Fairness Under Unawareness:A New Estimator Reduces Disparity When Protected Class Is Unobserved

Kaixin Du[1,a*+], Yanhao Ji[2,b+] , Fanxin Sun[3,c+], Yuanli Zhu[4,d]

*Corresponding author. Email: [a]dukx@alumni.shanghaitech.edu.cn, [b]a2383284877@gmail.com ,
[c]1328364904@qq.com, [d]1632093519@qq.com

[1]School of Information Science and Technology, Shanghai Tech University, Shanghai, 201210, China
[2]Cognitive science department, University of California, San Diego, 92093, San Diego
[3]Faculty of Science and Engineering, University of Nottingham Ningbo, 312000, China
[4]Faculty of Science and Engineering, University of Nottingham Ningbo, 312000, China

+These authors contributed equally to this work and should be considered co-first authors.

Abstract. Fairness has become an important topic in recent years. Since the boosting development of the society and economy, discrimination against certain disadvantaged groups or communities has been considered a great issue that needs to be detected and prevented seriously. Some real-world examples that might be affected by fairness could be job offers, university admissions and loan approvals. The absence of fairness towards a specific protected class, such as when an advantaged group is more likely to receive favorable outcomes than a disadvantaged one, can lead to significant societal issues. Thus, for both ethical and legal reasons, it is important to conduct estimation methods based on an individual's protected class information and the outcome of a binary decision to measure the fairness of such decisions. This paper will discuss some existing methods for predicting protected information and put forward a new method named switch estimator method, which can improve the traditional method by calculating the bias separately depending on their different conditions. The results will be illustrated using real-world dataset, evaluating the different estimator's performance based on the same protected variable prediction.

Keywords: fairness, protected class, switch estimator, demographic disparity, probablistic proxy model.

1. Introduction

Fairness is a topic that has been paid increasing attention to over recent years. In the real world, discrimination over certain protected class has been an ongoing issue, taking place in various aspects [1]. Typical examples are when it comes to giving offer of a job or entrance to a university [2], or giving approval of a request (e.g. loan) [3,4], to individuals with different classes or different genders. For example, the examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants [5]. Huge social problems will occur if fairness is absent toward a certain protected class, say one advanced group of people is more likely to get the favorable outcome than the other disadvantaged group. According to the previous work

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345325

(Barocas et al, 2016), the Title VII's prohibition of discrimination in employment in American antidiscrimination law illustrated the importance of ensuring equality in the employment process [6]. In other words, decision making algorithms have to ensure their impartiality even though they do not have direct access to protected information such as age, gender, or race [7]. Thus, for both ethical and legal needs, it is of vital importance to construct certain estimating method that can predict individuals' protected class information as well as the outcome of the binary decision to assess fairness of the decision. Efforts has been made towards detecting and studying unfairness. Previous works includes Hutchinson and Mitchell's article, where unfairness under different circumstances has been compared and standardized. They also used mathematical methodologies to measure fairness, which paved the way for future studies [8].

1.1 Evaluating fairness with unknown protected class information

Developing a method with protected class and binary decision outcome information may be feasible in theory. However, in reality, it is very likely for investigators to encounter the situation where the protected class information is inaccessible, which means not being authorized to have access to the information, or doing this might be illegal. Under this circumstance, in order to remedy the situation, some methods need to be applied to estimate the missing information using other related variables to ensure the practicability of the fairness assessing process. Previous studies, for example the Bureau of Consumer Financial Protection (BCFP), have applied probabilistic proxy models with naïve Bayes [9] method to achieve it.

1.2 Problems in fairness assessing methods

Bias has been a long-standing issue among a wide range of estimating methods. For example, when estimating race, geolocation is one of the most typical proxy variables with which the assumption goes that a person living in the geolocation where most residents are of race a will be very likely to be labeled as race **a** (which is somehow regarded as "advantaged"), although he or she might actually be of race **b** (regarded as "disadvantaged"). This misclassification is one important source of bias that contributes to the overestimation of demographic disparity. Another source of bias is the removal of the unclassified portion in some estimating methods that will lead to bias because of change in sample distribution. Besides, there might also exist other biases whose causes are unknown. Figuring out the source of bias and reducing their influence on final outcome has become a difficult task that need further investigation into in the future.

1.3 Contributions

Motivated by the above flaws and limitations of the existing methods, efforts are made towards several topics and the key contributions of this paper are as follows:

(1) Random forest is used as a classification model in estimating the probability P for classes conditioning on proxy variables, with partly unknown protected class information. The weights for different groups are readjusted to alleviate the problem of sample unevenness, which achieved a high accuracy in testing.

(2) A new switch estimator is proposed as a combination of the thresholded estimator and the weighted estimator. It reduces the bias caused by removal of unclassified individuals in thresholded estimator method and balances positive and negative bias generated by different

estimators to achieve a smaller bias. Grid search method is then used to find the overall best **q** and comparison of the bias for different methods on the same dataset confirms that the new estimator is efficient in cutting down bias.

2. State of the art

This section aims to review on some basic definitions and methods used to evaluate the fairness of a binary decision in the previous paper [10]. Some existing limitations in these methods that lead to possible bias will also be pointed out, and solutions to cutting down the bias will be presented in section 2.3.

2.1. Basic variables and indicators

The main variables of interest in evaluation of fairness under a binary decision are listed as follows:

Binary decision Y: For example loan approval; $Y = 1$ stands for a favorable outcome and $Y = 0$ stands for an unfavorable outcome.

Protected class A: Such as race or gender; $A = 1$ represents the advantaged group and $A = 0$ represents the disadvantaged group. Only binary case $A \in \{0,1\}$ is discussed for simplicity.

Proxy variable Z: A set of covariates taking values $z \in Z$ which are used to predict protected class A in a probabilistic proxy model.

Proxy variable X: the input features included in Z, which is used to represent Y using the output of some function $f(X)$ in order to satisfy that Y is independent of A conditionally on Z.

And two indicators, mean group outcome and demographic disparity, are used to evaluate the fairness of a certain binary decision:

Definition 2.1. The mean group outcome for the group $A = a$ ($a \in \{0, 1\}$) is:

$$\mu(a) = E(Y \mid A = a) \tag{1}$$

where E is the usual expectation with respect to population distribution.

Definition 2.2. The demographic disparity is the difference in mean group outcomes between the advantaged and disadvantaged groups:

$$\delta = \mu(1) - \mu(0) \tag{2}$$

where $\delta > 0$ means the advantaged group ($A = 1$) tends to get more favorable outcome than the disadvantaged group ($A = 0$) averagely.

The demographic disparity can be directly calculated if the protected class labels are known. Otherwise, a probability proxy model can be applied to estimate the probabilities $P(A = 1 \mid Z)$ and $P(A = 0 \mid Z)$ of membership to the different groups within A. In the following sections, two different estimators for demographic disparity based on the probability proxy model will be introduced.

2.2 Estimating methods

Before defining the thresholded estimator, the definition of thresholded estimated membership A_i will be given first. Suppose there are N independent and identically distributed (i.i.d.) samples $(Y_i, Z_i)_{i=1}^{N}$. Then the known probability proxy estimates $\{ P(A_i = a \mid Z_i): a \in \{0, 1\}, i \in \{1, \dots, N\}\}$, which is obtained by applying proxy model to observed proxy variable Z_i, can be used to predict the true membership A_i of the ith sample (i.e. predicted label $\hat{A}_i$). Definition 2.3. Let threshold $q \in [\frac{1}{2}, 1)$. Then the thresholded estimated membership $\hat{A}_i$ for the ith unit is:

$$\hat{A}_i = \begin{cases} 1, & P(A_i = 1 \mid Z_i) > q \\ 0, & P(A_i = 0 \mid Z_i) > q \\ NA, & otherwise \end{cases} \tag{3}$$

where NA represents an unclassified observation that is excluded from the subsequent outcome disparity evaluation. A graphic summarization for a single unit $A_i = 1$ is shown in Figure 1, where $A = \hat{A}$ represents correctly classified, and $A \neq \hat{A}$ represents wrongly classified:

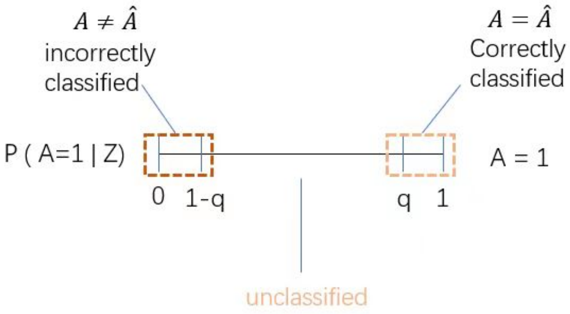

Figure 1. The graphical illustration of the thresholded estimated membership $\hat{A}_i$.

These predicted labels are then used to impute the unknown A_i's in estimating the mean group outcome as well as demographic disparity between two groups. Two different estimators are put forward: thresholded estimator with a hard classification rule that categorize individuals straight into two different groups and remove the unclassified part; and Weighted estimators with a soft classification rule that use the conditional probability $P(A_i = a \mid Z_i)$ directly in the estimation of μ and δ. Further detailed information about these two estimators will be given in the next section.

2.3 Limitation and Solution

The weighted estimator method is an improvement of thresholded method that it cuts down some bias generated by the hard classification rule. However, there exist a practical problem in both methods that the probability $P(A_i = a|Z_i)$ is not always accessible in the real examples. Under this circumstance, with some A_i's unknown, both methods would fail. In order to remedy the situation, random forest method will be used to construct a model to estimate $P(A_i = a|Z_i)$ in section 4.

3. Different estimators for indicators

In this section, detailed information and formal definition of the two existing estimators (thresholded estimator and weighted estimator) will be given, and they will then be combined together to form the new estimator (switch estimator). The bias of all three methods and analyze the advantage and limitations of the new method will also be discussed.

3.1 Thresholded Estimator

The thresholded estimator is based on the definition of thresholded estimated membership $\hat{A}_i$ in section 2.2. Definition 3.1. Let $(Y_i, Z_i)_{i=1}^N$ be N i.i.d. samples, $\{\hat{A}_i\}_{i=1}^N$ be the estimated labels according to Definition 2.3, and I(S) is the indicator function for some set S. The thresholded estimators for mean group outcomes and demographic disparity are: $(a \in \{0, 1\})$

$$\hat{\mu}_q(a) = \frac{\sum_{i=1}^N I(\hat{A}_i = a)Y_i}{\sum_{i=1}^N I(\hat{A}_i = a)}$$

$$\hat{\delta}_q = \hat{\mu}_q(1) - \hat{\mu}_q(0) \tag{4}$$

3.2 Weighted Estimator

Despite its wide use, the threshold estimator method has a remarkable flaw. Based on a hard classification rule, this method inevitably generates some bias because of both the misclassification of some individuals, and the removal of unclassified portion. The weighted estimator, using soft classification instead, is a new estimator proposed in Chen et al.'s paper in order to avoid these problems. Definition 3.2. Let $(Y_i, Z_i)_{i=1}^N$ be defined as above. Then, the weighted estimators for mean group outcomes and demographic disparity are: $(a \in \{0, 1\})$

$$\hat{\mu}_w(a) = \frac{\sum_{i=1}^N P(A_i = a|Z_i)Y_i}{\sum_{i=1}^N P(\mathrm{A}_i = a|Z_i)}$$

$$\hat{\delta}_w = \hat{\mu}_w(1) - \hat{\mu}_w(0) \tag{5}$$

3.3 Switch Estimator

Although weighted estimator cut down the estimating bias generated by thresholded estimator, bias still exists in this method. Chen et al. analyzed the bias sources of both methods and gave numerical examples in their paper. And in section 4.2 in the paper, the bias for thresholded estimator method and weighted estimator method have reversed signs (i.e. $\hat{\delta}_w < 0$ and $\hat{\delta}_q > 0$ for a same dataset, see Figure 2). That gives the inspiration to combine the two methods together to create a new estimator which will balance out the positive and negative bias and achieve a smaller absolute value of bias than any single method. The base theory of switch estimator is to use thresholded estimator as base and then use weighted estimator to handle the unclassified part to decrease the influence of bias. The definition as well as the graphic explanation are presented below:

Definition 3.3. Let $(Y_i, Z_i)_{i=1}^N$ be defined as above. Then, the switch estimators for mean group outcomes and demographic disparity are: $(a \in \{0, 1\})$

$$\hat{\mu}_S(a)\begin{cases}\dfrac{\sum_{i=1}^{N} I(\hat{A}_i = a)Y_i}{\sum_{i=1}^{N} I(\hat{A}_i = a)}, & \hat{A}_i \in \{0,1\}\\ \dfrac{\sum_{i=1}^{N} P(A_i = a|Z_i)Y_i}{\sum_{i=1}^{N} P(\mathrm{A}_i = a|Z_i)}, & \hat{A}_i = NA\end{cases}$$

$$\hat{\delta}_s = \hat{\mu}_s(1) - \hat{\mu}_s(0) \tag{6}$$

The corresponding graphical presentation is shown in Figure 2, where weighted estimating method is applied after the application of thresholded estimating method:

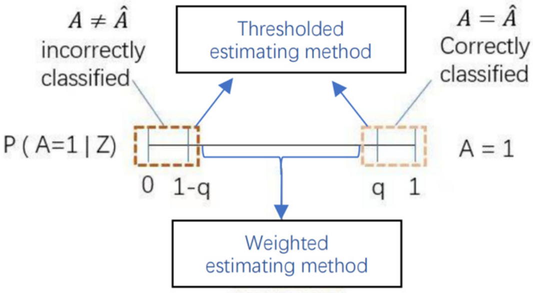

Figure 2. The graphical presentation of the switch estimator.

Here a special annotation on the threshold q is made. The target is to find the best value of q that minimizes the estimating bias, thus the value of q should be chosen in the case that the absolute value of the final bias ($\hat{\delta}_s - \delta_s$) turn out to be the smallest, instead of being the best q only for the thresholded estimator step. Grid search method is used here to traverse every q and determine the best one. The advantage of this combined switch estimator method is that it can handle the bias of thresholded estimator generated by the unclassified (and removed) part, offsetting positive and negative bias from two different estimating methods to get good experiment result. However, its limitation still exists that it cannot handle the problem of misclassification in the thresholded estimator step.

4. EXPERIMENTS and results

4.1 Dataset

In the experiment section, the same dataset used by previous paper (Chen et al., 2018) is used, which is provided in the website Public HMDA - LAR Data Fields | HMDA Documentation (cfpb.gov). Initially over 980k unique values from California and Alabama states are collected. The dataset contains information including race, state code, credit type, and income etc. The aim here is to examine this dataset and to search for an efficient method to demonstrate the decrease of bias when predicting protected variables.

4.2 Data Cleaning

In order to ensure a clean dataset for the experiment, several data cleaning processes are conducted on this dataset. The first step is to remove missing labels in this dataset. Due to the excessiveness of the data (over 980k values are included in the dataset), those rows that have missing values are deleted directly. The next action is to reduce the size of the dataset since the dataset is still too large to analyze or operate even after the removal of rows containing missing values. Since only interested in a binary experiment between protected variables, the race columns are limited to only contain "White" and "African American" values. Additionally, the "White" is labeled as advantage group "1", and "African American" as disadvantage group "0". It is also noticeable that there are different types of labels in the dataset: while the race, state code columns contain category information, income and property value contain numerical information. In order to handle the problem, a partial normalization which normalized numerical information to from 0 to 1 is conducted.

4.3 Thresholded Estimator Implement

4.3.1 RandomForest Classifier

Firstly Random Forest is used as a classification model. Features are all the data except "derived_race" (the aforementioned A) and "action_taken" (the aforementionedY) column, and labels are the sensitive attributes "derived_race" in the dataset. The processed dataset is then divided into a train set and a test set in the ratio of 80% and 20% to train the model. However, the dataset has a very serious problem of sample unevenness (about 90% of the samples in the dataset for A=="White" and 10% for A=="African American"), it is also necessary to resample the dataset to solve the problem of sample unevenness. The parameter class_weight="balance" is used to adjust the weights of the input to $\frac{N_{samples}}{N_{classes} * np.bincount(y)}$, in other words, giving each class different weights to average their impact on the model. The output of the network is the probability of predicting $A == 1$. The final accuracy of the model on the test set is 84.78%.

4.3.2 Setting Q to calculate indicators μ and δ

Next, Q-values are set and the predicted probabilities are classified into three categories (1,0 and NA) according to the classification criteria defined in 3.1. $\hat{\mu}_q(1)$ and $\hat{\mu}_q(0)$ are calculated for the samples classified as 1,0 and $\hat{\delta}_q$ are obtained by subtracting them. Three values (Q=0.7, 0.8 and 0.9) are tried respectively as well as the corresponding $\hat{\mu}_q$ and $\hat{\delta}_q$, which will be shown in 4.6.

4.4 Weighted Estimator Implement

4.4.1 RandomForest Classifier

The weighted estimator is implemented with the same preprocessing method of the data and the same random forest classifier is used to ensure that the train set and the test set are all consistent with those used in the Thresholded Estimator, in order to compare the results more informatively. The output of the network is the probability of predicting $\mathrm{A} == 1$.

4.4.2 Using predicted probabilities to calculate indicators μ and δ

Next, the predicted $P(A_i = 1 \mid Z_i)$ are used as the weights of the samples, and $\hat{\mu}_q(1)$ and $\hat{\mu}_q(0)$ are calculated for the samples classified as 1,0 based on definition 3.2. Finally $\hat{\delta}_q$ is obtained by subtracting, whose results will be shown in 4.6.

4.5 Switch Estimator Implement

4.5.1 RandomForest Classifier

The switch estimator is implemented with also the same preprocessing method of the data and randomforest classifier to ensure that the train set and the test set are all consistent with those used in the above two estimators, in order to compare the results more informatively.

4.5.2 Grid Search for Q

Next, the grid search method is applied to traverse the search for the globally optimal solution Q and constantly calculating $\hat{\mu}_S$ and $\hat{\delta}_s$, finding the solution with the smallest absolute value of $\hat{\delta}_s$. The results will be shown in 4.6.

4.6 Outcome

Our outcome is presented as follows:

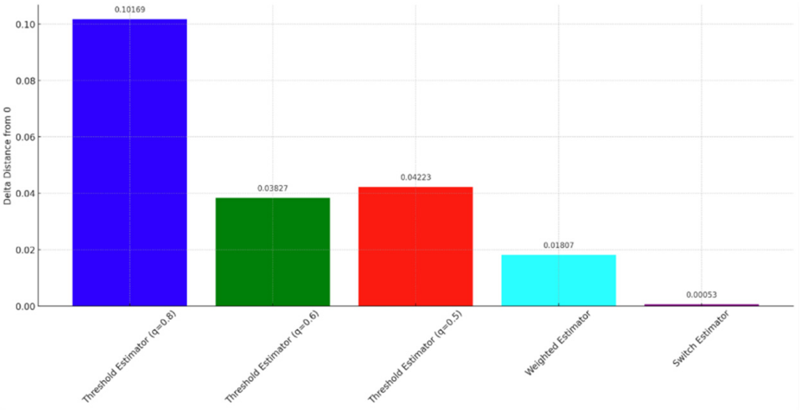

Figure 3. Delta Distance from 0 for Different Estimators.

As shown in Figure 3, it is evident that different estimation methods output varying delta values, implying distinct interpretations of the data. Specifically, in the threshold estimator visualization, as the q value increases from 0.5 to 0.8, there is a noticeable rise in the delta value with q = 0.8 exhibiting the highest deviation from 0. This could suggest threshold estimator ignore about high portion of values and consider them as unclassified section at this q setting. In comparison, while both the Weighted Estimator and Combined Estimator display relatively low delta values, the latter is almost negligible, hinting at its potential stability and resistance to bias or fluctuations. It can be figured out from this bar chart that switch estimator has the lowest bias delta that is relatively close to the actual 0.

5. Conclusion

This paper illustrates a new method when it comes to predicting protected variables using proxy variables that is different from using threshold estimator or weighted estimator solely. The new switch estimator proposed in this paper ease the bias source from the unclassified section abandonment of using threshold estimator. Meanwhile, the switch estimator considers the reweighting method for unclassified portion and combines these two portions together, successfully reducing the bias from predicting protected values. The grid search method is also introduced, which is deployed in the process of finding optimal overall threshold in this paper. In general, the newly introduced switch estimator flexibly considered each portion of the probability and their matching bias, merging two different estimators to solve for different types of bias variously. At the same time, it is important to mention different distributions in the dataset or reality may vary the outcome, such as one of the variables would be relatively small or difficult to observe. The future work includes the following three aspects. Firstly, efforts can be made in using the doubly-robust estimator to minimize the cross-entropy loss in estimating the probability for $A_i = 1$ conditionally on Z to improve the efficiency and correctness of estimating. Secondly, the problem that individual with which features will be more willing to give their protected class information (for example, whether people with higher income is more willing to give their race information) can be further looked into. Finally, targeting at the switch estimator, it can be investigated that how two estimators can be combined in a better way, for example designing a self-adapted weighter that give different weight to the two estimators based on different situations.

REFERENCES

[1] Cecilia Munoz, Megan Smith, and DJ Patil. (2016). Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Technical Report May.

[2] Delores A Conway and Harry V Roberts (1983). Reverse Regression, Fairness, and Employment Discrimination. Journal of Business & Economic Statistics., Vol. 1, No. 1: 75-85.

[3] William H. Greene. (1984). Reverse regression: The algebra of discrimination.Journal of Business and Economic Statistics. Vol. 2, No. 2: 117–120.

[4] Jiahao Chen. (2018). Fair Lending Needs Explainable Models for Responsible Recommendation. In Proceedings of the 2nd FATREC Workshop on Responsible Recommendation (FATREC’18). ACM.

[5] P. J. Bickel, E. A. Hammel, and J. W. O’Connell (1975). Sex Bias in Graduate. Admissions: Data from Berkeley. Science,Vol. 187, No. 4175: 398–404.

[6] U.S. Equal Employment Opportunity Commission. (2009) Federal Equal Employment Opportunity (EEO) Laws. https://www.eeoc.gov/fact-sheet/federal-laws-prohibiting-job-discrimination-questions-and-answers

[7] Solon Barocas, Andrew Selbst (2016). Big Data’s Disparate Impact. California Law Review,Vol. 104, No. 3: 671–732.

[8] Ben H. and Margaret Mitchell (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: Fairness, Accountability, and Transparency (FAT*'19). Atlanta. pp. 49–58.

[9] Lacramioara Mazilu, Norman W. Paton, Nikolaos Konstantinou, and Alvaro A. A. Fernandes. (2022). Fairness-aware Data Integration. J. Data and Information Quality, Vol. 14, No.28: 1-26.

[10] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, Madeleine Udell (2019). Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In: Fairness, Accountability, and Transparency (FAT*' 19). Atlanta. pp. 339–348