Hindawi
Complexity
Volume 2019, Article ID 5964068, 10 pages
https://doi.org/10.1155/2019/5964068

Research Article

Big Data Market Optimization Pricing Model Based on Data Quality

Jian Yang,[1] **Chongchong Zhao**,[1] **and Chunxiao Xing**[2]

[1]*School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China*
[2]*Research Institute of Information, Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Jian Yang; yangjian2015@xs.ustb.edu.cn

Received 3 February 2019; Accepted 7 April 2019; Published 23 April 2019

Guest Editor: Thiago Christiano Silva

In recent years, data has become a special kind of information commodity and promoted the development of information commodity economy through distribution. With the development of big data, the data market emerged and provided convenience for data transactions. However, the issues of optimal pricing and data quality allocation in the big data market have not been fully studied yet. In this paper, we proposed a big data market pricing model based on data quality. We first analyzed the dimensional indicators that affect data quality, and a linear evaluation model was established. Then, from the perspective of data science, we analyzed the impact of quality level on big data analysis (i.e., machine learning algorithms) and defined the utility function of data quality. The experimental results in real data sets have shown the applicability of the proposed quality utility function. In addition, we formulated the profit maximization problem and gave theoretical analysis. Finally, the data market can maximize profits through the proposed model illustrated with numerical examples.

1. Introduction

With the rapid development of information technology, big data has become the core resource of all walks of life. Government departments, research institutions, IT companies, financial institutions, etc. have generated massive amounts of data during operations. In addition, due to the rise of mobile networks and smart terminals, a large proportion of people now have smart phones with sensors, which can easily collect data beyond the past possible range using GPS, cameras, microphones, etc. The storage and calculation of big data are no longer the sole purpose. By using data mining and machine learning to analyze data, it provides an opportunity to bring about breakthroughs in processing video, images, and speech [1]. Unfortunately, only a small amount of data is currently being fully utilized and its use is limited as well. The reuse of these data can create huge commercial value, which is the true meaning of big data. Therefore, in order to make profits and provide data utilization, data can be resold to other organizations [2].

Marketplaces are enablers for the exchange of data. Therefore, data trading has become an innovative business model that has driven the advent of DT (Data Technology) era. In this era, data has become an important asset for companies, from the exclusive internal data to the sharing between companies. However, due to the lack of standardized data sharing channels and unified transaction specifications, big data trading platforms and data markets have emerged as the times require in this context.

Nowadays, data products and related services are increasingly being provided to the online data market, which carries the data publisher's data and provides it to data consumers. Figure 1 [3] presents an intuitive description of the formation and flow of data products. Firstly, the initial seller is the original data provider. For example, Xignite [4] sells financial data, Gnip [5] publishes social network data, and Factual [6] deals with geographical data. Secondly, the data market provides a centralized management platform for data providers to upload, store, and sell data in order to support online transactions of the data. The current

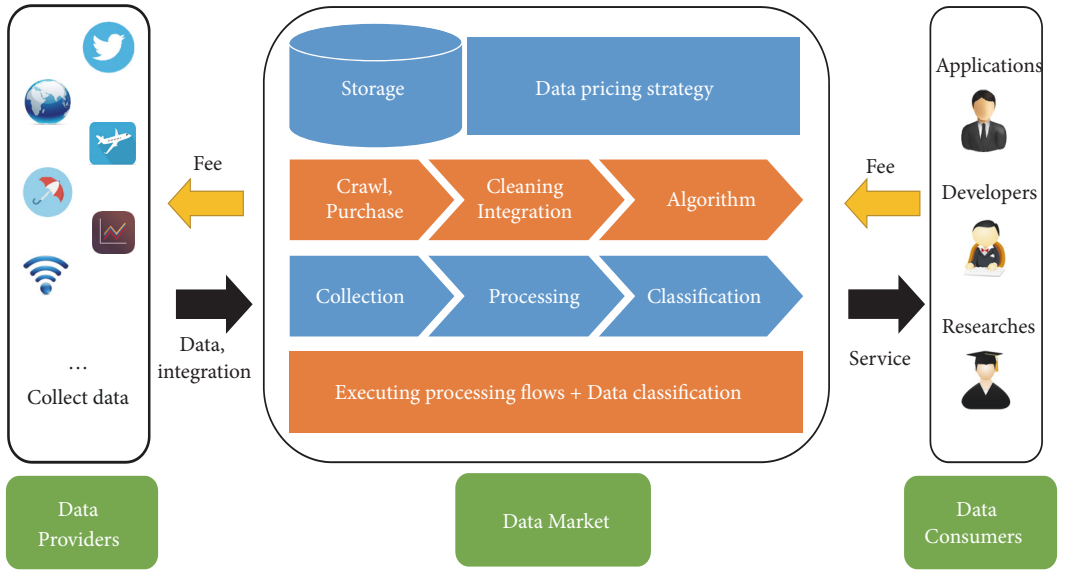

Figure 1: A typical big data market model.

data platforms include Factual, Infochimps [7], Xignite, and Windows Azure Data Marketplace [8]. The last is the terminal buyer, which is the consumer of big data products. There are generally three types of consumers who have demand for big data products, i.e., enterprises, government departments, and research institutions. These consumers need the data and corresponding services provided by the online market in order to innovate products, optimize decisions, or conduct research.

However, the big data market has not formed a unified pricing mechanism yet, and various pricing strategies are still not perfect; i.e., different data markets offer different pricing mechanisms. Currently, the major pricing mechanisms in the data market include subscription, bundling, and discrimination. However, the impact of data quality on the pricing mechanism has rarely been studied. Many literatures [9, 10] indicate that data quality is very important for the evaluation of data value. Hence, in this paper, we have proposed a pricing model based on quality utility to optimize data market pricing.

The key contributions of this paper can be summarized as follows:

(i) We first summarized several dimensional indicators that affected data quality and established a linear model to calculate the quality scores. Based on this, a hierarchical division method of the square root of the quality score is proposed.

(ii) We proposed a utility model based on the quality level and verified it with real-world datasets, using machine learning algorithms. The results have proved the applicability of this utility model.

(iii) From the perspective of economics, we considered the consumers' willingness to pay and formulated an optimized pricing scheme based on the quality utility function. Numerical experiments have shown that the owners of data platform can maximize profits by determining the quality level and subscription fee.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Then, the dimension of data quality and the method of level division are presented in Section 3. Section 4 describes the utility function of data quality and the suitability of the model is verified by machine learning algorithm. Section 5 formulates the profit maximization problem and gives theoretical analysis. Section 6 presents and analyzes the numerical experimental results. Section 7 concludes the paper.

2. Literature Review

The valuation of intangible assets, such as cloud computing services [11–13] and network information services [14, 15], is not a new challenge for practitioners and researchers. Relevant scholars have done a lot of work on the pricing of information products and services.

Before studying data pricing, we first review the representative work of these methods. The information service market usually involves three commonly used pricing mechanisms:

(1) Subscription mechanism: Windows Azure Data Market [8] is a decent example for subscription pricing scheme. Azure has monthly subscriptions of two types, limited and unlimited. Balasubramanian et al. [16] consider the difference between usage frequency associated with payment model and consumer psychological cost. They believe that the two pricing mechanisms of information products, i.e., fixed cost and pay-per-payment mechanisms, may affect the profit of the seller.

(2) Bundling pricing: this strategy originates from capital data market, and it represents an aggregation technique [17]. In the capital data market, data vendors

bundle multiple types of products in accordance with certain strategies and allocate different prices for them to be selected by heterogeneous consumers. Niyato et al. [18], considering smart data pricing methods to solve the problem of IoT data management, adopt a binding strategy that allows multiple providers to form alliances and provide bundled services to attract more users and achieve higher revenues.

(3) Version control pricing mechanism: the strategy is a widespread differentiation strategy used in information-product markets. Wei et al. [19] inspect the versioning strategy where consumers differ in individual tastes for quality. They found that if groups have mutually exclusive characteristics, they are the values associated with the shared features; then, versioning strategy is optimal. Li et al. [20] defined a nonlinear function to describe the "willingness to pay" and the utility of consumers with specific quality requirements and developed a hybrid steady state evolution algorithm. They observed that monopolies can obtain more profits by using multiversion strategy.

There are also some scholars who have studied the pricing model of data products from different perspectives. Koutris et al. [21] studied query-based pricing, and they designed a pricing algorithm that satisfies no-arbitrage and no-discount allowing the price of any query to be exported automatically. Shen et al. [22] proposed a big personal data pricing model based on tuple granularity. By investigating the data attributes that affect the value of data, this model is proposed to implement a positive rating and reverse pricing for big personal data. By dynamically adjusting the model parameters, the users can enjoy improved benefits. Yang et al. [23] studied the pricing model of personal privacy data. They proposed a framework to compensate for privacy loss. This method can compensate for privacy loss based on user's preferences and allow users to control their data through financial means.

Through extensive review of the literature, we can conclude that existing data pricing literature either investigates published data pricing methods or studies new approaches that focus on relevance and privacy. Data quality is a key factor affecting data assessment and has been ignored so far.

In the entire data life cycle, such as data creation, transformation, transmission, and application, each stage may cause various data quality problems. Liu et al. [24] summarized the problems faced by current big data research in data collection, processing, and analysis, namely, the collection of unreal data, information incompleteness, consistency, and reliability. In [25], there are a total of 21 quality standards. Ding et al. [26] summarize relevant quality dimensions and review their applicability to the data market.

Data quality is characterized by multidimensionality and complexity. Therefore, in this paper, we consider an optimized pricing model based on data quality, hoping to provide data platform owners with useful pricing decision recommendations.

3. Data Value Evaluation Based on Quality

When the data market owner wants to sell data at a reasonable price, the first thing to consider is to evaluate the value of data. On the one hand, data value can be measured by the size of data [27], on the other hand, it can be measured based on the quality of data. In this paper, we evaluate data from the perspective of data quality. First, we introduce different dimensions of data quality. Then, we establish a linear model based on these dimensions to evaluate data value. Finally, we adopt the square root mapping function and divide the quality level.

3.1. Dimensions of Data Quality. Data quality includes multiple dimensions. The measurement of dimensions will vary according to the type of data, so quality has to be evaluated using the criteria that the data has to comply with. In [28, 29], the applicability of the quality dimension to the data market has been reviewed, especially the concept of version control, i.e., the data seller creates different quality versions of the data product to suit the needs and tastes of heterogeneous consumers. Literature [30] summarizes seven quality standards, which were expressed as Q_d = {*accuracy*, *completeness, redundancy, data volume, latency, response time, timeliness*}. These quality dimensions allow continuous versioning. In other words, we can create any number of quality levels based on them. For simplicity, only three measures in Q_d are all scaled in interval [0, 1] and will be demonstrated here in detail, i.e., *accuracy*, *completeness,* and *redundancy*. Table 1 [31] contains the metrics, report names, and description definitions for each quality attribute and lists the calculation formulas.

Several other quality dimensions also have their calculation methods. However, due to space limit, we omit them from the paper.

3.2. Data Quality Level Division. Creating a universal data quality assessment standard can be an arduous task for all types of data. Without loss of generality, a linear model is presented as below, but other options may exist.

$$\begin{aligned} Qualityscore &= w_1 * accuracy + w_2 * completeness \\ &\quad + \cdots + w_n * redundancy \\ s.t. &\quad w_1 + w_2 + \cdots + w_n = 1 \end{aligned} \tag{1}$$

where $w_1, w_2, \ldots, w_n$ are related weight factors, which can be set by users in practice.

We adopt the method of dividing the quality level in [30]. In this paper, the quality score is defined in interval (0,1), and we first scale it to the sector of the appropriate function domain $[s_{min}, s_{max}]$, e.g., $[0, 100]$. Then, since the square root function can produce more reasonable level intervals, we adopt the square root of the quality score to rank on the basis of the previous step. For instance, a domain of $[0, 100]$ and quality levels of $Q_l = 10$, as done in this paper,

Table 1: Metric definitions, description, and calculation.

Attributes	Metric	Description	Variables	Formula
Accuracy	Proportion of accurate cells	Indicate the proportion cells in a data source that has correct value according to the domain and the type of information of the data source.	*nce*: Number of cells with errors *ncl*: Number of cells	$pac=1-\frac{nce}{ncl}$
Completeness	Proportion of complete cells	Indicate the proportion of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e., a value coherent with the domain of the column).	*nr*: Number of rows *nc*: Number of columns *ic*: Number of incomplete cells *ncl*: Number of cells	$ncl = nr * nc$ $pcc=1-\frac{ic}{ncl}$
Redundancy	Proportion of duplicate records	Redundancy expresses the proportion of duplicate records in the data source. Since this factor is the cost-indicator, we convert it to the benefit-indicator.	*nr*: Number of rows *red*: Number of duplicate records	$pdc=1-\frac{red}{nr}$

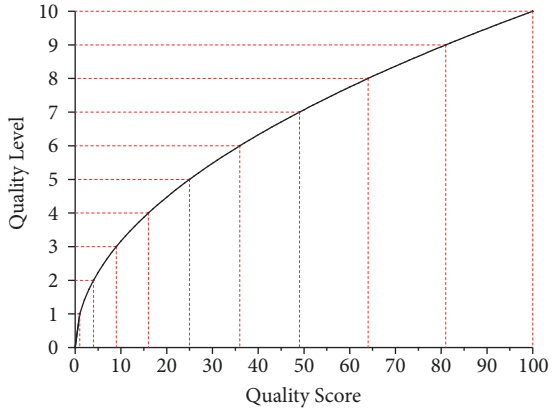

Figure 2: Mapping of quality scores and levels.

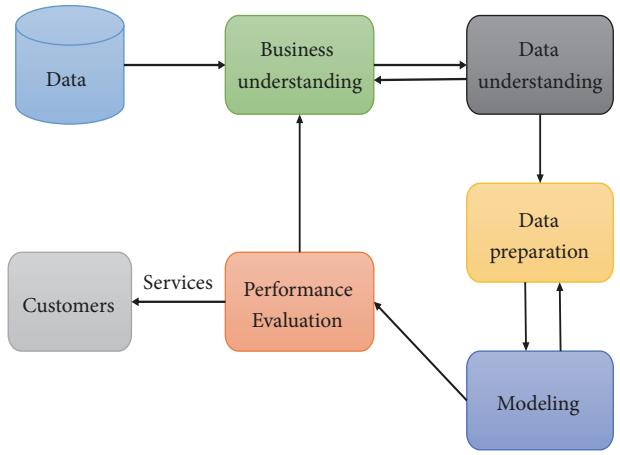

Figure 3: Big data business intelligence service.

is that examples are more illustrative. Figure 2 presents such a mapping relationship.

4. The Utility of Data Quality

4.1. Utility Functions. In current big data business applications, it is usually big data sets that adopt model-based methods to extract knowledge and information to solve complex business applications. Figure 3 shows the process of big data business intelligence.

It can be seen that data plays an important role in the entire business analysis. The quality of data directly determines the accuracy of the machine learning model [32] and ultimately affects business decisions. In order to explain this certain phenomenon, the usefulness of quality must be measured on a new scale. Therefore, the usefulness of quality is $U(q)$, which is the utility of quality. According to the experience of machine learning and data mining, under the condition of the same amount of data, the higher quality information is input into the classifier, the better the classification effect will be. Therefore, this utility function can be considered as the quality of the model. For example, the utility is the accuracy of classifying input into a discrete-value output.

We suppose that a utility function $U(q)$ has the following three basic properties:

(1) $U(q)$ is nonnegative.

(2) $U(q)$ is an increasing function of q.

(3) $U(q)$ is a concave function of q.

Usually we assume that the function $U(q)$ is nonnegative and twice differentiable; then, (2) and (3) state that $U'(q) > 0$ and $U''(q) < 0$.

The first attribute is rational as quality utility cannot be negative. The second attribute is the obvious requirement that the higher the quality, the better. Several reasons are given for the third property. One way to justify it is to require that the marginal utility $U(q)$ is a decreasing function [33] of data quality q.

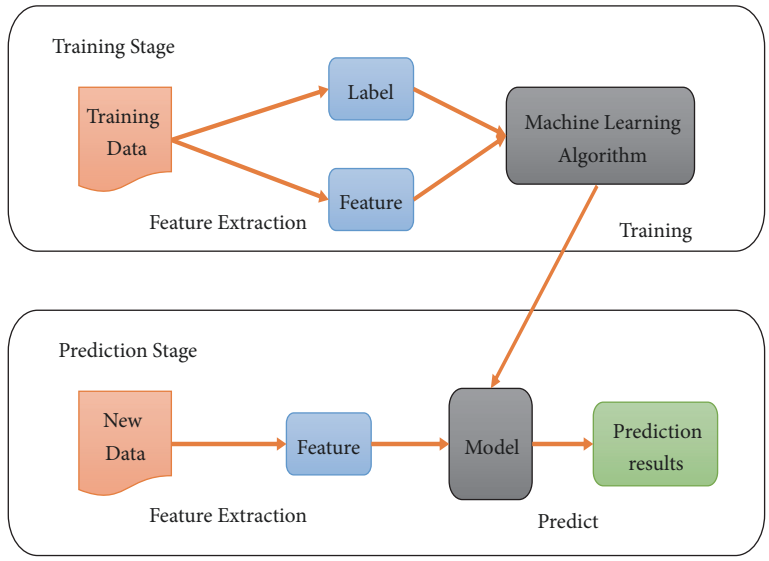

Figure 4: A basic machine learning workflow.

4.2. Estimating Utility Functions. In order to determine the utility function of data quality in big data analysis, we consider the study from the perspective of classification-based machine learning.

Next, we describe the process of classification. For a data set D, it can be expressed as a $k \times l$ matrix $\overline{X}$, where each row corresponds to an item, and the first l elements of each row correspond to the l property values of the item. A machine learning task can be divided into two phases, as shown in Figure 4.

(i) Training Stage. The data is subjected to feature extraction to generate data features and prediction targets(*Label*) and then trained by machine learning algorithms to generate the model.

(ii) Predicting Stage. Input testing dataset: after feature extraction, produce data features, using the trained model to make predictions and finally producing prediction results.

As shown in (2). A is a label column, which is the real category attribute. The last column is the category predicted by the classifier, denoted by Pr. Evaluating a classifier can be judged by minimizing the error between A and Pr, i.e., $\min \sum_{i=1}^{k} \|A - Pr\|^2$.

$$\overline{X} = \begin{array}{c} \\ m_1 \\ m_2 \\ \vdots \\ m_k \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & \cdots & X_l & A & Pr \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1l} & \check{y}_1 & \hat{y}_1 \\ x_{21} & x_{22} & \cdots & x_{2l} & \check{y}_2 & \hat{y}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kl} & \check{y}_k & \hat{y}_k \end{pmatrix} \end{array} \tag{2}$$

Suppose that the classification accuracy for each item m_i is defined as a. But beyond that, we also assume the utility, i.e., accuracy $U = ACC$. To estimate the utility function, we use data of different quality levels during the model training. Specifically, the experimental point $(q_{s1}, a_1), \ldots, (q_{sj}, a_j), \ldots, (q_{sk}, a_k)$ is a nondecreasing sequence, where q_{sj} is the quality level of the corresponding data and satisfies $q_{sj} \leqslant q_{s(j+1)}$. These points are then used to find a set of optimal parameters of the utility function $U(q_s; \beta)$ by nonlinear least squares, where β is an optimal parameter. The optimal parameter of the utility function $U(q_s; \beta)$ by minimizing the sum of square errors is as follows:

$$\min \sum_{j=1}^{k} \left\| ACC_j - U\left(q_{sj}; \beta\right) \right\|^2 \tag{3}$$

In this paper, for simplicity, we consider the following exponential-based utility function:

$$\begin{aligned} U\left(q_s; \beta = [\beta_1, \beta_2, \beta_3]\right) &= \beta_1 - \beta_2 \exp\left(\beta_3 q_s\right) \\ s.t. \quad & \beta_1, \beta_2 > 0 \\ & \text{and } \beta_3 < 0. \end{aligned} \tag{4}$$

where q_s is the quality level and β_1, β_2, and β_3 are the curve fitting parameters of the utility function to real-world experiments, i.e., the ground truth. In order to find a utility function that satisfies the corresponding condition, we can adjust the parameter β, so that the sum of the squared errors between the experimental and the estimated points is minimized.

4.3. Experimental Evaluation Based on Real Datasets. In order to prove the rationality of the proposed utility function, we use a real dataset called MNIST [34], which contains a variety of hand-written digital pictures and contains the labels for each picture. We use Artificial Neural Networks (ANNs) model for classification training. ANNs use nonlinear mathematical equations to successively develop meaningful relationships between input and output variables through a learning process. Specifically, we applied convolutional neural network (*cnn*) for classification training.

Table 2: Frequently used notations.

Notation	Description
q_s	The quality level of the data products and services
p	Subscription fees for data products and services
WTP	Customers' Willingness to Pay
η	Customer sensitivity to quality level
$\mathcal{G}(q_s, p)$	Profit resulting from the separate sales of the data product and service under p and q_s
$U(q_s, \beta)$	Data utility with curve fitting parameter p and quality level q_s
N	The number of customers willing to buy a data product or service
c	The unit price of the data quality
$\mathcal{L}(\cdot)$	Lagrangian of the profit function $U(q_s, \beta)$

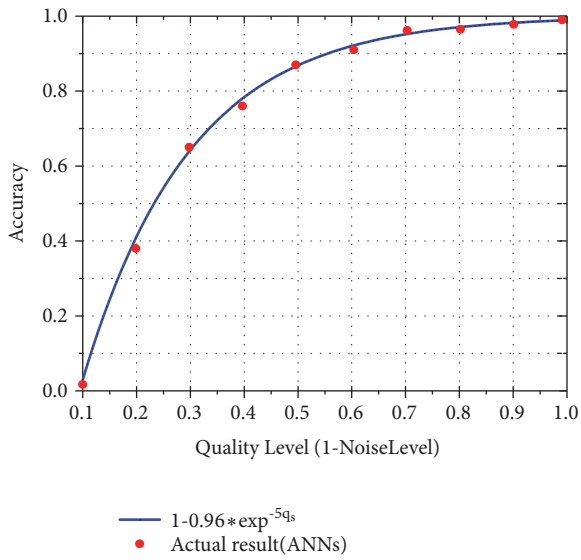

Figure 5: Accuracy trends under different quality levels ($\beta_1 = 1$, $\beta_2 = 0.96$, $\beta_3 = -5$).

Due to the multidimensionality and complexity of data quality, it would be a difficult task if all quality dimensions were taken into account to classify quality levels. Our goal is to illustrate the effect of different quality levels of a given data on model classification capabilities. For simplicity, and in order to reflect our motivation, in the experimental design stage, we draw on the experience of the concept of signal-noise ratio (SNR) [35] in the electronic information field. Specifically, we use the method of adding noise to the label data to express the effect of different quality levels on the accuracy of the model. In the experiment, we assume that the original MNIST training set and testing set labels are all noise-free. Use the following steps to add noise to the label:

(i) Select K samples from N total samples according to the given noise ratio NoiseLevel, $K = N*$ NoiseLevel.

(ii) For each sample of the selected K samples, replace its original label with a random number between 0 and 9 except the original label.

The quality level is the inverse image of the noise level. For simplicity, Figure 5 shows the trend of accuracy at different levels of quality. Obviously, as the quality level increases, the accuracy also increases, and the higher the quality level, the smaller the increase in accuracy. The accuracy of growth is getting smaller and smaller. Further, the proposed utility function can well fit the actual accuracy result and rationalize the concave function. It also facilitates the derivation of optimal pricing, which will be described in the next section.

5. Optimal Pricing

In this section, we first analyzed consumers' willingness to pay from the perspective of consumer behavior. Then, we introduced the profit maximization model with data quality level. Finally, the closed-form solutions of the subscription fee and quality level were derived and proved to be globally optimal. The key notations and description used throughout the paper were defined in Table 2.

5.1. Customers' Willingness to Pay. Every consumer in the market has personal preferences and interests. They make purchasing decisions based on their own needs, preferences, and prices by a self-selection process. This self-selection is described by a consumer's Willingness To Pay (WTP) [36]. WTP refers to the price that a customer is willing to pay in order to purchase a certain number of data products or services. This price is also referred to the customer's reservation price. In other words, they are willing to pay the highest price of the product. We assume that the data platform knows the customer's willingness to pay obeys the probability distribution, and the data platform is faced with a choice dilemma, i.e., loss of customer because of high price or consumer surplus due to low price. Each arriving consumer has a specific subjective price for a certain product, i.e., reservation price, and only if the consumer's reservation price is greater than the value of the product, the customer will purchase it.

We assume customers' sensitivities of quality level by $\eta = \{\eta_1, \eta_2, \ldots, \eta_M\}$, which is randomly distributed from 0 to 1. Note that the higher the quality of data provided, the more the willingness of customers to pay for the data product, which is

$$\frac{\partial WTP}{\partial U(q_s)} > 0 \tag{5}$$

where $U(q_s)$ is the data quality utility function mentioned in Section 4. Customers who would like to obtain the best experience need to pay more. Assume the WTP function is linear, which is

$$WTP = \eta U(q_s) \tag{6}$$

5.2. Profit Function of Data Platform. In Section 1, we described a typical big data market model. The data platform purchases raw data from the data publishers and pays for the data providers. The data platform needs to process, convert, and store the collected data, or to establish application-level services (e.g., business analysis, visualization). This results in fixed costs (purchased from raw data) and variable costs (data processing, deep processing), which are collectively referred to costs in this paper. The data platform can set the subscription fee based on the quality level of the provided data and service to determine its profit maximization. The data consumer decides whether or not to purchase according to their willingness to pay and consumption. We assume that the probability density of all customers' willingness to pay is $f(p)$, and its cumulative distribution function is $F(p)$, which indicates the probability that consumers' willingness to pay is less than the value of products, i.e., the probability that the customer is unwilling to purchase the product. Then, the expected profit of the data platform is computed as follows:

$$\begin{aligned}\mathcal{G}(q_s, p) &= pNPr - cq_s = pN\int_0^{WTP} f(w)\,dw - cq_s \\ &= pN\left(\int_0^{p} f(w)\,dw + \int_p^{WTP} f(w)\,dw\right) \\ &\quad - cq_s = pN(F(p) + WTP - p) - cq_s\end{aligned} \tag{7}$$

where N is the number of potential customers, p is the subscription fee of the data product, q_s is the data quality level, and c is the data cost of the unit quality purchased from the data provider. Profit $\mathcal{G}(\cdot)$ is the difference between subscription revenue and total data cost. The costs of the service (such as calculation cost) are ignored.

In the above equation, $F(p)$ is equal to 0, and substituting (6) into (7) we have

$$\mathcal{G}(q_s, p) = pN(\eta U(q_s) - p) - cq_s \tag{8}$$

where, without loss of generality, we assume that η=1, and (4) and (8) are merged and organized as follows:

$$\mathcal{G}(q_s, p) = pN(\beta_1 - \beta_2 \exp(-\beta_3 q_s) - p) - cq_s \tag{9}$$

The profit maximization problem can be formulated as follows:

$$\begin{aligned}\text{maximize} \quad & \mathcal{G}(q_s, p) \\ s.t. \quad & C_1\colon q_s \geqslant 0; \\ & C_2\colon p \geqslant 0\end{aligned} \tag{10}$$

The goal of (10) is to maximize the profitability of the data platform by jointly optimizing q_s and p. For constraints C_1 and C_2, they ensure nonnegative solutions of q_s and p. Next, we will provide closed-form solutions $(\overline{q_s}, \overline{p})$ to this profit maximization problem and prove their global optimality.

5.3. Optimal Pricing and Quality Level. We use Karush-Kuhn-Tucker (KKT) [37] conditions to optimize the profitability of the data platform. The KKT condition is an important idea for solving Lagrangian duality problem. It is widely used in operations research, convex and nonconvex optimization, machine learning, and other fields. Based on (10), we describe the Lagrangian dual problem as follows:

$$\begin{aligned}\mathcal{L}(q_s, p, \lambda_1, \lambda_2) &= \mathcal{G}(q_s, p) + \lambda_1 q_s + \lambda_2 p \\ s.t. \quad & \lambda_1 \geqslant 0, \\ & \lambda_2 \geqslant 0\end{aligned} \tag{11}$$

where λ_1 and λ_2 are called Lagrange multipliers and they are related to constraints C_1 and C_2, respectively.

Proposition 1. *The closed-form solutions of $\overline{q_s}$ and $\overline{p}$ exist. Equation (10) has the following two roots:*

$$\overline{p} = \frac{\beta_1 \pm \sqrt{\beta_1^2 + 8c/N\beta_3}}{4} \tag{12}$$

and

$$\overline{q_s} = \frac{\ln\left(\left(\beta_1 \pm \sqrt{\beta_1^2 - 8c/N\beta_3}\right)/2\beta_2\right)}{\beta_3} \tag{13}$$

where $\lambda_1 = 0$ and $\lambda_2 = 0$.

Proof. To get this result, we first need to find (11) the first derivative of q_s and p. Then set both derivatives to zero and set the constraint to $(\lambda_1 = \lambda_2 = 0)$. In this way, a closed-form solution can be derived by a set of equations consisting of (14) and (15).

$$\frac{\partial \mathcal{L}(\cdot)}{\partial q_s} = -Np\beta_2\beta_3 \exp(\beta_3 q_s) - c + \lambda_1 = 0 \tag{14}$$

$$\frac{\partial \mathcal{L}(\cdot)}{\partial p} = N(\beta_1 - \beta_2 \exp(\beta_3 q_s) - 2p) + \lambda_2 = 0 \tag{15}$$

□

Next, we can consider two special cases where the data quality level q_s is fixed or the subscription fee p is fixed. The former corresponds to a situation where the data product has a fixed quality level, and the data platform owner only optimizes the subscription fee. In contrast, the latter corresponds to a fixed subscription fee and the data platform owner only optimizes the quality level of the data. We have the following proposition.

Proposition 2. *On the one hand, if q_s is fixed, the solution $\overline{p}$ of the problem in (10) is globally optimal. On the other hand, if p is fixed, the solution $\overline{q_s}$ of the problem in (10) is globally optimal.*

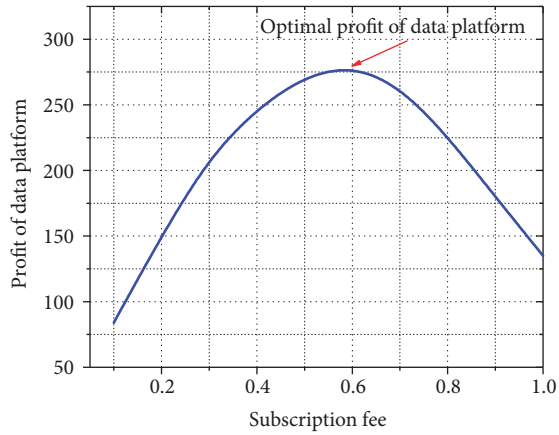

Figure 6: Data platform profit under different subscription fees.

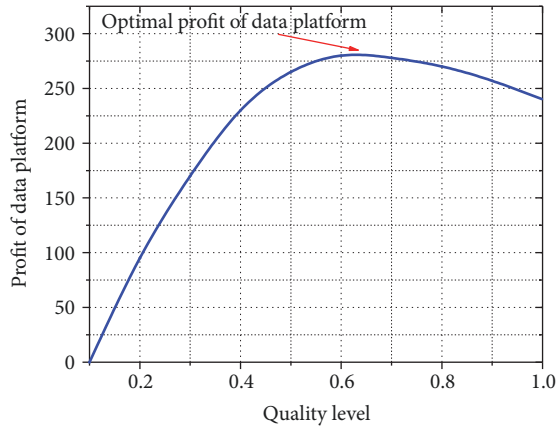

Figure 7: Data platform profits under different quality levels.

Proof. We solve the second derivatives of $\mathcal{G}(q_s, p)$ for q_s, p, respectively.

$$\frac{\partial^2 \mathcal{G}(q_s, p)}{\partial q_s^2} = -pN\beta_2\beta_3^2 \exp(\beta_3 q_s) < 0 \quad (16)$$

$$\frac{\partial^2 \mathcal{G}(q_s, p)}{\partial p^2} = -2N < 0 \quad (17)$$

which are nonpositive. Therefore, the solutions of the special cases are globally optimal. □

6. Numerical Experiment

In this section, we consider using the previous utility function $U(q_{sj}; \beta)$ to obtain the numerical results of the optimal pricing scheme. From this, we can further provide data platform owners with useful decision strategies. We adopt the fitted parameters as shown in Figure 5. In addition, we assume that the number of consumers is 1000. For verification purpose, we standardize the data quality level from 0 to 1.

Figures 6 and 7 show the profit of the data platform under parameters β, p, q_s. In Figure 6, we set the fixed data quality level $q_s = 0.6$ and, at the same time, change

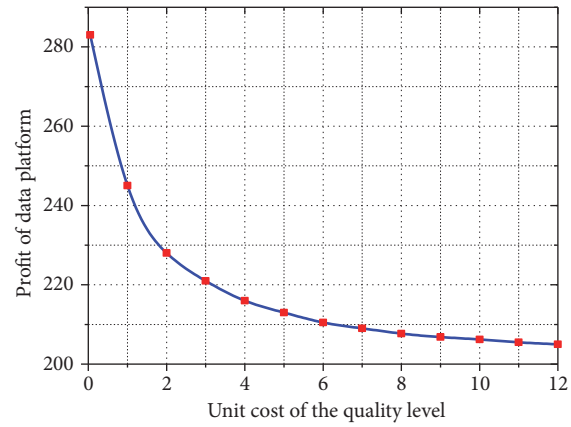

Figure 8: Data platform profit under different quality level costs.

subscription fee. Obviously, when the price of data is low, it will stimulate consumer spending and bring profit growth to the data platform. When the price of data is high, it will reduce the profit of the data platform. The possible influence factor is that too high price affects the willingness to pay of consumers, resulting in the loss of data platform profits. Obviously, the optimal subscription price to maximize profits can be calculated by (10).

In Figure 7, we fix $c = 1$ to investigate the effect of different levels of data quality on platform profits. Obviously, when the quality level is lower, the utility of the data and optimized subscription fee are also lower, resulting in less profit for the data platform. However, if the data quality level is high, the cost of the data platform will also increase (i.e., the data platform needs to pay more for the data publisher), which will lead to lower profits. The curve in Figure 8 shows the result. The profit of the data platform decreases as the data price of per unit quality increases. Obviously, the maximum profit can be achieved when applying the best requested data quality.

7. Conclusions

In this paper, we proposed a data pricing and profit maximization model based on data quality levels. We first constructed a linear model of the quality score based on the data quality dimension and used the square root to divide the quality level. Then we established a quality level utility model and verified the applicability of the model with machine learning algorithms. Finally, we proposed an optimized pricing mechanism allowing data platform owners to optimize quality levels and subscription fees to maximize profits.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported in part by National Nature Science Foundation of China (Grant no. 91646202) and National Key R&D Program of China (SQ2018YFB140235).

References

[1] S. A. Fricker and Y. V. Maksimov, "Pricing of data products in data marketplaces," in *Proceedings of the International Conference of Software Business*, vol. 304, pp. 49–66, 2017.

[2] Y. Jiao, P. Wang, D. Niyato, M. Abu Alsheikh, and S. Feng, "Profit maximization auction and data management in big data markets," in *Proceedings of the 2017 IEEE Wireless Communications and Networking Conference, WCNC*, pp. 1–6, San Francisco, CA, USA, 2017.

[3] A. Muschalle, F. Stahl, A. Laser, and G. Vossen, "Pricing approaches for data markets," in *Proceedings of the Workshop Business Intelligence for the Real Time Enterprise*, pp. 129–144, 2012.

[4] Xignite, http://www.xignite.com/.

[5] Gnip, http://support.gnip.com/.

[6] Factual, https://www.factual.com/.

[7] Infochimps, http://www.infochimps.com/.

[8] Microsoft windows azure marketplace, https://azuremarketplace.microsoft.com/en-us/marketplace/.

[9] H.-T. Moges, V. Vlasselaer Van, W. Lemahieu, and B. Baesens, "Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes - An exploratory study," *Decision Support Systems*, vol. 83, pp. 32–46, 2016.

[10] M. Barnabishvili, T. Ulrichs, and R. Waldherr, "Data on the descriptive overview and the quality assessment details of 12 qualitative research papers," *Data in Brief*, vol. 8, pp. 1059–1068, 2016.

[11] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais, and I. Ahmad, "Cloud computing pricing models: a survey," *International Journal of Grid and Distributed Computing*, vol. 6, no. 5, pp. 93–106, 2013.

[12] F. Teng and F. Magoulès, "Resource Pricing and Equilibrium Allocation Policy in Cloud Computing," in *Proceedings of the 2010 IEEE 10th International Conference on Computer and Information Technology (CIT)*, pp. 195–202, Bradford, United Kingdom, June 2010.

[13] H. Shah-Mansouri, V. W. S. Wong, and R. Schober, "Joint optimal pricing and task scheduling in mobile cloud computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5218–5232, 2017.

[14] X. Liu, M. Dong, K. Ota, P. Hung, and A. Liu, "Service pricing decision in cyber-physical systems: insights from game theory," *IEEE Transactions on Services Computing*, vol. 9, no. 2, pp. 186–198, 2016.

[15] D. Niyato, X. Lu, P. Wang, D. I. Kim, and Z. Han, "Economics of internet of things: an information market approach," *EEE Wireless Communications*, vol. 23, pp. 136–145, 2016.

[16] S. Balasubramanian, S. Bhattacharya, and V. V. Krishnan, "Pricing information goods: a strategic analysis of the selling and pay-per-use mechanisms," *Marketing Science*, vol. 34, no. 2, pp. 218–234, 2015.

[17] Y. Bakos and E. Brynjolfsson, "Aggregation and disaggregation of information goods: Implications for bundling, site licensing, and micropayment systems," *Lectures in E-Commerce*, pp. 103–122, 2001.

[18] 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.

[19] X. Wei and B. R. Nault, "Monopoly versioning of information goods when consumers have group tastes," *Production Engineering Research and Development*, vol. 23, no. 6, pp. 1067–1081, 2014.

[20] M. Li, H. Feng, F. Chen, and J. Kou, "Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers," *Computers & Operations Research*, vol. 40, no. 10, pp. 2374–2386, 2013.

[21] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proceedings of the Symposium on Principles of Database Systems*, vol. 62, pp. 167–178, 2012.

[22] 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.

[23] J. Yang and C. Xing, "Personal data market optimization pricing model based on privacy level," *Information*, vol. 10, no. 4, p. 123, 2019.

[24] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134–142, 2016.

[25] F. Stahl and G. Vossen, "Data quality scores for pricing on data marketplaces," in *Asian Conference on Intelligent Information and Database Systems*, vol. 9621, pp. 215–224, 2016.

[26] X. Ding, H. Wang, D. Zhang, J. Li, and H. Gao, "A fair data market system with data quality evaluation and repairing recommendation," in *Asia-Pacific Web Conference*, vol. 9313, pp. 855–858, 2015.

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

[28] F. Stahl, *High-quality Web information provisioning and quality-based data pricing [Ph.D. thesis]*, University of Münster, 2015.

[29] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Computers & Industrial Engineering*, vol. 112, suppement 1, pp. 1–10, 2017.

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

[31] J. Yang and C. Xing, "Data source selection based on an improved greedy genetic algorithm," *Symmetry*, vol. 11, no. 2, p. 273, 2019.

[32] R. Blake and P. Mangiameli, "The effects and interactions of data quality and problem complexity on classification," *Journal of Data and Information Quality*, vol. 2, no. 2, pp. 1–28, 2011.

[33] J. Greene and J. Baron, "Intuitions about declining marginal utility," *Journal of Behavioral Decision Making*, vol. 14, no. 3, pp. 243–255, 2001.

[34] The mnist database, http://yann.lecun.com/exdb/mnist/.

[35] R. Kieser, P. Reynisson, and T. J. Mulligan, "Definition of signal-to-noise ratio and its critical role in split-beam measurements," *ICES Journal of Marine Science*, vol. 62, no. 1, pp. 123–130, 2005.

[36] K. Wertenbroch and B. Skiera, "Measuring consumers' willingness to pay at the point of purchase," *Journal of Marketing Research*, vol. 39, no. 2, pp. 228–241, 2002.

[37] G. Gordon and R. Tibshirani, "Karush–Kuhn–Tucker conditions," *Optimization*, pp. 1–26, 2012, https://www.cs.cmu.edu/ggordon/10725-F12/slides/16-kkt.pdf.