CERIAS

The Center for Education and Research in Information Assurance and Security

Valuation-based Data Acquisition to Improve Machine Learning Fairness

Ekta and Romila Pradhan {elnu, rpradhan}@purdue.edu

Motivation and Introduction

- ML models, while renowned for their accuracy, may introduce bias against certain groups, highlighting the need for fairness in algorithms.
- Prioritizing fairness in algorithm design is paramount to address potential biases inherent in ML models.
- Fairness of downstream ML models can be improved by acquiring the right set of data points. Data acquisition techniques have been studied well in the context of improving model accuracy, not so much for fairness.
- **Problem statement:** Determine additional data points that must be acquired to improve the model bias by the most

Background

Explored the Following existing approaches for data acquisition to improve model performance are explored to improve model fairness:

1. **Coverage based acquisition** [1]: Use data coverage as a proxy for model performance, determining the least amount of additional data to resolve the lack of coverage. Explored the impact of coverage to improve fairness of the datasets by computing the MUPs corresponding to a suitable threshold.
2. **SliceTuner** [4]: Provides the algorithm to selectively acquire data for the predefined slices to improve model accuracy. When considered for fairness, the blocker was the identification of predefined slices.

Our Solution

- **Key idea**: Acquire additional data points based on their discerned impact on model fairness or their potential to mitigate bias within the dataset.
- **Challenge**: To compute the impact of data points, there is a need to retrain the model iteratively after potentially acquiring each additional data point.
- **Solution**: Approximate the impact of data points by using data valuation techniques such as influence functions [2]:

$$\mathcal{I}_{\text{pert,loss}}(z, z_{\text{test}})^{\top} \overset{\text{def}}{=} \nabla_{\delta} L(z_{\text{test}}, \hat{\theta}_{z_\delta, -z})^{\top}\Big|_{\delta=0}$$
$$= -\nabla_{\theta} L(z_{\text{test}}, \hat{\theta})^{\top} H_{\hat{\theta}}^{-1} \nabla_x \nabla_{\theta} L(z, \hat{\theta})$$

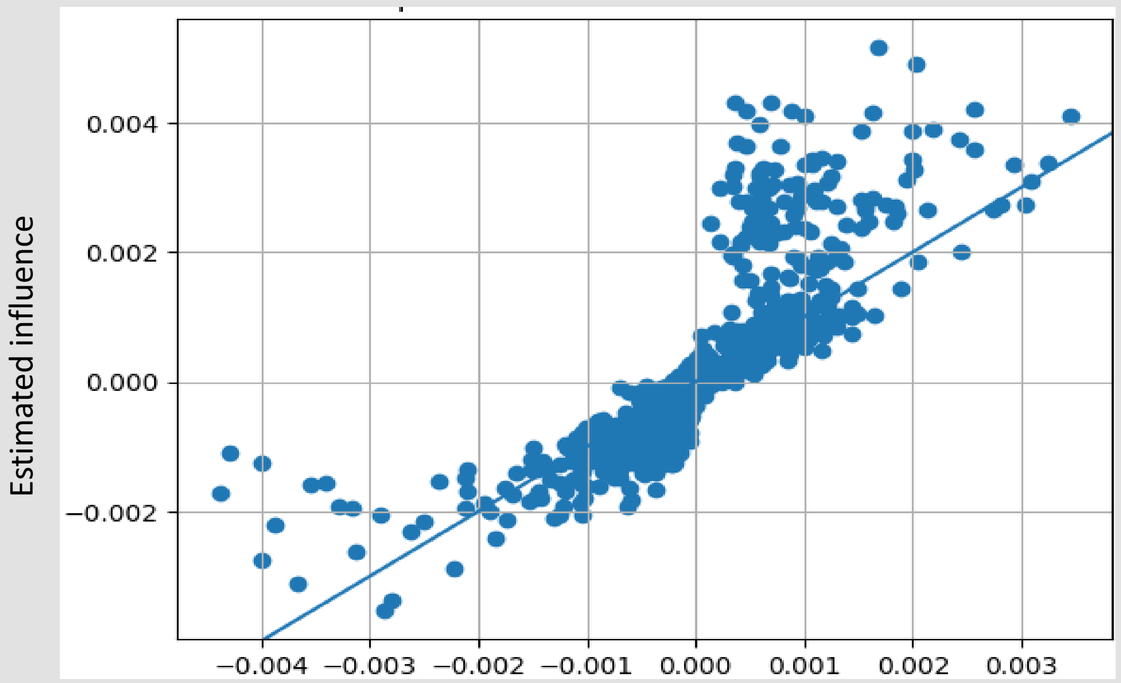
- **Assumption**: Access to a loss function that is convex and twice-differentiable
- **Overall approach**:
  1. Determine the valuation of a data point in terms of its effect on improving fairness using influence functions
  2. Arrange the data points in descending order based on their estimated impact.
  3. Acquire the data points in the order of their ranking, starting with the most impactful.

Experimental Setup

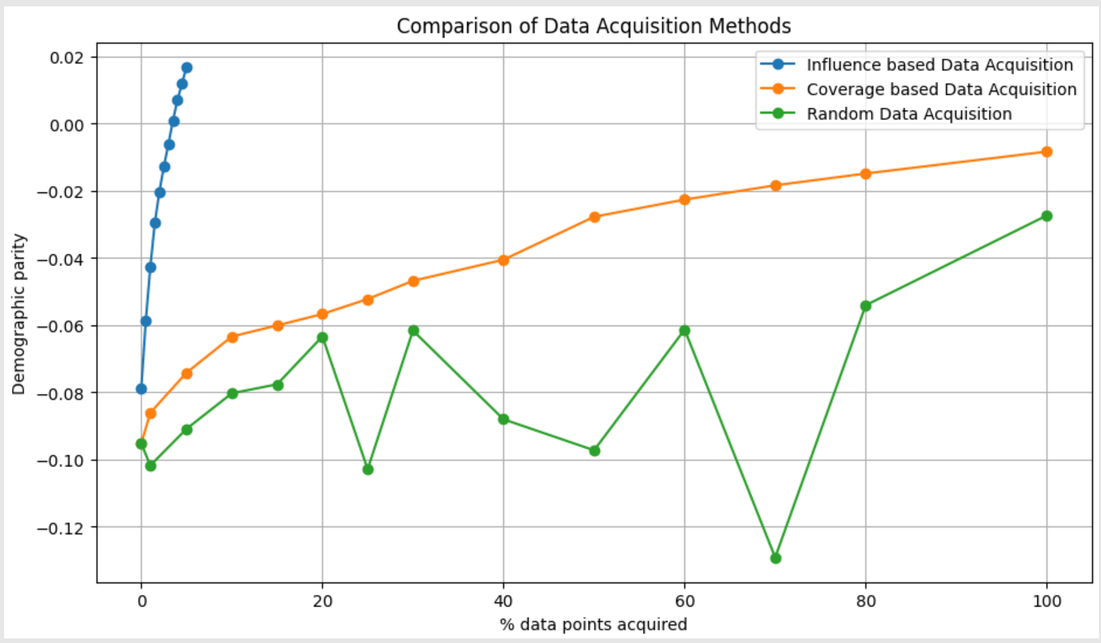
Dataset	Size	Sensitive Attribute	Prediction Task
German Credit [3]	1,000	Age	Good or Bad Credit
Adult Census	48,000	Gender	Income > 50K or <=50K
COMPAS	7,000	Race	Will an individual recidivate?

- Fairness metric – Demographic Parity [5]
- Model – Logistic Regression
- Prediction task – Binary Classification

Results: Can influence functions estimate fairness?



Experimental Results



Conclusions

- Valuation-based approach are effective in determining additional data points that should be acquired to reduce model bias without retraining the model.
- Compared to coverage-based methods, valuation-based methods are highly effective and efficient in acquiring additional data.
  - Coverage based acquisition shows a considerable decline in a demographic parity even after acquiring ~100% of data.
- Valuation-based solutions based on influence functions are limited to a special class of machine learning models.

References

[1] *Asudeh, A., Jin, Z., & Jagadish, H. (2019). Assessing and remedying coverage for a given dataset. In 2019 ieee 35th international conference on data engineering (icde) (pp. 554–565).*

[2] *Koh, P. W., & Liang, P. Understanding black-box predictions via influence functions. In D. Precup & Y. W. Teh (Eds.), Proceedings of the 34th international conference on machine learning (Vol. 70, pp. 1885–1894). PMLR. Retrieved from https://proceedings.mlr.press/v70/koh17a.html*

[3] *Dheeru Dua and Casey Graff. Uci machine learning repository, 2017*

[4] *Tae, K. H., & Whang, S. E. (2021). Slice tuner: A selective data acquisition framework for accurate and fair machine learning models. In Proceedings of the 2021 international conference on management of data (pp. 1771–1783).*

[5] *Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Proceedings of the international workshop on software fairness (pp. 1–7).*

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