A framework for benchmarking discrimination-aware models in machine learning

Rodrigo L. Cardoso,[1] Wagner Meira Jr,[1] Virgilio Almeida,[1] Mohammed J. Zaki[2]
[1]Federal University of Minas Gerais
[2]Rensselaer Polytechnic Institute
{rodrigolc,meira,virgilio}@dcc.ufmg.br, zaki@cs.rpi.edu

Abstract

Discrimination-aware models in machine learning are a recent topic of study that aim at minimizing the adverse impact of machine learning decisions for certain groups of people due to ethical and legal implications. We propose a benchmark framework for assessing discrimination-aware models. Our framework consists of systematically generated biased datasets that are similar to real world data, created by a Bayesian network approach. Experimental results show that we can assess the quality of techniques through known metrics of discrimination, and our framework is flexible and can be extended to most real datasets and fairness measures to support a diversity of assessments.

1 Introduction

Discrimination-aware learning is a topic of research that aims at minimizing the impact of bias against certain groups due to ethical reasons and legal implications. Given a labeled dataset whose records represent individuals, let its attributes be divided into non-protected and protected attributes, such as race and gender. The problem consists of building a classifier that takes the non-protected attributes of an individual and maps them to a class label, so that the classifier has minimum discrimination and maximum accuracy.

There are several recent works that tackle the issue of discrimination-aware learning; see for example the survey by (Hajian, Bonchi, and Castillo 2016) on fairness in machine learning. However, it is interesting to note that there is no consensus on the best technique for a given application scenario, since it also depends on the level of discrimination that is inherent to the dataset being targeted. The recent study by (Kleinberg, Mullainathan, and Raghavan 2016) discusses the impossibility of a classifier to satisfy multiple notions of fairness. This is one of the reasons why researchers have put their efforts on minimizing the effect of biased data on the predictions. We argue that an effective strategy to assess discrimination-aware learning models is by running them on scenarios where the discrimination-related parameters differ, so that we can observe how well the models behave. The main contribution of this work is a framework for comparing discrimination-aware learning models. To the best of our knowledge, there is no other such benchmark framework in the literature. Our framework comprises systematically generated biased datasets that are sampled from Bayesian networks learned from real world data. Our main concern is to explore alternative discrimination scenarios, that is, we want to learn from data representing different levels of bias for the purpose of analyzing the behavior of techniques when bias levels differ. This work focuses on two metrics for assessing discrimination: disparate impact and disparate mistreatment (Zafar et al. 2017). However, our approach may be extended to other metrics (see (Zliobaite 2017)) as well.

2 Related work

Most of the studies in this area may be divided into two groups (Hajian, Bonchi, and Castillo 2016): *discrimination discovery*, which focuses on studying metrics and identifying how much discrimination there is in a dataset, and *discrimination prevention*, focused on building classification models that are less likely to produce discriminatory results.

For *discrimination discovery* there are a number of works that aim at identifying patterns of discrimination in data (Pedreschi, Ruggieri, and Turini 2008; Ruggieri, Pedreschi, and Turini 2010). Some of these works propose and study metrics to quantify the amount of discrimination. Zliobaite (Zliobaite 2017) surveys several such metrics; she defines the *mean difference* as the most commonly used metric in early works, also called *slifd* (Pedreschi, Ruggieri, and Turini 2009) or *disparate impact*. Zafar et al. (Zafar et al. 2017) propose a metric called *disparate mistreatment*, which measures the difference between misclassifications.

Discrimination prevention methods can be divided into three groups: *pre-processing*, *in-processing* and *post-processing* techniques (Hajian, Bonchi, and Castillo 2016). We focus more on *Pre-processing techniques* (Kamiran and Calders 2011; Feldman et al. 2015), since they modify the training set in order to make it as discrimination-free as possible, so that a classifier becomes less prone to exhibit bias. *In-processing techniques* (Calders and Verwer 2010; Kamiran, Calders, and Pechenizkiy 2010; Zafar et al. 2017) work by changing the classifier to produce less discriminating models, whereas *Post-processing techniques* (Hajian et al. 2014) change the outcome of classification models.

Regarding discrimination-related assessment, several pre-

vious works on discrimination-aware learning generate synthetic data for evaluation (Hardt, Price, and Srebro 2016; Zliobaite, Kamiran, and Calders 2011). However, they create simple models with a few variables and their conditional dependencies, containing protected and non-protected artificial attributes and a binary class. In contrast, the main difference in our approach is that we generate biased data sampled from Bayesian networks learned from real world data, instead of purely synthetic data.

There are works that employ Bayesian networks to deal with discrimination discovery and prevention task. Bonchi et al. (2017) address the problem of learning probabilistic causal structures of discrimination from datasets. Mancuhan et al. (2014) use Bayesian networks for the task of bias prevention. Our work learns these structures, identifies discrimination patterns and uses them to sample data at different levels of bias. While previous approaches focused on discovering and mitigating discrimination, we employ sampled data from Bayesian networks in order to evaluate the quality of discrimination prevention techniques.

3 Definitions

A discrimination-aware learning model usually exploits the trade-off between accuracy and discrimination. The task is usually framed as a binary classification problem, where not only the accuracy should be maximized, but also discrimination has to be minimized. The classification score is defined according to the probabilities of outcomes for different groups of individuals. These groups are sets of individuals who share the same value of a specific attribute and, by contrasting their outcomes, we assess the model discrimination.

Formally, we are given a labeled dataset D where each record represents an individual. For each individual there are two sets of attributes $X = \{x_1, ..., x_n\}$ and $S = \{s_1, ..., s_m\}$. We call the set X the legally usable attributes, or *non-protected attributes*, which, in theory, may be used in decision making without any legal implications, e.g., annual income. We call the set S the non-legally usable attributes, or *protected attributes*. We are not supposed to use any attribute from S in a decision making process because they are protected by law, for instance, race, sex.

The class label $y \in \{+, -\}$ is the variable the model tries to predict for each individual. A positive class label $y = +$ (also denoted y^+) expresses a favorable outcome for the individual; alternatively, a negative class label $y = -$ (also denoted y^-) expresses an unfavorable outcome. An example is in credit scoring, where a bank wants to decide whether a customer has good credit score or not. A good credit score is represented by $y = +$, while a bad credit score is represented by $y = -$.

In order to measure discrimination with respect to a protected attribute s for individuals that belong to a group g, we quantify the difference between probabilities of positive outcomes for individuals that belong to g, that is, $s = g$, and those who don't belong to g, that is, $s \neq g$ or $s = \bar{g}$. We assume here that individuals belonging to g usually suffer from discriminatory conditions, and are called the **deprived group**. Individuals not belonging to g usually have an unfair advantage in models, and they are called **favored group**.

Definition 3.1. Let D be a dataset with labeled binary classes $y \in \{+, -\}$, a protected attribute s, an attribute value g for attribute s that defines individuals belonging to a group and $\bar{g}$ that defines individuals not belonging to the same group. The discrimination $disc_{D,s,g}$ in D with respect to the attribute s for individuals from g is defined by the following equation:

$$disc_{D,s,g} = P(y^+|s = \bar{g}) - P(y^+|s = g) \quad (1)$$

That is, the difference between the probability of positive class for the favored ($\bar{g}$) versus the disfavored group (g). For example, members of the favored group may have a higher probability of a good credit score. Definition 3.1 is widely used for measuring discrimination (Kamiran and Calders 2011; Hajian et al. 2014) and it is often referred to as **disparate impact**.

Another way of measuring discrimination is through the two metrics presented in Zafar et al. (Zafar et al. 2017) as **disparate mistreatment**. These are the metrics for misclassification that take into account the difference between false-positive and false-negative rates for individuals of different groups, defined as follows:

Definition 3.2. Let a classifier f assign for every record in D a class label $\hat{y}$. The disparate mistreatment of the classifier f in the dataset D with respect to the attribute s between the groups g and $\bar{g}$ is defined as:

$$D_{fp(D,s,g,f)} = P(\hat{y}^+|s = \bar{g}, y^-) - P(\hat{y}^+|s = g, y^-) \quad (2)$$

$$D_{fn(D,s,g,f)} = P(\hat{y}^-|s = g, y^+) - P(\hat{y}^-|s = \bar{g}, y^+) \quad (3)$$

The favored group $\bar{g}$ may have a higher false-positive rate (called **overestimation**) and a lower false negative rate (called **underestimation**) compared to the disfavored group g. These metrics seek to measure the extent of this problem for a given classifier f.

4 Benchmark Methodology

Our benchmark framework comprises systematically generated biased datasets that are derived from Bayesian networks learned from real-world data. We learn an approximate network structure that describes a dataset, which in turn depends on the conditional probabilities between the attributes. Modified Bayesian networks with different degrees of bias are used to generate new datasets that are used for evaluating discrimination-aware models.

4.1 Estimating Bayesian Networks

A Bayesian network is a probabilistic graphical model that maps conditional dependencies of random variables into a directed acyclic graph. We use an estimated Bayesian network to generate synthetic data, which allow us to quantify how much influence a protected attribute has on the outcome of a classification model.

To learn the structure of a Bayesian network in order to study causal relationships between the variables, we use the popular `R` library `bnlearn` (Margaritis 2003). Knowing the

Bayesian network that represents a given data allows us to reproduce the characteristics inherent to the original data and to also adjust specific parameters to generate diverse scenarios. By learning these structures we can modify any node's conditional probabilities and are thus able to calibrate the bias of an outcome with respect to a protected attribute. Currently, we imply two methods for learning Bayesian Networks, Hill-Climbing greedy search and Tabu Search algorithm, both described in (Margaritis 2003).

4.2 Modifying probabilities

Once we have learned a Bayesian network that represents the data, we can change the conditional probabilities of specific nodes. By doing this, we create a scenario where some of the nodes have different degrees of influence on other nodes. The modification is performed by selecting a specific attribute of our interest, that is, a protected attribute s. This attribute is represented by a node and it will have a direct or indirect influence on the outcome. The influence is observed in the conditional probability table of the outcome node.

Suppose we have a target attribute $s \in \{\bar{g}, g\}$ and the outcome $y \in \{-, +\}$. Recall that g represents a deprived group and $\bar{g}$ represents a favored group. We modify the Bayesian network by changing values in the probability table. Let $0 \leq \beta \leq 1$ be the level of artificial bias we want to insert on this node. The new probability $P'(y = +|s = g)$ is defined as:

$$P'(y = +|s = g) = P(y = +|s = g)(1 - \beta) \qquad (4)$$

We only change the probabilities for the deprived group ($s = g$) because, if we change the conditional probabilities of the other group, we are inserting twice the amount of bias. Our methodology consists of generating n Bayesian networks for each dataset. Each Bayesian network is generated with an increasing β (from 0.0 to 1.0).

Table 1: Toy conditional probability table

	$s = \bar{g}$	$s = g$
$P(y = -\|s)$	0.63	$0.80 \rightarrow 0.90$
$P(y = +\|s)$	0.37	$0.20 \rightarrow 0.10$

Table 1 represents a toy example of conditional probabilities of the outcome given a variable s. Using the disparate impact score we can see that the discrimination against the group g is 0.17 since $P(y^+|\bar{g}) - P(y^+|g) = 0.37 - 0.20 = 0.17$. Suppose we want to insert a bias level β of 0.5. In order to do this, we change the probability under the column $s = g$, as shown in Table 1. Now we can observe that the discrimination against the group g is 0.27 $(0.37 - 0.10)$, which means that we have increased the resulting discrimination.

4.3 Sampling & Evaluation

After learning a Bayesian network with the conditional probabilities on its nodes, we can then sample data from it. Since we have the frequency of each attribute in the dataset, we can sample this structure, so that samples remain similar to the real data regarding the probability of each individual attribute. A sample consists of randomly generated observations, where each the attribute values in an observation are generated according the probability table learned from the original data. For each bias level introduced, we generate random samples that form the corresponding biased training dataset. These systematically biased datasets are used for evaluating discrimination models and metrics. Those techniques either pre-process the input data and generate a new data or use the input data without any pre-processing and provide a discrimination-aware classification model.

5 Experiments

We compare well-known discrimination-aware techniques by testing them on the systematically biased datasets. We mainly test pre-processing techniques and some in-processing techniques. When a pre-processing technique is employed, it outputs a modified dataset that is used by conventional classification models. The test set is the original real data that was used to generate the Bayesian networks. Evaluation is performed in two different ways. First we use the most common way to evaluate a discrimination-aware model, that is, by comparing the discrimination of the prediction to the accuracy on the test set. The second way consists of measuring overestimation (eq. (2)) and underestimation (eq. (3)) of the resulting predictions.

5.1 Discrimination-aware Techniques

The techniques that we compared are mainly pre-processing ones that aim to produce datasets that are supposed to generate less biased classifiers.

Baseline: It consists of removing the protected attribute from the training set. It has been argued that such removal may even increase discrimination (Pedreschi, Ruggieri, and Turini 2008). This is due to the fact that some of the attributes may describe the protected one, for example, the neighborhood information may carry racial information about individuals. This problem is known as *redlining*.

Calders et al. (Calders and Verwer 2010): propose several pre-processing approaches for dealing with the problem of discrimination aware learning. These are:

- *Massaging* changes the class label of individuals in order to balance positive outcomes between groups. Individuals of the deprived group from the negative class are reassigned to the positive class, and individuals of the favored group having a positive class are reassigned to the negative class. Instances are selected for class reassignment based on a score learned by a ranker.
- *Re-weighting* assigns higher weights to individuals of the deprived group that have a positive class label and to individuals of the favored group that have a negative class label.
- *Uniform Sampling* applies the following rule on a randomly chosen instance: if the instance is from the deprived group with negative class, it is removed, otherwise, if it is from the positive class, it is duplicated. Likewise, if

the instance is of the favored group with a positive class is removed, but if it has a negative class it is duplicated.

- *Preferential Sampling* chooses instances based on a ranker like in the Massaging technique. The change rules are the same as in uniform sampling.

Black Box Auditing (Auditor) Black Box Auditing[1] is an implementation of Gradient Feature Auditing (GFA) introduced in (Adler et al. 2016). This technique works by repairing the dataset via a pre-processing technique, which means that it changes attribute labels. The resulting repaired dataset is expected to have lower discrimination. We run the data repairer described by (Feldman et al. 2015) at the repair levels of 0.25, 0.50 and 0.75.

5.2 Datasets

We test on the following original real-world datasets. The **Adult dataset**[2], also known as Census Income, is a widely used dataset in previous discrimination-aware learning studies. The task is to predict whether an individual has a yearly income greater than $50K or not (i.e., high vs. low income). It has 48,842 instances with 14 attributes. The protected attribute is sex and the original dataset has an inherent disparate impact against women (equal to 0.19). For generating biased data, we increase the influence of the feature *relationship* on the outcome, thus making it less likely that wives have high income, thus increasing the bias against women. The **Pro Publica COMPAS dataset**[3] records racial bias on recidivism scores. The data contains information from defendants such as race, age, criminal history and whether the defendant had committed a crime within a two-year window. It has 6,150 instances with 13 attributes. The sensitive attribute is `race`, which can be either "Caucasian" or "African-American". We modified the influence of the variable race on the outcome for generating biased data. **Dutch census** (Calders and Verwer 2010) is also a demographic census. We use it to make predictions of whether an individual has a "high level" occupation or not. This dataset has 11 attributes and we define sex as the protected attribute.

5.3 Experimental setup

The experiments are conducted by evaluating the various discrimination-aware techniques on our set of systematically biased datasets generated from the real-world data mentioned above. This set consists of 4 training datasets with increasing levels of artificial bias (β) against individuals of the defined deprived group. For the classifier, we use the Weka implementation of the C4.5 decision tree (Hall et al. 2009). We reproduced the experimental setup described by the authors of the techniques and used default parameters (Lichman 2013; Zliobaite, Kamiran, and Calders 2011). Thereafter, we measure the accuracy and discrimination observed when the original data is used as the testing data. Note that we refer to disparate impact as *discrimination* (eq. (1)), and measure disparate mistreatment via *overestimation* and *underestimation* (eq. (2) and eq. (3)).

[1]https://github.com/algofairness/BlackBoxAuditing

[2]https://archive.ics.uci.edu/ml/datasets/adult

[3]https://github.com/propublica/compas-analysis

6 Results and discussion

In this section we present the results of employing our framework to assess several techniques. Every experiment was run 30 times. We computed the confidence intervals as well as the variances for each set of experiments and most of them ranged between 90% and 95%.

6.1 Discrimination vs. Accuracy

Table 2 shows the Discrimination versus Accuracy on each dataset when we only remove the protected attribute. We perform this experiment for a couple of reasons. The first one is to make sure that removing the protected attribute does not reduce the resulting discrimination in the classification task. The second reason is to define upper bounds for both discrimination and accuracy. That is, the discrimination without any technique being employed must be higher than when some technique is used, otherwise the use of this anti-discrimination technique would be of no use. It is also expected that the accuracy will be higher than the resulting accuracy when a discrimination-aware technique is used because the technique must lower the discrimination at the cost of accuracy.

Table 2: Discrimination and Accuracy on biased datasets from Adult, COMPAS and Dutch data (classifier does not use the protected attribute)

β \ Dataset	Adult		COMPAS		Dutch	
	Disc.	Acc.	Disc.	Acc.	Disc.	Acc.
0.00	0.162	0.845	0.121	0.878	0.322	0.832
0.25	0.280	0.829	0.121	0.880	0.358	0.831
0.50	0.581	0.729	0.123	0.879	0.393	0.827
0.75	0.592	0.726	0.120	0.878	0.479	0.805
1.00	0.602	0.723	0.124	0.880	0.677	0.729

We can observe in Table 2 that, as we increase β, the discrimination on the Adult census increases and the accuracy decreases, which means that highly biased data is worse for the performance of a traditional classifier like C4.5. The COMPAS case shows better performance as it does not increase the discrimination as we raise β. In COMPAS the resulting decision trees for each β are similar, which explains its behavior. Dutch census is similar to Adult. As expected, removing the protected attribute contributes poorly in reducing discrimination because other attributes are highly correlated with the protected one.

Data pre-processing techniques Table 3 shows Discrimination vs. Accuracy results for different pre-processing techniques. Recall that we expect the discrimination and accuracy upper bounds to be preserved. We then highlight scenarios where the accuracy increased instead of decreasing.

We can see that in Table 3, on **Adult**, every technique except Massaging increases discrimination when using a more biased training set. Re-weighting and Uniform Sampling behave very similarly (this behavior is held true on

Table 3: Discrimination and Accuracy for data preprocessing techniques trained on artificially generated datasets learned from adult census.

	β \ Tech.	Massaging		Reweighting		Unif. Sampling		Pref. Sampling	
		Disc.	Acc.	Disc.	Acc.	Disc.	Acc.	Disc.	Acc.
Adult	0.00	0.046	0.830	0.111	0.842	0.112	0.842	-0.006	0.822
	0.25	0.009	0.816	0.140	**0.840**	0.139	**0.838**	-0.034	0.812
	0.50	0.095	**0.778**	0.260	**0.802**	0.257	**0.799**	0.044	**0.794**
	0.75	0.082	0.718	0.479	0.710	0.480	0.710	0.142	**0.756**
	1.00	0.003	0.676	0.473	0.690	0.474	0.690	0.416	0.673
COMPAS	0.00	0.095	**0.879**	0.104	0.876	0.110	0.877	0.076	0.872
	0.25	0.084	0.879	0.103	0.876	0.113	0.879	0.067	0.871
	0.50	0.070	0.875	0.102	0.875	0.111	0.878	0.054	0.870
	0.75	0.054	0.872	0.098	0.874	0.104	0.877	0.044	0.866
	1.00	0.057	0.871	0.107	0.877	0.108	0.878	0.045	0.866
Dutch	0.00	0.101	0.791	0.153	0.818	0.159	0.817	0.066	0.790
	0.25	0.013	0.768	0.154	0.818	0.166	0.818	0.018	0.774
	0.50	-0.052	0.748	0.150	0.816	0.161	0.816	-0.029	0.760
	0.75	-0.119	0.721	0.157	0.816	0.159	**0.813**	-0.062	0.749
	1.00	-0.182	0.693	0.483	0.720	-	-	-	-

other datasets). Preferential sampling has a steady decrease in accuracy and increase in discrimination, but its result for 100% bias is quite different. We hypothesize that, when the dataset becomes fully biased, its results are not really significant anymore, but are included for sake of providing a bound.

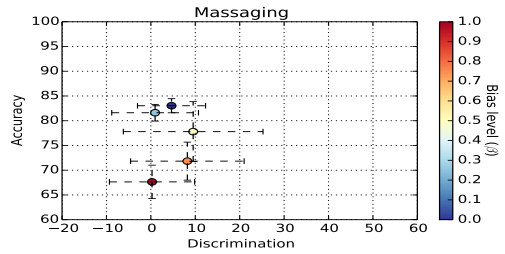

Figure 1: Discrimination vs Accuracy for data Massaging technique on Adult data

The most intriguing result is for the Massaging technique. Table 3 and Figure 1 show that it has lower discrimination for lower and higher β, but when we introduce moderate β it increases the resulting discrimination. In Figure 1, colored dots represent a set of 30 artificial datasets each and their colors vary according to β. Massaging works by defining a decision boundary between positive and negative outcomes through a Naive Bayes classifier, then it selects points from the dataset and changes their labels in order to balance the ratio of positives and negatives between the two classes of the sensitive attribute. Massaging works by changing labels, thus forcing the deprived group to have more positive outcomes and the favored group to have more negative outcomes. This is possibly the reason Massaging causes negative discrimination. We can observe that the results for the various techniques on **COMPAS** were also consistent with previous findings. Massaging and Preferential Sampling performed better in terms of lowering discrimination. Nevertheless, the highlighted scenarios where the accuracy increased suggests that data manipulation might actually improve both accuracy and discrimination on specific scenarios. It can be noticed on **Dutch** that Massaging and Preferential Sampling behave similarly, but the former has more spread values, which means that Preferential Sampling is more stable when dealing with this dataset. Re-weighting and Uniform Sampling, except on higher β, keep the discrimination and accuracy around the same level. Notice that we couldn't run Uniform or Preferential Sampling for higher β, since those techniques don't work well on very imbalanced scenarios.

Table 4: Discrimination and Accuracy for Auditor inprocessing technique.

	β \ Tech	Repair 0.25		Repair 0.50		Repair 0.75	
		Disc.	Acc.	Disc.	Acc.	Disc.	Acc.
Adult	0.00	0.149	0.844	0.064	0.818	0.039	0.805
	0.25	0.197	**0.841**	0.110	0.827	0.034	0.801
	0.50	0.492	**0.768**	0.242	**0.828**	0.031	**0.796**
	0.75	0.597	0.725	0.592	0.726	0.034	**0.795**
	1.00	0.605	0.723	0.605	0.723	0.576	**0.728**
COMPAS	0.00	0.121	**0.879**	0.123	**0.879**	0.122	**0.880**
	0.25	0.127	0.880	0.125	0.879	0.127	0.880
	0.50	0.124	0.879	0.129	0.880	0.127	0.879
	0.75	0.124	**0.879**	0.128	**0.879**	0.130	0.878
	1.00	0.125	0.879	0.129	0.879	0.129	0.877
Dutch	0.00	0.270	0.829	0.224	0.825	0.161	0.815
	0.25	0.270	0.831	0.211	0.824	0.154	0.814
	0.50	0.304	**0.829**	0.224	0.825	0.150	0.811
	0.75	0.390	**0.806**	0.295	0.798	0.143	0.798
	1.00	0.586	**0.730**	0.468	0.717	0.212	0.619

Auditor's technique Table 4 presents results for Auditor at repair levels of 0.25, 0.50 and 0.75. On **Adult**, we see it works well for lower β, but it performs poorly in more biased scenarios. It is worth noting that scenarios with very high discrimination are less realistic, which means that Auditor performance may be considered decent because it has minimal losses with respect to accuracy. When the repair level is set to 0.75, Auditor keeps the accuracy and discrimination at the same levels for every β with low deviation, except when maximum bias is inserted. On **COMPAS** Auditor performed similarly for each one of the repairing levels, but didn't quite remove the discrimination. The consistency of the results is also demonstrated by the similarity of the decision trees generated by the C4.5 in every repairing level. On **Dutch** census scenario, Auditor kept the accuracy around the 80%, but it didn't prevent the discrimination, and pre-processing techniques performed better. It maintained the best accuracy on overall for this scenario though. It can be observed that the 0.75 repair level performs slighly better than repair level 0.25 and 0.50. The highlighted numbers suggest that the Auditor could improve the accuracy in some cases. However, this improvement is not statistically significant, and, more importantly, there are cases where the discrimination increases.

6.2 Overestimation vs. Underestimation

Table 5 shows results on Overestimation and Underestimation for each dataset when no technique is applied. Again, it is important to consider these values as upper bounds, which means that the objective of each techniques is to reduce these

Table 5: Overestimation and Underestimation for Adult, COMPAS and Dutch (classifier does not use the protected attribute).

β \ Dataset	Adult		COMPAS		Dutch	
	Over.	Under.	Over.	Under.	Over.	Under.
0.00	0.065	0.105	0.012	0.046	0.194	0.072
0.25	0.154	0.399	0.015	0.044	0.239	0.108
0.50	0.471	0.712	0.015	0.047	0.268	0.162
0.75	0.480	0.745	0.013	0.043	0.312	0.346
1.00	0.480	0.831	0.014	0.049	0.356	0.870

(absolute) values. We can see that Adult and Dutch have high Overestimation and Underestimation at higher values of β. Curiously, COMPAS keeps the values close. This suggests that on COMPAS higher β may have little influence on the resulting Overestimation and Underestimation, which is also explained by the fact that the decision trees generated by the classifier are similar no matter the β.

Table 6: Overestimation and Underestimation for data preprocessing techniques.

	β \ Tech.	Massaging		Reweighting		Unif. Sampling		Pref. Sampling	
		Over.	Under.	Over.	Under.	Over.	Under.	Over.	Under.
Adult	0.00	-0.025	-0.217	0.030	-0.063	0.029	-0.062	-0.051	-0.324
	0.25	-0.062	-0.269	0.051	0.012	0.045	0.012	-0.084	-0.353
	0.50	0.011	-0.130	0.154	0.196	0.151	0.198	-0.023	-0.236
	0.75	0.005	-0.111	0.384	0.470	0.383	0.476	0.074	-0.124
	1.00	-0.086	-0.070	0.351	0.706	0.355	0.709	0.298	0.644
COMPAS	0.00	-0.017	0.021	-0.000	0.026	0.006	0.036	-0.031	0.003
	0.25	-0.024	0.006	-0.000	0.023	0.008	0.032	-0.038	-0.008
	0.50	-0.031	-0.012	-0.001	0.024	0.006	0.030	-0.044	-0.026
	0.75	-0.043	-0.031	-0.005	0.020	0.002	0.028	-0.051	-0.039
	1.00	-0.034	-0.032	0.001	0.030	0.006	0.031	-0.045	-0.040
Dutch	0.00	-0.065	-0.096	0.014	-0.095	0.023	-0.087	-0.091	-0.145
	0.25	-0.165	-0.159	0.015	-0.094	0.029	-0.078	-0.151	-0.170
	0.50	-0.239	-0.200	0.012	-0.098	0.024	-0.085	-0.197	-0.208
	0.75	-0.314	-0.236	0.019	-0.089	0.018	-0.083	-0.214	-0.247
	1.00	-0.384	-0.263	0.120	0.700	-	-	-	-

Data pre-processing techniques Table 6 shows the results for Overestimation and Underestimation for data preprocessing techniques. For **Adult**, Massaging reduces Overestimation and Underestimation in more biased scenarios. It is also interesting to note that almost every result of Massaging was a negative value. This technique keeps Overestimation closer to 0, meaning that it balances false positives between the two groups. The Underestimation has slightly higher negative values, but these values get closer to 0 as we insert more bias. This means that the technique underestimates the favored group on lower β, but it manages to balance the Underestimation between groups under higher β. Re-weighting and Uniform Sampling performed similarly by increasing both Overestimation and Underestimation, although they exhibit high deviation on both. Preferential Sampling presented hard to predict behavior; it has low Overestimation and its Underestimation started very negative and became closer to 0, except for the last biased sample. For **COMPAS**, we can see that Massaging and Preferential Sampling introduce negative Underestimation and Overestimation at higher β. Re-weighting and Uniform Sampling kept both Underestimation and Overestimation closer to 0, which is desirable. **Dutch** scenario suggests that Massaging and Preferential Sampling work similarly by inserting negative Overestimation and Underestimation when we increase β. Re-weighting and Uniform Sampling did a good job on keeping the values close (expect for the case of higher β on Re-weighting).

Table 7: Overestimation and Underestimation for Auditor in-processing technique.

	β \ Tech	Repair 0.25		Repair 0.50		Repair 0.75	
		Over.	Under.	Over.	Under.	Over.	Under.
Adult	0.00	0.056	0.096	0.010	0.006	0.001	-0.014
	0.25	0.087	0.269	0.032	0.092	0.001	-0.019
	0.50	0.368	0.644	0.131	0.386	0.002	-0.019
	0.75	0.484	0.750	0.482	0.750	0.003	-0.016
	1.00	0.483	0.829	0.484	0.821	0.470	0.727
COMPAS	0.00	0.015	0.046	0.019	0.047	0.019	0.046
	0.25	0.019	0.048	0.020	0.051	0.017	0.045
	0.50	0.018	0.049	0.020	0.052	0.021	0.044
	0.75	0.017	0.045	0.020	0.051	0.026	0.050
	1.00	0.019	0.051	0.022	0.052	0.018	0.053
Dutch	0.00	0.122	0.017	0.077	-0.027	0.013	-0.097
	0.25	0.128	0.031	0.071	-0.022	0.018	-0.093
	0.50	0.163	0.084	0.077	-0.015	0.018	-0.096
	0.75	0.196	0.265	0.094	0.173	0.019	-0.089
	1.00	0.223	0.796	0.109	0.685	0.030	0.314

Auditor technique Table 7 shows results for Overestimation and Underestimation of Auditor technique for each dataset on repair levels of 0.25, 0.50 and 0.75. For **Adult**, every technique performed similarly to Re-weighting and Uniform Sampling except Auditor at 0.75 repairing level, which kept almost every value close to zero for both Overestimation and Underestimation. For **COMPAS**, Auditor did not quite improve the result if compared to no technique used. In general, those techniques couldn't keep up with preprocessing techniques for this scenario. As for the **Dutch** case, Auditor improved the consistency of its results when we increase the repairing level to 0.75 as it has nearly 80% less overestimation compared to repair level of 0.25. It isn't pareto-dominated (which means that it isn't outperformed on both metrics) by any technique and provides a competitive result in most cases (if compared to the previous techniques).

7 Conclusions and Future Work

This work introduces a novel benchmark framework for validating discrimination-aware data mining and machine learning models using systematically biased datasets generated from real world data. The need for a benchmark is crucial due to the lack of a common ground for the evaluation of techniques. We demonstrated the applicability and effectiveness of the proposed benchmark through a comparative assessment among several models on three relevant datasets.

The value of our benchmark approach is apparent, when we observe that it is hard to define which technique is better than another. What is important is to decide which constraint is more relevant under a given scenario and then interpret the accuracy versus discrimination or over/under-estimation re-

sults in order to perform a trade-off between metrics. Our framework makes this type of decision easier. In the future, we intend to extend our work to a more complete coverage on techniques of disparate impact and disparate mistreatment removal, as well as other fairness metrics.

Acknowledgements

This research was partly supported by the Brazilian Research Agencies CNPq, CAPES, and FAPEMIG, by projects InWeb, MASWeb, EUBra-BIGSEA, INCT-Cyber and Atmosphere.

References

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2016. Auditing black-box models for indirect influence. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 1–10.

Bonchi, F.; Hajian, S.; Mishra, B.; and Ramazzotti, D. 2017. Exposing the probabilistic causal structure of discrimination. *I. J. Data Science and Analytics* 3(1):1–21.

Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21(2):277–292.

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, Sydney, NSW, Australia, August 10-13, 2015*, 259–268.

Hajian, S.; Monreale, A.; Pedreschi, D.; Domingo-Ferrer, J.; and Giannotti, F. 2014. Fair pattern discovery. In *Symposium on Applied Computing, SAC 2014, Gyeongju, Republic of Korea - March 24 - 28, 2014*, 113–120.

Hajian, S.; Bonchi, F.; and Castillo, C. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2125–2126.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10–18.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 3315–3323.

Kamiran, F., and Calders, T. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33(1):1–33.

Kamiran, F.; Calders, T.; and Pechenizkiy, M. 2010. Discrimination aware decision tree learning. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, 869–874.

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

Lichman, M. 2013. UCI machine learning repository.

Mancuhan, K., and Clifton, C. 2014. Combating discrimination using bayesian networks. *Artif. Intell. Law* 22(2):211–238.

Margaritis, D. 2003. Learning bayesian network model structure from data.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 560–568.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, 581–592.

Ruggieri, S.; Pedreschi, D.; and Turini, F. 2010. Data mining for discrimination discovery. *TKDD* 4(2):9:1–9:40.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, 1171–1180.

Zliobaite, I.; Kamiran, F.; and Calders, T. 2011. Handling conditional discrimination. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, 992–1001.

Zliobaite, I. 2017. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* 31(4):1060–1089.

A Appendix

In this appendix we depict a practical example of our Bayesian network based sampling method and show results regarding two more techniques created to reduce disparate impact and disparate mistreatment. Those techniques were both developed by Zafar et al in (Zafar et al. 2015) and (Zafar et al. 2017).

A.1 Practical example

We take the **adult** dataset (also known as **census income** dataset) from UCI Machine Learning Repository [4] (Lichman 2013) as an example for our methodology. This dataset is a demographic census and it is commonly used to predict whether an individual's income is greater than $50K per year or not given a list of attributes that include age, education level, marital status, occupation and sex.

A data cleansing step is performed by discretizing some of the attributes and removing others. The clean dataset is used as input for the `bnlearn` Hill-Climbing learner.

Figure 2 is a visual representation of the generated Bayesian network using the adult census dataset as input. It can be observed that there is conditional dependency between the variables *relationship* and *income_class* and also there is conditional dependency between *relationship* and *sex*. This evidence supports the fact that this data is biased with respect to sex, because sex is described by the relationship, e.g., a *wife* is a *woman*, and income class is also described by the relationship, e.g., being a *wife* may influence income.

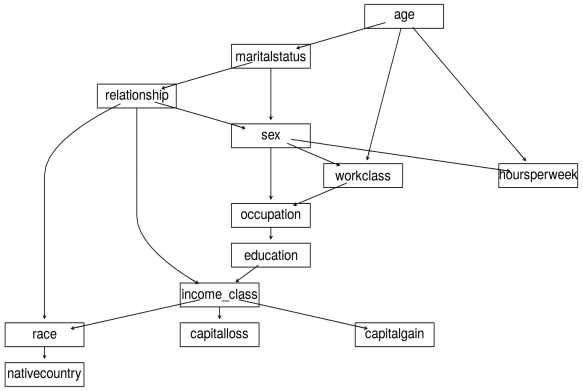

Figure 2: Bayesian network learned from the adult dataset.

For the synthetic data generation process, we have the corresponding BN for the Adult census dataset and its conditional probabilities table. Each row in this dataset is randomly generated by observing the BN. The list of dependencies is represented by the directed edges in Figure 2.

Firstly we observe the independent node in the BN, which is age. We sample age according to its distribution. After sampling age, we can sample the marital status. The conditional probability table for marital status is given in Table 8. Notice that age is divided into bins, which means that younger people are represented by bins identified by smaller numbers and older ones by larger numbers. We omit some of the bins due to space constraints.

[4] https://archive.ics.uci.edu/ml/index.php

Table 8: Conditional probabilities of marital status given some age categories (1-younger,10-older)

Marit. Status \ Age Cat.	2 (22-28)	5 (46-52)	8 (70-76)
Divorced	0.087	0.206	0.081
Married-AF-spouse	0.001	0.000	0.002
Married-civ-spouse	0.372	0.630	0.560
Married-spouse-absent	0.011	0.016	0.014
Never-married	0.490	0.077	0.060
Separated	0.032	0.033	0.014
Widowed	0.003	0.035	0.265

This process continues until we have sampled every feature. We then proceed to perform it again until we have enough rows sampled, thus generating a synthetic dataset that follows the same distribution as the original adult data.

In order to produce biased samples consider Table 9, which contains conditional probabilities for the income class node given education levels (Edu-level) and relationships (r). We can see that, according to this table, the probability of a positive outcome (the individual earns more than $50K per year), considering the fact that the individual is a wife is slightly better with higher education levels, but it is worse for lower education levels. In order to produce even more bias, we must decrease the probabilities in the first line of Table 9. For instance, we may decrease the probability of high income considering that the relationship is *wife*. This is done gradually, each time by 10% of the original probability, until the probability reaches 0, as described in Section 4.2. For example, the probability of an individual having high income, considering she is a woman and her education level is 10th grade, is 0.100. The first artificial dataset is sampled by lowering this probability to 0.090, the second one is 0.080 and so on.

The resulting sampled datasets are then used as training data for a classifier. We chose the C4.5 decision tree classifier due to its good results on previous works of discrimination-aware learning. Since our aim is to evaluate discrimination-aware techniques, the choice of the classifier is not that important.

Table 9: Conditional probabilities of high income (above 50$ a year) given education and relationship

Cond. prob \ Edu-level	1st-4th	10th	HS	Doctorate
$P(y = high \mid r = wife)$	0.000	0.100	0.343	0.850
$P(y = high \mid r = husband)$	0.085	0.128	0.311	0.837

A.2 Zafar et al.

We also consider the implementations of Zafar et al. (Zafar et al. 2017; 2015). (Zafar et al. 2015) designs a fair classifier that works by modifying the decision boundary of a classifier. It can maximize fairness under accuracy constraint

and maximize accuracy under fairness constraint. It is also possible to add a constraint on misclassification of positive outcome, which means that it only changes the label for the protected group from negative to positive class and doesn't change the label of non-protected groups to the negative outcome. This technique will be used for the disparate impact experiments (Accuracy vs. Discrimination). In (Zafar et al. 2017), they introduced the notion of disparate mistreatment and proposed a methodology that aims at reducing false positives and false negatives rates. The technique consists of training classifiers in such a way that their decision boundaries are modified in order to avoid those misclassification rates. We can adjust which constraints we want to use: correcting false positive, false negatives rates or both. This technique will be used for the disparate mistreatment experiments (Overestimation vs. Underestimation).

Discrimination vs Accuracy Table 10 presents results for Zafar's disparate impact technique for each one of its three constraints.

Table 10: Discrimination and Accuracy for Zafar's in-processing techniques trained on artificially generated datasets learned from adult census.

	Tech / β	Acc. cons.		Disc. cons.		Misclass.	
		Disc.	Acc.	Disc.	Acc.	Disc.	Acc.
Adult	0.00	0.128	0.837	-0.018	0.791	-0.066	0.590
	0.25	0.156	0.829	-0.032	0.773	-0.016	0.571
	0.50	0.181	**0.821**	-0.046	**0.760**	-0.015	0.530
	0.75	0.205	**0.814**	-0.046	**0.747**	-0.003	0.503
	1.00	0.258	**0.813**	-0.032	**0.743**	0.063	0.528
COMPAS	0.00	0.010	0.631	0.032	0.629	0.020	0.617
	0.25	0.014	0.630	0.014	0.628	0.029	0.615
	0.50	0.005	0.628	-0.002	0.622	0.022	0.610
	0.75	0.005	0.628	-0.027	0.620	0.040	0.610
	1.00	0.004	0.626	-0.043	0.615	0.033	0.607
Dutch	0.00	0.108	0.806	-0.031	0.746	-0.020	0.695
	0.25	0.091	0.801	-0.003	0.765	-0.059	0.647
	0.50	0.061	0.792	0.044	0.785	-0.105	0.596
	0.75	-0.025	0.762	0.127	**0.812**	-0.090	0.558
	1.00	-0.147	0.719	0.252	**0.820**	0.032	0.705

In Table 10 we present the results for Zafar's disparate impact techniques. In particular, the technique described in (Zafar et al. 2015) for fairness, accuracy and positive misclassification constraints. We can notice in the **Adult** scenario that the Discrimination constraint indeed keeps the discrimination around the same level at the cost of accuracy when the bias level β increases. On the other hand, for accuracy constraint, it keeps the accuracy around the same level and the discrimination increases with higher β. An interesting result can be observed when the positive misclassification constraint is applied, it causes lower discrimination levels but at a great cost of accuracy. In **COMPAS** case, Zafar's techniques tried to enforce less discrimination but also decreased the accuracy. As for the **Dutch** scenario, Zafar's techniques, except the one for positive misclassification, achieved competitive accuracy compared to pre-processing techniques and decreased accuracy. It provided a trade-off between accuracy and discrimination. As for the highlighted cases, when we apply Accuracy and Discrimination constraints on adult dataset the accuracy does not suffers very much for higher β. This means that Zafar's classifiers not only fix the bias inserted but also makes the classifier consistent with the original data. On Dutch dataset the same pattern can be observed for higher β.

Overestimation vs Underestimation Table 11 presents results for Overestimation and Underestimation of Zafar's technique for each dataset when False Positive, False Negative and Both constraints used.

Table 11: Overestimation and Underestimation for Zafar's in-processing techniques trained on artificially generated datasets learned from adult census.

	Tech / β	FP. cons.		FN. cons.		Both	
		Over.	Under.	Over.	Under.	Over.	Under.
Adult	0.00	0.040	-0.028	0.063	0.096	0.041	-0.015
	0.25	0.072	0.105	0.136	0.264	0.086	0.157
	0.50	0.158	0.348	0.355	0.477	0.266	0.399
	0.75	0.365	0.557	0.450	0.593	0.388	0.521
	1.00	0.479	0.764	0.466	0.748	0.470	0.751
COMPAS	0.00	0.076	0.117	0.101	0.157	0.106	0.168
	0.25	0.089	0.133	0.081	0.125	0.106	0.168
	0.50	0.051	0.086	0.067	0.109	0.075	0.132
	0.75	0.044	0.066	0.072	0.112	0.075	0.132
	1.00	0.048	0.072	0.067	0.108	0.005	0.014
Dutch	0.00	-0.103	-0.000	-0.114	-0.009	-0.113	-0.008
	0.25	-0.125	-0.021	-0.162	-0.056	-0.149	-0.044
	0.50	-0.142	-0.039	-0.202	-0.103	-0.164	-0.059
	0.75	-0.120	-0.028	-0.273	-0.279	-0.154	-0.082
	1.00	-0.040	-0.355	-0.453	-0.906	-0.108	-0.656

Table 11 in **Adult** scenario we can see that Zafar's technique with False Positive constraints improved the results from Table 5 by pareto despite having high Overestimation and Underestimation for higher β and it also isn't pareto-dominated by any pre-processing technique. When we use the False Negative constraint it also improves the results if compared to Table 5, but is pareto-dominated to Reweighting, Uniform Sampling and Preferential Sampling. Applying Both constraints it improves the results, isn't pareto-dominated by the previous techniques but is almost outperformed by the False Positive constraint. Considering the **COMPAS** scenario, Zafar's techniques for False Positive constraint, False Negative constraint and Both constraints are almost pareto-dominated by every pre-processing technique. They didn't improve the results when no technique is employed. As for **Dutch** census case, the False Positive constraint improves the results if compared to Table 5. The remaining constraints are pareto-dominated by FP constraint.