Improving Prediction Fairness via Model Ensemble

Dheeraj Bhaskaruni
Department of Computer Science
University of Wyoming
Laramie, Wyoming
Email: vbhaskar@uwyo.edu

Hui Hu
Department of Computer Science
University of Wyoming
Laramie, Wyoming
Email: hhu1@uwyo.edu

Chao Lan
Department of Computer Science
University of Wyoming
Laramie, Wyoming
Email: clan@uwyo.edu

***Abstract*—Fair machine learning is a topical problem. It studies how to mitigate unethical bias against minority people in model prediction. A promising solution is ensemble learning – Nina et al [1] first argue that one can obtain a fair model by bagging a set of standard models. However, they do not present any empirical evidence or discuss effective ensemble strategy for fair learning.**

In this paper, we propose a new ensemble strategy for fair learning. It adopts the AdaBoost framework, but unlike AdaBoost that upweights mispredicted instances, it upweights unfairly predicted instances which we identify using a variant of Luong's k-NN based situation testing method [2]. Through experiments on two real-world data sets, we show our proposed strategy achieves higher fairness than the bagging strategy discussed by Nina et al and several baseline methods. Our results also suggest standard ensemble strategies may not be sufficient for improving fairness.

I. Introduction

Machine learning models are ubiquitous today. However, studies show that many models are unethically biased against minority people. For example, they systematically score more innocent black defendants as high risk reoffender [3] (compared with non-black defendants), systematically underscore qualified female job applicants in Amazon's AI-hiring system [4] and job platform XING [5] (compared with male job applicants), or make higher mis-classification rates on black customers in facial verification tasks [6] (compared with non-black customers). There is an urgent need to mitigate these biases [7] to maintain public trust in machine learning applications in society.

In recent years, many approaches have been proposed to mitigate prediction bias, such as label preprocessing [2], [8] , feature preprocessing [9], [10], model regularization [11], [12] and model postprocessing [13].

We notice that a promising approach is ensemble learning. In 2017, Nina et al [1] first argue that one can obtain a fair model by bagging a set of standard models, since unethical biases in these models may be averaged out through bagging. However, their arguments remain at theoretical level and no empirical evidence is presented. Besides, their arguments are based on the standard bagging strategy, and there is no discussion on what ensemble strategy is more effective for learning fair models.

In this paper, we propose a new ensemble heuristic and show it achieves higher model fairness than several baseline methods on real-world data sets. Surprisingly, in our experiments, standard ensemble strategies such as bagging [1] do not effectively improve fairness. Comparatively, our properly designed ensemble heuristic significantly improves fairness.

Technically, our heuristic adopts the AdaBoost framework. But unlike AdaBoost that upweights mis-predicted instances, we upweight unfairly predicted instances by properly designed weights; unfairly predicted instances are identified using the k-NN based situation testing technique developed by Luong et al [2]. Our hypothesis is that, if a model learns to pay more attention on instances that can be easily unfairly predicted, the model can reduce (or avoid) bias in its prediction.

Through experiments on two real-world data sets [14], [15], we show the proposed heuristic achieves significantly higher model fairness than the bagging strategy discussed in [1] as well as several other baseline methods. We also observe that ensemble models can achieve higher fairness than single models, but *only* with properly designed ensemble strategies.

The rest of the paper is organized as follows: in Section II, we review related works; in Section III, we present the proposed ensemble strategy; experimental results are presented and discussed in Section IV and conclusions in Section V.

II. Related work

A. Overview of Fair Machine Learning

Fair machine learning is an emerging field that investigates and mitigates unethical bias in algorithm predictions [16].

Several fairness notions have been studied, such as statistical disparity [10], equalized odds [13], individual fairness [11] and preference [17]. The most common notion is statistical disparity, which states a model is fair if it has similar positive prediction rates across different demographic groups. It is rooted in the legal notion of 80%-rule in the United States labor law [18]. In this work, we focus on statistical disparity.

Many fair learning approaches are proposed. Label preprocessing [2], [8] assumes there are unfair labels in training data; they detect and correct these labels before standard learning is performed on training data. Feature preprocessing [9], [10] assumes a model is fair if it is built on fair features; they first learn fair features and then learn a standard model on them. Model regularization [11], [12] directly penalizes unethical bias during learning. Model postprocessing [13] learns a standard model and modifies its predictions to make them fair. Finally, model ensemble [1] assumes an ensemble of standard (unfair) models is fair as their biases may be averaged out.

2375-0197/19/$31.00 ©2019 IEEE
DOI 10.1109/ICTAI.2019.00273

B. Overview of Ensemble Methods

Ensemble learning is a powerful tool [19]. Its central idea is to ensemble a set of (typically weak) base models to obtain a strong and robust model.

There are many strategies to ensemble models. Two common ones are bagging and adaboost. In bagging, each base model is trained on a bootstrap of the training set; all base models are then averaged to generate a strong model. In adaboost, base models are trained in order, and each model is optimized by minimizing a weighted loss on the training set; an instance has higher weight if it is mis-predicted by the previous models; finally, all base models are averaged in a weighted fashion to generate a strong model, where a base model has higher weight if it performs more accurately on the training set.

C. Ensemble Method for Fair Machine Learning

Nina et al [1] is the first work that points out ensemble method can improve model fairness. Their key argument is that, even if each base model is biased, their averaged model may be fair as those biases can cancel each other. This is an interesting argument. We present an example below.

An Example of How Bagging can Improve Fairness

Consider the task of applying machine learning to predict whether an employee deserves to be promoted or not. Our goal is to remove gender bias in prediction. Let x be a random employee and f_1, f_2 be two base models. Suppose f_1 is biased against females with predicted promotion probabilities

$$\begin{aligned} p_{1a} &= \Pr\{x \text{ is promoted} \mid x \text{ is male}, f_1\} = 0.5 \\ p_{1b} &= \Pr\{x \text{ is promoted} \mid x \text{ is female}, f_1\} = 0.1, \end{aligned} \tag{1}$$

where 'x is promoted' means 'x' is predicted by the model as deserving to be promoted.

Then, by definition, the statistical disparity of f_1 is

$$SD(f_1) = p_{1a} - p_{1b} = 0.4. \tag{2}$$

Similarly, suppose f_2 is biased against male with

$$\begin{aligned} p_{2a} &= \Pr\{x \text{ is promoted} \mid x \text{ is male}, f_2\} = 0.1 \\ p_{2b} &= \Pr\{x \text{ is promoted} \mid x \text{ is female}, f_2\} = 0.5, \end{aligned} \tag{3}$$

and thus

$$SD(f_2) = p_{2a} - p_{2b} = -0.4. \tag{4}$$

Now, consider an ensemble model f that randomly picks f_1 or f_2, each with 50% chance, to make prediction.[1] Its predicted promotion probability for males is

$$\begin{aligned} p_a &= \Pr\{x \text{ is promoted} \mid x \text{ is male}\} \\ &= p_{1a}p(f_1) + p_{2a}p(f_2) \\ &= 0.5 \cdot 50\% + 0.1 \cdot 50\% = 0.3. \end{aligned} \tag{5}$$

Similarly, its predicted promotion probability on female is

$$p_b = p_{1b}p(f_1) + p_{2b}p(f_2) = 0.3. \tag{6}$$

[1]This is the theoretical model discussed in Nina et al [1]. We can easily verify the expected prediction of this model equals the prediction of bagging model (that averages all model outcomes as the final outcome).

Thus, the statistical disparity of this ensemble model is

$$SD(f) = p_a - p_b = 0.3 - 0.3 = 0. \tag{7}$$

Comparing $SD(f)$ with $SD(f_1)$, $SD(f_2)$, we see each base model has large statistical disparity (thus unfair); however, the ensemble model has zero statistical disparity (thus fair).

Limitations in Nina et al's Work

While Nina et al points out the potential of using ensemble to improve model fairness, their work has two limitations.

First, their arguments remain at the theoretical level, and no empirical evidence is presented in the work. *How much fairness can ensemble models achieve on real-world data sets*? This is the first question we aim to address in this paper.

Second, their arguments are based on a standard ensemble strategy named bagging. *Are standard ensemble strategies optimal for learning fair models*? *Can we design new ensemble strategies to improve fairness more efficiently*? We will also address these questions in the paper.

D. K-NN Situation Testing

Luong et al [2] proposes a k-NN based technique to detect unfair labels in a data set. Motivated by the legal technique of situation testing, they make the assumption that an instance has unfair label if (i) the instance lies in a neighborhood that has large statistical disparity, (ii) the instance belongs to the minority group such as female and (iii) the label is disadvantaged such as 'not promoted'. They propose to first detect and correct these unfair labels in the training set, and then perform standard machine learning on the set. They empirically show this improves model fairness.

In this paper, we modify Luong et al's method and apply it to identify unfairly predicted labels. There are modifications. First, we perform detection based on predicted labels instead of true labels. Second, we identify unfairly predicted labels only based on (i) but not (ii) or (iii). We will elaborate the detailed detection algorithm in the methodology section.

III. Proposed Ensemble Strategy for Fair Learning

A. Preliminaries

We will describe an instance using a tripe (x, s, y), in which x is non-sensitive feature, s is sensitive demographic feature (e.g., gender or race) and y is label.

Let there be a training set $L = \{(x_i, s_i, y_i)\}_{i=1,\ldots,n}$, where (x_i, s_i, y_i) is the i_{th} instance in the set.

Let f be an ensemble model mapping from x to y.[2] Let f_t be the t_{th} base model of f, and assume $t = 1, \ldots, m$.

B. The Proposed Ensemble Strategy

We begin with the AdaBoost framework. It defines the ensemble model as

$$f(x) = \sum_{t=1}^{m} \alpha_t f_t(x), \tag{8}$$

[2]We can also assume f maps from (x, s) to y w.l.o.g..

where α_t is the model weight for f_t. The base models will be trained sequentially, and model f_t will be trained by minimizing a weighted loss on L, i.e.,

$$f_t = \arg\min_h \sum_{i=1}^{n} w_i^{(t)} \cdot \text{loss}(h(x_i), y_i), \tag{9}$$

where loss$(\cdot)$ is any loss function and $w_i^{(t)}$ is the weight for x_i. Instance weights of the first model are initialized to one.

We define a function $\delta_i^{(t)}$ to indicate whether x_i is unfairly predicted by f_t, i.e.,

$$\delta_i^{(t)} = \begin{cases} 1, & \text{if } x_i \text{ is unfairly predicted by } f_t \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

We will elaborate how to identify unfair predictions later. For now, motivated by AdaBoost, we design model weights as

$$\alpha_t = \ln\left\{\frac{1-\epsilon_t}{\epsilon_t}\right\}, \tag{11}$$

where

$$\epsilon_t = \frac{\sum_{i=1}^{n} w_i^{(t)} \delta_i^{(t)}}{\sum_{i=1}^{n} w_i^{(t)}}. \tag{12}$$

Intuitively, model f_t receives higher weight in the ensemble if it makes fair prediction on more instances. Indeed, being fair on more instances means more $\delta_i^{(t)}$ will be zero, implying smaller ϵ_t and thus larger α_t.

Then, we design update rule for instance weight as

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp\{\alpha_t \cdot \delta_i^{(t)}\}. \tag{13}$$

Intuitively, an instance receives higher weight if it is unfairly predicted by the previous model – it is our hypothesis that paying more attention to learning such data could improve fairness of the ensemble model more efficiently.

Now, it remains to design the algorithm to identify unfairly predicted instances in (10).

C. Identify Unfairly Predicted Instances

We will modify the k-NN based situation testing method [2] and apply it to detect unfairly predicted instances.

For training instance x_i, let $N_{i,k}$ be the set of its k-nearest neighbors (identified based on Euclidean distance between instances). From the context in Section II.C., we first define the statistical disparity of model f_t in $N_{i,k}$ as

$$\begin{aligned} SD(f_t; N_{i,k}) = &\Pr\{x \text{ is promoted} \mid x \text{ is male}, x \in N_{i,k}\} \\ &- \Pr\{x \text{ is promoted} \mid x \text{ is female}, x \in N_{i,k}\} \end{aligned} \tag{14}$$

Then, we identify x_i as unfairly predicted if $SD(f; N_{i,k})$ is bigger than a threshold r (and vice versa). Thus (10) becomes

$$\delta_i^{(t)} = \begin{cases} 1, & SD(f_t; N_{i,k}) > r \\ 0, & SD(f_t; N_{i,k}) \le r. \end{cases} \tag{15}$$

To better understand the detection process, we give an example in Figure 1. Let $*$ be the instance being examined and set $k = 5$. We see that $\Pr\{x \text{ is promoted} \mid x \text{ is male}, x \in N_{i,k}\} = 2/3$ and $\Pr\{x \text{ is promoted} \mid x \text{ is female}, x \in N_{i,k}\} = 1/2$. Therefore, $SD(f_t; N_{i,k}) = 2/3 - 1/2 \approx 0.17$. If we set threshold $r = 0.1$, then $SD(f_t; N_{i,k}) > r$ and instance $*$ will be considered as unfairly predicted by f_t.

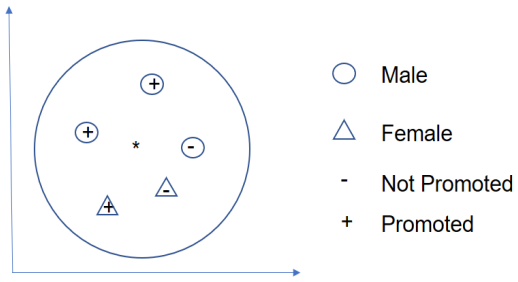

Fig. 1. K-NN situation testing

D. Summary of Ensemble Learning Algorithm

The proposed ensemble fair learning method is summarized in Algorithm 1. It is built on AdaBoost, but differs at steps 3 and 4. At Step 3, AdaBoost increases model weight α_t if f_t is accurate, but we increase it if f_t is fair. At Step 4, AdaBoost increases instance weight $w_i^{(t+1)}$ if x_i is mis-predicted by f_t, but we increase it if x_i is unfairly predicted by f_t.

Algorithm 1 Proposed Ensemble Fair Learning Method

Input: training set L, ensemble size m, neighborhood size k and identification threshold r.
Initialize: instance weight $w_i^{(1)} = 1/n$, $i = 1, \ldots, n$
for $t = 1, \ldots, m$ **do**
1: train base model f_t based on (9).
2: compute identifier $\delta_i^{(t)}$ based on (14), (15)
3: compute model weight α_t based on (11), (12)
4: update instance weight $w_i^{(t+1)}$ based on (13)
end for
Output: ensemble model $f = \sum_{t=1}^{m} \alpha_t f_t$.

IV. Experiment

A. Experiment Data

We experimented on two public data sets that are commonly used for fair learning: the Credit Default data set[3] and the Community Crime data set[4].

The Credit Default data set contains 30,000 instances and 23 features (e.g.,default payments, gender, history of payment). We treated education degree as the sensitive attribute and binarized it into 'higher' and 'lower' in the same way as [20]. We treated default payment (1=yes, 0=no) as the binary label. We down-sampled the data set from 30,000 to 20,000.

The Communities Crime data set contains 1,993 instances, each described by 101 features (e.g. percent of the population under the poverty line, percent of divorced males in the

[3]https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[4]http://archive.ics.uci.edu/ml/datasets/communities+and+crime

community, violent crimes per population). We treated the fraction of African-American residents as the sensitive feature, and binarized it so that a community is considered minority if the fraction is above 0.5 and majority otherwise.The predictive variable is the community crime rate that is binarized into high if the rate is above 0.5 and low otherwise.

Both our preprocessed data sets are available at https://uwyomachinelearning.github.io/.

B. Experiment Design

On each data set, we randomly chose 75% instances for training and used the rest for testing. We performed each experimented method over 50 random trials and reported the average performance.

We compared the proposed ensemble model with two popular ensemble models, namely, bagging (discussed in Nina et al) and AdaBoost. Logistic regression is used as the base model.

We also compared with two single models: standard logistic regression and the existing fair logistic regression [21]. For fair logistic regression, we used the same set of hyper-parameters as used in the original paper.

We used three evaluation metrics: classifier error, statistical disparity (SD) and equalized odds (EO) [13]. EO is a variant of SD that further conditions SD on instances with the same true label. Following the context in Section II.C., EO(f) is the difference between the following two probabilities

$$\Pr\{f(x) = \text{promoted} \mid x \text{ is male, actually promoted}\}$$

$$\Pr\{f(x) = \text{promoted} \mid x \text{ is female, actually promoted}\}.$$

C. Results and Discussions

Experimental results on all examined methods on two data sets are shown in Table I and Table II respectively.

In Table I, we see our proposed method achieves significantly lower statistical disparity and equalized odds than other methods, suggesting its strength in learning fair models. But we also observe this method achieves higher prediction error, suggesting a trade-off between fairness and accuracy. (Other methods cannot achieve the same fairness as our method, even when we re-balanced their accuracy-fairness trade-off.)

Surprisingly, we see both standard ensemble strategies do not naturally improve fairness (as argued in previous study) – they perform similarly to standard logistic regression and worse than fair logistic regression. Comparatively, our properly designed ensemble strategy outperforms both standard and fair single models. These results suggest that ensemble learning can improve fairness only through proper ensemble strategies.

To double-check if the standard ensemble strategy could improve fairness in certain random trials, we plot the error bars of bagging and our method in Figure 2. We see that bagging, even in its best case, still has very high statistical disparity (0.11). Comparatively, our method, even in its worst case, still achieves much lower disparity (0.04). This suggest our proposed ensemble strategy is more effective and robust for learning fair models.

Finally, our proposed ensemble of logistic regression significantly outperforms the existing fair logistic regression method.

Method	SD	EO	Error
LR	.1000 ± .0000	.4695 ± .0000	.1883 ± .0000
FairLR [21]	.0898 ± .0971	.0620 ± .0882	.1166 ± .0189
Bagging [1]	.2267 ± .2025	.2855 ± .1867	.1187 ± .0987
AdaBoost	.0746 ± .0124	.3712 ± .0889	**.1013 ± .0117**
Our Method	**.0239 ± .0247**	**.0593 ± .0367**	.1604 ± .0445

TABLE I
CLASSIFICATION PERFORMANCE ON THE COMMUNITY CRIME DATA SET. LR STANDARDS FOR LOGISTIC REGRESSION.

Method	SD	EO	Error
LR	.1531 ± .0000	.1161 ± .0000	.3438 ± .0000
FairLR	.0779 ± .0571	.1256 ± .0112	**.2412 ± .0469**
Bagging	.1915 ± .1766	.1267 ± .1024	.3066 ± .0277
AdaBoost	.1697 ± .0335	.1104 ± .0625	.2877 ± .0238
Our Method	**.0213 ± .0171**	**.0019 ± .0000**	.4486 ± .0589

TABLE II
CLASSIFICATION PERFORMANCE ON THE CREDIT CARD DATA SET

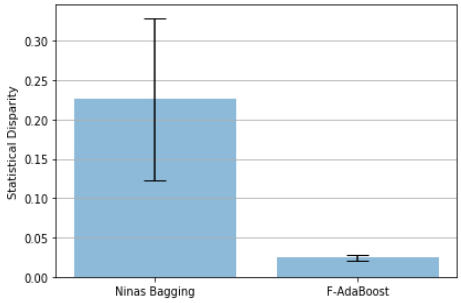

Fig. 2. Statistical Disparity of 50 Random Trials on Community Crime

This suggests that ensemble learning is indeed a promising approach for fair learning (with the right ensemble strategy).

Similar observations are found in Table II.

D. Sensitivity Analysis

In this section, we performed sensitivity analysis of the proposed ensemble fair learning method.

In Figure 3 and 4, we show statistical disparity and prediction error versus the number of base models on two data sets respectively. First, we see that fairness is improved as more base models are added, and the improvement starts to converge when the number passes 50 (on both data sets). Second, we see prediction error increases as more base models are added. This may be explained by the fact that our method is not designed to boost accuracy – it pays more attention to unfairly predicted instances and fair models. Combing both observations, we see a trade-off between accuracy and fairness in the proposed method. How to reduce this trade-off remains an open question.

In Figure 5, we show statistical disparity versus the neighborhood size (k in Algorithm 1). We see that neither small k or large k gives the highest fairness – the optimal values are achieved at 11 and 15 on two data sets respectively. This makes sense, because too small of k may not include enough examples

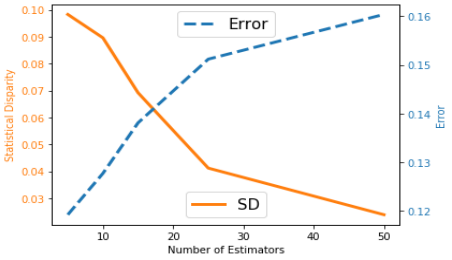

Fig. 3. Performance on Community Crime

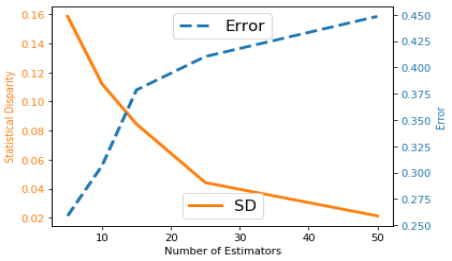

Fig. 4. Performance on Credit Card

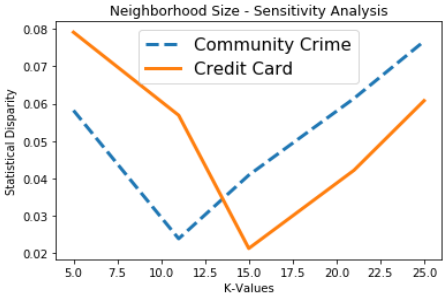

Fig. 5. Statistical Disparity versus k

for accurate estimation of disparity in the neighborhood, while too large of k may include neighbors that are generated from a different distribution and thus not representative of the disparity in the neighborhood. How to identify the optimal k remains an open question.

Finally, in Figure 6 we show fairness versus the threshold r. The performance is similar to that in Figure 5. This also makes sense – too small of r will identify every instance as unfairly predicted (thus increasing weights of all instances), and too large r will identify no instance as unfairly predicted (thus increasing no weight). How to identify the optimal r remains an open question.

V. Conclusions and Future Work

In this paper, we propose a new ensemble strategy to learn fair models. It adopts the AdaBoost framework, and upweights unfairly predicted instances when learning base models. We show our method achieves higher fairness than the prior work as well as several baseline methods. In particular, our results suggest that standard ensemble strategies do not naturally improve fairness, and one should carefully design ensemble strategies for learning fair models.

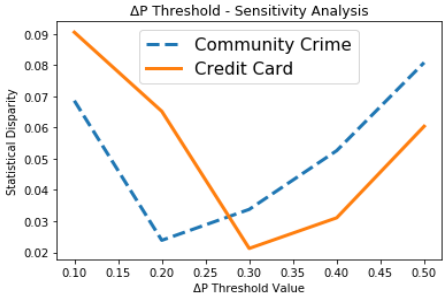

Fig. 6. Statistical Disparity versus Threshold r

References

[1] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "On fairness, diversity and randomness in algorithmic decision making," *CoRR*, 2017.

[2] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *KDD*, 2011.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and its's biased against blacks." in *ProPublica*, 2016.

[4] "Amazon reportedly killed an ai recruitment system because it couldn't stop the tool from discriminating against women," in *Fortune*, 2018.

[5] P. Lahoti, K. P. Gummadi, and G. Weikum, "ifair: Learning individually fair data representations for algorithmic decision making," in *ICDE*, 2019.

[6] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, 2012.

[7] R. Courtland, "Bias detectives: the researchers striving to make algorithms fair," *Nature*, 2018.

[8] L. Zhang and X. Wu, "Anti-discrimination learning: a causal modeling-based framework," *Int. Journal of Data Science and Analytics*, 2017.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*, 2015.

[10] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, 2013.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ACM Innovations in Theoretical Computer Science Conference*, 2012.

[12] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017.

[13] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

[14] I.-C. Yeh and C. hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, 2009.

[15] M. Redmond and A. Baveja, "A data-driven software tool for enabling cooperative information sharing among police departments," *European Journal of Operational Research*, 2002.

[16] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," *CoRR*, 2018.

[17] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," in *NIPS*, 2017.

[18] D. Biddle, *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000.

[20] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," in *NIPS*, 2018.

[21] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *ICDM Workshops*, 2011.