Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness?

Manh Khoi Duong(✉) and Stefan Conrad

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
{manh.khoi.duong,stefan.conrad}@hhu.de

Abstract. Fairness is a critical consideration in data analytics and knowledge discovery because biased data can perpetuate inequalities through further pipelines. In this paper, we propose a novel pre-processing method to address fairness issues in classification tasks by adding synthetic data points for more representativeness. Our approach utilizes a statistical model to generate new data points, which are evaluated for fairness using discrimination measures. These measures aim to quantify the disparities between demographic groups that may be induced by the bias in data. Our experimental results demonstrate that the proposed method effectively reduces bias for several machine learning classifiers without compromising prediction performance. Moreover, our method outperforms existing pre-processing methods on multiple datasets by Pareto-dominating them in terms of performance and fairness. Our findings suggest that our method can be a valuable tool for data analysts and knowledge discovery practitioners who seek to yield for fair, diverse, and representative data.

Keywords: fairness · bias · synthetic data · fairness-agnostic · machine learning · optimization

1 Introduction

Data analytics has grown in popularity due to its ability to automate decision-making through machine learning. However, real-world data can contain biases that produce unfair outcomes, making fairness in data pipelines involving machine learning a pressing concern. Fairness in machine learning typically deals with intervening algorithms providing equitable outcomes regardless of protected characteristics such as gender, race, or age group.

The existing related works can be divided into three categories [5,8,20]. The first category of methods are pre-processing methods, which aim to reduce bias in the data. Examples of such methods include data augmentation and data

This work was supported by the Federal Ministry of Education and Research (BMBF) under Grand No. 16DHB4020.

R. Wrembel et al. (Eds.): DaWaK 2023, LNCS 14148, pp. 176–190, 2023.
https://doi.org/10.1007/978-3-031-39831-5_17

balancing [2]. The second category of methods are in-processing methods, which aim to enforce fairness constraints during the training procedure [15]. Examples of in-processing methods include regularization techniques and constrained optimization [31]. The last category are post-processing methods that allow the improvement of fairness after training by correcting the outputs of the trained model [14].

The goal of this paper is to introduce a pre-processing method that achieves fairness by including generated data points. This is done by utilizing a statistical model that learns the distribution of the dataset, enabling the generation of synthetic samples. Additionally, a discrimination measure is employed to evaluate the fairness when incorporating the generated data points. Our method treats the discrimination measure as a black-box, making it able to optimize any discrimination measure defined by the user. We refer to this property of our algorithm as *fairness-agnostic*. This makes it suitable for cases where a specific fairness notion is required.

For the experimentation, multiple datasets known to be discriminatory were used. The experiments were performed by firstly loading the datasets and then pre-processing them using different pre-processing techniques. The pre-processed datasets were then fed into several classifiers. The performance of each classifier was then evaluated in terms of performance and fairness to assess the effectiveness of the pre-processing methods. Our experiments have empirically shown that our technique effectively lessens discrimination without sacrificing the classifiers' prediction qualities. Moreover, it is compatible with any machine learning model. Of the pre-processors tested, none were able to meet all of these conditions. The scope and application of our method is not necessarily limited to tabular data and classification tasks, even though experiments were conducted on them. The method is more broadly suitable for supervised learning tasks where the data, label, and protected attribute are available. Only the appropriate discrimination measures have to be derived for the right task. Generally, our primary contributions are:

- The introduction of a novel pre-processing technique that can optimize any given fairness metric by pre-selecting generated data points to include into the new fair dataset.
- We carry out a comprehensive empirical study, comparing our method against three widely recognized pre-processors [9,13,31], using multiple datasets commonly found in fairness literature.
- We present interesting and valuable properties, such as the empirical evidence that our method consistently improved fairness in comparison to the unprocessed data.

2 Related Work

Many pre-processing algorithms in literature alter the dataset to achieve fairness [4,9,31]. Because the methods simply return a fair dataset, they can be used with any estimator. However, such approaches cannot be used with ease: They often require a parameter setting that sets how aggressive the change should be. As the approaches differ in their methodology, it is hard to interpret the parameter's setting and their unexpected effects on the data. Data alteration methods also have a higher risk of producing data that do not resemble the original data distribution in any ways.

Other approaches return a weight for each sample in the dataset that the estimator should account for when fitting the data [1,13]. While the approaches seem promising [1,13], they require estimators to be able to handle sample weights. A way to account for this is to replicate samples based on their sample weights. However, this is not computationally scalable for larger datasets or for larger differences between the sample weights.

Another related approach is removing data samples that influence estimators in a discriminatory way [28]. Nevertheless, this approach does not seem feasible for smaller datasets.

Differently from related works, we present an algorithm that does not come with the above mentioned drawbacks. Further, our approach is able to satisfy any fairness notion that is defined for measuring discrimination or bias in the dataset. While the work of Agarwal et al. [1] also features this property, the fairness definitions must be formalizable by linear inequalities on conditional moments. In contrast, our work requires the fairness definitions to quantify discrimination in a numeric scale where lower values indicate less discrimination. This can be as simple as calculating the differences of probabilistic outcomes between groups.

While there exist works that train fair generative models to produce data that is fair towards the protected attribute on images [7,24,27] or tabular data [12,23], our approach can be seen as a framework that employs generative models and can therefore be used for any data where the protected attribute is accessible. Specifically, our research question is not *"How can fair generative models be constructed?"*, we instead deal with the question *"Using any statistical or generative model that learns the distribution of the dataset, how can the samples drawn from the distribution be selected and then included in the dataset such that fairness can be guaranteed?"*. Other works that generate data for fairness include generating counterfactuals [26] and generating pseudo-labels for unlabeled data [6].

3 Measuring Discrimination

In this section, we briefly present *discrimination measures* that assess the fairness of data. For that, we make use of following notation [5,8,20]: A *data point* or *sample* is represented as a triple (x, y, z), where $x \in X$ is the *feature*, $y \in Y$ is the ground truth *label* indicating favorable or unfavorable outcomes, and $z \in Z$ is the *protected attribute*, which is used to differentiate between groups. The sets X, Y, Z typically hold numeric values and are defined as $X = \mathbb{R}^d$, $Y = \{0, 1\}$, and $Z = \{1, 2, \ldots, k\}$ with $k \geq 2$. For simplicity, we consider the case where protected attributes are binary, i.e., $k = 2$. Following the preceding notation, a *dataset* is defined as the set of data points, i.e., $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{n}$. Machine learning models $\phi : X \times Z \rightarrow Y$ are trained using these datasets to predict the target variable $y \in Y$ based on the input variables $x \in X$ and $z \in Z$. We call the output $\hat{y} := \phi(x, z)$ *prediction.*

Based on the work of [32], we derive *discrimination measures* to the needs of the pre-processing method in this paper. To make our algorithm work, a *discrimination measure* must satisfy certain properties which we introduce in the following.

Definition 1. *A* discrimination measure *is a function $\psi : \mathbb{D} \rightarrow \mathbb{R}^{+}$, where $\mathbb{D}$ is the set of all datasets, satisfying the following axioms:*

1. *The discrimination measure $\psi(\cdot)$ is bounded by [0, 1]. (Normalization)*
2. *Minimal and maximal discrimination are captured with 0, 1 by $\psi(\cdot)$, respectively.*

The first and second axiom together assure that the minimal or maximal discrimination can be assessed by this measure. Furthermore, through normalization it is possible to evaluate the amount of bias present and its proximity to the optimal solution. As achieving no discrimination is not always possible, i.e., $\psi(\mathcal{D}) = 0$, we consider lower discrimination as better and define a fairer dataset as the one with the lower discrimination measure among two datasets.

Literature [2,5,8,19,20,32] on fairness-aware machine learning have classified fairness notions to either representing group or individual fairness. We subdivide the most relevant fairness notions into two categories which are *dataset* and *prediction notions* and derive discrimination measures from it as suggested by [32]. From now on, we denote x, y, z as random variables describing the events of observing an individual from a dataset $\mathcal{D}$ taking specific values.

Dataset notions typically demand the independency between two variables. When the protected attribute and the label of a dataset are independent, it is considered fair because it implies that the protected attribute does not influence or determine the label. An example to measure such dependency would be the *normalized mutual information* (NMI) [29] where independency can be concluded if and only if the score is zero. Because it is normalized as suggested by the name, it is a discrimination measure.

Definition 2 (Normalized mutual information). *Let $H(\cdot)$ be the entropy and $I(y; z)$ be the mutual information [25]. The normalized mutual information score is defined in the following [30]:*

$$\psi_{NMI}(\mathcal{D}) = 2\frac{I(y;z)}{H(y) + H(z)}.$$

Statistical parity [15,31] and *disparate impact* [9] are similar notions that also demand independency, except they are specifically designed for binary variables. Kang et al. [16] proved that zero mutual information is equivalent to statistical parity. To translate statistical parity to a discrimination measurement, we make use of differences similarly to Žliobaitė [32].

Definition 3 (Statistical parity). *Demanding that each group has the same probability of receiving the favorable outcome is statistical parity, i.e.,*

$$p(y = 1 \mid z = 1) = p(y = 1 \mid z = 0).$$

Because we want to minimize discrimination towards any group, we measure the absolute difference between the two groups to assess the extent to which the dataset fulfills statistical parity. This is also known as (absolute) statistical disparity (SDP) [8]. A value of 0 indicates minimal discrimination:

$$\psi_{SDP}(\mathcal{D}) = |p(y = 1 \mid z = 1) - p(y = 1 \mid z = 0)|. \tag{1}$$

Because disparate impact [9] essentially demands the same as statistical parity but contains a fraction, dividing by zero is a potential issue that may arise. Therefore, its use should be disregarded [32]. Note that dataset notions can also be applied to measure the fairness on predictions by exchanging the data label with the prediction label.

Parity-based notions, fulfilling the *separation* or *sufficiency* criterion [2], require both prediction and truth labels to evaluate the fairness. Contrary to the category before, measuring solely on datasets is not possible here. Despite this, it is still essential to evaluate on such measures to account for algorithmic bias. Here, the discrimination measure takes an additional argument, which is the prediction label $\hat{y}$ as a random variable. According fairness notions are, for example, *equality of opportunity* [10], *predictive parity* [2], and *equalized odds* [2].

Definition 4 (Equalized odds). *Equalized odds is defined over the satisfaction of both* equality of opportunity *and* predictive parity *[10],*

$$p(\hat{y} = 1 \mid y = i, z = 1) = p(\hat{y} = 1 \mid y = i, z = 0) \ \ \forall i \in \{0, 1\},$$

where equality of opportunity is the case of $i = 1$ and predictive parity is the case of $i = 0$, correspondingly. Making use of the absolute difference, likewise to SDP (1), we denote the measure of equality of opportunity as $\psi_{EO}(\mathcal{D}, \hat{y})$ and predictive parity as $\psi_{PP}(\mathcal{D}, \hat{y})$.

To turn equalized odds into a discrimination measure, we can calculate the average of the absolute differences for both equality of opportunity and predictive parity. This is referred to as *average odds error* [3]:

$$\psi_{ODDS}(\mathcal{D}, \hat{y}) = \frac{\psi_{EO}(\mathcal{D}, \hat{y}) + \psi_{PP}(\mathcal{D}, \hat{y})}{2}. \tag{2}$$

4 Problem Formulation

Intuitively, the goal is to add an amount of synthetic datapoints to the original data to yield for minimal discrimination. With the right discrimination measure chosen, it can be ensured that the unprivileged group gets more exposure and representation in receiving the favorable outcome. Still, the synthetic data should resemble the distribution of the original data. The problem can be stated formally in the following: Let $\mathcal{D}$ be a dataset with cardinality n, let $\tilde{n}$ be the number of samples to be added to $\mathcal{D}$. The goal is to find a set of data points $S = \{d_1, d_2, \ldots, d_{\tilde{n}}\}$ that can be added to the dataset, i.e., $\mathcal{D} \cup S$ with $\mathcal{S} \sim P(\mathcal{D})$, that minimizes the discrimination function $\psi(\mathcal{D} \cup S)$. Hence, we consider the following constrained problem:

$$\begin{aligned} \min \quad & \psi(\mathcal{D} \cup \mathcal{S}) \\ \text{subject to} \quad & \mathcal{S} \sim P(\mathcal{D}) \\ & |\mathcal{S}| = \tilde{n}. \end{aligned} \tag{3}$$

The objective (3) suggests that the samples d_i that are added to the dataset $\mathcal{D}$ are drawn from $P(\mathcal{D})$. To draw from $P(\mathcal{D})$, a statistical or generative model P_G that learns the data distribution can be used. Therefore generating data samples and bias mitigation are treated as sequential tasks where the former can be solved by methods from literature [22]. Because the discrimination measure ψ can be of any form, the optimization objective is treated as a black-box and is solved heuristically.

5 Methodology

Our algorithm relies on a statistical model, specifically the Gaussian copula [22], to learn the distribution of the given dataset $P(\mathcal{D})$. Gaussian copula captures the relationship between variables using Gaussian distributions. While assuming a Gaussian relationship, the individual distributions of the variables can be any continuous distribution, providing flexibility in modeling the data.

Still, the type of model for this task can be set by the user as long as it can sample from $P(\mathcal{D})$. Because discrimination functions are treated as black-boxes, the algorithm does not require the derivatives of ψ and optimizing for it leads to our desired *fairness-agnostic* property: It is suitable for any fairness notion that can be expressed as a discrimination function. Our method handles the size constraint in Eq. (3) as an upper bound constraint, where a maximum of $\tilde{n}$ samples are added to $\mathcal{D}$.

Our method, outlined in Algorithm 1, begins by initializing $\hat{\mathcal{D}}$ with the biased dataset $\mathcal{D}$. Then $\hat{n}$ is set as a multiplicative $r > 1$ of the original dataset's size. Lastly in the initialization, the distribution of $P(\mathcal{D})$ is learned by a generative model P_G. The algorithm then draws m samples from the generative model P_G which are referred to as the set of candidates C. The next step is decisive for the optimization (Line 9): The candidate which minimizes the discrimination most when included in the dataset $\hat{\mathcal{D}}$ is added to $\hat{\mathcal{D}}$. The steps of drawing samples and adding the best candidate to the dataset is repeated till $\hat{\mathcal{D}}$ has a cardinality of $\hat{n}$ or the discrimination is less than the fairness threshold ϵ. Because ϵ is set to 0 by default, the algorithm can stop earlier before the dataset reaches its requested size if the discrimination cannot be further reduced, i.e., $\psi(\hat{\mathcal{D}}) = 0$. Because calculating $\psi(\hat{\mathcal{D}} \cup \{c\})$ (Line 9) does not involve retraining any classifier and solely

Algorithm 1. Pseudocode of MetricOptGenerator

```
Input: $\mathcal{D}, r = 1.25, m = 5, \epsilon = 0$
Output: $\hat{\mathcal{D}}$
    Initialization:
    $\hat{\mathcal{D}} \leftarrow \mathcal{D}$
    $\hat{n} \leftarrow \lfloor r \cdot |\mathcal{D}| \rfloor$
    $P_G \leftarrow$ learn distribution of $P(\mathcal{D})$
    Generating fair samples:
    for $i = 1$ to $\hat{n} - |D|$ do
        if $(\psi(\hat{\mathcal{D}}) \leq \epsilon)$ then
            return $\hat{\mathcal{D}}$
        end if
        $C \leftarrow$ sample $m$ candidates from $P_G$
        $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{\text{argmin}_{c \in C}\, \psi(\hat{\mathcal{D}} \cup \{c\})\}$
    end for
    return $\hat{\mathcal{D}}$
```

evaluates the dataset, this step is practically very fast. In our implementation, we generate a set of synthetic data points prior to the for-loop, eliminating the sampling cost during the optimization step. We refer to Appendix A for the proof outlining the polynomial time complexity of the presented method.

6 Evaluation

To evaluate the effectiveness of the presented method against other pre-processors in ensuring fairness in the data used to train machine learning models, we aim to answer following research questions:

- **RQ1** What pre-processing approach can effectively improve fairness while maintaining classification accuracy, and how does it perform across different datasets?
- **RQ2** How stable are the performance and fairness results of classifiers trained on pre-processed datasets?
- **RQ3** How does pursuing for statistical parity, a data-based notion, affect a prediction-based notion such as average odds error?
- **RQ4** Is the presented method fairness-agnostic as stated?

To especially address the first three research questions, which deal with effectiveness and stability, we adopted the following experimental methodology: We examined our approach against three pre-processors on four real-world datasets (see Table 1). The pre-processors we compare against are *Reweighing* [13], *Learning Fair Representation* [31] (LFR), and *Disparate Impact Remover* [9] (DIR). The data were prepared such that categorical features are one-hot encoded and rows containing empty values are removed from the data. We selected sex, age, race, and foreign worker as protected attributes for the respective datasets. Generally, the data preparation was adopted from AIF360 [3].

Table 1. Overview of datasets.

Dataset	Protected Attribute	Label	Size	Description
Adult [17]	Sex	Income	45 222	Indicates individuals earning over $50 000 annually
Bank [21]	Age	Term Deposit	30 488	Subscription to a term deposit
COMPAS [18]	Race	Recidivism	6 167	Arrested again for a new offense within a period of 2 years after initial arrest
German [11]	Foreign Worker	Credit Risk	1 000	Creditworthiness of loan applicants

All hyperparameter settings of the pre-processsors were kept as they are, given the implementation provided by AIF360 [3]. For the case of LFR, we empirically had to lower the hyperparameter of optimizing for fairness. It was initially set too high which led to identical predictions for all data points. For our approach, we set $r = 1.25$ which returns a dataset consisting of additional 25% samples of the dataset's initial size. The discrimination measure chosen was the absolute difference of statistical parity (1), which all other methods also optimize for. Further, we set $m = 5$ and $\epsilon = 0$ as shown in Algorithm 1.

The experimental methodology for a single dataset is visualized in Fig. 1 as a pipeline. The given dataset is firstly split into a training (80%) and test set (20%). Afterwards, the training set is then passed into the available pre-processors. Then, all debiased data are used to train several classifiers. We employed three different machine learning algorithms—*k-nearest neighbors* (KNN), *logistic regression* (LR), and *decision tree classifier* (DT)—to analyze the pre-processed datasets and the original, unprocessed dataset for comparison. The unprocessed dataset is referred to as the baseline. Finally, the performance and fairness is evaluated on the prediction of the test set. It is noteworthy to mention that the test sets were left untouched to demonstrate that by pre-processing the training data, unbiased results can be achieved in the prediction space even without performing bias mitigation in the test data. Due to stability reasons (and to handle **RQ2**), we used Monte Carlo cross-validation to shuffle and split the dataset. This was done 10 times for all datasets. The results from it set the performance-fairness baseline. While our optimization focuses on SDP, we address **RQ3** by assessing the error of average odds. To answer **RQ4**, we refer to Sect. 6.2 for the experimentation and discussion.

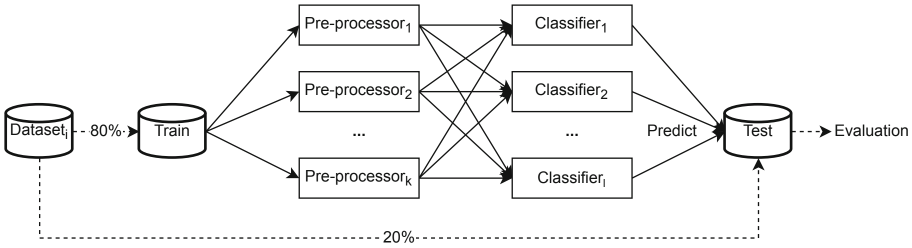


Fig. 1. Experimental methodology visualized as a pipeline

6.1 Comparing Pre-processors

Table 2 presents the performance-fairness test results of pre-processors on different datasets (**RQ1**). For the discrimination, the table displays SDP and average odds error of the predictions on the test sets. To assess the classifier's performances, we used *area under the receiver operating characteristic curve* (AUC). An estimator that guesses classes randomly would produce an AUC score of 0.5. Here, higher scores imply better prediction performances. Means and standard deviations of the Monte Carlo cross-validation results are also displayed to evaluate the robustness (**RQ2**). We note that all classifiers except of KNN were able to handle sample weights in training, which are required for Reweighing. Therefore Reweighing was not able to mitigate bias in KNN and performed as well as the baseline in contrast to other approaches including ours.

Because all pre-processors aim to reduce statistical disparity (or the equivalent formulation), we compare the SDP scores between the pre-processors: In most cases, our approach produced Pareto optimal solutions with respect to both SDP and AUC. Generally, only Reweighing and our approach appear to consistently improve fairness without sacrificing notable prediction power. In direct comparison, LFR improved the fairness at most across all experiments but at the same time sacrifices prediction quality of all classifiers to such a great extent that the predictions become essentially useless. In experiments where LFR attained standard deviations of 0 across all scores (Table 2b, 2d), we investigated the pre-processed data and found that LFR had modified almost all labels to a single value. As a result, the estimators were unable to classify the data effectively, as they predicted only one outcome. The results of DIR are very inconsistent. DIR sometimes even worsens the fairness, as seen in the COMPAS and German datasets, where SDP and average odds error are increased in most settings. This situation arises when there is an excessive correction of the available discrimination for the unprivileged group, leading to discrimination against the privileged group. If the discrimination measures are defined such that the privileged or unprivileged groups do not matter (similarly to this paper), reverse discrimination would not mistakenly occur by our approach. This extra property renders our method more suitable for responsible use cases.

When comparing the average odds error rates (**RQ3**), our approach has successfully reduced algorithmic bias without aiming for it under nearly all experiments. The increase in the average odds error rate (mean), albeit negligible, was observed only when training DT on the Banking data and LR in the German dataset. In all other ten model and dataset configurations, our approach did reduce the error rate without particularly optimizing for it. This can be expected in practice as the independency of the label with the protected attribute (SDP) is a sufficient condition for average odds.

Table 2. The tables displays each classifier's mean test performance and discrimination when trained on different pre-processed training sets. The best performing statistic for each classifier is marked in bold. Minimal standard deviations are marked bold, too. All values displayed are percentages.

(a) Adult

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	81.28	0.53	19.65	1.04	24.50	1.27
	LFR	50.18	1.20	**0.18**	**0.46**	**0.16**	**0.37**
	Our	78.91	0.52	9.68	1.19	10.55	1.26
	Original	**81.35**	0.52	19.77	0.96	24.66	1.21
	Reweighing	78.95	**0.48**	4.96	1.22	1.37	0.76
KNN	DIR	75.35	0.86	20.94	2.35	22.31	3.32
	LFR	51.80	4.96	**1.14**	3.26	**0.70**	**2.02**
	Our	75.26	**0.60**	18.84	2.91	19.80	3.54
	Original	**75.53**	0.85	21.09	**2.16**	22.33	3.05
	Reweighing	**75.53**	0.85	21.09	**2.16**	22.33	3.05
LR	DIR	80.12	0.59	17.84	0.46	22.80	0.49
	LFR	55.35	8.64	**1.33**	3.05	**0.80**	2.02
	Our	76.96	0.52	3.60	0.82	1.30	0.63
	Original	**80.13**	0.59	17.75	**0.45**	22.71	**0.48**
	Reweighing	77.29	**0.51**	4.63	0.57	1.90	0.71

(b) Bank

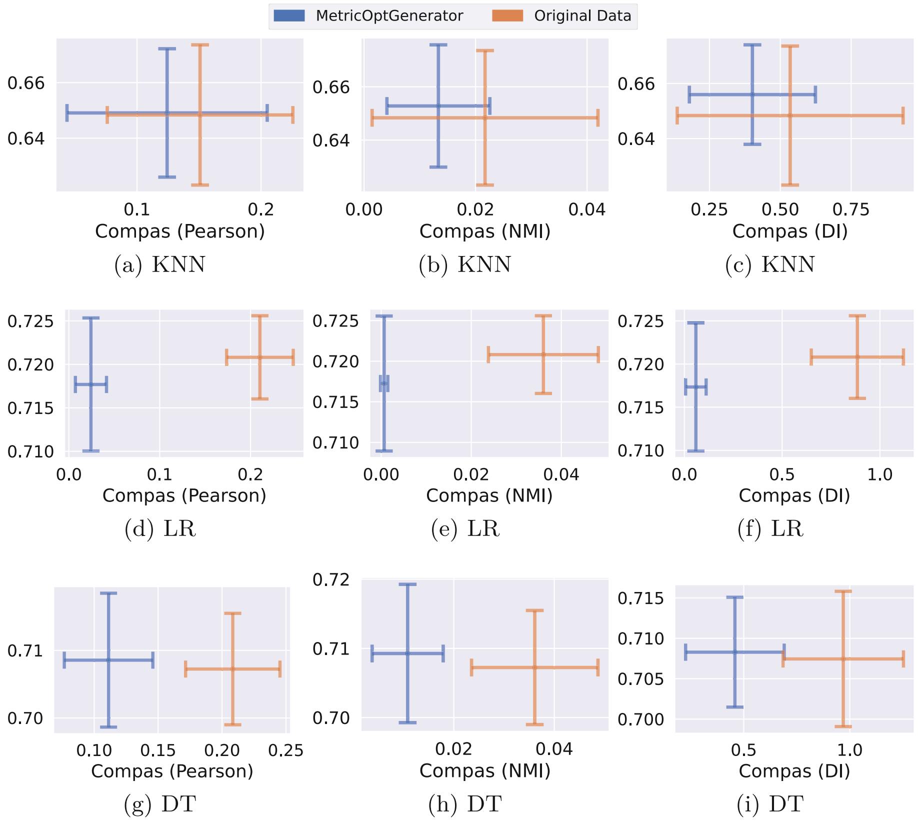
		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	67.66	1.24	3.30	2.03	7.81	3.66
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	72.68	0.96	10.36	3.75	6.93	3.42
	Original	**72.94**	1.15	10.69	1.70	6.49	2.97
	Reweighing	72.81	1.16	9.58	1.99	5.93	3.07
KNN	DIR	81.42	0.82	8.43	3.69	6.56	3.16
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	**86.98**	0.66	9.05	3.55	5.00	3.74
	Original	**86.98**	0.65	9.05	3.55	5.00	3.75
	Reweighing	**86.98**	0.65	9.05	3.55	5.00	3.75
LR	DIR	91.48	0.42	3.90	1.08	3.60	2.59
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	91.25	0.51	6.72	2.24	2.87	2.00
	Original	**92.14**	0.34	7.00	2.85	4.37	2.57
	Reweighing	92.11	0.36	5.82	2.47	3.37	1.81

(c) COMPAS

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	70.75	**0.53**	23.58	4.70	22.06	4.71
	LFR	50.26	3.97	**8.32**	21.12	**8.05**	21.17
	Our	**70.91**	0.85	10.55	4.31	8.63	4.06
	Original	70.76	0.82	21.16	**3.64**	19.67	3.77
	Reweighing	70.35	0.97	10.22	4.02	8.98	**2.80**
KNN	DIR	**65.78**	2.88	21.74	7.64	20.58	7.03
	LFR	53.58	6.15	**2.29**	**3.67**	**3.00**	**4.20**
	Our	65.13	**1.49**	12.56	7.98	11.94	7.48
	Original	64.84	2.52	15.62	7.88	14.55	8.16
	Reweighing	64.84	2.52	15.62	7.88	14.55	8.16
LR	DIR	**72.28**	0.54	23.20	3.41	21.31	3.77
	LFR	56.94	9.07	**1.95**	3.44	**2.55**	3.57
	Our	71.78	0.68	2.31	**1.51**	5.38	**1.79**
	Original	72.08	**0.48**	21.74	3.76	20.01	4.06
	Reweighing	71.52	0.76	3.89	2.46	5.61	2.17

(d) German

		AUC		SDP		AVG Odds Error	
		mean	std	mean	std	mean	std
Model	Preprocessor						
DT	DIR	61.15	4.02	23.22	13.70	28.23	14.30
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	**63.08**	3.47	16.61	10.90	26.78	11.09
	Original	62.76	3.95	15.07	11.29	27.12	10.83
	Reweighing	62.71	4.65	17.53	15.03	33.15	6.73
KNN	DIR	**55.42**	4.44	18.70	9.28	23.91	7.47
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	54.21	4.34	12.95	2.95	13.81	3.26
	Original	54.08	3.86	16.99	4.63	18.27	4.95
	Reweighing	54.08	3.86	16.99	4.63	18.27	4.95
LR	DIR	78.05	2.15	20.05	11.84	29.79	14.16
	LFR	50.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Our	77.60	1.73	15.49	10.46	31.00	9.50
	Original	**78.10**	1.67	16.79	11.10	29.83	10.54
	Reweighing	78.05	1.88	16.42	11.10	30.32	10.56



(a) KNN (b) KNN (c) KNN

(d) LR (e) LR (f) LR

(g) DT (h) DT (i) DT

Fig. 2. Results of optimizing different discrimination objectives with our method on the COMPAS dataset. Objectives are ordered by columns, classifiers by rows. The y-axis displays AUC.

6.2 Investigating the Fairness-Agnostic Property

To demonstrate the fairness-agnostic property of our algorithm (**RQ4**), we evaluated our method against the baseline dataset on multiple measures and examine whether the objective was improved (see Fig. 2). The COMPAS dataset was used for this experiment. The chosen objectives are: the absolute value of Pearson's ρ, NMI (2), and the objective of disparate impact (DI) as given by [9]. All other experimental settings remained the same as described prior, except that other pre-processing methods were not used.

It can be observed that all discrimination measures were lowered significantly. Generally, our method was able to optimize on any fairness notion, as evidenced here and Sect. 6.1. It was even able to outperform algorithms that were specifically designed for a single metric, demonstrating its adaptability.

7 Conclusion

Machine learning can be utilized for malicious purposes if estimators are trained on data that is biased against certain demographic groups. This can have an incredibly negative impact on the decisions made and the groups that are being discriminated against.

The presented pre-processing method in this work is a sampling-based optimization algorithm that firstly uses a statistical model to learn the distribution of the given dataset, then samples points from this distribution, and determines which one to add to the data to minimize the discrimination. This process continues until the predefined criteria set by the user are satisfied. The method can optimize any discrimination measure as it is treated as a black-box, making it more accessible for wider use cases.

The results of our experiments demonstrate that our technique is reliable and significantly reduces discrimination while not compromising accuracy. Although a few other methods performed similarly in a few experiments, they were not compatible with certain estimators or even added bias to the original data. Because fairness was improved among the experiments and our method adds samples, it indicates that representativeness can be achieved with our method. Our research underscores the importance of addressing bias in data and we hope to contribute such concerns in data analytics and knowledge discovery applications.

8 Discussion and Future Work

The results of our approach demonstrate that it is possible to achieve fairness in machine learning models using generated data points. Despite our approach showing promise, it is important to acknowledge that our results rely heavily on the quality of the statistical model used to generate synthetic data. For tabular data, Gaussian copula [22] seems to be a good choice.

In future work, we aim to explore the potential of our method in making pre-trained models fairer with our method. While retraining large models using debiased datasets may not always be feasible from a cost-effective perspective, our approach allows using generated data to fine-tune the model for fairness, which provides a more efficient alternative.

Additionally, our evaluation deals with datasets where the protected attribute is a binary variable, which leaves some use cases untreated. Neglecting to recognize non-binary groups can lead to overlooking those who are most in need of attention. Similarly, research on dealing with multiple protected attributes at the same time could be done. This is to make sure that no protected group is being disadvantaged. Previous studies have touched on this subject [1,4,32], but we hope to reformulate these issues as objectives that work with our approach.

A Proof of Time Complexity

Theorem 1 (Time complexity). *If the number of candidates m and fraction r are fixed and calculating the discrimination $\psi(\mathcal{D})$ of any dataset $\mathcal{D}$ takes a linear amount of time, i.e., $\mathcal{O}(n)$, Algorithm 1 has a worst-case time complexity of $\mathcal{O}(n^2)$ where n is the dataset's size when neglecting learning the data distribution.*

Proof. In this proof, we will focus on analyzing the runtime complexity of the for-loop within our algorithm as the steps before such as learning the data distribution depends heavily on the used method. The final runtime of the complete algorithm is simply the sum of the runtime complexities of the for-loop that is focus of this analysis and the step of learning the data distribution.

Our algorithm firstly checks whether the discrimination of the dataset $\hat{\mathcal{D}}$ is already fair. The dataset grows at each iteration and runs for $\lfloor rn \rfloor - n = \lfloor n(r-1) \rfloor$ times. For simplicity, we use $n(r-1)$ and yield,

$$\begin{aligned}\sum_{i=0}^{n(r-1)-1} n+i &= \sum_{i=1}^{n(r-1)} n+i+1 \\ &= \sum_{i=1}^{n(r-1)} n + \sum_{i=1}^{n(r-1)} i + \sum_{i=1}^{n(r-1)} 1 \\ &= n^2(r-1) + \frac{(n(r-1))^2 + (n(r-1)+1)}{2} + n(r-1) \in \mathcal{O}(n^2),\end{aligned}$$

making the first decisive step for the runtime quadratic.

The second step that affects the runtime is returning the dataset that minimizes the discrimination where each of the m candidates $c \in C$ is merged with the dataset, i.e., $\psi(\hat{\mathcal{D}} \cup \{c\})$. The worst-case time complexity of it can be expressed by

$$\begin{aligned}\sum_{i=1}^{n(r-1)} m(n+i) = m \cdot \sum_{i=1}^{n(r-1)} n+i &= m \cdot \left(\sum_{i=1}^{n(r-1)} n + \sum_{i=1}^{n(r-1)} i \right) \\ &= m \cdot \left(n^2(r-1) + \frac{(n(r-1))^2 + (n(r-1))}{2} \right) \in \mathcal{O}(n^2),\end{aligned}$$

which is also quadratic. Summing both time complexities makes the overall complexity quadratic. □

Although the theoretical time complexity of our algorithm is quadratic, measuring the discrimination, which is a crucial part of the algorithm, is very fast and can be assumed to be constant for smaller datasets. Conclusively, the complexity behaves nearly linearly in practice.

In our experimentation, measuring the discrimination of the Adult dataset [17], which consists of 45 222 samples, did not pose a bottleneck for our algorithm.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69. PMLR (2018)
2. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019). http://www.fairmlbook.org
3. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. CoRR arxiv:1810.01943 (2018)
4. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf
5. Caton, S., Haas, C.: Fairness in machine learning: a survey. arXiv preprint arXiv:2010.04053 (2020)
6. Chakraborty, J., Majumder, S., Tu, H.: Fair-SSL: building fair ML software with less data. arXiv preprint arXiv:2111.02038 (2022)
7. Choi, K., Grover, A., Singh, T., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1887–1898. PMLR, 13–18 July 2020
8. Dunkelau, J., Leuschel, M.: Fairness-aware machine learning (2019)
9. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29 (2016)
11. Hofmann, H.: German credit data (1994). https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
12. Jang, T., Zheng, F., Wang, X.: Constructing a fair classifier with generated fair data. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 7908–7916 (2021)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
14. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929. IEEE (2012)
15. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 35–50. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_3
16. Kang, J., Xie, T., Wu, X., Maciejewski, R., Tong, H.: InfoFair: information-theoretic intersectional fairness. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 1455–1464. IEEE (2022)
17. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: KDD 1996, pp. 202–207. AAAI Press (1996)

18. Larson, J., Angwin, J., Mattu, S., Kirchner, L.: Machine bias, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
19. Makhlouf, K., Zhioua, S., Palamidessi, C.: Machine learning fairness notions: bridging the gap with real-world applications. Inf. Process. Manage. **58**(5), 102642 (2021). https://doi.org/10.1016/j.ipm.2021.102642. https://www.sciencedirect.com/science/article/pii/S0306457321001321
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
21. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst. **62**, 22–31 (2014)
22. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410, October 2016. https://doi.org/10.1109/DSAA.2016.49
23. Rajabi, A., Garibay, O.O.: TabfairGAN: fair tabular data generation with generative adversarial networks. Mach. Learn. Knowl. Extr. **4**(2), 488–501 (2022)
24. Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J. Res. Dev. **63**(4/5), 1–3 (2019)
25. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
26. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: a common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 166–172 (2020)
27. Tan, S., Shen, Y., Zhou, B.: Improving the fairness of deep generative models without retraining. arXiv preprint arXiv:2012.04842 (2020)
28. Verma, S., Ernst, M.D., Just, R.: Removing biased data to improve fairness and accuracy. CoRR arXiv:2102.03054 (2021)
29. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1073–1080 (2009)
30. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., DATA, M.: Practical machine learning tools and techniques. In: Data Mining, vol. 2 (2005)
31. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333. PMLR (2013)
32. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Min. Knowl. Disc. **31**, 1060–1089 (2017)