01009425

1. Introduction

This report sets out an experimental investigation into the effect of regularisation on both accuracy and fairness. To perform the experiments, we used the object-orientated AI Fairness 360 library, AIF360, a relatively new open source Python toolkit for machine learning fairness [1]. We assess the effectiveness of AIF360's bias mitigation algorithms and apply the empirical observations to come up with a model selection strategy for fair classification. We also report a short study into the effect of the penalty type.

2. Effect of Regularisation

2.1. Experimental Setup

Table 1 shows a summary of the experimental setup. Three fairness metrics were analysed alongside accuracy. *Sex* was used as the protected attribute in each experiment (common to all three of AIF360's out-of-box datasets), with the value *Male* defined as privileged and *Female* unprivileged. Each dataset was randomly split into training (70%) and test (30%) partitions and subsequently standardised. For each value of c, the inverse of regularisation strength, a different classifier was trained to generate predicted (binary) class labels. All metrics were computed with AIF360's `ClassificationMetric` class which takes two datasets as input, the test dataset containing the ground-truth labels and the predicted dataset. For the second stage of the experiment, the training datasets were transformed into fairer datasets according the `fit` and `transform` methods in AIF360's `Reweighing` class. Following the transformation, accuracy, alongside the three chosen fairness metrics, were re-calculated with new classifiers for each value of c to enable a before-and-after comparison.

Datasets	Adult, Compas, German
Metrics	Balanced Accuracy Mean Difference Average Odds Difference Equal Opportunity Difference
Classifier	Logistic Regression
Penalty Type	l2
Penalty Strength (c)	$\{0.000001, 0.00001, ..., 100\}$
Fairness Algorithm	Reweighting

Table 1: A summary of the experimental setup.

2.2. Results Pre-Bias Mitigation

Figure 1 shows the result of our experiments for the *Adult*, *Compas* and *German* datasets. We firstly note that the metrics respond differently to a changing regularisation strength depending on the dataset. For the Adult dataset, a lower value of c (higher strength) leads to increased accuracy, but at the expense of less fair predictions. For Compas, we also observe a negative relationship between fairness and accuracy, but the effect of c, particularly at small values, is reversed: strong regularisation leads to poorer accuracy but a fairer outcome. The picture is less clear for German, the smallest dataset of the three (700 examples).

Adult Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.803	10^{-5}
Mean Difference	0.192	10^{0}
Avg. Odds Difference	0.097	10^{-1}
Equal Opp. Difference	0.107	10^{-1}
Compas Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.666	10^{-2}
Mean Difference	0.268	10^{-5}
Avg. Odds Difference	0.258	10^{-5}
Equal Opp. Difference	0.232	10^{-5}
German Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.712	10^{1}
Mean Difference	0.032	10^{-3}
Avg. Odds Difference	0.045	10^{-3}
Equal Opp. Difference	0.003	10^{0}

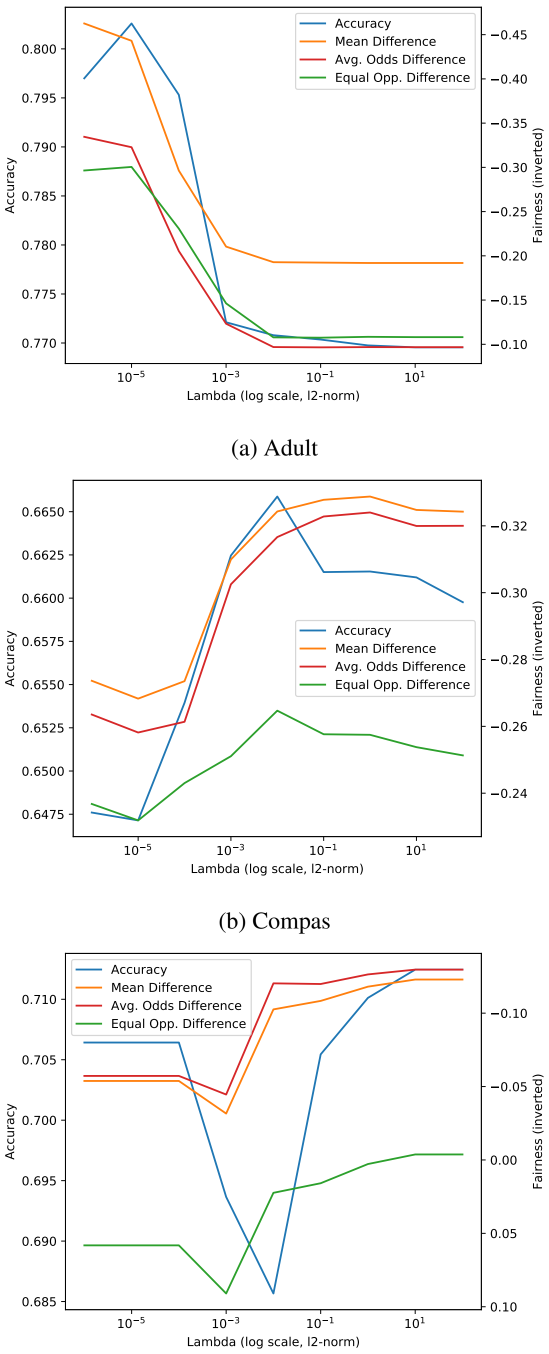
Table 2: For each dataset, we show the 'best' value achieved on the test partition and the corresponding (inverse) regularisation strength, c. For accuracy, the best value is the maximum value. For fairness, a value of 0 implies both groups have equal benefit, hence, we take the absolute value and define the minimum as the best.

To highlight the accuracy-versus-fairness tradeoff, Table 2 shows, for all possible values of c, the best value of each metric achieved alongside the corresponding amount of regularisation. As is also clear from figure 1, the same value of c never leads to an optimum in both accuracy and fairness which has implications for model selection.

To summarise, the results from the Adult and Compas datasets suggest that the value of c that optimises accuracy also worsens fairness, particularly for Mean Difference and Average Odds Difference, and less so for Equal Opportunity Difference.

2.3. Results Post-Bias Mitigation

Figure 2 shows the result of a repeat of the regularisation experiment after having applied a bias mitigation algorithm, specifically, the re-weighting algorithm. The re-weighting algorithm is a type of pre-processing algorithm that ensures



(a) Adult

(b) Compas

(c) German

Figure 1: The effect of c on accuracy and fairness (inverted) pre-bias mitigation.

Adult Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.802	10^{-6}
Mean Difference	0.088	10^{-2}
Avg. Odds Difference	0.038	10^{-4}
Equal Opp. Difference	0.001	10^{-6}
Compas Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.666	10^{-2}
Mean Difference	0.076	10^{-6}
Avg. Odds Difference	0.064	10^{-5}
Equal Opp. Difference	0.044	10^{-1}
German Dataset		
Metric	Best	Argbest ($c \in C$)
Balanced Accuracy	0.712	10^{-6}
Mean Difference	0.018	10^{-6}
Avg. Odds Difference	0.012	10^{-6}
Equal Opp. Difference	0.022	10^{-1}

Table 3: The best metric value achieved post-bias mitigation and the corresponding (inverse) regularisation strength, c.

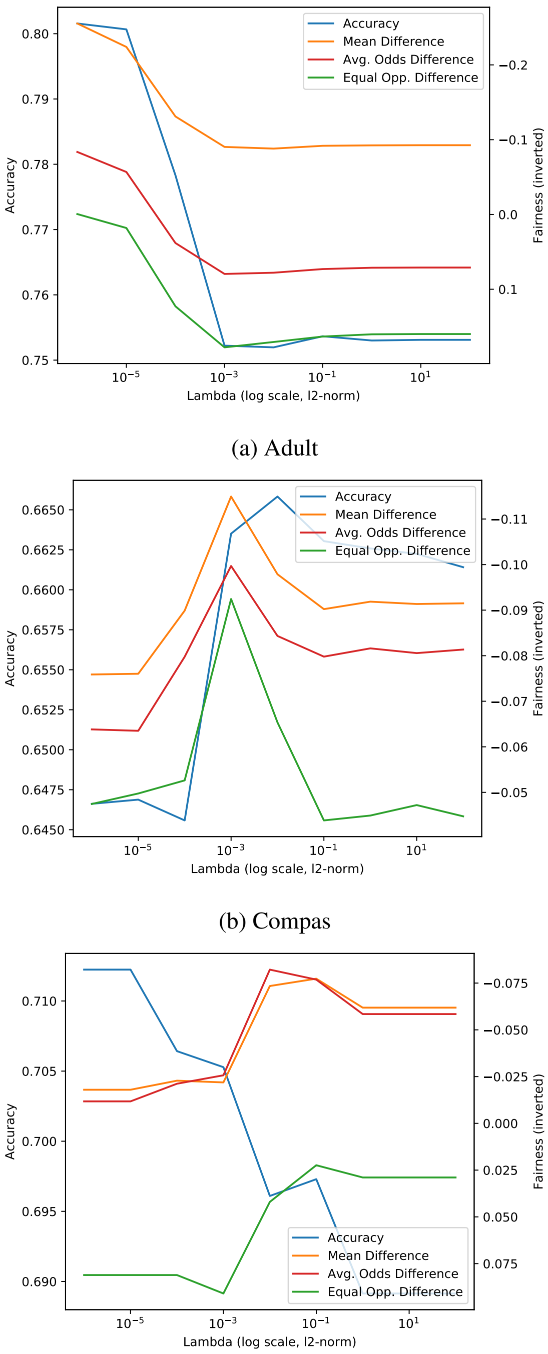
fairness before classification by altering the training data, so that is becomes bias-free [2]. This is in contrast to in-processing, or post-processing algorithms. As anticipated, the predictions are less biased as measured by the fairness curves (they span a narrower range of values). However, in some cases, bias has been introduced in the other direction, i.e., towards the unprivileged group. This highlights the fact that re-weighting does indeed lead to fairer predictions, but it is not possible to guarantee perfect fairness. In terms of the effect of c post-bias mitigation, it remains the case that, with a particular value of c, we are not, in general, able to precisely optimise for both accuracy and fairness at the same time. The exception is with the German dataset, where strong regularisation ($c = 0.000001$) leads to the best outcome for accuracy and two of the three fairness metrics analysed, Mean Difference and Average Odds Difference. This can be seen both in Figure 2 and in the table of optima post-bias mitigation (Table 3).

3. Model Selection

In an attempt to build a fair classification pipeline, we performed model selection over all possible classifiers (one for each value of c), repeating the experiment on each of AIF360's included datasets.

3.1. Experimental Procedure

We used the same setup as in Ttable 1, but further split the non-test partition into training (70%) and validation (30%) datasets. After applying the re-weighting algorithm, we cycled through each value of c, trained a new classifier on the de-biased examples, and chose the best classifier according the both the accuracy and fairness of the predictions on the validation dataset. This best classifer was subsequently used to generate predictions on the held-out dataset.



(a) Adult

(b) Compas

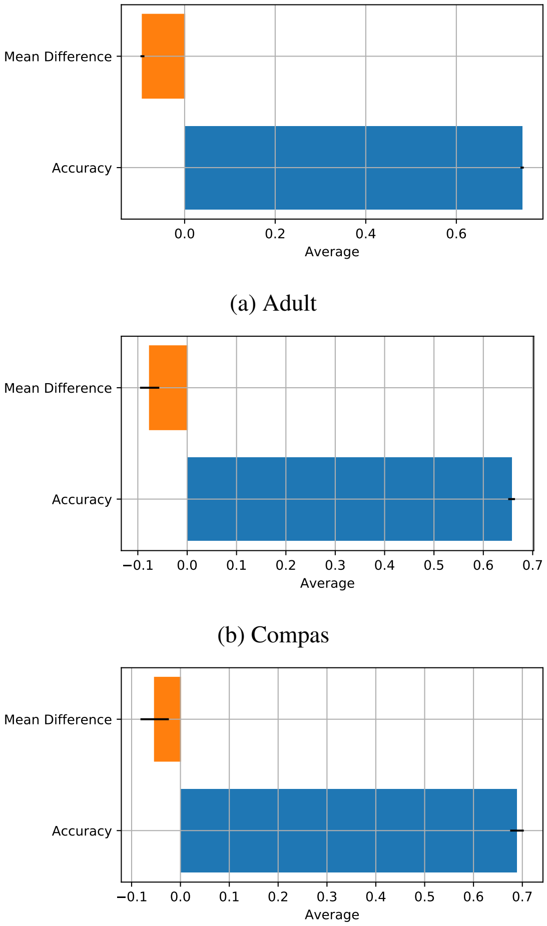
(c) German

Figure 2: The effect of c on accuracy and fairness (inverted) post-bias mitigation.

We repeated this for 5 random training/validation/test splits.

3.2. Objective Function

In order to select the best model, we used a simple weighted function of the accuracy and a fairness measure to generate an overall score. We used accuracy as-is, but for the fairness metric we took the inverse of the absolute value in order to favour values closer to 0 (equal benefit to both groups). We applied an equal weight to both metrics to encapsulate a balance between accuracy and fairness.



(a) Adult

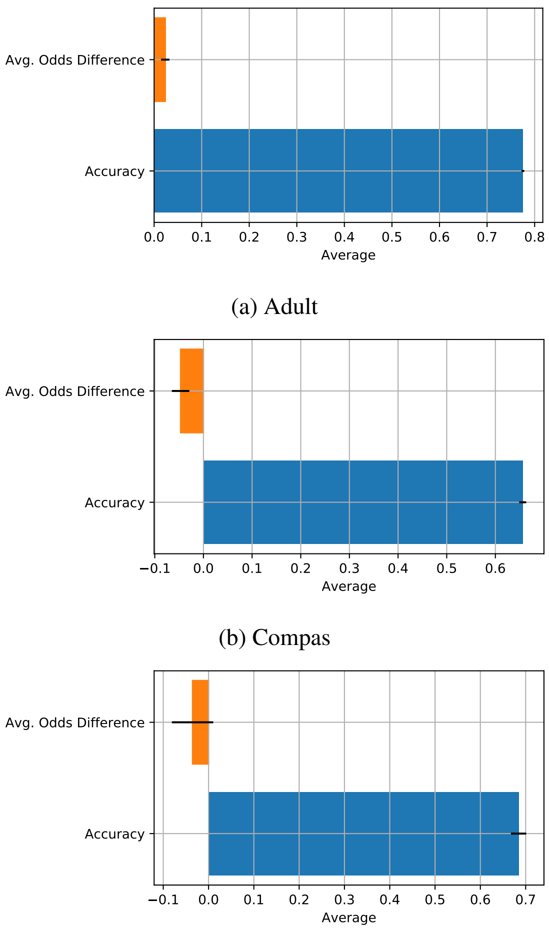
(b) Compas

(c) German

Figure 3: Average Mean Difference and accuracy (with experimental error) evaluated on the test set after performing model selection.

3.3. Results & Discussion

Figures 3-5 show the average accuracy and fairness on the held-out dataset after model selection. For each experiment, we evaluated fairness using the same metric that was included in the objective function. In almost all cases, the results show a significant reduction in discrimination with



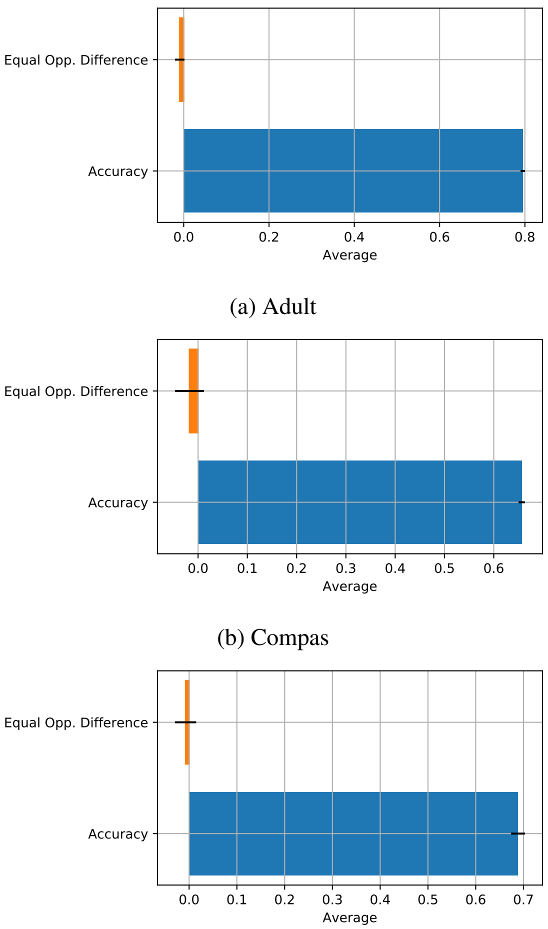
(a) Adult

(b) Compas

(c) German

Figure 4: Average Average Odds Difference and and accuracy (with experimental error) evaluated on the test set after performing model selection.

only minor cost ($1-2\%$) to accuracy when comparing accuracy to its optimum in Table 3. In only one case (Mean Difference, Adult) do we observe a notable (5%) drop in accuracy. This leads us to conclude that it is indeed possible to come up with a fair classification regime if we are willing to accept a small drop in accuracy. While we don't present their exact values, we also observe a good if not perfect agreement in the the regularisation strength, c, that led to the best model during each random partitioning. This leads to our second key conclusion that, without the flexibility in being able to vary the regularisation strength, we would find it significantly harder to find a model that balances both accuracy and fairness. Furthermore, as our analysis reveals, the particular strength cannot be anticipated in advance as it is both metric- and dataset-dependent.



(a) Adult

(b) Compas

(c) German

Figure 5: Average Equal Opportunity Difference and accuracy (with experimental error) evaluated on the test set after performing model selection.

4. Remarks on Penalty Type

Throughout the analysis we used l2-regularisation as the penalty type, also known as *ridge* (logistic) regression. As we've demonstrated, l2-regularisation helps prevent overfitting, but can also lead to fairer predictions, given the right calibration of c (model complexity). Instead of adding the sum-of-squares of the vector of weight coefficients to the data error, as in l2-regularisation, it is possible to generalise this second error term, E_W, to any p-norm [3]:

$$E_W = c\sum_{j=1}^{M} |w_j|^p. \tag{1}$$

With this in mind, we experimented with the effect of l1-regularisation. However, we weren't able fully generalise our model selection process (and objective function).

Firstly, we had to heavily restrict the range of values of c, with the strongest strength possible $c = 0.1$ for Adult and German and $c = 0.01$ for Compas. Otherwise, the selection process would select the model with perfectly fair but random predictions. Secondly, we could replicate (but not improve upon) the results for Adult and Compas in terms of accuracy and fairness. However, for German, l1-regularisation led to reduced accuracy versus the l2-based solution (similar fairness). Thirdly, model selection was unable to find one setting of c in each random splitting of the data. Lastly, for large datasets such as Adult ($31,655$ examples), l1-regularisation drastically increased the required computation time (from seconds to minutes). For all these reasons, if regularisation is to be used to satisfy both accuracy and fairness objectives, we believe that the simplier l2-regularisation is a more robust approach.

5. Conclusion

In this report we have performed a detailed analysis of the effect of regularisation on both fairness and accuracy using the AIF360 library. Our key conclusion is that, with the right regularisation strength, it is possible to optimise for both accuracy and fairness after applying a pre-processing bias mitigation algorithm. Our natural next steps would be to look at whether in-processing and post-processing algorithms are also relevant in assessing the effect of regularisation, and perhaps how AIF360 compares to other commercial fairness toolkits.

References

[1] R. K. E. Bellamy et al. 2018. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. arXiv:1810.01943. 1

[2] T. Calders, F. Kamiran. 2012. *Data preprocessing techniques for classification without discrimination*. Knowledge and Information Systems. 33(1):1–33. 2

[3] C Bishop. 2006. *Pattern Recognition and Machine Learning*. New York: Springer. 4