CERIAS
The Center for Education and Research in Information Assurance and Security

Data Acquisition to Improve Machine Learning Fairness through Multi-Armed Bandit

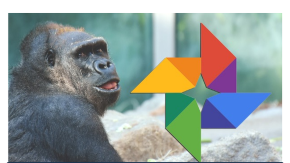
Jahid Hasan
Ph.D. Student
Computer and Information Technology
Purdue University

Dr. Romila Pradhan
Assistant Professor
Computer and Information Technology
Purdue University

Motivation

Over the past few decades, the extensive use of machine learning (ML) has shifted our focus from its implementation to its consequences. There have been several instances indicating bias in ML-based systems deployed in sensitive fields (e.g., law, finance, HR, etc.). These issues introduced the notion of fairness in ML models; as we continue to emphasize on equity, fairness is now deemed as important as precision. A key step toward achieving *fair* models is ensuring *fair* training data, which can be achieved by focusing on fair data acquisition.

Amazon's AI-based recruitment tool rejected female candidates despite having the same qualifications.



Google's image recognition algorithm labeled African Americans as gorilla.

COMPAS, a law enforcement AI tool, classified Hispanic females as highly likely to re-offend than others.

Images Source: www.google.com/serach?

Introduction

Machine learning has seamlessly integrated into every aspect of our lives, including healthcare, law, finance, and beyond. Due to ethical considerations and contemporary debate, the study of fairness in machine learning has become paramount[*1*]. Numerous studies have shown that biases in ML models often originate from biased training data, making data the root cause of the issue[*2*]. We can address this issue with existing data preparation studies[*3*].

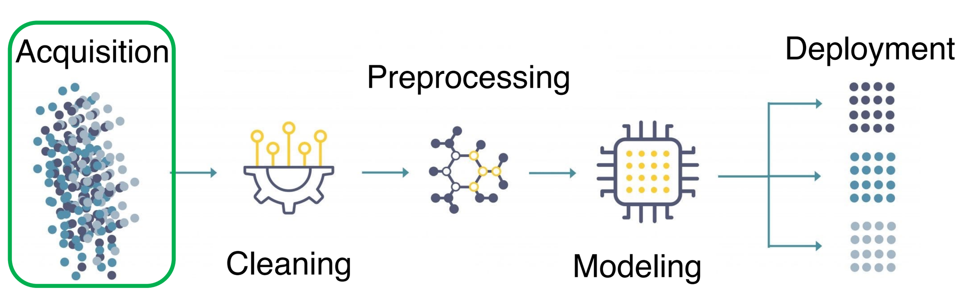


Figure 1: Machine learning Workflow

However, these approaches are problem-specific and can negatively impact downstream data usage. A more efficient approach would be to focus on earlier stages in the data science pipeline such as data acquisition (Figure 1), which can significantly improve the quality of downstream analyses. So far, in most research, data acquisition has only focused on accuracy[*4*]; we took the initiative to consider data acquisition for ML fairness. We address the following research question: **can ML model fairness be enhanced through data acquisition?** We employ a comprehensive solution for fair data acquisition that includes data source selection, merging sources, clustering data instances, and finally, adopting an approach based on multi-armed bandits to acquire data for improved model fairness.

Framework

Consider that we have a trained model M, with training dataset T with the statistical parity S_p. Our focus is to acquire data points for T to improve S_p. After selecting candidate datasets $D_1, D_2, \dots D_n$, we merge the dataset into D. Then cluster the dataset D into the optimal number of clusters $C_1, C_2, \dots C_m$. Later, Multi-Armed Bandit (MAB) has been used to acquire data points from each cluster and queried them. In each round, MAB selects a cluster based on its reward score and randomly collects a mini-batch of data, b from the cluster. MAB then merges this batch b with the existing training data as, $T \cup b$ and retrains the model. After that, it evaluates the value of S_p'. If S_p' improves than the past, update the training dataset as $T = T \cup b$. Otherwise, it will keep the training dataset unchanged.

The **Multi-Armed Bandit** is a classic exploration-exploitation tradeoff in reinforcement learning, where an agent must choose between exploiting the arms(options) with known high rewards and exploring other arms to discover even higher rewards potentially. These algorithms iteratively update their strategies based on observed rewards, gradually converging towards the most rewarding arm while still exploring to refine their knowledge.

**Statistical Parity: $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$*

Results

We used the Adult Income Dataset, which classifies whether an individual's income exceeds $50K$. This dataset has around $48K$ data points. As shown in Figure 2, we split the dataset through a smart-sampling(i.e., to get a biased model) approach to illustrate our method's efficiency. We used the Gaussian Model Mixture (GMM) to cluster the data pool. After training a logistic regression model, we observed the initial S_p.

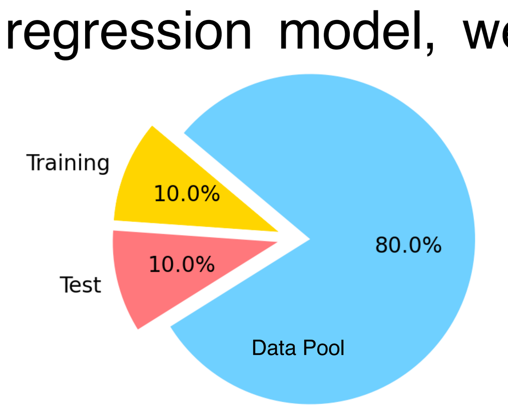


Figure 2: Dataset Split Illustration

Then, MAB was applied to the data acquired to improve fairness. As a baseline, we consider Random data acquisition. Data acquisition through MAB outperforms Random acquisition, as Figure 3 shows.

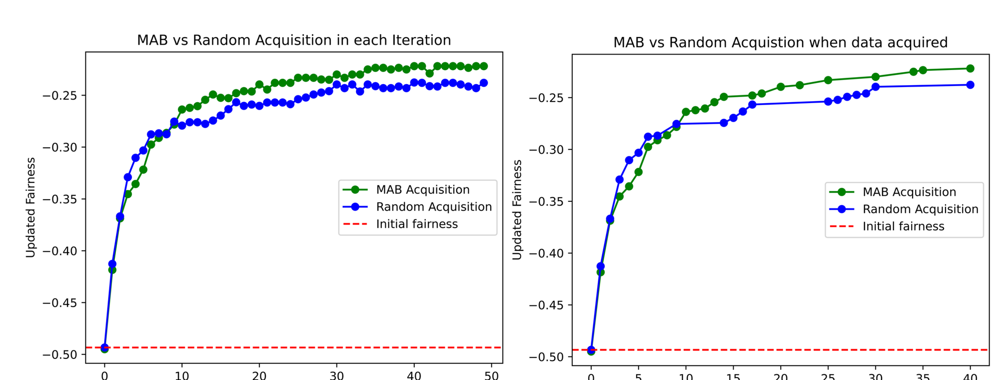


Figure 3: Observed fairness through MAB vs Random data acquisition

The left figure shows the overall observed fairness for each round, while the right graph shows at which points MAB acquired a mini-batch of data. Its monotonic growth indicates MAB only acquires data when it improves fairness compared to the initial and last data acquisition steps.

Discussion

Fair data acquisition is crucial to dealing with machine learning (ML) model biases. Our proposed approach involves clustering datasets based on their similar characteristics, which provides options for the algorithm. We use Multi-Armed Bandit to determine the exploration and exploitation probability after each iteration and select clusters based on their prior. This approach differs from random acquisition in that it takes prior knowledge into account. As shown in Figure 3, MAB is more consistent throughout the process. Each iteration updates its decision based on previous knowledge and seeks continuous improvements. While a mini-batch may not improve fairness, it tries to shift the cluster to have a new variant. In contrast, Random acquisition cannot maintain consistent improvement (for example, iteration 17 to 25) since it makes decisions agnostically.

Further Work

The proposed data acquisition approach exhibits superior performance compared to random acquisition. However, since existing clustering techniques often produce imbalanced clusters, the Multi-Armed Bandit (MAB) approach acquires data only from the dominant cluster, rendering it marginally better than random acquisition. To address this issue, we employ fair-balance clustering techniques to ensure fair and balanced data acquisition and to evaluate the acquired data, we focus on collecting batches of data according to their influence on fairness instead of randomly acquiring mini-batches from clusters.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD IConference* in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[2] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

[3] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl Inf Syst*, vol. 33, no. 1, pp. 1–33, Oct. 2012, doi: 10.1007/s10115-011-0463-8.

[4] C. Chai, J. Liu, N. Tang, G. Li, and Y. Luo, "Selective data acquisition in the wild for model charging," *Proc. VLDB Endow.*, vol. 15, no. 7, pp. 1466–1478, Mar. 2022, doi: 10.14778/3523210.3523223.

PURDUE UNIVERSITY
NSF
Google Research
CENTER FOR ADVANCING SAFETY OF MACHINE INTELLIGENCE
A collaboration between Northwestern University and UL Research Institutes
CERIAS