Effect of Feature Hashing on Fair Classification

Ritik Dutta*
dutta.ritik@iitgn.ac.in
IIT Gandhinagar, India

Varun Gohil*
gohil.varun@iitgn.ac.in
IIT Gandhinagar, India

Atishay Jain*
atishay.jain@iitgn.ac.in
IIT Gandhinagar, India

ABSTRACT

Learning new representations of data to reduce correlation with sensitive attributes is one method to tackle algorithmic bias. In this paper, we explore the possibility of using feature hashing as a method for learning new representations of data for fair classification. Using Difference of Equal Odds as our metric to measure fairness, we observe that using feature hashing on the Adult Dataset leads to 5.4x improvement in metric score while losing an accuracy of 6.1% compared to when the data is used as is.

ACM Reference Format:
Ritik Dutta*, Varun Gohil*, and Atishay Jain*. 2020. Effect of Feature Hashing on Fair Classification. In *7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020), January 5–7, 2020, Hyderabad, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3371158.3371230

1 INTRODUCTION

Machine learning is being increasingly used in settings where it might introduce algorithmic bias. Real-world data often has sensitive information that can be used (intentionally or otherwise) by companies to discriminate against specific groups of people. For example, if the learning problem is to predict the salary of a person, the use of sensitive features such as race or gender so that one group gains an unfair advantage is undesirable. While sensitive features can often improve the accuracy of machine learning models, regulations might prevent the use of these features when the systems are deployed in the real-world.

Weinberger et. al. [6] suggested feature hashing, also known as the "hashing trick" as a dimensionality reduction method and demonstrated the feasibility of their approach. Feature hashing allows for significant storage compression for parameter vectors, which is extremely useful whenever a large number of parameters with redundancies need to be stored within bounded memory capacity. In this paper we analyze the effect of feature hashing on the fairness constraints proposed by the authors of [5].

2 BACKGROUND

Feature hashing can be thought of as a mapping $\phi : \chi \rightarrow \mathbb{R}^m$ which *hashes* high dimensional input vectors x into a lower dimensional feature space $\mathbb{R}^m$. Further explanations for feature hashing can be found in [2, 6].

To measure fairness, we use the fairness constraint defined by the metrics Difference of Equal Opportunity (DEOp) and Difference of Equal Odds (DEOd) [5]. DEOp is defined for each output label. In a binary classification problem, we want the true positive rates of all classes of the sensitive variable to be equal. DEOp$^+$ sums the differences between each individual class's true positive rate and the average true positive rate of all classes. Similarly, DEOp$^-$ sums the difference over the true negative rates. DEOd is the average of DEOp$^+$ and DEOp$^-$. In an ideal condition, we want DEOd to be 0.

$$DEOp^* = \sum_{t\in S} \left| P\{\hat{y} = y | s = t, y = *1\} - \frac{1}{|S|} \sum_{t'\in S} P\{\hat{y} = y | s = t', y = *1\} \right|$$

where, y $\in \{-1, +1\}$ is the binary output label
$S \in$ {1,...,k} represents the sensitive feature
x belongs to the input space
$* \in \{-, +\}$ and
$\hat{y}$ = sign(f(x,s))

$$DEOd = \frac{DEOp^+ + DEOp^-}{2}$$

3 EXPERIMENTS AND RESULTS

For our experiments, we use the ADULT dataset[3] to perform a binary classification task of predicting if a person makes over $50,000 per year. Each sample contains 14 features like age, education and occupation, of which two of the features – namely race and gender – can be considered sensitive. We use Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel as our model. Classification accuracy is calculated for the train and the test set for the original data as well as the feature hashed dataset. The results are presented in Table 1.

Metric	Without Hashing	With Hashing
Test DEOd	0.10315	0.01904
Train DEOd	0.07200	0.05187
Test Accuracy	83.152	77.065
Train Accuracy	87.346	78.869

Table 1: Results (Lower is better for DEOd)

The results on the test set show that the DEOd metric score is 5.4x better than the case when the original data is used as is, while the classification accuracy is reduced by 6.1%. Prior works [1, 4, 7] suggest that learning new representations of original data can serve as a legitimate method to reduce algorithmic bias. Our initial results suggest that feature hashing might also serve as an acceptable method to reduce bias, while leading to only a minor drop in accuracy. In future, we plan to further concretize this intuition by conducting extensive experiments with multiple datasets and other techniques for learning feature representations.

4 ACKNOWLEDGEMENT

We would like to thank Prof. Anirban Dasgupta, IIT Gandhinagar for his help in this project.

*Authors contributed equally to this research.

CoDS COMAD 2020, January 5–7, 2020, Hyderabad, India

ACM ISBN 978-1-4503-7738-6/20/01…$15.00
https://doi.org/10.1145/3371158.3371230

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing Black-box Models for Indirect Influence. *Knowl. Inf. Syst.* 54, 1 (Jan. 2018), 95–122. https://doi.org/10.1007/s10115-017-1116-3

[2] Wenlin Chen, James T. Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. 2015. Compressing Neural Networks with the Hashing Trick. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 2285–2294. http://dl.acm.org/citation.cfm?id=3045118.3045361

[3] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[4] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[5] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking Advantage of Multitask Learning for Fair Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. ACM, New York, NY, USA, 227–237. https://doi.org/10.1145/3306618.3314255

[6] Kilian Q. Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford, and Alexander J. Smola. 2009. Feature Hashing for Large Scale Multitask Learning. *CoRR* abs/0902.2206 (2009). arXiv:0902.2206 http://arxiv.org/abs/0902.2206

[7] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 325–333. http://proceedings.mlr.press/v28/zemel13.html