Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck

Jin-Young Kim and Sung-Bae Cho

Department of Computer Science, Yonsei University, Seoul 03722, South Korea
{seago0828, sbcho}@yonsei.ac.kr

Abstract

Algorithmic bias indicates the discrimination caused by algorithms, which occurs with protected features such as gender and race. Even if we exclude a protected feature inducing the unfairness from the input data, the bias can still appear due to proxy discrimination through the dependency of other attributes and protected features. Several methods have been devised to reduce the bias, but it is not yet fully explored to identify the cause of this problem. In this paper, non-discriminated representation is formulated as a dual objective optimization problem of encoding data while obfuscating the information about the protected features in the data representation by exploiting the unbiased information bottleneck. Encoder learns data representation and discriminator judges whether there is information about the protected features in the data representation or not. They are trained simultaneously in adversarial fashion to achieve fair representation. Moreover, the algorithmic bias is analyzed in terms of bias-variance dilemma to reveal the cause of bias, so as to prove that the proposed method is effective for reducing the algorithmic bias in theory and experiments. Experiments with the well-known benchmark datasets such as Adults, Census, and COMPAS demonstrate the efficacy of the proposed method compared to other conventional techniques. Our method not only reduces the bias but also can use the latent representation in other classifiers (i.e., once a fair representation is learned, it can be used in various classifiers). We illustrate it by applying to the conventional machine learning models and visualizing the data representation with t-SNE algorithm.

Introduction

Discrimination is the unfair treatment of individuals based on specific features, also called sensitive attributes such as gender and race [Helm, 2016; Shipman and Griffiths, 2016]. It has been found that machine learning, which has significantly led to constructing a model capable of deciding the labels of novel data, can lead to unexpected results with bias [Dwork et al., 2012; Feldman et al., 2015; Kamiran and Calders, 2010; Koene, 2017; Luong et al., 2011; Zemel et al., 2013]. Even machine learning algorithms have amplified algorithmic bias [Dressel and Farid, 2018]. However, algorithmic bias cannot be solved by removing the sensitive variables from the input. This problem is called 'proxy discrimination'. For example, even if we delete race information, it is possible to derive race by zip code.

The methods of mitigating bias are divided into three categories, as shown in Figure 1: pre-processing, in-processing, and post-processing [Calmon et al., 2017]. Pre-processing is to solve the problem by eliminating the bias present in the training data itself [Calmon et al., 2017; Edwards and Storkey, 2016; Grgic-Hlaca et al., 2018; Hajian, 2013; Louizos et al., 2017]. In-processing is to reduce the bias by adding a constraint to the learning algorithm even if there is a bias in the data [Fish et al., 2016; Kamishima et al., 2011; Zafar et al., 2017; Zhang et al., 2018]. Post-processing is to ensue decisions themselves [Hardt et al., 2016]. A more in-depth study is needed because of the lack of analysis on the cause of the algorithmic bias, suggesting only a fragmentary solution like fair representation or classifier. Some studies analyzed data as biased and proposed metrics to evaluate algorithmic bias but lack the theoretical background of the cause [Zafar et al., 2017; Zhang et al., 2018].

We formulate the problem of the algorithmic bias as a dual objective optimization that confuses the protected feature in latent space and fairly encodes the data for unbiased representation by using the information bottleneck [Alemi et al., 2017]. We propose an unbiased information bottleneck method to reduce the association of the protected attributes with the latent variable as shown in equation (1).

$$\max \mathcal{L}_{IB} = I(Z, X) - I(Z, A) \tag{1}$$

where I represents the mutual information, X is the data, Z is the latent variable and A means the protected attribute. Through the equation (1), we can maximize the relationship between representation and data features and at the same time minimize the relationship with features that cause bias. To minimize the $I(Z, A)$, we construct two models: encoder to learn data representation by projecting data into the latent space and discriminator to judge whether there is information about the sensitive attributes in the data representation. They are trained in adversarial and

achieve fair representation with the output of the encoder. To prevent the data from being distorted, i.e., to maximize the $I(Z, X)$, the decoder reconstructs the data with the output of the encoder so that the encoder should be trained to better represent the information in the data on the latent space. Besides, the algorithmic bias is analyzed from the bias-variance dilemma, a typical error analysis metric in machine learning. We also demonstrate theoretically that the difference in bias of each protected feature can be solved by achieving a fair representation, proving the validity of our method. To evaluate the performance of the model, we conduct several experiments on the well-known benchmark datasets.

The main contribution of this paper is summarized as follows.

- It is the first attempt to analyze the cause of the algorithmic bias from the bias-variance dilemma perspective to the best of our knowledge.
- We prove theoretically that the algorithmic bias can be solved by learning fair representation.
- With the proposed unbiased information bottleneck method with adversarial learning, we are able to train the fair representation that can be used for transfer learning.
- Achieving the highest performance in various datasets, the proposed algorithm is proved to be fair without data-dependency compared to other known techniques.

The rest of this paper is organized as follows. In Section 2, we introduce the research for solving the algorithmic bias in three aspects. The work we have done in this paper and the proposed model are presented in Section 3 and the evaluation is performed in Section 4. Section 5 presents the summary and some discussion.

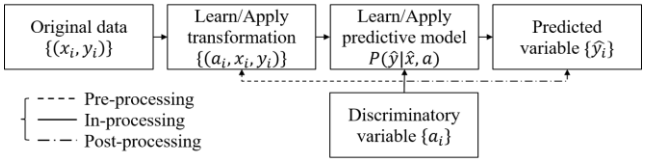

Figure 1. Schematic diagram of the approaches to solve the algorithmic bias.

Related Works

Several works have been conducted to recognize and solve the problem of algorithmic bias. The approach to solve it can be categorized into three types as mentioned in the previous section. We summarize the studies for each approach in Table 1.

The solution through pre-processing is to preemptively block the algorithmic bias by eliminating the biases present in the data. Hajian proposed a method to solve indirect bias as well as algorithmic bias with multiple metrics for computing discrimination [Hajian, 2013]. Edwards and Storkey presented a method to learn the fair representation to solve the algorithmic bias and applied it to the image anonymization to remove the text in the image [Edwards and Storkey, 2016]. Louizos et al. achieved a fair representation through variational autoencoder, one of the methods of learning data representation [Louizos et al., 2017]. Calmon et al. proposed a solution of the algorithmic bias by introducing three important criteria to be addressed in the preprocessing approach [Calmon et al., 2017]. Grgic-Hlaca et al. proposed a method to select features for procedurally fair learning [Grgic-Hlaca et al., 2018].

The above studies have the advantage of pre-blocking the bias, but they have the disadvantage of losing the characteristics of the data and lowering the performance. Therefore, research has been carried out to achieve fairness in the learning algorithm. Kamishima et al. proposed a method of solving algorithmic bias through a regularization approach based on analyzed unfairness [Kamishima et al., 2011]. Fish et al. achieved fairness through a technique of shifting decision boundaries [Fish et al., 2016]. Zafar et al. proposed a method to add a regularization term called disparate mistreatment to the learning algorithm [Zafar et al., 2017]. Zhang et al. used the prediction results to teach the model to be fair [Zhang et al., 2018].

Hardt et al. defined a metric for evaluating the degree of unfairness to use post-processing techniques to solve algorithmic bias in a post-hoc manner [Hardt et al., 2016]. Although most of the studies proposed a method to solve algo-

Category	Description	Authors	Pros & Cons
Pre-processing	Preprocessing training data	Hajian (2013), Edwards (2016), Louizos (2016), Calmon (2017), Grgic-Hlaca (2018)	• Prevent bias through sophisticated preprocessing • Possible to resolve bias present in data
			• Occur individual distortion that loses feature of the data
In-processing	Modifying learning algorithm with constraints	Kamishima (2011), Fish (2016), Zafar (2017), Zhang (2018)	• Use complete data • Applicable regardless of data
			• Instability due to constraints added to the learning algorithm • Impossible to resolve bias present in data
Post-processing	Post-hoc	Hardt (2016)	• Achievable of zero bias • Applicable to all data and models
			• User intervenes to analyze results and control decision boundary

Table 1. The summary of the related works.

rithmic bias, they lack the cause of the problem and the evaluation of the validity of the method. In this paper, we attempt to analyze the algorithmic bias in terms of bias-variance dilemma, and theoretically prove that we can solve this problem through fair representation learning. We also show that the algorithmic bias can be solved by constructing a model in adversarial learning.

Proposed Method

Overview

As shown in equation (2), bias-variance dilemma shows the conflict in trying to minimize the components of error at the same time; bias and variance [Geman et al., 1992; Vijayakumar, 2007]. The bias term evaluates how similar the approximate function is to the real function, and the variance term evaluates how complex the approximation is. We will demonstrate that the algorithmic bias occurs because of the difference in bias values of each protected feature. We will also prove theoretically that we can reduce the difference in bias for each sensitive feature by making the representation fair. If we construct an objective function that reduces such a difference as equation (3), it becomes intractable to learn from the data that is not separated. We modify this so as to prove that the fair representation reduces the difference between bias values of protected features. Fair representation is achieved through adversarial learning of the encoder and discriminator by using the unbiased information bottleneck, and detailed learning process will be followed.

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{\mathrm{x}\sim\mathcal{X}}\left[\left(y-\hat{f}(x)\right)^2\right] \\ &= \mathbb{E}_{\mathrm{x}\sim\mathcal{X}}\left[f(x)-\hat{f}(x)\right]^2 \\ &\quad +\mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}_{x\sim\mathcal{X}}\left[\hat{f}(x)\right]\right)^2\right], \\ &= \mathrm{Bias}(\hat{f})^2 + Variance(\hat{f})\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{\mathrm{x}\sim\mathcal{X}}\left[\left(y-\hat{f}(x)\right)^2\right] \\ &\quad +\left[\mathrm{Bias}^{+}(\hat{f})-Bias^{-}(\hat{f})\right]^2,\end{aligned} \tag{3}$$

where $\mathcal{X}$ is the space of data, x is a sample of data, y is real label corresponding to x, f is a real function, $\hat{f}$ is an approximated function, and Bias^{+} and Bias^{-} are expectations of the difference between approximated and real functions with only positive and negative protected features, respectively. Chen et al. also provided the approach to define the algorithmic bias from the viewpoint of bias-variance dilemma, but they just demonstrated the relationship between them [Chen et al., 2018]. We not only associate them but also derive the solution based on it.

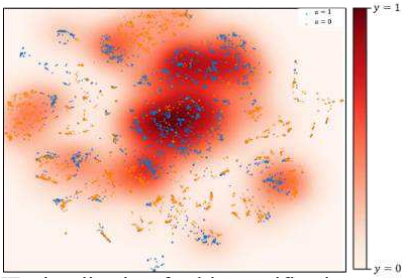

Figure 2. t-SNE visualization for bias verification present in data itself. Blue color corresponds to positive sensitive features whereas orange color corresponds to negative protected features. The more reddish the background, the denser the positive real label.

Analysis on Algorithmic Bias

It is demonstrated from the perspective of bias-variance dilemma that algorithmic bias can be solved through fair representation, unlike previous studies. Figure 2 shows the relationship between the bias for each protected feature and the algorithmic bias. The data representation is shown using the t-SNE algorithm [Maaten and Hinton, 2011]. We confirm that many points whose protected feature a is positive exist in a space where the actual label y is positive, so the trained model predicts that $y = 1$ for data with $a = 1$ (i.e., $\mathrm{Bias}^{+}(\hat{f})$ is bigger than $\mathrm{Bias}^{-}(\hat{f})$). We formulate an objective function as equation (4) because the algorithmic bias can be solved by learning the same bias value for each sensitive feature. However, since it is intractable, we derive equation (8) from equation (4).

$$\begin{aligned}&\mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}\left[\hat{f}(x)-f(x)\right] \\ &\quad\approx \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{-}}\left[\hat{f}(x)-f(x)\right]\end{aligned} \tag{4}$$

$$\begin{aligned}&\Leftrightarrow \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}\left[\hat{h}\circ\hat{g}(x)-h\circ g(x)\right] \\ &\quad\approx \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{-}}\left[\hat{h}\circ\hat{g}(x)-h\circ g(x)\right]\end{aligned} \tag{5}$$

$$\begin{aligned}&\Leftrightarrow \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}\left[\hat{h}\circ\hat{g}(x)\right]-\mathbb{E}_{x\sim\mathcal{X}^{-}}\left[\hat{h}\circ\hat{g}(x)\right] \\ &\quad\approx \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}[h\circ g(x)]-\mathbb{E}_{x\sim\mathcal{X}^{-}}[h\circ g(x)]\end{aligned} \tag{6}$$

$$\Leftrightarrow \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}\left[\hat{h}\circ\hat{g}(x)\right]\approx \mathbb{E}_{x\sim\mathcal{X}^{-}}\left[\hat{h}\circ\hat{g}(x)\right] \tag{7}$$

$$\Leftarrow \mathbb{E}_{\mathrm{x}\sim\mathcal{X}^{+}}[\hat{g}(x)]\approx \mathbb{E}_{x\sim\mathcal{X}^{-}}[\hat{g}(x)] \tag{8}$$

where g and h are real functions of feature extractor and classifier, respectively, functions with hat symbol represent approximated functions, and $\mathcal{X}^{+}$ and $\mathcal{X}^{-}$are the spaces of data with positive and negative sensitive features, respectively. Because functions of classifier are not always invertible, we represent the left arrow in equation (8).

As a result, if the feature extractor results (i.e., data representation) have similar distributions for each of the protected features, the difference in bias is reduced, resulting in mitigating algorithmic bias.

Fair Representation via Adversarial Learning of Information Bottleneck

There have been many issues of intractable formula when achieving fair representation through equation (1) [Chen et al., 2016; Alemi et al., 2017]. We can obtain the similar results by maximizing the lower bound to maximize $I(X,Z)$, which represents the association between data and

latent variables. Inspired by Chen et al., we derive the lower bound of $I(X,Z)$ as shown in equations (9) to (12).

$$I(Z,X) \tag{9}$$
$$= \mathbb{E}_{z\sim q(z|x)}\left[\mathbb{E}_{x'\sim p(x|z)}[log\ P(x'|z)]\right] + H(x) \tag{10}$$
$$\geq \mathbb{E}_{z\sim q(z|x)}\left[\mathbb{E}_{x'\sim p(x|z)}[log\ Q(x'|z)]\right] + H(x) \tag{11}$$
$$= \mathbb{E}_{x'\sim P(x'),z\sim q(z|x)}[log\ Q(x'|z)] + H(x) \tag{12}$$

where p and q represent the decoder and encoder, respectively, P and Q mean real and approximated distributions, and H is the entropy. From equation (10) to equation (11), we use the fact that Kullback-Leibler divergence is non-negative. Let the first term of equation (12) be $\mathcal{L}_{\mathrm{XZ}}$.

At the same time, we should minimize the $I(Z,A)$ to achieve the equation (1). Inspired by Alemi et al., we can minimize the $I(Z,A)$ by reducing the upper bound of it, and we derive it from equations (13) to (17).

$$I(Z,A) \tag{13}$$
$$= \int p(a|z)p(z)\ log\frac{p(z|a)}{p(z)} \tag{14}$$
$$= \int p(a|z)p(z)\ log\ p(a|z) - \int p(z)\ log\ p(z) \tag{15}$$
$$\leq \int p(a|z)p(z)\ log\ p(a|z) - \int p(z)\ log\ m(z) \tag{16}$$
$$= \int p(a|z)p(z)\ log\frac{p(z|a)}{m(z)} \tag{17}$$

where $m(z)$ is a variational approximation. However, one critical issue in this approach is the difficulty of choosing the proper approximator $m(z)$. We propose another formulation of the upper bound based on information bottleneck theory [Tishby et al., 2000; Tishby and Zaslavsky, 2015]. As shown in Figure 3, we define the additional model t, called latent transfer, that takes latent variable z from encoder and outputs an intermediate representation r. Consequently, a modified upper bound of $I(Z,A)$ can be obtained as:

$$I(Z,A) = I\left(Z, D\big(t(Z)\big)\right) \tag{18}$$

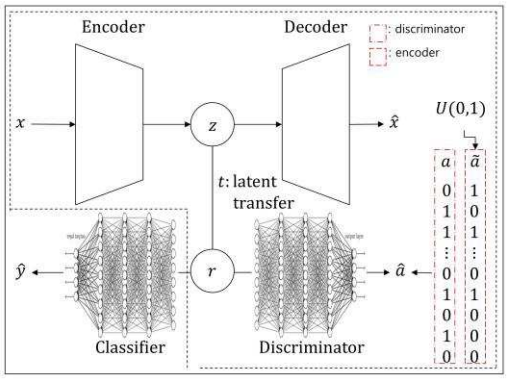

Figure 3. The architecture of the proposed method. To train a discriminator, sensitive feature a is sampled from $\mathcal{A}$. Encoder is trained to project data into fair representation with $\tilde{a}$ from $U(0,1)$. When the fair representation is obtained, only the models inside the dashed line are learned. When solving the classification problem fairly, the output value from the encoder becomes input to the classifier.

$$\leq I\big(Z, t(Z)\big) \tag{19}$$
$$\leq \int t(r|z)p(z)\ log\frac{t(r|z)}{s(r)} \tag{20}$$

where $s(r)$ is a variational approximation. The first inequality holds thanks to the Markov property [Tishby and Zaslavsky, 2015]. Let the equation (20) be $\mathcal{L}_{\mathrm{AZ}}$, and we can rewrite the equation (1) as follows:

$$\max \mathcal{L}_{IB} \leq \mathcal{L}_{XZ} - \mathcal{L}_{AZ} \tag{21}$$

Using this approach, to solve the algorithmic bias through fair representation learning as we have proved in the previous section, the encoder and discriminator are learned in adversarial. The architecture of the proposed method is illustrated in Figure 3. The objective function for the three models is as follows.

$$\mathcal{L}_{f_e,f_d} = \mathbb{E}_{x\sim\mathcal{X}}\left[\begin{matrix}-\mathcal{L}_{XZ} + \mathcal{L}_{AZ} \\ +\mathbb{E}_{\tilde{a}\sim U(0,1)}\left[l\left(\tilde{a}, D\big(f_e(x)\big)\right)\right]\end{matrix}\right] \tag{22}$$

$$\mathcal{L}_{\mathrm{D}} = \mathbb{E}_{(x,a)\sim\mathcal{X}\times\mathcal{A}}\left[l\left(a, D\big(f_e(x)\big)\right)\right], \tag{23}$$

where f_e, f_d, and D are functions of encoder, decoder, and discriminator, respectively, $\mathcal{A}$ is the space of sensitive features, l is a binary function of measuring the reconstruction or categorical loss, $\tilde{a} \sim U(0,1)$ means random sampling of zero or one, and $\mathcal{D}_{\mathrm{KL}}$ is a Kullback-Leibler divergence to make the result distribution as simple as we want.

Several works were proposed to achieve the fairness by modifying the learning process of generative adversarial autoencoder and adversarial autoencoder, similar to the method proposed in this paper [Edwards and Storkey, 2016; Mardras et al., 2018; Wadsworth et al., 2018; Zhang et al., 2018; Goodfellow et al., 2014; Makhzani et al., 2015]. They proposed a method of training encoder to make discriminator to classify the sensitive features as opposite with data representation, which can learn the information about the protected features. However, we first propose an unbiased information bottleneck method and train an encoder to make discriminator to randomly classify the protected features with $\tilde{a} \sim U(0,1)$ term so that the information about the sensitive attribute disappears on data representation. Discriminator classifies projected latent variables from data with protected features, so that the encoder learns to construct a fair representation independent of the sensitive attribute for fooling the discriminator. Besides, their approaches are in-processing, resulting in unavailable fair representation. Our optimization also converges since it is equivalent to the original objective function as shown in equation (24).

$$\min_{\mathrm{G}}\max_{\mathrm{D}}\left[\begin{matrix}\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}\left[\mathbb{E}_{a\sim U(0,1)}\left[\big(a - D\big(G(z)\big)\big)\right]\right]\end{matrix}\right]$$
$$\Leftrightarrow \min_{\mathrm{G}}\max_{\mathrm{D}}\left[\begin{matrix}\mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] \\ +\mathbb{E}_{z\sim p_z(z)}\left[\big(1 - D\big(G(z)\big)\big)\right]\end{matrix}\right] \tag{24}$$

The objective function converges when $p_d \approx p_g$, where p_d and p_g are distributions of real and generated data, respectively [Goodfellow et al., 2014]. We can map pairs of real and fake in the generative adversarial networks to pairs of positive and negative of sensitive features, so the convergence point is $\mathbb{E}_{x\sim X^+}(x) \approx \mathbb{E}_{x\sim X^-}(x)$. The algorithm for learning the fair representation $f_e(x)$ is as follows.

Algorithm 1 Training fair representation

Input: Data $\mathcal{X}$ and corresponding sensitive feature $\mathcal{A}$
Output: Fully trained encoder f_e
1: Initialize f_e, f_d, D
2: **for** epochs **do**
3: **for** batches **do**
4: Sample $x, a, \tilde{a}$ from $\mathcal{X}, \mathcal{A}$, U(0,1) respectively;
5: $\theta_{f_e} \leftarrow \theta_{f_e} - \eta \frac{\partial \mathcal{L}_{f_e,f_d}}{\partial \theta_{f_e}}(x, \tilde{a})$;
6: $\theta_{f_d} \leftarrow \theta_{f_d} - \eta \frac{\partial \mathcal{L}_{f_e,f_d}}{\partial \theta_{f_d}}(x)$;
7: $\theta_D \leftarrow \theta_D - \eta \frac{\partial \mathcal{L}_D}{\partial \theta_D}(x, a)$;
8: **end for**
9: **end for**
10: **return** f_e

where $\theta_{f_e}, \theta_{f_d}$, and θ_D are parameters of encoder, decoder, and discriminator, respectively, and η is a learning rate. As the generative adversarial network converges when $p_d \approx p_g$, our model can converge when $p_{X^+} \approx p_{X^-}$.

In addition, the fair representation, which is the result of the proposed method in this paper, is not subject to the model in classifying the actual label in representation because there is no term associated with the classifier in the learning process. The algorithm for classifying data representations into real labels with other classifiers C is as follows.

Algorithm 2 Classifying data with fair representation

Input: Data $\mathcal{X}$, corresponding label $\mathcal{Y}$, and classifier C
Output: Fully trained classifier C
1: Initialize C
2: **for** epochs **do**
3: **for** batches **do**
4: Sample x, y from $\mathcal{X}, \mathcal{Y}$ respectively;
5: $\theta_C \leftarrow \theta_C - \eta \frac{\partial l}{\partial \theta_C}\left(y, C\left(f_e(x)\right)\right)$;
8: **end for**
9: **end for**
10: **return** C

where l is a binary function measuring the difference between real label y and calculated label $\hat{y} = C\left(f_e(x)\right)$. If we use a machine learning classifier instead of deep learning, we use an algorithm that learns the classifier instead of the parameter update method on line 5.

Experiments

Dataset and Experimental Settings

We use the well-known benchmark datasets such as Adults, Census, and COMPAS to evaluate the performance of the proposed method compared to other known techniques.[1] The details of the datasets are described in Table 2. We show the size of the final features by one-hot encoding of the categorical attributes in each data in the third column. In Census dataset, class-imbalance is solved by under-sampling technique. In the COMPAS dataset, "Caucasian" and "African-American" are used as the protected features. If the decile score is greater than 5, we set the actual label to a positive value; otherwise, we set it to a negative value. The layers used for encoders, decoders, and discriminators are fully-connected and use rectified linear unit as an activation function [Nair and Hinton, 2010]. We set the binary function l as mean squared error when used in reconstruction and as cross-entropy when used in classification. We compare the Base model, which is the same to our method without adversarial learning, the variational fair autoencoder (VFAE) proposed by Louizos, and the adversarial model proposed by Edwards to demonstrate the superiority of the proposed method [Edwards and Storkey, 2016; Louizos et al., 2017].

Data	# of data	# of features	Protected features	Real label
Adults	32,561	102	Gender	Income
Census	37,136	389	Gender	Income
COMPAS	11,742	16	Race	Risky score

Table 2. The details of datasets used in this paper

Bias Reduction

To quantitatively evaluate the algorithmic bias, we use the well-known metrics such as the equality of opportunity and the equality of odds as shown in equations (25) and (26) [19]. The Opportunity measures devaluation as the difference in True Positive Rate (TPR) for each protected feature. The Odds measures over-evaluation as the difference in TPR and False Negative Rate (FNR) for each protected feature.

$$\text{Opportunity} = \left|\frac{\Sigma_{x\in X^+}\mathbb{I}_{f(y|x)=1}}{\Sigma_{x\in X^+}\mathbb{I}_{y=1}} - \frac{\Sigma_{x\in X^-}\mathbb{I}_{f(y|x)=1}}{\Sigma_{x\in X^-}\mathbb{I}_{y=1}}\right| \quad (25)$$

$$\text{Odds} = \left|\frac{\Sigma_{x\in X^+} f(y|x)}{\Sigma_{x\in X^+}\mathbb{I}} - \frac{\Sigma_{x\in X^-} f(y|x)}{\Sigma_{x\in X^-}\mathbb{I}}\right| \quad (26)$$

[1] Adult & Census datasets: https://archive.ics.uci.edu/ml/datasets/
COMPAS dataset: https://kaggle.com/danofer/compass

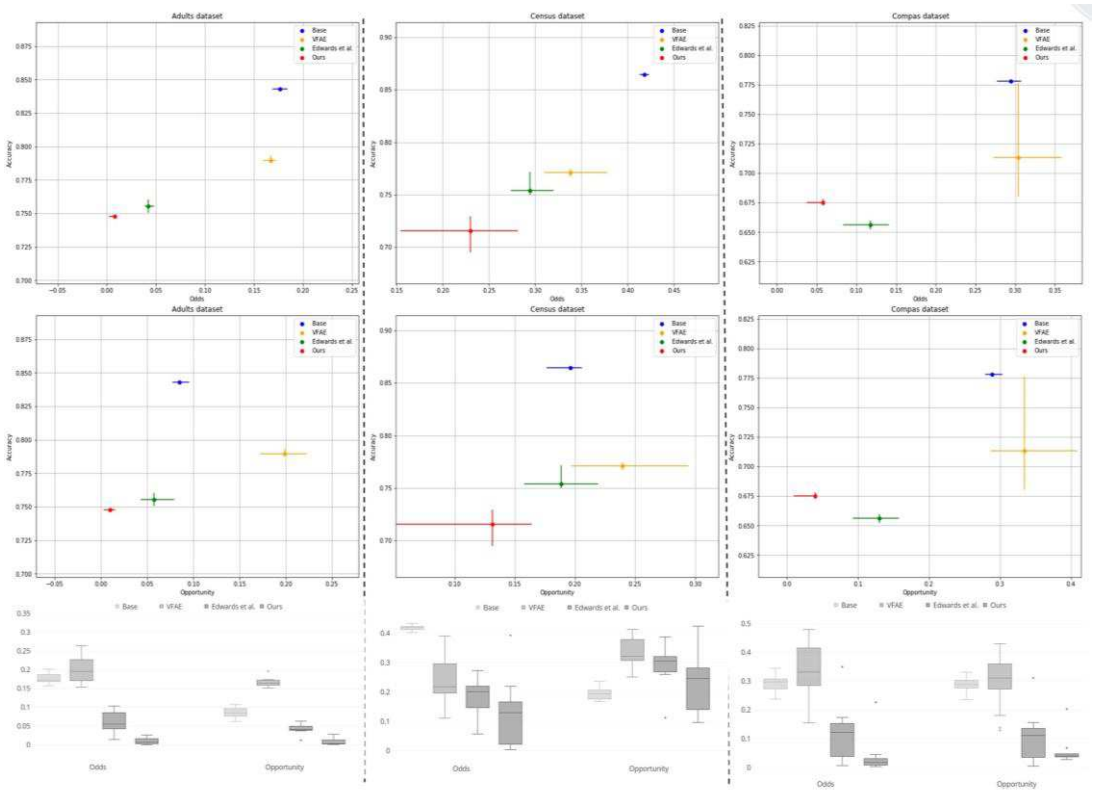

Figure 4. The experimental results of base, variational fair autoencoder, model proposed by Edwards et al. and the proposed model with respect to three datasets. We compare them with two metrics; equality of opportunity and odds, and our model has the best fairness compared to others. Look at the figures after zoon-in.

We conduct the experiment with 10-fold cross validation, and the result of evaluation with each metric is shown in Figure 4. The proposed model achieves better fairness on all datasets relative to the different models. The numerical results are shown in Tables 3 and 4. The performance of the proposed model was statistically significant at the 95% significance level except for four cases, and all the results were significant at the 90% significance level. The trade-off between discrimination and utility is also likely to occur in the proposed model and this problem will be addressed for future research. In this paper, we analyze that the algorithmic bias occurs due to the difference of the

Models	Base	VFAE	Edwards et al.	Ours
Adult dataset				
Average	0.1765	0.1669	0.0421	**0.0077**
Std. dev.	0.0128	0.0119	0.0129	0.0091
p-value	1.11×10^{-17}	1.43×10^{-17}	1.87×10^{-6}	-
Census dataset				
Average	0.4203	0.3357	0.2944	**0.2297**
Std. dev.	0.0090	0.0497	0.0720	0.0955
p-value	1.06×10^{-4}	5.20×10^{-3}	6.09×10^{-2}	-
COMPAS dataset				
Average	0.2942	0.3153	0.1076	**0.0577**
Std. dev.	0.0277	0.0287	0.0844	0.0497
p-value	2.88×10^{-9}	3.89×10^{-11}	7.19×10^{-2}	-

Table 3. The numerical results of experiments. We verify the performance with Odds metric and conduct t-test.

Models	Base	VFAE	Edwards et al.	Ours
Adult dataset				
Average	0.0851	0.1987	0.0576	**0.0098**
Std. dev.	0.0132	0.0328	0.0274	0.0084
p-value	1.20×10^{-11}	6.12×10^{-9}	2.02×10^{-4}	-
Census dataset				
Average	0.1961	0.2448	0.1884	**0.1314**
Std. dev.	0.0190	0.0770	0.0600	0.1114
p-value	5.83×10^{-2}	1.09×10^{-2}	9.91×10^{-2}	-
COMPAS dataset				
Average	0.2891	0.3038	0.1184	**0.0400**
Std. dev.	0.0239	0.0250	0.0959	0.0636
p-value	1.40×10^{-7}	3.55×10^{-9}	2.01×10^{-2}	-

Table 4. The numerical results of experiments. We verify the performance with Opportunity metric and conduct t-test.

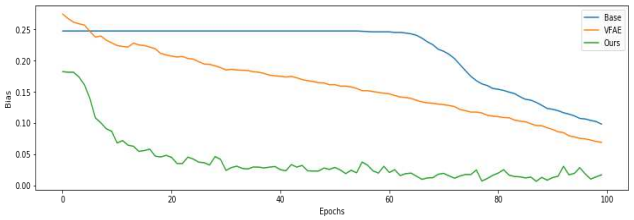

Figure 5. Evaluation of model performance by bias reduction result. Y-axis represents a difference between bias values for each protected feature.

biases (i.e., $|Bias^{+} - Bias^{-}|$) by the protected feature in the bias-variance. Therefore, we show in Figure 5 that the proposed method reduces the difference of bias by protected feature, which means indirectly lowering the intractable loss in equation (2) with our learning algorithm.

Fairness Latent Variable

The results of the proposed method for solving algorithmic bias through fair representation are depicted in Figure 6 using the t-SNE algorithm. Note that the model proposed by Edwards et al. belongs to in-processing approach so that we could not use fair representation using it. The distribution of raw data exists in an unknown pattern and is separated by protected features, whereas the distribution of data representation becomes simple using VFAE. However, in data representation via VFAE, data can be distinguished according to a protected feature. We can observe that the data representation independent of the protected feature is trained by using our method.

Since the fair representation learned through the proposed model has non-discrimination characteristics in itself, the classification result is not biased regardless of the classifier. We classify the real labels by inputting the learned fair representation to various machine learning algorithms such as logistic regression (LR), decision tree (DT), and random forest (RF), and the results are shown in Figure 7. Surprisingly, even if we bring the pre-trained fair representation by the proposed method, results for all the machine learning algorithms used in the experiment has the best fairness in the proposed method. In addition, except for DT, accuracy is reduced by only about 2%. Comparisons and analysis using more classifiers will be proceeded further in the future.

Conclusions

In this paper, we address the need for solving algorithmic bias problem. Fairness is achieved to some degree by adversarial learning of encoder and discriminator for fair representation through the proposed method. We analyze the cause of the algorithmic bias from the perspective of bias-variance dilemma and prove that it can be solved through fair representation. The proposed method learns the representation independent of the classifier including deep learning, because the bias has the lowest value when the output of the encoder is input to classifiers.

We demonstrate that the cause of algorithmic bias is due to the difference of biases for each sensitive feature in bias-variance, and we apply adversarial learning as a way to reduce the difference. However, since this process is dual optimization, it leads to unstable learning. In the future, we will try to find a way to reduce bias differences more stably. We also need to evaluate the scalability of the proposed method for fair representation in other domains such

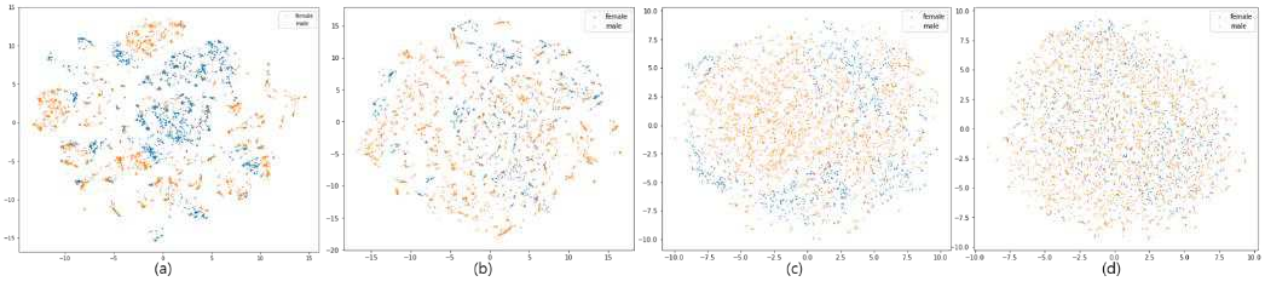

Figure 6. t-SNE visualization from the Adults dataset on: (a) data itself, (b) representation with base model, (c) representation with VFAE model, and (d) the proposed model. It can be seen that the representation of method except ours can be easily separated with protected features.

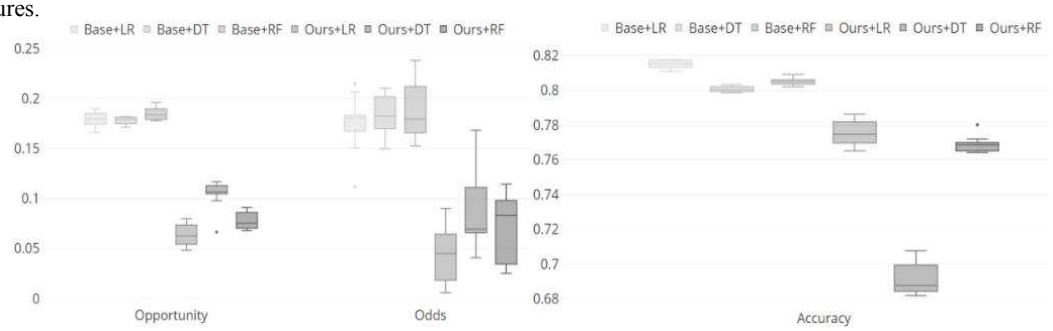

Figure 7. Performance of learned fair representation in the proposed method.

as image and text.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [2016-0-00562(R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly]. J. Y. Kim has been supported by NRF (National Research Foundation of Korea) Grant funded by the Korean Government (NRF-2019-Fostering Core Leaders of the Future Basic Science Program/Global Ph.D. Fellowship Program).

References

Alemi A. A.; Fischer I.; Dillon V. J.; and Murphy K. 2017. Deep Variational Information Bottleneck. *Int. Conf. on Learning Representation*: 1-19.

Calmon P. F.; Wei D.; Vinzamuri B.; Ramamurthy N. K.; and Varshney R. K. 2017. Optimized Pre-Processing for Discrimination Prevention. *In Advances in Neural Information Processing Systems*: 3992-4001.

Chen I.; Johansson F. E.; and Sontag D. 2018. Why is My Classifier Discriminatory? *In Advances in Neural Information Processing Systems*: 3539-3550.

Chen X.; Duan Y.; Houthooft R.; Schulman, J.; Sutskever, B.; and Abbeel, P. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *In Advances in Neural Information Processing Systems*: 2172-2180.

Dressel J., and Farid H. 2018. The Accuracy, Fairness and Limits of Predicting Recidivism. *Science Advances* 4(1): 1-6.

Dwork C.; Hardt M.; Pitassi T.; Reingold O.; and Zemel R. 2012. Fairness Through Awareness. *In Proc. of the 3rd Innovation in Theoretical Computer Science Conf.*:214-226

Edwards H., and Storkey A. 2016. Censoring Representation with an Adversary. *Int. Conf. on Learning Representation*: 1-14.

Feldman M.; Friedler A. S.; Moeller J.; Scheidegger C.; and Venkatasubramanian S. 2015. Certifying and Removing Disparate Impact. *In Proc. of the 21st ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*: 259-268.

Fish B.; Kun J.; and Lelkes D. A. 2016. A Confidence-based Approach for Balancing Fairness and Accuracy. *In Proc. of Int. Conf. on Data Mining*: 144-152.

Geman S.; Bienenstock E.; and Doursat R. 1992. Neural Networks and the Bias/Vaiance Dilemma. *Neural Computation* 4(1): 1-58.

Goodfellow I.; Pouget-Abadie J.; Mirza M.; Xu B.; Warde-Farley D.; Ozair S.; Courville A.; and Bengio Y. 2014. Generative Adversarial Nets. *In Advances in Neural Information Processing Systems*: 2672-2680.

Grgic-Hlaca N.; Zafar B. M.; Gummadi P. K.; and Weller A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *In AAAI Conf. on Artificial Intelligence*: 51-60.

Hajian S. 2013. Simultaneous Discrimination Prevention and Privacy Protection in Data Publishing and Mining. *arXiv preprint* arXiv: 1306.6805.

Hardt M.; Price E.; and Srebro N. 2016. Equality of Opportunity in Supervised Learning. *In Advances in Neural Information Systems*: 3315-3323.

Helm T. 2016. David Cameron Calls on David Lammy to Investigate Race Bias in UK Courts. *The Guardian*. Retrieved: 23.08.2019.

Kamiran F., and Calders T. 2010. Classification with no Discrimination by Preferential Sampling. *In Proc. 19th Machine Learning Conf Belgium and the Netherlands*.: 1-6.

Kamishima T.; Akaho S.; and Sakuma J. 2011. Fairness-Aware Learning through Regularization Approach. *IEEE 11th Int. Conf. on Data Mining Workshops*: 643-650.

Koene A. 2017. Algorithmic Bias: Addressing Growing Concerns. *IEEE Technology and Society Magazine* 26(2): 31-32.

Louizos C.; Swersky K.; Li Y.; Welling M.; and Zemel R. 2017. The Variational Fair Autoencoder. *Int. Conf. on Learning Representation*: 1-11.

Luong T. B.; Ruggieri S.; and Turini F. 2011. k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. *In Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*: 502-510.

Maaten L., and Hinton G. 2011. Visualizing Data Using t-SNE. *Journal of Machine Learning Research* 9(11): 2579-2605.

Madras D.; Creager E.; Pitassi T.; and Zemel R. 2018. Learning Adversarially Fair and Transferable Representations. *arXiv preprint* arXiv: 1802.06309.

Makhzani A.; Shlens J.; Jaitly N.; and Goodfellow I. 2015. Adversarial Autoencoders. *Int. Conf. on Learning Representations*: 1-16.

Nair V., and Hinton G. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proc. of the 27th Int. Conf. on Machine Learning*: 807-814.

Shipman T., and Griffiths S. 2016. A Young Black Man is More Likely to be in Prison than at a top University. *The Times*. Retrieved: 21.08.2019.

Tishby N.; Pereira, C. F.; and Bialek W. 2000. The Information Bottleneck Method. *arXiv preprint* arXiv:0004057.

Tishby N., and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle. *IEEE Information Theory Workshop*: 1-5.

Vijayakumar S. 2007. The Bias-Variance Tradeoff. *University Edinburgh Lecture notes*: 1-2.

Wadsworth C.; Vera F.; and Piech C. 2018. Achieving Fairness through Adversarial Learning: An Application to Recidivism Prediction. *arXiv preprint* arXiv: 1807. 00199.

Zafar B. M.; Valera I.; Rodriguez G. M.; and Gummadi P. K. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *In Proc. of the 26th Conf. on World Wide Web*: pp. 1171-1180.

Zemel R.; Wu Y.; Swersky K.; Piitassi T.; and Dwork D. 2013. Learning Fair Representations. *Int. Conf. on Machine Learning*: 325-333.

Zhang H. B.; Lemoine B.; and Mitchell M. 2018. Mitigating Unwanted Biases with Adversarial Learning. *In AAAI Conf. on Artificial Intelligence*: 335-340.