Strategies Selection for a Fair Classification in Logistic Regression: A Comparative Analysis

Murilo V. Pinheiro[1], **Maria de Lourdes M. Silva**[1], **Javam C. Machado**[1]

[1]Department of Computer Science – Federal University of Ceará (UFC)
CEP 60.440-900 – Fortaleza – CE – Brazil

{murilo.pinheiro, malu.maia, javam.machado}@lsbd.ufc.br

Abstract. *The increasing use of technology leads society to a new concern: the use of machine learning models on personal data and the potentially biased classification. A new definition, called* ***fairness****, emerged to mitigate and combat discrimination in algorithms. Fairness literature includes several techniques to guarantee fair outputs for different demographic groups. There are three phases where an algorithm can achieve fairness. We explore some methods of each stage to construct a comparative analysis that evaluates fairness and utility metrics. Our analysis aims to understand the many ways to achieve fairness using logistic regression in the three most popular datasets in fairness literature. We include several experiments to compare five fairness techniques and select the best for each application.*

1. Introduction

Several companies use technology for decision-making, optimizing time and resources to solve problems. However, using algorithms to categorize individuals may bring social concerns about the process. Machine Learning algorithms can propagate discrimination caused by unfair correlations between sensitive information and the classification. In addition, those algorithms can also generate bias when the input data has unrepresentative samples or even by its design choices.

An effect of algorithmic unfairness is enhancing discrimination based on specific personal characteristics, called *protected attributes*. It refers to features protected by law from discrimination or harassment, such as gender or ethnicity [Government 2015]. The classification models have to attend fairness to guarantee that protected groups are free of discrimination. A model can achieve non-discrimination in three degrees: modifying training data, algorithm design, or algorithm outputs [Pitoura et al. 2021].

We address solving the fairness problem for the logistic regression, a classic classification model and compare several methods free of discrimination. We aim to determine the best strategies to achieve fairness in logistic regression.

In literature, numerous works address fairness. The most significant to our research are summarized as follows. Feldman et al. design an approach that repairs the non-protected features and returns the transformed data [Feldman et al. 2015]. Similarly, Kamiran and Calders propose a method that modifies the data but uses a different technique. Their approach weights each instance, depending on the combinations of group and label, and returns the weighted data [Kamiran and Calders 2012].

On the other hand, some works propose modifications in existing Machine Learning models to ensure fairness. For instance, Kamishima et al, add a regularization term to the objective function to remove the bias in the model predictions [Kamishima et al. 2012]. Narasimhan designed an approach that optimizes a model based on a fairness constraint [Narasimhan 2018]. To ensure fairness after modifying the model's outcomes, Kamiran et al. propose a method that estimates the best threshold of a given output [Kamiran et al. 2012].

Using the logistic regression model, we propose a comparative analysis that selects the best strategies for a fair classification. Our analytical experiments consider the three most popular datasets in literature [Fabris et al. 2022] and compare the fairness methods to define the best for each application.

2. Background

2.1. Logistic Regression Classifier

The logistic regression model is a Machine Learning algorithm for classification analysis. It is a supervised algorithm that seeks the probability of an event occurrence. Machine Learning literature generally uses logistic regression for binary classification but also can adjust the model for multiclass problems. Linear regression is the basis for logistic regression, a statistical method that aims to find the optimal coefficients of a linear function. The optimization considers the linear relationship between variables and the target and produces a numerical output.

The binary logistic regression uses a non-linear function, logistic or sigmoid, to adapt the optimization phase of this algorithm. It maps the linear regression outputs into values between 0 and 1, representing the probability p of a tuple x belonging to a class. Mathematically, $p(x) : \mathbb{R} \rightarrow [0, 1]$. Finally, the model uses the probabilities functions and a given threshold value to predict, for values greater or equal to the threshold are classified into a group, otherwise is another group.

2.2. Group Fairness

Group fairness is one of the various concepts of fairness. It aims to guarantee that all groups of a protected feature have the same probability of being classified in a particular class. This concept is quantifiable using fairness metrics, such as disparate impact or statistical parity difference. In fairness literature, statistical parity is also called independence, group fairness, and demographic parity. Statistical parity requires that the opportunity for individuals of all protected groups to receive positive classifications be similar to the opportunity for the entire population [Dwork et al. 2011].

Considering a dataset of individuals $X = \{x_1, ..., x_n\}$, a protected attribute $A \in \{0, 1\}$, and a classifier prediction $\hat{Y} \in \{0, 1\}$, where each individual x_i has a protected value $a \in A$ and a prediction $\hat{y} \in \hat{Y}$. The statistical parity difference quantifies the treatment difference of a classifier over two groups, a privileged and an unprivileged. The privileged group represents the individuals with $A = 1$ and the unprivileged, the individuals with $A = 0$.

Definition 1 *Given a set of individuals X, a protected attribute A, and the model's prediction $\hat{Y}$, the statistical parity difference measures the divergence between the condi-*

tional probability of $\hat{Y} = 1$ *given the protected attribute* A *of a dataset* $D = (X, A)$.

$$SPD = P(\hat{Y} = 1|A = 1) - P(\hat{Y} = 1|A = 0) \quad (1)$$

2.3. Bias Mitigation Methods

Pitoura et al. defined the three possible phases to apply fairness: pre-processing, in-processing, and post-processing [Pitoura et al. 2021].

Pre-processing methods modify the input data before the model runs. It aims to remove any underlying bias or discrimination in data.

In-processing techniques create new algorithms or update existing classification models to include a step that guarantees fairness. Those procedures generally add fairness constraints or regularization terms to adapt the model's objective function.

Post-processing strategies make no adjustments to the data or the algorithm. They modify the predicted labels, or probabilities outcomes, to ensure fairness classification.

3. Fair Strategies for Logistic Regression

We selected five approaches and grouped them into the methods defined in Subsection 2.3.

3.1. Pre-processing Techniques

Feldman et al. propose the Disparate Impact Remover (DIR), a group fairness identifier and bias remover. Their method creates a repairer that modifies the data to reduce the demographic disparity between sensitive groups. It maintains the internal ranking within a protected group, meaning that individuals holding the highest positions will still retain those positions in the repaired data. However, the method not necessarily preserves the rankings between groups. The approach has a parameter $\lambda \in [0, 1]$ that specifies the amount of repair in data.

Feldman's method is a two-phase algorithm; the first phase calculates the disparity between the demographic groups. In the second phase, it calculates the changes in the data values aiming to reduce the disparity, changing the features to benefit the unprivileged group. The method uses the repair parameter and the median of all given features from each group to minimize the disparity.

Kamiran and Calders also designed a pre-processing technique that weights data to ensure fairness. Their method assigns each possible combination of target and group to a different weight. The approach computes the weights using an estimation that considers the observed and the expected joint probabilities of the target and the protected attribute. This approach is called reweighing (RW). The weights are assigned to each individual, creating a dataset free of discrimination. A model trained with the weighted dataset performs better regarding fairness than the original dataset. However, not all classifiers can incorporate weights in the learning process. The authors demonstrate the method works better in a balanced dataset [Kamiran and Calders 2012].

3.2. In-processing Techniques

Narasimhan formulates the Convex Optimization (CO), an optimization problem that aims to minimize the (utility) loss function while maintaining a fairness constraint. He

considers a fairness parameter proportional to Statistical Parity Difference that imposes a penalty on the model's utility when unfair. This algorithm handles F1-Score and Accuracy in addition to other complex loss functions. It accepts other complex constraints besides fairness. We adapt Narasimhan's method for optimizing the F1-Score under Statistical Parity constraint, and, for parameter tuning, we designate four optimal values explicitly used by the author for evaluation and use 0.1 as relaxation for Statistical Parity Difference [Narasimhan 2018].

Kamishima et al. adapt the regularization idea for the fairness criteria and propose the Prejudice Remover (PR). They modify the logistic regression model with a discrimination-aware regularization term for the learning objective creating a new classifier. Their approach allows controlling the possible trade-off between utility and fairness, adjusting the regularization term, which is proportional to the fairness metric [Kamishima et al. 2012].

3.3. Post-processing techniques

Kamiran et al. design the Reject Option Classifier (ROC), a post-processing approach based on the concept of critic region and probabilities values obtained from probabilistic models, i.e., a model which returns the probabilities of classification. The authors define the critic region as the outcomes probability values bounded by the classification threshold. The individuals in the critic region, called rejected instances, are considered highly uncertain and biased. Their approach classifies rejected instances based on their belonging to the protected group, i.e., if they belong to the protected group, the approach classifies them positively; otherwise, negatively. The technique iterates between the boundaries of the critic region to find the model with a greater utility metric while the fairness metric is below a given value. This method has a lot of parameters, such as the highest and lowest threshold of the critic region, the number of critic regions, and others [Kamiran et al. 2012].

4. Experiments and Results

4.1. Data

For our experiments, we select three real worlds datasets previously used in the free bias classification: Adult Income [Becker and Kohavi 1996], German Credit Risk [Hofmann 1994] and COMPAS [Larson et al. 2016]. We apply a common preprocess for all these three to prepare the data for a Logistic Regression. The datasets represent three areas highly affected by machine learning fairness: creditworthiness, income classification, and criminal recidivism. Table 1 briefly describes the three selected datasets. Each dataset instance/row contains an individual's data, including the protected and target values.

dataset	#rows	#columns	protected	target
Adult	48842	15	gender	income
COMPAS	4744	54	race	recidivism
Credit	1000	20	sex	payer classification

Table 1. Description of the datasets.

4.2. Methodology

Running each method five times and took the mean of the evaluation metrics. We use Statistical Parity Difference to measure fairness and F1 Score to measure utility. A important phase is prepare the data to apply Logistic Regression following the same protocol for the three datasets: (i) removing missing data, (ii) normalizing numerical values, and (iii) encoding categorical values to one-hot encoding.

We split the dataset into 70% train, 15% test, and 15% validation. Then, we selected some fairness parameters for each method defined by the corresponding authors. Finally, after the tuning parameters, the training and validation phase, and the test phase with cross-validation tests, we obtained the results for each method and dataset.

We use the default values defined by each author for the parameters of the methods. We also operate another cross-validation with two different criteria to choose the best parameters between multiple default values defined by the authors.

i. The classifier with higher utility and fairness below a threshold.
ii. The classifier with a greater normalized mean between justice and utility.

Narasimhan's method selects the best parameters using the first method, which the author has defined. In Kamishima's and Kamiran's methods, we choose the best with the second criterion. In the training phase and the final analysis, we evaluate the outcomes using Statistical Parity Difference and F1-Score, a well-known utility metric of classification tasks.

4.3. Results

The Figure 1 shows the F1-Score and the Statistical Parity Difference, where the bar value refers to the mean value, besides the line in the center of the bar, representing the standard deviation, in six different cases as the Baseline Logistic Regression (BLR) without modify the data and the other five defined approaches: Reweighing (RW), Disparate Impact Remover (DIR), Convex Optimization Method (CO), Prejudice Remover (PR) and Reject Class Option (ROC).

Observing the results, the first step is to understand that neither approach surpasses all the others in all cases. The second step is analyze which strategies have worked well in each dataset, showing promising results compared to BLR, with lower unfairness and low damage to utility values; these methods are more independent of data distribution. Those techniques are RW, CO, and PR. The others, ROC and DIR, have bad results in at least one data set, with problems like high decrease in utility or lower fairness gain.

Going through the results, we can observe that the impact of the methods varies according to the dataset, i.e., the effectiveness of a strategy depends on the distribution, size, and data types of a dataset. Looking exclusively at SPD values, all methods increase the fairness in the model, but some of these approaches bring harsh losses to utility in some datasets.

Some approaches perform well, having higher utility loss and lower unfairness values in all datasets compared to the Baseline Logistic Regression, demonstrating the applicability of these methods for general purposes.RW considers the observed probability, so a good result in this method is highly dependent on the data distribution to achieve

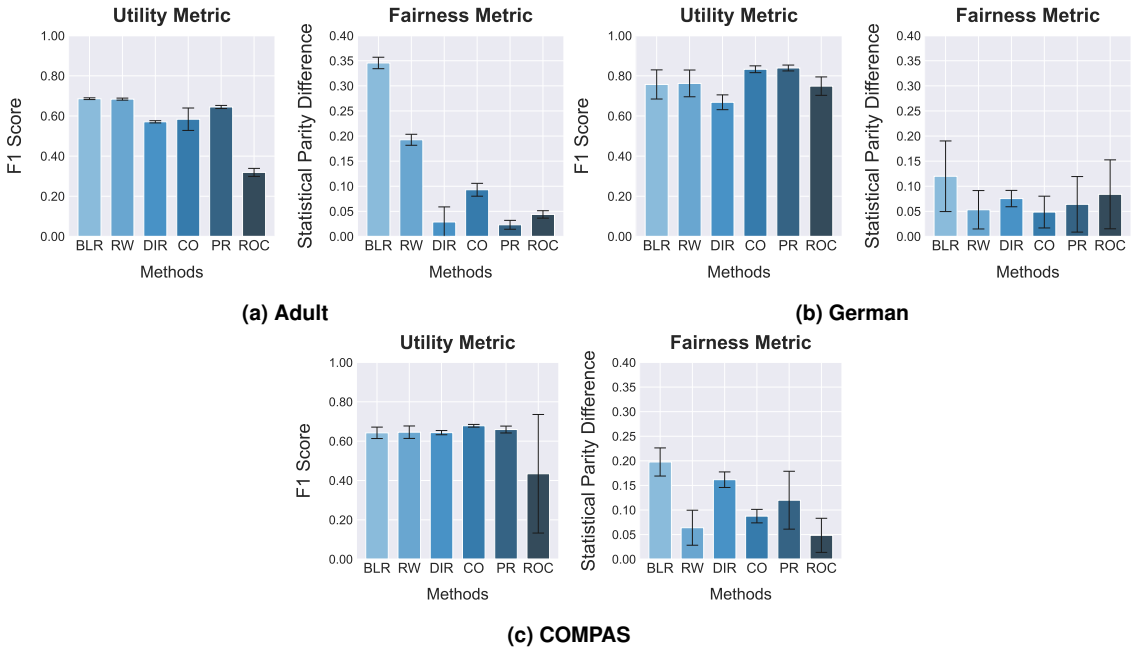

Figure 1. Methods results for each dataset.

good results. The DIR method is the pre-processing technique that needs just numerical features, which explains the drop in the utility values since we need to prune categorical features.

The two in-processing approaches, CO and PR, use an optimizer to find the better result, considering the utility and fairness values. When we look at the two metrics, these two methods achieve better results, bringing a better balance between the two metrics. The post-processing technique ROC had a high negative impact on utility for all datasets, likely due to its blind approach to the data and model; considering just the outcome makes maintaining utility challenging, especially with newly observed data.

Analyzing by dataset, we can define the best approachs for each one. For the Adult dataset the PR method has the best approach for this data without significantly affecting utility and yielding good fairness results. In German, the in-processing techniques perform well, even improving utility when applied. The COMPAS dataset works better with RW and CO, yielding good results for the two metrics.

5. Conclusion

The increasing adoption of automated decision-making has a crescent impact on people's lives. Comprehending the construction of fair classifiers and understanding the costs and gains of this task is the objective of our work. Our results show that preserving justice and equality demonstrates how these methods can work in different behaviors and adapt to the different sets of laws and requirements.

In this comparative study, we find the optimal approach for each dataset in five selected methods for fair classification. With the plurality of good ways to achieve fairness, choosing the best technique depends on available tools and data characteristics. Testing and analyzing different methods are crucial for determining the optimal strategy, considering the non-linear relationship between fairness and utility.

References

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2011). Fairness through awareness. *CoRR*, abs/1104.3913.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152.

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*, pages 259–268.

Government, U. (2015). Equality act 2010: What do i need to know? quick start guide to discrimination by association and perception for voluntary and community oranisations. https://www.gov.uk/guidance/equality-act-2010-guidancehttps://www.gov.uk. Last accessed: May 23, 2023.

Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, pages 924–929. IEEE.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer.

Larson, J., Roswell, M., and Atlidakis, V. (2016). Compas. `https://github.com/propublica/compas-analysis`. July 29, 2022.

Narasimhan, H. (2018). Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654. PMLR.

Pitoura, E., Stefanidis, K., and Koutrika, G. (2021). Fairness in rankings and recommenders: Models, methods and research directions. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2358–2361. IEEE.