Bachelor's Thesis

Radboud University Nijmegen

Evaluating the Reductions Intervention for Enforcing Fairness in Classification

Author:
M. de Jong, s4813111
M3.deJong@student.ru.nl

Supervisor:
prof. dr. T.M. Heskes
Department of Computer Science

June 28, 2019

1 Abstract

Fairness is a topic in machine learning concerned with removing biases present in machine learning systems. One method for removing such biases is the reductions approach by Agarwal et al.[1], which has the notable advantage of being usable for a wide range of fairness definitions. Past work shows that this method gives a good fairness-accuracy tradeoff compared to other methods that enforce fairness. In this paper, the performance of this method will be further investigated. To do this, I will compare the method against a preprocessing method for the demographic parity fairness definition.

2 Introduction

The usage of machine learning (ML) systems has become widespread in recent years, now being used in topics as wide as supply chain forecasting[3] to predicting whether individuals will suffer from PTSD[18]. While computers typically tend to be viewed as objective and unbiased, machine learning systems do not necessarily hold these properties, as their behavior may depend on data biased by humans. Machine learning systems are trained to predict based on a certain dataset. A problem is that any biases present in this dataset might thereby also be present in the predictions of the system. In the existing literature, cases where bias against a certain social group (e.g. based on race, gender, etc.) exist, are typical examples of unfairness. However, it is important to remember that, in theory, the problem occurs for any kind of bias that might exist in the behavior of these systems.

An illustrative example is the chatbot Tay, developed by Microsoft. Tay was trained on data obtained through the social media platform Twitter; but due to the racist nature of some texts on Twitter, Tay started making racist claims itself[16]. This is a clear example of how bias in a machine learning system may arise from human data. An example that has received significant attention due to its societal relevance, is that of the COMPAS algorithm for recidivism[5]. Judges in the US used this system to predict whether a criminal was likely to re-engage in crime. This information was used in the process of making sentence judgments. A disturbing property of the system is that it made much more false positives for black individuals than white individuals; in other words, it more often falsely predicted that black individuals would recidivate than it made such false predictions for white people. This example demonstrates a clear need for fairness in machine learning systems.

Fairness is the subfield within machine learning that is concerned with removing the effects of unfair biases in machine learning systems. The field has gained a lot of traction in recent years; at ICML 2018, two out of the five best papers were on the topic of fairness[21], and several workshops on the topic are given[11]. So far, the focus in the field has mostly been on statistical definitions of fairness or algorithms that enforce such definitions, although findings from a practitioner survey[11] suggest that more focus is required on the process of fair data collection.

Research into the process of creating fair machine learning systems can be thought of as falling within one of three lines:

1. Research into informal definitions of fairness. These informal definitions should be justified by legal, moral and philosophical considerations, and the implications reach beyond the topic of fairness in machine learning.

2. Research into formalizations of fairness. These formalizations should define fairness in precise, mathematical terms, while effectively representing informal definitions of fairness.

3. Research into interventions on ML-systems that can enforce fairness in these systems. These interventions may be methodical (e.g. certain steps taken in the data collection

process that ensure fairness) or technical (e.g. utilizing certain algorithms that guarantee fairness). These interventions should try to minimize the violation of the formalizations of fairness. If they are to be considered useful, these interventions also should not disturb the other functions of the system.

This research falls strictly into the third category, and the focus will be on technical interventions. In this research, two interventions that enforce fairness on a classification task will be compared.

2.1 Related Work

In the literature, many technical interventions for enforcing fairness have already been proposed. Such interventions can roughly be divided into three categories[2]:

1. Pre-processing interventions. These methods transform the data the machine learning models are trained on.
2. In-processing interventions. These methods alter the training process of machine learning models.
3. Post-processing interventions. These methods alter the machine learning model post-hoc; the decisions made by the model are adapted to be fairer.

A wide range of pre-processing interventions have already been proposed, using e.g. sampling, re-weighting, messaging of data[13] or using more advanced methods such as t-closeness[17]. A state-of-the-art approach proposed by Calmon et al.[2] formulates an optimization problem for transforming the data that seeks to control discrimination, while limiting distortion in individual data samples and preserving utility. A notable preprocessing intervention that enforces demographic parity through shifting the feature distributions for different groups towards the median was proposed by Feldman et al.[6]. Some methods for enforcing counterfactual fairness, also do so by preprocessing the data[14].

Perhaps most of the research has inquired into in-processing interventions for fairness. The most promising approaches will be listed here. Zafar et al. propose an approach for avoiding both disparate impact and disparate treatment[19] and later an approach for avoiding disparate mistreatment[20]. Donini et al.[4] propose an intervention for enforcing the 'Equality of Opportunity' fairness measure. A method proposed by Fish et al.[7] focuses on explicitly balancing fairness and accuracy through shifting the decision boundary for sensitive groups. Johnson et al.[12] propose an approach for enforcing fairness that works for any machine learning model, for an 'impartiality score' they define. Enforcing counterfactual fairness can also be done at training time[15].

Likely the least researched of the three are postprocessing interventions. Only one approach within this category seems notable. This is the method used by Hardt et al.[10], which was tested for the Equalized Odds and Equality of Opportunity fairness measures.

As should be clear now, many interventions for enforcing fairness have already been proposed. Research that is much rarer, however, is research wherein different interventions are compared. In theory, all the methods for enforcing fairness on a classification task can be compared on two things: the extent to which an intervention succeeds in enforcing fairness and the effect the intervention has on the accuracy of the classifier. However, in practice, this is very difficult as many methods are specific to:

- One fairness measure/A group of fairness measures. Many of the in-processing methods mentioned above are specific to one fairness measure.
- One machine learning model/A group of machine learning models. For instance, the method proposed by Zafar et al.[19] only works for convex-based classifiers.
- One machine learning problem/A group of machine learning problems. The reductions

approach tested in this paper only works for binary classification problems, for instance. Other measures can be applied to any kind of classification problems[2].

Only one paper so far contains a notable comparison on many different methods for enforcing fairness. In this paper by Friedler et al.[9] methods were compared not only by the tradeoff they gave between accuracy and fairness but also the stability of such methods, the degree of dependency for such measures on preprocessing and how different fairness measures correlate. Much more research wherein different interventions are compared seems necessary.

In this paper, I will be investigating the reductions approach proposed by Agarwal et al.[1]. I consider this intervention particularly interesting as this approach can be used on a wide range of fairness metrics. This is in contrast to many other interventions, which typically only work for one particular metric of fairness. Focusing on an intervention that can enforce multiple fairness metrics is beneficial, as it means research into what metrics best represent fairness, and research about enforcing fairness may occur independently.

2.2 Research Formulation

As stated above, many different fairness interventions have been proposed, but there exists very little research that compares these different methods. This hinders advancements in the field, as it complicates judgements about which approaches are truly the most promising, and might be deserving of further investigation/development. With this research, I wish to contribute to the research comparing the methods. The question addressed in this research is:

How does the reduction intervention for enforcing fairness compare to other interventions for enforcing fairness?

The reductions method will be compared to the repair preprocessing method of Feldman et al.[6], as it is a simple approach that has been shown to be effective at enforcing fairness.

I will compare the two methods on the demographic parity fairness formalization, often also called 'disparate impact'. Explanations of this formalization will follow later.

The experiments will be performed for two classifiers and two datasets. The results of the interventions will also be compared against conditions without intervention. There are then a total of three independent variables: The classifier, the dataset and the type of fairness (if any) enforced on the classifier.

The results will be analyzed on both how well the accuracy/fairness tradeoffs can be made, and the maximum fairness that can be obtained using the intervention. Based on the earlier comparisons by Agarwal et al., I predict that the reductions method for enforcing fairness can both maintain a higher accuracy while enforcing fairness, and that it can further enforce fairness than the preprocessing approach. As both techniques measure the same fairness definition in slightly different ways (although the metrics are roughly equivalent), fairness will be evaluated using both metrics.

3 Fairness and Methods for Enforcing Fairness

In this section, the demographic parity fairness formalization is discussed, along with the two methods for demographic parity.

3.1 Notation

The notation used here will be used throughout the rest of this report.

Let X be the input vector (of attributes) to a classifier, and Y the true class that is to be predicted by the classifier. In this case, as the classification problems are binary classification, $Y \in \{0, 1\}$. For a (trained) classifier h, the prediction is denoted by $h(X)$. The group attribute is denoted by G and may be used as a feature

in X. In the datasets utilized here, the group attribute is also binary, so therefore, $G \in \{0, 1\}$. The function err(h) returns the error of a classifier h, and the the function $f(h)$ returns the unfairness of a classifier h.

3.2 Demographic Parity

The fairness definition that is used in the experiment is demographic parity. The underlying idea behind demographic parity is that the true class Y and the group class G should be statistically independent. In formal notation:

$$P[h(X)] = P[h(X)|G].$$

However, this definition is much too strict to be enforceable. In practice, it will almost never be the case that the two probabilities are exactly equal. This is why alternative metrics are used. The idea behind these metrics is that they indicate the extent to which the above definition is violated. As will become clear in a later section, there are various ways to implement metrics that all implement the same underlying definition of statistical independence.

3.3 Reductions Approach

In the paper by Agarwal et al.[1], the problem of enforcing fairness on a machine learning system is formalized as a constrained optimization problem; we want to optimize our classifier while satisfying certain fairness constraints. In this section, we explain how this is achieved using the reductions approach. A more elaborate explanation may be found in their paper.

An important concept to the Reductions approach is the use of *randomized classifiers*. Randomized classifiers can be seen as classifiers that combine the predictions of a set of base classifiers. In this set, each base classifier has its own weight. The randomized classifier predicts a probability for the class label, and this probability is the weighted average of the predictions of the base classifiers.

The reductions approach can best be explained as involving three steps:

1. Formulating the optimization problem of a randomized classifier that is to be solved.

2. Applying the method that solves the optimization problem by repeatedly adding base classifier to the randomized classifier.

3. Training the individual base classifiers using cost-sensitive training.

The reductions approach is called the *reductions* approach, as the constrained optimization problem can be reduced to a sequence of cost-sensitive classification problems.

3.3.1 Optimization Problem Formulation

The reductions approach seeks to minimize the error of a randomized classifier, while satisfying constraints over the fairness of that randomized classifier. These constraints specify that the unfairness function cannot exceed a particular value. Formally, for a particular unfairness value c which the randomized cannot exceed, we want to find:

$$\min_Q \text{err}(Q) \text{ subject to } f(Q) < c.$$

This particular constrained minimization problem can be solved by finding the saddle points for a Lagrange function $L(Q, \lambda)$. In the case of a function with two parameters, a saddle point occurs when there is a relative maximum for one of the parameters, and a relative minimum for the other parameter. In this case, the partial minimum of the Q parameter in the Lagrange function $L(Q, \lambda)$, corresponds to the minimum of the error for the randomized classifier. Likewise, the partial maximum for the λ parameter in the Lagrange function $L(Q, \lambda)$, corresponds to points where the fairness constraints are satisfied.

3.3.2 Method for Solving the Optimization Problem

An approximated saddle point is found using the standard method proposed by Fruend et al.[8]. This method can be described as two players playing a zero-sum game in turns. The Q-player seeks to minimize $L(Q, \lambda)$ by changing the Q parameter in its turn. In other words, every turn the Q-player seeks to minimize the error by changing the randomized classifier. The λ-player seeks to maximize $L(Q, \lambda)$ by changing the λ parameter in its turn. In other words, every turn the Q-player finds the constraint that is the most violated, if any. When neither player can make a turn that increases/decreases $L(Q, \lambda)$ by more than a constant, then an approximate saddle point is found.

In the turns of the Q-player, the randomized classifier is changed by adding a new base classifier to it. Initially, the set of base classifier of the randomized classifier is empty; it is filled by the Q-player adding a new base classifier h to the set every turn. All the base classifiers in the set are given equal weights.

3.3.3 Training the Base Classifiers

The new base classifiers that are added to the set are first optimized on the data under the current λ-value/constraint. It turns out that the optimization process under this constraint can be transformed into a cost-sensitive training process, which is thus how the base classifier is trained. A cost-sensitive training process entails that each data sample has its own cost; certain samples have more effect on training the classifier.

The result of the reductions approach is the randomized classifier Q which satisfies the fairness constraints. Unlike the base classifiers, which predict a class label of either 0 or 1, the randomized classifier gives a probability with which it predicts the probability of the sample being of class 1. This probability is obtained by averaging the predictions of the base classifiers of the randomized classifier.

3.4 Preprocessing Approach

The preprocessing method enforces fairness through transforming the dataset. In the transformed dataset, the distributions of the input attributes for certain groups are shifted to the median for those attributes for all the groups. This is how the intervention achieves fairness.

In more detail, take for a certain feature X_i and a certain group class $g \in G$. Let $X_{i,g}$ be the distribution of X_i conditioned on group g. Let $F_{X_{i,g}}$ be the cumulative distribution for $X_{i,g}$. Let $F^{-1}_{X_{i,g}}$ be the quantile function for $X_{i,g}$.

The median distribution for a particular feature A_i is defined by its quantile function $F^{-1}_{A_i}$. Its quantile function, in turn, is defined as the median of the quantiles of the different group classes: $F^{-1}_{A_i}(u) = \text{median}_{g \in G}(F_{X_{i,g}}(u))$.

The new dataset is created by shifting the distributions of the features per group class. The quantile points for a data point, conditioned on a certain class, are shifted to the quantile points of the median over the classes. That is, if a given data point has a quantile value of 0.9 conditioned on its certain class, after shifting, it will have a quantile function of 0.9 over the median of the classes. Formally, for a given feature value $x \in X_i$ that is in group g, its value after shifting $\hat{x}$ will be $\hat{x} = F^{-1}_{A_i}(F_{X_{i,g}}(x))$.

The procedure described above is only capable of fully shifting the quantiles. However, this may have a large negative effect on accuracy. For this reason, a partial repair algorithm is also proposed. The partial repair algorithm takes a repair degree $\lambda \in [0, 1]$, and shifts the quantiles proportionally to this degree. For $\lambda = 0$, no changes will be made to the dataset. Repair with $\lambda = 1$ performs a full repair.

The workings of partial repair are relatively simple. In this case, *combinatorial repair* was used. Combinatorial repair works simply by partially moving the ranks towards the median. For $\lambda = 0.5$, a data point is shifted half the nominal ranks to the median.

It should be noted that in general, the method of repair only works on numerical data, as it is

not possible to define a cumulative or quantile function over a categorical distribution.

4 Experimental Methods

4.1 Datasets

The two datasets that were used in the experiment are:

- The COMPAS recidivism dataset[a]. This dataset contains information about convicts, and how likely they are to recede (to commit another crime). In the experiment, the classification task was to predict whether a convict will recede. The group attribute was taken to be the race of the individual. Because we know from previous studies that there exists unfairness between the 'Caucasian' and 'African-American' categories, these two categories were taken as the groups. The entries for other races were dropped. Certain small preprocessing steps were performed as they were done in the original analysis.
- The adult income dataset[b]. This dataset contains information about the attributes (among which the income) of adults in the United States. In the experiment, the classification task was to predict whether a particular individual earned more than $50,000$ dollar a year. The group attribute was taken to be the gender of the individual.

For both datasets, numerical features were normalized. Categorical data, when possible, was made ordinal, and otherwise removed. This is so that the preprocessing approach (which cannot handle categorical data) could be compared on equal grounds to the reductions approach. The accuracy of the classifiers was observed with and without this step, to check that this step would not cause a dramatic decrease in accuracy. The group attribute was included among the attributes in the datasets, as the preprocessing intervention requires access to the group class at test time, and a comparison would be unfair if the reductions approach did not have this access. Both datasets were split into a training/test set with ratio 4:1.

4.2 Classifiers

Two cost-sensitive classifiers were used in the experiments:

- A logistic regression classifier. No regularization was required. The classifier used was an implementation from the Sklearn package[c].
- A decision tree classifier. A max-depth was set for either classification task to prevent overfitting. The classifier used was an implementation from the Sklearn package[d].

Both classifiers are relatively simple, but they were chosen because more complex classifiers showed to be prone to overfitting. I noticed this in initial testing with a random forest classifier.

4.3 Fairness Interventions

For every dataset/classifier combination, the condition was subjected to three fairness interventions:

- No intervention being applied. In this case, the classifier is just trained on the dataset, as normally.
- The reductions intervention being applied. The allowed constraint violation was set to the values 0.005, 0.01, 0.05, 0.1 and 0.5. Further reducing this number would have

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

[b] `https://archive.ics.uci.edu/ml/datasets/Adult`

[c] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[d] `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`

further increased fairness on the training set, but it would be unlikely to increase fairness considerably further on the test set. Furthermore, increasing fairness has diminishing returns, where increasing fairness comes at an increasingly higher cost of accuracy. The implementation I used is the implementation of the authors themselves[e].

As described earlier, the reductions approach results in the probability of a class instead of a prediction of the class. Intuitively, it makes sense to binarize this probability using a threshold of 0.5 to obtain the class predictions. Another approach would be to sample from the probability to obtain the class predictions. In the original paper, their choice for either approach was not mentioned.

While binarizing the probability did give better accuracy, it significantly hindered the algorithm in reducing fairness. In one condition it would stop fairness from improving at all. For this reason, I chose to sample from the probability, which reduced unfairness without problems. For every fairness degree, the accuracy and fairness of the resulting classifier were averaged 30 times, to average out the variance introduced by the additional random aspect.

- The preprocessing intervention being applied. The repair degree was set to the values 0, 0.25, 0.5, 0.75 and 1. As stated in the previous section, 1 signifies a full repair. The implementation is the implementation of the authors themselves[f]. As this implementation is stochastic (for partial repair), the intervention was run 30 times and the accuracy and fairness were averaged.

  The intervention was run with the *kdd* option set to true, which means the the group attribute is removed proportional to the repair degree.

  With this approach, the entire dataset was transformed using the preprocessing intervention and consequently split into a train/test set.

[e] `https://github.com/Microsoft/fairlearn`

[f] `https://github.com/algofairness/BlackBoxAuditing`

4.4 Fairness Metrics

As mentioned before, there are multiple ways to implement demographic parity as a metric. Both papers measure demographic parity using different metrics. In the experiment, I have tested both interventions on both methods, to see whether results also hold between different metrics. Intuitively, I expect this to be the case, as the demographic parity fairness definition underlies both metrics.

1. In the reductions paper, demographic parity was measured using

   $$\max_{g \in G} |P[(h(X)|G = g] - P[(h(X)]|.$$

   This is also the measure that is directly minimized by the method. The authors chose this metric as it's mathematical form is easy to optimize. We will call this metric *max disparity*.

2. In the Preprocessing paper, demographic parity was measured as

   $$\frac{P[h(X)|G = 1]}{P[h(X)|G = 0]}.$$

   Where group 0 is the 'discriminated' group. In the experiment, we didn't specify which specific group would be discriminated against, but instead computed the score for both groups as being the discriminated group, and took as unfairness metric the maximum of the two. This means I implemented the metric as

   $$\max_{g \in G} \frac{P[h(X)|\neg(G = g)]}{P[h(X)|G = g]}.$$

   We will call this metric *disparate impact*.

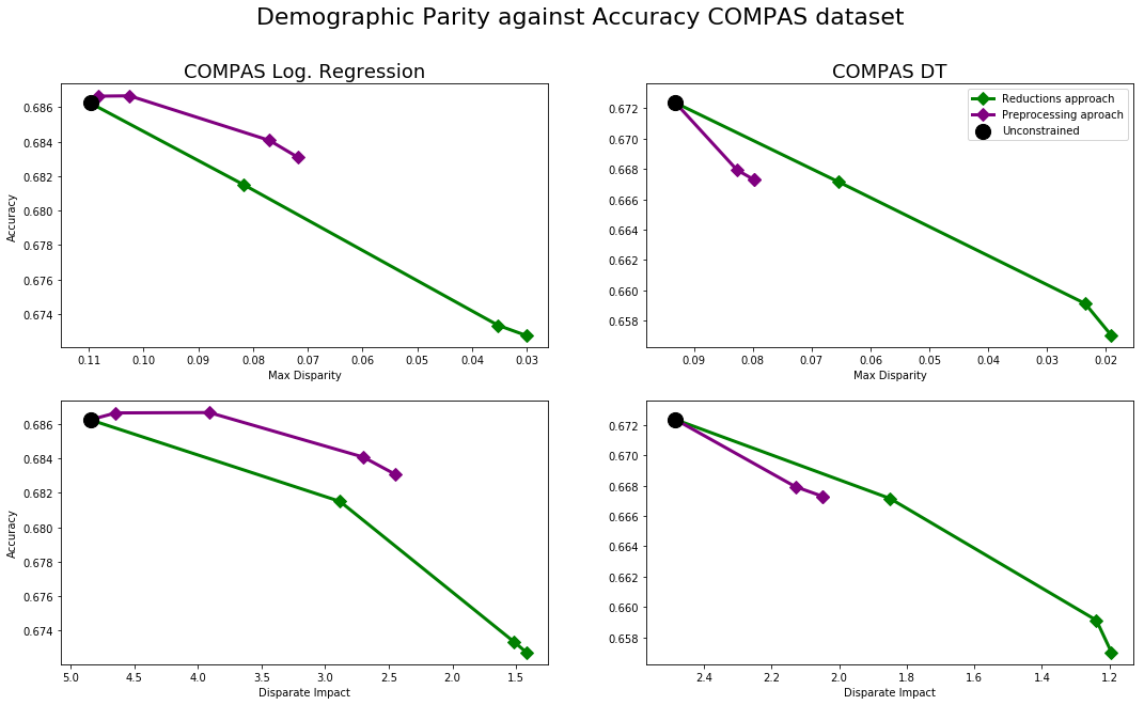

Figure 1: Experimental results for the COMPAS dataset for the reductions and preprocessing intervention, and without intervention. The first row shows the results for the 'max disparity' fairness metric, the second row the results for the 'disparate impact' metric. It should be noted that the axis ranges differ per subplot.

These metrics will be applied to the predictions of the trained classifiers on the test sets.

5 Experimental Results

Figure 1 contains the experimental results as obtained for the COMPAS dataset. Figure 2 contains the results for the Adult dataset.

The trained classifiers gave similar accuracies on our datasets as previous studies obtained. Using an unconstrained logistic regression classifier, Agarwal et al. obtained an accuracy of roughly 0.70 (an error of 0.30) on the COMPAS dataset and an accuracy of roughly 0.85 (an error of 0.15) on the adult dataset. My unconstrained classifiers obtained accuracies of 0.685 and 0.672(for the logistic regression and DT classifier respectively) on the COMPAS dataset and accuries of 0.825 and 0.828 on the adult dataset. The slightly lower accuracies obtained here can likely be explained by the removal of certain categorical attributes.

The logistic regression classifier gave a better accuracy on the COMPAS dataset than the decision tree classifier, but also a higher degree of unfairness. The opposite is true for the Adult dataset.

The most striking result is that in all cases, the reductions intervention is able to further increase fairness than the preprocessing intervention. This does not only hold for the fairness metric that it directly minimizes (max disparity), but also for the fairness metric that was

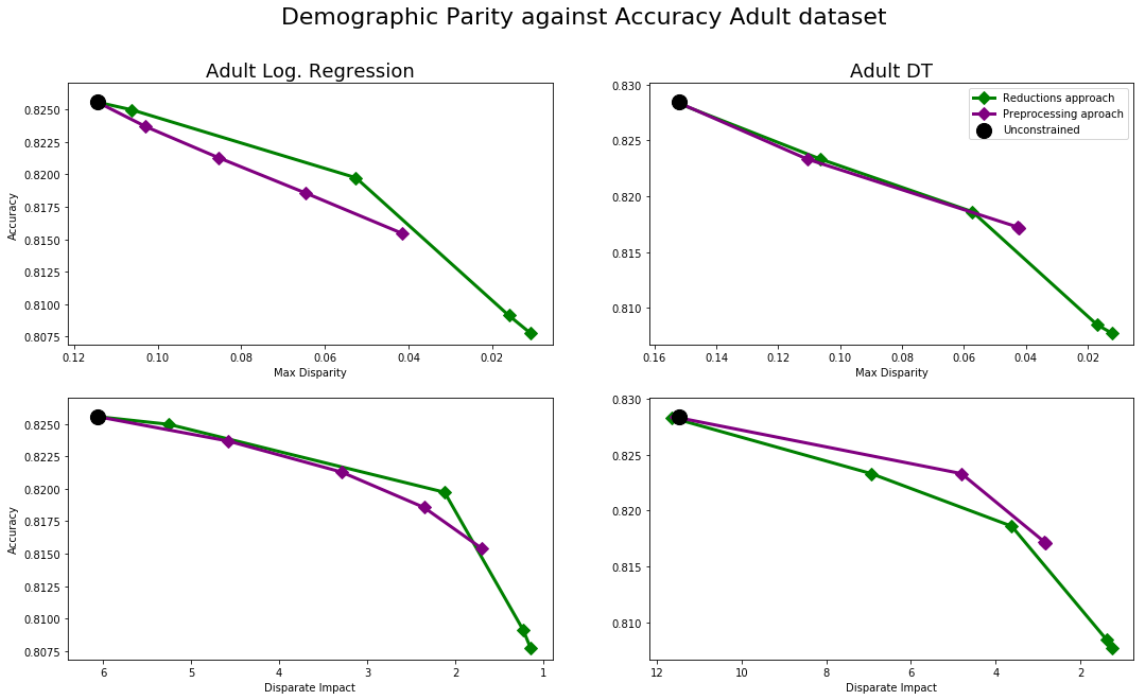

Figure 2: Experimental results for the Adult dataset for the reductions and preprocessing intervention, and without intervention. The first row shows the results for the 'max disparity' fairness metric, the second row the results for the 'disparate impact' metric. It should be noted that the axis ranges differ per subplot.

originally used to evaluate the preprocessing intervention.

In all dataset/classifier combinations, the reductions approach is able to obtain somewhat similar fairness; roughly 0.02 for max disparity and roughly 1.25 for disparate impact. The preprocessing approach is more variable in how far it can enforce fairness; its results range between 0.08 and 0.04 for max disparity, and between 3 and 1 for disparate impact.

It is difficult to conclude something about the fairness/accuracy tradeoff between the two classifiers. In two of the dataset/classifier combinations, the reductions intervention performs better; in one, the preprocessing intervention performs better. In one, it depends on the metric.

Figures 3 and 4 shows the same results but with the 95% confidence intervals added. It should be noted that these confidence intervals may not be very accurate, because in many cases the distribution of the results did not resemble a normal distribution. Still, in many cases, the confidence intervals overlap, which means a conclusion cannot confidently be drawn.

When the fairness constraint is set to only accept a very low violation, the cost of fairness seems to be between a 1-2% difference in accuracy. However, this did differ considerably between datasets and classifiers.

The different metrics seem to give similar results to each other. In three out of four conditions, the intervention that performed best on one metric also performed best on the other. Only in one condition, on the Adult dataset with

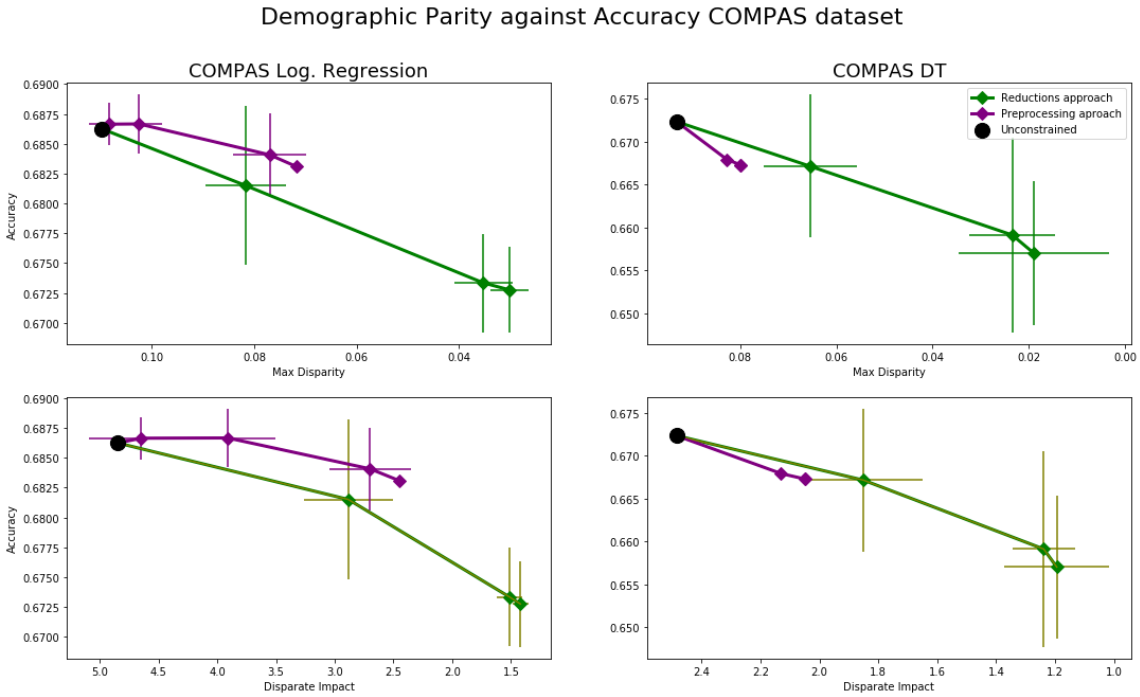

Figure 3: The same results as figure 1, but with 95% confidence intervals added. The confidence intervals were computed using two times the standard deviation. A problem is, that, particularly for the decision tree classifier, the results did not closely resemble a normal distribution. Inspection of certain data points found that they sometimes had two distinctive, remote and very narrow peaks. The 95% confidence intervals computed may thus very poorly represent the true 95% confidence intervals.

a decision tree classifier, did the reductions approach score best on the max disparity metric, while the preprocessing approach scored best on the disparate impact metric.

6 Discussion

One weakness of my experiment was that both classifiers used were relatively simple. These simple classifiers were chosen because more complex classifiers tended to overfit. However, it is of interest to know whether similar results in fairness could be obtained on more complex classifiers. Classifiers that can more effectively utilize the interactions between input attributes, may bring about unfairness through these complex interactions. From the experiment that I performed, we cannot conclude that the interventions can remove such interactions as sources of unfairness.

The fact that the other intervention required certain preprocessing steps also limits the conclusions we can draw about the reductions intervention. As stated, we transformed into ordinal/removed categorical attributes because the preprocessing intervention could not utilize them. This means that both interventions were tested on datasets with only ordinal attributes. However, this makes the results about the reductions approach less generalizable; perhaps different results would have been achieved on a dataset

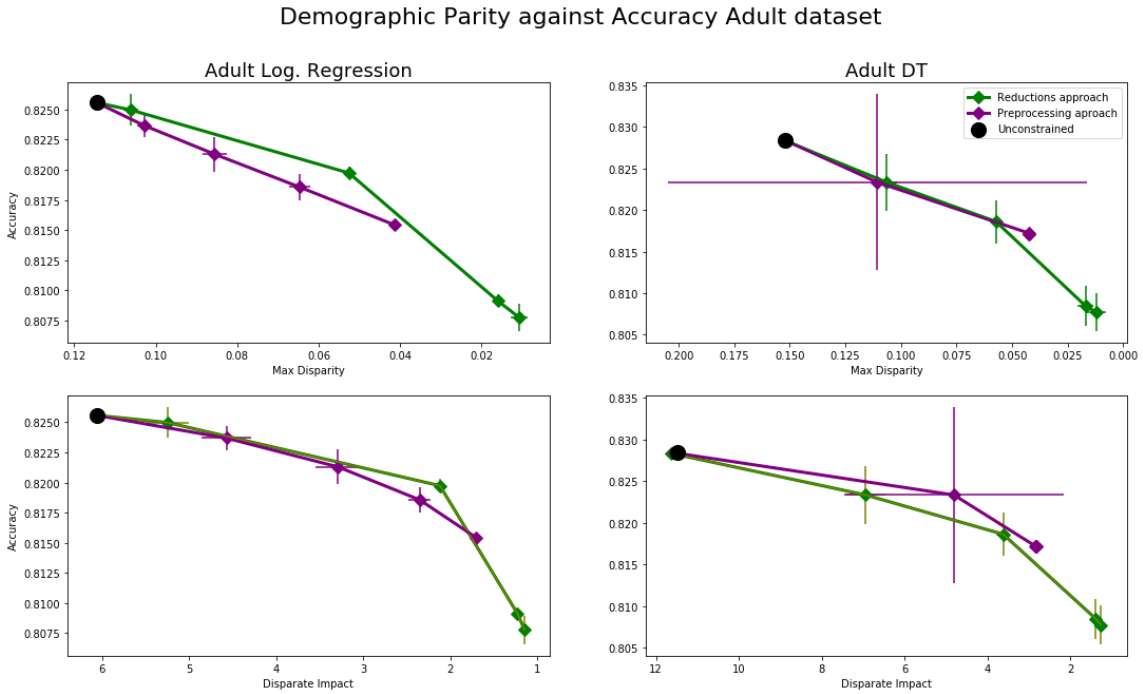

Figure 4: The same results as figure 2, but with 95% confidence intervals added, following the explanation with figure 3.

that also includes categorical attributes.

There are multiple research directions which could expand upon the current research:

- A comparison of the reductions intervention against other interventions. As explained in the introduction, the number of proposed interventions is already very high. To strengthen the certainty of the conclusions obtained in this research, comparisons with other interventions are necessary.
- A comparison of interventions on more complex classifiers, such as support vector machines or random forest classifiers.
- A comparison of interventions for multiple protected groups. In real applications, discrimination on a lot of factors (gender, race, religion, etc.) should be avoided. Therefore, for practical applications, it is important that approaches can prevent discrimination on more than factor.

To make a broader claim about the required advances in the field, an agreement on common terminology and metrics seems of necessity. As indicated in the introduction, a lot of research has already focused on interventions to increase fairness, but these methods are very difficult to compare as they focus on different fairness definitions. More so, even when the same fairness definition is used, it is often given a different name, or different metrics for measuring it are used. To give an example of the former, the measure here named 'demographic parity' also goes by the names 'disparate impact' and 'independence' in other sources. Although comparisons in the field are inherently difficult because interventions vary in focus on fairness definitions/ML models/ML problems, the lack of common ter-

minology further complicates the evaluation between different fairness interventions.

Lastly, it is important to always keep practitioner needs in mind for future research directions. Holstein et al.[11] have performed an investigation in the needs and challenges in teams that develop commercial products. They draw the conclusion that "although the fair ML literature has overwhelmingly focused on algorithmic 'de-biasing', future research should also support practitioners in collecting and curating high quality datasets in the first place".

7 Conclusion

The experiment indicates that the reductions approach is very effective in enforcing demographic parity. For all experimental conditions and demographic parity metrics, the reductions approach could more extensively increase fairness than the preprocessing intervention.

From the experiment, no clear conclusion can be drawn about which intervention can make a better accuracy/fairness tradeoff. The differences in the tradeoff between the two interventions are generally small. Which intervention gave a better tradeoff ultimately differed between conditions.

Research that could expand upon this research should focus on comparisons with other interventions, using more complex classifiers or protecting multiple groups. Further, in the broader field of fairness in machine learning, extra focus is necessary on common terminology and the gathering of fair datasets in the first place.

Acknowledgements

I would like to thank my peers Marjolein Laan, Vincent W. Philips and Tiemen de Jong for giving me feedback on preliminary version of this thesis. Your help was very much appreciated.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A Reductions Approach to Fair Classification. *CoRR*, abs/1803.02453, 2018.

[2] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized Pre-Processing for Discrimination Prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[3] Rl Carbonneau, R Vahidov, and Kevin Laframboise. Forecasting supply chain demand using machine learning algorithms. *Distributed Artificial Intelligence, Agent Technology, and Collaborative Applications*, pages 328–365, 01 2008.

[4] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints, 2018.

[5] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, jan 2018.

[6] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

[7] Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. *CoRR*, abs/1601.05764, 2016.

[8] Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. 1996.

[9] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A

comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 329–338, New York, NY, USA, 2019. ACM.

[10] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *CoRR*, abs/1610.02413, 2016.

[11] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? *CoRR*, abs/1812.05239, 2018.

[12] Kory D. Johnson, Dean P. Foster, and Robert A. Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models, 2016.

[13] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.

[14] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[15] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness, 2018.

[16] H Reese. Why Microsoft's 'Tay' AI bot went wrong. https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/, mar 2016. Accessed on February 25th 2019.

[17] Salvatore Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy*, 7(2):99–129, August 2014.

[18] Glenn Saxe, Sisi Ma, Jiwen Ren, and Constantin Aliferis. Machine learning methods to predict child posttraumatic stress: A proof of concept study. *BMC Psychiatry*, 17, 07 2017.

[19] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification, 2015.

[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. In *Proceedings of the 26th International Conference on World Wide Web - WWW 17*. ACM Press, 2017.

[21] Z. Zhong. A Tutorial on Fairness in Machine Learning. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb, oct 2018. Accessed on February 25th 2019.