Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 6217495, 12 pages
https://doi.org/10.1155/2023/6217495

Research Article

A Multidimensional Data Utility Evaluation and Pricing Scheme in the Big Data Market

Yuling Chen,[1,2] **Rui Bai**,[1,3] **Yongtang Wu**,[2,4] **Tao Li**,[1] **and Hui Zhou**[1]

[1]*State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou 550025, China*
[2]*Blockchain Laboratory of Agricultural Vegetables, Weifang University of Science and Technology, Weifang, Shandong 262700, China*
[3]*Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China*
[4]*Weifang Key Laboratory of Blockchain on Agricultural Vegetables, Weifang, Shandong 262700, China*

Correspondence should be addressed to Rui Bai; bair0412@163.com

Received 29 July 2022; Revised 5 October 2022; Accepted 15 October 2022; Published 21 February 2023

Academic Editor: Xu Zheng

Big data as a derivative of information technology facilitates the birth of data trading. The technology surrounding the business value of big data has come into focus. However, most of the current research focuses on improving the performance of big data analytics algorithms. Data pricing is still one of the main issues in data trading. Therefore, we aim to tackle the problem of evaluating the utility of data in the big data trading market and the problem of maximizing the profits of the various roles involved in data trading. To this end, we propose a Multidimensional Data Utility Evaluation (MDDUE) method through three data quality dimensions, namely, data size, availability, and completeness. Next, we propose a big data trading market model including data providers, service providers, and service users. An optimal data-pricing scheme based on a three-party Stackelberg game is proposed to maximize the participants' profits. Finally, a machine learning model is used to verify the rationality and validity of the MDDUE. The results show that MDDUE can evaluate the utility of data more accurately than previous work. The existence and uniqueness of the Nash equilibrium are demonstrated through numerical experiments.

1. Introduction

With the rapid development of the Internet of Vehicles (IoV), Internet of Things (IoT), and other information technologies, the amount of data generated globally on a daily basis is heavily aggregated and exploding [1, 2]. The huge volumes and diverse sources of big data also make its application in various fields increasingly widespread [3, 4]. Big data is gradually becoming a basic resource alongside land and oil. However, emerging privacy concerns have prevented data owners from sharing their datasets [5]. In addition, the collection and storage of a large amount of data lead to some security issues in IoT [6]. Specifically, data is stored and maintained independently in different departments and isolated from each other, which hinders the effective circulation of data and prevents the value of big data from being fully realized. These problems lead to isolated data islands. To consider data resources as a commodity to be shared and circulated, it is important to establish an efficient data trading market [7]. There are three roles in the three-party traditional big data trading market: data owners, data trading platforms, and data users such as the Data Marketplace, Big Data Exchange, and Microsoft Azure Marketplace. Data owners collect data through IoT sensors and the Internet, etc. [8]. They consign the raw data to data trading platforms for data users to choose from and purchase [9]. But there are some problems with the traditional model of data trading, such as malicious data trading platforms or data consumers who may illegally cache and resell the data of data owners. It is not only an issue of copyright but also of privacy [10, 11]. It has to ensure high value and accuracy when sharing the sensing data [12]. As a result, how to

promote the secure development of big data trading and building a digital economy with data as a key element has become a major challenge in the big data era. The introduction of the concept of Big Data as a Service (BDaaS) [13] has enabled big data services to become a commodity instead of raw data. In the big data service-trading market, what the service users need is not the entire raw dataset, but the data-processing results or data services based on the dataset. Instead of sending the raw data to the service providers, the data providers process the raw data into various data products and data services for the data consumers to purchase and use [14]. For example, service providers can provide users with personalized recommendations based on big data services [15–17].

Few existing studies have considered the big data trading market from an economic perspective. The pricing strategies of data products or data services are difficult to define, and various pricing strategies are still incomplete. The big data trading market has not yet developed a uniform pricing standard. The literature [18] investigates the quantification of data utility from a data science perspective and proposes an optimal pricing scheme based on a big data trading market model. The literature [19] introduces the concept of the signal-noise ratio in the electronic information fields to evaluate the data utility. Based on [18, 19], the literature [14] evaluates the data utility in terms of two dimensions: data size and data noise level. However, we cannot overlook the impact that the data class-balanced ratio has on the utility of the data. Game theory is used in many scenarios as a practical tool to solve optimal problems (e.g., IoV) [20]. We complement the data utility evaluation method by introducing a class-balanced ratio. Meanwhile, we use game theory to maximize the profits of the three roles in the big data trading market.

In the big data trading market, the lifecycle of big data is mainly divided into five stages: data collection and uploading, data analysis, data pricing, data trading, and data protection [21, 22]. We can use blockchain to ensure the fairness of data trading [23–25]. Cryptographic methods can be used to guarantee the security of data [26, 27]. Our work mainly lies in the third stage.

The main contributions of this paper are as follows.

(1) We propose the Multidimensional Data Utility Evaluation (MDDUE) method that considers data size, availability, and completeness. The MDDUE takes into account more data quality dimensions and is more accurate than the literature [14]

(2) We introduce a service-based big data trading market model. In addition, a pricing scheme based on the Stackelberg game is proposed to maximize the profits of the three parties in the big data trading market

(3) Validating the rationality and applicability of the MDDUE through a machine learning model. The existence and uniqueness of the Nash equilibrium are proven using backward induction, and the numerical experiments show that the proposed pricing scheme can maximize the profits of the three parties

The rest of the paper is organized as follows. Section 2 presents the related work about data pricing. We introduce our scheme in Section 3, including the Multidimensional Data Utility Evaluation (MDDUE) method and the optimal data-pricing scheme. Section 4 shows the experiment results and analysis. Finally, we conclude this paper in Section 5.

2. Related Work

The emergence of data trading has facilitated the effective flow of data and provided a channel to fully exploit the value of big data. Information products are different from traditional goods in that they are easy to copy, modify, and spread. These characteristics make data trading different from traditional commodity trading and require appropriate specifications to address the specific features of data products. How to establish a uniform pricing strategy is one of the challenges facing the big data trading market. There are three main challenges to data pricing: diverse data sources, the complexity of data management, and the diversity of data [21]. The authors in [21] summarised the current data-pricing strategies and pricing models in the big data market. For example, data-pricing strategies are classified into six main categories: Free Data Strategy, Usage-Based Pricing Strategy, Package Pricing Strategy, Flat Pricing Strategy, Two-Part Tariff Strategy, and Freemium Strategy. Based on the strategies above, there are two main pricing models: the economic-based pricing model and the game theory-based pricing model.

Some scholars have tried to study data pricing from other perspectives. Koutris et al. [28] proposed query-based pricing which allows generating the price of any query automatically, and the pricing algorithm satisfies no arbitrage and no discount. Shen et al. [29] proposed a tuple granularity-based pricing model for personal big data. The model can be automatically adjusted according to the attributes that affect the value of the data. Inspired by information entropy, Li et al. [30] proposed a new data-pricing method based on data information entropy and gave a pricing function based on the results of the method mentioned above. They have done a lot of experiments to verify the method. The paper inspires research concerning the pricing mechanism of big data. Cai et al. [31] proposed a new privacy-preserving data trading framework for web-browsing histories. The framework takes into account the privacy preferences of different users and compensates the users for the privacy of their data according to the degree of privacy leakage. To reduce the heavy burdens and private leakage of data exchange in the IoT, Cai et al. [32, 33] proposed a novel framework for range-counting trading over IoT networks by jointly considering data utility, bandwidth consumption, and privacy preservation. However, all of the above methods do not take into account the optimal profits of the participants.

Data quality is one of the factors that influence the quality of machine learning models, which opens up a new way

of thinking about data pricing. Stahl and Vossen [34] summarised seven metrics for evaluating data quality: accuracy, amount of data, availability, completeness, latency, response time, and timeliness. Niyato et al. [18] and Yang et al. [19] built a data utility evaluation function via data size and noise level, respectively, in other words, the amount of data and availability. Xiao et al. [14] combined both dimensions to quantify the value of data. But in some cases, the method cannot evaluate the data utility accurately. It has been found that the completeness of the data also has a relatively large impact on the accuracy of machine learning models, also known as the imbalance ratio [35]. Therefore, we introduce completeness into the data utility evaluation function which we called the class-balanced ratio.

3. System Model

We first describe the data utility evaluation method MDDUE in detail. Then, we describe a big data trading market and formulate an optimal pricing problem based on the Stackelberg game to maximize the profits of the participants. This study uses the method of Xiao et al., and the description of the method partly reproduces their wording [14]. The symbols used in this paper commonly are shown in Table 1 below.

Table 1: List of commonly used symbols.

Symbol	Definition
D	Dataset
n	Data size
α	Data nonnoise ratio
β	Data class-balanced ratio
u	Data utility
P_s	Data service subscription price per unit of data utility
P_u	Data price per unit of data utility
$\delta(\cdot)$	Accuracy function for data size
$\mu(\cdot)$	Accuracy function for data nonnoise ratio
$\varphi(\cdot)$	Accuracy function for data class-balanced ratio
$U(\cdot)$	Data utility evaluation function
$\mathrm{Pro}_{\mathrm{DP}}(\cdot)$	Profit function of the data provider
$\mathrm{Pro}_{\mathrm{SP}}(\cdot)$	Profit function of the service provider
$\mathrm{Inc}_{\mathrm{SU}}(\cdot)$	Reward function of the service user
$\mathrm{Pro}_{\mathrm{SU}}(\cdot)$	Profit function of the service user
C_{DP}	Cost per unit of data utility

3.1. MDDUE: Multidimensional Data Utility Evaluation. To price data, it is necessary to evaluate the utility of unstructured big data. CNN-based machine learning algorithms are increasingly used in a wide range of applications, such as face recognition, intrusion detection, and natural language processing [36, 37]. As an important technique for data analysis, machine learning is also an effective tool for evaluating the value of data. The process of machine learning providing data services to service users is shown in Figure 1.

The quality of raw data is very important for machine learning models. In the machine learning model, a raw dataset that has n tuples can be presented as $D = \{(\overrightarrow{x_i}, y_i) | i = 1, \cdots, n\}$, where $\overrightarrow{x_i}$ is a feature set of data samples. y_i is the class label of $\overrightarrow{x_i}$. Supervised learning is widely used in classification and prediction problems [38].

We will introduce three dimensions that affect the data quality, namely, data size, availability, and completeness, next. We adopt the nonnoise ratio of data instead of availability and the class-balanced ratio of data instead of completeness. By changing these data quality dimensions, different data versions can be customized to meet user demand for data service quality.

Data size n: we assume that the accuracy of the machine learning model is a_i when the dataset size is n_i where i is the index of the experimental datasets. To determine the accuracy function $\delta(n)$, we set different data sizes $n_1, n_2, \cdots, n_m$ under the same other conditions. We can get a set of experimental points $\{(n_1, a_1), (n_2, a_2), \cdots, (n_m, a_m)\}$ after training the machine learning model using a series of datasets. We can apply least squares to minimize the mean-squared error to find the accuracy function $\delta(n)$ to fit these points.

$$\min \frac{1}{m} \sum_{i=1}^{m} (a_i - \delta(n_i))^2. \tag{1}$$

Data nonnoise ratio α: in the real world, labeling requires a certain amount of expertise. For example, in medical imaging, even experts have different opinions on the labels. So there will be noisy labels in the dataset [39]. The existence of noisy labels causes significant performance degradation of machine learning models and hence the availability of the dataset [40]. That is why we use the data nonnoise ratio instead of availability.

The data nonnoise ratio is the proportion of the total data set without noisy data. The data nonnoise ratio is denoted as

$$\alpha = 1 - \frac{z}{n}, \tag{2}$$

where z is the number of data with noisy labels and n is the data size.

We construct datasets with different data nonnoise ratios by the following method:

(1) Select z samples from the dataset randomly

$$z = n(1 - \alpha). \tag{3}$$

(2) Replace the labels of these z samples with other random labels

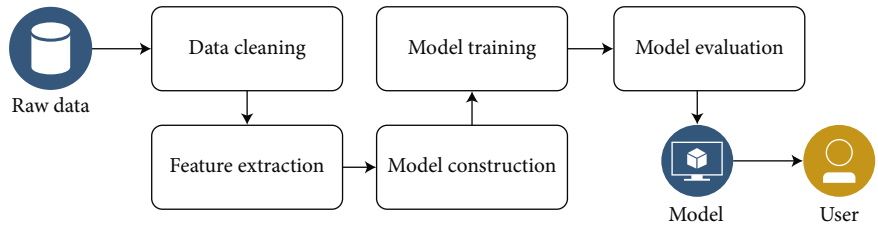

Figure 1: Machine learning in big data service.

Similarly, we assume that the accuracy of the machine learning model is b_i when the data nonnoise ratio is α_i with the method above. We can also acquire the set of a series of experimental points $\{(\alpha_1, b_1), (\alpha_2, b_2), \cdots, (\alpha_d, b_d)\}$ with different data nonnoise ratios. Then, we can determine the accuracy function $\mu(\alpha)$ by minimizing the mean-squared error.

$$\min \frac{1}{d}\sum_{i=1}^{d}(b_i - \mu(\alpha_i))^2. \tag{4}$$

Data class-balanced ratio β: in the real-world dataset, the number of samples in one class in the dataset is too small compared to those in other classes due to various reasons such as sampling difficulties, which also affect the accuracy of machine learning models [35]. The data class-balanced ratio is the degree of balance in the number of samples in the dataset. We define the data-balanced ratio as the inverse of the imbalanced ratio in [35].

$$\beta = \frac{\min_j \{n_1, n_2, \cdots, n_i\}}{\max_k \{n_1, n_2, \cdots, n_i\}}. \tag{5}$$

Other things being equal, we assume that the accuracy of the machine learning model can reach c_i when the data class-balanced ratio is β. We determine the accuracy function $\varphi(\beta_i)$ of machine learning using the set of experimental points $\{(\beta_1, c_1), (\beta_2, c_2), \cdots, (\beta_h, c_h)\}$ obtained by conducting experiments at different class-balanced ratios as above.

$$\min \frac{1}{h}\sum_{i=1}^{h}(c_i - \varphi(\beta_i))^2. \tag{6}$$

All other things being equal, the higher the quality of the dataset, the higher the accuracy of the model. Therefore, the accuracy of the machine learning model can be equated with data utility to some extent [14]. We can obtain datasets with different data utilities by changing the data size, nonnoise ratio, and class-balanced ratio. The function of data utility can be expressed as

$$u = U(n, \alpha, \beta). \tag{7}$$

It is known that machine learning models are more accurate with a larger dataset size, nonnoise ratio, and class-balanced ratio. But the accuracy stops getting bigger when it gets big enough. This is where it is time to optimize the model itself [40]. It is not the focus of our work.

We guess that the data utility function has the following properties:

(1) Monotonically increasing: $\delta'(n) > 0, \mu'(\alpha) > 0, \varphi'(\beta) > 0$

(2) The diminishing marginal efficiency: $\delta''(n) \le 0, \mu''(\alpha) \le 0, \varphi''(\beta) \le 0$

So we assume the data utility functions for the three impact factors individually as follows:

$$\begin{aligned} \delta(n) &= \theta_1 - \theta_2 \exp(-\theta_3 n), \\ \mu(\alpha) &= \theta_4 - \theta_5 \exp(-\theta_6 \alpha), \\ \varphi(\beta) &= \theta_7 - \theta_8 \exp(-\theta_9 \beta), \end{aligned} \tag{8}$$

where $\vec{\theta} = \{\theta_i | = 1, \cdots, 9\}$ are the fitting parameters.

The aggregated data utility evaluation function can be expressed as

$$U = \omega_1 - \omega_2 \exp(-(\omega_3 * n + \omega_4 * \alpha + \omega_5 * \beta)), \tag{9}$$

where $\vec{\omega} = \{\omega_i | i = 1, \cdots, 5\}$ are the fitting parameters.

3.2. Optimal Pricing Based on Stackelberg Game. In this section, we first describe a big data market model for selling big data services. Then, we formulate an optimal data-pricing scheme based on the *Stackelberg* game to maximize the profits of each trading participant.

We consider the big data market where a data provider provides the data and a service provider provides the data service to the service user as shown in Figure 2.

The data provider uses different tools or technologies (e.g., IoT sensors, multitarget detection in smart IoT, social media, smart devices, and social network) to collect data [41–43]. And the data provider processes the raw data so that it has the data utility the service user needs and charges the service provider. The service provider buys data from the data provider. Then, the service provider uses the dataset to train different machine learning models to provide data services to the service user. We argue that the utility of the raw data can be equated to the value of the machine learning models [44]. The service user determines the optimal

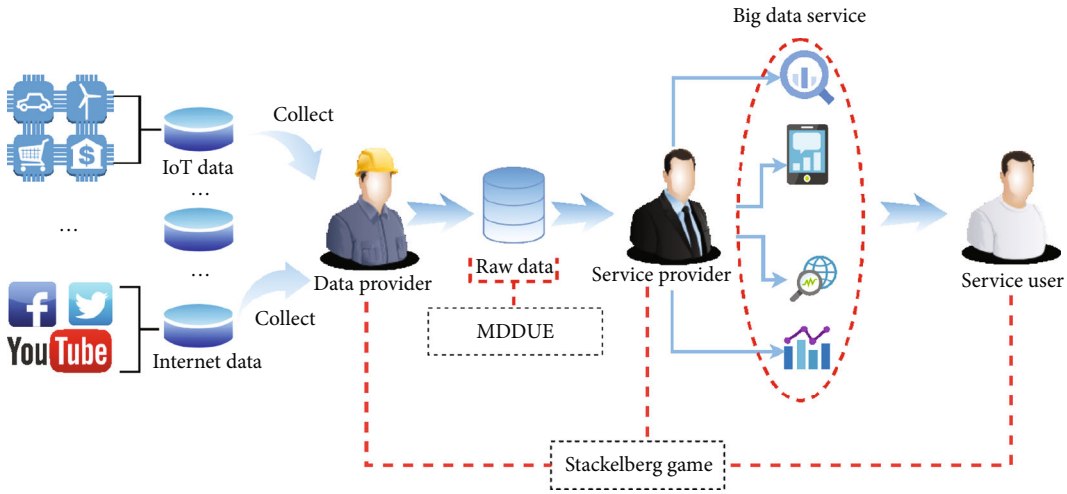

Figure 2: The big data market model.

demand for data utility based on the profit function to maximize its profits.

In the big data trading market, service users have different data utility demands for different machine learning models. Uniform pricing for all data services is unreasonable. Therefore, service providers need to set prices separately for different service users. We will explain the roles of the big data market model and its profit functions below.

(i) *Data provider*: the raw data collected by the data provider will incur storage costs, communication costs, maintenance equipment costs, etc. For easy calculation, we assume that the cost to the data provider increases linearly with the utility of the data. The cost of data processing is denoted by C_{DP} per unit of data utility. This costing approach is widely used in studies of cloud computing, the IoT, and Internet services [45]. The price of per unit data utility is denoted by P_u. We conclude that the profit function of the data provider can be expressed as

$$Pro_{DP}(u, P_u) = (P_u - C_{\mathrm{DP}})u, \tag{10}$$

where u is the data utility demand of the service user.

(ii) *Service provider*: the service provider buys data of a certain data size, nonnoise ratio, and class-balanced ratio from the data provider. Then, the service provider evaluates the utility of data through MDDUE. Finally, the service provider trains the machine learning model with the data. The service provider provides the data services or products to the service user and charges the service user at a price P_s per unit data utility. As we all know, a higher data utility means higher P_s. We can get the profit function of the service provider as

$$\mathrm{Pro}_{\mathrm{SP}}(u, P_s) = (P_s - P_u)u. \tag{11}$$

(iii) *Service user*: we assume that the service user is rational and only subscribes to the data services if the profit function is positive. Service users derive economic value from the use of data services. The higher the data utility, the higher the reward. But the reward function should be of diminishing marginal utility. Specifically, the rate of increase in rewards decreases as the data utility increases. So we assume the reward function for the service user is

$$\mathrm{Inc}_{\mathrm{SU}}(u) = \theta_1 \ln(1 + \theta_2 u), \tag{12}$$

where θ_1 and θ_2 are the experience parameters that are set by the service users.

Therefore, the profit function of service users can be expressed as

$$\mathrm{Pro}_{\mathrm{SU}}(u) = \mathrm{Inc}_{\mathrm{SU}}(u) - P_s u. \tag{13}$$

The strategies made by the three roles influence each other. The interactions of the data provider, the service provider, and the service user can be modeled as a three-stage Stackelberg game [46, 47]. In the traditional Stackelberg game, the player that makes the first decision is called the leader. After the leader, the remaining players make decisions according to the leader's decision, which are called the followers, and so on until a Nash equilibrium is reached [48].

We consider it to be a variant of the Stackelberg game. As shown in Figure 3, the game model in this paper can be expressed as three stages. In stage 1, the data provider sets the unit data utility price P_u^* to maximize profits. Then, the

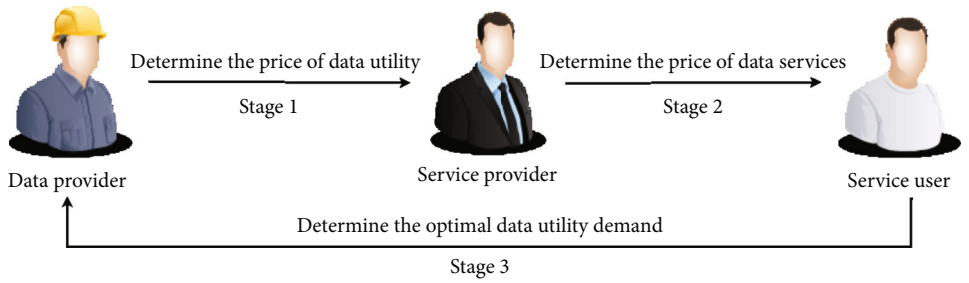

Figure 3: The three-stage Stackelberg game.

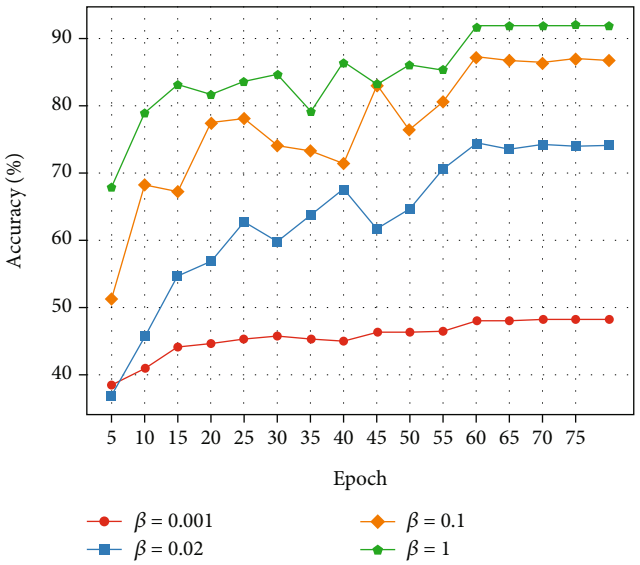

Figure 4: Training curves of WRN-34 on Cifar10. $n = 0.5, \alpha = 1$.

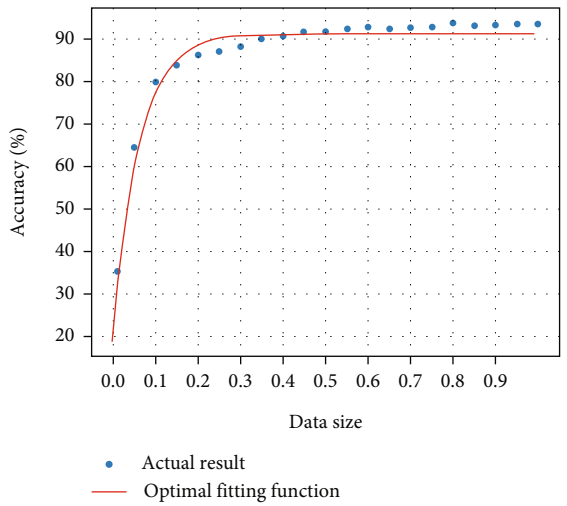

Figure 5: Accuracy of the model for different data sizes: $\alpha = 1, \beta = 1$.

service provider sets the price of service P_s^* in stage 2. And in stage 3, the service user decides the demand of data utility u^*. The three stages of the game are as follows:

Stage 1. The data provider determines the unit price of data utility P_u^*:

$$\text{Subgame } P_u^* = \arg \max_{P_u > C_{DP}} \text{Pro}_{\text{DP}}(u, P_u). \tag{14}$$

Stage 2. Given the optimal data price P_u^*, the service provider determines the price of the data service per unit data utility P_s^*:

$$\text{Subgame } P_s^* = \arg \max_{P_s > 0} \text{Pro}_{\text{SP}}(u, P_s). \tag{15}$$

Stage 3. The service user as the follower determines the data utility demand u^* according to the optimal service subscribe price P_s^* to maximize the profits:

$$\text{Subgame } u^* = \arg \max_{0<u<1} \text{Pro}_{\text{SU}}(u). \tag{16}$$

The subgame perfect equilibria of the Stackelberg game are usually solved by backward induction. At the Nash equilibrium, each player's strategy is optimal under the other player's strategies, so no players will change their strategies. The Stackelberg game is a dynamic game with full information. We assume that each player has complete information about other players in the game model.

Theorem 1. *When $\theta_1 > 0$ and $\theta_2 > 0$, the Stackelberg game has a unique Nash equilibrium (u^*, P_s^*, P_u^*).*

$$\begin{aligned} u^* &= \frac{2\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)}{3C_{DP}} - \frac{1}{\theta_2}, \\ P_s^* &= \frac{3\theta_1 C_{DP}}{2\left(\left|\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right|\right)}, \\ P_u^* &= \frac{9k^2 C_{DP}^2}{4\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)^2}, \end{aligned} \tag{17}$$

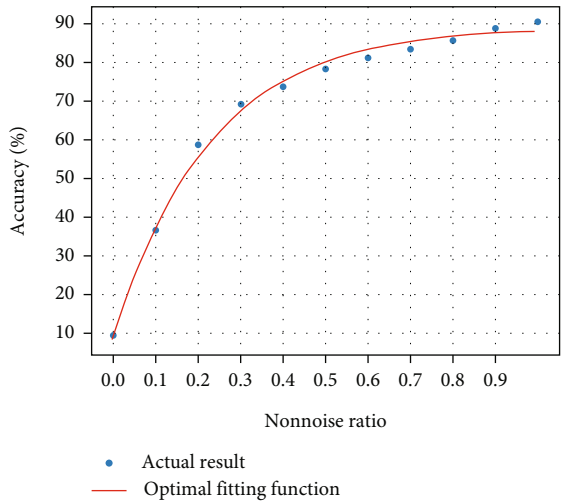

FIGURE 6: Accuracy of the model for different data nonnoise ratios: $n = 0.5$, $\beta = 1$

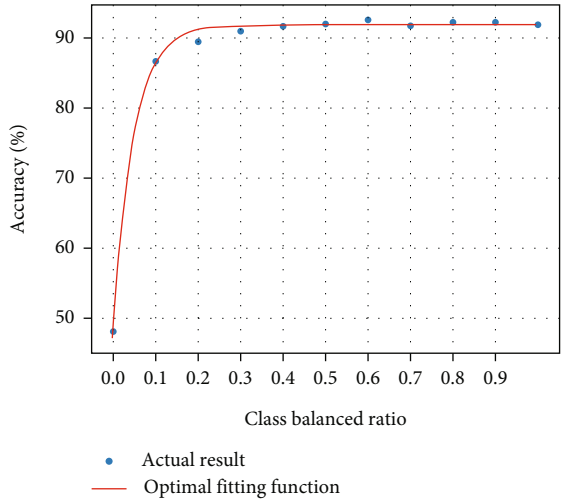

FIGURE 7: Accuracy of the model for different data class-balanced ratios: $n = 0.5$, $\alpha = 1$

where

$$Y_{1,2} = \frac{3}{8}kC_{DP}\left(-\frac{9kC_{DP}}{\theta_2} \pm \sqrt{\frac{81k^2C_{DP}^2}{\theta_2^2} + 3k^4C_{DP}}\right),$$
$$k = \frac{\theta_1}{\sqrt{\theta_1\theta_2}}. \tag{18}$$

Proof. According to the profit function (12) and the reward function (13) of the service user, we can get the finally profit function as follows:

$$\mathrm{Pro}_{\mathrm{SU}}(u) = \theta_1 \ln(1 + \theta_2 u) - P_s u. \tag{19}$$

According to (19), the first derivative of $\mathrm{Pro}_{\mathrm{SU}}(u)$ is

$$\frac{\partial \mathrm{Pro}_{\mathrm{SU}}}{\partial u} = \frac{\theta_1\theta_2}{1 + \theta_2 u} - P_s. \tag{20}$$

The second derivative is

$$\frac{\partial^2 \mathrm{Pro}_{\mathrm{SU}}}{\partial u^2} = -\frac{\theta_1\theta_2^2}{(1 + \theta_2 u)^2}. \tag{21}$$

Obviously, $\partial^2 \mathrm{Pro}_{\mathrm{SU}}/\partial u^2 < 0$ when $\theta_1 > 0$ and $\theta_2 > 0$. Therefore, the subgame of the stage has the optimal demand u^*. Let $\partial Pro_{SU}/\partial u = 0$.

$$\frac{\theta_1\theta_2}{1 + \theta_2 u} - P_s = 0. \tag{22}$$

We can get the optimal demand u^*:

$$u^* = \frac{\theta_1}{P_s} - \frac{1}{\theta_2}. \tag{23}$$

Given u^*, the service provider sets the optimal service price of per data utility P_s^* as response. We use u^* to replace u in (11). We can acquire

$$\mathrm{Pro}_{\mathrm{SP}}(u^*, P_s) = (P_s - P_u)\left(\frac{\theta_1}{P_s} - \frac{1}{\theta_2}\right). \tag{24}$$

Then, we calculate the partial derivative of (24) with respect to P_s.

$$\frac{\partial \mathrm{Pro}_{\mathrm{SP}}}{\partial P_s} = \frac{\theta_1 P_u}{P_s^2} - \frac{1}{\theta_2}. \tag{25}$$

The second derivative of (24) for P_s is, namely,

$$\frac{\partial^2 \mathrm{Pro}_{\mathrm{SP}}}{\partial {P_s}^2} = -\frac{\theta_1 P_u}{P_s^3}. \tag{26}$$

Clearly, $\partial^2 \mathrm{Pro}_{\mathrm{SP}}/\partial {P_s}^2 < 0$ when $\theta_1 > 0$. Therefore, there is an optimal solution to the subgame. We can get P_s^* through calculating $\partial \mathrm{Pro}_{\mathrm{SP}}/\partial P_s = 0$:

$$P_s^* = \sqrt{\theta_1\theta_2 P_u}. \tag{27}$$

Given u^* and P_s^*, the data provider sets the optimal data price P_u^* as the optimal response.

We take the u^* and P_s^* into (10). Then, we can get

$$\mathrm{Pro}_{\mathrm{DP}}(u^*, P_u) = (P_u - C_{\mathrm{DP}})\left(\frac{\theta_1}{\sqrt{\theta_1\theta_2 P_u}} - \frac{1}{\theta_2}\right). \tag{28}$$

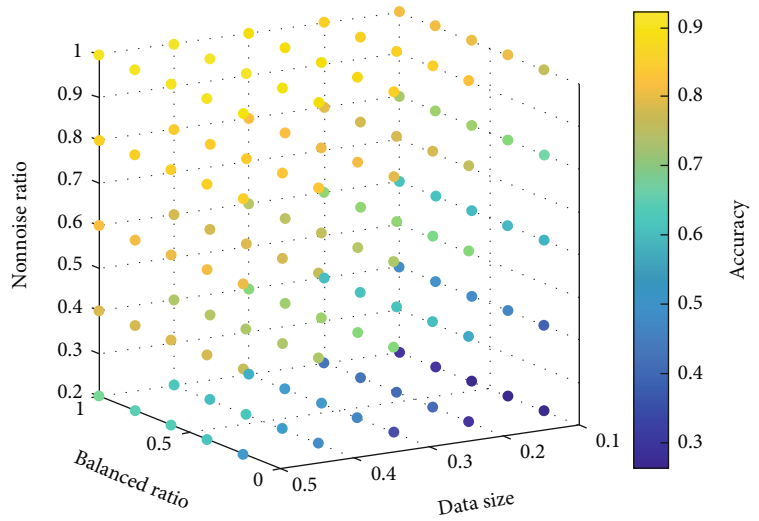

Figure 8: Actual accuracy of the model with different data size n, nonnoise ratio α, and class-balanced ratio β.

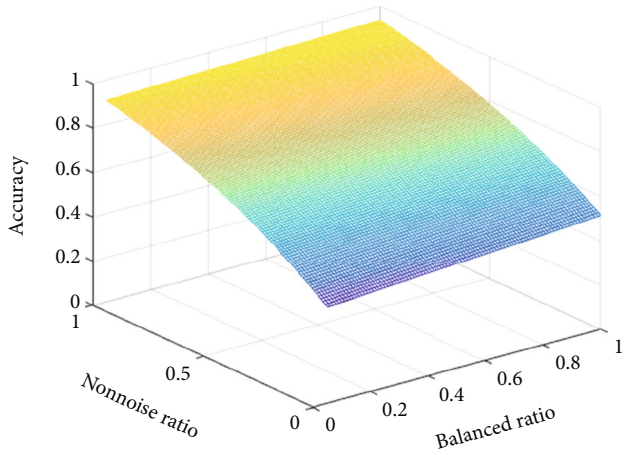

Figure 9: Data utility evaluation function with fixed $n = 0.5$.

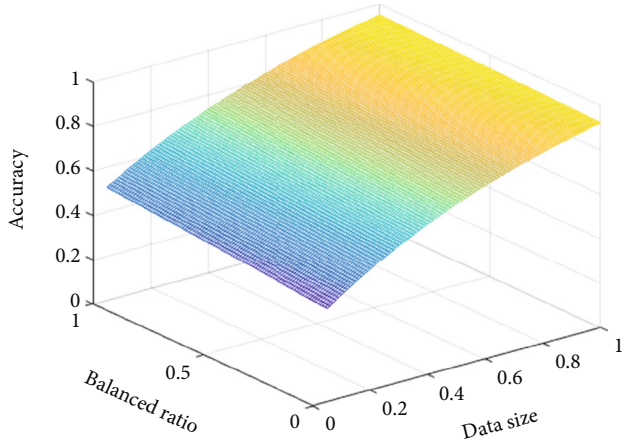

Figure 10: Data utility evaluation function with fixed $\alpha = 0.5$.

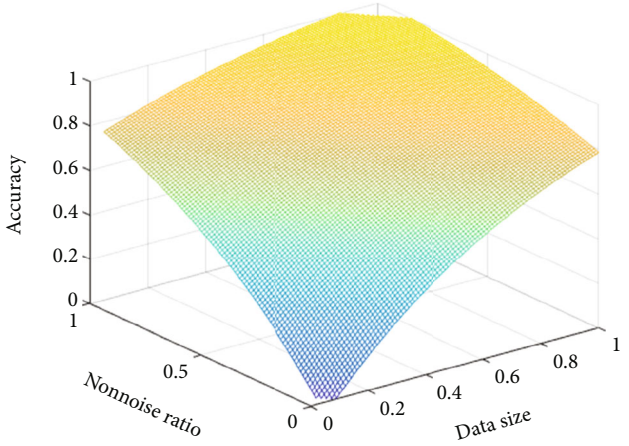

Figure 11: Data utility evaluation function with fixed $\beta = 0.5$.

The first derivative of (28) with respect to P_u is

$$\frac{\partial \mathrm{Pro}_{\mathrm{DP}}}{\partial P_u} = \frac{1}{2} k P_u^{-1/2} + \frac{1}{2} P_u^{-3/2} k C_{\mathrm{DP}} - \frac{1}{\theta_2}. \tag{29}$$

The second derivative of (28) with respect to P_u is

$$\frac{\partial^2 \mathrm{Pro}_{\mathrm{DP}}}{\partial P_u^2} = -\frac{1}{2} k P_u^{-3/2} - \frac{3}{4} P_u^{-5/2} k C_{\mathrm{DP}}, \tag{30}$$

where $k = \theta_1 / \sqrt{\theta_1 \theta_2}$.

There is an optimal data price P_u^* because (30) is negative. Let $\partial \mathrm{Pro}_{\mathrm{DP}} / \partial P_u = 0$.

$$\frac{1}{2} k P_u^{-1/2} + \frac{1}{2} P_u^{-3/2} k C_{\mathrm{DP}} - \frac{1}{\theta_2} = 0. \tag{31}$$

We conclude by observation that (31) is a cubic equation in one unknown about $P_u^{-1/2}$. We can derive the three roots

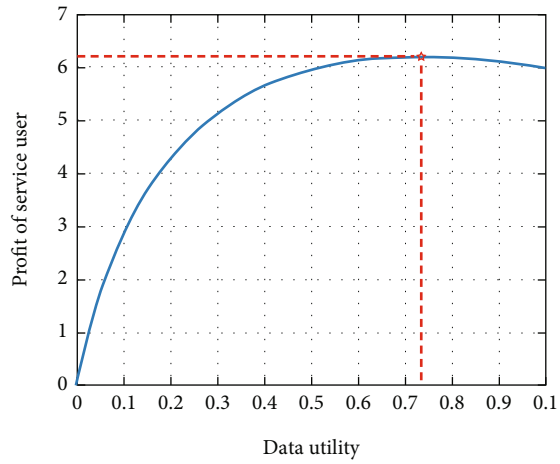

Figure 12: Variation curve of service user's profits with data utility u.

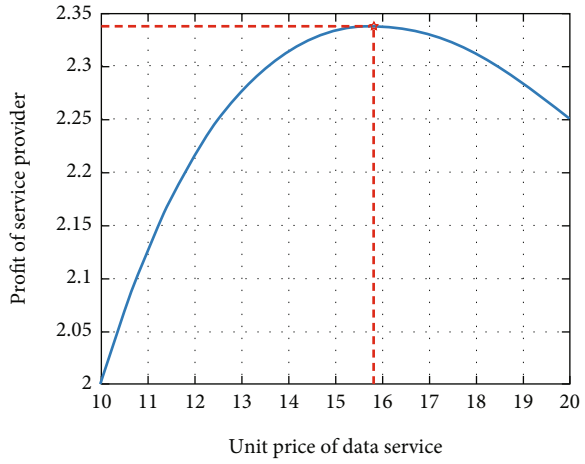

Figure 13: Variation curve of service provider's profits with the price of data service P_s.

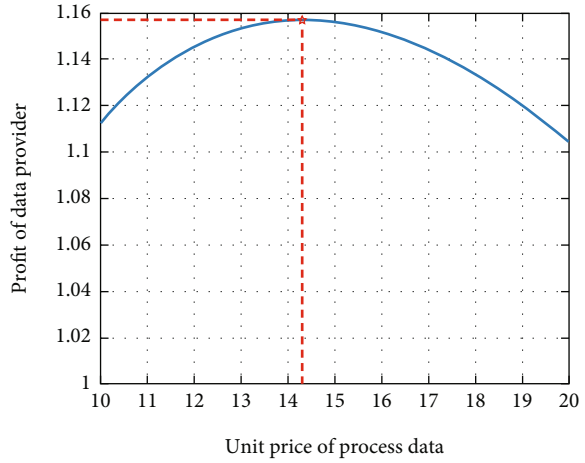

Figure 14: Variation curve of data provider's profits with data price P_u.

of (31) according to *Shengjin's formulas* [49].

$$t_1^* = \frac{-2\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)}{3kC_{\mathrm{DP}}},$$
$$t_{2,3}^* = \frac{\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right) \pm \sqrt{3}\left(\sqrt[3]{Y_1} - \sqrt[3]{Y_2}\right)i}{3kC_{\mathrm{DP}}}, \tag{32}$$

where $Y_{1,2} = 3/8kC_{\mathrm{DP}}(-9kC_{\mathrm{DP}}/\theta_2 \pm \sqrt{81k^2C_{\mathrm{DP}}^2/\theta_2^2 + 3k^4C_{\mathrm{DP}}})$.

We round it off as t_2^* and t_3^* are imaginary roots. Because of $|Y_1| < |Y_2|$, there is

$$\sqrt[3]{Y_1} + \sqrt[3]{Y_2} < 0. \tag{33}$$

We can get

$$P_u^{*-1/2} = \frac{-2\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)}{3kC_{\mathrm{DP}}}. \tag{34}$$

According to (34),

$$P_u^* = \frac{9k^2C_{\mathrm{DP}}^2}{4\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)^2}. \tag{35}$$

Finally, we replace P_u in (23) and (27) with (35).

$$u^* = \frac{2\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)}{3C_{\mathrm{DP}}} - \frac{1}{\theta_2},$$
$$P_s^* = \frac{3\theta_1 C_{\mathrm{DP}}}{2\left|\left(\sqrt[3]{Y_1} + \sqrt[3]{Y_2}\right)\right|}. \tag{36}$$

The theorem is proven. □

4. Experiment

The experiment is divided into two parts. First, we design experiments to prove the rationality and validity of the MDDUE, and then, we prove the existence and uniqueness of the Stackelberg game Nash equilibrium through numerical experiments.

4.1. Parameter Fitting Based on Cifar10. To verify the rationality and validity of the MDDUE, the public dataset in our experiments is Cifar10 [50]. The Cifar10 dataset consists of 60,000 images in 10 classes. It has 6000 images per class. The machine learning model can identify which class an image is. The machine learning model in our experiment is Wide ResNet (WRN) [51]. The WRN is a variant of the ResNet network, with a dropout layer added between the two convolutional layers, increasing the width of the ResNet and improving the training speed.

To argue that changes in the class-balanced ratio have an impact on the accuracy of the machine learning model, we constructed datasets with different class-balanced ratios based on Cifar10. However, the traditional approach to constructing an unbalanced dataset in literature [35] would

change the data size which is a factor that affects the accuracy of the model. We assume that the number of samples in each class is N. To meet the requirements of the rest of the conditions being constant, we take $n_{\min}$ samples in the first five classes and $n_{\max}$ samples in the remaining classes as the training set.

$$n_{\min} = \frac{N}{2} - \frac{N}{2} * \left(\frac{1-\beta}{1+\beta}\right),$$
$$n_{\max} = \frac{N}{2} + \frac{N}{2} * \left(\frac{1-\beta}{1+\beta}\right), \tag{37}$$

where β is the class-balanced ratio.

Figure 4 shows the training curves of different data class-balanced ratios under data size $n = 0.5$ and data nonnoise ratio $\alpha = 1$. It is shown that the smaller the class-balanced ratio, the lower the accuracy of the model. So the introduction of the class-balanced ratio will make the data utility evaluation function more reasonable.

For simplicity of calculation, we normalize the data size to a value in the interval 0-1. Figure 5 shows the model accuracy for different data sizes with the other two dimensions fixed. As we can see, as the data size increases, the accuracy of the machine learning model increases. However, as the data size increases to a certain level, the growth rate of the model accuracy then becomes smaller. Similarly, Figures 6 and 7 show the variation in the accuracy of the model for different class-balanced ratios and different nonnoise ratios, respectively, for the same two other dimensions. All three fitting functions are closer to the actual results.

As shown in Figure 8, the actual accuracy of the model with different data size n, nonnoise ratio α, and class-balanced ratio β is displayed. The average accuracy is the average value of accuracies under 10 experiments.

Using the above actual accuracy data, we can determine the optimal parameters of the proposed data utility evaluation function on the WRN-34 and Cifar10 by minimizing the mean-squared error. To verify the correctness of the function and to better demonstrate the data utility function, we have selected a few special cuts of the function image as in Figures 9–11.

By comparing Figure 8 with Figures 9–11, it can be seen that the data utility function is closer to the actual value. At this point, the values of $\omega_1, \omega_2, \omega_3, \omega_4$, and ω_5 are 1.1262, 1.3484, 1.3121, 1.1592, and 0.1343, respectively. We will use the data utility function shown below in the following numerical experiments.

$$U = 1.1262 - 1.3484 * e^{(-(1.3121n+1.1592\alpha+0.1343\beta))}. \tag{38}$$

4.2. Numerical Experiments. We conduct numerical experiments to demonstrate the existence and uniqueness of the Nash equilibrium. We set the parameters of the service user's reward function $\theta_1 = 5$ and $\theta_2 = 10$, and we set the subscription price of data service $P_s = 6$ and the price per unit data utility $P_u = 5$. As shown in Figure 12, we can see that the profits of the servicer user increase with the increase of data utility. However, when the data utility increases to a certain degree, the profits of the service user reduce because of the increase of subscription price. We can obtain that there is an optimal data utility that maximizes the profits of the service user.

Figure 13 shows that the profits of the service provider increase with the rise of the subscription price. It is clear to see that the profits rise gradually with the increase of P_s. But the data utility demand reduces due to higher prices leading to the reduced profits of the service provider. So there is an optimal price strategy to maximize the service provider's profits.

We set the fixed cost per data utility $C_{DP} = 1$. We know from Figure 14 that as P_u increases, the profits of the data provider increases. However, the profits of the data provider start to decrease when it reaches a certain value because of the fact that the increase in P_u led to an increase in P_s. It leads to a reduction in the demand of the data utility. Therefore, the optimal profits of the data provider can be acquired if the optimal price P_u^* is applied.

5. Conclusions

First, we propose MDDUE to evaluate the utility of data. Then, we advance an optimal data-pricing scheme based on the Stackelberg game in this paper. Specifically, we construct a data utility evaluation function through three data quality dimensions. We are the first to introduce the class-balanced ratio into the data utility evaluation function to make it more accurate and more reasonable. Then, we propose an optimal data-pricing scheme based on the three-stage Stackelberg game. The profits of the three roles in the data trading market can be maximized by using the scheme. Finally, we verify the rationality and validity of MDDUE through a specific machine learning model WRN and a real-world dataset Cifar10. Meanwhile, we prove the existence and uniqueness of the Nash equilibrium, then demonstrate results through numerical experiments. In the future work, we will improve the universality of MDDUE. And we will be working on building fairer and more secure data trading solutions using technologies such as cryptography and blockchain.

Data Availability

The Cifar10 dataset is found at http://www.cs.utoronto.ca/~kriz/cifar.html.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research is funded by the National Natural Science Foundation of China (62202118, 61962009), Top Technology Talent Project from Guizhou Education Department (Qianjiao Ji [2022]073), and Foundation of Guangxi Key

Laboratory of Cryptography and Information Security (GCIS202118).

References

[1] X. Xu, Z. Fang, J. Zhang et al., "Edge content caching with deep spatiotemporal residual network for IoV in smart city," *ACM Transactions on Sensor Networks*, vol. 17, pp. 1–33, 2021.

[2] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *Enabling Real-Time Business Intelligence,vol. 154*, pp. 129–144, Springer, Berlin Heidelberg, 2013.

[3] Y. Wang, T. Li, M. Liu, C. Li, and H. Wang, "STSIIML: study on token shuffling under incomplete information based on machine learning," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 11078–11100, 2022.

[4] X. Zhou, W. Liang, K. I. Kai, R. H. Wang, and Q. Jin, "Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 1, pp. 246–257, 2021.

[5] Y. Chen, J. Sun, Y. Yang, T. Li, X. Niu, and H. Zhou, "PSSPR: a source location privacy protection scheme based on sector phantom routing in WSNs," *International Journal of Intelligent Systems*, vol. 37, no. 2, pp. 1204–1221, 2022.

[6] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, and K. I.-K. Wang, "Hierarchical adversarial attacks against graph-neural-network-based IoT network intrusion detection system," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9310–9319, 2022.

[7] W. Dai, C. Dai, K.-K. R. Choo, C. Cui, D. Zou, and H. Jin, "SDTE: a secure blockchain-based data trading ecosystem," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 725–737, 2020.

[8] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[9] X. Cao, Y. Chen, and K. J. Ray Liu, "Data trading with multiple owners, collectors, and users: an iterative auction mechanism," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 268–281, 2017.

[10] L. Kong, L. Wang, W. Gong, C. Yan, Y. Duan, and L. Qi, "LSH-aware multitype health data prediction with privacy preservation in edge environment," *World Wide Web*, vol. 25, no. 5, pp. 1793–1808, 2022.

[11] T. Jung, X.-Y. Li, W. Huang et al., "AccountTrade: accountability against dishonest big data buyers and sellers," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 223–234, 2019.

[12] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[13] E. Xinhua, J. Han, Y. Wang, and L. Liu, "Big data-as-a-service: definition and architecture," in *2013 15th IEEE International Conference on Communication Technology*, pp. 738–742, Guilin, China, 2013.

[14] Z. Xiao, D. He, and D. Jiayi, "A Stackelberg game pricing through balancing trilateral profits in big data market," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12658–12668, 2021.

[15] L. Qi, W. Lin, X. Zhang, W. Dou, X. Xu, and J. Chen, "A correlation graph based approach for personalized and compatible web APIs recommendation in mobile app development," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[16] Y. Suresh, J. Senthilkumar, K. Saraswathi, and V. Mohanraj, "Deep learning enabled social media recommendation based on user comments," *Computer Systems Science and Engineering*, vol. 44, no. 2, pp. 1691–1702, 2023.

[17] X. Zhou, W. Liang, K. I.-K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 171–178, 2021.

[18] D. Niyato, M. A. Alsheikh, P. Wang, D. In Kim, and Z. Han, "Market model and optimal pricing scheme of big data and Internet of Things (IoT)," in *IEEE International Conference on Communications (ICC)*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[19] J. Yang, C. Zhao, and C. Xing, "Big data market optimization pricing model based on data quality," *Complexity*, vol. 2019, Article ID 5964068, 10 pages, 2019.

[20] X. Xiaolong, Q. Jiang, P. Zhang et al., "Game theory for distributed IoV task offloading with fuzzy neural network in edge computing," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 11, pp. 4593–4604, 2022.

[21] F. Liang, Y. Wei, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018.

[22] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyberphysical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[23] T. Li, Y. Chen, Y. Wang et al., "Rational protocols and attacks in blockchain system," *Security and Communication Networks*, vol. 2020, Article ID 8839047, 11 pages, 2020.

[24] T. Li, Z. Wang, G. Yang, Y. Cui, Y. Chen, and X. Yu, "Semi-selfish mining based on hidden Markov decision process," *International Journal of Intelligent Systems*, vol. 36, no. 7, pp. 3596–3612, 2021.

[25] T. Li, Z. Wang, Y. Chen, C. Li, Y. Jia, and Y. Yang, "Is semi-selfish mining available without being detected?," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 10576–10597, 2022.

[26] Y. Zuo, Z. Kang, X. Jian, and Z. Chen, "BCAS: a blockchain-based ciphertext-policy attribute-based encryption scheme for cloud data security sharing," *International Journal of Distributed Sensor Networks*, vol. 17, no. 3, 2021.

[27] Y. Chen, S. Dong, T. Li, Y. Wang, and H. Zhou, "Dynamic multi-key FHE in asymmetric key setting from LWE," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 5239–5249, 2021.

[28] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *Journal of the ACM*, vol. 62, no. 5, pp. 1–44, 2015.

[29] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang, "A pricing model for big personal data," *Tsinghua Science and Technology*, vol. 21, no. 5, pp. 482–490, 2016.

[30] X. Li, J. Yao, L. Xue, and H. Guan, "A first look at information entropy-based data pricing," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2053–2060, Atlanta, GA, USA, 2017.

[31] H. Cai, F. Ye, Y. Yang, Y. Zhu, and J. Li, "Towards privacy-preserving data trading for web browsing history," in *2019*

IEEE/ACM 27th International Symposium on Quality of Service (IWQoS), pp. 1–10, Phoenix, AZ, USA, 2019.

[32] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 144–153, Dallas, TX, USA, 2019.

[33] X. Zhipeng Cai, J. W. Zheng, and Z. He, "Private data trading towards range counting queries in Internet of Things," *IEEE Transactions on Mobile Computing*, 2022.

[34] F. Stahl and G. Vossen, "Fair knapsack pricing for data marketplaces," in *Advances in Databases and Information Systems, vol. 9809*, pp. 46–59, Springer International Publishing, Cham, Switzerland, 2016.

[35] Y. Cui, M. Jia, T.-Y. Lin, S. Yang, and S. Belongie, "Class-balanced loss based on effective number of samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, Long Beach, CA, USA, 2019.

[36] W. Liang, H. Yiyong, X. Zhou, Y. Pan, and K. I.-K. Wang, "Variational few-shot learning for microservice-oriented intrusion detection in distributed industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5087–5095, 2022.

[37] Y. Chen, X. Yang, T. Li, Y. Ren, and Y. Long, "A blockchain-empowered authentication scheme for worm detection in wireless sensor network," *Digital Communications and Networks*, 2022.

[38] M. A. Alsheikh, D. T. Hoang, D. Niyato, D. Leong, P. Wang, and Z. Han, "Optimal pricing of Internet of Things: a machine learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 4, pp. 669–684, 2020.

[39] G. Algan and I. Ulusoy, "Meta soft label generation for noisy labels," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7142–7148, Milan, Italy, 2021.

[40] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[41] X. Zhou, X. Xuesong, W. Liang, Z. Zeng, and Z. Yan, "Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart IoT," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12588–12596, 2021.

[42] H. Jianwen, Y. Chen, X. Ren, Y. Yang, X. Qian, and X. Yu, "Blockchain-enhanced fair and efficient energy trading in industrial Internet of Things," *Mobile Information Systems*, vol. 2021, Article ID 7397926, 13 pages, 2021.

[43] X. Zhou, W. Liang, K. I.-K. Wang, and S. Shimizu, "Multi-modality behavioral influence analysis for personalized recommendations in health social media environment," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 888–897, 2019.

[44] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowledge and Information Systems*, vol. 64, no. 6, pp. 1417–1455, 2022.

[45] Z. Yang, Z. Xiong, D. Niyato, P. Wang, and J. Jin, "Joint optimization of information trading in Internet of Things (IoT) market with externalities," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Barcelona, Spain, 2018.

[46] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, "Smart data pricing models for the Internet of Things: a bundling strategy approach," *IEEE Network*, vol. 30, no. 2, pp. 18–25, 2016.

[47] K. Liu, X. Qiu, W. Chen, X. Chen, and Z. Zheng, "Optimal pricing mechanism for data market in blockchain-enhanced Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9748–9761, 2019.

[48] C. Xu, K. Zhu, C. Yi, and R. Wang, "Data pricing for blockchain-based car sharing: a Stackelberg game approach," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–5, Taipei, Taiwan, 2020.

[49] S. Fan, "A new extracting formula and a new distinguishing means on the one variable cubic equation," *Natural Science Journal of Hainan Teachers College*, vol. 2, no. 2, pp. 91–98, 1989.

[50] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.

[51] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12, York, UK, 2016.