Bias Mitigation in Machine Learning

WRJX35

***Abstract*—Machine learning models have become increasingly involved in the decision-making processes by companies in all industries. Whilst the use of these models can be of great value to their users, models trained on human centric datasets can inherit its potential underlying historical bias and thus have a discriminatory effect on some demographic groups.**

Keywords—Algorithmic Bias, Binary Classification, Discrimination, Machine Learning

I. Introduction

This essay aims to discuss the contribution of the paper [1] by Kamiran and Calders to the study of algorithmic bias and to present the writers views on the topic of algorithmic bias.

II. Paper Review

The paper by Kamiran and Calders [1] focuses primarily on the practical aspect of algorithmic bias within the scope of binary classifiers, proposing solutions which they claim mitigate the bias within a classifier and taking a theoretical study into the accuracy-discrimination trade off which occurs as a result of attempting to mitigate bias.

The authors proposed and discussed three bias mitigating solutions; massaging the dataset (a solution they proposed in an earlier paper [2]), reweighing the dataset and resampling the dataset. All of these approaches include modifying the training data in some way based on either probability values predicted, or observations made by a classifier already trained on this data that does not attempt to mitigate bias. The modification of the dataset does lead to a reduction in model accuracy, as is demonstrated in the authors' experimental results. However, it is also shown that the solutions are effective in reducing discrimination by a notable degree, making them a valuable contribution to this area of study.

A second valuable contribution made by the paper, which has already been briefly mentioned, is its theoretical study into the accuracy-discrimination trade-off by which these solutions are bounded. In this study it is proven that when using non-scoring-based classifiers, the best-case scenario is a linear trade-off between accuracy and discrimination. However, using rankers (scoring-based classifiers), which rank items according to the probability that they belong to the positive class, can produce models with a lesser trade-off by using different thresholds to classify the items based on their probabilities, depending on the value of the sensitive attribute, i.e. the underprivileged class would not need to be ranked with as high a probability as the privileged class in order to be classified with a positive class label. The conclusions drawn from this study provide a useful insight into the limitations of discrimination mitigating measures in machine learning, providing a foundation on which to base future attempts to create mitigation methods which aim to minimise the accuracy-discrimination trade-off.

The usefulness of the solutions proposed by the authors within real world applications is demonstrated within their experimental procedure, in which they test their models on various human-centric datasets. One such dataset, which is not extensively covered in the paper, is the German Credit Dataset [3]. This dataset contains records of German credit applicants labelled by whether they were assessed as *good* or *bad* risk and containing sensitive features such as *age* and *gender*. Assuming that the proposed solutions create models that perform in a similar manner on this dataset as they do on the Census Income Dataset, there could be a practical application for their use in consumer credit checks within the retail and banking industries. Such models are needed in these industries as there are many features related to the sensitive features *age* and *gender* that could affect the risk assessments. For example, the status of an applicant's checking and savings account is likely to be affected by their income, which in the case of people over the age of 30 can be expected to be higher on average than those under the same age due to progress within careers, higher wages etc. However, the classifier should not discriminate based on age, which it likely will if no bias mitigating measures are implemented, hence the importance of the solutions proposed in [1].

III. Personal view of Algorithmic Bias

With regards to classification models such as credit predictors and bail risk assessors, there are many effective methods of mitigating bias already available and have been for some time with [1] having been published as far back as 2012†. Since then there have been countless more papers released by other authors further studying the issue and attempting to create new mitigation methods, focusing on a variety of machine learning applications. However, there is still evidence of algorithmic bias in such classifiers having a serious negative impact on the lives of people in underprivileged demographics, one notable example being the Ofqual A-Level scandal[4] in 2020 which lead to thousands of students A-Level result being marked down as a result of bias against students from lower income areas. Given the amount of available research and literature around the topic, it seems reasonable to question whether instances like these are due to social issues as opposed to a lack of engineering solutions, i.e. are the developers of algorithms such as Ofqual's A-level's grade predictor building said algorithms with bias mitigation in mind or simply focusing on the models accuracy whilst remaining ignorant (or indifferent) to the historical bias they are subjecting their models to. Going forward it seems necessary to encourage developers to take an active approach to identify potential sources of bias within their data and minimise its effect. To achieve this it may be necessary to introduce further legislation and regulation into the machine learning industry, and more strongly enforce existing legislation such as the Equality Act 2010 [5] with increased fines for breaching it.

Of course, there are other areas of AI which are significantly more complex and are from having been researched exhaustively, such an example being that of natural language processing. As further progress is made into these areas of study, it is of vital importance that research is carried out with the intention of identifying and mitigating bias kept very much in mind. This will hopefully lead to more in-processing techniques for handling bias with regards to these problems, something that appears lacking in current NLP solutions.

† April 2021 at the time of writing

REFERENCES

[1] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1-33, 2011.

[2] F. Kamiran and T. Calders, "Classifying without discriminating", *2009 2nd International Conference on Computer, Control and Communication*, 2009.

[3] D. Newman, S. Hettich, C. Blake, and C. Merz, "(uci) repository of machine learning databases," 1998.

[4] B. Staton and L. Hughes, "Ministers under pressure to review A-level results", *Ft.com*, 2021. [Online]. Available: https://www.ft.com/content/12dad2d8-380c-435f-b689-8fddad19311d. [Accessed: 29- Apr- 2021]

[5] "Equality Act 2010", *Legislation.gov.uk*, 2010. [Online]. Available: https://www.legislation.gov.uk/ukpga/2010/15/contents. [Accessed: 29- Apr- 2021]