Fair Classification with Counterfactual Learning

Maryam Tavakol
Technical University of Dortmund, Dortmund, Germany
maryam.tavakol@tu-dortmund.de

ABSTRACT

Recent advances in machine learning have led to emerging new approaches to deal with different kinds of biases that exist in the data. On the one hand, counterfactual learning copes with biases in the policy used for sampling (or logging) the data in order to evaluate and learn new policies. On the other hand, fairness-aware learning aims at learning fair models to avoid discrimination against certain individuals or groups. In this paper, we design a counterfactual framework to model fairness-aware learning which benefits from counterfactual reasoning to achieve more fair decision support systems. We utilize a definition of fairness to determine the bandit feedback in the counterfactual setting that learns a classification strategy from the offline data, and balances classification performance versus fairness measure. In the experiments, we demonstrate that a counterfactual setting can be perfectly exerted to learn fair models with competitive results compared to a well-known baseline system.

KEYWORDS

Fairness-aware learning; counterfactual reasoning; classification

ACM Reference Format:
Maryam Tavakol. 2020. Fair Classification with Counterfactual Learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401291

1 INTRODUCTION

Often, machine learning methods highly depend on factual reasoning, which means that the existent observations are considered the facts. The collected data, the sampling (or logging) policy, the environmental constraints, and many other factors are the main components for learning various models. However, these conditions bring different kinds of biases with themselves into the frameworks, that effectively shape the resulting models learned from these scenarios [16]. One of the main concerns in recent artificial intelligence research is that the data-driven approaches preserve the unfairness available in the collected/offline data in the resulting models. For instance, if historically, gender was a compelling factor to indicate having a lower or higher income, this effect still has a significant role in designing new decision support systems. Therefore, fairness-aware learning has emerged to eliminate this effect by taking different measures of fairness into the optimization process. These measures are mostly defined based on sensitive attributes existing in the data and aim at balancing the decisions made for protected and non-protected groups.

On the other hand, counterfactual methods are designed to learn unbiased policies from logged bandit data via counterfactual reasoning. Counterfactual reasoning is introduced for evaluation and learning from offline data, which takes into account the conditions that could have happened if the data was created, sampled, or labeled differently [4, 18]. In these scenarios, only the partial labels (aka bandit feedback) are available, that means, the feedback is solely provided for the chosen policy/decision at the time of sampling. Hence, counterfactual learning aims to model all the other policies using counterfactual reasoning which leads to learning unbiased policies from sampled data.

Therefore, both concepts (fairness-aware and counterfactual learning) are correlated in terms of removing biases from the available data. Fairness is concerned about biases in the data against minority groups, and counterfactual methods learn unbiased policies from sampled data. Intuitively, we draw a remarkable connection between two concepts by assuming that the sampling policy in the counterfactual setting is equivalent to the unfair decisions in the fairness setting. As a result, we are able to model fairness-aware learning in a counterfactual framework, and show how counterfactual learning can move from the unfair decisions in the data toward learning models that are more fair.

In this paper, we model fairness-aware multi-class classification problems in a counterfactual setting, using a definition of fairness based on **equalized odds**, that can be easily extended to other definitions of fairness. Equalized odds [7] is an indicator to determine an equal opportunity for various groups by considering the difference of true classified instances between protected and non-protected groups in all classes. We further utilize this measure to turn it into a reward function for the optimization problem in the counterfactual framework. Empirically, we illustrate that our approach is perfectly able to balance fairness versus AUC (Area Under the ROC Curve) on a real-world dataset.

2 BACKGROUND

2.1 Fairness-Aware Learning

Generally speaking, fairness is defined as the absence of any discrimination against individuals or groups. Mehrabi et al. [16] enumerate more than twenty different biases from several perspectives which result into variant definitions and/or formulations for fairness, from disparate treatment and disparate impact [3], to equalized odds and counterfactual fairness [7, 14]. These definitions are employed in several tasks for fairness-aware learning that among them, we aim at addressing the classification problems.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401291

One of the first attempts to have a fair classifier employs a regularization approach in the logistic regression method for multiple sources of unfairness [11]. Some classifiers are designed to only satisfy a certain definition of fairness [6], while other methods tend to provide a more general framework for discrimination-free classification [1]. Zafar et al. [20] present an optimization approach to maximize the classification performance subject to a fairness constraint and vice versa which leads to a trade-off between accuracy and fairness. Building upon that, other methods as well integrate a fairness measure into the optimization problem such as boosted trees [9] and adversarial neural networks [15]. Moreover, we would like to draw a line between "counterfactual fairness" and fair learning in "counterfactual setting" to avoid confusion. Counterfactual fairness [14] is one definition of fairness that considers the same decision for actual as well as counterfactual situations [12, 13]. However, counterfactual setting provides a learning framework to learn and evaluate unbiased policies from logged data that we employ in this paper for fairness-aware learning.

2.2 Counterfactual Learning

Counterfactual reasoning is introduced as a means of learning from logged bandit feedback [4, 18], and has been studied in interactive systems such as recommendation. The main problem in these systems is that the data is collected with a sampling (or logging) policy and learning any new policy would be highly biased toward that policy. The line of research in this domain is mainly based on inverse propensity scoring [8], where the samples are re-weighted according to the relation of an actual policy to the sampling policy. In the recent years, several counterfactual estimators have been introduced for off-policy evaluation of new policies that also cope with the bias-variance problem. Some examples include direct model [5], doubly robust estimators [5], self-normalized [19], and so on. Most recently, Su et al. [17] have introduced a general family of estimators in the contextual bandit setting, which trades-off the bias versus variance, and any of the above estimators can be instantiated from that. Additionally, due to its differentiable property, this estimator can be used in gradient-based learning algorithms like POEM [18] for learning the optimal policy. POEM is an efficient algorithm for structured output prediction, in which, predictions are characterized using the linear function of a joint feature space of the policy. Alternatively, Kallus [10] brings in a new algorithm that optimizes the policy and weights simultaneously using balancing methods from the causal inference. Nonetheless, their algorithm is computationally expensive, we thus benefit from POEM algorithm.

3 FAIR CLASSIFICATION MODEL

Counterfactual methods are a reliable technique to remove the decision biases from the logged data in order to learn impartial policies. Additionally, many available datasets that are used to design new decision support systems suffer from discrimination against certain individuals or groups. Therefore, we connect the two concepts, and in this section, present a counterfactual framework for fairness-aware learning which can be effectively employed in real-world scenarios.

3.1 Preliminaries

We begin by formulating the problem of fair classification in which the resulting model is impartial toward various groups. We render a binary classification task that is easily extendable to multi-class classification problems. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a feature vector of size d, $y \in \{-1, 1\}$ be the corresponding class label, and for every sample, s be an additional *sensitive* attribute such as gender or race that we consider is binary for simplicity, i.e., $s \in \{0, 1\}$. As a result, the collection of n i.i.d. samples form a training set of $\{(\boldsymbol{x}_i, s_i, y_i)\}_{i=1}^n$.

Standard approaches do not distinguish between the sensitive attribute and all the other attributes and consider the entire feature vector as $\bar{\boldsymbol{x}} = [\boldsymbol{x}, s]$. Hence, any classification algorithm can be applied on the data $\{(\bar{\boldsymbol{x}}_i, y_i)\}_{i=1}^n$ in order to optimize a performance measure such as accuracy or AUC. However, the obtained models are highly biased toward the information existing in the data which is discriminative. For instance, if historically, women had a lower hiring rate than men, the resulting models would be still unfair against women. We thus eliminate the sensitive attribute from the samples and design a model that optimizes a measure of classification performance as well as a measure of fairness.

In this paper, we use the definition of **equalized odds** for evaluating fairness, and we show later in this section that any other definition of fairness is applicable in our framework. Equalized odds [7] declares that both protected and non-protected groups should have equal true positive rates and false positive rates. In a formal notation, a prediction $\hat{y}$ satisfies equalized odds if

$$P(\hat{y} = 1|s = 0, y) = P(\hat{y} = 1|s = 1, y), \quad y \in \{0, 1\}, \tag{1}$$

where in terms of optimization problem, we are interested in maximizing their ratio and a ratio of one is the optimal. Therefore, we aim at finding a function f that maximizes the described fairness measure as follows

$$\max_f \quad \min \left(\frac{P(f(\boldsymbol{x}) = 1|s = 0, y)}{P(f(\boldsymbol{x}) = 1|s = 1, y)}, \frac{P(f(\boldsymbol{x}) = 1|s = 1, y)}{P(f(\boldsymbol{x}) = 1|s = 0, y)} \right). \tag{2}$$

Note that $P(f(\boldsymbol{x}) = 1|s = 1, y)$ and $P(f(\boldsymbol{x}) = 1|s = 0, y)$ are the positive rates for the protected and non-protected groups, respectively. In addition, we still need to deal with the classification task which can be only evaluated from offline data. let L be a chosen loss function for the classification task, function f should also optimize the following objective function

$$\min_f \quad \sum_{i=1}^n L\left(y_i, f(\boldsymbol{x}_i)\right). \tag{3}$$

Given these two objective functions, we propose a novel framework which optimizes both Equations (2) and (3) to balance fairness measure versus classification performance without combining them in a single objective as [20] does. Therefore, our approach is flexible for different fairness constraints/definitions.

3.2 Counterfactual Framework

In the counterfactual setting, a context $\boldsymbol{s} \in \mathcal{S}$, drawn from an unknown distribution $P(\mathcal{S})$, provides the information that is required to make a decision, and let $a \in \mathbb{A}$ be an action/decision which is chosen from the possible set of actions $\mathbb{A}$. A policy $\pi : \mathcal{S} \to \mathbb{A}$ gives a probability distribution over possible actions in a given context and $\pi_0(a|\boldsymbol{s})$ is the sampling policy. Additionally, for every sample (s, a) there is a partial feedback r as a numerical reward signal which leads to the logged data in the form of $\{(s_i, a_i, r_i)\}_{i=1}^{n}$.

Given the set of all available policies Π, we aim to find an optimal policy π^* which minimizes the loss of prediction on offline data, while it is unbiased with respect to π_0. To do so, we first need an unbiased estimator of a new policy π to estimate the loss R

$$R(\pi) = \mathbb{E}_s \mathbb{E}_{a \sim \pi(a|s)} \mathbb{E}_r[r], \tag{4}$$

which is used in the following objective function

$$\pi^* = \arg\min_{\pi \in \Pi} [R(\pi)]\,. \tag{5}$$

As mentioned in Section 2.2, different estimators have been proposed to compute an unbiased estimate of Equation (4), and several learning algorithms to optimize Equation (5). Therefore, we model the fairness-aware learning in this counterfactual setting in order to learn a policy that guarantees a bounded bias and variance.

In our model, we define $s_i := x_i$ which indicates that a feature vector in the classification task is equivalent to a context in the counterfactual setting. We further consider that the class labels are the decisions derived from a particular policy, which has two implications. First, in terms of learning the optimal policy π^*, we aim at re-labelling the samples in order to additionally comply with the fairness constraint. Hence, optimizing the objective function in Equation (3) translates into learning the optimal policy in the counterfactual setting such that $f(x_i) \simeq \pi^*(a_i|x_i)$. Note that the value of a_i is not limited to binary values and the model thus can be used in multi-class classification problems as well. Second, in terms of existing offline data, the decisions have been already made taking the sensitive attribute into account which leads to the sampling policy $\pi_0(a_i|x_i) = y_i$. Therefore, the sampling policy is known and is deterministic. Nevertheless, we are interested in a stochastic policy to approximate the likelihood of all the decisions in a given context which later can be used in characterizing the feedback. Accordingly, we estimate π_0 from the data via logistic regression in order to determine the decisions with low probability. Consequently, $\hat{\pi}_0$ is learned from $\{(\bar{x}_i, y_i)\}_{i=1}^{n}$ as the unfair/biased sampling policy in which the data includes the sensitive attribute.

Furthermore, we compute the rewards of the counterfactual model from the fairness measure introduced in Equation (1). Note that the reward can be interpreted as risk in this context, since it is employed in the minimization task of Equation (4). Let $s = 1$ denotes the protected group, in order to satisfy the equalized odds measure, we determine a number k such that

$$\frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1 \wedge s_i = 1\} + k}{\sum_{i=1}^{n} \mathbb{1}\{s_i = 1\}} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1 \wedge s_i = 0\} - k}{\sum_{i=1}^{n} \mathbb{1}\{s_i = 0\}} \tag{6}$$

holds for the logged data, where $\mathbb{1}\{.\}$ is the indicator function. The above equation implies that swapping the label of k positive samples from the non-protected group with the label of k negative samples from the protected group will lead to equal positive ratios, or in other words, a fair classification. We utilize the learned unfair (sampling) policy $\hat{\pi}_0$ to penalize the k samples with the lowest probabilities from each group that belong to the incorrect class in terms of fairness. Therefore, the reward value is given by

$$r_i = \begin{cases} 0 & i \in \{\mathbb{B}_k^+ \vee \mathbb{B}_k^-\} \\ -1 & \text{otherwise} \end{cases} \tag{7}$$

where $\quad \mathbb{B}_k^+ = \left\{\arg\min_{i,s=0} \quad \hat{\pi}_0(y_i = 1|\bar{x}_i)\right\} \times k,$
and $\quad \mathbb{B}_k^- = \left\{\arg\min_{i,s=1} \quad \hat{\pi}_0(y_i = 0|\bar{x}_i)\right\} \times k.$
Note that "$\times k$" stands for taking the *arg min* for k times. Consequently, Equation (7) guides the counterfactual learning toward more fair models by substituting the less likely samples of non-protected group in the positive class with low probable samples of protected group devoted to the negative class.

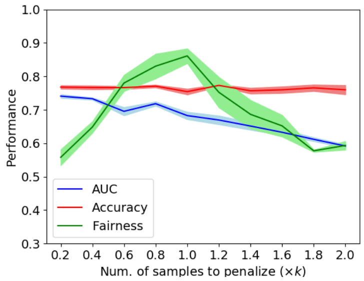

Figure 1: Classification performance vs. fairness measure.

4 EMPIRICAL STUDY

We conduct our experiments on the Adult income dataset [2] that contains a total of ~45k subjects with a binary label to indicate a high or low income. We consider the attribute "gender" as the binary sensitive feature separating the protected group from the non-protected one. The data is prepared using the pre-processing technique in [20] and the shuffled data is split into train, validation, and test sets with ratios of (60%, 10%, 30%). We use POEM algorithm [18] to train a counterfactual model in which the reward is computed from Equation (7). Model selection is performed on the validation set that leads to a self-normalized estimator with a variance regularizer [19]. The sampling policy π_0 is estimated on a fraction of training data via logistic regression algorithm with LBFGS solver and l_2-norm regularizer. For evaluating the performance, we compute the Area Under the ROC curve (AUC) which gives a more strong metric for classification scenarios compared to accuracy, particularly, when there is class imbalance. Additionally, fairness is measured with the ratio given by the *min* operator in Equation (2), in which the value of 1 satisfies the equalized odds. We repeat all the experiments over several runs and report the average results with their standard error.

4.1 Performance Results

In the first experiment, we evaluate the effectiveness of our counterfactual approach in terms of the value of k to find a trade-off point between classification performance and fairness. Recall that k determines the number of samples to penalize from the training data in order to satisfy equalized odds. Nevertheless, this number might be strong with respect to fairness such that it causes reverse fairness (discrimination against non-protected group), which also can lead to the loss of performance. Hence, we modify the number of samples from each group to penalize in the range of $[0.2k, \ldots, 2k]$ and evaluate the performance.

Figure 1 shows that the value of k itself is the right amount of samples to penalize for gaining the maximum fairness. However, the degree of fairness is declined in both directions: in the left part, due to reducing the fairness cost, and in the right part, due to

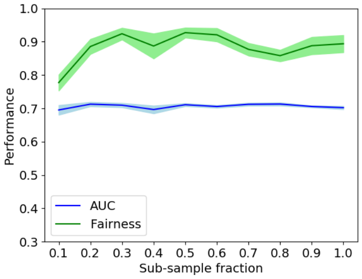

Figure 2: Performance w.r.t. different sub-sample fractions.

reverse fairness. On the other hand, by increasing the number of samples to penalize, the sampling policy, which contains the actual class labels is more undermined and the classification performance diminishes. Note that we also add the results in terms of accuracy to show that accuracy might be misleading in classification tasks and does not provide the true effectiveness of a method.

In addition, we study the effect of sampling fraction for learning the logging policy $\hat{\pi}_0$. That means we vary the size of sub-sampled data from the training set to estimate $\hat{\pi}_0$. Figure 2 exhibits the performance results in terms of both AUC and fairness for sub-sample fractions from 0.1 to 1. The figure shows that changing the fraction data does not have a significant outcome on the performance.

We further evaluate our counterfactual framework compared to a qualified method for fair classification introduced by [20] as a baseline. We utilize the code and setup that are available online to compute the performance of their method. The baseline employs the p%-rule to measure fairness where $p = 100$ is equivalent to equalized odds. Figure 3 represents the performance of both methods in terms of AUC and fairness. We compare the baseline with the results from our approach with 0.8k and sub-sample fraction of 0.2 as the best performing configuration (see Figure 2) in terms of multiplicative loss factor (γ). This factor balances classification performance and fairness measure in the optimization process of the baseline method. The figure demonstrates that our model is in-line with the baseline method. There is only a raise in the fairness over our model for $\gamma = 0.2$, however, the AUC is lower than our classification performance. Additionally, the intersection point, where the AUC and fairness are balanced, is almost the same for both approaches, about 0.71 (comp. Figure 1). Moreover, the counterfactual framework is computationally much more efficient than the baseline and thus is an excellent method for learning fair models.

5 CONCLUSIONS

In this paper, we presented a counterfactual framework tuned for fairness-aware learning. Leveraging the ideas of counterfactual reasoning to learn unbiased policies from offline data, we designed a model in which the unfair decisions are now rectified to balance fairness versus classification performance. In our setting, we considered the biased decisions (or class labels) as the sampling policy, and utilized a fairness measure for specifying the partial feedback for those decisions. Our empirical results showed that our counterfactual framework is able to effectively cope with the fairness issue, and increases the measure of fairness while maintaining an acceptable classification performance for the decision systems.

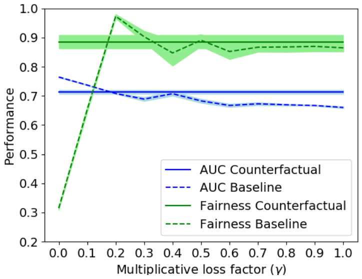

Figure 3: Performance compared to baseline.

ACKNOWLEDGMENTS

Part of this research is supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876/B3.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning.*

[2] A. Asuncion and D.J. Newman. 2007. UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* (2016).

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* (2013).

[5] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning.*

[6] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

[7] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems.*

[8] Daniel G Horvitz and Donovan J Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* (1952).

[9] Vasileios Iosifidis and Eirini Ntoutsi. 2019. AdaFair: Cumulative Fairness Adaptive Boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.*

[10] Nathan Kallus. 2018. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems.*

[11] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer.

[12] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. 2019. The sensitivity of counterfactual fairness to unmeasured confounding. *arXiv preprint arXiv:1907.01040* (2019).

[13] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems.*

[14] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems.*

[15] Gilles Louppe, Michael Kagan, and Kyle Cranmer. 2017. Learning to pivot with adversarial networks. In *Advances in neural information processing systems.*

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[17] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. 2019. CAB: Continuous Adaptive Blending for Policy Evaluation and Learning. In *International Conference on Machine Learning.*

[18] Adith Swaminathan and Thorsten Joachims. 2015. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research* (2015).

[19] Adith Swaminathan and Thorsten Joachims. 2015. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems.*

[20] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics.*