International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217251

PCIV method for Indirect Bias Quantification in AI and ML Models

Ashish Garg*, Dr. Rajesh SL
CSIT, JAIN (Deemed University), Bengaluru, Karnataka, India

Article Info
Volume 5, Issue 3
Page Number: 687-693

Publication Issue :
May-June-2019

Article History
Accepted : 01 May 2019
Published : 13 May 2019

ABSTRACT

Data Scientists nowadays make extensive use of black-box AI models (such as Neural Networks and the various ensemble techniques) to solve various business problems. Though these models often provide higher accuracy, these models are also less explanatory at the same time and hence more prone to bias. Further, AI systems rely upon the available training data and hence remain prone to data bias as well. Many sensitive attributes such as race, religion, gender, ethnicity, etc. can form the basis of unethical bias in data or the algorithm. As the world is becoming more and more dependent on AI algorithms for making a wide range of decisions such as to determine access to services such as credit, insurance, and employment, the fairness & ethical aspects of the models are becoming increasingly important. There are many bias detection & mitigation algorithms which have evolved and many of the algorithms handle indirect attributes as well without requiring to explicitly identify them. However, these algorithms have gaps and do not quantify the indirect bias. This paper discusses the various bias detection methodologies and various tools/ libraries to detect & mitigate bias. Thereafter, this paper presents a new methodical approach to detect and quantify indirect bias in an AI/ ML models.

Keywords : Artificial Intelligence, Machine Learning, Biased Model, AI Ethics, Fair AI, Computer Science

I. INTRODUCTION

Victor Hugo, a famous French poet, once said "Being good is easy, what is difficult is being just". This quotation was not given in the context of AI & ML models but it appears to fit very well there. It might be easier to build a good AI/ ML model but building a fair model is much more challenging. The fairness of AI/ ML models has remained a topic of conceptual debate for decades and organizations as well as governments had acknowledged the need for imbibing ethics in predictive models for a long. But, the importance and urgency of making policies and frameworks on AI ethics have increased tremendously in the last 10 years. Bias can be noticed in many predictive models where selection/ rejection decision is unfair and tilted in favor of privileged groups based on attributes such as gender, color, race, etc. Fairness in AI models means that the output produced by the AI algorithm must be free of intentional or unintentional bias against the unprivileged groups. Unethical Bias can be due to any

of the legally protected attribute (such as race, gender, color, etc.) or attributes which are not legally protected but considered sensitive by the organization or society based on the context. AI and ML algorithms have created new challenges for being prone to discrimination. Though human decision-making has historically remained prone to discrimination and it was earlier thought that machines cannot discriminate. Though machines may not have emotions to discriminate they are also not free of discrimination. AI & ML algorithms work like a black-box and at times their workings are not understandable to human beings AI & ML algorithms may act upon classes of people based on historical data which may victimize certain groups of people without them ever being aw1are of the discriminative automated decision-making. Hence, it is critically important that bias in AI models be detected and mitigated.

In Financial industry, financial data modeling is often served with challenges with regards to various types of data bias (such as Gender bias, Racial bias, Age bias) as well as imbalanced data [9]. AI models producing unfair/ unjust output is not rare. The infamous incident at Google where 'Timnit Gebru' (who highlighted the risk of bias in large language models) had to abruptly leave the company is well known. COMPAS system (Correctional Offender Management Profiling for Alternative Sanctions) discrimination is another infamous incident. COMPAS system has been used by the United States to assess the probability of a defendant becoming a recidivist. It was later found to be largely discriminatory against African American ethnicities [17]. Gender bias can be noticed in high- income jobs. Generative systems such as GAN have also been found to exacerbate the biases in the generated data. All such incidents stress the critical need for organizations to ensure that their AI & ML systems are unbiased.

Bias in AI models may be caused by the data itself or the algorithm. Usage of incomplete/ imbalanced training data may result in the model getting trained in a biased way and hence produce biased output. Black box AI/ ML algorithms can also cause bias. In an endeavor to maximize the overall accuracy, these algorithms can unnoticeably get biased against certain groups in the protected attributes. Algorithmic bias need not be intentional on the part of the algorithm developer. It can be unintentional as well.

A common misconception among novice data scientists is that the removal of protected attributes from the training dataset will make the model fair. This is not true. Removal of protected/ sensitive attributes from the training dataset does not automatically ensure fairness. There might be some non-protected attributes in the dataset which are strongly correlated with one or more of the protected attributes. These attributes can also cause bias in the model.

Discrimination can be of two types - Direct or Indirect [12]. If discrimination is caused directly by the protected attributes, then it is called direct discrimination. On the other hand, if discrimination is caused by attributes that are not directly protected but strongly correlated with the directly protected attributes, then it is called indirect discrimination.

Explainability of AI models is an inter-related aspect that is important to ensure model fairness as well.

Explainable AI refers to methodologies to ensure model internal workings and output is explained in a way such that the results of the solution can be understood by not only the AI practitioners but also by the business and even the consumers. For example, if a credit card approval model rejects an application for a new credit card, the bank should be able to tell

the applicant about the factors resulting in rejection of his/ her application (such as current income, profession, past income, credit record, etc.)

A. Bias Detection Algorithms

Various algorithms to check for fairness in AI algorithms have got evolved. Some of the popular metrics to detect fairness in AI algorithms are as follows:

1. Equal Opportunity: 'Equal Opportunity' states that each group in the attribute under consideration should get True Positives at identical rates. This metric calculates the difference between the true positive rates (TPR) for underprivileged groups and the privileged groups. This metric ignores the False positives. It should be used as a definition of fairness only if the problem statement/ context requires to focus only or largely on the True positives and False Positives are not costly.

2. Equalized Odds: Just like 'Equal Opportunity', this metric state that each group in the attribute under consideration should get True Positives at identical rates. However, this metric also requires that the model should correctly identifies the false-positives at equal rates across groups. Many libraries use the average difference of FPR and TPR between unprivileged and privileged groups to calculate the 'Average Odds' difference as a metric to check for fairness/ bias. This metric is a more restrictive definition. It aims to equalize the 'True Positive Rate' (TPR) and 'False Positive Rate' (FPR) for each group. However, this may lead to model performance being degraded as it may fail to optimize accuracy on the majority group

3. Conditional Demographic Disparity: Before understanding Conditional Demographic Disparity (CDD), let's have a look at the definition of demographic disparity. Demographic disparity checks the proportion of the rejected candidates in the dataset along with the proportion of the selected candidates. If the proportion is unequal, then a bias is indicated. For instance, in case of a job vacancy, out of all candidates selected, black people may comprise 30% but out of all candidates rejected, black people comprise 50%, then we say that there is bias because the selection rate and rejection rate are unequal. This metric suggests that a predictor is unbiased if the prediction y is independent of the protected attribute p. The lowest value for this metric is -1 and the highest value is +1. To obtain a consolidated measure for disparities across all sub- groups, we can perform a weighted average on them (based on the proportion of the number of observations in each sub-group). This will give us what we can call 'Co nditional Demographic Disparity' or CDD [20].

Apart from these, there are many other fairness/ bias detection metrics used by many of the AI frameworks/ libraries. For example, IBM AIX supports metrics like Disparate impact, Theil Impact, and distance-based metrics like 'Euclidean Distance', 'Mahalanobis Distance', 'Manhattan Distance' to check for bias in AI/ ML models. Various open-source libraries in programming languages like R/ python are available to detect bias in AI/ ML models.

Demographic Disparity metric looks at the rejections and acceptance in the model output for each subgroup and determines whether any subgroup has a larger proportion of the rejected outcomes than the accepted outcomes. To make a useful inference, the demographic disparity may need to be seen for all subgroups. Conditional Demographic disparity (CDD) gives a single measure for all of the disparities found in the subgroups defined by an attribute of a dataset by averaging them. It is defined as the weighted average of demographic disparities (DD) for each of the subgroups, with each subgroup disparity weighted in proportion to the number of observations in contains.

Many data scientists, scholars, and courts treat it as a 'Gold Standard' for evaluating discrimination in automated systems. Amazon has also adopted the

usage of CDD as the metric to check for fairness in their AWS platform for AI & ML model development. However, the CDD metric ignores the bias within subgroups or among individuals as is the case with other group algorithms. CDD metric may also give encouraging results (indicating very little bias) in case the bias is there in some small sub-groups due to the small weightage assigned based on the proportion of the number of observations in the sub-group.

B. Bias Mitigation Algorithms

For mitigation of bias in AI/ ML models, various bias-mitigation matrices have got evolved. These algorithms can be applied at the 'pre-processing' stage, 'in-processing' stage, or 'post-processing' stage. 'Reweighing' [16], 'Optimized preprocessing' [13], 'Disparate Impact Remover' [7], and LFR [6] are some of the popular pre-processing bias mitigation algorithms. 'Adversarial Debiasing' [2] and 'Prejudice Remover' [14] are some of the popular in-processing bias mitigation algorithms. 'Equalized odds postprocessing' [8], 'Calibrated equalized odds postprocessing' [5], and 'Reject option classification' [15] are some of the popular post-processing bias-mitigation algorithms. Similar to bias detection, various open-source tools in programming languages such as python are available to mitigate bias for AI & ML models.

C. Bias Detection and Mitigation Toolsets

There are many toolkits/ libraries available which help in detecting and mitigating bias through various metrics. Some of the toolkits/ libraries focus on only the bias detection part and some are more comprehensive and cover bias mitigation and even explainability part as well.

FairML [1] is a toolkit that helps to perform an audit for the predictive models by analyzing the independent variables of the model and then quantifying their relative significance. With the help of FairML, predictive models can be easily audited to assess fairness.

FairTest [11] detects bias in a dataset by looking for correlations between the sensitive attributes and the dependent variable. The toolkit also provides access to a catalog of datasets.

Aequitas [3] is another promising toolkit. It supports multiple fairness metrics. It also provides a "fairness tree" which can assist users to find the appropriate metric based on the problem statement and the context. Themis [18] is a toolkit that produces automated test suites to detect two types of bias – Group, and Causal.

Themis-ML [19] is one such toolkit that provides various metrics for bias detection along with various bias mitigation algorithms (such as 're-labeling', 'reject option classification' etc.).

AIF360 [4] is a framework that comprises a comprehensive set of bias detection metrics and bias mitigation algorithms. It is an extensible framework. There are many real-world datasets referred by AIF 360 and other toolsets to help the data scientists better understand the metrics/ algorithms and also to prove their efficacy. Some of such datasets are Ricci, Adult Income, German Credit, and ProPublica Recidivism [10].

II. METHODS AND MATERIAL

It is critically important for a data scientist to correctly identify the direct & indirect attributes responsible for bias in the AI/ ML model. Identification of direct attributes is largely based upon local & global laws, guiding principles, organizational practices and the domain knowledge of the data scientist. However, it is also contextual and it is this contextual nature of discrimination that makes it difficult to define.

Detection and mitigation of bias caused by indirect attributes is an even bigger challenge.

While many algorithms have evolved to detect and mitigate bias in AI & ML models, removal of bias caused by indirect attributes has remained a challenge. Though, many of the algorithms handle bias caused by indirect attributes, these algorithms tend to ignore the small amount of bias caused by indirect variables.

However, when considered in totality, it becomes significant. This paper presents here a methodological approach to detect and quantify bias caused by indirect attributes in AI/ ML based systems.

Following steps may be performed to detect & quantify the bias using this methodology:

1. Identify Direct Protected/ Sensitive Attributes: Based on the legal/ social/ organization and contextual factors, find out the applicable sensitive/ protected attributes which are there in the modeling dataset (which is to be used for building the model) as well as the ones which are not even there in the modeling dataset. Include those attributes which are legally protected as well as those which are not legally protected but important from the perspective of the organization, social & ethical perspective.

2. Identify Attributes correlated with Direct Attributes: Find out the attributes in the dataset which are strongly correlated with the identified sensitive/ protected attributes. These are referred to as indirect sensitive/ protected attributes.

3. Calculate protected attribute indirect correlation value (PCIV): To start with, treat each of the protected attribute as confounding variable and find the strength of association (using Cramer's V value) between the Protected attribute and the dependent variable as well as each of the other attributes in the training dataset. For each of the protected attribute, attributes correlating with the protected attribute (p) beyond a defined statistically significant threshold (t) need to be identified. Cramer's V value of >= 0.1 may be taken as suggestive threshold (t) value. These are to be labeled as statistically significant indirect attributes (a). Now, the sum of Cramer's V correlation (wc) between (a) and the protected attribute (p) of the model needs to be calculated. Next, Cramer's V (dc) between (p) and the dependent variable (d) is calculated and correlation (pc) among each of the elements in (a) is calculated. Finally, Cramer's V (c) between each of the elements in (a) and the dependent variable is calculated

Now, we calculate the protected attribute indirect correlation value (PCIV) as weighted Cramer's V for the protected attribute against the dependent variable as

$\sum(wci) * (ci)$

Where i ranges from 0 to number of statistically significant indirect variables

PCIV value of 0.1 (or other cut-off value as per problem statement) will indicate indirect bias in respect to the said protected attribute

III. RESULTS AND DISCUSSION

In case of a Machine Learning algorithm for 'Loan Approval' for individual customers, assume that the following variables are there in the modeling dataset:

- Profession
- Gender
- Risk Rating
- Education

Here, 'Gender' is a legally protected attribute. After analysis, a strong correlation (say Cramer's V value of 0.3) was found between 'Gender' and 'Loan Approval'. On further analysis, we find statistically significant correlation between 'Gender' and 'Profession' as well as 'Gender' and 'Education'.

Following statistics is obtained:

- p = 'Gender'
- a = ['Profession', 'Education']

- d = 'Loan Approval'
- t = 0.1
- wc = [0.3, 0.4]
- dc = 0.2
- c = [0.2, 0.3]

PCIV = (0.3 * 0.2) + (0.4*0.3)
=0.2 + 0.06 + 0.12 = 0.18

As PCIV is greater than 0.1, we can say that bias indirectly caused by 'Gender' is present. Hence, appropriate bias mitigation algorithm needs to be applied to mitigate the indirect bias.

IV. CONCLUSION

Bias refers to an adverse act committed against a protected individual or group. Therefore, identification of attributes which are directly or indirectly cause of bias is important. While Identification of direct attributes is largely based upon local & global laws, guiding principles, organizational practices and the domain knowledge of the data scientist, identification of indirect attributes has remained a challenge. This paper has presented a methodological approach to detect indirect bias caused by protected attributes in AI/ ML models. The methodology is expected to yield superior results as it does not ignore the small bias caused by many indirect attributes and also does not require knowledge of privileged/ under-privileged attribute values. Apart from indirect bias detection, this methodology quantifies the indirect bias as well.

V. REFERENCES

[1]. Adebayo 2016, "FairML: ToolBox for Diagnosing Bias in Predictive Modelling", Massachusetts Institute of Technology (June 2016)

[2]. Zhang, Lemoine et al. 2018, "Mitigating Unwanted Biases with Adversarial Learning", AIES (Feb 2018)

[3]. Saleiro, Kuester et al. 2019, "Aequitas: A Bias and Fairness Audit Toolkit", arXiv (Apr 2019)

[4]. Bellamy, Dey et al. 2018, "AI Fairness 360: An Extensible Toolkit for detecting, understanding, and mitigating unwanted algorithmic bias", arXiv (Oct 2018)

[5]. Pleiss, Raghavan 2017, "On Fairness and Calibration", arXiv (Nov 2017)

[6]. Zemel, Wu et al. 2013, "Learning Fair Representations", PMLR (2013)

[7]. Feldman, Friedler et al. 2015, "Certifying and Removing Disparate Impact", International Conference on Knowledge Discovery and Data Mining (August 2015)

[8]. Hardt, Price et al. 2016, "Equality of Opportunity in Supervised Learning", arXiv (Oct 2016)

[9]. Zhang, Zhou 2019, "Fairness Assessment for Artificial Intelligence in Financial Industry", arXiv (Dec 2019)

[10]. Friedler, Scheidegger et al. 2018, "A comparative study of fairness-enhancing interventions in machine learning", arXiv (Feb 2018)

[11]. Tramer, Atlidakis et al. 2017, "FairTest: discovering unwarranted associations in data-driven applications", IEEE European Symposium on Security and Privacy (2017)

[12]. Sara Hajian et al. 2013, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining"

[13]. Calmon, Wei et al. 2017, "Optimized Pre-Processing for Discrimination Prevention", NIPS (Dec 2017)

[14]. Kamishima, Akaho et al. 2012, "Fairness-Aware Classifier with Prejudice Remover Regularizer", Springer-Verlag Berlin Heidelberg (2012)

[15]. Kamiran, Karim et al. 2012, "Decision Theory for Discrimination-Aware Classification", IEEE 12th International Conference on Data Mining (2012)

[16]. Kamiran & Calders 2011, "Data preprocessing techniques for classification without discrimination", Springerlink.com (2011)

[17]. Mehrabi, Morstatter et al. (2019), "A Survey on Bias and Fairness in Machine Learning", arXiv (Sep 2019)

[18]. Galhotra, Brun et al. (2017), "Fairness Testing: Testing Software for Discrimination", ESEC/FSE (2017)

[19]. Bantilan 2017, "Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation", arXiv (Oct 2017)

[20]. Wachter, Mittelstadt et al. 2020, "Why Fairness cannot be automated", arXiv (2020)

Cite this article as :

Ashish Garg, Dr. Rajesh SL, "PCIV method for Indirect Bias Quantification in AI and ML Models", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5, Issue 3, pp. 687-693, May-June-2019. Available at doi : https://doi.org/10.32628/CSEIT217251
Journal URL : https://ijsrcseit.com/CSEIT217251