Simultaneous Improvement of ML Model Fairness and Performance by Identifying Bias in Data

Aakash Agarwal, Bhushan Chaudhari, Dr. Tanmoy Bhowmik
AI Garage Mastercard
(aakash.agarwal,bhushan.chaudhari,tanmoy.bhowmik)@mastercard.com

Abstract

Machine learning models built on datasets containing discriminative instances attributed to various underlying factors result in biased and unfair outcomes. It's a well founded and intuitive fact that existing bias mitigation strategies often sacrifice accuracy in order to ensure fairness. But when AI engine's prediction is used for decision making which reflects on revenue or operational efficiency such as credit risk modeling, it would be desirable by the business if accuracy can be somehow reasonably preserved. This conflicting requirement of maintaining accuracy and fairness in AI motivates our research. In this paper, we propose a fresh approach for simultaneous improvement of fairness and accuracy of ML models within a realistic paradigm. The essence of our work is a data preprocessing technique that can detect instances ascribing a specific kind of bias that should be removed from the dataset before training and we further show that such instance removal will have no adverse impact on model accuracy. In particular, we claim that in the problem settings where instances exist with similar feature but different labels caused by variation in protected attributes, an inherent bias gets induced in the dataset, which can be identified and mitigated through our novel scheme. Our experimental evaluation on two open-source datasets demonstrates how the proposed method can mitigate bias along with improving rather than degrading accuracy, while offering certain set of control for end user.

1 Introduction

AI powered predictive modeling techniques have been widely adopted by business verticals in different domains such as finance, healthcare, sports, banking, etc, often for making sensitive decisions ranging from personalized marketing, loan application approval (Mukerjee et al. 2002) [3] to dating and hiring process(Bogen 2018, Cohen 2019)[4,5]. Unsurprisingly, with the continuous evolution and ever-increasing complexity, there have been several recent high-profile examples of machine learning (ML) going wrong in terms of bias, fairness and interpretability.

The presence of unintended demographic disparities or differential/disproportionate impact on individuals by machine learning models is demonstrated by (Calders 2013) [29]. Unfairness can be imparted in models because of bias present in training data. Various types of bias such as annotation bias, historical bias, prejudice bias, etc may lead to unfair models and selective bias towards a particular group. In (Mehrabi et al. 2019)[7], the authors have provided a comprehensive coverage of such biases supportaed by real life examples.

The definitions used to understand the bias in models can be broadly categorized into three types: independence, separation and sufficiency (Sharma et al.2020)[10]. Specifically, a classifier satisfies independence if the protected attribute (such as race or gender) for which the model may be biased is independent of the classifier decision. Separation is satisfied if the classifier decision is independent

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

of the protected attribute conditioned on the true label. Sufficiency is satisfied if the true label is independent of the protected attribute conditioned on the classifier prediction. Details on these fairness criteria, both mathematically and with respect to different worldviews, may be found in (Barocas 2019, Yeom 2018) [27,28] along with definitions of fairness metrics, such as statistical parity difference for independence and average odds difference for separation, from (Garg et al. 2020, Bellamy et al. 2019, Sharma et al. 2020) [8,9,10]. We are taking statistical parity difference and average odds difference metrics into consideration for this paper while being aware of the fact that there are various fairness metrics which are relevant to gauge biasness of models. Determining the right measure to be used must consider the proper legal, ethical, and social context.

Using these fairness metrics, several bias mitigation algorithms are developed to satisfy the various criteria of fairness for machine learning models to reduce bias. Methods to mitigate bias generally fall into three categories. Pre-processing techniques transform the data so that the underlying discrimination is removed (Alessandro 2017)[21]. If the algorithm is allowed to modify the training data, then pre-processing can be used (Bellamy et al. 2018)[22]. Our proposed methodology falls under this category as explained in the subsequent sections. In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process (Alessandro 2017) [21]. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase (Bellamy et al. 2018)[22].

The literature extensively discusses the inherent trade-off between accuracy and fairness – as we pursue a higher degree of fairness, we may compromise accuracy (see for example (Kleinberg et al. 2017) [12]). Many papers have empirically supported the existence of this trade-off (Be-chavod 2017, Friedler et al. 2019) [13, 14]. Generally, the aspiration of a fairness-aware algorithm is to develop a model that is fair without significantly compromising the accuracy or other alternative notions of utility.

In this paper, we are proposing a method to find bias inducing samples in the dataset and then dropping these samples such that the pre-processed dataset represents a more equitable world. In an equitable world, model outcome is independent of the protected attributes (such as gender, race, etc). Our novel approach can be described as follows: given a dataset that contains a protected attribute (such as gender, race, etc), samples with similar attributes but different protected attributes and different outcomes are flagged. For example – Credit Risk dataset contains 2 samples: 1 male and 1 female such that male and female sample have same attributes but model predicted low risk for male and high risk for female. We establish that these instances induce bias as the attributes are same but the outcome is different due to dependence on protected attributes (such as male, female) and thereby result in unfair treatment by the model. Further, protected attributes are not used for modelling, so such samples can confuse the model as they have nearly have the same attributes but different label and thus can be viewed as pseudo label noise.

Hence our objective boils down to detect and remove such instances before training to make sure the resultant model is fairer. In the process we also show that such close instance removal does not compromise on the model performance, rather on the contrary, it improves the accuracy. This simultaneous improvement of model fairness and accuracy which are in contrast of each other, although seems to be astonishing, but we could provide rationale for this achievement using prior work of Frénay et. al. [16]. This prior art explains the affect of such noisy instances on model performance and claim that label noise hampers the performance of the classifier which is also backed by (Long 2008)[15].

To summarize, in this paper, we make the following contributions :

1. We propose a systematic way of identification of bias inducing instances as per our definition in the previous section, and their subsequent removal from training data.
2. We show that how the bias inducing instances removal ensures model fairness using the standard fairness metric.
3. Further we show improvement in model accuracy trained on the bias eliminated data along with justification.
4. We offer control in terms of adjustable hyperparameters to adjust fairness and accuracy as per the dataset and business requirements.

2 Proposed Methodology

We are proposing a method to filter out potential bias creating samples from the dataset based on certain similarity criteria. We present the complete methodology for the proposed solution in algorithm 1.

Algorithm 1 Function to get Unbiased Data

```
    GetUnbiasedData(data, protected attribute , privileged group, label, favourable label)
Input : Training data where protected attribute is one of the feature and label is target variable which will be used for prediction.
Output : Unbiased data/ Unbiased Model
Steps :
• Remove correlated features with protected attribute.
• Normalize the continuous features and one hot encode the categorical features.
• Prepare two different groups based on protected attribute and output label
    1)Samples with privileged protected attribute and favourable output label
    2)Samples with un-privileged protected attribute and non-favourable output label
• Calculate cosine similarity between group 1 samples to group 2 samples.
• Flag similar samples from both groups based on cosine similarity threshold.
• Rank these similar samples (flagged from previous step) as per count of samples it is similar to with opposite group such that samples with higher count are at the top.
• Remove the top k% similar samples from both groups.
• Apply reweighing technique on remaining instances to ensure same base rates with respect to protected attribute and output.
• Drop protected attribute and Fit any model of your choice.
```

All the respective evaluation and fairness metrics are calculated on the model outcomes of the test dataset. As discussed, hyperparameters – top k% instances to remove and minimum similarity score are tuned to generate the results.

3 Experiments

This section describes the experiment design and performance of our proposed methodology when tested on two open-source datasets. Our focus here is on the improved fairness and accuracy obtained when the filtered dataset is used to train standard prediction algorithms. To perform the experiments, we have selected two datasets commonly used in the fairness research: UCI Adult dataset [1] and German Credit dataset [2]. We have studied both these datasets for demonstrating bias with respect to gender as a protected attribute and male as the privileged group.

For the Adult dataset we had taken one group as all female instances having <50k income and other group as all male instances having >=50k income as here privileged group is male and favourable outcome is to have >=50k income. For German credit dataset we had taken one group as all female instances with bad risk value and other group as all male instances with good risk value. In this dataset, male is a privileged group and good credit risk is favourable outcome.

In the experiments performed, we had taken cosine similarity threshold as 0.99 to flag similar samples from both the groups. After that we had ranked all the flagged samples according to count of similar samples from other group and removed top k% instances. For values of k=1,2 results are illustrated in table 2. We have experimented with various classification algorithm such as XGBoost, LighGBM, Random forest and logistic regression to check the effectiveness of our proposed methodology. For each of the mentioned algorithm we found that the proposed method was able to increase accuracy as well as decrease biasness as shown in table 1.

		Adult Dataset			**German Dataset**		
Algorithm	Data	Accuracy	AOD	SPD	Accuracy	AOD	SPD
XGBoost	Raw	0.84	0.19	0.18	0.73	0.12	-0.03
	1 % removal	**0.85**	**-0.01**	**0.09**	0.74	-0.05	0.04
	2 % removal	0.85	-0.03	0.10	**0.76**	**0.01**	**0.04**
LightGBM	Raw	0.84	0.19	0.18	0.72	0.07	0.04
	1 % removal	0.84	-0.05	0.09	**0.74**	-0.06	0.01
	2 % removal	**0.85**	**-0.04**	**0.09**	0.73	**0.04**	**0.03**
Random Forest	Raw	0.82	0.19	0.18	0.70	0.03	0.05
	1 % removal	0.83	**-0.02**	**0.12**	**0.72**	**0.01**	**0.03**
	2 % removal	**0.84**	-0.15	0.14	0.70	0.09	0.05
Logistic Regression	Raw	0.76	0.22	0.09	0.69	0.15	0.22
	1 % removal	0.76	0.03	0.01	**0.69**	**0.08**	**0.04**
	2 % removal	0.76	**0.008**	**0.01**	0.69	0.16	0.06

Table 1 : Fairness and accuracy results on two open source dataset

	Adult Dataset		**German Dataset**	
Instance	Male	Female	Male	Female
Total	21790	10771	690	310
1 % removal	217	107	6	3
2 % removal	435	215	13	6

Table 2 : Number of samples removed from each dataset

4 Conclusion

From the results mentioned in the previous section, it is evident that after the removal of pseudo label noise i.e., instances with similar features but with different output label, model has become fairer. Pseudo label noise is one of the potential bias imparting instances that might have been generated due to bias at the time of data annotation. As we are removing the biased data from the training set which is responsible for making the model biased, hence a model built on such filtered dataset is unbiased and fair model. We believe that pseudo label noise instances are the ones that are responsible for confusing the model and thereby leading to the distortion and shift of decision boundaries. These instances can be viewed through the lens of label noise as discussed in the survey paper (Frénay et. al. 2014)[16] where authors mentioned that label noise is responsible for the decrease in performance of classifiers and removal of such instances is one of the methods to improve performance of such models.

In this paper, we present an advanced but simplistic data pre-processing and filtering based method to remove bias from the data accompanied by contrasting upswing in the machine learning model performance. It overcomes the limitation of existing methods for bias mitigation at the cost of model accuracy. Promising experimental results on two publicly available open-source datasets makes our research well grounded.

5 Future Scope

As part of the future scope of research, we intend to develop similar in-processing algorithm to remove psuedo label noise. In our current set-up, we have used the cosine similarity metric to find out the close instances. A survey of comprehensive set of similarity metrics and comparison of their impact on bias mitigation can be carried out further. We intend to apply the same methodology to image and other data modalities. Another future direction of exploration would be to try out parallel strategies other than flagging and removing similar instances from training data, such as changing labels of such instances and explore which strategy is suitable for a given context.

References

[1] UCI Machine learning data repository. 2021.Adult dataset - https://archive.ics.uci.edu/ml/datasets/Adult

[2] UCI Machine learning data repository. 2021. German credit dataset- https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[3] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. Interna-tional Transactions in operational research 9, 5 (2002), 583–597.

[4] Miranda Bogen and Aaron Rieke. 2018. Help wanted: an exam-ination of hiring algorithms, equity. Technical Report.and bias. Technical report, Upturn.

[5] Lee Cohen, Zachary C. Lipton, and Yishay Mansour. 2019. Efficient candidate screening under multiple tests and implica-tions for fairness. (2019). arXiv:cs.LG/1905.11361

[6] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In Advances in neural in-formation processing systems. 3315–3323.

[7] Ninaresh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristi-na Lerman and Aram Galstyan.2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[8] Garg, Pratyush Villasenor, John Foggo, Virginia. (2020). Fairness Metrics: A Comparative Analysis.

[9] R. K. E. Bellamy et al. 2019. "AI Fairness 360: An extensi-ble toolkit for detecting and mitigating algorithmic bias," in IBM Journal of Research and Development, vol. 63, no. 4/5.

[10] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf,Vinod Muthusamy, and Kush R. Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20).

[11] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venka-tasubramanian. 2016. On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016).

[12] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[13] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. , 1733–1782 pages.

[14] Sorelle A Friedler, Carlos Scheidegger, Suresh Venka-tasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Con-ference on Fairness, Accountability, and Transparency. ACM, 329–338.

[15] Philip M. Long, Rocco A. Servedio.2008. Random Classi-fication Noise Defeats All Convex Potential Boosters.

[16] Benoît Frénay, Michel Verleysen. 2014. Classification in the Presence of Label Noise: a Survey.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Aware-ness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). ACM, New York, NY, USA, 214–226.

[18] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Sil-va. 2017. Counterfactual Fairness. In Advances in Neural In-formation Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Gar-nett (Eds.). Curran Associates, Inc., 4066–4076.

[19] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fair-ness in learning: Feature selection for fair decision making. In NIPS Symposium on Machine Learning and the Law, Vol. 1. 2.

[20] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learn-ing. arXiv preprint arXiv:1901.10002 (2019).

[21] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to dis-crimination-aware classification. Big data 5, 2 (2017), 120–134.

[22] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018).

[23] Kamiran, F. and Calders, T. 2012. Data preprocessing tech-niques for classification without discrimination. Knowledge and Information Systems, 33(1):1–33.

[24] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing systems.

[25] B. H. Zhang, B. Lemoine, and M. Mitchell. 2018. "Miti-gating UnwantedBiases with Adversarial Learning," AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2018.

[26] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf,Vinod Muthusamy, and Kush R. Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 7 pages.

[27] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org.

[28] Samuel Yeom and Michael Carl Tschantz. 2018. Discrimi-native but Not Discriminatory: A Comparison of Fairness Defi-nitions under DifferentWorldviews. arXiv preprint arXiv:1808.08619 (2018).

[29] Toon Calders and Indr˙e Žliobait˙e. 2013. Why unbiased computational processes an lead to discriminative decision procedures. In Discrimination and privacy in the information society. Springer, 43–57.

[30] Pratik Gajane and Mykola Pechenizkiy. 2017. On formaliz-ing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017).