ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240570

Making Fair Classification via Correlation Alignment

Jingran Yang[a], Lingfeng Zhang[b] and Min Zhang[c,*]

[a,b,c]East China Normal University
ORCID (Jingran Yang): https://orcid.org/0009-0008-6406-9222, ORCID (Lingfeng Zhang): https://orcid.org/0000-0002-6427-9587, ORCID (Min Zhang): https://orcid.org/0000-0002-3152-4347

Abstract. Machine learning learns patterns from data to improve the performance of the decision-making systems through computing, and gradually affects people's lives. However, it shows that in current research machine learning algorithms may reinforce human discrimination, and exacerbate negative impacts on unprivileged groups. To mitigate potential unfairness in machine learning classifiers, we propose a fair classification approach by quantifying the difference in the prediction distribution with the idea of correlation alignment in transfer learning, which improves fairness efficiently by minimizing the second-order statistical distance of the prediction distribution. We evaluate the validity of our approach on four real-world datasets. It demonstrates that our approach significantly mitigates bias w.r.t demographic parity, equality of opportunity, and equalized odds across different groups in a classification setting, and achieves better trade-off between accuracy and fairness than previous work. In addition, our approach can further improve fairness and mitigate the fair conflict problem in debiased networks.

1 Introduction

Machine learning algorithms are increasingly integrated into daily lives, and sometimes they are even applied to high-stakes decision-making. However, existing research suggests that machine learning may replicate and exacerbate human discrimination and bias in certain scenarios, such as financial credit [30][4], employment services [49][17][31][26], medical diagnosis [35][22][43], recommendation systems [18][53][39], college standardized test [10][11] and so on, with potentially negative effects on groups or individuals in society. The promotion of machine learning applications depends on the improvement of people's trust in them, so ethically aligned machine learning is an inevitable development direction [32][48].

In this context, algorithmic ethics, especially machine learning fairness, has attracted widespread attention [42] [38] [2]. For example, the Draft Ethics Guidelines for Trustworthy AI issued by the European Union emphasizes that trustworthy AI should meet the conditions of transparency and fairness [42]. Machine learning fairness refers to the neutrality of a system in decision-making and resource allocation, ensuring no bias based on inherent or acquired characteristics [40]. Such characteristics are called sensitive attributes, which are features related to people, typically including race, gender, age, etc. Depending on the value of the sensitive attribute, samples can be divided into different groups (e.g., male and female groups), which may be treated differently by machine learning algorithms. For example, a hiring model may make unfair decisions about candidates of a certain gender or race [36].

The existing machine learning fairness is divided into individual and group fairness. Dwork et al. [21] first proposed the definition of individual fairness, that is, intuitively, the system should give similar prediction results to similar individuals. As research continues, people gradually pay attention to group fairness, whose purpose is to ensure that two or more groups are treated similarly [45], that is, the predictions of groups with different sensitive attributes have similar probability distributions in terms of classification performance metrics. It coincides with the goal of transfer learning, which aims to improve distribution similarity. Therefore, based on this commonality, we tend to use the strategy of transfer learning to punish the distribution difference across groups to achieve group fairness.

While existing fairness research has made some progress in pre-processing [25][23][41][9], in-processing [7][33][52] and post-processing [29][24] stages, there are still some challenges. First, different metrics of group fairness may contradict each other [6][20]. They may emphasize different fair principles or values, so pursuing one fairness metric may adversely affect others. Next, the relationship between fairness and predictive performance needs to be carefully balanced when implementing group fairness. Too much emphasis on fairness may compromise the accuracy of the model [27].

In this paper, inspired by the strategy of transfer learning and the intuition that the predicted probabilities of different groups should be similar, we propose an in-processing debiasing framework via correlation alignment (CAF). Instead of modifying the data and network, our method imposes independence constraints directly on the model outputs to eliminate machine learning bias. This is achieved by enforcing a smaller correlation alignment distance called CORAL [44] between the model output distributions corresponding to groups with different sensitive attributes. Our main contributions include:

- We propose an in-processing debiasing method CAF which achieves significant fairness optimization through a simple and general algorithm. It achieves good experimental results on multiple real-world datasets and fairness metrics without conflict. [1]
- Our method ensures a better trade-off between fairness and accuracy than previous work while improving group fairness.
- Our method can be combined with other debiasing methods to further optimize fairness metrics that have been improved, and even solve the problem of fairness conflict brought by these methods.

* Corresponding Author. Email: mzhang@sei.ecnu.edu.cn.

[1] https://github.com/jryang100/CAF

2 Related work

In this section we will introduce some work related to this paper.

Bias mitigation. In order to achieve group fairness, researchers have proposed a series of metrics and debiasing methods to ensure the model has fair predictions across different groups. According to different stages of training, these methods can be divided into pre-processing, in-processing and post-processing methods.

Pre-processing methods try to transform data before training so that the underlying discrimination is removed [37]. An intuitive method is to delete sensitive attribute information [25]. Feldman et al. [23] introduced a feature-adjusting disparate impact remover to equalize marginal distributions across different sensitive attributes. Instead of modifying data features, Burnaev et al. [9] improved fairness by balancing the data distribution.

In-processing methods consider fairness during training, which eliminate discrimination by modifying objective functions or imposing constraints [34]. Beutel et al. [7] adopted a method called absolute correlation, which minimize the correlation between subgroup identity and negative example prediction. Madras et al. [33] took representation learning as the key to mitigate the downstream unfair prediction results. Zhao et al. [52] proposed the CFair (Conditional Learning of Fair Representations) algorithm based on balanced error rate and conditional alignment of representations.

Post-processing methods improve fairness by modifying the prediction after training [13]. Kamiran and Calders [29] adjusted the leaf labels of the decision tree to obtain an unbiased classifier. Fish et al. [24] achieved post-processing fairness by changing the decision boundaries of protected groups.

Transfer learning. Machine learning usually assumes similar distributions for training and test data [1]. Violating this assumption may degrade the performance on new domains, and even require rebuilding models from scratch with newly collected data [47]. However, it is expensive or infeasible in many applications. In this case, transfer learning is proposed to solve this problem by transferring knowledge from different but related source domain [54].

Bousmalis et al. [8] built a reconstruction network to learn domain-invariant representations. Ajakan et al. [3] obtained domain-insensitive features by minimizing label prediction error and maximizing domain classification error. However, these methods may require additional network components and complex training processes. In practice, researchers propose to implement transfer learning by minimizing divergence [44][14][15][16]. For example, Das and Lee [14] introduced hyper-graph smoothness and hyper-graph sparsity constraints for better performance on the target domain. Sun et al. [44] proposed a general and simple method called CORAL (correlation alignment), which measures the distribution difference by calculating second-order statistics between source and target domains, and achieves domain alignment by minimizing this distance.

Transfer learning improves data consistency, which helps to narrow the prediction difference across groups in machine learning fairness. Among them, CORAL can directly quantify and explain discrimination, and can be used as a regularization term to punish the distribution differences without changing the network structure. In addition, its simplicity, generality and ease of implementation allow us to achieve significant fairness improvement in a short time.

3 Preliminary

In this section, we will introduce the notations, fair definitions, and problem settings explored in this paper.

3.1 Notation

We consider the general learning task whose goal is to build a mapping from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$ based on a training set $\{(x_i, y_i)\}_{i=1}^N$. Among the features x of any sample, the partial attribute $a \in \mathcal{A}$ is denoted as the sensitive attribute, and $s \in \mathcal{S}$ refers to the part that does not contain the sensitive attribute, that is, $\mathcal{S} \cup \mathcal{A} = \mathcal{X}$ and $\mathcal{S} \cap \mathcal{A} = \emptyset$. On this basis, we use X, S, A and Y to represent the random variables of x, s, a and y respectively. To simplify the representation, we assume that $A, Y \in \{0, 1\}$.

In this paper, we define a binary classification model as $f : \mathcal{X} \mapsto \mathcal{P}$, where $\mathcal{P}$ is the probability space of the prediction. That is, f maps a input x to a two-dimensional vector $p = [p_0, p_1]$, where p_0 and p_1 denote the probability that x belongs to the negative and positive classes respectively. Therefore, the prediction label of variable X is

$$\hat{Y} = \arg\max f(X). \tag{1}$$

In order to optimize the training process, we need to calculate the gap between the model output and the ground truth value, and optimize through back-propagation. We can define the prediction loss like

$$\mathcal{L}_{pred} = CE(f(X), Y), \tag{2}$$

where CE represents the cross-entropy loss function [19].

3.2 Problem setup

Research on fairness in machine learning usually evaluates models from two aspects: utility and fairness. For the utility metric, the closer the accuracy of the model to 1, the better the model performs. Moreover, the closer the fairness metric is to 0, the fairer the model is. In this paper, we mainly study the problem of group fairness, where group members are determined by sensitive attribute A. We next briefly review the definitions most relevant to this work.

Definition 1. *Given true positives TP, true negatives TN, false positives FP and false negatives FN, the utility metric is defined as* $Acc = (TP + TN)/(TP + FP + FN + TN)$.

Definition 2. *If the prediction result and sensitive attribute of a model are statistically independent, it satisfies demographic parity, which is defined as* $P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)$.

We use the absolute value of the difference across the prediction probabilities of the two groups as the criterion for evaluating the fairness of the model, denoted as

$$\Delta DP = |P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0)|. \tag{3}$$

Definition 3. *If both false positive rate and true positive rate of predictions of the unprivileged group and the privileged group are the same, we can say that the model satisfies equalized odds:* $P(\hat{Y} = 1|A = 1, Y) = P(\hat{Y} = 1|A = 0, Y), Y \in \{0, 1\}$.

The fairness evaluation metric of equalized odds is denoted as

$$\begin{aligned} &\Delta Eodds \\ &= \frac{1}{2}(|P(\hat{Y} = 1|A = 0, Y = 0) - P(\hat{Y} = 1|A = 1, Y = 0)| \\ &+ |P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1)|). \end{aligned} \tag{4}$$

Definition 4. *As a relaxation of equalized odds, equality of opportunity can be defined as* $P(\hat{Y} = 1|A = 0, Y = 0) = P(\hat{Y} = 1|A = 1, Y = 0)$.

Based on the work done by Beutel et al. [7], we give the evaluation metric of equality of opportunity as follows

$$\begin{aligned}&\Delta FPR\\&= |P(\hat{Y}=1|A=0,Y=0) - P(\hat{Y}=1|A=1,Y=0)|.\end{aligned} \tag{5}$$

4 Correlation alignment for improving group fairness

The purpose of group fairness is to ensure that groups with different sensitive attributes are treated similarly, that is, the model produces the same predictive distribution for them. This goal coincides with the task of transfer learning, which is to reduce the distribution difference between the source and target domains to improve the similarity. Therefore, in order to measure and reduce bias in the model, we propose a CAF (Framework via Correlation Alignment) algorithm. Based on the real training data, this algorithm calculates the prediction distribution difference across different groups in the training process, adds the difference as fairness loss into the objective function, and updates the parameters by gradient descent algorithm to improve the group fairness of the model.

4.1 Calculate distribution difference from a statistical perspective

CORAL [44] is proposed to reduce the distribution difference between source domain and target domain by aligning their feature distributions, so as to achieve better domain adaption effect. According to the inspiration of CORAL, we align the feature distributions by calculating the statistical properties of the domain features, including first-order statistics (mean value) and second-order statistics (covariance matrix), in order to make the feature distributions as close as possible after the transformation. Because only simple statistical calculation and linear transformation are needed, the difference calculation method is simple and efficient.

To explain how to calculate the distribution difference, we use the following example to illustrate. Assume there are two datasets D_S and D_T we will align, where $D_S = \{v_i\}$, $v_i \in \mathbb{R}^d$ of the source domain, and the dataset $D_T = \{u_j\}$, $u_j \in \mathbb{R}^d$ of the target domain. The calculation process is shown in Algorithm 1. For each data distribution, we calculate the first-order statistical feature μ and second-order statistical feature C, and measure the difference of the data distribution by calculating the square of the Frobenius norm of the second-order statistical features, so as to facilitate the subsequent distribution alignment.

We can perform a feature alignment process by punishing distribution difference dis_{ST} calculated by Algorithm 1. This process minimizes the difference between the statistical properties of the source domain and the target domain, thereby reducing domain adaptation and helping to align the distributions of the two domains.

4.2 Mitigate bias by distribution alignment

One view of group fairness metrics is that the output distribution should be consistent across groups [50]. Beutel et al. [7] establish this idea by minimizing the absolute correlation of the prediction with group membership. We take a similar idea, using the distribution alignment goal of transfer learning to achieve a better match by minimizing the difference in predictions between different groups. Specifically, the CAF algorithm trains the model from two aspects:

Algorithm 1: Algorithm CORAL to calculate distribution difference

Input: Source data $D_S = \{v_1, v_2, ..., v_{N_S}\}$, Target data $D_T = \{u_1, u_2, ..., u_{N_T}\}$
Output: Distribution difference dis_{ST}
`/* Calculate mean values of data */`
$\mu_S = \frac{1}{N_S}\sum_{i=1}^{N_S} v_i$
$\mu_T = \frac{1}{N_T}\sum_{i=1}^{N_T} u_i$
`/* Align first-order statistics by zero-centering source and target data */`
$D'_S = D_S - \mu_S$
$D'_T = D_T - \mu_T$
`/* Covariance matrices of the centered source and target data */`
$C_S = \frac{1}{N_S - 1} D'^{\top}_S D'_S$
$C_T = \frac{1}{N_T - 1} D'^{\top}_T D'_T$
`/* Calculate distribution difference */`
$dis_{ST} = \|C_S - C_T\|_F^2$
return dis_{ST}

The first aspect is to complete the classification task from the point of view of utility by constantly narrowing the gap between the prediction and the ground truth label. Specifically, we input the samples into the deep neural network for forward propagation, obtaining prediction from the output layer of the network, calculating the gap between the prediction and the true label by cross-entropy function, and finally getting the label prediction loss $\mathcal{L}_{pred}$ as Equation (2).

On the other hand, from the perspective of fairness, we expect that similar groups will not produce different prediction due to different sensitive attributes, that is, they will not receive unfair treatment due to sensitive attributes, so we achieve group fairness by constantly reducing the gap in prediction results of groups with different sensitive attributes. Specifically, we input the samples X_a and $X_{a'}$ corresponding to the groups a and a' into the deep neural network to obtain their respective prediction probabilities p and p'. Then, we use the strategy of CORAL to calculate the difference between the prediction probability p and p', take the difference value as a fairness loss, and punish it in the training process:

$$\mathcal{L}_{fair} = CORAL(f(X_a), f(X_{a'})). \tag{6}$$

Based on regularization technique, we use the fairness regularization parameter λ to trade-off the above two losses, and finally get the objective function of the training model as

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{pred} + \lambda \mathcal{L}_{fair}\\ &= CE(f(X), Y) + \lambda CORAL(f(X_a), f(X_{a'})).\end{aligned} \tag{7}$$

Then, by using Adadelta gradient descent method [51], we reduce label prediction loss and fairness loss respectively, and continuously optimize parameters until convergence. The parameter optimization process is as follows

$$\theta = \theta - lr\frac{\partial(\mathcal{L}_{pred} + \lambda \mathcal{L}_{fair})}{\partial \theta}, \tag{8}$$

where lr is learning rate in training. The CAF algorithm is shown in Algorithm 2. We quantify the difference of the model's output distribution on different groups through Algorithm 1 and introduce it as a regularization term into the objective function. By minimizing

Algorithm 2: Algorithm CAF to improve group fairness

Input: Train dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, Maximum training epoch max_epoch, Size of each batch $batch\_size$, Learning rate lr, Fairness regularization parameter λ

Initialize the model parameters θ

Divide the training set D into $M = \lceil N/batch\_size \rceil$ batches, where $D^{(i)}$ consists of $X^{(i)}$ and $Y^{(i)}$

for $e \leftarrow 1$ **to** max_epoch **do**

  $\mathcal{L}_{pred} = 0$

  $\mathcal{L}_{fair} = 0$

  for $i \leftarrow 1$ **to** M **do**

    /* Calculate the class label prediction loss of the model */

    $\mathcal{L}_{pred} += CE(f(X^{(i)}), Y^{(i)})$

    /* Calculate the difference in predictions across different groups as the fairness loss term */

    $\mathcal{L}_{fair} += CORAL(f(X_a^{(i)}), f(X_{a'}^{(i)}))$

    /* Integral training loss */

    $\mathcal{L} = \mathcal{L}_{pred} + \lambda \mathcal{L}_{fair}$

    /* Update parameters according to the above objective function */

    $\theta_e = \theta_{e-1} - lr \nabla_\theta \mathcal{L}$

  end

end

this difference, we aim to reduce the model's dependence on sensitive attributes to achieve fairer predictions.

In Algorithm 2, $CORAL$ is the algorithm used for calculating prediction distribution difference across groups with different sensitive attributes, whose details can be found in Algorithm 1.

5 Experiment

In this section, we conduct experiments to evaluate the effectiveness of our CAF algorithm. The experiment focuses on three primary questions:

- How does the CAF algorithm affect the utility and fairness metrics of the model, as well as their trade-off?
- For datasets with multiple sensitive attributes, can CAF simultaneously optimize their fairness metrics?
- For debiased networks, can CAF further improve the fairness of them?

5.1 Benchmark datasets

We perform experiments on four popular real-world datasets in the literature of machine learning fairness, including German Credit dataset [28], Adult dataset [5], MEPS (Medical Expenditure Panel Survey) dataset [12] and Law School dataset [46]. The details of these datasets are described below.

German Credit: This dataset is used to predict whether an individual is likely to repay a loan. The German Credit dataset contains data on 1,000 individuals who applied for a loan and includes 20 features such as loan purpose, loan duration, age, job, residence, and a label 'Default' that judges the likelihood of an individual repaying the loan. The label 'Default' and sensitive attribute 'age' ($\leq$30 years old and $>$30 years old) are binary.

Adult: This dataset is used to predict whether an individual's annual income exceeded \$50k and contains 48,842 census records. It includes 14 features such as marital status, education, sex, occupation, where 'sex' is the sensitive attribute, and a label 'income' to determine whether an individual's annual income is more than \$50k. Race is also in the Adult dataset, which is generally used when studying the effects of multiple sensitive attributes [23].

MEPS: This dataset is used for utilization of health services and contains 15,730 samples. The MEPS dataset contains 138 features, including region, marital status, pregnancy status, race, ARTHTYPE (type of arthritis), ASTHDX (asthma diagnosis), and a label called utilization that determined whether an individual would use health care services frequently during a given time period. In MEPS dataset, 'race' is sensitive attribute, and 'sex' will be used when studying the fairness on multiple sensitive attributes.

Law School: This dataset is used to predict whether an individual will pass the bar exam to become a legal practitioner in the United States. It contains 18,692 data records and consists of 11 features of applicants, including LSAT scores, sex, family income, the level of law school they applied to, and a label 'pass_bar' that indicates whether the student ultimately passed the exam. The feature 'male' is the sensitive attribute, and in the context of studying multiple sensitive attributes, race would be taken into account.

The fairness metrics adopted in this paper are defined for tasks with binary sensitive attributes. In the settings of multi-value sensitive attributes, we may transform them into binary attributes through binarization. In future, we will explore more fairness definitions on multi-value sensitive attributes and extend the method in this paper to tasks with such attributes.

5.2 Experimental models

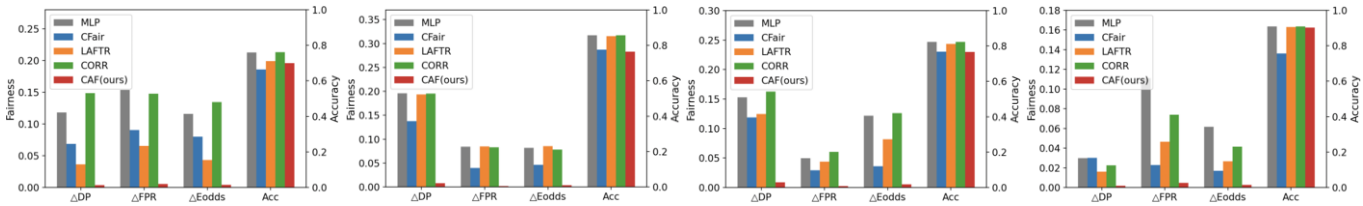
To verify the effect of our approach on different fairness metrics, we conducted the controlled experiment with a fixed baseline network architecture for each dataset. The MLP model in our work consists of three layers: input layer, hidden layer and output layer, where the first layer is set as the encoder and the remaining two layers are used as the classification head. The target prediction loss function is the cross entropy loss function. We call the proposed method CAF, which aligns the prediction results of different groups with correlation by domain adaptation method, and minimizes the regularized loss function to improve fairness.

We compare our framework against baselines such as MLP (Multilayer Perceptron) without debiasing and some popular debiasing methods CORR (Correlation Loss) [7], LAFTR [33], CFair [52].

CORR [7] : Regularization model, that is, a model that combines specific regularization methods on the basis of common classification tasks. CORR encourage the model to maintain a low absolute correlation between the predictive output and the sensitive attribute of the negative examples, thereby reducing the model's dependence on the sensitive attribute to achieve a fairer prediction.

LAFTR [33] : A multi-task adversarial model, that is, a network with a shared hidden layer and two distinct classification layers. LAFTR adds a new classification layer to the MLP to predict sensitive attributes. LAFTR is designed to pit two classifiers against each other, one to minimize label prediction loss which used to improve accuracy, and one to maximize sensitive attribute prediction loss which used to reduce the impact of sensitive attributes.

CFair [52] : A multi-task adversarial model. On the basis of LAFTR model, CFair increases the number of adversarial layers, that is, one adversarial layer (sensitive attribute prediction layer) is built



(a) German Credit dataset (b) Adult dataset (c) MEPS dataset (d) Law School dataset

Figure 1. Comparisons of fairness and accuracy of different algorithms on benchmark datasets.

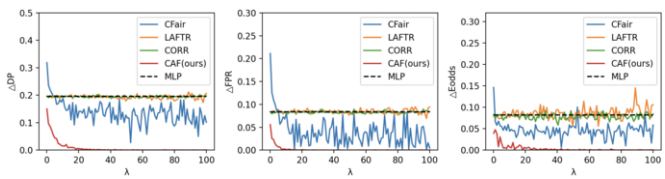


Figure 2. Comparisons of fairness on the Adult dataset.

for one group, so that there is finally a shared hidden layer, a label classification layer and $|\mathcal{A}|$ adversarial layers.

5.3 *Performance analysis on bias mitigation*

Performance on utility and fairness metrics. We compare the bias mitigation performance of CAF with other competing methods and illustrate their results in Figure 1. Each subfigure shows how different algorithms perform on accuracy and fairness metrics for each dataset. For each dataset, we evenly select a range of fairness coefficients λ to obtain the value of each metric, and then calculate the mean of the values to plot figures.

On average, CAF achieves significant fairness optimization on multiple datasets without conflicts between fairness metrics. In contrast, although the accuracy of another regularization method CORR does not decrease significantly, the debiasing effect was not obvious, and sometimes may exacerbate bias, such as on the MEPS dataset. Moreover, CORR cannot simultaneously achieve multiple fairness, such as the ΔDP and $\Delta Eodds$ metrics on the German Credit dataset. LAFTR and CFair are debiasing methods using adversarial strategy, and their debiasing effects are better than CORR. But LAFTR and CFair, like CORR method, do not guarantee all fairness metrics simultaneously. For example, LAFTR can't guarantee fairness metric $\Delta Eodds$ on the Adult dataset, CFair can't guarantee fairness metric ΔDP on the Law School dataset. In addition, the debiasing effect of CFair is at the expense of much model utility, which is worse than that of other networks including CAF.

In addition, we also plot the fairness curves under different values of coefficient λ. For space reasons, we only show the results on the Adult dataset in Figure 2, the other datasets are similar. In the process of adjusting λ, the fairness gap of CAF is always smaller than those of the baseline network MLP and other debiasing networks. With the increase of λ, the fairness gap after debiasing by CAF algorithm shows a decreasing trend and finally approaches 0, which reflects the effectiveness of our method to repair fairness. However, we find that perfect fairness may result from the model simply classifying all the data into the majority class, so the choice of fairness coefficient needs to trade off the pursuit of utility, and we advise readers not to use a particularly large λ.

Fairness-utility trade-off. In order to fully discuss the trade-off between fairness and accuracy, we compute the Pareto front on the basis of the results in Figure 1 and further plot Figure 3.

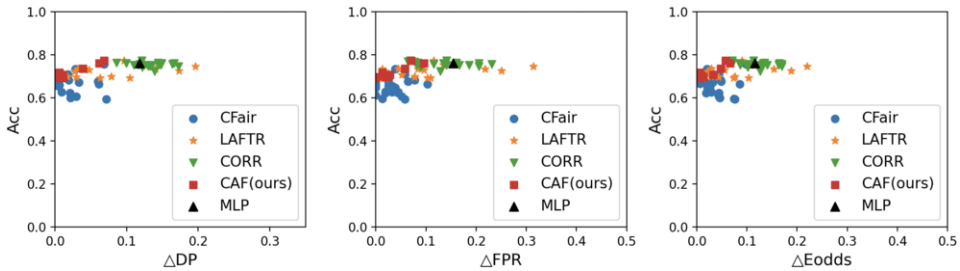
It is known that the closer the Acc value is to 1, the more accurate the prediction is, and the closer the fairness gap value is to 0, the fairer the classification is. Therefore, the closer the line is to the top left corner in the figure, the better trade-off between fairness and accuracy is achieved. Obviously, the red areas corresponding to the CAF method are more concentrated on the upper-left corner in each figure. In comparison, the accuracy of data points of the other methods is not significantly better than that of CAF, but the values of fairness metrics are too scattered and unstable, and some methods even worsen the fairness issues of the original MLP. It suggests that the CAF method performs better than the comparative work in the trade-off on different datasets and different fairness metrics.

Multiple sensitive attributes. We discuss whether CAF can achieve fairness improvement in the case of multiple sensitive attributes. On the basis of Algorithm 2, we calculate the prediction distribution difference of multiple sensitive attributes respectively, and penalize the mean of these differences. The experimental results obtained on the Adult dataset, MEPS dataset, and Law School dataset are plotted in Figure 4. Through the analysis of the first three columns in Figure 4, we can find that CAF can improve fairness on multiple sensitive attributes simultaneously, and there is no obvious conflict between different sensitive attributes. The last column in Figure 4 show that fairness optimization of multiple sensitive attributes at the same time may result in a slight loss of joint accuracy. However, in some cases, the joint accuracy will be higher than the results obtained by optimizing for certain sensitive attributes individually. For example, in the Law School dataset, the joint accuracy is higher than the result of optimizing attribute 'race' individually. In the MEPS dataset, the joint accuracy is close to the results of the other two cases, and sometimes slightly better than the result of optimizing attribute 'sex' individually. In addition, as the case of optimizing one sensitive attribute, accuracy for the joint optimization finally converges to the stage where the majority class label is simply predicted to attain the absolute fairness, when the coefficient λ is extremely increased.

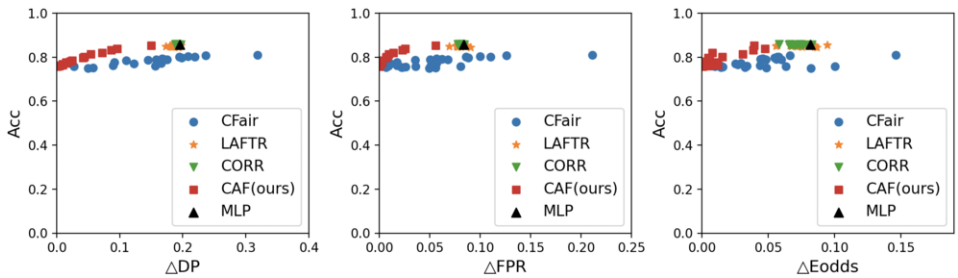
In conclusion, CAF can improve multiple fairness metrics without conflicting with each other, and in the case of the same hyperparameter λ, the debiasing effect of our method is more obvious. This conclusion also applies to cases where there are multiple sensitive attributes. At the same time, our algorithm has a better trade-off between fairness and utility, that is, compared with other methods, we can obtain much better model fairness with slight accuracy loss.

5.4 *CAF with debiased network*

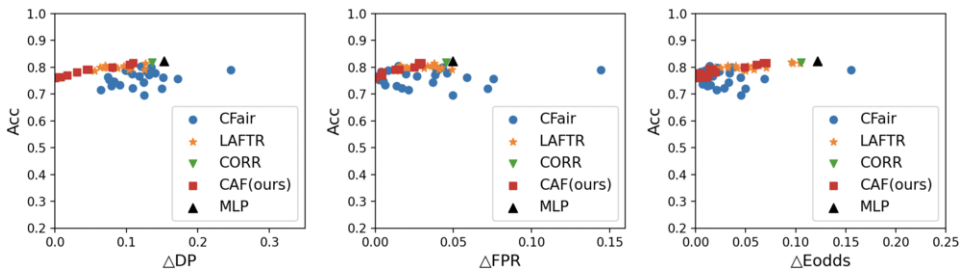
In the discussions so far, we have reported the performance of CAF on MLP, a network without any debiasing strategy. In this section,



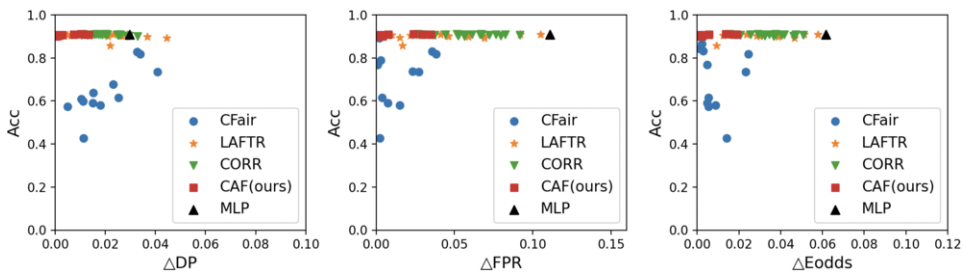
(a) German Credit dataset



(b) Adult dataset



(c) MEPS dataset



(d) Law School dataset

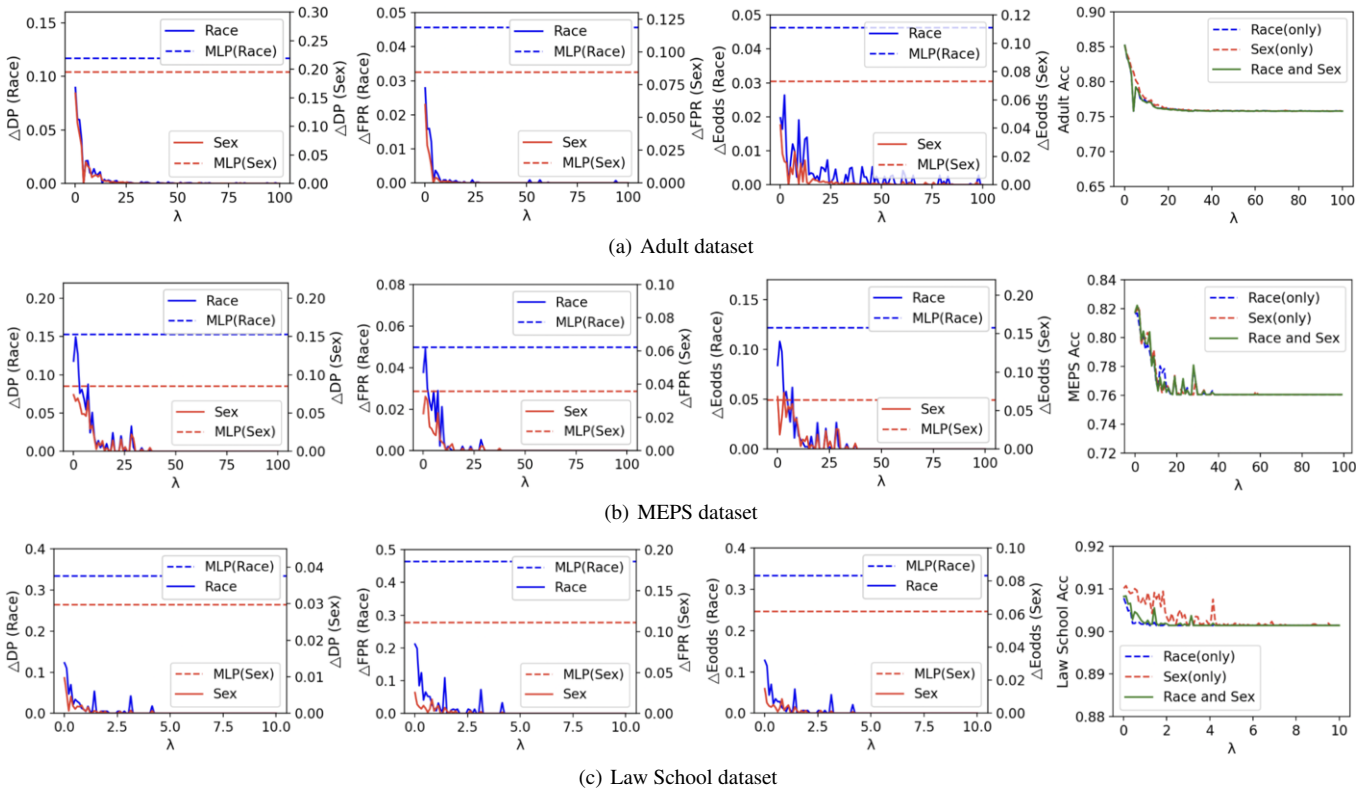
Figure 3. Fairness-accuracy tradeoff comparisons of CAF with other baselines on the German Credit dataset, Adult dataset, MEPS dataset and Law School dataset. We sample 100 values of λ evenly spaced from the ranges [1,1000],[0.1,100],[0.1,100] and [0.01,10] respectively.

we will further discuss whether CAF can improve the fairness of debiased networks and whether it can solve the problem of fairness conflicts of adversarial networks.

In order to answer these questions, we combine the CAF algorithm with LAFTR and CFair respectively, to conduct experiments on benchmark datasets. We first use adversarial training to learn the debiased encoder, which seeks to yield fair representations [33]. Then we use it as the primary network, combined with CAF, to update the encoder to see if it can get better fairness performance compared with MLP network and the original encoder. We take MEPS dataset as an example to show the experimental results, where we set λ_LAFTR=1 and λ_CFair=10. We evenly select 100 values of λ_CAF in the range of [0.1, 100] to obtain values of metrics, and calculate the average of these records to plot Figure 5.

A similar trend can be observed on different debiased networks, the CAF algorithm can better maintain and even optimize the group fairness of different baseline networks. In addition, the optimized model can significantly improve the fairness of the model with a small loss of utility. With the CAF algorithm, LAFTR can convert 6.3% accuracy loss into 94.4% improvement of ΔDP, 95.5% improvement of ΔFPR, and 98.5% improvement of $\Delta Eodds$. Similarly, the CFair model can also significantly improve group fairness with a small loss of accuracy after being combined.

In addition, we also observe that the network combined with CAF can solve the problem of conflicting fairness metrics caused by the original network. Taking the CFair network as an example, in the



(a) Adult dataset

(b) MEPS dataset

(c) Law School dataset

Figure 4. Fairness and accuracy curves after discrimination mitigation for multiple sensitive attributes on the Adult dataset, MEPS dataset and Law School dataset. We sample 100 values of λ evenly spaced from the ranges [0.1,100],[0.1,100] and [0.01,10] respectively.

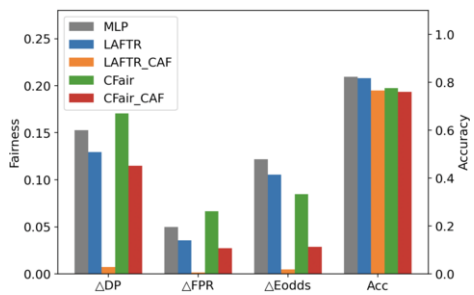


Figure 5. Fairness performance of debiased network combined with CAF algorithm on MEPS dataset.

MEPS dataset, compared with MLP network, both the fairness metrics ΔDP and ΔFPR show the problem of increased discrimination. After combining with CAF, CFair_CAF not only improves three fairness metrics, but also alleviates the conflict problem, so that the fairness metrics after bias mitigation are better than them from MLP.

In conclusion, this experiment shows that CAF can not only improve the fairness of MLP, but also further improve the performance of debiased networks, and even bypass the contradiction of different fairness metrics of adversarial networks. The integration of CAF boost the performance of the original encoder in terms of fairness, which only has a little loss in utility.

6 Conclusion and future work

In this paper we propose a novel in-processing method CAF to improve group fairness by punishing second-order statistical differences in prediction probabilities between groups with different sensitive attributes. By comparing bias mitigation effects of different models on different datasets, our experiments confirm that CAF can improve multiple group fairness metrics simultaneously with little accuracy loss, and can improve fairness of multiple sensitive attributes simultaneously without conflicting with each other. In addition, CAF can also be combined with the debiased networks to further improve fairness and mitigate the fair conflict problem in them.

The CAF algorithm provides a better trade-off between group fairness and utility compared to other methods, but the accuracy still has a certain loss. In addition, CAF as a lightweight regularization method, is suitable for differentiable models in a variety of application scenarios, which in turn guides social decision-making towards fairness. However, for non-parametric models, such as decision trees, a direct application may not be feasible. Therefore, future work will focus on combining pre-processing and post-processing methods to improve utility and expand the range of applications.

Acknowledgements

This work is funded by the “Digital Silk Road” Shanghai International Joint Lab of Trustworthy Intelligent Software (No. 22510750100) and NSFC Programs (No. 62161146001, No. 62372176).

References

[1] N. Agarwal, A. Sondhi, K. Chopra, and G. Singh. Transfer learning: Survey and classification. *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS 2020*, pages 145–155, 2021.

[2] T. AI. Oecd digital economy papers. 2023.

[3] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[4] V. Aseervatham, C. Lex, and M. Spindler. How do unisex rating regulations affect gender differences in insurance premiums? *The Geneva Papers on Risk and Insurance-Issues and Practice*, 41:128–160, 2016.

[5] A. Asuncion and D. Newman. Uci machine learning repository, 2007.

[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[7] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.

[8] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

[9] E. Burnaev, P. Erofeev, and A. Papanov. Influence of resampling on accuracy of imbalanced classification. In *Eighth international conference on machine vision (ICMV 2015)*, volume 9875, pages 423–427. SPIE, 2015.

[10] T. A. Cleary. Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. *ETS Research Bulletin Series*, 1966(2):i–23, 1966.

[11] T. A. Cleary. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968.

[12] J. W. Cohen, S. B. Cohen, and J. S. Banthin. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. *Medical care*, pages S44–S50, 2009.

[13] B. d'Alessandro, C. O'Neil, and T. LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[14] D. Das and C. G. Lee. Unsupervised domain adaptation using regularized hyper-graph matching. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3758–3762. IEEE, 2018.

[15] D. Das and C. G. Lee. Graph matching and pseudo-label guided deep unsupervised domain adaptation. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 342–352. Springer, 2018.

[16] D. Das and C. G. Lee. Sample-to-sample correspondence for unsupervised domain adaptation. *Engineering Applications of Artificial Intelligence*, 73:80–91, 2018.

[17] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[18] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[19] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005.

[20] M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[22] B. Fang, M. Jiang, P.-y. Cheng, J. Shen, and Y. Fang. Achieving outcome fairness in machine learning models for social decision problems. In *IJCAI*, pages 444–450, 2020.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[24] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 144–152. SIAM, 2016.

[25] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

[26] R. M. Guion. Employment tests and discriminatory hiring. *Industrial Relations: A Journal of Economy and Society*, 5(2):20–37, 1966.

[27] C. Haas. The price of fairness-a framework to explore trade-offs in algorithmic fairness. 2019.

[28] H. Hofmann. Uci statlog (german credit data) data set. *UCI Machine Learning Repository*, 2000.

[29] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[30] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34 (11):2767–2787, 2010.

[31] S. Kiritchenko and S. M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

[32] M. Loi and M. Christen. How to include ethics in machine learning research. *Ercim News*, 116(3):5, 2019.

[33] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[35] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[36] O. A. Osoba, W. Welser IV, and W. Welser. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

[37] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[38] S. Plan. The national artificial intelligence research and development strategic plan. *National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee*, 2016.

[39] K. Ross and C. Carter. Women and news: A long and winding road. *Media, Culture & Society*, 33(8):1148–1165, 2011.

[40] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

[41] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[42] N. A. Smuha. The eu approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4):97–106, 2019.

[43] E. K. Spanakis and S. H. Golden. Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports*, 13:814–823, 2013.

[44] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[45] M. Wan, D. Zha, N. Liu, and N. Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

[46] L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[47] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[48] A. Yapo and J. Weiss. Ethical implications of bias in machine learning. 2018.

[49] L. Yarger, F. Cobb Payton, and B. Neupane. Algorithmic equity in the hiring of underrepresented it job candidates. *Online information review*, 44(2):383–395, 2020.

[50] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

[51] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[52] H. Zhao, A. Coston, T. Adel, and G. J. Gordon. Conditional learning of fair representations, 2020.

[53] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[54] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.