Fairness Guarantees Under Demographic Shift

Shengxin Zhang[1a*], Ruiyang Wang[2b], Yushan Li[3c], Jiarui Yang[4d], Ruichao Zhuang[5e]

*Corresponding author's e-mail: [a] sz68@illinois.edu; [b] wry20031124@163.com;
[c] Liysh75@mail2.sysu.edu.cn; [d] scyjy9@nottingham.edu.cn; [e] Rachel-zhuang1@hotmail.com

[1]Grainger College of Engineering, University of Illinois at Urbana Champaign, Champaign, 61820, United States
[2]Department of Mathematics, University College London, London, WC1E 6BT, United Kingdom
[3]School of Software Engineering, Sun Yat-Sen University, Guangdong, Zhuhai, 519000, China
[4]Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo, 315100, China
[5]College Of Liberal Arts, The International Department Affiliated High School of SCNU, Guangzhou, PRC. 510630, China

Abstract: Contemporary research has established that machine learning applications, especially in social contexts, can inadvertently lead to unfair model predictions that manifest as racism, sexism, or discrimination. Such biases stem from factors such as population growth and economic changes. Typically, a model is trained and later deployed to predict relevant problems. However, this approach usually assumes that the training dataset is reflective of the data expected in real-world deployment. As the distribution shift caused by demographic changes remains unknown, Giguere, Metevier et al. utilized a student t-test to compute the upper confidence bound. However, we identified several areas of potential improvement within this algorithm. In this paper, we propose methods to optimize the Shifty Algorithm by enhancing its robustness and refining its loss function. To evaluate the performance of the modified Shifty Algorithm, we used the UCI Adult Census dataset and a real-world dataset on university admissions exams and subsequent student achievement. Through these experiments, we demonstrate that models trained using our method successfully mitigate bias when faced with demographic shifts. Our experimental results validate the robust fairness assurances of our algorithm under real-world conditions. Moreover, they highlight the ability of the new Shifty Algorithm to train models effectively, ensuring fairness in the event of demographic shifts, while making fewer assumptions.

Keywords: Machine Learning, Fairness, Demographic Parity, Sensitive.

1 Introduction

In recent years, the swift evolution of Machine Learning (ML) models has propelled their widespread integration into various aspects of our lives. These potent algorithms are used to perform daily tasks like recommending shopping options, filtering loan applicants, deploying police officers, and informing bail and parole decisions [1]. However, due to issues such as label shift, demographic shift, or covariate shift, these advanced models can sometimes generate unfair predictions that significantly impact us. Therefore, it is crucial to develop ML algorithms that are not just accurate but also impartial and fair, as they currently influence many facets of our lives. In the context of decision-making, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347147

characteristics [2]. Currently, numerous researchers are striving to address the diverse elements contributing to a model's unfair behavior, with demographic shift being a key factor among them. During the model training process, many tend to assume that the distribution used for prediction remains consistent between the training and deployment processes. However, evaluation data often comes from a different distribution than the one on which the model will be deployed, for instance, due to data collection procedures or distributional shifts over time [3]. Furthermore, those involved in model development often adopt a passive stance during the evaluation process, missing an opportunity to apply their knowledge of anticipated distributional changes and identify the shifts they want their model to withstand. Consequently, rectifying unfair prediction in the model's decision-making process has become a trending topic of discussion recently.

Consider, for example, a model that uses the Scholastic Assessment Test (SAT) scores and High School GPA of students to predict their academic success in college. If we apply a dataset from twenty years ago during the training phase, the distribution of the data would be distinctly different from the present, due to demographic shifts among students and the increasing number of applicants over time. This phenomenon, known as demographic shift, significantly impacts the performance of the model since many sensitive attributes (such as sex and race) change. However, the transformation of the distribution is often unknown during training, as we cannot predict future demographics.

To address the issue of fairness guarantee under demographic shift, Giguere et al. introduced an algorithm named Shifty [4]. This algorithm primarily comprises three sections: Data Partitioning, Candidate Selection, and Fairness Testing. However, upon evaluating this algorithm and reproducing the experiment, we identified several potential issues such as complex notation, a complicated loss function, and its restriction to data under normal distribution. To enhance the performance of this algorithm in the context of demographic shift, we proposed several ways to optimize it, including refining its loss function, removing restrictions, or improving its robustness.

Our main contributions are:

1.Optimizing the error function and penalty term used in the Shifty algorithm.

2.Removing the restriction that fairness testing has to be written in conditional events.

3. Enforcing fairness constraints so the model can predict fairly.

2 Background and Related Work

We provide examples of our method for fair categorization, while the techniques we provide can be simply applied to various issue scenarios. A set of features and a corresponding label make up a data instance in this context. When evaluating its fairness, features can be divided into three parts, $X, S, \tilde{S}$. X stands for the features which do not affect the fairness, S stands for the sensible variable, a categorical protected fairness attribute, such as race or sex, $\tilde{S}$ stands for demographic feature which represents difference the distributions of training and deployment data, whose domains are denoted by $\mathcal{X}, \mathcal{S}, \tilde{\mathcal{S}}$ separately, and we denote labels by $Y \in \mathcal{Y}$.

In general, a model, $\theta: \mathcal{X} \to \mathcal{Y}$ is used to predict the label associated with $\mathcal{X}$ without knowing the true label, whose quality is measured by loss function ℓ. We consider that such classification ignores the fairness attribute. Hence, in order to obtain an accurate classifier, we typically choose a training algorithm a that aims to minimize the chosen loss and provide it with a dataset composed by n independent observation samples, $D = \{X_i, Y_i, S_i\}_{i=1}^{n}$, where $P(X_i, Y_i, S_i) \coloneqq P(X, Y, S)$ for all $i \in \{1, \ldots, n\}$.

2.1 Assessment of the fairness of an algorithm

The user specifies a function based on the particular fairness requirements of a given application, g, which accepts a model θ and is adjusted so that $g(\theta) > 0$ if and only if θ behaves unfairly. Here, we consider the illustrative case where

$$g_{DP}(\theta) := \left|\mathrm{E}[\theta(X)|S = s_0] - \mathrm{E}[\theta(X)|S = s_1]\right| - \epsilon_{DP} \quad (1)$$

2.2 Fairness classification under demographic shift

The instance observed during training and instances after the model has been put into use are denoted by $(X, Y, S, \tilde{S})$ and $(X', Y', S', \tilde{S}')$ separately. In order to formalize demographic shift, two following conditions are required:

(1) the demographic attribute may change between training and deployment.

$$\exists \tilde{S} \in \tilde{\mathcal{S}} \quad s.t. \quad P(\tilde{S} = \tilde{s}) \neq P(\tilde{S}' = \tilde{s}) \quad (2)$$

(2) the pre- and post-shift joint distributions over the instances are essentially the same.

$$\forall (x, y, s, \tilde{s}),\ P\left(X = x, Y = y, S = s \,\middle|\, \tilde{S} = \tilde{s}\right) = P\left(X' = x, Y' = y, S' = s \,\middle|\, \tilde{S}' = \tilde{s}\right) \quad (3)$$

2.3 Related Work

Admittedly, the shifty algorithm provides high fairness guarantee under demographic shift. However, it applied complicated procedures which can be simplified for better understanding. One idea that is close to our approach is the prejudice remover regularizer in AIF 360 package where the authors introduce a prejudice remover regularizer directly tries to reduce the prejudice index denoted as R(D, Θ) regularizer to enforce fair classification [5], and it will penalize the model when it behave unfairly. However, in comparison, the prejudice remover regularizer only decreases prejudice in the dataset and does not solve the problem of demographic shift. Another approach that is similar to our approach is the paired-consistency method proposed by Horesh, etc, in 2020 where they proposed a consistency score metric embedded within the loss function as a fairness regularization term, to make the model consistency aware [6]. However, their approach requires access to a human fair domain expert, which is often more inefficient compared with our algorithm which solve the problem using the penalty term.

3 Methodology

Figure 1 shows an overview of shifty_pro. The shifty_pro algorithm inspired by shifty algorithms consists of three core parts: data partitioning, candidate selection, and a fairness test [7]. First, data partitioning divides the data into two parts, which are used separately in

Candidate Selection and in Fairness Test. Next, the purpose of candidate selection is to minimize the loss of models. Once a candidate model is found, Fairness Test preforms based on the rest of the data and fairness specifications given by the user to return fair candidate model or no solution found.

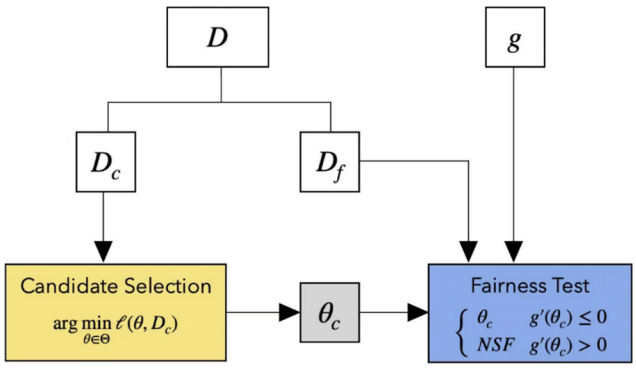

Figure 1. Shifty_pro accepts training data, D, which is separated into D_c and D_f. D_c is used to select a candidate model, θ_c. Then, D_f and definitions of behaviors, g, are used in Fairness Test to guarantee the fairness of θ_c after deployment.

3.1 Candidate Selection

We perform candidate selection by minimizing the loss consisting of two terms: one estimates the worst-case classification error on the deployment distribution, and another penalizes models when unfair behaviors appear.

$$\ell(g, D_c, \theta) := \text{Error}(D_c, \theta) + \text{Penalty}(\theta, \lambda) \quad (4)$$

$\text{Error}(D_c, \theta)$ estimates the worth case classification error, where indicator function, $\mathbb{I}[\cdot]$, returns 1 if its argument is TRUE and 0 otherwise, and $\phi(\tilde{s})$ acts as a reweighting scaling factor.

$$\text{Error}(D_c, \theta) := \frac{1}{|D_c|}\sum_{(x,y,\tilde{s})\in D_c} \mathbb{I}[\theta(x) \neq y]\phi(\tilde{s}) \quad \text{where} \quad \phi(\tilde{s}) := \frac{P(\tilde{S}'=\tilde{s}')}{P(\tilde{S}=\tilde{s})} \quad (5)$$

$\text{Penalty}(\theta, \lambda)$ penalizes models which behave unfairly after deployment, where $b(\theta)$ is the regularization term, and λ is Used to control the strength of regularization.

$$\text{Penalty}(\theta, \lambda) := \lambda b(\theta) \text{ where } b(\theta) := \max(0, g(\theta)) \quad (6)$$

3.2 Fairness Test

Since the fairness of the model, $g(\theta)$, depends on X, Y and S, the fairness guarantees based on g may fail after the model is deployed. We denote the unfair behavior after deployment of the candidate model by $g'(\theta_c)$, which depends on X', Y' and S'. By the fairness definition, if its value is below zero, shifty_pro returns θ_c, and otherwise returns NSF.

4 Evaluation

In line with our recent modifications to Shifty algorithm's candidate selection process, we have evaluated our method's performance regarding fairness constraints, applying a new penalty term and error function. Specifically, we have transitioned from the $max(0, sup\ q \in Q\ \hat{U}ttest(g, Dc, \theta, \delta;\ q))$ penalty term to one defined by $\lambda * b(\theta)$ where λ represents a hyperparameter that is similar to regularization term that decide the strength of regularization. Meanwhile, we also simplified the error function so it no longer calculates the infimum of the loss function. To verify our proposed improvements, we have conducted a variety of tests on different datasets, aiming to establish whether our model performs more efficiently using the revised loss function.

Our findings primarily address three crucial research questions (RQ):

(RQ1) In practical applications, do models trained using the new loss function adhere to robust fairness guarantees under demographic shifts, compared to those trained using previous methodologies?

(RQ2) Can this updated loss function incorporating this regularization term be utilized to train models with accuracy levels comparable to those achieved using prior techniques that do not account for demographic shifts?

While implementing a dataset is a relatively straightforward task, answering the research questions above necessitates the identification and utilization of datasets that consistently demonstrate a demographic shift extractable from the features. Consequently, in our experiments, we relied primarily on previously used dataset like famous UCI adult dataset. Furthermore, we undertook meticulous data cleaning to ensure their appropriateness for both the training and testing phases.

UCI Adult Census Dataset: The UCI Adult Census Dataset is composed of data of 48,842 individuals in the 1884 census [1], along with various features including sensitive attributes like race and sex. Hence, we choose this dataset and set race as our sensitive attribute while the sex as the demographic attribute in order to assess the fairness under demographic shift in the data partitioning process. After doing some data cleaning, we set the label y hat as the income level of whether an individual has annual income of over $50,000 and start our multilayer perceptrons classifier training to predict.

To resolve the research question mentioned above, our group generates training and deployment datasets by resampling using bootstrap sampling method from a fixed population while keeping either training set or test set to include the demographic variable. This ensures the handling of imbalance data between sensitive groups, enhances the model's robustness through generating various train and validation sets, and assessing how well the fairness constraints generalize under different data distribution. In addition, since we are curious about the performance of the new penalty term on its accuracy, we conducted an experiment on whether the new penalty term led to a more accurate and fair result. Moreover, considering the hyperparameter in this algorithm, we mainly utilized the Grid Search Optimization method to tune our parameter so it has the lowest loss value and demographic parity metric close to zero which means all samples have the same selection rates.

5 Results

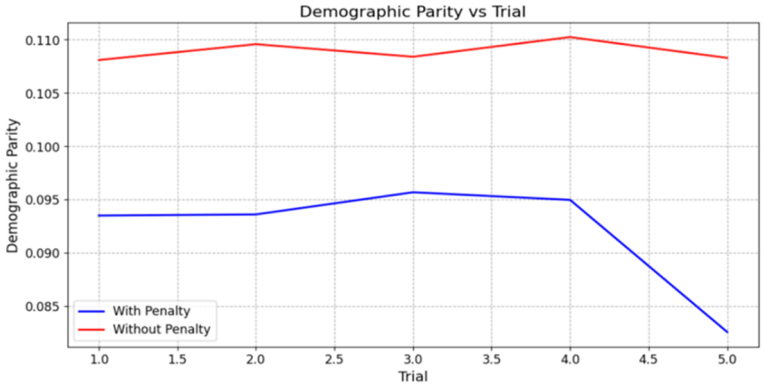

Figure 2: Model's Demographic Parity Graph where blue line represent loss function with penalty and red line represent loss function without penalty where Demographic Parity of 0 represent all group have the same selection rate.

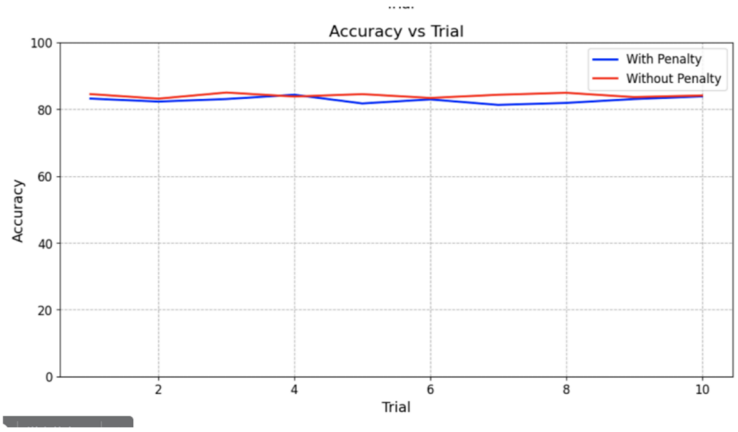

Figure 3: Model's Accuracy Graph where blue line represent loss function with penalty and red line represent loss function without penalty.

Our results indicate that updating the penalty term significantly improves the accuracy and fairness of the Multi-Layer Perceptron (MLP) model when it comes to adjusting the lambda and learning rate parameters(see Figure 2). In our experiment, we primarily used the demographic parity metric as our g(θ). However, users have the flexibility to choose from a range of other options based on their specific interests and requirements. Options include metrics such as equalized odds, equal opportunity, or disparate impact, each of which assesses fairness from a unique perspective and may be more suited to particular contexts or scenarios.

According to the results depicted in the figure 3, there is a noticeable difference between models that incorporate a penalty term in their loss function and those that do not. The models with a penalty term show a better balance between fairness and accuracy. This is likely due to the penalty term nudging the model towards making predictions that are not only accurate but

also fair. It helps the model recognize and minimize the impact of certain features that might lead to unfair predictions, thereby ensuring a fairer outcome.

Models without a penalty term, however, show a marked decline in fairness, even though they may achieve a high level of accuracy. This is because without the penalty term, these models optimize purely for accuracy and may overemphasize certain features that introduce bias. Additionally, most of the time there is a trade-off as indicated that “if there is a difference in separability (accuracy of the best classifiers) on two groups of people, any attempt to change the classifiers to attain fairness can affect the accuracy for one or both the groups” [8]. Consequently, while these models may perform well in terms of raw predictive power, they risk making unfair predictions by disproportionately affecting certain demographic groups. Hence, the introduction of a penalty term into the model's loss function is a critical step towards achieving fair machine learning outcomes. It ensures that the model does not overly prioritize accuracy at the expense of fairness, which is especially important in socially sensitive applications where unjust predictions could have significant repercussions.

6 Conclusion

In this paper, we develop shifty-pro algorithm, which modify Shifty algorithm’s candidate selection process, simplify the error function, and optimize the penalty term. We successfully reduce the strictness of fairness metrics requirements of the shifty algorithm, which is supposed to be a normal distribution, and expanding the range of the metric while improving accuracy and fairness. Except for our contribution, there are various ideas to further enhance the robustness, and performance of the algorithm. Firstly, according to the author, the Shifty algorithm does not assure fair model prediction when users are enforcing individual fairness constraints. However, to resolve this issue, we can refer to the idea of which apply Lapacian transformation to optimize our model, ensuring both group fairness and individual fairness [9]. Graph Laplacian can be used for post-processing in algorithmic fairness to address individual fairness concerns. It involves casting the post-processing problem as a graph smoothing problem with graph Laplacian regularization, which preserves the goal of treating similar individuals similarly. We propose using graph Laplacian before the candidate training step as a future direction for optimization, aiming to enhance individual fairness alongside group fairness.

Besides, we observe that the current algorithm of demographic shift is limited, which is only apply for single feature such as gender or race, and it may not be sufficient to address fairness concerns in scenarios involving marginal shifts or covariate shifts. Therefore, a crucial direction for future research is to investigate alternative methods to ascertain fairness when faced with these complex scenarios. One promising aspect for further exploration is the application of Group-Dro, a novel fairness evaluation technique that accounts for the effects of marginal shifts and covariate shifts. By incorporating Group-Dro into our framework when computing penalty and loss function, we can potentially achieve a more comprehensive and robust fairness assessment in a broader range of real-world situations. Another is a novel adversarial approach for seeking fair decision making called Batch Fair Robust Log-Loss learning under covariate shift [10]. This model builds on robust classification method of under covariate shift, where the target distribution is estimated by a worst-case adversary that

maximizes the log-loss while matching the feature statistics under source distribution. Therefore, if the separating set of features is known, it can incorporate them as constraints for the adversary. It's a possible way to combine this algorithm with our fairness test to deal with the scenarios under covariate shift.

References

[1] Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82-89.

[2] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

[3] Sun, X., Wu, B., Zheng, X., Liu, C., Chen, W., Qin, T., & Liu, T. Y. (2021). Recovering latent causal factor for generalization to distributional shifts. Advances in Neural Information Processing Systems, 34, 16846-16859.

[4] Giguere, S., Metevier, B., Brun, Y., da Silva, B. C., Thomas, P. S., & Niekum, S. (2022, April). Fairness guarantees under demographic shift. In Proceedings of the 10th International Conference on Learning Representations (ICLR).

[5] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2012. Lecture Notes in Computer Science(), vol 7524. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33486-3_3.

[6] Horesh, Y., Haas, N., Mishraky, E., Resheff, Y. S., & Meir Lador, S. (2020). Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning. In Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I (pp. 590-604). Springer International Publishing.

[7] Giguere, S., Metevier, B., Brun, Y., da Silva, B. C., Thomas, P. S., & Niekum, S. (2022, April). Fairness guarantees under demographic shift. In Proceedings of the 10th International Conference on Learning Representations (ICLR).

[8] Dutta, S., Wei, D., Yueksel, H., Chen, P. Y., Liu, S., & Varshney, K. (2020, November). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In International conference on machine learning (pp. 2803-2813). PMLR.

[9] Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for individual fairness. Advances in Neural Information Processing Systems, 34, 25944-25955.

[10] Rezaei, A., Liu, A., Memarrast, O., & Ziebart, B. D. (2021, May). Robust fairness under covariate shift. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 11, pp. 9419-9427).