Data Processing Techniques to Enhance Algorithmic Fairness

Jiale Shi

Abstract— **The problem of discrimination caused by machine learning algorithms has received increasing attention. How to avoid the perpetuation and amplification of the discrimination from machine learning systems has become a fundamental issue of machine ethics. In fact, the discrimination presented by machine learning algorithms are often caused by the biases existed in training data. This suggests that eliminating biases existed from the training data, especially the biases caused by sensitive attributes, is an important technique of improving the fairness of the algorithm. The existing data bias reduction algorithms can be categorized into two kinds, causality based methods and association based methods. The causality based methods need the expert knowledge of the underlying causal structure in the dataset. The association based methods require applying heuristic restrictions in bias reduction process, without considering the influence of attributes that correlated with sensitive attributes. In this paper, we propose a data pre-processing method considering the effects of the attributes correlated with sensitive attributes to enhance the algorithm fairness by combining the association based bias reduction method. We evaluated our proposed method on public dataset. The evaluation results show our proposed method can identify the sensitive attributes exactly and the fairness of the machine learning algorithms can be improved compared to the existing methods.**

Index Terms—**Algorithmic fairness, Bias mitigation, Sensitive attribute, Indirect sensitive attributes.**

I. INTRODUCTION

The application of machine learning algorithms has brought significant progress to various public affairs, such as finance, anti-terrorism, taxation, justice, medical care, and insurance, directly impacting the well-being of citizens. However, in recent years, issues of unfairness and discrimination have caused by widely applied machine learning algorithms in areas such as credit scoring, crime prediction, and loan evaluation. As a result, the ethics of algorithm, especially concerning the fairness of machine learning algorithms, has gained considerable attention from the public and the government[1]. The problem of algorithmic fairness may exacerbate the bias to the groups that have historically been discriminated against. For example, in 2014, a team at Amazon developed an automated hiring system to screen the resumes of the job applicants. According to Reuters[2], the hiring system was trained based on 10 years of Amazon's hiring data and it gives a score from 1 to 5 to each job applicant. However, in 2015, the team realized that the system showed a significant gender bias for male candidates and female candidates due to historical discrimination (bias) in the training data. Although Amazon improved the system to hide gender attributes, there was no guarantee that there are biases still in other ways. Therefore, the project was abandoned entirely in 2017. Furthermore, similar examples include gender bias in online advertising and Google image search for occupations. Based on the above examples, it is known that the analytical judgments supported by machine learning systems may influence the decision maker. The discrimination presented in these machine learning systems are caused by the bias in training data. And, this discrimination will be reinforced and legitimized by the increasing deployment of machine learning algorithms. How to avoid perpetuating and amplifying the discrimination by machine learning systems have become a critical issue of the algorithmic fairness.

Data bias correction algorithms, also known as algorithmic fairness pre-processing methods, there are two main bias reduction algorithms, causality based methods[3] and association based methods[4]. The causality based methods need the expert knowledge of the underlying causal structure in the dataset. This approach is not practical for applying in different areas without domain knowledge. The association based methods require applying heuristic restrictions in bias reduction process, without considering the influence of attributes that correlated with sensitive attributes. When performing bias reduction operations on sensitive attributes, two different strategies can be applied. One is horizontal method[5], which performs operations on the tuples of the dataset. The other is vertical method[6], which performs operations on the attributes of the dataset. However, the horizontal method can be considered invasive because it changes the distribution of the dataset. In practice, the vertical method is the common way to remove the identified sensitive features directly[7]. Doing by this can ensure fairness without tampering the dataset. However, there are multiple attributes correlated with identified sensitive features. If we do not consider the impact of indirect sensitive attributes and remove their effects on fairness, the discrimination reduction operation cannot achieve the expected effectiveness.

In summary, finding a unique method to optimize the original dataset and maintain the accuracy and fairness of machine learning algorithms is a challenge. In order to reduce the discrimination of machine learning algorithms at the root and increase their fairness, in this paper, we combine the method of pre-processing algorithm in algorithmic fairness, and optimize the correlation method in bias reduction algorithm to improve fairness.

Manuscript received March 17, 2024
Jiale Shi, School of Computer Science and Technology, TianGong university, TianJin, China

II. Related Work

A. Related Research on Algorithmic Fairness

In this section, we will provide a review of relevant works of algorithmic fairness, including different definitions about algorithm fairness and data bias reduction methods.

Fairness through unawareness (FTU) is achieved when the sensitive attribute is not used in the algorithm for classification and prediction. Individual fairness, on the other hand, was proposed by Dwork et al.[4] in 2012 and is achieved when the algorithm predicts the same outcome for similar individuals. In other words, if two individuals are similar according to a certain metric, their predictions should also be similar. Kim et al. improved on this concept by introducing preference-informed individual fairness (PIIF), which allows for some deviation from individual fairness to meet personal preferences and provide more favorable solutions for individuals.

In legal contexts, fairness of decision-making processes is typically evaluated based on two main criteria: differential treatment and differential impact. The above definitions have inspired various researchers to explore ways to promote fairness in decision-making processes. For instance, Zafar et al[7] have investigated how to remove sensitive attributes from decision-making to avoid differential treatment, and how to add fairness constraints to eliminate differential impact. They have also introduced covariance to transform non-convex problems into convex shapes and examined the sensitive attributes of multi-classification and the analysis of multiple sensitive attributes. On the other hand, Beretta et al[8] have combined different democratic ideals with the concept of fairness to propose evaluation criteria for fairness that are suitable for different democratic backgrounds. They have suggested that counterfactual fairness, unconscious fairness, and fairness based on group conditional fairness are more suitable for competitive democracy, while individual fairness is more appropriate for liberal democracy, and preference-based fairness is more fitting for egalitarian democracy.

Salimi et al[5] introduced a user-centric approach for feature classification by allowing users to categorize features as sensitive, acceptable, or unacceptable. Acceptable features are those that the user allows to influence the classifier's predictions, while unacceptable features are those that may introduce biases based on sensitive attributes. They also proposed a Capuchin (CA) system that can repair data that does not conform to the user's feature classifications by adding or removing tuples. This system is designed to provide users with greater control over the fairness of the model by allowing them to specify which features are considered sensitive and ensuring that the model is not influenced by them. The CA system can also help to reduce the impact of biases by repairing the data that may leak sensitive attribute biases.

Although the existing methods can improve the fairness of the algorithm to some extent, there are still some problems. For example, Unawareness proposed by Zafar et al., which overemphasizes the constraint of sensitive attributes, but this method have few enhancements to model fairness. The individual fairness proposed by DWork et al. cannot properly quantify the gap between individuals. The bias reduction algorithm, Capuchin (CA), breaks the causal chain of these attributes by adding and removing tuples. However, this approach can be considered invasive because it alters the data distribution.

B. FairLearn

FairLearn[9] is a Python library for fair learning. It is designed to help machine learning practitioners identify and correct unfairness in model predictions, as well as to facilitate the construction of fair and balanced machine learning models. The FairLearn library consists of two main modules which are fairness metrics and fairness correction. The fairness metrics are mainly used in the fairlearn.metrics module under FairLearn.

III. Mathods

A. Definition of the problem

In this paper, the goal of this study is to use preprocessing methods in algorithmic fairness to improve the classification fairness of machine learning models. To this end, this study improves a traditional preprocessing model for algorithmic fairness: the unconscious fairness model (FTU). The traditional unconscious fairness model only emphasizes the constraints on sensitive attributes, but such constraints do ignore an important issue, i.e., sensitive attributes do affect other attributes. This situation leads to the fact that constraining only the sensitive attributes is often unsatisfactory for improving the fairness of the algorithm, because the sensitive attributes can be represented by those attributes that are affected by them (attributes that are highly correlated with the sensitive attributes).

For this reason this study proposes the concept of indirect sensitive attributes.Assuming that given an original dataset D and its sensitive attribute S, this study uses to determine the indirect sensitive attributes. We refer to attributes that are highly correlated with sensitive attribute S as indirect sensitive attributes I.

In addition in order to better measure fairness, a concept of fairness, called relative fairness. Assuming that the original dataset D has two subsets D_1 and D_2, the predictor should have the similar probability of getting a positive result when predicting both subsets.

B. Data Preprocessing

This study began with data preprocessing, preprocessing is very important in constructing a classification model and can often determine the results of model training. First, the given data need to be divided into training and test sets, the data segments that do not contain anomalies as training data, and the rest of the data with anomalies as test data. Then in order to improve the robustness of the model, it is also necessary to standardize the training and test sets. After first standardizing the training set, the test set is then standardized using the standardized parameters (mean and variance) of the training set.

C. Algorithmic Process

The traditional unconscious fairness model (FTU) is a preprocessing method for algorithmic fairness, which will exclude the influence of sensitive attributes on the dataset by deleting the sensitive attributes in the dataset, and then

exclude the influence of the sensitive attributes on the classifiers, so as to improve the fairness of the machine learning model. In this study, the traditional unconscious fairness model (FTU) is improved, and the FTU_{ISA} model is proposed, which deletes the sensitive attributes of the original dataset and then deletes the indirect sensitive attributes to improve the fairness of the model to improve the performance of the algorithm, and the algorithmic process is shown in Table 1.

Table 1. Specific Algorithm Flow

Algorithm 1: Fairness Improvement Process

Input： Original dataset, target attributes, sensitive attributes, indirect sensitive attributes

Output： Outcomes with increased fairness

1： **repeat**

2: Remove sensitive attributes (traditional FTU model)

3: **for the number of indirect sensitive attributes do:**

4: Remove Indirect Sensitive Attributes

5: **if**(Reduction in F1 score)

6: rollback

7: **break**

8: **end for**

IV. Experiment and Result

A. Dataset Introduction

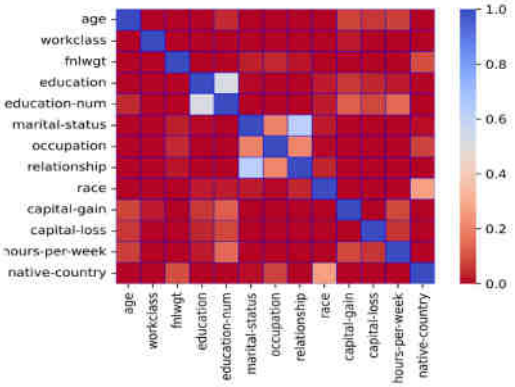
The Adult dataset[10] contains information from U.S. Census of the 1994. The prediction task is to determine whether a person's annual income exceeds 50k dollars. In the above section, we have used the adult dataset to demonstrate how to identify sensitive attributes, and improving the fairness of the prediction model.

The COMPAS[11] dataset is the dataset associated with this tool for training and evaluating crime risk assessment algorithms. It is often used for prediction tasks related to crime, such as whether the offender will reoffend within two years, whether the offender will return to violent crime within two years and whether the defendant will evade court when he appears in court. Afterwords, we used Compas dataset to validate our method again.

B. Sensitive Attributes Identification

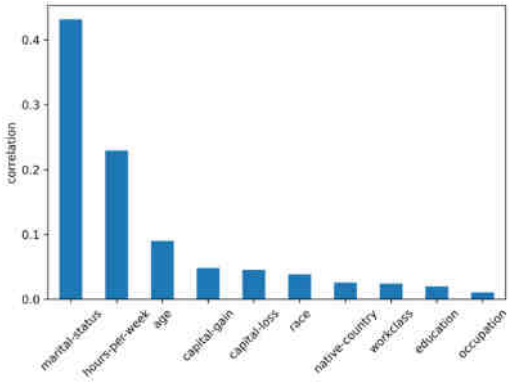
Before identifying indirect sensitive attributes, this study needs to analyze the correlation between attributes other than sensitive attribute S. The purpose of this step is to find similar attributes with high correlation, and in the subsequent analysis, this study analyzes only one of these similar attributes. Then, the results of the analysis of that attribute are used to represent these similar attributes, and using this method can help this study to reduce the amount of data that needs to be analyzed. Taking the Adult dataset as an example, it can be seen from the above analysis that the sensitive attribute in the Adult dataset is the gender attribute. This study does not consider the gender attribute and analyzes the correlation between other attributes and the results of the analysis are shown in Figure 1. In this study, it can be found that the correlation between education attribute and years of education attribute, and the correlation between relationship attribute and marital status attribute is higher than other attributes. Therefore, in the following analysis, this study considers only one attribute from each of these two sets of attributes separately. Meanwhile, the fnlwgt attribute is a sequence number, which is of little value for further analysis and will not be considered in the following study.

Figure 1. Correlation between other attributes



After analyzing the correlation of each attribute, this study needs to determine the indirect sensitive attributes through the sensitive attributes. In this study, the attributes that are highly correlated with the sensitive attributes are called indirect sensitive attributes. The order of correlation with sensitive attributes is shown in Figure 2. In Figure 2, this study can find that marital-status attribute and hours-per-week attribute have higher correlation with gender attribute than other attributes. This is because the attribute of relationship is similar to the attribute of marital status. This study considers relationship attributes, marital-status and hours-per-week as possible indirect sensitive attributes.

Figure 2. Indirect Sensitive Attributes Ordering



C. Machine Learning Model Fairness Improvement

This study uses the UCI Adult dataset to train a predictor. The Adult dataset is initially a categorization

dataset that is used to predict whether a person will earn more than $50,000 per year. In this paper, in order to make the experimental objectives clearer and more intuitive, this study converts the classification problem of the Adult dataset into a decision problem for loans. In this study, the value of the target attribute is expressed as whether each individual has repaid a loan in the past. The income attribute is the target attribute in the Adult dataset. This study takes the value of the income attribute to be greater than $50,000 to indicate a person who has repaid a loan in the past, and less than $50,000 to indicate a person who has repaid a loan in the past. This study uses this data to train a predictor to predict whether an individual will repay a loan. This study uses a logistic regression model to train the dataset, logistic regression is a generalized linear regression analysis model that belongs to the supervised learning of machine learning, which is similar to regression in terms of its derivation process and computation, but is actually mainly used for solving binary classification problems.

This study uses the Metrics Module module in the Fairlearn library to evaluate the fairness of the model, and the metrics evaluated include the accuracy score and the selection rate, with the accuracy score representing the accuracy of the prediction model, and the selection rate representing the percentage of individuals in the able group who are able to repay the loan. The overall prediction results were first evaluated, and then the male and female groups were evaluated separately according to the gender of the sensitive attribute. The results are shown in Table 2. In Table 2, this study evaluates the fairness of the model using the ratios regarding the selection rates of males and females, and it can be found that there is a significant difference between the selection rates of females and males. This situation indicates that there is a problem of unfairness in this prediction model.

Table 2. Projected results without consideration of fairness

Group	accuracy	recovery rate
All group	0.851	0.196
Female group	0.925	0.074
Male group	0.814	0.258

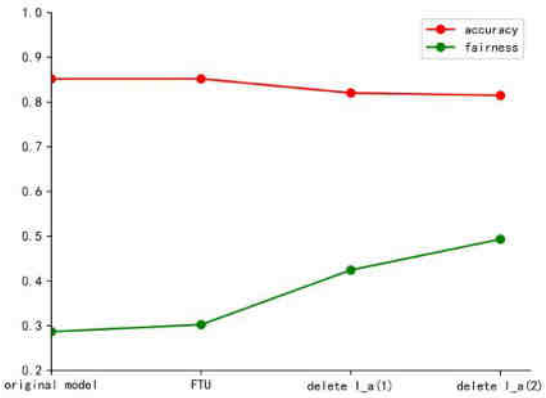
After that, the unconscious fairness model was used in this study to enhance the fairness of the machine learning model and the results are shown in Table 3. The enhancement of the fairness of the prediction results are not obvious.

Table 3. Predicted results using the FTU method

Group	accuracy	recovery rate
All group	0.851	0.196
Female group	0.925	0.076
Male group	0.814	0.254

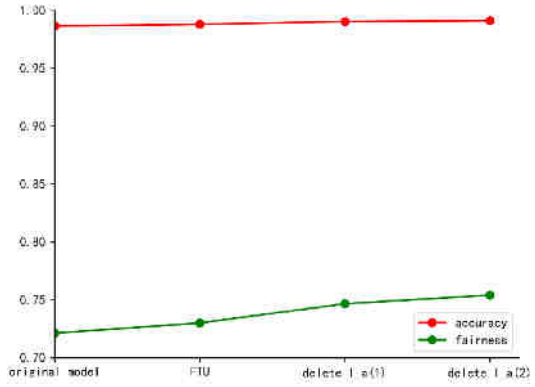
Finally, this study uses the FTU_{ISA} model proposed in this chapter to improve the fairness of the prediction model, as shown in Figure 3. Where the red part represents the prediction correctness, the green part represents the prediction fairness, the prediction fairness is the ratio of the female repayment rate to the male repayment rate, and the second point of the horizontal axis is the performance of the traditional FTU method. Comparing the original FTU method, it can be found that the FTU_{ISA} model improves the fairness of the machine learning model significantly.

Figure 3. Fairness enhancement of the FTUISA model



This study also conducted experiments on the COMPAS dataset to test the performance of the FTU_{ISA} model. The classification model used was a logistic regression model with a prediction target of whether an individual would commit a crime within two years and the sensitive attribute was the racial attribute. The comparison of FTU_{ISA} model with the original FTU model as shown in Figure 4.

Figure 4. Experimental results on the COMPAS dataset



V. CONCLUSION

This study focuses on an improved algorithmic fairness preprocessing model FTU_{ISA}, which improves on the original unawareness fairness model (Fairness Through Unawareness, FTU) to better enhance the fairness of the classification model. This study introduces the definition of problems related to indirectly sensitive attributes and the algorithmic process of the FTU_{ISA} model. After that, this study conducts experiments on two public datasets, they are the UCI Adult dataset and the COMPAS recidivism prediction dataset. From the experimental results, this study is able to improve the fairness of the machine learning model with guaranteed accuracy, and the comprehensive experiments on the public datasets show that this study's method performs better than the traditional FTU model.

Different from the causality based methods, our method need not to satisfy the predetermined inference result, the

predetermined outcome of this inference may existed biases in itself. At the same time, we reducing the fairness improving problem into a dataset amendment problem, making our method better applicable.

REFERENCES

[1] M. Kearns and A. Roth, The ethical algorithm: The science of socially aware algorithm design, Oxford University Press, 2019.
[2] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, in: Ethics of data and analytics, Auerbach Publications, 2018, pp. 296–299.
[3] S. Galhotra, Y. Brun and A. Meliou, Fairness testing: testing software for discrimination, in: Proceedings of the 2017 11th Joint meeting on foundations of software engineering, 2017, pp. 498–510.
[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.
[5] B. Salimi, L. Rodriguez, B. Howe and D. Suciu, Interventional fairness: Causal database repair for algorithmic fairness, in: Proceedings of the 2019 International Conference on Management of Data, 2019, pp. 793–810.
[6] R. Salazar, F. Neutatz and Z. Abedjan, Automated feature engineering for algorithmic fairness, Proceedings of the VLDB Endowment 14(9) (2021), 1694–1702.
[7] N. Grgic-Hlaca, M.B. Zafar, K.P. Gummadi and A. Weller, The case for process fairness in learning: Feature selection for fair decision making, in: NIPS symposium on machine learning and the law, Vol. 1, Barcelona, Spain, 2016, p. 11.
[8] E. Beretta, A. Santangelo, B. Lepri, A. Vetrò and J.C. De Martin, The invisible power of fairness. how machine learning shapes democracy, in: Canadian Conference on Artificial Intelligence, Springer, 2019, pp. 238–250.
[9] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach and K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).
[10] D. Dua, C. Graff et al., UCI machine learning repository (2017).
[11] J. Larson, S. Mattu, L. Kirchner and J. Angwin, How we analyzed the COMPAS recidivism algorithm, ProPublica (5 2016) 9(1) (2016), 3–3.