Surviving in Diverse Biases: Unbiased Dataset Acquisition in Online Data Market for Fair Model Training

Jiashi Gao[1], Ziwei Wang[1,2], Xiangyu Zhao[3], Xin Yao[4], Xuetao Wei[1]*

[1]Southern University of Science and Technology (SUSTech), Shenzhen, China
[2] University of Birmingham (UoB), UK
[3] City University of Hong Kong (CityU), Hong Kong SAR, China
[4] Lingnan University (LU), Hong Kong SAR, China
12131101@mail.sustech.edu.cn, 12250053@mail.sustech.edu.cn, xy.zhao@cityu.edu.hk, xinyao@ln.edu.hk, weixt@sustech.edu.cn

Abstract

The online data markets have emerged as a valuable source of diverse datasets for training machine learning (ML) models. However, datasets from different data providers may exhibit varying levels of bias with respect to certain sensitive attributes in the population (such as race, sex, age, and marital status). Recent dataset acquisition research has focused on maximizing accuracy improvements for downstream model training, ignoring the negative impact of biases in the acquired datasets, which can lead to an unfair model. *Can a consumer obtain an unbiased dataset from datasets with diverse biases?* In this work, we propose a fairness-aware data acquisition framework (FAIRDA) to acquire high-quality datasets that maximize both accuracy and fairness for consumer local classifier training while remaining within a limited budget. Given the biases of data commodities remain opaque to consumers, the data acquisition in FAIRDA employs explore-exploit strategies. Based on whether exploration and exploitation are conducted sequentially or alternately, we introduce two algorithms: the knowledge-based offline data acquisition (KDA) and the reward-based online data acquisition algorithms (RDA). Each algorithm is tailored to specific customer needs, giving the former an advantage in computational efficiency and the latter an advantage in robustness. We conduct experiments to demonstrate the effectiveness of the proposed data acquisition framework in steering users toward fairer model training compared to existing baselines under varying market settings.

Introduction

The online data markets enable data transactions and allow data providers to profit from publishing data resources for sale (Koesten 2018; Azcoitia and Laoutaris 2022; Fruhwirth, Rachinger, and Prlja 2020; Chen, Li, and Xu 2022), rather than hoarding it in a monopolistic manner (Maher and Khan 2022). This benefits data consumers, such as machine learning (ML) researchers, who can access a wider range of heterogeneous datasets. However, datasets from different data providers may exhibit varying levels of bias (Tommasi et al. 2017; Mehrabi et al. 2021; Roselli, Matthews, and Talagala 2019) with respect to certain sensitive features in the population (e.g., race, sex, age, marital status). For example, datasets collected from hospitals in predominantly Caucasian neighborhoods may not adequately represent symptoms and outcomes relevant to African-American populations. The biases in the datasets are caused primarily by (1) different collection methods: data providers may use different devices, environments, standards, or techniques to collect and annotate data, resulting in different data distributions and biases; (2) different data characteristics: data providers may represent different individuals, enterprises, scenarios, or domains, leading to data that reflect different preferences, needs, behaviors, or characteristics; (3) different processing methods: data providers may apply preprocessing steps on the datasets, such as cleaning, transformation, padding, encoding, to improve the quality and usability of the data, which may affect the bias of the dataset. Since the ML models extract insights from training data, the biases in the datasets may potentially lead to model fairness concerns when the training data are collected from online markets. Ideally, an ethically unbiased classifier should minimize statistical disparities across different groups, each with unique sensitive attributes—a crucial consideration in social applications.

One direct strategy for consumers (e.g., model owners) to mitigate the bias is de-biasing (Geirhos et al. 2018) by dropping the sensitive attribute in the acquired datasets. However, this method may only be effective if other dataset features (proxy variables) are uncorrelated with the sensitive attribute. Other sampling-based bias migration strategies include reweighing (Calders, Kamiran, and Pechenizkiy 2009), the removal of data points (downsampling), or the addition of new data points (upsampling) (Salimi et al. 2019; Amend and Spurlock 2021). While the existing sampling-based strategies can address imbalanced class distributions, they may not effectively address other types of bias, such as feature bias (He et al. 2021) or measurement bias (Fahse, Huber, and van Giffen 2021). Additionally, there is a risk of over-fitting, as these methods may remove a significant portion of the data. Overall, previous work on model-based bias mitigation methods only tried to make tradeoffs between demographic parity (Feldman et al. 2015) and joint utility, which may deteriorate the performance of ML classifiers.

However, acquiring unbiased datasets to ensure the fairness and ethics of the trained ML model in online data markets is challenging due to the **lack of transparency** in the

*Corresponding Author.

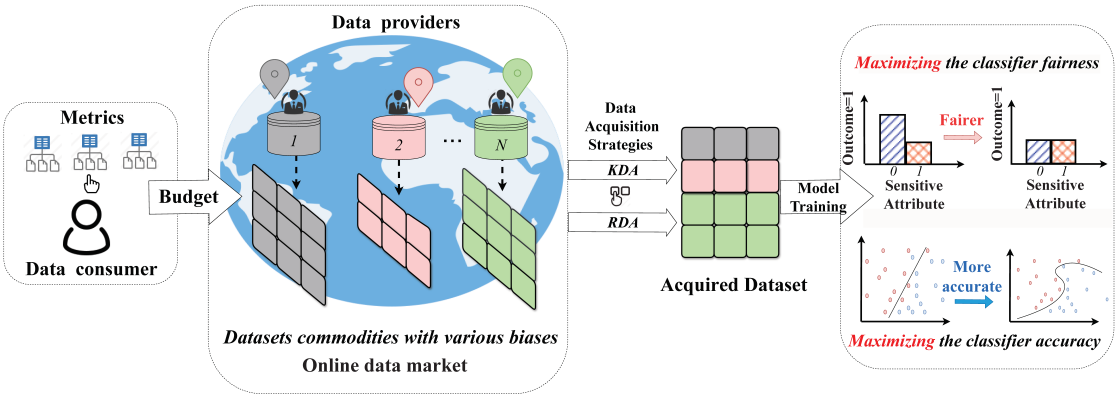

Figure 1: An overview of our proposed fairness-aware data acquisition framework, FAIRDA. This framework is designed to guide budget-constrained consumers within the online data market in selecting datasets from available dataset commodities so that the required dataset maximizes both the accuracy and fairness improvement for the consumer's model training.

shopping mode (Van Miegroet, Alexander, and Leunissen 2019; Newman et al. 2021). Consumers cannot inspect the data before buying it, so they cannot assess how much bias the dataset may introduce to ML model training. Buying all the available datasets to evaluate their bias is impractical, especially when the data consumers usually have a limited budget. The divisible property of data products (Zheng et al. 2017; Chen, Li, and Xu 2022; Fernandez, Subramaniam, and Franklin 2020) and the online shopping mode allows data consumers to query partial data from multiple data providers in different interactions (Hayashi and Ohsawa 2020; Bajoudah, Dong, and Missier 2019; Li et al. 2022), resulting in a diverse combination of acquired subsets for their specific task. In this context, we propose a fairness-aware data acquisition framework, FAIRDA (as shown in Figure 1), to acquire high-quality datasets that bring maximum accuracy and fairness improvement for the consumer's classifier within limited budget costs.

We propose both offline and online unbiased data acquisition algorithms: a knowledge-based data acquisition (KDA) and a reward-based data acquisition (RDA). Both algorithms are based on the principle of exploration-exploitation. The distinctions between the two algorithms are: (1) In KDA, we initially employ a knowledge-mining stage to explore the potential utilities of datasets towards enhancing model accuracy and fairness. Upon determining the utilities, we formulate the fair data acquisition as a nonlinear knapsack problem (NLKP) with model accuracy as the optimizing objective and fairness requirement as the constraint. This approach involves executing exploration and exploitation phases in a sequential manner; (2) In RDA, we model unbiased data acquisition as a Bernoulli bandit problem. For each dataset, we establish a reward function to evaluate its utility for improving the model's accuracy and fairness. In each iteration, a small subset of data from the dataset with the highest reward is acquired, and the reward functions are iteratively updated based on the utility of the acquired data. When the reward functions have not yet converged, exploration and exploitation occur alternately within the iterations. As the reward functions approach convergence, the data acquisition process naturally shifts from exploration to a more exploitation-oriented approach. These algorithms are tailored for different scenarios; KDA offers greater efficiency in terms of runtime, whereas RDA showcases enhanced robustness to fluctuations or disturbances within market environments.

In summary, this paper makes the following contributions.

- We formally investigate an unexplored issue: unbiased data acquisition for ML model training in the online data market. The study's findings have practical implications for real-world data markets, enabling strategic consumers to obtain ethical ML models from data markets through innovative solutions.
- We propose a fairness-aware data acquisition framework (FAIRDA) to guide budget-limited data consumers to acquire unbiased data for fairer ML model training. The framework includes two practical algorithms: KDA, which offers a highly efficient run-time solution, and RDA, which is robust to market setting changes.
- We conduct extensive experiments to validate our proposed framework's effectiveness in various scenarios, such as different budget settings and dataset heterogeneity levels. Compared to the optimal baseline, our framework can achieve up to 77%, 25% and 76% fairness improvement for the ML model under limited budgets, $1K$, $5K$, and $10K$, respectively.

Related Work

Online Data Market

Several commercialized data transaction platforms facilitate online data acquisition, notable examples being Dawex [1], Bloomberg [2], LexisNexis [3], and Acxiom [4]. These platforms

[1]https://www.dawex.com/en. Accessed on July 29, 2024.
[2]https://www.bloomberg.com/explore/data-that-is-made-for-more-performance. Accessed on July 29, 2024.
[3]https://www.lexisnexis.com. Accessed on July 29, 2024.
[4]https://www.acxiom.com. Accessed on July 29, 2024.

require intermediary intervention to perform transactions, employing bilateral and negotiated contractual relationships, distinguishing them from conventional open markets. The growing interest in rich and high-value data has prompted numerous endeavours (Wang and Tsang 2023; Figueredo, Seed, and Wang 2022; Firouzi et al. 2022; Dixit et al. 2023; Li et al. 2023; Byeon et al. 2023; Fanchang et al. 2023; Pandey, Pinson, and Popovski 2023) aimed at designing open markets for data trading. Fernandez et al. (2020) have proposed highly practical data market platforms, emphasizing the necessity to design interaction protocols that allow buyers and arbitrators to communicate what data is needed. A distinguishing characteristic of data as an asset is that it can be divided and combined arbitrarily (Fernandez, Subramaniam, and Franklin 2020; An et al. 2021), and in practice, sometimes multi-source datasets combinations may better satisfy buyers' needs. Our work is consistent with this vision and addresses the ethical issue in this context.

Data Acquisition

As machine learning research has long focused on models rather than datasets, Mazumder et al. (2023) have elucidated that this oversight of data quality severely limits the scope of ML models' effectiveness, often resulting in inaccuracies, biases, and a lack of robustness upon real-world application. This refocuses our attention on data acquisition, which is fundamentally important in developing data-centric algorithms. Gong et al. (2023) proposed an online data selection framework for federated learning model training with limited on-device storage. Chai et al. (2022) proposed Auto-Data, which selects points from candidate datasets to obtain higher model accuracy improvement. However, the above work did not consider the budget constraint and only focused on data acquisition in the wild. In data market contexts, Li et al. (2021) were the first to define and study the problem of selecting subsets from a dataset to improve the accuracy of classifier models. Tae et al. (2021) proposed a slice tuner framework to optimize data acquisition strategies across multiple data slices within a dataset, ensuring equitable model performance across different demographic groups or data segments. However, the effectiveness of slice selection depends on accurately estimating the learning curves for these slices, a task that can be challenging without sufficient initial data or clearly defined characteristics for each slice, particularly in a market context. Li et al. (2024) proposed *Bulk Acquisition* and *Sequential Acquisition* strategies to improve machine learning model confidence by optimizing data collection. Unlike the existing works, we focus on strategically purchasing datasets that may contain unknown heterogeneous biases, ensuring that the combination of these datasets is unbiased and suitable for training ethically fair machine learning models.

Preliminary

Data and Model Notions

Our focus is on consumers who use acquired data to train classifiers. A data provider in the data market supplies data from a specific domain $\Gamma\{\mathcal{X}, \mathcal{Z}, \mathcal{Y}\}$, where $\mathcal{X}$ denotes the general feature space, $\mathcal{Y}$ the label space, and $\mathcal{Z}$ the sensitive feature space. Assume a conditional distribution $P(y|x, z)$ over Γ, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$. Let $\mathcal{D}_{init}$ represent the data initially possessed by the consumer and the initial classifier trained on this data as $\mathcal{M}_{init}$. Let $\mathcal{D} \sim \Gamma$ denote the newly acquired dataset from the data market. The model trained on the combination of $\mathcal{D}_{init}$ and $\mathcal{D}$ is represented as $\mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}$. The objective in training this classifier is to align the distribution of outcomes $P\left(\hat{y} = \mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}(x, z)|x, z\right)$ closely with $P(y|x, z)$, where $\hat{y} = \mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}(x, z)$ denotes the predicts of the model when given the data point (x, z).

Market Participants

We establish follow-up data acquisition solutions for data markets that involve interactions between a data consumer and multiple data providers. These solutions can serve as a building block for multiple data consumers.

Data Consumer The data consumer is defined as the owner of the classifier model $\mathcal{M}_{\mathcal{D}_{init}}$, which may have been trained on the consumer initially owned data $\mathcal{D}_{init}$. The consumer desires to obtain richer datasets from the data market for model training but is limited by budget constraints $\mathbf{B}$.

Data Provider The data providers, who can be individuals or companies, offer a collection of datasets in the data market and earn profit when their datasets are subscribed to or bought by consumers.

A desirable data market should enable interactions that convey consumer task specifications and metadata containing dataset descriptions provided by data providers (Fernandez, Subramaniam, and Franklin 2020; Brickley, Burgess, and Noy 2019). These interactions enable consumers to discover a pool of candidate datasets, denoted as $\mathbf{D}_{pool} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$. The description of each dataset in the candidate pool is closely related to the downstream task that the consumer aims to achieve. Such interactions are already permitted on commercial data market platforms such as Amazon Web Services (AWS) [5], where customers can search for datasets relevant to their needs. When a candidate pool has at least two datasets that are not i.i.d., we refer to it as a heterogeneous scenario, which is denoted as $\exists i, j \in \{1, 2, \ldots, N\}, i \neq j$ and $\mathcal{D}_i \not\sim \mathcal{D}_j$.

Accuracy and Fairness Metrics

For a dataset $\mathcal{D}$ in the online data market, we focus on its two aspects of data utility: the accuracy improvement and the fairness improvement that $\mathcal{D}$ can contribute to the classifier training. The *accuracy* after obtaining the dataset $\mathcal{D}$ can be calculated as follows:

$$Acc = \frac{\sum_{(x,z,y)\in\mathcal{D}_{\text{val}}} \mathbb{I}[\hat{y} = y]}{|\mathcal{D}_{\text{val}}|}, \hat{y} = \mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}(x, z), \quad (1)$$

where $\mathbb{I}$ is an indicator function that outputs 1 only when $\mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}$ predicts correctly.

To measure the improvement in fairness that dataset $\mathcal{D}$ can bring to the classifier, We employ two representative fairness measures:

[5] https://aws.amazon.com/marketplace. Accessed on July 29, 2024.

Definition 1 *(Equalized Odds (EO) (Hardt, Price, and Srebro 2016)) The fairness metric requires that the classifier's true positive rate (or false positive rate) is equal across different sensitive groups.*

$$EO = \left|P\left(\hat{y}=1 \mid \mathrm{z}=1,\mathrm{y}=y\right)-P\left(\hat{y}=1 \mid \mathrm{z}=0,\mathrm{y}=y\right)\right|, \quad \hat{y}=\mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}(x,z), y=\{0,1\},\{x,y,z\}\in\mathcal{D}_{val}. \tag{2}$$

Definition 2 *(Demographic Parity (DP) (Feldman et al. 2015)) The fairness metric requires the classifier's positive prediction rate to be equal across different sensitive groups.*

$$DP = \left|P\left(\hat{y}=1 \mid \mathrm{z}=1\right)-P\left(\hat{y}=1 \mid \mathrm{z}=0\right)\right|, \quad \hat{y}=\mathcal{M}_{\mathcal{D}_{init}\cup\mathcal{D}}(x,z),\{x,z\}\in\mathcal{D}_{val}. \tag{3}$$

The classifier is strictly fair when the fairness measure is 0.

Proxy Accuracy and Fairness Metrics

The utility functions in Equation (1)∼(3) require an available validation dataset $\mathcal{D}_{\text{val}}$. However, data consumers frequently lack access to validation datasets that accurately represent real-world application scenarios (Yoon, Arik, and Pfister 2020). For example, suppose a financial institution trains a model to identify loan default risk. If the available validation dataset consists primarily of historical or regionally limited data, it may fail to capture economic shifts or demographic changes in the future, resulting in imprecise model performance evaluation in actual operational environments. To address this issue, we propose proxy metrics that are independent of a validation process.

Definition 3 *(Proxy accuracy metric U_{Acc} (Li, Yu, and Koudas 2021)) The proxy accuracy metric $U_{Acc}(\mathcal{D})$ is defined as the error rate of the model $\mathcal{M}_{\mathcal{D}_{init}}$ tested on the newly acquired dataset $\mathcal{D}$,*

$$U_{Acc}(\mathcal{D}) = \frac{\sum_{(x,z,y)\in\mathcal{D}}\mathbb{I}[\hat{y}\neq y]}{|\mathcal{D}|}, \hat{y}=\mathcal{M}_{\mathcal{D}_{init}}(x,z), \tag{4}$$

where $|\mathcal{D}|$ denotes the size of the $\mathcal{D}$.

The proxy accuracy metric assigns a higher value when $\mathcal{M}_{\mathcal{D}_{init}}$ has a higher incorrect classification rate on the newly acquired dataset $\mathcal{D}$. The design intuition is that introducing data with different distributions from the existing training set can result in more significant performance improvements for the model than introducing i.i.d. data with respect to the existing training set, because diverse datasets help in better capturing the underlying variability in the data, thereby improving the model's generalization capabilities. Since $\mathcal{M}_{\mathcal{D}_{init}}$ is only trained on $\mathcal{D}_{init}$, a higher incorrect classification rate means that $\mathcal{D}$ is drawn from a distribution that is more different than $\mathcal{D}_{init}$ that consumers already have and is more encouraged to acquire.

Definition 4 *(Proxy fairness metric $U_{EO}(\mathcal{D})$ and $U_{DP}(\mathcal{D})$) The proxy fairness metrics $U_{EO}(\mathcal{D})$ and $U_{DP}(\mathcal{D})$ are defined as the EO and DP of the model $\mathcal{M}_{\mathcal{D}_{init}}$ tested on the newly acquired dataset $\mathcal{D}$, respectively.*

$$\begin{aligned} &U_{EO}(\mathcal{D}) \\ &= \left|P\left(\hat{y}=1 \mid \mathrm{z}=1,\mathrm{y}=y\right)-P\left(\hat{y}=1 \mid \mathrm{z}=0,\mathrm{y}=y\right)\right| \\ &= \left|\frac{\sum_{(x,z,y)\in\mathcal{D}_{\mathrm{z}=1,\mathrm{y}=y}}\mathbb{I}[\hat{y}=1]}{|\mathcal{D}_{\mathrm{z}=1,\mathrm{y}=y}|}-\frac{\sum_{(x,z,y)\in\mathcal{D}_{\mathrm{z}=0,\mathrm{y}=y}}\mathbb{I}[\hat{y}=1]}{|\mathcal{D}_{\mathrm{z}=0,\mathrm{y}=y}|}\right|, \\ &\hat{y}=\mathcal{M}_{\mathcal{D}_{init}}(x,z), y=\{0,1\},\{x,y,z\}\in\mathcal{D}. \end{aligned} \tag{5}$$

$$\begin{aligned} &U_{DP}(\mathcal{D}) = \\ &\left|P\left(\hat{y}=1 \mid \mathrm{z}=1\right)-P\left(\hat{y}=1 \mid \mathrm{z}=0\right)\right| \\ &= \left|\frac{\sum_{(x,z,y)\in\mathcal{D}_{\mathrm{z}=1}}\mathbb{I}[\hat{y}=1]}{|\mathcal{D}_{\mathrm{z}=1}|}-\frac{\sum_{(x,z,y)\in\mathcal{D}_{\mathrm{z}=0}}\mathbb{I}[\hat{y}=1]}{|\mathcal{D}_{\mathrm{z}=0}|}\right|, \\ &\hat{y}=\mathcal{M}_{\mathcal{D}_{init}}(x,z),\{x,z\}\in\mathcal{D}. \end{aligned} \tag{6}$$

A lower U_{EO} or U_{DP} indicates that the decision bias of $\mathcal{M}_{\mathcal{D}_{init}}$ towards different sensitive groups learned from $\mathcal{D}_{init}$ is less consistent with that of $\mathcal{D}$, which suggests that $\mathcal{D}$ can help mitigate the bias present in $\mathcal{D}_{init}$ and is encouraged to acquire. For instance, when a biased model $\mathcal{M}_{\mathcal{D}_{init}}$ predicts a lower probability of *income exceeding* $50K$ for the sensitive group where *race=Black* and exhibits reduced income bias for the sensitive group where *race=Black* on $\mathcal{D}$, it implies that other general attributes distributions of the *race=Black* sensitive group in $\mathcal{D}$ differ from that in $\mathcal{D}_{init}$, i.e., in $\mathcal{D}$, the proportion of the *race=Black* group with the property attribute *has house* being *Yes* is higher compared to $\mathcal{D}_{init}$. Thus, the addition of data from $\mathcal{D}$ can help increase the probability of the group *race=Black* being classified as having an *income exceeding* $50K$ and reduce the bias present in the original dataset.

We take EO as a fairness metric to implement the following discussion of unbiased data acquisition, and evaluate both EO and DP separately in the experiments.

Problem Definition

Commonly, the online data markets respond to data query by randomly sampled subset, $RandomSampling(\mathcal{D}_i, S_i)$, where S_i denotes sampling size on dataset $\mathcal{D}_i$.

Unbiased Data Acquisition Problem Taking U_{EO} as an example, we formally define the budget allocation problem for the fair classifier within the data marketplace. The goal is to optimize the sampling size $\vec{S} = \{S_i, i=1,...,N\}$, so that the subset combination $\cup_{i=1}^{N}\bar{\mathcal{D}}_i = \cup_{i=1}^{N} RandomSampling(\mathcal{D}_i, S_i)$, achieving maximum proxy accuracy while minimizing the proxy fairness.

$$\begin{aligned} \vec{S}^* = &\underset{S_i\in[0,|\mathcal{D}_i|]}{\arg\max}\ U_{Acc}\left(\cup_{i=1}^{N}\bar{\mathcal{D}}_i\right), \\ &\text{s.t. } U_{EO}\left(\cup_{i=1}^{N}\bar{\mathcal{D}}_i\right)\leq\varepsilon, \sum_{i=1}^{N} S_i\times \mathrm{p}_i \leq \mathbf{B}, \\ &\varepsilon\to 0, i=1,...,N, \end{aligned} \tag{7}$$

where $\cup_{i=1}^{N}\bar{\mathcal{D}}_i$ represents the entire collection of data that consumers obtain from the online data market, $\{\mathrm{p}_i, i=1,...,N\}$ denotes the unit price vector of datasets commodities, ε denotes the acceptable upper bound for bias, and $\mathbf{B}$ denotes the total amount of consumer's budget.

Fairness-aware Data Acquisition Framework

Framework Overview

Our framework FAIRDA, shown in Figure 1, helps budget-constrained consumers in the online data market choose datasets that improve accuracy and fairness for model training. Initially, the consumer identifies datasets pertinent to the task at hand within the data marketplace, utilizing the provided metadata descriptions. Subsequently, the dataset identifiers/indexes/IDs are aggregated to constitute a *candidate datasets pool*. To obtain an optimal data combination from the candidate datasets pool, consumers deploy a data acquisition algorithm that strategically allocates their budget across various dataset commodities. Within FAIRDA, we introduce two fairness-aware data acquisition strategies: KDA and RDA, which are detailed in the following. Finally, the consumer uses the acquired data to train their models.

Knowledge-based Data Acquisition

Datasets in online data markets are often opaque to data consumers before purchase. A direct way to make wise dataset selections is to consume some budget to investigate each candidate dataset's utility. In this section, we present a two-stage algorithm called knowledge-based data acquisition (KDA). The first stage of the algorithm is called *knowledge-mining*. During this stage, we propose a multi-step iterative sampling strategy that allows consumers to explore the knowledge about the potential accuracy and fairness improvement the data can bring to the model while minimizing budgetary costs. Based on the acquired knowledge, we reformulate the unbiased data acquisition problem in Equation (7) as a variant of the nonlinear knapsack problem (NLKP) and give an explainable solution for the remaining budget allocation across datasets.

Knowledge-mining Stage Sampling estimation is a direct method for consumers to gain insights into the utility of a dataset. However, a low sampling size can lead to inaccurate estimation of the actual utility of the data, while a high sampling rate can result in excessive budget costs. To balance the trade-off, we propose a minimum sampling size $\bar{S}_{min}$ as in Proposition 1, which ensures that the estimation error is acceptable while minimizing budget costs during this stage. We first clarify the notions about the estimation error.

Definition 5 *(Estimation Error Under Metric $U(\cdot)$) Let's consider a subset $\bar{\mathcal{D}}$ that is sampled from the dataset $\mathcal{D}$. The estimation error of evaluation metric $U(\mathcal{D})$ that results from this sampling can be defined as:*

$$e = \left| U(\mathcal{D}) - U(\bar{\mathcal{D}}) \right|. \tag{8}$$

Definition 6 *($\epsilon - \delta$ Approximation) $U(\bar{\mathcal{D}})$ is said to be an $\epsilon - \delta$ approximation of $U(\mathcal{D})$ if $Pr(\left|U(\mathcal{D}) - U(\bar{\mathcal{D}})\right| \geq \epsilon) \leq \delta$, where ϵ denotes the acceptable estimation error tolerance, δ denotes the significance level, and $U(\cdot)$ denotes any evaluation metric.*

Inspired by the work of Li et al. (2021), which introduces the minimum sampling size necessary for estimating the improvement in accuracy that a dataset can provide, we propose a tightly minimum sampling size required to comprehensively estimate both the accuracy and fairness improvements that a dataset can offer.

Proposition 1 *(Minimum Sampling Size $\bar{S}_{min}$) To achieve $\epsilon_{acc} - \delta_{acc}$ and $\epsilon_{eo} - \delta_{eo}$ approximations to the true values of $U_{Acc}(\mathcal{D})$ and $U_{EO}(\mathcal{D})$ using samples drawn from $\mathcal{D}$, the minimum required sampling size is given by:*

$$\begin{aligned} \bar{S}_{min} &= \max\left(\bar{S}^{EO}_{min}, \bar{S}^{Acc}_{min}\right) \\ &= \max\left\{ \begin{matrix} \max_{i\in[Z], j\in[Y]} \left(F^{EO}_{i,j} \cdot V^{EO}_{i,j} / \epsilon_{eo}\right)^2, \\ \left(F^{Acc} \cdot V^{Acc} / \epsilon_{acc}\right)^2 \end{matrix} \right\} \end{aligned} \tag{9}$$

where,

$$\begin{aligned} F^{EO}_{i,j} &= -\mathrm{PCT}_{|\bar{\mathcal{D}}|-1}\left(\tfrac{\delta_{eo}}{4}\right), \\ V^{EO}_{i,j} &= \sqrt{p_{i,j}(\bar{\mathcal{D}})\left(1 - p_{i,j}(\bar{\mathcal{D}})\right)}, \\ F^{Acc} &= -\mathrm{PCT}_{|\bar{\mathcal{D}}|-1}\left(\tfrac{\delta_{acc}}{2}\right), \\ V^{Acc} &= \sqrt{U_{Acc}(\bar{\mathcal{D}})\left(1 - U_{Acc}(\bar{\mathcal{D}})\right)}, \\ \bar{\mathcal{D}} &= RandomSampling(\mathcal{D}, \bar{S}_{min}). \end{aligned} \tag{10}$$

Here, $\mathrm{PCT}_{|\bar{\mathcal{D}}|-1}$ denotes the percentile function of the t-distribution with $|\bar{\mathcal{D}}| - 1$ degrees of freedom, and $p_{i,j}(\mathcal{D})$ represents $P(\hat{y} = 1 \mid z = i, y = j)$ for all $i \in [Z], j \in [Y]$ in the calculation of U_{EO}.

However, the minimum sampling size $\bar{S}_{min}$ in Proposition 1 cannot be solved analytically. It is important to note that the sample's standard deviations ($V^{Acc}, V^{EO}_{i,j}$) are stable and asymptotically converge to the population standard deviation (Li et al. (2021)). Additionally, as pointed out by Fisher et al. (1992), the values of ($F^{Acc}, F^{EO}_{i,j}$) also asymptotically converge to the opposite values of the $\frac{\delta_{acc}}{2}$-percentile and $\frac{\delta_{eo}}{4}$-percentile of standard normal distribution, respectively. These convergence properties facilitate the use of a multi-step iterative approach to estimate the approximate value of $\bar{S}_{min}$: As illustrated in Algorithm 1, the multi-step iterative approach involves two parallel calculation flows, $\bar{S}^{Acc}_{min}$ and $\bar{S}^{EO}_{min}$. In each iteration, if both $\bar{S}^{Acc,t}_{min}$ and $\bar{S}^{EO,t}_{min}$ have converged, the $\bar{S}_{min}$ can be determined by taking the maximum of the current $\bar{S}^{Acc,t}_{min}$ and $\bar{S}^{EO,t}_{min}$. If they have not converged, we continue to acquire additional data records of size $\triangle s$ and append them to the current dataset $\bar{\mathcal{D}}^{t+1}$. It is important to note that the calculation of $\bar{S}_{min}$ may converge and stop before the acquired data reaches the required minimum sampling size. In such cases, the data sampling with size $\triangle s$ is continued until the size requirement is met. The final obtained subset $\bar{\mathcal{D}}$ can then be used to achieve $\epsilon_{acc} - \delta_{acc}$ and $\epsilon_{eo} - \delta_{eo}$ approximations to the true values of $U_{Acc}(\mathcal{D})$ and $U_{EO}(\mathcal{D})$. The above process is applied to each dataset $\mathcal{D}_i$ in the candidate dataset pool, and the total budget cost for this process is $\sum_{i=1}^{N} \bar{S}_{min,i} \times \mathrm{p}_i$.

Knowledge-utilizing Stage Given that the data consumer has an initial budget of $\mathbf{B}$, the remaining budget after the knowledge-mining stage is $\mathbf{B} - \sum_{i=1}^{N} \bar{S}_{min,i} \times \mathrm{p}_i$. In the following knowledge-utilizing stage, our goal is to optimize

Coefficient	Value	Coefficient	Value
a_0^i	U_{Acc}^i	c_0^i	$\sum_{i=1}^N \bar{S}_{min,i} \times U_{Acc}^i$
a_1^i	$p_{1,1}^i \times \frac{\left\|\mathcal{D}_{i,(z=1,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_1^i	$\sum_{i=1}^N p_{1,1}^i \times \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=1,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_2^i	$\frac{\left\|\mathcal{D}_{i,(z=1,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_2^i	$\sum_{i=1}^N \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=1,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_3^i	$p_{0,1}^i \times \frac{\left\|\mathcal{D}_{i,(z=0,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_3^i	$\sum_{i=1}^N p_{0,1}^i \times \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=0,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_4^i	$\frac{\left\|\mathcal{D}_{i,(z=0,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_4^i	$\sum_{i=1}^N \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=0,y=1)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_5^i	$p_{1,0}^i \times \frac{\left\|\mathcal{D}_{i,(z=1,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_5^i	$\sum_{i=1}^N p_{1,0}^i \times \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=1,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_6^i	$\frac{\left\|\mathcal{D}_{i,(z=1,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_6^i	$\sum_{i=1}^N \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=1,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_7^i	$p_{0,0}^i \times \frac{\left\|\mathcal{D}_{i,(z=0,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_7^i	$\sum_{i=1}^N p_{0,0}^i \times \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=0,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_8^i	$\frac{\left\|\mathcal{D}_{i,(z=0,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$	c_8^i	$\sum_{i=1}^N \bar{S}_{min,i} \times \frac{\left\|\mathcal{D}_{i,(z=0,y=0)}\right\|}{\left\|\mathcal{D}_i\right\|}$
a_9^i	p_i	c_9^i	$\mathbf{B} - \sum_{i=1}^N \bar{S}_{min,i} \times \mathrm{p}_i$

Table 1: The detailed coefficients in Equation (11). U_{Acc}^i, $p_{0,0}^i$, $p_{0,1}^i$, $p_{1,0}^i$ and $p_{1,1}^i$ represent the estimates of $U_{Acc}(\mathcal{D})$, $P\left(\hat{y}=1 \mid \mathrm{z}=0, \mathrm{y}=0\right)$, $P\left(\hat{y}=1 \mid \mathrm{z}=0, \mathrm{y}=1\right)$, $P\left(\hat{y}=1 \mid \mathrm{z}=1, \mathrm{y}=0\right)$ and $P\left(\hat{y}=1 \mid \mathrm{z}=1, \mathrm{y}=1\right)$ for the $\mathcal{D}_i$, respectively, obtained from the minimum sample subset in knowledge-mining stage, respectively. The ratio between $\left|\mathcal{D}_{i,(\mathrm{z,y})}\right|$ and $|\mathcal{D}_i|$ can be easily determined if the data market provides an aggregation query. Otherwise, these sizes can be estimated using the group ratios within $\bar{\mathcal{D}}_{min,i}$. Random sampling techniques can be employed to maintain a statistically unbiased distribution.

data acquisition with the remaining budget across the N candidate datasets in the online data market so that all the samples obtained in both stages maximize the accuracy and fairness improvement of the classifier. We define the optimization variable in this stage as $\vec{S}' = \{S_i', i = 1, ..., N\}$, which represents the further sampling size for candidate datasets in the knowledge-utilizing stage. Based on the definitions of the utility metrics in Equation (4) $\sim$ Equation (6), and the unbiasedness of the sampled subset obtained by random sampling from the original data distribution, we modify the optimization objective in Equation (7) as follows:

$$\begin{aligned} \vec{S}'^* &= \arg\max \textstyle\sum_{i=1}^N a_0^i \times S_i' + c_0 \\ \text{s.t.} &\left| \frac{\sum_{i=1}^N a_1^i S_i' + c_1}{\sum_{i=1}^N a_2^i S_i' + c_2} - \frac{\sum_{i=1}^N a_3^i S_i' + c_3}{\sum_{i=1}^N a_4^i S_i' + c_4} \right| \leq \varepsilon, \\ &\left| \frac{\sum_{i=1}^N a_5^i S_i' + c_5}{\sum_{i=1}^N a_6^i S_i' + c_6} - \frac{\sum_{i=1}^N a_7^i S_i' + c_7}{\sum_{i=1}^N a_8^i S_i' + c_8} \right| \leq \varepsilon, \\ &\textstyle\sum_{i=1}^N a_9^i \times S_i' \leq c_9, \end{aligned} \tag{11}$$

The problem in Equation (11) is a variant of nonlinear knapsack problems (NLKP) (D'Ambrosio and Martello 2011), where the bias constraints are expressed as the sum of non-convex multivariate functions. Unlike the well-known (linear) knapsack problem (Dudziński and Walukiewicz 1987; Balas and Zemel 1980) with the linear objective and constraint functions, NLKP is NP-hard. Several algorithms are specifically tailored to the case of convex or concave functions, which is not assumed in our case (Hochbaum 1995; Granmo and Oommen 2011). Our study incorporates a meta-heuristic approach, the differential evolution (DE) method, to find a solution of the optimal sampling size $\vec{S}'^*$ within a continuous solution space. By leveraging the meta-heuristic solvers, we are able to efficiently identify the optimal $\vec{S}'^*$ during the knowledge-utilizing stage.

After completing the knowledge exploration and knowledge utilization stages in KDA, we can obtain the optimal data acquisition plan for consumers. This involves sampling the i-th dataset $\mathcal{D}_i$ in the candidate data pool $\mathbf{D}_{pool}$ at a sampling size of $S_i' + \bar{S}_{min,i}$, and allocating a corresponding budget of $(S_i' + \bar{S}_{min,i}) \times p_i$ for $\mathcal{D}_i$.

Reward-based Online Data Acquisition

While the deterministic solution in the KDA algorithm have strengths in terms of interpretability and runtime efficiency, it has been observed that in some scenarios, datasets with low utility still require sampling at a minimum sampling size $\bar{S}_{min}$, resulting in loss on budget. Moreover, the offline mechanism is not suitable in the case where interdependencies exist between the candidate datasets, as previously acquired data may impact and alter the potential performance improvements that datasets yet to be collected can contribute to model training. To tackle these issues, we propose reward-based data acquisition RDA in this section, an alternative solution that uses Bayesian probabilistic methods to optimize data acquisition. RDA approaches the data acquisition process as a Bernoulli bandit problem and implements many rounds of interactions, acquiring subsets with a small sampling size in each round. It dynamically updates rewards from different datasets to decide which dataset to sample next, enabling parallel knowledge mining and utilization.

Action space A consumer samples $\triangle S$ data in each iteration. There are N possible actions $\mathcal{A} = \{a_i\}, i = 1, ..., N$, each a_i representing sampling from i-th dataset in the candidate pool.

Algorithm 1: Multi-Step Iterative Calculation of minimum Sampling Size $\bar{S}_{min}$

Input: Dataset $\mathcal{D}$, acceptable estimation error tolerance: ϵ_{acc} and ϵ_{eo}, significance level δ_{acc} and δ_{eo};

Parameter: Current iteration: $t \leftarrow 0$; sampling size in each step: $\triangle s$;

Output: minimum sampling size $\bar{S}_{min}$, obtained subset: $\bar{\mathcal{D}}$;

Acquire subset: $\bar{\mathcal{D}}^t \leftarrow RandomSampling(\mathcal{D}, \triangle s)$.
while $\bar{S}_{min}^{Acc,t} - \bar{S}_{min}^{Acc,t-1} > 0$ or $\bar{S}_{min}^{EO,t} - \bar{S}_{min}^{EO,t-1} > 0$ **do**
  if $\bar{S}_{min}^{Acc,t} - \bar{S}_{min}^{Acc,t-1} > 0$ **then**
    Calculate U_{Acc} based on Equation (4).
    Calculate $\bar{S}_{min}^{Acc,t}$ based on Equation (9).
    $\bar{S}_{min}^{Acc,t} \leftarrow \text{Int}\left(\bar{S}_{min}^{Acc,t}\right)$.
  end if
  if $\bar{S}_{min}^{EO,t} - \bar{S}_{min}^{EO,t-1} > 0$ **then**
    Calculate $p_{i,j}$ based on Equation (5).
    Calculate $\bar{S}_{min}^{EO,t}$ based on Equation (9).
    $\bar{S}_{min}^{EO,t} \leftarrow \text{Int}\left(\bar{S}_{min}^{EO,t}\right)$.
  end if
  Acquire and append new acquired subset:
  $\bar{\mathcal{D}}^{t+1} \leftarrow \bar{\mathcal{D}}^t \oplus RandomSampling(\mathcal{D}, \triangle s)$.
  $t \leftarrow t + 1$.
end while
$\bar{S}_{min} \leftarrow \max\left(\bar{S}_{min}^{Acc,t}, \bar{S}_{min}^{EO,t}\right)$.
return: minimum sampling size $\bar{S}_{min}$, obtained subset: $\bar{\mathcal{D}}$.

Fairness-aware Reward The action reward calculation aligns with the optimization objective in Equation (7). Table 2 defines the reward for a newly acquired data point (x, z, y). The reward is only given if a new data point can improve both accuracy and fairness. First, the predicted label must be different from the true label to ensure accuracy improvement. Second, to enhance fairness, the dependency between the sensitive attribute and the outcome of the new data point must be opposite to that of the existing model. For example, when $U_{EO}^{y=1}(\mathcal{D}_{init}) > 0$, indicating that $P(\hat{y} = 1 \mid y = 1, z = 1) > P(\hat{y}(x) = 1 \mid y = 1, z = 0)$, the classifier's positive label output is more dependent on the sensitive attribute "$z = 1$." In this case, adding data points with $\hat{y} = 0, y = 1, z = 1$ should be rewarded as it increases $P(\hat{y} = 0 \mid y = 1, z = 1)$ and helps to mitigate the classifier's bias. A data point will not receive a reward if it does not meet these criteria.

Reward Updated Mechanism The reward r for a data point is a random variable that follows a Bernoulli distribution $r \sim Be(\theta)$, taking a binary output of 1 with probability θ, and 0 with probability $(1 - \theta)$. The reward r_i of action a_i can be regarded as a repeated process of selecting a single data point from $\mathcal{D}_i$ for S_i times. This process follows a binomial distribution $r_i \sim P_r(r_i|a_i, \theta_i) = Bin(S_i, \theta_i)$.

y	z	$U_{EO}^{y=1}(\mathcal{D}_{init})$	$U_{EO}^{y=0}(\mathcal{D}_{init})$	$\hat{y}$	r
1	1	> 0	$-$	0	1
1	0	< 0	$-$	0	1
0	0	$-$	> 0	1	1
0	1	$-$	< 0	1	1
1	$-$	0	0	0	1
0	$-$	0	0	1	1
Else					0

Table 2: Reward r of of a newly acquired data point (x, z, y)

In each interaction, the consumer randomly draws a value r_i from $P_r(r|a_i, \theta_i)$ for each $i \in 1, ..., N$, and chooses action $a^* = \arg\, max_{i \in \{1,...,N\}} r_i$. After receiving the reward for action a_i, we update the distribution $P_r(r_i|a_i, \theta_i)$ by updating the values of θ_i based on the reward received. The distribution $P_r(r_i|a_i, \theta_i), i \in \{1, ..., N\}$ before and after the update are the prior and posterior distribution, respectively. Following the operation in Li et al.'s work (2021), we consider the priors to be beta-distributed with parameters $\alpha = (\alpha_1, ..., \alpha_N)$ and $\beta = (\beta_1, ..., \beta_N)$, which is the conjugate prior probability distribution for binomial distributions. As observations are gathered, the distribution is updated according to Bayes' rule. Beta distributions are particularly convenient to work with because of their conjugacy properties. Specifically, considering the sampled subset in each iteration t is $\bar{\mathcal{D}}_t$, each action's posterior distribution is also beta with parameters that can be updated according to a simple rule:

$$(\alpha_i, \beta_i) \leftarrow \begin{cases} (\alpha_i, \beta_i), & \text{if } a_t \neq i \\ (\alpha_i, \beta_i) + \left(n_r, \left|\bar{\mathcal{D}}_t\right| - n_r\right), & \text{else} \end{cases} \tag{12}$$

where n_r is defined as the number of data points in $\bar{\mathcal{D}}_t$ which are rewarded based on the criterion in Table 2.

$$n_r = \textstyle\sum_{(x_i, y_i, z_i) \in \bar{\mathcal{D}}_t} r(x_i, y_i, z_i). \tag{13}$$

Non-stationary Systems During the iterations, the initial dataset $\mathcal{D}_{init}$ changes dynamically due to the addition of new data. As a result, the reward distribution becomes non-stationary with time-varying parameters θ_i^t. A solution is to systematically reduce the effect of past observations, for example, by limiting the influence of historical observations beyond a certain number of time periods, τ, in the past. At each time t, the consumer generates a posterior distribution based on the prior and conditioned only on the most recent τ actions and observations. However, the limited number of observations fails to concentrate and forces the probability model to explore indefinitely. Instead, we modify the posterior computation in Equation (12) in a way that involves modeling the evolution of a belief distribution in a manner that discounts the relevance of past observations and tracks a time-varying parameter θ^t, as inspired by previous research on dealing with nonstationary reward distributions (Raj and Kalyani 2017). Let the algorithm update parameters to identify the belief distribution of θ^t conditioned on the history $\mathbb{H}_{t-1} = ((a_1, n_r(1)), \ldots, (a_{t-1}, n_r(t-1)))$ according to:

$$(\alpha_i, \beta_i) \leftarrow (\alpha_i, \beta_i) + (S_i, F_i). \tag{14}$$

Algorithm 2: Online Reward-based Data Acquisition (*RDA*)

Input: Unit price of candidate datasets: $\mathbf{p}$; consumer-owned initial dataset: $\mathcal{D}_{init}$; remained budget: $b = \mathbf{B}$;
Output: $\mathcal{D}_{init}$.

for each action a in action space $\mathcal{A} = \{1, ..., N\}$ **do**
  Set α_a and β_a initially to 1.
end for
while $b > 0$ **do**
  for each action a in action space $\mathcal{A} = \{1, ..., N\}$ **do**
    $r_a \sim Beta(\alpha_a, \beta_a)$.
  end for
  Choose $a^* = \arg\, max_{a \in \{1,...,N\}} r_a$.
  Perform action a^* and sample the subset $\bar{\mathcal{D}} \subset \mathcal{D}_{a^*}$ with sampling size $\triangle S$.
  Calculate the success amount n_r in the iteration by Equation (13).
  Update the α and β by Equation (14) $\sim$ Equation (16).
  Update $\mathcal{D}_{init} \leftarrow \mathcal{D}_{init} \oplus \bar{\mathcal{D}}$.
  Update $b \leftarrow b - \triangle S \times \mathrm{p}_{a^*}$.
end while
return: $\mathcal{D}_{init}$

The discount factor γ will be implemented on the cumulative reward and cumulative failure that is shown below:

$$S_i(t) = \begin{cases} \gamma S_i(t-1) + n_r(t), & a_t = i \\ \gamma S_i(t-1), & a_t \neq i, \quad \gamma \in (0, 1] \end{cases} \tag{15}$$

$$F_i(t) = \begin{cases} \gamma F_i(t-1) + \left|\bar{D}_t\right| - n_r(t), & a_t = i \\ \gamma F_i(t-1), & a_t \neq i, \quad \gamma \in (0, 1] \end{cases} \tag{16}$$

The algorithm of RDA is outlined in Algorithm 2. The algorithm starts by inputting the state variables. Then, we set reward distributions for each action with a Beta distribution $Beta(\alpha, \beta)$. During each interaction, a reward r_a is randomly sampled from the distribution of each possible action a. The action with the highest reward a^* is then selected, and a subset from the corresponding dataset $\mathcal{D}_{a^*}$ is sampled by sampling size $\triangle S$. The success amount n_r is calculated using Equation (13). Then, the values of α and β are updated based on Equation (14) $\sim$ Equation (16). Finally, the newly obtained subset $\bar{\mathcal{D}}$ is merged into the obtained datasets $\mathcal{D}_{init}$ and the remained budget b is updated. The algorithm terminates when the budget is exhausted.

Experiments

In this section, we establish a data acquisition environment to simulate the online data market and conduct thorough evaluations to verify the effectiveness of the proposed data acquisition algorithms, KDA and RDA.

Settings

Datasets We conduct experiments on the real-world dataset, AdultCensus (Kohavi 1996) $\mathcal{D}_{AdultCensus}$, which is extracted from a 1994 Census Bureau database of people records and is processed to include 12 individual features. We choose the label as whether a person's income is higher than (positive class) or less than (negative class) 50K per year. We remove the nan items from the dataset, leaving $32,561$ records remaining. Gender is chosen as a sensitive attribute.

Baselines We define two baseline strategies, called Uniform − Budget and Uniform − Amount, which represent flat data acquisition approaches for comparison with our own methods. Additionally, we create two variations of our methods, No − fair − KDA and No − fair − RDA, to serve as control groups for experiments that do not consider ethical fairness.

1. Uniform − Budget: The consumer allocates their budget evenly among different dataset commodities.
2. Uniform − Amount: The consumer acquires the same number of samples from different data commodities.
3. No − fair − KDA: No − fair − KDA is a variant of KDA, the knowledge-mining stage of No − fair − KDA only considers the estimation error under U_{Acc}; its knowledge-utilizing stage solves the data acquisition problem without fairness consideration.
4. No − fair − RDA: No − fair − RDA is a variant of RDA. In No − fair − RDA, a newly acquired data point is rewarded with 1 as long as it can improve accuracy.

Preprocess To create a simulated data market with various data sources, we preprocess each dataset in four steps: removing bias, splitting, injecting bias, and injecting noise.

1. Removing bias: To remove the dependence of the label on sensitive attributes in the AdultCensus, we create counterfactual datasets. These new datasets have the same features as the original datasets but have sensitive attributes that are opposite in meaning and are appended to the original datasets.
2. Splitting: We randomly select 10% samples from each of the bias-removed datasets to form the test datasets, and 5% samples from the original datasets to serve as the initial data for consumers. Then, we randomly select $50,000$ samples from the remaining data for the original AdultCensus. These samples are equally split into 5 parts to create multiple data commodities.
3. Injecting bias: We first introduce a concept of bias level ranging from 0 to 1. The bias level represents the proportion of samples from the original dataset; the remaining samples come from its counterfactual. A bias level of 0.5 represents a sensitive attribute-label balanced dataset, and a bias level of 0 and 1 represents high bias on opposite sensitive groups. We introduce different levels of bias to the split datasets.
4. Injecting noise: We randomly select a proportion of labels in each dataset for replacement, with a replacement ratio of $10\% \sim 20\%$ for each dataset commodity.

As a result, we obtain 5 heterogeneous datasets $\mathcal{D}_1 \sim \mathcal{D}_5$ that simulate scenarios where data comes from 5 different sources.

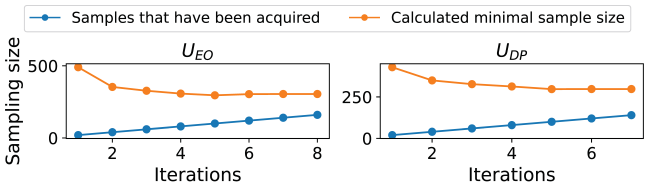

Figure 2: Illustration of iterations in minimum sampling size $\bar{S}_{min}$ calculation under U_{EO} fairness measure.

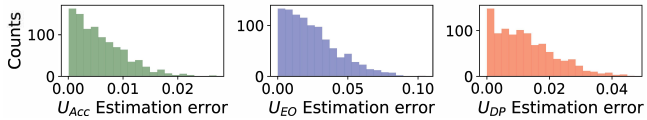

Figure 3: $\epsilon - \delta$ approximation in 1000 times with $\bar{S}_{min}$.

Hyperparameters The choice of ϵ_{acc} and ϵ_{eo} (ϵ_{dp}) in Algorithm 1 is crucial for balancing the representativeness of the minimum subset and the budget cost in the knowledge-mining stage. We choose ϵ_{acc} and ϵ_{eo} (ϵ_{dp}) to be one order of magnitude smaller than the upper bounds of U_{Acc} and U_{EO} (U_{DP}), respectively. For instance, since U_{Acc} varies from 0 to 1, we set ϵ_{acc} to $1e-1$. Moreover, we select significance levels of δ_{acc} and δ_{eo} (δ_{dp}) to be $1e-3$ to ensure high confidence in the results.

KDA: Knowledge-mining Stage

In this experiment, we verify the effectiveness of the knowledge mining, which includes: (1) the convergence of the multi-step iterative calculation of minimum sampling size $\bar{S}_{min}$; and (2) the minimum sampling set $\bar{\mathcal{D}}_{min}$ guarantees $U_{Acc}(\bar{\mathcal{D}}_{min})$, $U_{EO}(\bar{\mathcal{D}}_{min})$ and $U_{DP}(\bar{\mathcal{D}}_{min})$ with ϵ-δ approximation to $U_{Acc}(\mathcal{D})$, $U_{EO}(\mathcal{D})$ and $U_{DP}(\mathcal{D})$, respectively. In data settings, we generate an initial $\mathcal{D}_{init}$ and a provided dataset $\mathcal{D}$ using the pre-processing method described above. We run Algorithm 1 under U_{EO} and U_{DP} respectively, and the minimum sampling size $\bar{S}_{min}$ calculated in each step is shown by the orange line in Figure 2, respectively. The calculation process performs convergently. The blue line represents the samples collected during the multiple steps of the iterative process. In each step, we set the increment of sampling size to be small enough, 20 in our case, to avoid acquiring samples that exceeded the final computed $\bar{S}_{min}$, preventing unnecessary sample waste in the knowledge-mining stage. After calculating the minimum sampling size $\bar{S}_{min}$, we use random sampling to fill the minimum dataset $\bar{\mathcal{D}}_{min}$, which is a subset of $\mathcal{D}$ with size $\bar{S}_{min}$. We measure the statistical estimation error of $\bar{\mathcal{D}}_{min}$ to the true $\bar{\mathcal{D}}$ in terms of performance in 1000 trials, presented in Figure 3. We observe that in 1000 trials (significance level $\delta = 1e-3$), the estimation error of both $\left|U_{Acc}(\mathcal{D}) - U_{Acc}(\bar{\mathcal{D}}_{min})\right|$), $\left|U_{EO}(\mathcal{D}) - U_{EO}(\bar{\mathcal{D}}_{min})\right|$) and $\left|U_{DP}(\mathcal{D}) - U_{DP}(\bar{\mathcal{D}}_{min})\right|$) are less than the pre-defined estimation error tolerance $\epsilon_{acc} = 0.1$, $\epsilon_{eo} = 0.1$ and $\epsilon_{dp} = 0.1$, respectively.

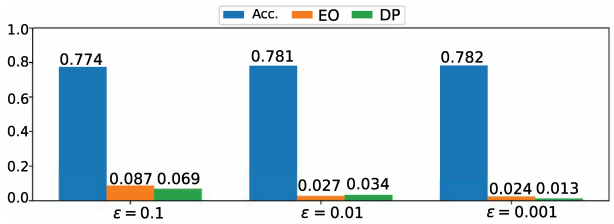

Figure 4: The effect of fairness constraint bound ε.

KDA: Effect of Fairness Constraint ε

We verify the impact of different fairness constraint bound ε in Equation (11) on the fairness utility of the acquired data. Figure 4 shows the classifier performance, including accuracy $Acc.$, and EO and DP values , after training on the data obtained under different settings of $\varepsilon = [0.1, 0.01, 0.001]$. The results indicate that ε controls the fairness utility of the acquired data. Specifically, the datasets obtained under a lower ε resulted in a fairer classifier. Based on these findings, we selected $\varepsilon = 0.001$ for subsequent experiments.

KDA vs. RDA under Different Budgets

In this experiment, we evaluate the performance of KDA and RDA under different budget amounts $\mathbf{B} = [1K, 5K, 10K]$ for the data consumer. The prices of data products from different sources follow a uniform distribution $\mathrm{p} \sim U(0.9, 1.1)$. According to Table 3, when the consumer has a very low budget ($\mathbf{B} = 1K$), the datasets acquired by KDA can achieve at least 77% fairness improvement when using EO metric and 81% fairness improvement when using DP metric for classifier training, compared to all baselines; the datasets acquired by RDA can achieve at least 29% fairness improvement when using EO metric and 43% fairness improvement when using DP metric for classifier training, compared to all baselines; when the consumer has a middle budget ($\mathbf{B} = 5K$), the datasets acquired by KDA can achieve at least 25% fairness improvement when using EO metric and 42% fairness improvement when using DP metric for classifier training, compared to all baselines; the datasets acquired by RDA can achieve at least 3% fairness improvement when using EO metric and 5% fairness improvement when using DP metric for classifier training, compared to all baselines; when the budget is sufficient ($\mathbf{B} = 10K$), the datasets acquired by KDA can achieve at least 76% fairness improvement when using EO metric and 76% fairness improvement when using DP metric for classifier training, compared to all baselines; the datasets acquired by RDA can achieve at least 56% fairness improvement when using EO metric and 69% fairness improvement when using DP metric for classifier training, compared to all baselines.

KDA vs. RDA under Different Dependence Degrees

In this experiment, we evaluate the performance of KDA and RDA under different dependence degrees of the datasets in the data market. The dependence degree measures the similarity of the distributions of the datasets, which we simulate by varying the proportion of overlap between candi-

Method	Budget								
	1K			5K			10K		
	Acc.	EO	DP	Acc.	EO	DP	Acc.	EO	DP
Uniform − Budget	0.604	0.179	0.215	0.741	0.102	0.051	0.770	0.050	0.027
Uniform − Amount	0.602	0.189	0.377	0.751	0.076	0.044	0.772	0.054	0.030
No − fair − KDA	0.727	0.133	0.123	0.771	0.035	0.019	0.782	0.070	0.026
No − fair − RDA	0.695	0.250	0.314	0.775	0.027	0.027	**0.784**	0.055	0.027
KDA	0.728	**0.030**	**0.023**	0.774	**0.020**	**0.011**	0.776	**0.012**	**0.006**
RDA	**0.759**	0.094	0.070	**0.777**	0.026	0.018	0.778	0.022	0.008

Table 3: The effect of budget in KDA and RDA.

Method	Dependence Degree								
	10%			30%			50%		
	Acc.	EO	DP	Acc.	EO	DP	Acc.	EO	DP
Uniform − Budget	0.771	0.060	0.020	0.765	0.073	0.116	0.757	0.212	0.075
Uniform − Amount	0.768	0.065	0.033	0.778	0.051	0.019	0.758	0.101	0.041
No − fair − KDA	**0.787**	0.042	0.017	**0.784**	0.040	0.023	0.767	0.084	0.049
No − fair − RDA	0.777	0.076	0.035	0.783	0.075	0.026	0.781	0.043	0.024
KDA	0.775	**0.013**	**0.014**	0.776	0.037	0.017	0.773	0.060	0.024
RDA	0.779	0.017	0.016	0.776	**0.035**	**0.015**	**0.786**	**0.022**	**0.013**

Bold: Best performance compared to all baselines.
Line: Second-best performance compared to all baselines.

Table 4: The effect of dependence degree in KDA and RDA.

Classifier Models	Runtime(s)	
	KDA	RDA
LogisticRegression	5.961	1.609
MLP(100)	5.674	237.952
MLP(200)	8.243	505.079

Table 5: Run-time comparison of KDA and RDA.

date datasets: $Overlap = [10\%, 30\%, 50\%]$. Table 4 shows that KDA is affected by increases in dependence, with the fairness utility of the acquired data decreasing. This is because KDA solves unbiased data acquisition problem in the knowledge-utilizing stage based on the separated utility of each dataset without considering the interdependence among them and the combination of datasets. RDA achieves higher robustness under various degrees of dependence, as it dynamically updates the reward of choosing each candidate dataset during iterations, taking into account the adding order and combination of the datasets.

KDA vs. RDA in Runtime

As in above, we verify that RDA has better robustness than KDA under different dependence degrees. However, The robustness comes at the expense of runtime cost, as the reward function in RDA requires retraining the classifier on newly acquired dataset in each iteration. We compare the runtime of KDA and RDA under different classifier models on CPU: Intel® Xeon® Gold 5318Y CPU@ 2.10Ghz in Table 5. We use different classifiers and as complexity increases, the runtime cost also increases significantly in RDA. Based on our experimental findings, we conclude that KDA has a higher runtime efficiency and is more suitable for complex model training task, while RDA is more adaptable to the variation or perturbation of the market settings.

Conclusion

In this paper, we formally analyzed the potential fairness issues related to acquiring datasets from online data markets. We proposed a fairness-aware data acquisition framework, FAIRDA, designed to obtain high-quality datasets that maximize both accuracy and fairness for local classifier training within a limited budget. We formally defined the unbiased data acquisition problem and introduced two efficient algorithms: knowledge-based data acquisition (KDA) and reward-based data acquisition (RDA). These algorithms enable the development of an optimal data acquisition strategy through interactions supported by online dataset markets. We conducted comprehensive experiments to validate the effectiveness of our framework across various scenarios, including different budget constraints and datasets heterogeneity. The experimental results demonstrated that our methods can effectively guide consumers in acquiring fairer and more ethical ML models from online data markets.

Acknowledgments

This work was supported by Key Programs of Guangdong Province under Grant $2021QN02X166$. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

Amend, J. J.; and Spurlock, S. 2021. Improving machine learning fairness with sampling and adversarial learning. *Journal of Computing Sciences in Colleges*, 36(5): 14–23.

An, B.; Xiao, M.; Liu, A.; Xie, X.; and Zhou, X. 2021. Crowdsensing Data Trading based on Combinatorial Multi-Armed Bandit and Stackelberg Game. In *Proceedings of the 37th IEEE International Conference on Data Engineering*, 253–264. IEEE.

Azcoitia, S. A.; and Laoutaris, N. 2022. A Survey of Data Marketplaces and Their Business Models. *SIGMOD Record*, 51(3): 18–29.

Bajoudah, S.; Dong, C.; and Missier, P. 2019. Toward a Decentralized, Trust-Less Marketplace for Brokered IoT Data Trading Using Blockchain. In *Proceedings of the 2019 IEEE International Conference on Blockchain*, 339–346. IEEE.

Balas, E.; and Zemel, E. 1980. An algorithm for large zero-one knapsack problems. *Operations Research*, 28(5): 1130–1154.

Brickley, D.; Burgess, M.; and Noy, N. 2019. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In *The World Wide Web Conference*, 1365–1375. ACM.

Byeon, H.; Kumar, P.; Manu, M.; Maaliw, R. R.; Singh, P. P.; and Maranan, R. 2023. Editable Blockchain for Secure IoT Transactions. In *Proceedings of the 2nd International Conference On Smart Technologies For Smart Nation*, 255–263.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, 13–18. IEEE.

Chai, C.; Liu, J.; Tang, N.; Li, G.; and Luo, Y. 2022. Selective Data Acquisition in the Wild for Model Charging. *Proceedings of the VLDB Endowment*, 15(7): 1466–1478.

Chen, J.; Li, M.; and Xu, H. 2022. Selling Data To a Machine Learner: Pricing via Costly Signaling. In *Proceedings of the 39th International Conference on Machine Learning*, 3336–3359. PMLR.

Dixit, A.; Singh, A.; Rahulamathavan, Y.; and Rajarajan, M. 2023. FAST DATA: A Fair, Secure, and Trusted Decentralized IIoT Data Marketplace Enabled by Blockchain. *IEEE Internet of Things Journal*, 10(4): 2934–2944.

Dudziński, K.; and Walukiewicz, S. 1987. Exact methods for the knapsack problem and its generalizations. *EJOR*, 28(1): 3–21.

D'Ambrosio, C.; and Martello, S. 2011. Heuristic algorithms for the general nonlinear separable knapsack problem. *Computers & Operations Research*, 38(2): 505–513.

Fahse, T.; Huber, V.; and van Giffen, B. 2021. Managing bias in machine learning projects. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, 94–109. Springer.

Fanchang, F.; Yadav, A.; Khan, A.; Azam, S.; and Yadav, D. 2023. Trusted Computing Data Transaction Scheme based on Blockchain. In *Proceedings of the 7th International Conference on IoT in Social, Mobile, Analytics and Cloud*, 276–281.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Fernandez, R. C.; Subramaniam, P.; and Franklin, M. J. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proceedings of the VLDB Endowment*, 13(12): 1933–1947.

Figueredo, K.; Seed, D.; and Wang, C. 2022. A Scalable, Standards-Based Approach for IoT Data Sharing and Ecosystem Monetization. *IEEE Internet of Things Journal*, 9(8): 5645–5652.

Firouzi, F.; Farahani, B.; Barzegari, M.; and Daneshmand, M. 2022. AI-Driven Data Monetization: The Other Face of Data in IoT-Based Smart and Connected Health. *IEEE Internet of Things Journal*, 9(8): 5581–5599.

Fisher, R. A. 1992. *Statistical methods for research workers*. Springer.

Fruhwirth, M.; Rachinger, M.; and Prlja, E. 2020. Discovering business models of data marketplaces. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 5738–5747.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gong, C.; Zheng, Z.; Wu, F.; Shao, Y.; Li, B.; and Chen, G. 2023. To Store or Not? Online Data Selection for Federated Learning with Limited Storage. In *Proceedings of the ACM Web Conference 2023*, 3044–3055. ACM.

Granmo, O.-C.; and Oommen, B. J. 2011. Learning automata-based solutions to the optimal web polling problem modelled as a nonlinear fractional knapsack problem. *Engineering Applications of Artificial Intelligence*, 24(7): 1238–1251.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th Advances in neural information processing systems*, 3315–3323. Curran Associates, Inc.

Hayashi, T.; and Ohsawa, Y. 2020. TEEDA: an interactive platform for matching data providers and users in the data marketplace. *Information*, 11(4): 218.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

Hochbaum, D. S. 1995. A nonlinear Knapsack problem. *Operations Research Letters*, 17(3): 103–110.

Koesten, L. 2018. A User Centred Perspective on Structured Data Discovery. In *Companion Proceedings of the 27th World Wide Web*, 849–853. IW3C2.

Kohavi, R. 1996. Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 202–207. AAAI Press.

Li, H.; Zhao, G.; Johari, R.; and Weintraub, G. Y. 2022. Interference, Bias, and Variance in Two-Sided Marketplace Experimentation: Guidance for Platforms. In *Proceedings of the 31st ACM Web Conference*, 182–192. ACM.

Li, Q.; Li, Z.; Zheng, Z.; Wu, F.; Tang, S.; Zhang, Z.; and Chen, G. 2023. Capitalize Your Data: Optimal Selling Mechanisms for IoT Data Exchange. *IEEE Transactions on Mobile Computing*, 22(4): 1988–2000.

Li, Y.; Yu, X.; and Koudas, N. 2021. Data Acquisition for Improving Machine Learning Models. *Proceedings of the VLDB Endowment*, 14(10): 1832–1844.

Li, Y.; Yu, X.; and Koudas, N. 2024. Data Acquisition for Improving Model Confidence. *Proceedings of the ACM on Management of Data*, 2(3).

Maher, M.; and Khan, I. 2022. From sharing to selling: challenges and opportunities of establishing a digital health data marketplace using blockchain technologies. *Blockchain in Healthcare Today*, 5.

Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; et al. 2023. DataPerf: Benchmarks for Data-Centric AI Development. In *Advances in Neural Information Processing Systems*, 5320–5347. Curran Associates, Inc.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6).

Newman, A.; Bavik, Y. L.; Mount, M.; and Shao, B. 2021. Data collection via online platforms: Challenges and recommendations for future research. *Journal of Applied Psychology*, 70(3): 1380–1402.

Pandey, S. R.; Pinson, P.; and Popovski, P. 2023. Strategic Coalition for Data Pricing in IoT Data Markets. *IEEE Internet of Things Journal*, 11(4): 6454–6468.

Raj, V.; and Kalyani, S. 2017. Taming non-stationary bandits: A Bayesian approach. *arXiv preprint arXiv:1707.09727*.

Roselli, D.; Matthews, J.; and Talagala, N. 2019. Managing Bias in AI. In *Companion Proceedings of 28th World Wide Web Conference*, 539–544. ACM.

Salimi, B.; Rodriguez, L.; Howe, B.; and Suciu, D. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, 793–810.

Tae, K. H.; and 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*, 1771–1783. ACM.

Tommasi, T.; Patricia, N.; Caputo, B.; and Tuytelaars, T. 2017. A deeper look at dataset bias. *Domain adaptation in computer vision applications*, 37–55.

Van Miegroet, H. J.; Alexander, K. P.; and Leunissen, F. 2019. Imperfect Data, Art Markets and Internet Research. *Arts*, 8(3).

Wang, S.; and Tsang, D. H. 2023. A Dynamic Data Trading Marketplace With Externalities. *IEEE Internet of Things Journal*, 11(7): 12745–12754.

Yoon, J.; Arik, S.; and Pfister, T. 2020. Data Valuation using Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, 10842–10851.

Zheng, Z.; Peng, Y.; Wu, F.; Tang, S.; and Chen, G. 2017. An Online Pricing Mechanism for Mobile Crowdsensing Data Markets. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 1–10. ACM.