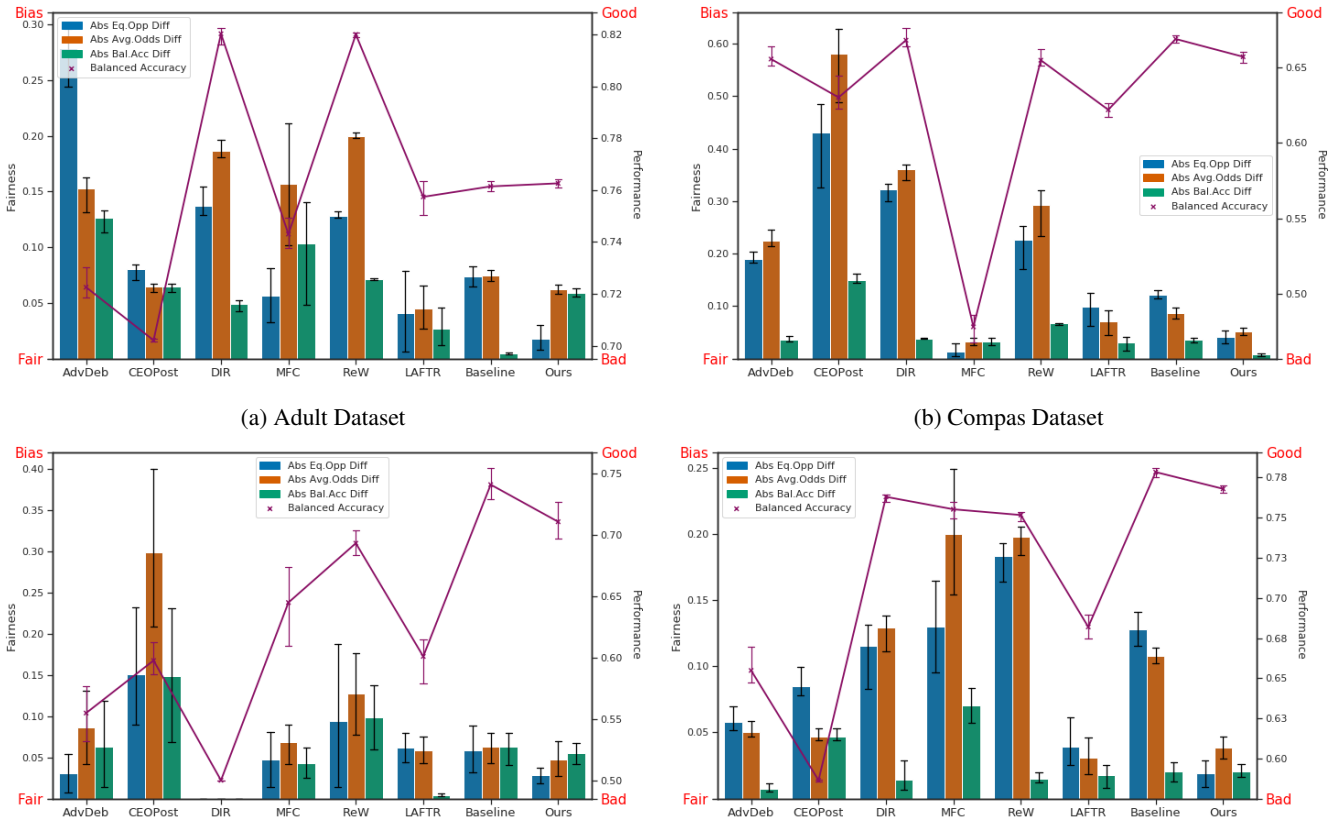
Each dataset is randomly split with the ratio of training, validation, and test sets being 70%, 15%, and 15%. We report the results on test dataset with 5 repetitions. Detail of the datasets is in the Supplementary material.

To verify the effectiveness of our method, we compare with the following related state-of-the-art methods: **Adversarial de-biasing** (abbreviated as AdvDeb) (Zhang, Lemoine, and Mitchell 2018) is an in-processing method that address conflicting gradient directions between accuracy and fairness objectives by projecting one gradient to another. **Calibrated equal odds post-processing** (abbreviated as CEOPost) (Pleiss et al. 2017) is a post-processing model that proposes relaxation method which minimizes the disparity in certain fairness metric to the preferred class among different sensitive groups, while maintaining the calibration condition. **Disparate impact remover** (abbreviated as DIR) (Feldman et al. 2015) is a pre-processing model that minimizes the demographic disparity between different sensitive groups. **Meta fair classifier** (abbreviated as MFC) (Celis et al. 2019) is an in-processing method that optimizes the classifier constrained on multiple metrics *e.g.,* statistical parity, or false discovery. **Re-weighing** (abbreviated as ReW) (Kamiran and Calders 2012) is a pre-processing model that mitigates data bias by re-sampling or re-weighing to data from different sensitive groups based on the size of the certain sensitive group. and **Learning Adversarially Fair and Transferable Representations** (abbreviated as LAFTR) (Madras et al. 2018) is a fair representation learning model that adopts fairness metrics as the adversarial objectives and analyze the balance between accuracy and fairness. We also include the **Baseline** method in the comparison, which is the classifier that has the same structure with ours and no other fairness methods applied. The purpose of this comparison is to show that transfer learning can contribute better to fairness with small sacrifice in performance compared to simply balance in the original dataset by discarding or duplicating samples. By adopting this strategy we can achieve the results that outperform or comparable to the comparing methods in both fairness and performance in most datasets.

As most of the datasets are unbalanced, we adopt balanced accuracy (average of true positive rate and true negative rate) to measure the accuracy rather than directly choosing accuracy. To measure fairness, we adopt three well-known fairness metrics: Absolute Balanced accuracy difference (balanced accuracy difference in different protected group), Absolute average odds difference (absolute difference in true positive rate and false positive rate between different protected groups), Absolute equal opportunity rate difference (absolute difference in the prediction accuracy of the preferred label in different protected group). These metrics are abbreviated as Abs Bal Acc Diff, Abs Avg. Odds Diff, and Abs Eq.Opp Diff respectively in the following context.

In our model, we design the encoder E with three dense layers with layer normalization (Ba, Kiros, and Hinton 2016) and ReLU followed by two residual layers. Decoder F has symmetric structure to the encoder. Discriminator D consists of three dense layers with leaky ReLU activation. The classifier C consists of three dense layers with ReLU and dropout with the probability of 0.7. We first train classifier with synthetic data only until validation accuracy converges. After that, we fine-tune the model with the real data and corresponding generated pairs with flipped values in sensitive attribute and target label. We conducted the experiments on a Quadro RTX 6000 GPU and Intel I9-9960X on Pytorch and Tensorflow framework.

[1]https://github.com/propublica/compas-analysis
[2]https://meps.ahrq.gov/mepsweb/



(a) Adult Dataset

(b) Compas Dataset

(c) German Dataset

(d) MEPS Dataset

Figure 2: Comparison *w.r.t.* fairness (left vertical axis) via three fairness metrics: absolute equal opportunity difference, absolute average odds difference, and absolute balanced accuracy difference, and comparison *w.r.t.* accuracy (right vertical axis) via balanced accuracy. Lower values for all three fairness metrics indicates better performance in fairness. Higher values for balanced accuracy indicates better performance in classification.

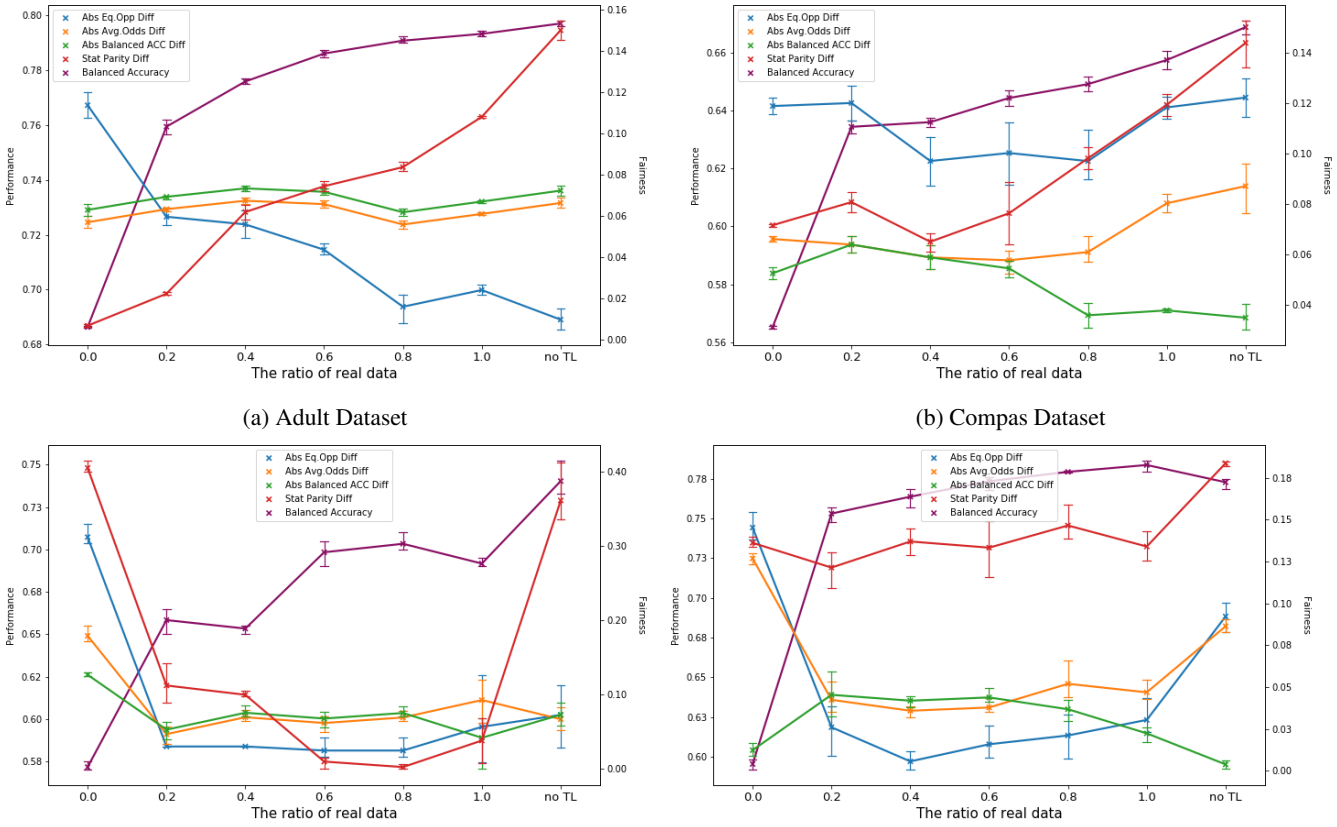
Quantitative Evaluation

In this subsection, we evaluate in both performance and fairness and summarize the results in Fig. 2. We could observe that the preliminary training on synthetic data helps the classifier to be more fair. As we can generate the synthetic data that is perfectly balanced, it achieves better results in fairness when the classifier is trained on the real data with its synthetic counterparts. The reported results of our transfer learning strategy is conducted with a 50:50 real-to-synthetic data ratio. This ratio is uniformly applied to all datasets and not manually selected. We chose this pairing mechanism since our motivation is to balance the dataset.

Especially in Meps and Compas datasets, we could achieve a classifier with small discrimination that has large gap between other comparing methods while retaining the performance. To achieve better performance, we fine-tuned our classifier with balanced real data with their synthetic pair *w.r.t.* to both target label and sensitive attribute. By transfer learning, we compensate the performance and get better or equivalent performance in all datasets. At the end, we could get best results in fairness with comparable performance.

Moreover, we executed transfer learning with downsampled dataset to balance the data *w.r.t.* both target labels and sensitive attribute by discarding samples that exceeds the number of the smallest group and adding its synthetic pairs. By doing so, we observed that we get better result *w.r.t.* both fairness and performance than merely using unbalanced original dataset. From this, we can infer the impact of the balance or under-representation of dataset on fairness is significant.

In order to evaluate the quality of synthetic data, we measure the similarity between real and generated distributions with FID score (Lucic et al. 2018). For the comparison, we calculate the baseline which measures FID score between real dataset and the dataset itself with additive normal distribution noise (zero mean and small std: 0.25(Adult), 0.12(Compass)). Note that the dataset is pre-processed that the maximum absolute value in each feature is normalized to 1.0. While the baseline FID score is 3.216 (Adult), and 4.056 (Compas), our synthetic dataset get 3.564 (Adult), and 3.581 (Compas) respectively. Since lower FID score indicates better synthetic data, it verifies the quality of our synthetic data and achieve comparable similarity with the real dataset compared to small noise injected data.



(a) Adult Dataset

(b) Compas Dataset

(c) German Dataset

(d) MEPS Dataset

Figure 3: Trade-off between fairness and accuracy when we transfer the pretrained model to different ratio of real data. Specially, we use "no TL" to show the results when we directly train the classifier with balanced real samples without pretraining with synthetic data. Lower values in fairness metrics (Abs Eq.Opp Diff, Abs Avg. Odds Diff, Bal Acc Diff, and Stat Parity Diff) indicate more fair results. Higher values in performance metric (balanced accuracy) indicate better performance.

Fairness performance trade-off

In this subsection, we evaluate how synthetic pairs contribute to the model fairness when we fine-tune the pretrained model pre-trained by varying the proportion of synthetic data in a batch. In Fig. 3, we plot the change of the fairness and performance *w.r.t.* the ratio of real and synthetic data for fine-tuning. In the fairness comparison, statistical parity difference indicates $|P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0)|$ to measure balanced prediction.

The classifier which is pre-trained with perfectly balanced synthetic data, we could achieve outperforming fairness when we implement transfer learning with certain ratio of real and synthetic data comparing with the classifier trained on only real data (no TL). We could empirically find the optimal ratio by sweeping the proportion of the real data to achieve fair classifier. For example, in the Compas dataset, we can get the best average fairness at the ratio of 0.8, considering the fairness metrics while preserving comparable performance. To compensate the performance, when we gradually increase the proportion of the real data, the performance improves and fairness deteriorates correspondingly.

Note that there is no consensus in fairness, there are some metrics moving opposite to the expected direction as some metrics conflict to each other. Therefore, this supports our assumption that merely training on original dataset is biased and synthetic pair improves the fairness by balancing the data.

Conclusion

In this paper, we introduce a novel strategy to achieve fair classification model utilizing synthetic data that can overcome the defect or short of dataset. In order to learn a fair classifier, we propose to train with perfectly balanced data *w.r.t.* target label and sensitive attribute and fine-tune with real data to guarantee the performance. Hereby, we can minimize the data discrimination from unbalanced and unfair dataset that recent machine learning or decision making models suffer from. This could resolve the problem of data acquisition which is hugely resource consuming. We demonstrate evidence on well-known datasets to support our statement that with synthetic pair to original data, we can achieve improvement fairness with minimum loss in accuracy.

Ethics Statement

In this paper, we take an advantage of perfectly balanced synthetic data to learn a fair classifier. Also this can be followed by diverse applications in different areas, it is adaptable to different downstream tasks.

However, there also exist potential limitation in generating synthetic data in some scenarios. When the dataset to train the VAE-GAN is imbalanced is an example. Similar to the problem that imbalanced dataset cause unfair classifier, it has potential risk of generating unreliable synthetic data on the smaller group. Therefore, we have some open research topics that how to obtain more robust synthetic data with imbalanced dataset and measure the quality of the fair synthetic data,

References

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-network adversarial fairness. In *AAAI*, volume 33, 2412–2420.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica 2016.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* .

Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *NeurIPS*, 3992–4001.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *ACM FAccT*, 319–328.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? In *NeurIPS*, 3539–3550.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 8789–8797.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2): 153–163.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4(1): eaao5580.

Du, M.; Yang, F.; Zou, N.; and Hu, X. 2019. Fairness in Deep Learning: A Computational Perspective. *arXiv preprint arXiv:1908.08843* .

Dua, D.; and Graff, C. 2019. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*, 214–226.

Escalera, S.; Torres Torres, M.; Martinez, B.; Baró, X.; Jair Escalante, H.; Guyon, I.; Tzimiropoulos, G.; Corneou, C.; Oliu, M.; Ali Bagheri, M.; et al. 2016. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *CVPR Workshops*, 1–8.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268.

Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; and Greenspan, H. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *ISBI*, 289–293. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*, 2672–2680.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NeurIPS*, 3315–3323.

Jiang, H.; and Nachum, O. 2019. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966* .

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *KAIS* 33(1): 1–33.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *ICML*, 1885–1894. JMLR. org.

Kohavi, R. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, 202–207.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *NeurIPS*, 4066–4076.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *MANAGE SCI* 65(7): 2966–2981.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* .

Lee, M. S. A.; and Floridi, L. 2020. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Available at SSRN* .

Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2018. Are gans created equal? a large-scale study. *Advances in neural information processing systems* 31: 700–709.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* .

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. In *NeurIPS*, 5680–5689.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *FAT*, 469–481.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM J. Res. Dev.* 63(4/5): 3–1.

Wang, X.; and Huang, H. 2019. Approaching Machine Learning Fairness through Adversarial Network. *arXiv preprint arXiv:1909.03013* .

Woodworth, B.; Gunasekar, S.; Ohannessian, M. I.; and Srebro, N. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* .

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, 570–575. IEEE.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 1171–1180.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*, 325–333.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AIES*, 335–340.