Valuation-based Data Acquisition for Machine Learning Fairness

Ekta
Purdue University
West Lafayette, IN, USA
elnu@purdue.edu

Romila Pradhan
Purdue University
West Lafayette, IN, USA
rpradhan@purdue.edu

ABSTRACT

Machine learning algorithms are increasingly being used in critical application domains such as healthcare, finance, and criminal justice. Because of their opacity, fairness in machine learning has become a huge concern. While the value of data in model fairness has been extensively studied, recent research has highlighted the importance of intermediate stages of data science pipelines in ensuring model fairness. In this work, we focus on data acquisition, one of the earlier stages in the data science pipeline, as a potential bias mitigation technique. We present Inf-acq, a data acquisition approach based on the idea of data valuation, that determines the order in which additional data points should be acquired to reduce model bias. Inf-acq is based on the concept of influence functions to identify data points that have the maximum impact on model fairness. We empirically evaluate Inf-acq on three real-world and synthetic datasets and show that with a one-time offline computation, the fairness of machine learning models can be significantly improved while acquiring very few data points.

VLDB Workshop Reference Format:
Ekta and Romila Pradhan. Valuation-based Data Acquisition for Machine Learning Fairness. VLDB 2024 Workshop: International Workshop on Quality in Databases.

VLDB Workshop Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/ekta0596/DataAcquisition-valuation.

1 INTRODUCTION

Machine learning algorithms are becoming increasingly prevalent in critical domains such as finance, healthcare, and criminal justice and prevention. Due to the black-box nature of these algorithms, concerns continue to mount around the fairness of their decisions [1, 2, 13]. In particular, these systems perpetuate and amplify existing biases in the data used to train them. The need to quantify the discrimination and debias these systems has given rise to a number of fairness metrics [16, 20] and bias mitigation techniques (see [16] for a recent survey on fairness and bias in machine learning).

With the current wave of data-centric AI, researchers have recognized the importance of *data quality* for the fairness of machine learning decisions. Prior research mostly focused on pipelines comprising of training data, machine learning models, and test data. Recent research on data science pipelines [4, 5] has highlighted the importance of different stages of data science pipelines in combating fairness violations. In particular, earlier stages in the pipeline have been shown to significantly impact the downstream model fairness. Data science practitioners also reported feeling the most control over their data during the collection and curation stages, and resorted to collecting more data to offset any modeling issues [12].

This work focuses on *data acquisition*, which is one of the earlier stages in the data science pipelines, as a potential bias mitigation technique. The success of machine learning has taught us that the more the data, the more accurate the learned model. However, if a learned model generates unfair decisions, acquiring more data might not necessarily improve the model fairness. To ensure that models are not perpetuating existing biases, it is imperative to carefully consider the data used to train them.

Existing work in this space assumes knowledge of particular groups for which more data should be acquired [3, 19]. [19] begins with a few predefined slices, estimates their learning curves reflecting the benefit of potential data acquisition, and acquires more data for slices that have the *best* learning curves. Their definition of group fairness is based on the difference in model accuracies for the slices, and is not tailored to popular notions of associative group fairness. [3] views data acquisition through the lens of *data coverage* that attributes model bias to lack of representation of particular groups in the underlying training data. The hypothesis is that resolving the problem of coverage will reduce model bias; however, we show in our experiments that even after guaranteeing coverage for certain groups, model fairness is not guaranteed.

Our approach presents an alternative strategy that does not tie down model bias to data coverage and is applicable to existing notions of fairness. The core idea follows *goal-oriented* data acquisition which evaluates data points based on their direct impact on downstream model fairness. The naïve approach computes the impact of each training data point on model fairness by acquiring another such data point, adding it to the current training dataset, retraining a model on the updated data, computing the fairness of the new model and comparing it with fairness of the model trained on the original data. It then ranks data points in decreasing order of this computed impact and acquires the one having the maximum impact on model fairness. This approach is computationally expensive for large datasets and complex models.

We present Inf-acq, an algorithm that assesses data points by computing their *estimated* impact on downstream model fairness. Inf-acq uses the concept of *influence functions* [8, 15] to approximate the first-order impact of acquiring a data point on model fairness. It then selects a data point acquiring which maximizes the gain in model fairness. Influence functions have been successfully used for training data debugging for poor model performance (in terms of accuracy [22] and fairness of model decisions [17]); however, their

Proceedings of the VLDB Endowment. ISSN 2150-8097.

effectiveness in data acquisition for addressing model fairness has not been studied before.

The paper is organized as follows. Section 2 introduces the data notation and preliminaries for the rest of the paper. Section 3 presents our method for data acquisition to mitigate model bias. We present the results of our experimental evaluation in Section 4. Section 5 discusses the existing related work. We discuss our conclusions and future work in Section 6.

2 PRELIMINARIES

We present relevant background information on classification and algorithmic fairness.

Classification. We consider the problem of binary classification. Consider a training dataset $\mathsf{D} = d_i^n = \{x_i, y_i\}_{i=1}^n \in \mathsf{Dom}(\mathbf{X}) \times \mathsf{Dom}(Y)$ where $x_i \in \mathsf{Dom}(\mathbf{X})$ with $\mathbf{X}$ denoting a set of features, and $y_i \in \mathsf{Dom}(Y) = \{0, 1\}$ denoting a binary label to be predicted. The objective of binary classification is to train a classifier $h : \mathsf{Dom}(\mathbf{X}) \rightarrow \hat{Y}$ on D such that each data point $\mathbf{x}$ has an associated predicted label $\hat{y} = h(\mathbf{x}) \in \{0, 1\}$. Classifiers use a learning algorithm that trains on D to learn the optimal parameters $\theta^* \in \mathbb{R}^p$ that minimize the empirical loss $\mathsf{L}(\mathsf{D}, \theta) = \frac{1}{n} \sum_{i=1}^n \mathsf{L}(d_i, \theta)$. We consider learning algorithms that use a loss function L that is strictly convex and is twice-differentiable. In this paper, we focus on logistic regression, which is one of the simplest such classifiers.

Algorithmic group fairness. Given a binary classifier $h : \mathbf{X} \rightarrow \hat{Y}$ and a protected attribute $S \in \mathbf{X}$ (such as gender, race, age etc.), we interpret $\hat{Y} = 1$ as a favorable (positive) prediction and $\hat{Y} = 0$ as an unfavorable (negative) prediction. We assume the domain of S, $\mathsf{Dom}(S) = \{0, 1\}$ where $S = 1$ indicates a privileged and $S = 0$ indicates a protected group (e.g., males and non-males, respectively). Algorithmic group fairness mandates that individuals belonging to different groups must be treated similarly. The notion of similarity in treatment is captured by different associative notions of fairness such as demographic parity, predictive parity and equalized odds [7, 16, 20]. We focus on demographic parity (also known as *statistical parity*), which is a widely used notion of group fairness. A classifier h satisfies statistical parity if both the protected and the privileged groups have the same probability of being predicted the positive outcome i.e., $P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1)$. We denote the chosen fairness metric by f and quantify the fairness in the predictions of a model trained on D by f_{D}. For example, in the case of demographic parity, $f_{\mathsf{D}} = P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)$ quantifies the difference in the probabilities of the protected and privileged groups having a positive outcome. If $f_{\mathsf{D}} < 0$, then the model is biased against the protected group, while $f_{\mathsf{D}} > 0$ indicates that the model is biased against the privileged group. The higher $|f_{\mathsf{D}}|$, the more unfair the model's predictions.

Problem Statement. Given classifier h trained on D and fairness metric f, we address the problem of determining the order in which additional data points $\mathsf{D}_{acq} \subset \mathsf{D}$ should be acquired such that the model learned on $\mathsf{D} \cup \mathsf{D}_{acq}$ is fairer than that learned on D (i.e., $f_{\mathsf{D} \cup \mathsf{D}_{acq}} < f_{\mathsf{D}}$).

3 VALUATION-BASED RANKING

In this section, we describe our approach to rank data points in D in the order they should be acquired to improve model fairness. Our approach is built upon the idea of the impact of acquiring data points on the fairness of the downstream model and ranking data points in decreasing order of their computed impact.

We denote the impact of data point $d_i \in \mathsf{D}$ as I_i and naïvely quantify it as the difference in the fairness metrics when a model is trained on $\mathsf{D} \cup d_i$ and when the model is trained on D, i.e.,

$$\mathsf{I}_i = f_{\mathsf{D}} - f_{\mathsf{D} \cup d_i} \tag{1}$$

Acquiring a data point with a positive impact $\mathsf{I}_i > 0$ results in a model with lower disparity (unfairness). On the other hand, a data point with a negative impact $\mathsf{I}_i < 0$ indicates that the model learned by acquiring d_i has higher unfairness. A data point with a higher positive impact lowers the disparity more than one with a lower positive impact and thus is more desirable for acquisition. Thus, by simply ranking data points in decreasing order of their impact, we can determine which data points are valuable for acquiring.

Note that to compute I_i (as in Eq. (1)), we need to compute fairness on a new model trained on $\mathsf{D} \cup d_i$. Ranking data points in decreasing order of their impact requires training $|\mathsf{D}|$ new models (one for each data point in D). This process is, therefore, computationally prohibitive for large datasets and complex machine learning models. In the following, we present our approach that ranks data points without training a new model for each data point.

3.1 Impact Estimation without Model Training

To approximate the impact of a training data point on model fairness without retraining the model, we utilize the concept of *influence functions* [8, 15] and present Inf-acq, an algorithm that ranks data points for acquisition using influence functions.

Given that θ^* minimizes empirical risk, i.e., $\theta^* = \mathrm{argmin}_{\theta \in \Theta} \mathsf{L}(\mathsf{D}, \theta) = \mathrm{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \mathsf{L}(d_i, \theta)$, we denote the gradient of the loss function by $\nabla_\theta \mathsf{L}(\theta)$ and its Hessian matrix by $\mathsf{H}_\theta = \nabla_\theta^2 \mathsf{L}(\mathsf{D}, \theta) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta^2 \mathsf{L}(d_i, \theta)$. Since $\mathsf{L}(\mathsf{D}, \theta)$ is convex and twice-differentiable (Section 2), H_θ is positive definite and therefore, H^{-1} exists.

The influence of up-weighting training data point $d_i \in \mathsf{D}$ by a small amount ϵ on the model parameters θ is then computed as:

$$\mathsf{Inf}_\theta(d_i) = \frac{1}{n} \frac{d\theta_\epsilon^*}{d\epsilon} = \left(\frac{1}{n}\right) \left(-\mathsf{H}^{-1} \nabla_\theta \mathsf{L}(d_i, \theta^*)\right) \tag{2}$$

In other words, $\mathsf{Inf}_\theta(d_i)$ computes the difference in model parameters *after* and *before* up-weighting d_i. More details on influence functions can be found in [15].

We use influence functions to *estimate* the impact I_i of data point d_i by approximating the change in the fairness metric f due to addition of d_i through the chain rule of differentiation. Combining Equations 1 and 2, we can estimate the change in fairness as:

$$\mathsf{I}_i = -\frac{d(f(\theta_\epsilon^*))}{d\epsilon} = -\frac{d(f(\theta_\epsilon^*))}{d\theta} \frac{d\theta_\epsilon^*}{d\epsilon} = -\nabla_\theta f(\theta^*) \mathsf{Inf}_\theta(d_i) \tag{3}$$

Note that since we define impact (Equation 1) as the difference between fairness *before* and *after* acquisition, there is an additional negative sign in the computation of I_i above.

Having estimated the impact of data points, we rank them in decreasing order of their estimated impact on fairness and acquire additional data points in that order. Algorithm 1 outlines the steps for Inf-acq, a valuation-based data acquisition method through the use of influence functions as in Equation 3.

Algorithm 1 Inf-acq Algorithm

Input: Dataset D, binary classifier h, fairness metric f
Output: D_{ord}, ordered list of data points to acquire

for each $d_i \in D$ **do**
    Compute estimated impact I_i as in Equation 3
end for
I= SORT-REVERSE(I) /* sorted impact of data points */
$D_{ord} = \{i$ **for** $I_i \in I\}$ /* data points sorted acc. to impact */

Complexity. For each data point, Inf-acq estimates its first-order approximate impact on model fairness. Once Hessian matrix and its inverse are computed *offline* in $O(np^2 + p^3)$ operations, the impact of a data point is computed in $O(p)$ operations (for calculating the gradient of the loss function with respect to p model parameters).

4 EXPERIMENTAL EVALUATION

We evaluate the effectiveness and efficiency of Inf-acq through experiments on three real-world and synthetic datasets. Our objectives are: (1) to assess the effectiveness of valuation-based data acquisition in improving model fairness, and (2) to evaluate the efficiency of the approach compared to other methods.

4.1 Experimental Setup

Datasets. We evaluated our methods on the following standard datasets in the fair ML literature:

Adult Census Income [10]. This dataset contains demographic and financial information of 48,844 individuals, the sensitive attribute is sex, and the prediction task determines whether an individual has annual income $\leq 50k$ or $> 50k$.

German Credit [10]. The dataset contains financial information of 1,000 individuals, the sensitive attribute is age, and the prediction task is to determine whether an individual is a good credit risk or a bad credit risk.

COMPAS [13]. This dataset contains demographic and criminal information on 7,214 defendants. The sensitive attribute is race, and the prediction task determines whether a defendant is at a high/low to re-offend in the next two years.

Fairness metrics. We consider demographic parity [20]. However, our approach is applicable to other associative notions of fairness e.g., predictive parity, equalized odds, true predictive parity etc.

Competing methods. We consider the following three methods of data acquisition:

Random: This method randomly selects a data point for acquisition, and considers all data points as equally important.

DeepDiver [3]: This method determines *maximally uncovered patterns*, which are predicate-based subsets that do not have enough coverage (as indicated by threshold τ). DeepDiver implements a *hitting* algorithm to identify data points that should be acquired such that most of the maximally uncovered patterns are covered. (Table 1 shows the maximally uncovered patterns for a number of thresholds across different datasets).

Inf-acq (Algorithm 1): This method ranks data points in decreasing order of their estimated impact on the downstream model fairness and acquires data in that order.

Dataset	τ	Maximal Uncovered Patterns (# of data points needed)
German	10	XXXXXXXXXXXXXXX00 (3)
	50	XXXXXXXXXXXXXXX0X (13)
	72	XXXXXXXXXXXXXXX0X (35)
	75	XXXXXXXXXXXXXXX0X (38)
COMPAS	10	XXXXX1202 (2)
	50	XXXXXX201 (41)
	100	XXXXXX202, XXXXXX201, XXXXXX002, XXXXXX000 (189)
	150	XXXXXX2X2 (22)
	200	XXXXXXX02 (10)
	500	XXXXXXX02, XXXXXXX01 (409)
	1,000	XXXXXXX02, XXXXXXX01, XXXXXXX00 (1,605)
Adult	500	XXXXX001 (286)
	1,000	XXXXX0X1 (259)
	2,000	XXXXXX01 (340)
	8,000	XXXXXX11, XXXXXX01 (6,803)
	10,000	XXXXXXX1 (803)
	20,000	XXXXXXX1 (10,803)

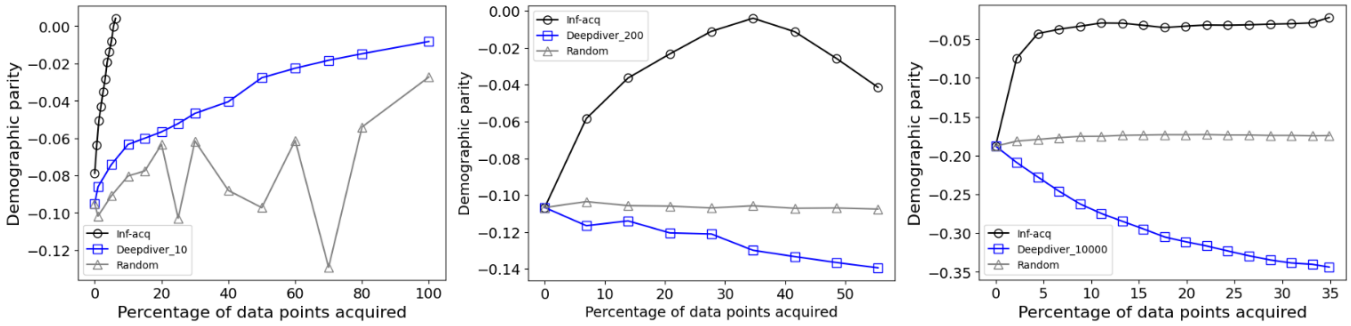
Table 1: DeepDiver: Maximal uncovered patterns identified for given threshold τ. Number of additional data points to be acquired to cover the identified patterns shown in brackets.

Performance metrics. *Effectiveness*: To evaluate the effectiveness of the proposed approach, we acquired data points individually in the order specified by a method and recorded the fairness metric upon each acquisition. A value closer to 0 indicates that the acquired data is fairer and less biased. *Efficiency*: To evaluate the efficiency of a method, we report the average offline and online time it takes to determine the next data point to be acquired.

4.2 Effectiveness of Inf-acq in reducing bias

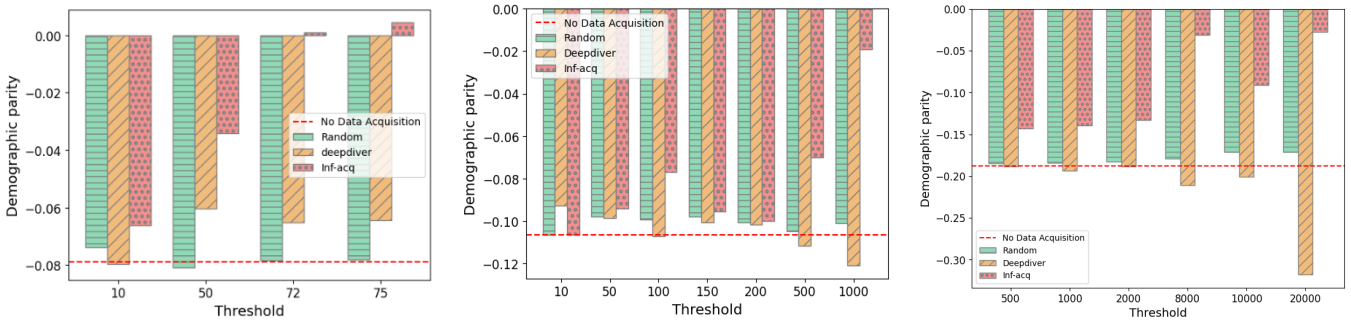
In this section, we evaluate the effectiveness of the aforementioned three competing methods in improving the fairness of the downstream machine learning model. To compare our method with DeepDiver and Random, we conduct two set of experiments. In the first experiment, we acquire data points through each method within a budget (e.g., x% of data points). In the second experiment, each method acquires the number of data points as indicated by the threshold in DeepDiver (Table 1). In both the experiments, we report the bias of the model trained on the data updated after acquisition.

4.2.1 Data acquisition on a budget. Our goal in this experiment is to assess the methods in terms of the number of additional data points acquired to reduce bias. We demonstrate in Figure 1, the gradual improvement in fairness (demographic parity in y-axis) for increasing number of acquired data points (x-axis) for all the methods. Initially, across datasets demographic parity is negative indicating that the learned model is biased against the protected group. The closer the line for a method is to the top horizontal axis, the fairer the learned model is. For example, in Figure 1a, Inf-acq reduces bias from -8% to 0% with around 4% of additional data points acquired. For DeepDiver, we choose a threshold of $\tau = 50$, 200 and 10,000 for German, COMPAS, and Adult respectively. From Table 1, note that DeepDiver identifies a single maximal uncovered pattern for each of these thresholds; we report the fairness metric after acquiring x% of data points satisfying the identified pattern. In case there are multiple maximal uncovered patterns (e.g., $\tau = 100$ for COMPAS), we acquire the respective number of data points satisfying either pattern. Across datasets, as more data points are acquired, Inf-acq consistently reduces bias (Figure 1). With German,



(a) German Credit (b) COMPAS (c) Adult Census Income

Figure 1: Comparison of different data acquisition techniques. While Inf-acq consistently reduces model bias, Random and DeepDiver do not always improve model fairness. DeepDiver's subscript denotes threshold τ — the minimum number of data points that a subset should have.



(a) German Credit (b) COMPAS (c) Adult Census Income

Figure 2: Comparison of different data acquisition techniques with different DeepDiver thresholds. As more data points are acquired, Inf-acq has considerably more reduction in model bias compared to Random and DeepDiver since it acquires data points in a targeted manner. While Random acquires data points in random fashion, the goal of DeepDiver is to resolve data coverage and hence it acquires those data points which do not have enough representation in the dataset.

model bias is completely removed by acquiring merely 4% of additional data points. COMPAS requires acquiring 35% of data points to remove bias but reduces bias by around 64% with acquisition of a mere 15% additional data points. The identified data points belong to the protected group with a favorable outcome, and by acquiring them, we ensure that the learned model breaks the dependency of the outcome on sensitive attribute race. In contrast, bias is not completely removed in Adult although there is a rapid decrease in bias (from −18% to −4%) with just 5% of additional data points. Model bias attains a minimum of −2% with the acquisition of 35% additional data points. This behavior is expected since Inf-acq estimates the effect of acquiring data points on model fairness, and its ranking thus identifies data points that positively impact bias. Random does not exhibit any steady improvement in bias; this behavior is expected since it randomly chooses a data point for acquisition and may select one that does not have change model bias at all. DeepDiver improves fairness for German but for the other two datasets, acquiring more data points worsens the model fairness. This observation is explained by DeepDiver's underlying hypothesis — data coverage is the indirect root cause of model bias — that does not guarantee that model bias will be reduced even after enough data points are gathered for the identified maximal uncovered pattern. Instead of focusing on patterns, Inf-acq identifies individual data points acquiring which would ensure that model bias will be reduced.

Takeaways: (1) Valuation-based data acquisition is effective in reducing model bias by acquiring very few data points. (2) Coverage-based data acquisition does not ensure that model bias will always be reduced by acquiring additional data points.

4.2.2 Data acquisition based on DeepDiver threshold τ. In this experiment, we compare Inf-acq with DeepDiver for varying levels of threshold τ. Our goal is to evaluate whether DeepDiver's coverage-based method has better reduction in model bias for any chosen threshold τ. For each dataset and τ, the number of additional data points that should be acquired to cover the maximally uncovered pattern is shown in Table 1. Correspondingly for Random and Inf-acq, we acquire the same number of data points and report the results in Figure 2. The x-axis represents the threshold and y-axis shows the fairness metric after the required number of data points (shown in Table 1) is acquired by each method. The dotted red line

denotes the base fairness when no data is acquired. We observe that across datasets, Inf-acq consistently exhibits superior performance compared to DeepDiver and Random. In particular, as the number of data points acquired increases (e.g., $\tau = 500, 1000$ in Figure 2b and $\tau = 8000, 10000, 20000$ in Figure 2c), Inf-acq shows the maximum reduction in bias (reducing from -18% to -3% in Adult). While Random rarely improves model fairness since data points are randomly chosen for acquisition, DeepDiver exhibits very small reduction in bias compared to Inf-acq. In some cases, acquiring data points identified by DeepDiver increase bias (e.g., $\tau = 1000$ in Figure 2b and $\tau = 20000$ in Figure 2c), thus reinforcing the observation that resolving data coverage does not equate to resolving model bias. Note that in some cases ($\tau = 72, 75$ in Figure 2a), the data points acquired by Inf-acq results in bias in the opposite direction, i.e., the learned model becomes biased *toward* the protected group (although the magnitude of the bias is very small — less than 1%). This flip of model fairness occurs because as more data points are acquired, those lower in the ranked list have a negative estimated impact on bias, and acquiring those data points increases bias by a small amount.

Takeaways: (1) As more data points are acquired, valuation-based data acquisition Inf-acq exhibits significant reduction in model bias. (2) Coverage-based data acquisition DeepDiver has little to no reduction in model bias in most cases.

4.3 Efficiency of Inf-acq

Our goal in this set of experiments is to evaluate the scalability of Inf-acq to larger datasets. In Table 2, we show the time taken by the different methods in selecting the appropriate data points for acquisition. Retraining denotes the method that computes the difference in model fairness before and after acquiring a data point, and acquires data in decreasing order of this difference (i.e., a data point that reduces model bias by the most is acquired first). The time for DeepDiver includes time taken to determine the maximally uncovered patterns and selecting data points matching the identified patterns. The time for Inf-acq presents the offline runtime metrics for Hessian calculation, which contributes the most to Inf-acq time.

Dataset	Retraining (s)	DeepDiver (s)	Inf-acq (s)
German	21.5	0.1 ($\tau = 10$)	3.7
COMPAS	567.8	1.2 ($\tau = 200$)	17.7
Adult	6915	4.8 ($\tau = 10000$)	132.2

Table 2: *Offline* time taken (in seconds) to acquire data points for the indicated τ. Time for DeepDiver includes the time taken to find the maximally uncovered patterns and find data points matching the patterns. Time for Inf-acq shows the one-time *offline* computation for the Hessian matrix.

We observe that while DeepDiver takes the least amount of time, it is not effective in reducing model bias (as seen in Figures 1 and 2). Inf-acq incurs a one-time cost of Hessian computation, which (as expected) is dependent on the dataset size. Once this offline computation is done, the impact of each data point is computed in less than a second and data points are ranked in decreasing order of their impact and acquired in that order. Note that Inf-acq is faster than Retraining by orders of magnitude — this performance and its effectiveness (Figures 1 and 2) bolster the importance of valuation-based data acquisition in reducing model bias.

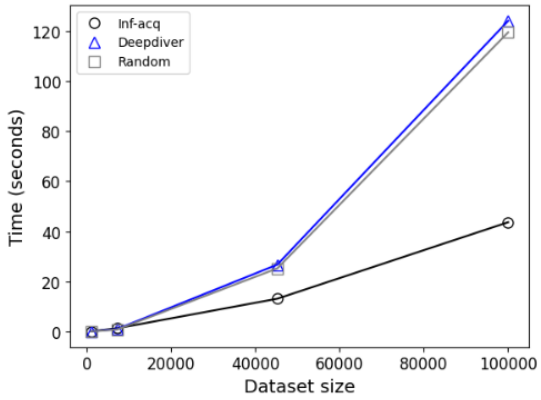


Figure 3: Efficiency of the methods for varying dataset sizes. Time shown includes the *online* computation time taken for data acquisition and computing model fairness.

In Figure 3, we present the efficiency of the competing methods for datasets of varying sizes in terms of both acquiring additional data points and computing the fairness of the updated model trained on the original and acquired data. Note that this plot shows the *online* computation times. We observe that as the dataset size increases, there is a general trend of increasing runtime across all methods. Inf-acq consistently exhibits the shortest runtime across all dataset sizes, which is because once the Hessian is computed offline, Inf-acq simply iterates over the data points in order of their estimated impact on model bias. Random shows a noticeable increase in runtime as the dataset size grows because for each additional data point, it chooses one at random from the remaining data points. DeepDiver takes the most time because it acquires data points corresponding to the maximally uncovered patterns for a given threshold which includes identifying data points that match the maximally uncovered pattern. **Takeaways:** (1) Inf-acq efficiently scales up to large datasets in the online computation of data points that should be acquired. (2) The one-time offline computation of Inf-acq, while slower than DeepDiver, is orders of magnitude faster than naïve model retraining which has an extreme computation cost and cannot be used on large datasets.

5 RELATED WORK

The study in this paper is related to the following three research areas: algorithmic fairness, data quality, and data valuation. While these areas were studied extensively, our approach of using data valuation for acquiring high-quality data to ensure fairer algorithmic systems is novel.

Algorithmic fairness. With the increasing prevalence of machine learning in critical domains, such as criminal justice, healthcare, and finance, ensuring fairness of machine learning decisions is of paramount importance. Recent examples of fairness violations or bias in algorithmic systems include discrimination based on race, skin tone, zipcodes, and perceived gender [1, 2, 9, 14]. A number of bias mitigation techniques have been introduced [16] that can be

categorized into pre-processing, in-processing, and post-processing techniques. Not only are pre-processing techniques among the easiest and the most effective solutions, data science practitioners have also reported data collection and curation stages to be their preferred solution to offset any modeling issues [12].

Data acquisition techniques. Data discovery [21], source selection [18], and data acquisition [19] are classical data management research problems that emphasize the importance of high-quality data in data-driven decision making. This work spotlights *data acquisition* – the process of acquiring high-quality data for downstream analyses – as a potential pre-processing bias mitigation technique. Since data acquisition is one of the earlier stages in the data science pipeline, recent efforts [3, 6, 19] that ensure the right data is collected holds promise for learning accurate and fair machine learning models. SliceTuner [19] acquires (possibly) different amounts of data for given few slices (data subsets) by estimating the learning curves of the slices and using them to compute the cost benefits of acquiring more data for those slices. DeepDiver [3] views representation bias as the root cause of model bias and seeks to ensure enough coverage of all slices. AutoData [6] adopts a multi-armed bandit approach and reinforcement learning to determine which data points should be acquired from a data pool curated from the wild; however, this approach necessitates retraining a learned model after each iteration to evaluate the utility of the acquired batch. Our work is different from these since we do not assume knowledge of slices that should be acquired, and we estimate the direct impact of acquisition on model fairness instead of model retraining or using data coverage as a proxy for fairness.

Data valuation. Data valuation has emerged as a powerful tool to explain the workings of black-box machine learning models [11, 15]. Data Shapley [11] offers a principled metric to quantify the value of each training datum toward model performance. While model agnostic, computing Data Shapley values is computationally expensive. Influence functions [8, 15] provide first-order approximations of the influence of training data points on model performance and have been used to identify data points responsible for model bias [17]. Influence functions incur an expensive one-time offline computation following the online computation of impact estimations is fast. To the best of our knowledge, this study is the first work to explore influence functions for the problem of data acquisition.

6 CONCLUSION

This paper presented Inf-acq, an approach based on data valuation to determine which additional points should be acquired such that a machine learning model learned on the updated data generates fairer decisions. To the best of our knowledge, Inf-acq is the first to leverage data valuation for machine learning models to the problem of data acquisition for fair machine learning. We assess data points by their impact on the model fairness and to expedite the computation of impact, we adopt influence functions that estimates the effect of adding a data point on model fairness without retraining the model. Our experimental evaluation on real-world datasets underscores the effectiveness of valuation-based data acquisition in reducing model bias without necessitating extensive model retraining. By prioritizing the acquisition of data points based on their valuation, Inf-acq exhibits superior performance compared to coverage-based approaches.

While incorporating data valuation provides a promising direction for improving data quality through data acquisition, influence functions are applicable to a niche class of machine learning models. Future work includes the development of methods that could be applied to black-box models. We also intend to explore data valuation for acquiring data subsets since sometimes acquiring related data might be easier than acquiring unrelated data points. Future work also includes incorporating data valuation for traditional data acquisition that assumes access to a data pool.

REFERENCES

[1] 2019. Housing Department Slaps Facebook With Discrimination Charge. lhttps://www.npr.org/2019/03/28/707614254/hud-slaps-facebook-with-housing-discrimination-charge.

[2] 2019. Self-driving cars more likely to hit blacks. lhttps://www.technologyreview.com/2019/03/01/136808/self-driving-cars-are-coming-but-accidents-may-not-be-evenly-distributed/.

[3] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 554–565.

[4] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 981–993.

[5] Sumon Biswas, Mohammad Wardat, and Hridesh Rajan. 2022. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In *Proceedings of the 44th International Conference on Software Engineering*. 2091–2103.

[6] Chengliang Chai, Jiabin Liu, Nan Tang, Guoliang Li, and Yuyu Luo. 2022. Selective data acquisition in the wild for model charging. *Proceedings of the VLDB Endowment* 15, 7 (2022), 1466–1478.

[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR* abs/1703.00056 (2017).

[8] R. Dennis Cook and Sanford Weisberg. 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics* 22, 4 (1980), 495–508.

[9] Jeffrey Dastin. 2018. RPT-INSIGHT-Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).

[10] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository.

[11] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 2242–2251.

[12] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[13] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.* ProPublica, May 23, 2016.

[14] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.

[15] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 1885–1894.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[17] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2022. Interpretable Data-Based Explanations for Fairness Debugging. In *Proceedings of the 2022 International Conference on Management of Data*. 247–261.

[18] Theodoros Rekatsinas, Amol Deshpande, Xin Luna Dong, Lise Getoor, and Divesh Srivastava. 2016. Sourcesight: Enabling effective source selection. In *Proceedings of the 2016 International Conference on Management of Data*. 2157–2160.

[19] Ki Hyun Tae and Steven Euijong Whang. 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*. 1771–1783.

[20] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[21] Gerhard Weikum. 2013. Data discovery. *Data Science Journal* 12 (2013).

[22] Weiyuan Wu, Lampros Flokas, Eugene Wu, and Jiannan Wang. 2020. Complaint-driven training data debugging for query 2.0. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1317–1334.