Faster Fair Machine via Transferring Fairness Constraints to Virtual Samples

Zhou Zhai[1], Lei Luo[2], Heng Huang [3], Bin Gu [1, 4*]

[1] School of Computer and Software, Nanjing University of Information Science and Technology, P.R.China
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, P.R.China
[3] Department of Electrical & Computer Engineering, University of Pittsburgh, USA
[4] Department of Machine Learning, MBZUAI, United Arab Emirates
zhouzhai@nuist.edu.cn, luoleipitt@gmail.com, henghuanghh@gmail.com, jsgubin@gmail.com

Abstract

Fair classification is an emerging and important research topic in machine learning community. Existing methods usually formulate the fairness metrics as additional inequality constraints, and then embed them into the original objective. This makes fair classification problems unable to be effectively tackled by some solvers specific to unconstrained optimization. Although many new tailored algorithms have been designed to attempt to overcome this limitation, they often increase additional computation burden and cannot cope with all types of fairness metrics. To address these challenging issues, in this paper, we propose a novel method for fair classification. Specifically, we theoretically demonstrate that all types of fairness with linear and non-linear covariance functions can be transferred to two virtual samples, which makes the existing state-of-the-art classification solvers be applicable to these cases. Meanwhile, we generalize the proposed method to multiple fairness constraints. We take SVM as an example to show the effectiveness of our new idea. Empirically, we test the proposed method on real-world datasets and all results confirm its excellent performance.

Introduction

Fair classification has been a topic of intense study in machine learning due to concerns to bias with respect to sensitive attributes, *e.g.*, against African-Americans while predicting future criminals (Flores, Bechtel, and Lowenkamp 2016) or NYPD stop-question-and-frisk program (Goel et al. 2016). Fairness means that there is no prejudice or favoritism based on the sensitive attributes (*e.g.* race, gender) of individuals or groups (Mehrabi et al. 2021). Since fairness is a complex and multi-faceted concept that depends on environment and culture, a number of notions are proposed to assess how fair a classifier is with respect to a sensitive group when compared to another, *e.g.*, disparate treatment (Xu et al. 2018), demographic parity (Barocas and Selbst 2016) and disparate mistreatment (e.g., average odds difference and equal opportunity difference) (Zafar et al. 2017a). A number of recent works have focused on designing the fair classification algorithms that satisfy one or multiple fairness constraints. According to whether to modify the underlying classifier, previous works in fairness measures and bias mitigation can be divided into two broad types.

Specifically, the first one treats the underlying classifier as a black box and implements a wrapper that works by pre-processing the data or post-processing the classifier's predictions (Calmon et al. 2017; Chzhen et al. 2019). Existing pre-processing approaches are specific to particular definitions of fairness and typically seek to come up with a single transformed data set that will work across all learning algorithms, which, in practice, leads to classifiers that still exhibit substantial unfairness. In contrast, post-processing allows a broader definition of fairness and produces provable fairness guarantees. However, post-processing would lead to unpredictable losses in accuracy and requires test-time access to the protected attribute.

The second one modifies existing fair classifiers through relaxing the desired definitions of fairness (Wu, Zhang, and Wu 2019; Donini et al. 2018; Cotter et al. 2019; Gu et al. 2022). For example, (Zafar et al. 2017b) learnt the convex margin-based classifiers by imposing linear constraints on the covariance between the predicted labels and the values of certain features. (Zafar et al. 2017a) utilized the discipline convex-concave program to handle the scenarios where disparate treatment and disparate mistreatment can be avoided. (Agarwal et al. 2018) provided a method to compute a nearly optimal fair classifier with respect to linear fairness constraints, like demographic parity or equalized odds, by the Lagrangian method.

Although a lot of fair classification algorithms have been proposed as mentioned above, all of them suffer from one or more of the following limitations. (i) Existing efficient solvers to original learning problems were not compatible to the fairness constraints which makes us to re-explore new tailored algorithms to accomplish fairness classification. However, the new tailored algorithms are normally not as efficient as the original ones. (ii) Existing fairness algorithms cannot support all types of fairness constraints as shown in Table 1. (iii) Fairness notions can be relaxed into linear and non-linear proxy fairness constraints. However, existing algorithms cannot handle both of them for each type of fairness notion. To sum up, it is still an open problem to design a faster and flexible classification algorithm that can handle all fairness notions regarding both linear and non-linear proxy functions.

*Corresponding Author

Method	DT	DP	EOP	OMR	FPR	FNR	FOR	FDR	MFC	TYPE
(Zafar et al. 2017b)	✓	✓	✓	×	×	×	×	×	✓	Linear
(Zafar et al. 2017a)	✓	×	×	✓	✓	✓	×	×	✓	Non-Linear
(Hardt, Price, and Srebro 2016)	×	×	×	×	✓	✓	×	×	✓	Linear
(Woodworth et al. 2017)	×	×	×	×	✓	✓	×	×	×	Linear
(Feldman et al. 2015)	✓	✓	✓	×	×	×	×	×	×	Linear
(Agarwal et al. 2018)	✓	×	×	×	✓	✓	×	×	×	Non-Linear
(Lohaus, Perrot, and von Luxburg 2020)	✓	✓	✓	×	×	×	×	×	×	Linear
Our	✓	✓	✓	✓	✓	✓	✓	✓	✓	Both

Table 1: Capabilities of different algorithms in handling different fairness notions. "MFC"denotes the multiple fairness constraints. "TYPE" refer to the forms of proxy fairness constraints used in the algorithms.

To address these challenging issues, in this paper, we reformulate the constrained optimization problem as a regularized optimization problem by making use of its dual form, in which the fairness constraints are moved into the objective and the corresponding Lagrange multipliers act as regularizers. Then, we provide a new framework for solving fairness classification problems using virtual samples to replace fairness constraints. We theoretically prove that all types of fairness with linear and non-linear covariance functions can be transferred to two virtual samples, which makes the existing state-of-the-art classification solvers be applicable to these cases. Meanwhile, we generalize the proposed method to multiple fairness constraints. We take SVM as an example to show the effectiveness of our new idea. Finally, the experimental results on real-world datasets show that our method achieves excellent performance compared with existing fair classification algorithms.

Contributions. The main contributions of this paper are summarized as follows.

1. We provide a new framework for solving fairness classification problems using virtual samples to replace fairness constraints. We take SVM as an example to demonstrate the effectiveness of our method on real-world datasets.
2. We theoretically prove that all types of fairness with linear and non-linear covariance functions can be transferred to two virtual samples and generalize the proposed method to multiple fairness constraints.

Fairness in Classification

Problem Setting

In this paper, we consider Support Vector Machine (SVM) for binary classification. Specifically, given a set of training samples $D = \{(x_i, y_i)\}_{i=1}^n$, where user feature vectors $x \in \mathbb{R}^d$ and ground truth labels $y \in \{-1, +1\}$. SVM tries to learn a linear discriminant function $f_\theta(x) = \langle w, \phi(x)\rangle + b$ [1] in the appropriately chosen kernel induced feature space to separate the training data. $\theta = (w, b)$ are the parameters of the model. Then, for a given unseen feature vector x, the classifier outputs the predicted label $\hat{y} = 1$ if $f_\theta(x) \geq 0$ and $\hat{y} = -1$ otherwise.

Assume that each user has an associate sensitive feature z (*e.g.*, gender, race). For ease of exposition, we assume z_i to be binary, *i.e.*, $z \in \{0, 1\}$. However, the classifier $f(x)$ cannot use the protected characteristic z at decision time, as it will constitute an unfair treatment. A number of notions have been used to determine how fair a classifier is with respect to a sensitive feature when compared to another, such as disparate treatment (DT) (Xu et al. 2018), demographic parity (DP) (Zafar et al. 2017b), equality of opportunity (EOP) (Lohaus, Perrot, and von Luxburg 2020), disparate mistreatment (include overall misclassification rate (OMR), false-positive rate (FPR), false-negative rate (FNR), false-omission rate (FOR) and false-discovery rate (FDR)) (Zafar et al. 2017a).

Fair Classification

In order to ensure that the learned discriminant function is fair, appropriate conditions can be incorporated into the classifier formulation. The general fair classification problems can be formulated as follows.

$$\begin{aligned} \min_{\theta \overset{\text{def}}{=} (w,b)} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} L(f_\theta(x_i), y_i) \\ s.t. \quad & |\Omega(f_\theta)| < \epsilon \end{aligned} \tag{1}$$

where C is the regularization parameter, $L(f_\theta(x_i), y_i)$ is the loss function to estimate the degree of inconsistency between the predicted label $\hat{y}$ and ground truth label y_i, $\Omega(\cdot)$ is conditional risk difference for specific fairness requirement (e.g., in terms of demographic parity, $\Omega(f_\theta) = P(\hat{y} = 1|z = 0) - P(\hat{y} = 1|z = 1)$), $\epsilon \in \mathbb{R}^+$ and the smaller ϵ is, the more fair the decision boundary would be.

However, it is difficult to directly solve (1) because measuring the explicit conditional probability is intractable. To overcome the difficulty, (Zafar et al. 2017b,a) proposed a tractable proxy by measuring fairness metrics using the covariance between the users sensitive attributes z and the signed distance $g(y, x)$ between the feature vectors and the classifier decision boundary:

$$\begin{aligned} Cov(z, g(y, x)) &= \mathbb{E}[(z - \bar{z})(g(y, x) - \bar{g}(y, x))] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})g(y_i, x_i) \end{aligned}$$

where $\bar{z}$ and $\bar{g}(y, x)$ are average values of z and $g(y, x)$ respectively, the term $\mathbb{E}[(z - \bar{z})\bar{g}_\theta(x)]$ cancels out since $\mathbb{E}[(z - \bar{z})] = 0$. Note that the $g(y, x)$ can be divided into linear and non-linear according to different fairness constraints.

[1] $\phi(\cdot)$ is transformation function from an input space to a high-dimensional reproducing kernel Hilbert space.

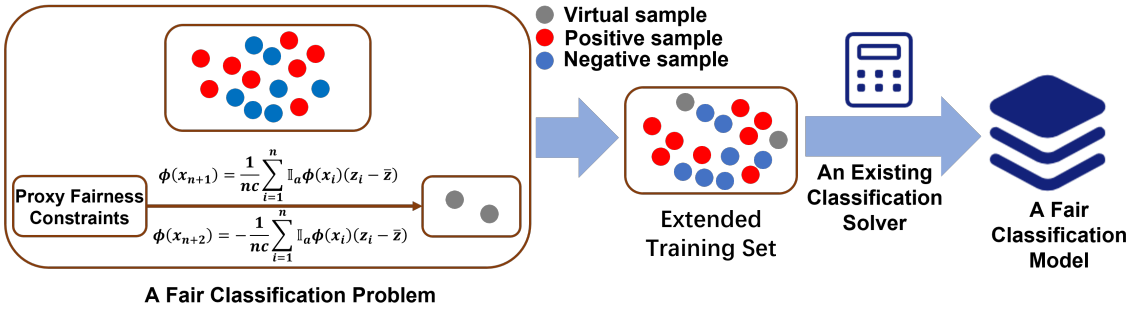


Figure 1: Our method transfers the fairness constraints to virtual samples, which makes the fair SVM problem be efficiently solved by the existing state-of-the-art classification solvers.

Thus, the formulation (1) can be rewritten as follows:

$$\min_{\theta \stackrel{\text{def}}{=} (w,b)} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} L(f_\theta(x_i), y_i) \qquad (2)$$
$$s.t. \quad \left|\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})g(y_i, x_i)\right| \leq c$$

where c is the covariance threshold which specifies an upper bound on the covariance between the sensitive attributes z and the signed distance $g(y, x)$. Since the formulation (2) does not need sensitive attributes to construct the decision hyperplane, we can directly remove disparate treatment by keeping the features x disjoint from sensitive attribute z.

Fair Classification Algorithm Using Virtual Samples

In this section, we take traditional SVMs (Cortes and Vapnik 1995; Gu et al. 2018; Zhai et al. 2020) as an example. Note that the results of SVMs can also be extended to other classification problems such as logistic regression (Kleinbaum et al. 2002).

Specifically, we first rewrite the above problem as a regularized optimization problem by making use of its dual, in which the fairness constraints are moved to the objective and considered as regularizers in the corresponding Lagrange function. Then, we transfer the fairness constraints to virtual samples, and use the existing solver to directly solve the fairness constrained classification problems. Specifically, we use the following virtual samples to replace fairness constraints:

$$\phi(x_{n+1}) = \frac{1}{nc}\sum_{i=1}^{n} \mathbb{I}_a \phi(x_i)(z_i - \bar{z}) \qquad (3)$$

$$\phi(x_{n+2}) = -\frac{1}{nc}\sum_{i=1}^{n} \mathbb{I}_a \phi(x_i)(z_i - \bar{z}), \qquad (4)$$

and $y_{n+1} = -1$, $y_{n+2} = 1$, where $\mathbb{I}_a$ is an indicator function for a specific fairness requirement that returns 1 when a is true and 0 otherwise. We summarize the flow diagram of our method in Fig. 1. The virtual samples only depend on if the derivative of $f_\theta(x)$ regarding to θ can be formulated by a linear combination of $\phi(x)$. As long as it holds, virtual samples which are a linear combination of $\phi(x_i)$ can be obtained correspondingly. Thus, our idea can also work for a lot of algorithms which include (kernelized) logistic regression, (kernelized) additive model, even boosting algorithms (Schapire and Singer 1998). In the following, we discuss the transformation of linear and non-linear fairness constraints respectively.

Linear Fairness Constraint

DP is defined as each group having the same probability of being classified as a positive outcome, *i.e.*, $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$. We can set $g(y_i, x_i) = f(x_i)$ in the problem (2) to obtain a fair SVM in terms of DP. Since we measure the average signed distance from all the samples to the decision boundary to measure the fairness, we construct two virtual samples to represent the average position of the samples. Then, we control the distance from the two virtual samples to the decision boundary to control the average signed distance from the real sample to the decision boundary. Specifically, we construct two virtual samples $\{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2})\}$ such that

$$\phi(x_{n+1}) = \frac{1}{nc}\sum_{i=1}^{n} \phi(x_i)(z_i - \bar{z}) \qquad (5)$$

$$\phi(x_{n+2}) = -\frac{1}{nc}\sum_{i=1}^{n} \phi(x_i)(z_i - \bar{z}), \qquad (6)$$

and $y_{n+1} = 1$, $y_{n+2} = 1$. Then, we transform the optimization problem (2) into the following minimization problem:

$$\min_{0 \leq \alpha \leq C} \quad \frac{1}{2}\sum_{i,j=1}^{n+2} y_i y_j \alpha_i \alpha_j \phi(x_i)\phi(x_j) - \sum_{i=1}^{n+2} \alpha_i \qquad (7)$$
$$s.t. \quad \sum_{i=1}^{n+2} y_i \alpha_i = 0$$

The problem (7) can be solved efficiently by many existing solvers (*e.g.* LIBSVM (Chang and Lin 2011) with a little modification and the quadprog function in MATLAB).

According to KKT conditions, for virtual samples $\{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2})\}$, the optimal solution satisfies:

$$y_{n+1} f_\theta(x_{n+1}) - 1 \geq 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i)(z_i - \bar{z}) \leq c$$

$$y_{n+2}f_\theta(x_{n+2}) - 1 \geq 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i)(z_i - \bar{z}) \geq -c$$

We show the equivalence between virtual samples $\{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2})\}$ and demographic parity in Theorem 1.

Theorem 1. *Given training set $D = \{(x_i, y_i)\}_{i=1}^n$, sensitive attributes $\{z_i\}_{i=1}^n$, fairness parameter c. In fair SVM, we can transform the demographic parity into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$, which are defined in (5)-(6).*

Proof. Considering the demographic parity, the fair SVM can be formulated as follows.

$$\begin{aligned} \min_{\theta \overset{\text{def}}{=} (w,b)} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} L(f_\theta(x_i), y_i) \qquad (8) \\ s.t. \quad & \left|\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z}) f_\theta(x_i)\right| \leq c \end{aligned}$$

Introducing Lagrangian multipliers $\alpha_{n+1}, \alpha_{n+2}, \alpha$ and v corresponding to inequality constraints in (8), we can write the Lagrangian of optimization problem (8) with linear fairness constraints as

$$\begin{aligned} \mathcal{L}(\theta, \xi, \alpha, v) \quad = \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i \qquad (9) \\ & -\alpha_{n+1}\left(-\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i)(z_i - \bar{z}) + c\right) \\ & -\alpha_{n+2}\left(\frac{1}{n}\sum_{i=1}^{n} f_\theta(x_i)(z_i - \bar{z}) + c\right) \\ & -\sum_{i=1}^{n}\alpha_i(y_i f_\theta(x_i) - 1 + \xi_i) - \sum_{i=1}^{n} v_i\xi_i \end{aligned}$$

where ξ is the slack variables, $\alpha_{n+1}, \alpha_{n+2}, \alpha$ and v are non-negative.

Then, we calculate the derivatives with respect to the primal variables, which yields

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} \quad = \quad & w - \sum_{i=1}^{n} y_i\alpha_i\phi(x_i) + \frac{\alpha_{n+1}}{n}\sum_{i=1}^{n}\phi(x_i)(z_i - \bar{z}) \\ & -\frac{\alpha_{n+2}}{n}\sum_{i=1}^{n}\phi(x_i)(z_i - \bar{z}) \\ \frac{\partial \mathcal{L}}{\partial b} \quad = \quad & -\sum_{i=1}^{n} y_i\alpha_i + \alpha_{n+1} - \alpha_{n+2} \end{aligned}$$

For simplifying the notation, we define two extra special examples x_{n+1}, x_{n+2} in an implicit manner:

$$\phi(x_{n+1}) \quad = \quad \frac{1}{nc}\sum_{i=1}^{n}\phi(x_i)(z_i - \bar{z}) \qquad (10)$$

$$\phi(x_{n+2}) \quad = \quad -\frac{1}{nc}\sum_{i=1}^{n}\phi(x_i)(z_i - \bar{z}) \qquad (11)$$

and we set $y_{n+1} = 1$, $y_{n+2} = 1$. Forcing the derivatives to zero, we can obtain:

$$w = \sum_{i=1}^{n+2} y_i\alpha_i\phi(x_i), \quad 0 = \sum_{i=1}^{n+2} y_i\alpha_i\,.$$

We transform the optimization problem (2) into the following minimization problem:

$$\begin{aligned} \min_{0 \leq \alpha \leq C} \quad & \frac{1}{2}\sum_{i,j=1}^{n+2} y_i y_j \alpha_i \alpha_j \phi(x_i)\phi(x_j) - \sum_{i=1}^{n+2}\alpha_i \qquad (12) \\ s.t. \quad & \sum_{i=1}^{n+2} y_i\alpha_i = 0 \end{aligned}$$

The proof is completed. □

Remark 1. *The $\phi(x_{n+1})$ and $\phi(x_{n+2})$ are difficult to be computed in high-dimensional reproducing kernel Hibbert space (RKHS). However, when training the SVM model, we only need to compute the inner product of any two vectors under RKHS, which is easily computed by the kernel function.*

Remark 2. *According to Theorem 1, we can directly extend the virtual samples to other linear fairness constraints, such as EOP. A classifier satisfies EOP if the probability of getting a true positive is independent of the value of the sensitive attribute: $P(\hat{y} = 1|z = 0, y = 1) = P(\hat{y} = 1|z = 1, y = 1)$. We can set $g(y_i, x_i) = \frac{1+y_i}{2} f(x_i)$ in the problem (2) to obtain a fair SVM in terms of EOP. Therefore, the EOP can be transformed to two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$ as defined in (3)-(4) with $\mathbb{I}_a = \mathbb{I}_{y=1}$.*

Non-Linear Fairness Constraint

A binary classifier does not suffer from disparate mistreatment if the over misclassification rates (OMR) for different groups of people having different values of the sensitive feature z are the same.

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1) \qquad (13)$$

We can set $g(y_i, x_i) = \min(0, y_i f(x_i))$ in the problem (2) to obtain a fair SVM in terms of OMR, which makes the fairness constraint in problem (2) non-linear. Because OMR only considers the fairness of misclassified samples to sensitive attributes, we can set $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y}$ in (3)-(4) and transform the OMR into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$. Since we do not know the corresponding predicted labels of the samples in advance, we first train a classifier without fairness constraints to obtain the predicted labels. We use these prediction labels to construct virtual samples to train the new model. We repeat the construction of new virtual samples to train the new model until all the predicted labels corresponding to the samples converge. At this point, the optimal solutions of the virtual samples satisfy the corresponding fairness constraint according to the KKT conditions. We show the equivalence between virtual samples $\{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2})\}$ and OMR in Theorem 2.

Theorem 2. *Given training set $D = \{(x_i, y_i)\}_{i=1}^n$, sensitive attributes $\{z_i\}_{i=1}^n$, fairness parameter c. We can transform the OMR to two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$ as defined in (3)-(4) with $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y}$.*

Proof. Considering the overall misclassification rate, the fair SVM can be formulated as follows.

$$\min_{\theta \overset{\text{def}}{=} (w,b)} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} L(f_\theta(x_i), y_i) \qquad (14)$$

$$s.t. \qquad \left| \frac{1}{n} \sum_{i=1}^{n} (z_i - \bar{z}) \min(0, y_i f_\theta(x_i)) \right| \leq c$$

Introducing Lagrangian multipliers $\alpha_{n+1}, \alpha_{n+2}, \alpha$ and v corresponding to inequality constraints in (14), we can write the Lagrangian of optimization problem (14) with non-linear fairness constraints as

$$\begin{aligned}
\mathcal{L}(\theta, \xi, \alpha, v) &= \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \\
&-\alpha_{n+1}\left(-\frac{1}{n}\sum_{i=1}^{n} \min(0, y_i f_\theta(x_i))(z_i - \bar{z}) + c \right) \\
&-\alpha_{n+2}\left(\frac{1}{n}\sum_{i=1}^{n} \min(0, y_i f_\theta(x_i))(z_i - \bar{z}) + c \right) \\
&-\sum_{i=1}^{n} \alpha_i (y_i f_\theta(x_i) - 1 + \xi_i) - \sum_{i=1}^{n} v_i \xi_i
\end{aligned} \tag{15}$$

where ξ is the slack variables, $\alpha_{n+1}, \alpha_{n+2}, \alpha$ and v are non-negative. We employ the Concave-Convex Procedure (CCCP) (Yuille and Rangarajan 2003) to solve the regularized optimization problem (15). The main mechanism of CCCP algorithm is to iteratively construct an optimized surrogate objective function which linearizes the concave part of the original Difference of Convex programming (Sriperumbudur and Lanckriet 2009) problem. The concave part of (15) is defined as follows:

$$\begin{aligned}
J_{cav}(\theta) = \quad & \alpha_{n+1} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{I}_{z_i=1} \min(0, y_i f_\theta(x_i))(z_i - \bar{z}) \\
& -\alpha_{n+2} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{I}_{z_i=0} \min(0, y_i f_\theta(x_i))(z_i - \bar{z})
\end{aligned}$$

In order to apply the CCCP update, we first have to calculate the derivative of the concave part with respect to θ:

$$\begin{aligned}
\frac{\partial J_{cav}(\theta)}{\partial \theta} =& \alpha_{n+1} \sum_{i=1}^{n} \mathbb{I}_{z_i=1} \frac{\partial J_{cav}(\theta)}{\partial f_\theta(x_i)} \frac{\partial f_\theta(x_i)}{\partial \theta} (z_i - \bar{z}) \\
& - \alpha_{n+2} \sum_{i=1}^{n} \mathbb{I}_{z_i=0} \frac{\partial J_{cav}(\theta)}{\partial f_\theta(x_i)} \frac{\partial f_\theta(x_i)}{\partial \theta} (z_i - \bar{z})
\end{aligned} \tag{16}$$

We introduce the notation

$$\begin{aligned}
\mu_i &= a_{n+1} \mathbb{I}_{z_i=1} \frac{\partial J_{cav}(\theta)}{\partial f_\theta(x_i)} - a_{n+2} \mathbb{I}_{z_i=0} \frac{\partial J_{cav}(\theta)}{\partial f_\theta(x_i)} \qquad (17) \\
&= \begin{cases} a_{n+1} & \text{if} \quad y_i f_\theta(x_i) < 0, z_i = 1 \\ a_{n+2} & \text{if} \quad y_i f_\theta(x_i) < 0, z_i = 0 \\ 0 & otherwise. \end{cases}
\end{aligned}$$

For simplifying the notation, we define two extra special examples x_{n+1}, x_{n+2} in an implicit manner:

$$\phi(x_{n+1}) = \frac{1}{nc} \sum_{i=1}^{n} \mathbb{I}_{\hat{y} \neq y} \phi(x_i)(z_i - \bar{z}) \tag{18}$$

$$\phi(x_{n+2}) = -\frac{1}{nc} \sum_{i=1}^{n} \mathbb{I}_{\hat{y} \neq y} \phi(x_i)(z_i - \bar{z}), \tag{19}$$

and we set $y_{n+1} = 1, y_{n+2} = 1$. Since $f_\theta(x_i) = w \cdot \phi(x_i) + b$ with $\theta = (w, b)$, and $\frac{\partial f_\theta(x_i)}{\partial \theta} = (\phi(x_i), 1)$, we can calculate the $J'_{cav}(\theta) \cdot \theta$ as follows

$$\begin{aligned}
J'_{cav}(\theta) \cdot \theta = \quad & \beta_{n+1} y_{n+1} [w \cdot \phi(x_{n+1}) + b] \\
- \quad & \beta_{n+2} y_{n+2} [w \cdot \phi(x_{n+2}) + b]
\end{aligned}$$

where $\beta_{n+1} = a_{n+1}|D_0|$ and $\beta_{n+2} = a_{n+2}|D_0|$, D_0 and D_1 as subsets of the training dataset D taking values $z_i = 0$ and $z_i = 1$ respectively. Then, define $\beta_i = 0$ for $1 \leq i \leq n$, we calculate the derivatives with respect to the primal variables, which yields

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w} &= w - \sum_{i=1}^{n+2} y_i (\alpha_i - \beta_i) \phi(x_i) \\
\frac{\partial \mathcal{L}}{\partial b} &= -\sum_{i=1}^{n+2} y_i (\alpha_i - \beta_i)
\end{aligned}$$

We can transform problem (14) into following convex inner loop problem of CCCP:

$$\begin{aligned}
\min_{0 \leq \alpha \leq C} \quad & \frac{1}{2} \sum_{i,j=1}^{n+2} y_i y_j (\alpha_i - \beta_i)(\alpha_j - \beta_j) \phi(x_i) \phi(x_j) - \sum_{i=1}^{n+2} \alpha_i \\
s.t. \quad & \sum_{i=1}^{n+2} y_i (\alpha_i - \beta_i) = 0
\end{aligned} \tag{20}$$

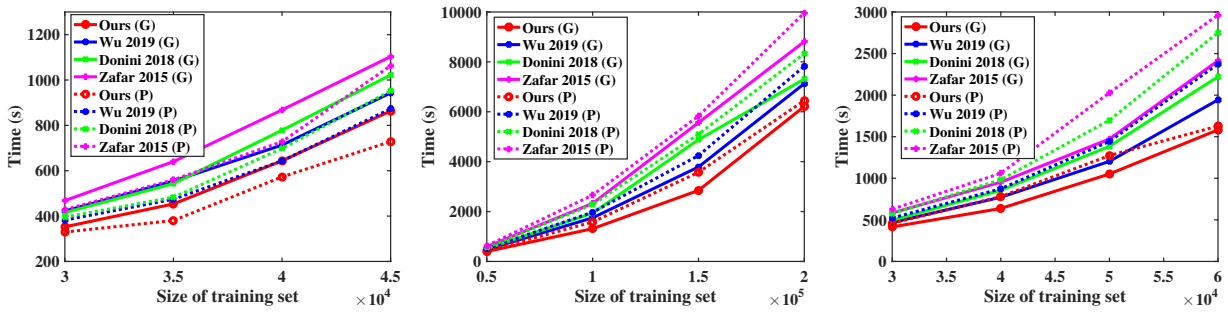
We solve the convex inner loop problem (20) iteratively until the the prediction label $\hat{y}$ converges. The proof is completed. □

Remark 3. *According to Theorem 2, we can directly extend the virtual samples to other non-linear fairness constraints as follows.*

1. *The FPR considers the fraction of misclassified samples in cases where the ground truth label is negative. We can set $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y \& y=-1}$ in (3)-(4) and transform the FPR into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$.*
2. *The FNR considers the fraction of misclassified samples in cases where the ground truth label is positive. We can set $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y \& y=1}$ in (3)-(4) and transform the FNR into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$.*
3. *The FOR considers the fraction of misclassified samples in cases where the predicted label is negative. We can set $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y \& \hat{y}=-1}$, $y_{n+1} = 1$ in (3)-(4) and transform the FOR into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$.*
4. *The FDR considers the fraction of misclassified samples in cases where the predicted label is positive. We can set $\mathbb{I}_a = \mathbb{I}_{\hat{y} \neq y \& \hat{y}=1}$ in (3)-(4) and transform the FDR into two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$.*

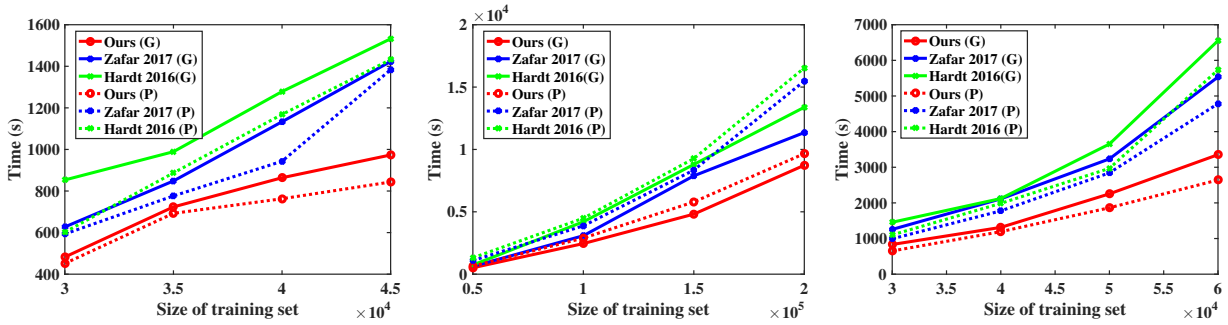
Satisfying Multiple Fairness Constraints Simultaneously

In certain application scenarios, it might be desirable to satisfy multiple fairness constraints simultaneously. Since the different proportion of samples with positive labels among groups with different sensitive attribute values, it is impossible to construct a classifier that satisfies the equal false discovery rate and false omission rate criterion, or also satisfies the equal false positive and false negative rate criterion (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2016). However, in practice, it may still be interesting to explore the best, even if imperfect, extent of fairness a



(a) Adult (b) CelebA (c) Dutch

Figure 2: Average running time of fair classification algorithm for demographic parity criterion.



(a) Adult (b) CelebA (c) Dutch

Figure 3: Average running time of fair classification algorithm for false-positive rate criterion.

classifier can achieve. For example, suppose z_1 is the average value of sensitive attributes in positive samples and z_2 is the mean of sensitive attributes in negative samples, since z_1 and z_2 are unlikely to be equal in practical problems, we need the average value of sensitive attributes in misclassified samples to be close to z_1 and z_2. Thus, we can set $\mathbb{I}_a = \mathbb{I}_{\hat{y}\neq y\&y=-1}$ and $\bar{z} = \frac{z_1+z_2}{2}$, and construct two virtual samples $\{(x_{n+1}, 1), (x_{n+2}, 1)\}$ as defined in (3)-(4) to satisfy FNR and FPR simultaneously. The classifier constructed in this way can meet FNR and FPR as much as possible.

Experiments

Experiment Setup

Design of Experiments: In this section, we conduct the experiments on several synthetic and real-world datasets to evaluate the effectiveness of our method in controlling fairness. To show the advantage of our method in computational efficiency, we compare the running time of our method with other fair classification algorithms. To show that our method can efficiently achieve comparable fairness compared to related fair algorithms, we compare the fairness of classifiers under different fairness constraints. Specifically, the compared algorithms are summarized as follows.

1. Zafar 2015 (Zafar et al. 2017b) and Zafar 2017 (Zafar et al. 2017a): They quantify fairness using the covariance between the users' sensitive attribute and the signed distance from the feature vectors to the decision boundary.
2. Donini 2018 (Donini et al. 2018): They present an approach based on empirical risk minimization.
3. Wu 2019 (Wu, Zhang, and Wu 2019): They propose a constraint-free criterion under which the learned classifier.
4. Hardt 2016 (Hardt, Price, and Srebro 2016): They operate by post-processing the outcomes of an unfair classifier.
5. Ours: We transfer the fairness constraints to virtual samples, which makes the existing state-of-the-art classification solvers be applicable to these cases.

Implementation: All of our experiments are conducted using SVM. We repeatedly split each dataset into a train (75%) and test (25%) set. The regularization parameter C in SVM is fixed at 10. The Gaussian kernel $K(x_1, x_2) = \exp(\kappa\|x_1 - x_2\|^2)$ with $\kappa = 0.5$ and Polynomial kernel $K(x_1, x_2) = (x_1 x_2 + 1)^2$ is used in all the experiments. For each dataset, we first calculate the fairness score $c^* = \left|\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})g(y_i, x_i)\right|$ of the model without fairness constraints. Then, we set the fairness constrain value $c = \frac{1}{2}c^*$ for our proposed methods. The compared methods are the same setting. The results are the average of 10 trials.
Datesets: Table 2 summarizes the datasets used in the experiments. We use two synthetic datasets (i.e., SynthOpp

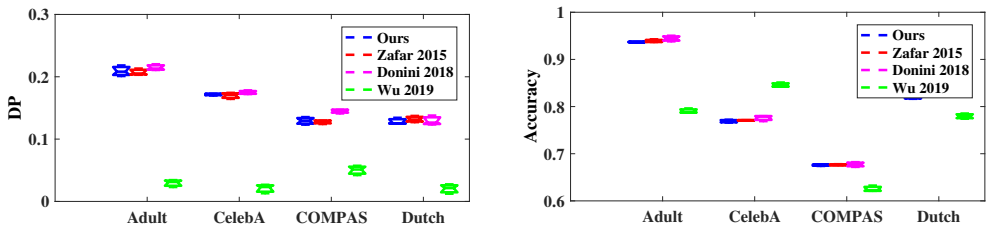


Figure 4: Performance of different methods for demographic parity criterion (Gaussian kernel).

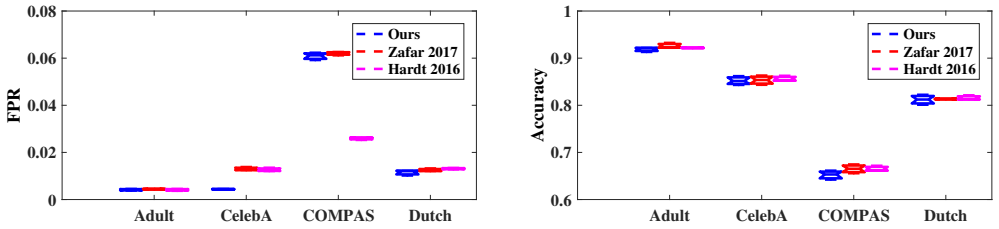


Figure 5: Performance of different methods for false-positive rate criterion (Gaussian kernel).

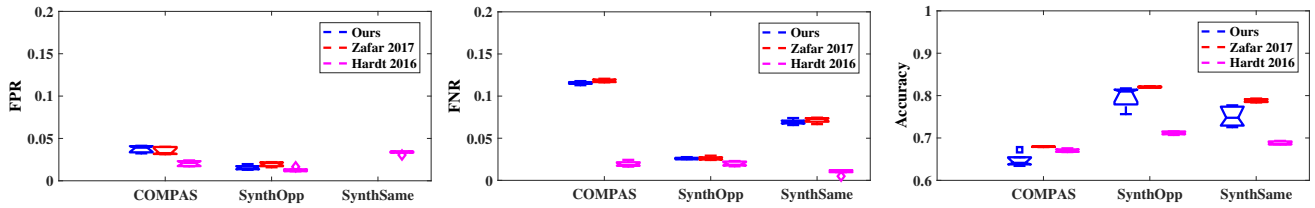


Figure 6: Performance of different methods for both FPR and FNR criterions (Gaussian kernel).

and SynthSame) whose details are provided in Appendix. The sensitive attribute is listed in the column "Sensitive Attribute" of Table 2.

Dataset	Size	Dimensionality	Sensitive Attribute
Adult	48,842	14	Gender
CelebA	202,599	40	Gender
Dutch	60,420	12	Gender
COMPAS	7,214	53	Race
SynthOpp	2,500	2	-
SynthSame	2,500	2	-

Table 2: The datasets used in the experiment.

Results and Discussions

Figure 2 presents the average run time (in seconds) of compared fair classification algorithms for DP criterion. Figure 3 presents the average run time (in seconds) of compared fair classification algorithms for FPR criterion. The results clearly demonstrate that our method is significantly faster than other fair classification algorithms with various types of fair constraints. This is because, our method transforms fairness constraints to virtual samples, which can make the existing fast classification solvers be applicable to the fairness constrained classification problems.

Figure 4 shows the performance of different methods for demographic parity criterion. Figure 5 shows the performance of different methods for false positive rate criterion. The results show that our method and baseline algorithms have the similar accuracies and fairness scores on fair SVM. These results confirm the superiority of our method, *i.e.*, much faster than existing fair classification algorithms, while retaining the similar accuracies and fairness scores as discussed above.

Figure 6 shows the performance of different methods for both false-positive rate and false-negative rate criterions. The results show that our method can effectively solve fairness classification problems with multiple constraints by constructing multiple virtual samples corresponding to fairness constraints.

Conclusion

We propose a novel method for fair classification via transferring the fairness constraints to virtual samples, which can make the existing state-of-the-art classification solvers be applicable to the fairness constrained classification problems. We use SVM as an example of machine learning models, and show that our new idea is working on traditional convex SVM. Empirically, we test our method on real-world datasets and all results confirm its excellent performance.

Acknowledgments

Bin Gu was partially supported by the National Natural Science Foundation of China (No:61573191)

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104: 671.

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

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3): 1–27.

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

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, 12760–12770.

Cortes, C.; and Vapnik, V. 1995. Support-vector networks. *Machine learning*, 20(3): 273–297.

Cotter, A.; Jiang, H.; Gupta, M. R.; Wang, S.; Narayan, T.; You, S.; and Sridharan, K. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *J. Mach. Learn. Res.*, 20(172): 1–59.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2791–2801.

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

Flores, A. W.; Bechtel, K.; and Lowenkamp, C. T. 2016. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80: 38.

Goel, S.; Rao, J. M.; Shroff, R.; et al. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1): 365–394.

Gu, B.; Yuan, X.; Chen, S.; and Huang, H. 2018. New Incremental Learning Algorithm for Semi-Supervised Support Vector Machine. In Guo, Y.; and Farooq, F., eds., *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, 1475–1484. ACM.

Gu, B.; Zhai, Z.; Li, X.; and Huang, H. 2022. Towards Fairer Classifier via True Fairness Score Path. In Hasan, M. A.; and Xiong, L., eds., *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3113–3121. ACM.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

Kleinbaum, D. G.; Dietz, K.; Gail, M.; Klein, M.; and Klein, M. 2002. *Logistic regression*. Springer.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lohaus, M.; Perrot, M.; and von Luxburg, U. 2020. Too Relaxed to Be Fair. In *International Conference on Machine Learning*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Schapire, R. E.; and Singer, Y. 1998. Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*, 80–91.

Sriperumbudur, B. K.; and Lanckriet, G. R. 2009. On the convergence of the concave-convex procedure. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 1759–1767. Curran Associates Inc.

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

Wu, Y.; Zhang, L.; and Wu, X. 2019. On Convexity and Bounds of Fairness-aware Classification. In *The World Wide Web Conference*, 3356–3362.

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.

Yuille, A. L.; and Rangarajan, A. 2003. The concave-convex procedure. *Neural computation*, 15(4): 915–936.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970.

Zhai, Z.; Gu, B.; Li, X.; and Huang, H. 2020. Safe Sample Screening for Robust Support Vector Machine. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 6981–6988. AAAI Press.