分类号 TP399　　密 级 公开
U D C 004.65　　编 号

# 中南财经政法大学

## 硕 士 专 业 学 位 论 文

# 信息熵与数据质量驱动的数据定价研究

## ——以机器学习数据集为例

研 究 生 姓 名：瞿　天　卓
校内教师姓名、职称：王 会 举 副 教 授
申 请 者 类 别：普 通 硕 士
校 外 导 师 姓 名：无
专 业 学 位 类 别：工程硕士专业学位
专 业 名 称：电 子 信 息
研 究 方 向：计 算 机 技 术
入 学 时 间：二〇二〇年九月

二〇二二年六月二日

# Information Entropy and Data Quality Based Data Pricing

## —Take Machine Learning Datasets as An Example

2022. 6 .2

# 摘　要

数据市场是目前研究的热点话题，在提出数据要素化的政策之后，国内各类数据交易平台迅速发展起来，针对数据市场的研究也越来越多。在 2021 年，北京成立了国际大数据交易所，同年，上海大数据交易所也宣布成立，这更加激发了对数据交易市场的研究热潮，同时也标志着我国数据要素市场的正处于蓬勃发展阶段。在数据交易市场的相关领域中，对数据定价体系的研究更是重中之重。现有的研究大多是从传统商品的定价方案进行迁移，但是数据资产具有独有的特性，例如数据可重复使用；数据可以在个体之间共享；数据复制简便等等，这些特性使得传统的定价方法并不适用于数据的定价，需要进一步的研究来探讨更加合理，更加符合数据资产特征的定价方法。本文从数据本身的属性出发，制定了数据定价策略，并进行了相关实验。

本文讨论了机器学习中数据的相关问题，提出通过数据市场来加快机器学习中数据的流通，在明确了数据市场的结构之后设计了数据市场中最为关键的定价策略。为了解决现有定价方案对消费者不透明、无法体现数据本身属性的问题，本文考虑通过数据的价值来制定价格，价值越高的数据具有越高的价格。在衡量数据价值的方法上，本文选择综合“质”（数据质量）和“量”（信息熵）两个指标来衡量数据价值，提出了计算数据集质量分数以及信息熵的数学方法，并且证明了评价方法的合理性。在得到数据价值的量化结果后，以此为基础制定了定价策略，实现数据的版本控制，并考虑了消费者的自由选择行为，建立了定价模型。实验部分针对了不同的数据成本类别以及不同的消费者分布分别进行了模拟实验，通过对实验结果的分析说明了本文制定的定价方案的优势所在，并且通过利润最大化和市场覆盖率来评价实验结果，此外，还讨论了该方案在实际运用上的可操作性。

本文的创新性成果主要集中在两个方面：第一，结合了信息熵和数据质量两个维度来衡量数据价值，改良了单一指标的评价结果；第二，在对模型的求解上，对传统的遗传算法进行了改进，提出了子代择优的遗传算法，避免算法陷入局部最优之中。本文的主要贡献在于针对数据资产领域的热门的数据定价问题进行研究，为数据定价策略提供了一种新思路。使用价值分数来制定数据的价格让现有数据市场上信息不对称的问题得到解决。并且，以机器学习中的数据集为例，面对具体的数据形式提出了实际的解决办法，并且实现了利润最大化和较高的市场覆盖率，消费者和数据提供者双方的行为，对数据市场的运营模式有一定的启示，同时也可以促进数据市场的良好发展。

**关键词：**数据市场；数据定价；信息熵；数据质量

# Abstract

Data market is a hot topic in the current research, after proposing the policy of data element, various types of data trading platforms in China have developed rapidly, and more and more research on the data market has been carried out. In 2021, Beijing established an international big data exchange, and in the same year, the Shanghai big data exchange was also announced, which further stimulated the research boom of the data trading market, and also marked that China's data element market is in a vigorous development stage. In the relevant field of the data trading market, the study of the data pricing system is even more important. Most of the existing research is migrated from the pricing scheme of traditional commodities, but data assets have unique characteristics, such as data reusability; data can be shared between individuals; data replication is easy, etc., these characteristics make the traditional pricing method not suitable for data pricing, and further research is needed to explore a more reasonable and more consistent pricing method with the characteristics of data assets. Starting from the properties of the data itself, this thesis formulates a data pricing strategy and conducts related experiments.

This thesis discusses the related issues of data in machine learning, and proposes to speed up the flow of data in machine learning through the data market. After clarifying the structure of the data market, the most critical pricing strategy in the data market is designed. In order to solve the problem that the existing pricing scheme is not transparent to consumers and cannot reflect the attributes of the data itself, this thesis considers the value of the data to set the price. The higher the value of the data, the higher the price. In the method of measuring the value of data, this thesis chooses to measure the value of data by combining the two indicators of "quality" (data quality) and "quantity" (information entropy). the validity of the evaluation method. After obtaining the quantification results of data value, a pricing strategy is formulated based on this, which realizes the version control of data, and considers the free choice behavior of consumers, and establishes a pricing model. In the experimental part, simulation experiments are carried out for different data cost categories and different consumer distributions. The analysis of the experimental results shows the advantages of the pricing scheme formulated in this thesis, and the experiments are evaluated through profit maximization and market coverage. As a result, in addition, the operability of this scheme in practical application is also discussed.

The innovative results of this thesis are mainly concentrated in three aspects: first, combining the two dimensions of information entropy and data quality to measure the value of data, and improving the evaluation results of a single indicator; second, in the solution of the model, the traditional evolutionary algorithm is improved, and an evolutionary algorithm that choose a better offspring to avoid the algorithm falling into a local optimum. The main contribution of this thesis is to study the popular data pricing problems in the field of data assets, and provide a new idea for data pricing strategies. The use of value scores to price data solves the problem of information asymmetry in existing data markets. And, taking the data set in machine learning as an example, a practical solution is proposed in the face of the specific data form, and the profit maximization and high market coverage are achieved. The operation mode of the data market has certain enlightenment, and it can also promote the good development of the data market.

**Key Words:** Data market; data pricing; information entropy; data quality

# 目　录

# 绪论

## 第一节　研究背景与意义

### 一、研究背景

随着信息技术的快速发展，以及云计算、人工智能等技术的涌现，大数据已经成为各行各业的核心推动力[1]。世界上的数据量目前一直处于快速增长之中，得益于移动网络和智能终端的兴起，除了传统的 IT 公司、研究机构之外，政府部门以及金融机构等各行各业在日常的运营过程中都会产生大量的数据。根据 IDC①的一项研究表明，到 2025 年，全世界的数据总量预计可以达到 175ZB②，并且，在这些数据中，有近 80%的数据为非结构化的数据[2]。近年来，在大数据商业化日渐加深的背景下，数据被公认为工作和生活的重要资源。一些传统的线下产品和服务如今都转为线上的形式，数据成为支撑业务的基础，此外，许多应用程序也都是基于数据的二次使用或重用而构建的[3][3]。世界经济论坛预测③，到 2022 年，全球 GDP 的 60%将实现数字化，基于数据驱动的数字化网络和平台，全世界将会产生有约 60-70%的新价值。数据的广泛共享和重用对经济有着深远的影响，在数字经济时代，数据通常被当作计算资源进行使用[5]，并且大数据在未来的发展前景也十分可观[6]。中国作为一个人口大国，产生的数据也是巨大的，如果有效的治理数据，使数据成为社会发展的动力是一个重要的话题。2020 年 4 月 9 日晚，中共中央发布了《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》，将数据纳入新的生产要素，与土地、劳动力、资本等传统生产要素并列，这意味着数据已经成为了继人力资源以及物力资源之后的又一重要资源，大数据的发展必将推动生活、工作以及社会进行一场大变革[7]。

在政策的支持下，数据作为生产要素为各行业赋能，加快了我国经济的数字化建设，数据在流通与共享的过程中创造了大量的价值，这使得数据被当作一种新的资产进行研究[8]。在之前的研究中，研究者普遍关注于数据计算的效率以及有效性，在数据为社会以及各个企业创造了大量的价值之后，数据本身的价值成为需要思考的问题。数据作为资产在市场上流通之后，必然会伴随着交易的进行，数据提供者可以通过向需求方转让数据来获取收益，这就产生了一些新的问题，例如数据价格如何衡量，怎么保护数据所有权等。目前国内有许多运营中的数据

① 国际数据（International Data Corporation）
② ZB 表示十万亿亿字节
③ World Economic Forum (December 2018) ‘Our Shared Digital future: Building an Inclusive, Trustworthy and Sustainable Digital Society’

交易所，例如最早的贵阳大数据交易所[④]、北京大数据交易支撑平台[⑤]等，但是从贵阳大数据交易所的交易数据来看，还有许多的问题需要改进[9]，数据的确权、定价以及管理等都存在一定的空白，需要进一步的研究来完善交易机制。

数据作为资产来讲具有一定的特殊性，与传统的资产不同，数据资产具有一些新的特性，例如数据的复制成本接近于零，数据的存储形式也与传统商品存在区别等，这些特性使得传统的定价策略很难适用于数据资产，必须制定出一种全新的定价方案。并且，对于个人来讲，数据的价值因人而异，在设计定价策略的时候，需要找到通用的方法来体现数据的一般价值，让市场中的大部分消费者接受制定出来的价格，在这种情况之下，考虑数据的本身特性，提出合理的定价方案是非常有必要的，一方面是对数据定价领域的完善进行探索；另一方面也可以起到一定的启发作用。

## 二、研究意义

在中央明确了数据在我国发展中的重要地位之后，意味着需要对数据定价领域进行深入的研究，这是我国要素市场建设的一项重要任务。目前市场中交易机制并没有统一规范，亟待更多的理论作为支撑。

### （一）学术意义

目前我国在数据定价领域还处于一个起步的阶段，各方学者都在对该领域进行探索，相对于国外的研究而言，国内的相关研究在数量以及研究内容上都存在明显的不足，更多的是对定价方案进行归纳总结，很少提出全新的定价方案。此外，大部分的研究将数据市场的结构与数据定价分离开来研究，没有综合两者进行考虑，因此，系统的定价理论比较少见，数据定价这一领域还有很大的开发空间。本文针对机器学习中的数据集提出了一种新的定价方案，是对目前现有定价理论的补充，同时也为数据定价模式提供一定的参考，具有比较重要的学术意义。

### （二）实践意义

在现存的数据交易市场中，交易的情况并不理想，其中有一个很大的问题是“柠檬市场”[10]的形成。在市场中，商家与消费者很容易形成信息不对称，一方面，数据一般包含了许多条目，消费者无法在短时间判断数据的好坏，商家很容易以此来欺瞒消费者；另一方面，数据市场的各项理论以及管理机制等都还不够成熟，数据对于买卖双方的透明度完全不同，消费者会逐渐对市场失去信任，这样的一种情况显然是不利于市场的发展的。因此，需要设计一种定价方案，体现数据的本身价值，尽可能的让消费者通过价格可以得到相应数据集的信息。

此外，完善的定价机制可以加快数据流通的效率，让消费者可以更加快速的购买到所需的数据。随着数据相关的业务量的增大，以及各种技术上的创新，社

[④] http://www.gbdex.com
[⑤] http://www.shujushichang.com

会上对于数据的需求量会增加，需要更为成熟的理论来维持供需平衡，扩大数据交易市场的受众，这样才能有利于我国数据领域的发展。定价方案可以运用到实际的数据市场中，让数据的定价变得更加合理，一方面，数据消费者会被成熟的定价机制吸引，另一方面，可以激发数据提供者提供数据，加快资源的流转。

综上所述，本文提出了信息熵与数据质量驱动的定价方案来解决用于机器学习中数据集的定价问题，从数据集的信息熵以及质量两个方面来衡量数据集的价值，并将此作为定价的重要依据，让数据集的价格体现数据本身的特性，这样才能让数据集对于消费者来说更加透明，一定程度上解决了买卖双方信息不对称的问题，从学术以及实践两个方面来说，都具有一定的价值。

## 第二节　国内外研究现状

数据市场领域的研究是新兴的研究热点，国外学者的研究早于国内学者，但是在目前为止，还未形成一套国际认可的标准理论体系，不论是基础的市场理论，还是定价的策略与标准，目前都处于探索的阶段。1998 年，Armstrong 在研究中第一次使用了“数据市场”这个词，将数据市场定义为买卖数据的平台，并且对数据市场的结构进行了介绍[11]。数据定价是数据市场中最为重要的一个研究方向，从定价的策略上面看，目前的研究主要分为两大类：一类是基于数据交易的方式，另一类是基于数据本身的性质。

从数据交易的方式上来制定定价策略，Muschalle 和 Stahl 等人将定价策略分成了以下几种：免费、按使用付费、套餐定价、统一费率定价、两阶段定价以及 Freemium 定价[12]。免费的定价机制可以吸引更多的消费者，但是不利于各方利益的分配。按使用付费符合一般的思维习惯，但是对于数据产品来说，边际成本几乎为零，从数量上来进行定价并没有优势。套餐定价类似于移动运营商的模式，提供几个可供选择的套餐模式，也有一些研究在针对其进行优化[19]。统一费率是单纯按使用时间来进行收费，两阶段定价是在统一费率的基础上针对不同消费者进行了个性化的分类。Freemium 定价是指免费提供基础的服务，消费者需要为增值服务支付相应的费用。上述定价模式都存在一些问题，数据作为交易的物品来说，消费者需要了解数据的相关信息，特别是对于机器学习来讲，数据的好坏直接决定了研究成果的有效性，这会影响消费者的购买选择，因此，需要有一种合理的定价方式来体现数据的透明度。

本文希望从数据本身出发来讨论数据定价问题，针对用于机器学习的数据集，提供一个系统的定价方案，让价格体现数据的相应价值，下文将会从三个角度来分析目前的相关研究情况，分别是传统的定价方法、基于数据质量的定价以及基于信息熵的定价。

## 一、传统定价方法的研究现状

传统的定价方法主要可以分为两类：一类是基于传统经济学的定价，另一类是基于博弈论的定价。

基于传统经济学的定价是从经济学的角度出发，将传统商品的定价策略运用于数据的定价上，实现数据资产的成本控制、利润最大化等。Balazinska 等人在 2011 年对数据市场进行了系统的研究，对数据市场的发展前景进行了展望[15]，提出了一些新兴的数据市场发展方向，并且提出了定价模型的两个属性：可理解性以及可预测性。可理解性是指数据提供者在提供了数据之后，如果数据出售给了消费者，获取的利润需要有一个合理的解释，同样对于消费者来说，他所支付的价格也需要有一个充分的说明，类似于消费清单。可预测性是指开发商可以粗略的估算出未来的收益情况，消费者在明确了自己的需求之后可以估算出后期需要的投入。对于数据来说，可预测性是比较难实现的，因为数据的形式、体量等方面与传统商品有着很大的区别。Koutris 等人在研究中提出了一个基于查询的数据定价的通用框架，实现了框架的无套利和无折扣，但是并没有给出具体的定价方案[17]。在后续研究上，利用了查询的时间复杂度来确定相应查询的价格[17]。Zheng,Peng 等人提出了一种新的基于在线查询的群体感知数据定价机制[17]，考虑了数据的不确定性、经济的稳健性，实现了模型的无套利以及收益的最大化，但是并不适用于普遍的场景，并且没有考虑市场的覆盖率。以上这些方法有一个共同的问题就是消费者需要为重复的数据来进行付费，因为查询的结果之间可能有重复的数据条目，也可能需要进行相同的几次查询，这就导致的定价策略的不合理性。除了为数据的查询进行定价，还有直接出售原始数据的形式[18]，通过使用数据获得的收益，提取部分作为数据提供者的收入，这种做法的缺点是提取收益的比例难以确定。

为了使得定价策略可以吸引到更多的消费者，有许多学者开始考虑数据市场中各参与者之间的关系。拍卖是一种市场的交易模式，买卖双方具有更多的自由度，类似于完全竞争的模式，经常用于寻找到标的物的合适价格，从这点上来说是非常适用于数据的定价的，因为数据的价值因人而异，可以通过拍卖找到最为合适的买家。王婷婷研究了运用拍卖对大数据进行定价的形式[19]，在研究中结合了传统信息产品的定价理论，并引入了评分机制，在模型的构建中加入了信用因素的影响，以此来改进对应的定价模型。该模型对消费者公开了一定的信息，但是并没有关注到数据本身的性质上，没有实质性的解决信息不对称的问题。

博弈论是定价和市场领域的一种有用方法，尤其是在数据商品的定价中。主要有三种不同的博弈论方案用于数据定价模型：非合作博弈、斯塔克伯格博弈，以及讨价还价博弈。Mazumdar 等人设计了一个定价模型来评估物联网传感数据[20]，在该模型中，所有供应商都以竞争的方式销售其数据，并将该模型定义为非

合作博弈。Niyato 等人建立的模型中[21]，将服务与数据绑定在一起，鼓励与消费者协商来指定价格，并且在之后的研究中利用斯塔克伯格博弈来实现商家的利润最大化。通过博弈论来实现数据定价虽然可以让数据趋于一个合理的价格，但是无法排除恶性竞争的存在，长期以来是不利于市场的良性发展的。

传统的定价方法基本上是从一般商品的定价策略迁移过来的，但是数据作为一种新的要素，具有很多与传统商品不同的性质，在制定价格的过程中，需要将数据的特性考虑进去。

## 二、基于数据质量定价的研究现状

质量可以在一定程度上决定价格，质量越高的产品往往具有更高的价格。数据同样如此，在过往研究中，已经表明了数据质量是一个多维的概念[22]，如果数据消费者认为数据的各项维度达不到期望的标准就不会购买该数据。

Pipino 在研究中提出了三种度量数据质量的函数形式：比率、最大或最小算子以及加权平均，并开发了一种主客观相结合的评价数据质量的方法[23]。但是对于数据定价领域来说，主观的信息指标收集十分困难，一方面由于市场的受众很广，信息收集过于繁琐，另一方面，考虑主观的因素可能会使通过质量制定出来的数据价格偏离部分消费者的预期。因此，研究基于数据质量的数据定价时，更多的需要考虑客观因素，将关注的重点集中到数据本身的属性之上。

Wang 和 Strong 对数据的各项维度进行了调查和分类研究[24]，以制定确定数据质量特征的分层框架，最终他们在总共收集的 179 个标准中确定了 15 个相关维度。Batini 等人在研究中根据过往学者提供的质量维度分类[25]，提出了一组基本维度。韩京宇等人也对此进行过综述[26]，为数据质量的研究提供了一个系统的思路。在这些研究中，虽然涉及到的数据质量维度众多，但是全部都认为数据的完整性、准确性、一致性以及及时性是最为重要的四个维度。在确定了质量维度之后，可以用具体的方法来计算数据质量。Heckman 等人在研究中建议关注数据的内在价值和质量[27]，而不是基础的信息价值，这样的作法可以提定价的高透明度和公平性。他们将数据年龄、数据量等因素纳入影响因素，建立了一个计算质量分数的线性模型，并且在该模型的基础之上评估数据的价值。但是文中并没有给出具体的定价方法，只是给出了一个简单的思路，没有对实际情况的模拟。除了简单的线性模型，也有学者考虑了质量维度之间的相互影响，Yu 等人研究了两项质量维度的相互影响[28]，研究认为数据在某项维度上的提升会导致数据在另一维度的表现下降，例如，当数据的及时性提升时，数据的完整性会受到影响，因此采用了非线性的计算方式来描述这一情况，并给出了定价方案，但是当质量的维度变多，维度之间的相互影响会变得难以描述，即使找到了合适的数学表示，求解上也会变得十分困难。

### 三、基于信息熵定价的研究现状

在 1948 年，香农借鉴了热力学当中熵的概念，并且将其在信息论之中进行了拓展[29]，并指出了熵的计算方法，熵的值越大，意味着不确定性越大。Holzinger 等人运用信息熵来进行数据挖掘的工作，并且讨论了近似熵、样本熵、模糊熵以及拓扑熵[30]。此外，在机器学习领域，也有学者通过熵值进行调整，优化实验的结果[31][32]。对于数据来讲，熵值可以反映数据集包含的信息量大小，通常来说，信息量越大的数据集对于研究来讲是更有价值的，因此，可以把熵值作为数据价值的衡量指标。

在姚建国等学者的研究中[33]，提出了基于数据集信息熵来进行数据定价的方法，从数据集中行与列的关系上来计算数据集所包含的信息量，以此为基础来制定数据的价格。Li 等人提出了三个基于数据信息量的定价函数[34]，充分证明了通过信息熵来进行数据定价的合理性。但是单纯以信息熵来决定价格太过于片面，因为特殊的数据会导致熵值的增加，在数据集中，异常数据的存在会导致研究结果的偏差，这点通过熵值是体现不出来的。有一些学者在研究中对信息量的表示做了改进，李希君提出了一个新的概念叫数据信息熵[35]，能够更好的表示数据的信息量，并且给出了相应的定价函数。Shen 等人在数据的信息熵的基础上[36]，实现了模型的无套利定价，但是主要针对的是个人数据。上述的研究中，还是没有解决影响因素过于单一的缺点。

## 第三节　研究内容与贡献

### 一、研究内容

为了解决当前数据要素交易流通中的核心难点问题之一——数据定价问题，本文对定价的方案进行了系统的研究，并提出了一种新的定价策略，该定价策略主要针对于机器学习中的数据集。本文从现有的定价方法上出发，总结出现存方法的不足之处，提出使用信息熵以及数据质量来衡量数据集本身的价值，计算出数据集对应的价值分数，以此为基础来实现定价，并证明了价值分数的合理性。在具体的定价方案上，考虑了消费者的行为以及支付意愿，构建出相应的定价模型，运用了版本控制的思想来实现模型利润的最大化，提升市场覆盖率。最后对本模型进行了模拟实验，改良了传统的遗传算法来求解模型，验证了该模型在市场中有着良好的表现，可以实现更加细致的市场划分，让数据出售方有利可图，数据消费者也更容易接受，此外，还使用了具体的数据集对该方案进行了实际的操作。

本文总体上可以分为两大部分：定价模型的构建以及实验验证。

### （一）模型构建

在模型的构建上，首先通过信息熵以及数据质量计算出数据集对应的价值分数，通过价值分数来实现多版本策略，消费者根据自己的购买意愿在各版本之间进行自由选择，模型的目标函数为利润的最大化。

### （二）实验验证

在模型构建完成之后，通过遗传算法对模型进行求解，通过模拟实验的总利润以及市场覆盖率来说明模型的有效性，并采用具体的数据集说明定价方案的实施过程。

## 二、贡献

第一，本研究针对数据资产领域的数据定价问题进行研究，为数据定价策略提供了一种新思路。

第二，在以往的研究中，有分别从信息熵或者数据质量来进行数据定价，文中将这两者结合起来考虑，提出了价值分数的概念，并且论证了价值分数的合理性。

第三，目前存在的大多数数据定价方法并没有考虑消费者的利益，本文通过研究数据集本身的价值，用价格来体现数据集的效用，这样可以为消费者的购买选择提供一定的参考，提高信息的透明度。

第四，本文以机器学习中的数据集为例，面对具体的数据形式提出实际的解决办法。

第五，本文提供的模型能够提高市场覆盖率，实现利润最大化，并且考虑了消费者的购买意愿，有利于市场的持续发展。

# 第四节　论文结构安排

本文主要分为五个部分来介绍文中涉及的理论、方法以及定价模型的构建与验证。首先论证综合了信息熵以及数据质量得出的价值分数的合理性，并根据价值分数设计定价策略，构建出相应的模型，并通过实验说明了模型的优势所在。文章的总体研究思路如图 0-1 所示。

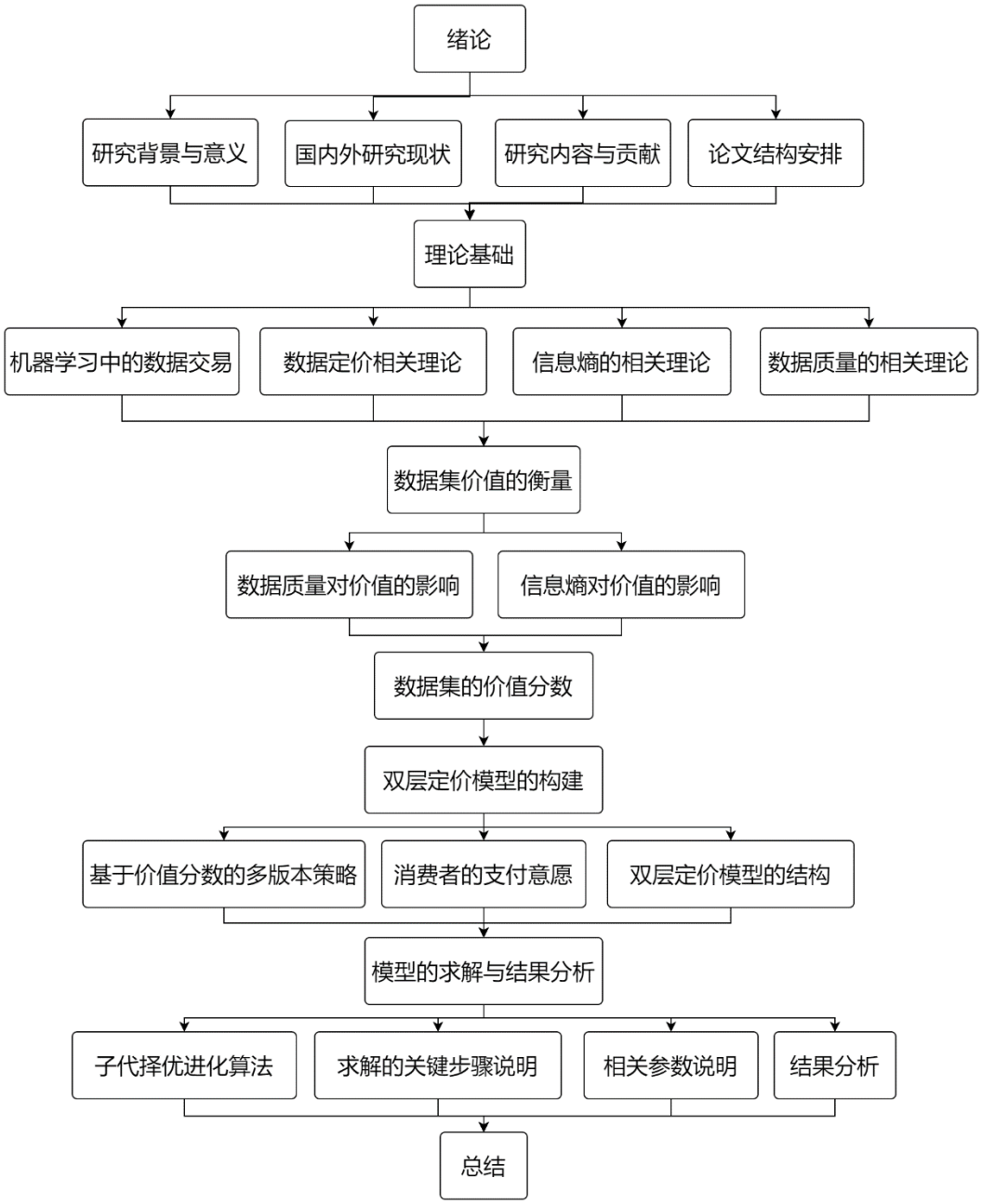


图 0-1 研究思路图

本论文的研究内容主要分为以下 6 个板块：

绪论部分详细的介绍了数据定价的国内外研究背景，阐述了文章选题的研究意义、主要研究内容和相关贡献，并概括了文章的基本结构。

第一章介绍了机器学习背景下的数据定价问题，阐述了数据市场的基本结构，并对数据定价的基本形式进行了介绍，还说明了信息熵与数据质量的相关理论知识。

第二章介绍了衡量数据价值的方法，并给出了数据集信息熵、质量分数的计算方法，综合得到数据集的价值分数，并且说明了根据数据价值形成的定价策略

的优势以及价值分数的合理性。

第三章介绍双层模型的构建过程，对模型的前提假设进行了阐述，介绍了基于价值分数的多版本策略以及消费者的非线性支付意愿函数。

第四章对本文建立的双层模型进行求解，算法上采用了子代调优的遗传算法，对不同消费者分布以及不同成本下的情况分别进行了实验，证明了模型可以实现利润最大化。

第五章对本文的主要贡献进行了总结，并且分析了文中设计理论的不足之处，讨论了相关领域未来的研究方向。

# 第一章 相关理论基础

## 第一节 机器学习中的数据交易

世界上的数据总量正在以每天大约 25 万亿字节的速度爆炸，并且，世界上几乎90%的数据都是在过去的几年之中创造出来的[37]。随着物联网越来越多地涉及到日常生活，各行各业的数据都在市场上流通，数据的来源变得是多样化[38][39]。企业运用相关的数据研究出众多的智能系统，并向市场提供服务[40][41][42]，多样化的数据源导致了庞大的数据量，同样也创造了巨大的商业价值，这一类数据被称为大数据。

为了充分发挥大数据的效用，各类的大数据应用被开发出来分析和挖掘大数据潜在的价值。在过去的几十年中，因为社会经济的发展和政策的帮助，促进了各领域数据量的增长，世界上各个领域都转向使用大数据技术，Oracle、IBM、Microsoft、Dell等公司在大数据管理和应用程序的开发方面投入了大量资金。此外，大数据应用行业每年 10%左右的速度增长，几乎是传统软件领域的两倍，在大数据众多的应用场景中，目前应用最广的两个领域是人工智能以及机器学习领域。

本文研究的场景为用于机器学习中的数据交易问题。机器学习在许多应用上都取得了突破性的成功，这导致了各行业对于机器学习应用需求的爆炸式增长[43]，Louis最近的一篇研究中预测，到 2024 年，全球机器学习市场预计将达到 208.3 亿美元[44]，这使得社会上对于数据的需求也在飞速的增长。构建一个成功的机器学习应用是一个复杂的过程，需要多方的合作。在应用开发的开始阶段，一方可能需要从另一方获取原始数据，并且通过数据标注服务来构建用于机器学习的训练数据，获取了数据之后，构建机器学习模型也需要多方的参与，一方训练好了模型之后，可以提供给另一方的来解决业务当中遇到的实际问题，或者基于模型来开发相应的应用程序。机器学习可以挖掘出数据背后的价值，是连接多方的管道，而数据对于机器学习至关重要。机器学习模型，尤其是深度模型，必须要通过大量的数据进行训练和测试，此外，将训练完毕的模型投入使用也需要数据，因为机器学习模型的输入也是由数据构成，并且通过机器学习构造的模型，在维护和更新的过程中仍然需要数据，可以看出，数据几乎贯穿整个机器学习管道，所以说数据是机器学习的关键所在，这也是文章将研究的对象定为机器学习中的数据集的主要原因。

但是在实际情况下，机器学习需要的数据往往是很难获取的，一方面，收集数据、创建标签和确保数据质量的成本往往是十分昂贵的，个人或者小公司可能无法负担起相应的费用；另一方面，在模型训练初期难以找到合适的数据，因为要使模型达到预期的效果，对数据要求是比较高的，不仅数据的各项属性需要对应需求，数据所提供

的信息量也很重要。因此，对于数据需求者来说，需要扩充获取数据的途径，数据需要在市场上流通起来，为了使得数据的流通交易更具有规范性，必须建立完整的市场体制，并且完善数据定价以及其他相关的策略。

考虑机器学习的场景，数据收集者从各种渠道收集数据，收集好的数据流向数据平台或者研究者手中，研究者运用数据进行研究开发，将成果给到企业或者其他机构进行实际的运用，形成应用程序或者制定发展方案，整个机器学习的过程可以用图 1-1 来表示。

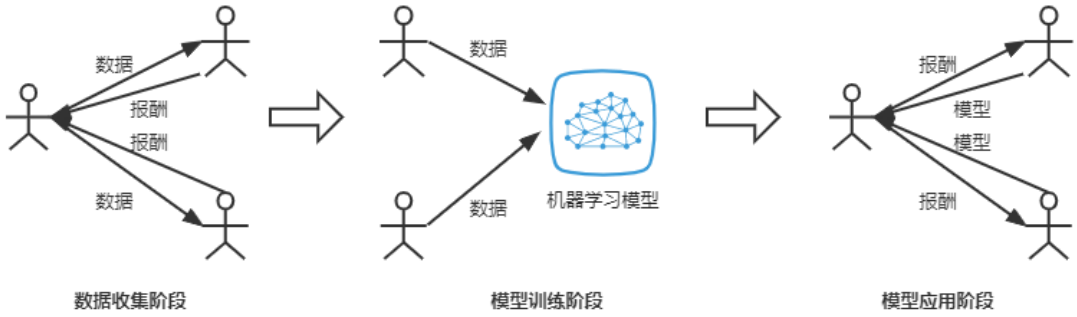


图 1-1 机器学习流程

图中可以看到，在机器学习的各个阶段都会有交易的产生，主要集中在数据收集阶段以及模型应用阶段：在数据收集阶段，数据从个人或者其他来源流向数据收集者，收集方向数据来源方支付一定的报酬以获取想要的数据；在模型应用阶段，为了得到模型所提供的服务，购买方需要支付模型对应的金额。本文将关注点聚焦到机器学习的第一个阶段，为机器学习中用到的数据集制定合理的定价策略。

## 第二节 数据定价的相关理论

数据定价是数据资产理论中的重要课题，是属于数据市场中的一项策略，数据市场的运营需要制定合理的定价模式。所以，在研究数据定价之前，需要了解数据市场的基本结构。

### 一、数据市场的基本结构

近年来，将数据作为一种商品进行评级已经变得越来越流行，数据市场已经成为一种新的商业模式，在市场之中，各种来源的数据可以被收集、处理、充实、购买和出售。对于传统的商品而言，建立完整的市场结构的作用是为了促进商品的交易，数据市场对数据这种特殊的商品而言，也可以起到同样的作用。数据市场的存在可以促进数据的交易，从个人或者组织收集的数据通过数据市场的中介作用出售给需要的人。如果没有数据市场的存在，数据的交易就会陷入混乱之中，各方的利益都会受到损失，

所以市场的规范与否决定了数据交易能不能往好的方向发展，在研究数据定价时，构建良好的市场体制是第一步。此外，数据市场除了规范各项制度，还可以加快数据的供需匹配，数据的买方和卖方通过数据市场来建立联系，使得数据在市场中更快地流通，有利于卖方通过数据来获利，也提高了数据的利用率。对于买方而言，能够更加快速、准确地获取需求的数据。所以，构建数据市场是十分重要的。

数据市场旨在促进数据交易，并作为数据代理从个人、公司和开放来源收集数据，并将其出售给数据消费者，如广告商、软件开发商、零售商、制造商、电信服务提供商等。数据市场在最初的时候是一种私人大规模信息交换，但是随着数据来源变得越来越广，各种数据存储和数据处理技术的提升，数据变成了一种公开的资源。数据市场结构的划分有多种方式，根据 Vomfell 等人的研究[45]，可以将数据市场的结构分为两类：

（一）单边市场

在单边市场中，市场只负责处理数据提供者或者数据消费者任意一边的业务，这样的市场结构称之为单边市场，并且可以分为两类，一类是面向卖方的单边市场，如图 1-2 所示，另一类是面向买方的单边市场，如图 1-3 所示。

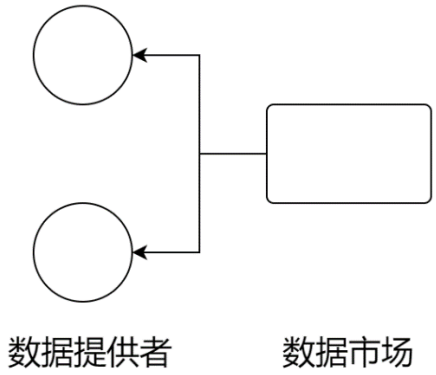


图 1-2 面向卖方的单边市场

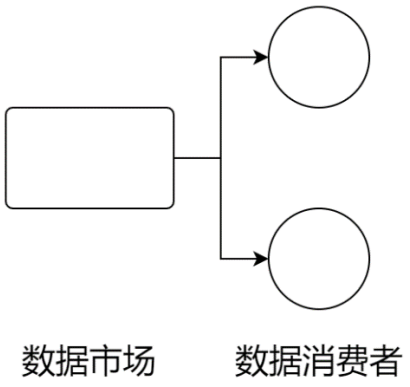


图 1-3 面向买方的单边市场

上述两种市场结构都较为简单，因为只需要考虑一边的情况，数据市场的管理较为简单，但是这种模式下，失去管理的另一边会变得比较混乱。并且，数据提供者与数据消费者并没有通过数据市场形成联系，很容易失去对市场的信任。

**（二）双边市场**

双边市场与单边市场的不同之处在于兼顾了买卖双方的管理。其中，有一种特殊的双边市场，买方和卖方直接进行交易，不通过任何官方机构的管理，这样的市场结构称之为分散的双边市场，如图 1-4 所示，通过建立严格数据市场模式来进行买卖双方的管理，这样的市场结构称之为集中的双边市场，如图 1-5 所示。

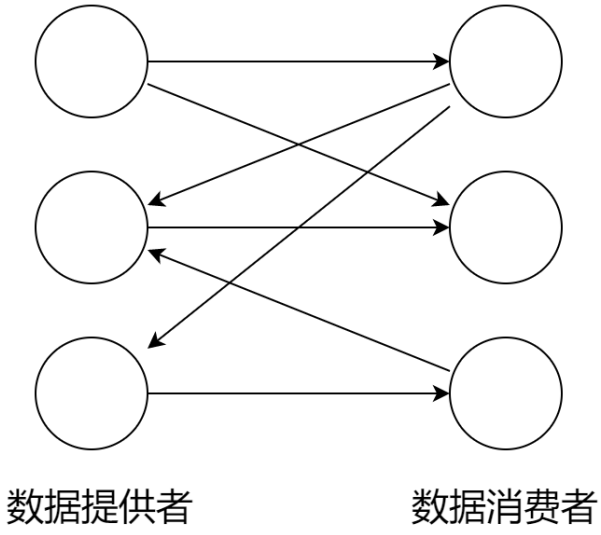


图 1-4 分散的双边市场

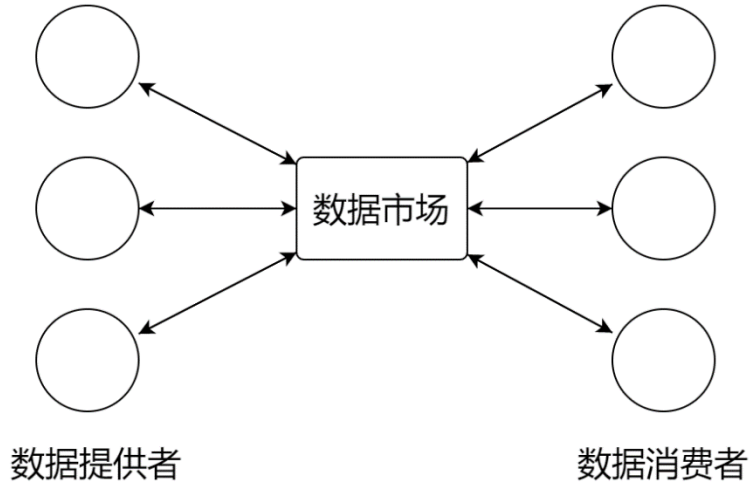


图 1-5 集中的双边市场

在上述的两种多边市场结构中，与单边市场最大的区别在于，买卖的双方直接或者间接形成了联系，这样的一种情况是非常有利于市场发展的。分散的双边市场中，

买卖双方可以随着交易的进行不断地进行信息交换，逐渐形成统一的运营标准，但是缺点在于，对于卖家来说，可能更倾向于选择自己信任的卖家，不利于市场接纳新卖家。并且，很容易形成私下的恶性交易，不利于市场的规范。在集中的双边市场中，买卖双方不进行直接的信息交换，而是通过数据市场来进行交易。数据市场从卖家手中获取所出售的数据，将数据的相关情况暴露给消费者进行参考，消费者自己选择想要购买的数据。这样一种做法便于对市场的交易行为进行系统的管理，有利于保护数据提供者的相关权益，并提供给消费者公平的交易环境。

## 二、数据资产的特性

根据 ISO 对于信息的定义来说[46]，信息是关于事实、过程以及思想等客体的知识，并且，当信息处于特定语境之中时，会有特定的含义，而数据则是信息的载体，是信息的一种形式化的体现方式，通过数据，可以达到交流、解释或处理所承载信息的目的。

在我国已经通过了数据要素化的相关政策，已经进入了数据要素化的时代，数据产品交易将会是促进经济发展的重要方式。随着数据资产交易的合法化，数据在日后将会变成一种产权可以界定，可以通过一定的手段进行交易的商品。在 Pei 等人的研究中，将数据分成了两类进行研究，分别是数字产品和数据产品[47]。在研究中，从四个方面分析了两者的区别：首先，数字产品往往是整体进行定价和销售的，例如电影等，而数据产品往往可以分割为一条一条的数据记录；其次是数据产品相比数字产品更加灵活，可以随意的聚合和分割；然后是用途方面也有区别，数字产品大多是是即时消费，而数据产品在购买了之后往往被用于模型的分析训练之中；最后，数据产品的转卖比数字产品更加灵活，因为数字产品只能整体售卖，而数据产品可以通过多种方式进行处理之后再进行转卖。

目前研究主要针对的是数据产品，因为只有数据产品才符合数据要素化的要求。数据需要在不同的行业发挥作用，通过与实际业务的结合来推动行业的发展，这样才能实现真正的数据资产化。对于传统的资产来讲，会计学中归纳了资产的五条特征[48]：资产有望为会计实体带来经济利益或者具有为会计实体服务的潜力；资产应为会计实体拥有或控制的资源；资产由会计实体过去的交易或事项形成；为企业创造与资源相关的经济利益资源的成本或价值能够可靠地计量。相较于传统的资产属性来说，数据资产除了具备上述的特点之外，还具有四个崭新的特点。

### （一）虚拟性

数据有别于传统资产，属于一种抽象的资源，虽然承载了很多的信息，但是只能通过一定的表现形式才能直观的看到，例如数据的可视化等手段。数据的虚拟特性使得数据不会像其他商品一样容易发生损耗，因为数据可以在各类电子存储设备之上储存，这也使得数据的运输成本相比传统资产要低很多。

### （二）边际成本为零

传统资产的每一次生产都需要一定的成本，即使掌握了生产技术，仍然需要购买生产所需的人力物力等。对于数据资产来说，在获取到一份数据之后，理论上来说可以对该数据进行无限的复制，也就是说数据资产的边际成本接近于零。

### （三）非竞争性和非排他性

正是由于数据资产具有虚拟性以及边际成本为零的特点，数据资产还表现出非竞争性和非排他性。非竞争性是指数据资产由于可以通过分享和复制交由他人使用，不像传统资产，资源是有限的，容易形成竞争。非排他性是指当一位消费者购买了数据之后，并不会影响其他消费者使用相同的数据，不会形成对资源的独占。

### （四）价值的不确定性

传统资产的价值大多是确定的，例冰箱、洗衣机之类的电器，消费者购买之后的用途基本一致。但是数据资产具有很强的主观性，同样的一份数据对于不同的人群有着截然不同的价值。例如对于发电厂风机的各类参数数据来说，研究这方面的专家对这样的数据会很感兴趣，但是对于普通人群来讲是没有任何价值的。其次，各行业对于数据的要求标准也会不一样，在金融行业，更注重数据的时效性，而对于研究行业来说，以往的数据也是重要的资产。最后，数据的价值还与其本身的质量、规模等有关，质量越高的数据会具有更高的价值，规模越大的数据 所带来的信息量也是更大的。所以，数据资产的价值具有很大的不确定性。

由于数据资产具有上述的这些特点，一些传统的定价方法可能并不适用于对数据资产的定价上，需要通过系统的研究开发出新的定价方案。

## 三、数据定价的基本形式

数据市场的关键在于数据定价策略。在现实生活中，有很多对数据进行定价的例子，分别采用隐式或显式的方式为数据进行定价。对于隐式定价的形式来说，互联网服务经常会收集用户的部分信息，用户在使用服务时，经常会要求同意服务的隐私条款，这些条款中就包括允许收集用户信息；一些商店会向消费者提供免费者购物卡，消费者用卡购物会享受一定的优惠，但是商店会保留这些消费者的购买历史记录，这些记录会提供给技术部门进行内部数据分析或者直接将其出售给第三方；还有一些公司在做营销的时候，会采用问卷的形式，通过问卷调查收集消费者的偏好信息，并给参与者优惠券或者小礼品等作为回报。上述这些无价值服务、折扣和优惠券都是对数据进行隐式定价的方式。另一方面，通过显式的方式进行数据定价通常是以交易平台的方式，例如数据公司 AggData[⑥]以固定价格出售位置数据，同时以折扣的方式出售数据集的捆绑包，还采用了订阅的方式允许消费者不受限制的访问各类数据，并提供通过协商定制数据的服务。另一家公司 Datacoup[⑦]采用了按月进行付费的模式，如果

⑥ https://www.aggdata.com/
⑦ http://datacoup.com/

月费由相应的数据属性决定。

在数据收集过程中，严格的补偿机制可以鼓励数据提供者共享其数据，并降低数据消费者补偿数据提供者的成本，在销售数据时，定价机制有助于吸引数据消费者并增加数据销售者的收入。然而，由于数据资产的特殊性，为数据定价并不容易。在目前的研究中，对于机器学习中的数据产品，可以分为三种主要的定价场景：第一种是为原始数据定价，即为收集所需的原始数据制定价格；第二种是为数据标签定价，数据标签的正确与否关系到训练得到的模型的准确性；第三种是为通过机器学习训练得到的模型定价。在本文中，定价的情形为机器学习所需的原始数据进行定价。Shapiro等人将定价的策略分为三大类[49]，分别是个性化定价、集体定价以及版本控制：个性化定价为每一个消费者提供不同的价格，但是这种模式需要收集每个客户的众多信息，并且由于法律等条件的约束，这种方式几乎是实现不了的；集体定价在个体与个体的联系上对市场进行了划分，不同的集体之间有价格差异；版本控制意味着不同版本的商品可以被区分，并以不同的价格出售，不同版本之间的区别可以是质量、技术支持以及服务等，商家提供给消费者一系列存在一定差异的产品，消费者可以进行自由选择。版本控制策略对于数据来说是一种较好的定价策略，因为即使是同样的一份数据，对于不同的消费者个体来讲，有着不一样的价值，并且，有些消费者可能只需要数据集的一部分数据。此外，对于数据产品来说，消费者购买了一份数据之后，一般不会出现再次购买同一份数据的情况，因此，如果想要占据更大的市场份额，需要提供一个数据集的多个版本让消费者进行选择，这样才能提高利润。

## 第三节 信息熵

信息的用途是消除不确定性，这同时也是信息的价值所在，一份信息可以消除的不确定性越多，这份信息所携带的信息量就是越大的，其他不能够消除不确定性的称为数据或者噪音。

可以用公式来描述信息量，假设信息量为 1 的信息用一个信号来表示，对于$n$个信号量来说，其所表示的信息量即为$n$。信号量可以表示事物的可能性关系，如公式(1-1)所示。

$$S = log_2 N \tag{1-1}$$

在公式(1-1)中，$S$表示总的信息量，$N$代表一个事件所有可能的个数。

公式(1-1)只适用于每个事件发生的概率都是相等的情况下，但是在实际情况下，大部分的时间发生的概率是不相等的，可以对公式(1-1)进行扩展，一般的形式如公式(1-2)所示。

$$S=\sum_{i=1}^{n} p_i \log \frac{1}{p_i}=-\sum_{i=1}^{n} p_i \log p_i \tag{1-2}$$

在公式(1-2)中，$S$表示各个事件发生的概率不相等的情况下，信息量的大小，$p_i$表示对应事件发生的概率。

公式(1-2)在所有的事件发生的概率都相等的情况下，信息量在数值上等于信息熵。信息熵是最早由香农[⑧]在信息论中所提出来的概念，通过信息熵可以定量的描述信息量，也就是信息不确定性的程度。在计量单位上，信息熵通用的单位是 bit，当公式中对数函数的底不同时，也会用到不同的计量单位。

在王小鸥等人的研究中[50]，认为在数据中出现频次更高的数据更加具有代表性，并且证明了信息熵的合理性。信息熵的计算方法如公式(1-3)所示。

$$H(X)=E[I(X)]=E[-\ln P(X)] \tag{1-3}$$

在公式(1-3)中，$H(X)$表示信息熵，$X$表示事件随机变量，$E$表示整体的期望函数，$P(X)$是$X$的概率分布，$I(X)$表示$X$携带的信息量。

当$X$的个数有限时，公式(1-3)可以转化为(1-4)：

$$H(X)=\sum_{i} P(x_i)I(x_i)=-\sum_{i} P(x_i)\log_b P(x_i) \tag{1-4}$$

在公式(1-4)中，$[x_1,x_2,,,x_n]$表示事件所有可能的结果。

除了计算公式之外，信息熵还具有四个性质：单调性、非负性、累加性以及连续性。单调性是指随着事件发生概率的增加，事件携带的信息量就越少。非负性是指熵值是非负的，当信息不能够消除任何的不确定性，也就是信息量为零，这时候熵值也为零。累加性是指总的不确定性可以通过每个事件的熵值累加得到。连续性是指表示信息熵的函数是连续函数。

## 第四节 数据质量

对于传统商品而言，质量一定程度上决定了价值，数据同样如此。数据质量是指将质量管理技术应用于数据的活动的开发和实施，以确保数据适合在特定环境中满足组织的特定需求。高质量数据是指符合预期需求的数据，当数据质量出现问题时，可能是出现了重复的数据、不完整的数据、不一致的数据、不正确的数据、定义不明确的数据、组织不严的数据或者是数据存在安全性。大数据行业是对数据质量要求较高的行业，研究人员需要通过对大数据进行统计分析来得出有效的结论。大数据被广泛接受的特征是“3V”模型，其中“3V”指的是数量、种类和速度 。大数据最为突出的特征是其庞大的数据量，Gao 等人在研究中表示，一个人的手机通话记录通常有

[⑧] 美国数学家、信息论的创始人，提出了信息熵的概念

7000 万个数据条目[51]，随着电子产品的多样化与交互性的提升，数据量的提升还在继续。大数据的种类指的是数据的多样性，这意味着大数据具有多样化的数据来源、数据结构和应用场景。速度则是指大数据具有实时更新的性质，例如，空气质量监测数据通常每天需要更新多次。如今，除了“3V”模型外，“4V”和“5V”模型也在出现，IBM 对大数据进行了全新的定义，将准确性纳入考虑范畴，以解释大数据带来的偏差问题，并且认为“4V”模型能够准确描述大数据 。

从“3V”模型到“4V”模型，增加了对数据质量的考虑。随着机器学习等技术的兴起，大数据不再一味的追求数据量，数据的质量成为了关注的焦点。一方面，数据的传输以及应用的速度变得极快，一些错误的数据很可能直接造成难以挽回的损失；另一方面，基于大数据的研究深入到社会上各个领域，其研究成果的优劣性与数据质量息息相关。因此，将大数据行业的焦点聚焦于数据质量是十分有必要的。

在大数据生命周期的各个环节，都会有相应的质量问题。在科学研究中，为了实验能够获得准确的结果，科学家们严格的确保数据的可靠性。但是如今随着大数据的商业化，对于商家来说，大数据应用的目的成了利润最大化，商家没有责任也无需确保数据的可靠性，并且，商家收集数据的方式、采用的算法等对于消费者来说都是不透明的，商家只需要保证自己的产品能够吸引主要的客户群体就可以实现盈利，这就导致数据的质量很难得到保证。

此外，数据的完整性也是一个影响数据质量的重要问题。大数据庞大的数据量并不一定等同于信息量，有可能一份数据包含的信息十分有限。例如手机呼叫数据，几乎每个人都会有产生可观的数据量，但是这种数据的应用范围却十分狭隘，在被用来研究人口的流动情况时具有一定的效用，但是这份数据中记录的位置只是运营商基站的位置，并不能准确地记录每个人的实时位置。此外，这份数据中的每一项并没有表现出相关个体的个性，没有每个人的社会属性，无法描述个体的行为特征。因此，如何收集到相对完整的数据对于提高数据质量来说是十分关键的，数据完整性的提升可以使得数据的运用范围更加广泛，创造更多的价值。除此之外，数据的代表性同样值得关注，在收集数据的阶段，由于设备、收集方式等问题的影响，很难做到收集到每一个用户的数据，因此，一般收集数据都采用抽样的形式，但是抽样就意味着收集到的数据只能代表一些个体的特征，并没有涵盖整个群体。随着大数据运用的商业化，商家为了利润可能抛弃掉一些群体而得出有失偏颇的结论，而对于数据本身而言，代表性越强，其可以带来的价值会越高。数据质量在整个大数据的生命周期中都扮演着重要的角色，只有高质量的数据才能创造出最大的价值，并且，在大数据从产生到运用的每一个阶段都可能产生相应的数据质量问题，除了上文提到的那些例子，还有数据的一致性、可靠性等等，甚至数据还会涉及到隐私问题。

# 第二章　数据集价值的衡量

本研究希望针对机器学习中通用的数据集制定一套系统的数据定价方案，使得价格可以体现数据的本身属性，让价格与数据集的价值形成密切的联系，解决目前数据市场交易中信息不对称的问题。

## 第一节 从信息熵衡量数据的价值

熵是信息量的量度，对于一个离散变量$X$来说，如果所有可能的取值为$n$个，记为$\{x_1, x_2, \dots, x_n\}$，并且$X$的概率分布函数为$P(X)$，则离散变量$X$的熵的定义如公式(2-1)所示。

$$E(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{2-1}$$

在公式(2-1)中，$E(X)$表示离散变量$X$的熵，$P(x_i)$表示$X$取值为$x_i$的概率，并且在$P(x_i) = \frac{1}{n}$时，$E(X)$取得最大值$\log_2 n$。

此外，当存在多个离散变量时，例如存在离散变量$Y$，可能的取值为$m$个，记为$\{y_1, y_2, \dots, y_m\}$，则两个变量的联合熵可以用公式(2-2)来表示。

$$E(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log_2 P(x_i, y_j) \tag{2-2}$$

在公式(2-2)中，$E(X,Y)$表示$X$，$Y$两个离散变量的联合熵，$P(X,Y)$为$X$，$Y$的联合概率分布函数。

对于机器学习的数据集来讲，通常的形式如式子 2-3 所示。

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \dots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \tag{2-3}$$

在公式(2-3)中，数据集$X$拥有$n$行数据，每一行数据有$m$个属性，将第$i$行数据记为$r_i = (x_{i1}, x_{i2}, \dots, x_{im}), i = 1,2, \dots, n$，第$j$列数据记为$s_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T, j = 1,2, \dots, m$。对于单个的属性$t_j$，该属性所有的取值为$k$个，记为$\{t_{j1}, t_{j2}, \dots, t_{jk}\}$，则该属性信息熵的计算如公式(2-4)所示。

$$E(t_j) = -\sum_{i=1}^{k} P(t_{ji})\log_2 P(t_{ji}) \tag{2-4}$$

在公式(2-4)中，$P(t_{ji})$为数据集中第$j$个属性取第$i$个值的时候的概率大小，其计算方法如公式(2-5)，(2-6)所示。

$$P(t_{ji}) = \frac{\sum_{i=1}^{n}\varphi(x_{ij}, t_{ji})}{n} \tag{2-5}$$

$$\varphi(x_{ij}, t_{ji}) = \begin{cases} 1, x_{ij} = t_{ji} \\ 0, x_{ij} \neq t_{ji} \end{cases} \tag{2-6}$$

公式(2-5)中，$\sum_{i=1}^{n} \varphi(x_{ij}, t_{ji})$表示每个取值出现的次数，将其在总数据行数中所占的比例作为概率。

通常，一个数据集具有多个属性，用集合$\rho_l = \{t_{j1}, t_{j2}, \dots t_{jl}\}$表示。$\rho_l$表示数据集具有$l$个属性，这些属性的联合熵可以用公式(2-7)表示。

$$E(\rho_l) = E(t_{j1}, t_{j2}, \dots t_{jl}) \tag{2-7}$$

联立公式(2-2)、(2-5)以及(2-6)可以求解出公式(2-7)中$E(\rho_l)$的值。为了便于计算，将其映射到(0,1)之间，由于当每个离散变量出现的概率相等时，熵取得最大值，记此时的值为$E_{max}$，则熵值的变换如公式(2-8)所示。

$$Entropyscore = \frac{E(\rho_l)}{E_{max}} \tag{2-8}$$

## 第二节　从数据质量衡量数据的价值

当数据成为交易的资产，数据市场所有者需要制定合理的定价策略，数据质量直接影响数据的价值，因此可以通过数据质量来评估数据的价值。一方面，可以通过数据量的大小来测量数据的价值，Niyato 等人在研究中运用了该方法实现了模型利润的最大化[52]；另一方面，可以根据数据的质量来测量数据的价值。对于机器学习来说，数据集的质量越高，更容易得到良好的训练效果，因此，从数据质量的角度来评估机器学习中数据集的价值是比较合理的。首先，需要确定衡量数据质量的不同维度，然后，基于选取的这些维度，建立了一个线性模型来评估数据价值。

数据质量包括多个维度，选取的衡量维度不同，计算的结果就会出现偏差，数据与数据之间也会有差异，因此必须在众多的质量维度中选取最具代表性的维度，并且，由于面向的是通用的数据集，维度的选取还必须具有普适性。在 Stahl 等人的研究中[53]，讨论并证明了考虑质量维度对数据定价的适用性，并且为推进通过数据质量来进行数据定价提供了一些研究的思路。在之后的研究中，Stahl 总结了七个质量维度[54]，分别是精确度、完整性、冗余、数据量、延迟、响应时间以及及时性，同时考虑了市场上消费者对于数据产品的需求，通过改变上述的某一项或者几项创造出不同的数据版本，调整数据质量的级别，实现连续的版本控制，以这样的做法来促使更多的消费者购买数据，并给出了具体的模型。结合相关学者对数据质量的研究，选取最具代表性的三项维度对机器学习中的数据集进行数据质量的衡量，分别是准确性、完整性以及冗余性。准确性是数据的基础特性，因为信息是信息的载体，如果数据不够准确，其所携带的信息也会是错误的，这样的数据是没有任何价值的，并且在机器学习中还会影响训练结果。完整性也是极为重要的一个维度，数据不完整也会造成信息的缺失。冗余性是指重复的数据，如果一份数据集中出现了大量的重复数据，不仅增加了研究

所需的时间，对研究结果也没有任何的正面效用。任何的数据集都会在这三项维度上形成差异，因此，这三项维度也是最为通用的，适用于衡量大多数的数据集的数据质量，上述三个质量维度的定义及其计算方法如表 2-1 所示。

表 2-1 质量维度的定义及其计算方法

| 质量维度 | 定义 | 参数 | 计算公式 |
| --- | --- | --- | --- |
| 准确性($pa$) | 根据数据源的域和数据的信息类型，表示数据源中具有正确值的单元格比例 | $nce$:错误单元格的数量<br>$ncl$:单元格总数 | $pa = 1 - \frac{nce}{ncl}$ |
| 完整性($pc$) | 表示数据集中完整单元格的比例。即单元格不是空的，并且单元格中的值有意义 | $nr$:数据行数<br>$nc$:数据列数<br>$ic$:不完整数据的数量<br>$ncl$:单元格总数 | $ncl = nr * nc$<br>$pc = 1 - \frac{ic}{ncl}$ |
| 冗余性($pr$) | 冗余表示数据集中重复记录的比例。 | $nr$:数据行数<br>$ndr$:重复的数据行数 | $pr = 1 - \frac{ndr}{nr}$ |

表 2-1 中质量维度的第一项为准确性，对于用于机器学习的数据集来讲，数据往往反映了现实世界中的某一项指标，如果数据出现了偏差，会影响机器学习的结果。错误单元格是指数据集中与数据源的域和数据的信息类型不符的数据，例如对于未成年人信息的数据集，年龄限制在 0-18 之间的整数，如果年龄对应的单元格中出现负数或者超过 18 的整数，这一类就属于与数据源的域冲突的数据，如果出现不是整数的数据，就属于与数据源的信息类型不符合，上述这两种情况全部都会纳入错误单元格的计算之中。完整性以及冗余性的定义易于理解，在计算方式上都是运用的简单的统计方法。

在表 2-1 中，由于各项维度的计算方式为比例，所以每一项的计算结果取值都在 0-1 之间，将所有的三项质量维度整合起来，用线性模型来计算总质量分数，如公式(2-9)所示：

$$Qualityscore = w_1 * accuracy + w_2 * completeness + w_3 * redundancy \tag{2-9}$$

在公式(2-9)中，$w_1 + w_2 + w_3 = 1$，这样就保证了总体的质量分数依然分布在 0-

1 的区间之中，具体的权重值可以由用户根据实际情况自行进行调整，对于对准确性要求较高的数据，可以提高相应的权重，其他两个维度也可以以同样的方式进行调整。此外，该计算方法还可以很容易的扩展到多个维度，如公式(2-10)所示。

$$Qualityscore = w_1 * accuracy + w_2 * completeness + w_3 * redundancy + \cdots + w_n * dimension_n \tag{2-10}$$

在公式(2-10)中，$w_1 + w_2 + w_3 + w_n = 1$，$dimension_n$表示设定的第$n$个衡量数据质量的维度。这对于一些特殊行业的数据是有意义的，比如金融行业非常重视数据的时效性，所以可以加上对该维度的考察。由于本文的研究对象为通用的机器学习的数据集，因此，以准确性、完整性以及冗余性对数据集的数据质量进行衡量。

## 第三节　数据集的价值分数

在上文中，介绍了分别从信息熵以及数据质量来衡量数据价值的方法，但是只从其中的一个角度来表示数据价值有点过于单一，下文将会把两者结合起来，从“质”（数据质量）和“量”（信息量）来计算数据的价值，并讨论了这种方法的合理性。

### 一、数据集价值分数的计算

信息熵是信息量的度量，对于一份数据来说，熵值($Entropyscore$)的大小代表了信息量的多少，并且，信息量越多表示这份信息的价值越大。质量同样如此，质量分数($Qualityscore$)越高，代表数据本身的价值越高。两者都可以影响数据的价值，并且都是正相关的形式，因此，可以将两者综合起来，形成数据集的价值分数，计算方式如公式(2-11)所示。

$$v = w_E * Entropyscore + w_Q * Qualityscore \tag{2-11}$$

公式(2-11)中，$w_E$表示熵值所占的权重，$w_Q$表示质量分数所占的权重，并且有$w_E + w_Q = 1$。

通过上述过程，可以计算出数据集的价值分数，通过数据集的价值分数可以制定相应的定价策略，并且，权重值可以进行调整，以达到最好的效果。价值分数是一个通用的标准，任意一个数据集都可以按照上述方法计算出相应的价值分数，并作为数据定价的基础。

### 二、价值分数的合理性证明

通过价值分数来衡量数据集的价值是合理的，下面将对其合理性进行验证。

验证过程使用了来自 UCI 的 glass 数据集[9]，该数据集是一个纯数字的数据集，并且没有缺失，该数据集的部分数据如图 2-1 所示。

[9] https://archive.ics.uci.edu/ml/datasets/Glass+Identification

```
1,1.52101,13.64,4.49,1.10,71.78,0.06,8.75,0.00,0.00,1
2,1.51761,13.89,3.60,1.36,72.73,0.48,7.83,0.00,0.00,1
3,1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0.00,0.00,1
4,1.51766,13.21,3.69,1.29,72.61,0.57,8.22,0.00,0.00,1
5,1.51742,13.27,3.62,1.24,73.08,0.55,8.07,0.00,0.00,1
6,1.51596,12.79,3.61,1.62,72.97,0.64,8.07,0.00,0.26,1
7,1.51743,13.30,3.60,1.14,73.09,0.58,8.17,0.00,0.00,1
8,1.51756,13.15,3.61,1.05,73.24,0.57,8.24,0.00,0.00,1
9,1.51918,14.04,3.58,1.37,72.08,0.56,8.30,0.00,0.00,1
```

图 2-1 glass 数据集部分数据

数据集一共有 214 条数据，属性数量为 10，数据的第一列为编号，第二列为玻璃的折射率，第三列为玻璃中含钠(Na)的百分比，第四列为玻璃中含镁(Mg)的毫克数，第五列到第十列分别为铝(Al)、硅(Si)、钾(K)、钙(Ca)、钡(Ba)以及铁(Fe)的含量，数据集的最后一列为玻璃的类型，一共有七种类型的玻璃，用数字 1-7 来表示，但是该数据集中不包含类型为 4 的数据条目。

为了计算简便，选取最后一列来计算该数据集的信息熵，通过公式(2-4)可以计算信息熵，用 python 代码实现如图 2-2 所示。

```python
def calcShannonEnt1(dataSet):
    num=len(dataSet)                        #数据集的样本数量
    labelCount={}                           #创建一个数据字典，它的键是数据集最后一列的数据，即样本的类别；它的值是该分类中的样本数量
    #计算每种类别下的样本数量，并将其放在字典中对应的键下
    for featureVec in dataSet:
        label=featureVec[-1]                #取样本中的最后一个值
        if label not in labelCount.keys():
            labelCount[label]=1
        else:
            labelCount[label]+=1
    #计算数据集的熵
    shannonEnt=0.0
    for key in labelCount.keys():
        pro=float(labelCount[key])/num
        shannonEnt-=pro*log(pro,2)
    return shannonEnt
```

图 2-2 计算数据集的信息熵

通过计算可得，glass 数据集的信息熵为 2.1763（保留四位小数，后续数据做同样处理），由于一共有 6 个类别，所以该种情况下信息熵的最大值为 2.5850，由公式(2-8)可得，该数据集的$Entropyscore$为 0.8419,由于该数据并没有缺失或者错误的数据，所以质量分数$Qualityscore$为 1。

通过公式 2-11 计算价值分数，假设$w_E$= $w_Q$=0.5，此时价值分数$v_1$=0.9209。

接下来改变数据集中的一些值，观察各项指标上的变化。将数据集最后一列的前 50 个数字从 1 变为 4，由于该数据集并没有类型为 4 的数据，所以这 50 个数据代表了错误的数据。在这种情况下再次计算数据集的熵值为 2.4589，由于类别总数增加到

了 7 个，所以此时信息熵的最大值为 2.8074，此时的$Entropyscore$为 0.8759，比数据全部正确时候的分数 0.8419 要高，但是从实际情况上来说，数据带来了错误的信息，数据集的价值应该下降，所以，仅仅通过熵值来评估数据的价值会有一定的缺陷。

当使用价值分数进行衡量时，由于有 50 个错误数据，通过公式 2-9 来计算质量分数，为了计算简便，令$w_1 = 0.4, w_2 = w_3 = 0.3$，由于数据集依旧完整并且没有重复的数据条目，所以$Qualityscore$计算出来的结果为 0.9066，所以综合得到的价值分数$v_2$=0.8913，可以看到与第一次的结果相比，$v_2 < v_1$，由此可见，相比于用信息熵单独评价数据的价值，综合数据质量可以减少异常数据的干扰，使得对数据集价值的衡量更加准确。所以，将信息熵与数据质量结合起来，是对信息熵对错误数据反映不足的补充，运用价值分数来衡量数据质量，相对于单一的评价指标是有一定的优势的。

# 第三章　双层定价模型的构建

## 第一节　模型的前提假设

为了构建合理的数据定价模型，首先需要明确数据市场的基本结构。本节将会对一些前提假设进行说明。

本文的市场环境假设为垄断市场：一方面，在竞争的市场环境下，定价只需要尽可能的接近于成本就可以获得优势，在垄断的环境下才需要讨论定价策略，使得利润可以最大化，并且提高市场覆盖率；另一方面数据资产与传统的资产不同，容易形成垄断的环境，像 Facebook、Tencent 等互联网公司，几乎垄断了用户的相关数据；其次，垄断市场也更有利于资源的集中以及分配，对于数据需求方来说，垄断市场可以缩短搜寻数据的时间成本。在垄断的条件下，数据市场所有者负责提供数据交易的平台，尽可能保证数据交易的公正性、透明性，统一管理的标准，避免市场的秩序被扰乱，提高数据提供者和消费者对数据市场的信任度，进而使得市场朝着良性的方向发展。

对于数据市场的结构，在之前的讨论中，将数据市场分为了两大类，一类是单边的数据市场，一类是多边的数据市场。为了规范数据提供者以及消费者的行为，加强数据市场的管理，本文的数据市场假设为集中的多边市场结构。详细的市场结构如图 3-1 所示。

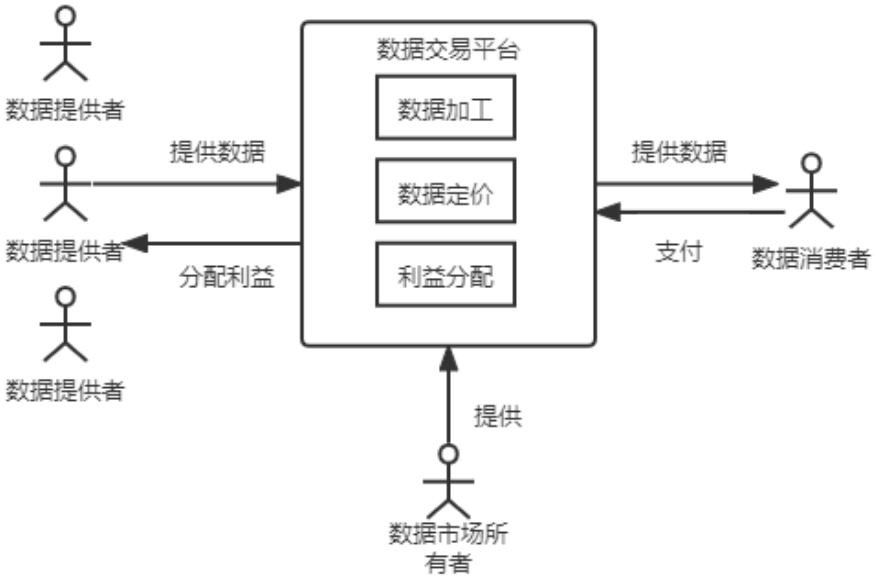


图 3-1 数据市场结构

在图 3-1 中，参与数据交易过程的主要涉及到了三个角色：数据提供者、数据市场所有者以及数据消费者。数据提供者负责收集和提供数据，是数据市场中数据的来

源，他们将手上拥有的数据提供给数据市场所有者，经过平台的加工处理，按照制定好的定价策略完成定价之后集中展示给消费者。数据市场所有者负责运营和管理数据市场，以及制定数据市场的相关策略，具体包括各项管理制度、定价策略以及交易规则等，更重要的充当中介的作用，建立起消费者与数据提供者的联系。消费者在官方平台上寻找合适的数据，支付相应的金额获得数据的所有权。在数据成功的销售之后，赚取的利润在数据市场所有者以及相应的数据提供者们之间分配。

在本文中，消费者被设定为对数据的价值分数敏感的群体。每一位消费者对数据的敏感程度不同，对于同一份数据，购买的意愿不相同，在多个数据产品存在的条件下，消费者会选择消费意愿最大的产品。

## 第二节 基于价值分数的多版本策略

商家在考虑产品销售策略时，实行多版本方针最常用的一种策略。例如在电器行业，生产商总是会生产几个版本的电器，不同的版本之间以质量水平作为区分，消费者根据自己的需求以及收入水平来选择自己需要的版本。计算机行业常用的开发工具 idea、pycharm 等，都会提供至少社区版以及专业版两个版本供市场上的消费者进行选择，社区版和专业版在功能上会存在着差异，一般的用户选择社区版进行使用，研究人员倾向于选择专业版，这样的策略使得他们拥有广大的用户群体。汽车界也是如此，由于生产成本过高，汽车厂商需要制定合理的生产方案，除了研究往期的销售数据之外，还有一种做法是提高生产的灵活性以避免滞销的损耗。现在每一款车几乎都会有不同的几个版本，用户还可以选择自己需要的配饰，汽车厂商通过灵活的版本控制来赚取更大的市场占比，提高销售利润。除此之外，生产商在设计以及生产最高质量版本的产品时，会产生固定的成本，由于信息产品的特殊性，在生产出高质量的产品之后，开发低质量版本的技术以及资源投入相对来说会减少，因为低版本的开发过程中并不存在挑战性的技术任务，因此，从成本控制的方面来讲，多版本策略也是合理的一种选择。经济学中对实物商品交易的早期研究表明，实物商品价格点的差异化主要受产品线特征差异化的影响，该模型由 Mussa 和 Rosen 在 1978 年提出，并命名为 V 垂直分割或质量分割[55]，在这个模型中，消费者更倾向于选择更高质量的商品而不是低质量的商品，为了满足不同的消费者，生产者通常提供不同质量水平的产品系列。

在本文中，考虑通过价值分数来区分数据产品，由于消费者对价值分数有着不同的敏感度，通过对数据集进行处理，改变数据集的价值分数，形成同一份数据的多个不同版本，并为每一个版本制定不同的价格，以满足市场中不同消费者的需求，如图 3-2 所示。

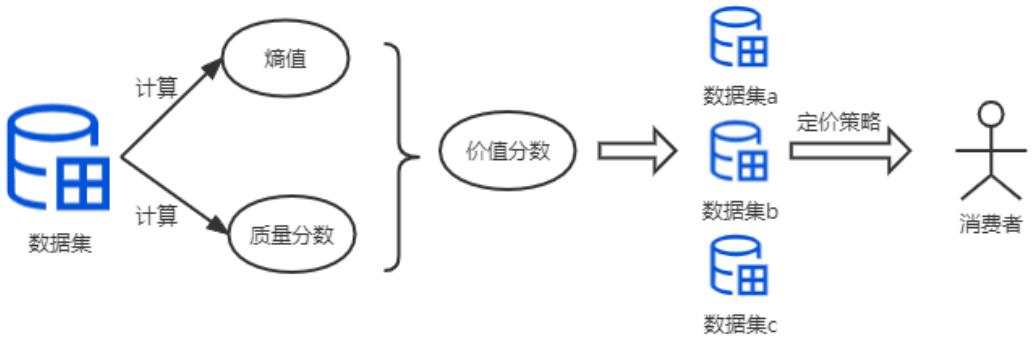


图 3-2 基于价值分数的多版本策略

在图 3-2 中，数据集经过数据平台的处理之后，以数据集 a、b、c 三个版本在市场中流通，三个版本之间通过价值分数形成差异。假设在市场中，提供了一系列价值分数不同的数据产品，记为$V=\{v_1,v_2,v_3,\dots,v_K\}$，其中$v_1<v_2<v_3<\cdots<v_K$，数值越大表示该级别的产品对应的价值分数越高，并且可以为消费者提供更多的功能需求。价值分数最高的版本记为$v_H$，所以有$v_K=v_H$，把价值分数最低版本的产品记为$v_L$，有$v_1=v_L>0$，所有的价值分数都分布在$[v_L,v_H]$之间，$K$表示该开发商提供给市场的版本数量，并且，在功能上，高价值分数的产品的包含了低价值分数版本的全部功能，所有版本的产品都是垂直兼容的，低版本产品在进行更新或者添加了功能模块之后可以升级到价值分数更高的版本，同样的，对于高价值分数的版本，通过一些删减处理，可以降低价值分数。

除了划分版本之外，还需要考虑每个版本的成本消耗。与传统的商品不同，对于大多数的数据产品来说，由于数据可以复制，并且可以在光盘等介质中存储，利用网络来进行传输，其边际成本可以忽略不记，在学术研究中，通常以零边际成本来处理。但是，对于很多产品，例如数据库开发工具等，高版本的产品除了在功能上有所区别之外，还代表了更好的售后服务等，这些会使得边际成本不再为 0，考虑到实际情况，将$K$个版本的产品对应的边际成本用$MC=[c_1,c_2,\dots,c_K]$表示，在本文中将会分三种情况来讨论，分别是零边际成本、线性边际成本以及二次递增边际成本。

零边际成本如公式(3-1)所示。

$$c_1=c_2=\cdots=c_K=0 \tag{3-1}$$

线性边际成本如公式(3-2)所示。

$$c_k=c_0+av_k,k=1,2,3,\dots,K \tag{3-2}$$

在公式(3-2)中，$c_k$表示第$k$个版本的成本，$c_0$以及 $a$都是常数，并且都大于 0。

二次递增边际成本如公式(3-3)所示。

$$c_k=c_0+aq_k+bv_k^2,k=1,2,3,\dots,K \tag{3-3}$$

在公式(3-3)中，$c_0$以及 $a,b$都是常数且大于 0。

此外，每一个版本的产品都有其对应的销售价格，记为$P=p_1,p_2,\dots,p_K$，并且有$p_1<p_2<p_3<\cdots<p_K$，在边际成本不为零的情况下，价值分数更高的的版本对应了

更高的成本，当成本增加的时候，该版本产品的售价也会随之增加。

## 第三节 消费者的非线性支付意愿函数

支付意愿(WTP)表示消费者对产品感兴趣的程度，在研究中，一般用消费者愿意为产品支付的价格来衡量。尽管对数据产品进行版本控制是一种很好的做法，但是在一些研究中，当消费者的支付意愿被定义为线性的函数时，模型的效果并不是最优的。线性的 WTP 函数形式如公式(3-4)所示。

$$W(e,v) = ev \tag{3-4}$$

在公式(3-4)中，$W$表示消费者的支付意愿，$e$表示消费者的类型，$v$表示产品在某一项指标上的级别。

消费者购买产品得到的效用如公式(3-5)所示。

$$U(e,v,p) = ev - p \tag{3-5}$$

在公式(3-5)中，$U$表示效用，$p$表示产品的售价。

消费者会倾向于选择对自己效用最大的产品。Chen 等人的研究[56]认为，这种情况下，实行多版本的策略并不一定是有利的，因为当引入低版本产品的时候，相对于购买更高价格的高版本产品，消费者通常会购买更加廉价的低版本产品。这就意味着刚开始选择高版本产品的消费者会转向购买低版本产品，降低了利润。例如，当市场上有级别为$v_1 = 0.2, v_2 = 0.85$的两个版本的产品，售价分别为$p_1 = 0.1, p_2 = 0.8$，对于类型$e = 1$的消费者来说，由公式(3-5)计算可得，购买$v_1$产生的效用$U_1$=0.1，购买$v_2$产生的效用$U_2$=0.05，由于$U_2 < U_1$，消费者选择低版本的产品进行购买。

在通常情况下，产品的版本越高对于消费者的吸引力就越大，因为高级别往往可以带来更好的体验。但是数据产品具有跟传统产品具有不一样的特性，一个最突出的特点就是数据产品具有更明显的划分。例如像 Idea 之类的专业软件，专业版相对于免费版有着更多的功能，需求专业版的消费者对于免费版是没有任何兴趣的，因为免费版提供不了他们所需的服务，而对于使用免费版的消费者来说，也根本没有必要购买专业版，因为功能已经过剩了。这就导致线性支付意愿函数可能并不适用于数据产品，需要开发出更符合上述情况的表现形式。

Sundararajan 等人使用了非线性函数对数据产品进行定价[57]，在消费者之间形成了价格歧视，取得了不错的效果。Krishnan 等人在用质量划分版本的情况下，提出了饱和质量的概念[58]，在质量达到饱和质量之后，WTP 的值增加的速率会发生变化，WTP 函数是一个分段的形式。本文将构造一个基于用户自由选择的非线性效用函数，以研究对数据产品的版本控制策略在不同消费者以及不同成本下的表现。

消费者对数据产品的价值分数敏感，当数据产品被提供给消费者之后，消费者按照一定的标准从中选择适合自己的产品。由于每一个消费者都是独立的个体，需要将每个人区分出来，本文中引入对于价值分数的敏感程度来区分消费者。敏感度高的消

费者对于数据集价值的追求要高于敏感度低的消费者，当需要在两个价值分数不同的版本之间做出选择时，敏感度高的消费者会倾向于选择价值分数更高的数据，用$e$来表示消费者对价值分数的敏感程度。为了便于描述消费者的自由选择过程，引入了效用函数如公式(3-6)所示。

$$U(e, v_k, p_k), k = 1,2,3, \dots, K \tag{3-6}$$

在公式(3-6)中，$U$表示产品对于消费者的效用，$k$表示数据产品的版本号，$v_k$表示第$k$个版本产品的价值分数，$p_k$表示第$k$个版本产品的出售价格。

当效用$U > 0$的时候，表示消费者购买产品能够带来一定的效用，有可能购买此产品，相反，当$U \leq 0$的时候，消费者一定不会购买此产品。当存在多个产品能够为消费者带来效用的时候，消费者会选择效用最大的产品。

对于敏感程度为$e$的消费者，选择效用最大的产品对应的价值分数版本，如公式(3-7)所示。

$$e^* = \underset{k}{\operatorname{argmax}}\{ U(e, v_k, p_k), k = 1,2,3, \dots, K\} \tag{3-7}$$

公式(3-7)中，$e^*$表示敏感程度为$e$的消费者对应的效用最大的数据产品。

售出数据后商家获取的利润如公式(3-3)所示。

$$\varphi(e, v_k, p_k) = p^* - c^* \tag{3-8}$$

公式(3-8)中，φ表示售出数据后的收益，$p^*$表示售出数据的价格，$c^*$表示相应的成本。

由于每一个消费者对价值分数的敏感程度不同，也存在着多个价值级别的产品选择，因此，不能简单的通过线性的函数来描述消费者的消费行为，在本文中，采取了支付意愿（WTP）来反映消费者的自由选择行为，在数值上表现为愿意为产品支付的预期价格。

显然，敏感度不同的消费者对价值分数的需求不同。假设消费者敏感程度为$e$，与该敏感度对应的数据产品价值分数如公式(3-9)所示。

$$v_e = \frac{e}{e_{max}} v_H \tag{3-9}$$

公式(3-9)中，$v_e$表示消费者期望的数据价值分数水平，$e_{max}$为最高的敏感度，$v_H$表示最高的价值分数。

当产品的价值分数没有达到$v_e$的时候，随着价值分数的提高，用户会倾向于选择价值分数较高的版本，当价值分数明显超过$v_e$之后，支付意愿的提升速度会减缓，消费者可能由于产品价格的提升选择产品质量略低的版本，这样是更符合实际情况的。WTP 函数定义如公式(3-10)所示。

$$W(e,v)=\begin{cases} ev_e\left[1+\theta_1\left(\dfrac{v-v_e}{\theta_1 v_e}\right)^{\alpha}\right], & v\geq v_e \\ ev\left[1-\left(\dfrac{v_e-v}{(1-\theta_2)v_e}\right)^{\alpha}\right], & \theta_2 v_e\leq v<v_e \\ 0, & v<\theta_2 v_e \end{cases} \tag{3-10}$$

在公式(3-10)之中，$W(e,v)$表示当消费者敏感程度$e$，数据集的价值分数为$v$的时候，消费者的支付意愿，$\theta_1,\theta_2,\alpha$都是常数，并且取值都在 0-1 之间。

当$v<\theta_2 v_e$时，由于与消费者心理预期的价值分数$v_e$相差太多，消费者不愿意购买此产品，所以支付意愿$W$为 0；当$\theta_2 v_e\leq v<v_e$的时候，可以计算出一阶导数$\frac{\partial W}{\partial v}>0$，二阶导数$\frac{\partial^2 W}{\partial v^2}>0$，为凹函数；同样的，当$v\geq v_e$时，可以得到$\frac{\partial W}{\partial v}>0$，$\frac{\partial^2 W}{\partial v^2}<0$，为凸函数。取$\theta_1=0.1$，$\theta_2=0.5$，$\alpha=0.5$，$e=0.8$画出图像如图 3-3 所示。

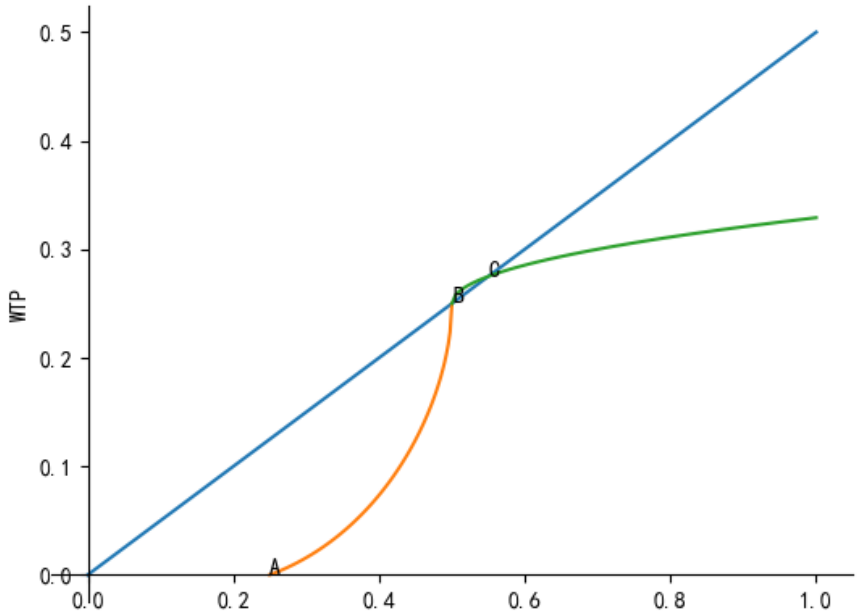


图 3-3 支付意愿(WTP)函数

图 3-3 中包含了一条直线，一条曲线，直线为$W=ev$，表示消费者的支付意愿线性增长而增长。在达到 A 点之前，即$v<\theta_2 v_e$的时候，支付意愿 WTP 的值为 0，表示消费者不愿购买价值分数过低的产品；从 A 点到 B 点之间，即$\theta_2 v_e\leq v<v_e$的时候，产品的价值分数低于消费者的心理预期($v_e$)，随着价值分数的逐渐增加，消费者的支付意愿也会增加，并且，相比于线性增加的速率，非线性的情况下，在价值分数接近于消费者预期的值($v_e$)的时候，速率会有显著的提升；在图像经过 B 点之后，即$v\geq v_e$的情况下，随着价值分数再次提高，支付意愿增加的速率开始减缓，因为对于消费者来说，产品的功能开始过剩，价值分数低的产品也可以满足自己的需求，购买的意愿逐渐减弱。

显然，相比于简单的线性函数，上述提出的分段函数更贴合于实际情况，符合消费者的心理以及数据产品的特性，能够更好的拟合消费者在面对价值分数不同的多个版本的情况下自我选择的情景。

## 第四节 双层定价模型的结构

结合上文的假设以及定价策略的讨论，为了定量的分析定价策略的合理性，需要建立相应的数学模型。同时考虑数据的版本划分以及消费者的自由选择行为，将定价模型分为两层，如图 3-4 所示。

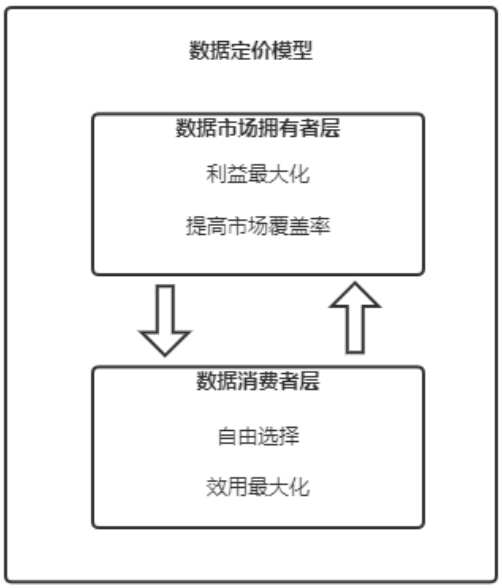


图 3-4 数据定价模型

在图 3-4 中，第一层最大化数据市场拥有者的利润，也是模型的目的函数，在传统的策略上，考虑成本是商品定价的唯一原则，尤其是数字商品，但是实际上，只考虑成本是一种常见的缺陷，成本应该只是合理定价的一个因素，一个成熟的定价策略应该是实现商家的利润最大化，而不是单纯的降低成本。模型的第二层用来表示消费者在市场中的行为，每一个消费者有着独特的个性，消费者在市场上流通的数据产品中选择对自己效用最大的版本，因此，消费者的自由选择直接影响到数据产品的销售情况，从而决定了数据市场拥有者的利润。通过模型两层之间的互相影响，以利润的大小反应当前定价策略的优劣性，从而对定价策略进行调整以得到该种情况下的最优解。

作为垄断者，需要决定具体的定价策略，提供具体的版本数量，以及每个版本所对应的价格，以此来达到利润的最大化。假设市场上总共有$M$个消费者，以敏感度来区分用户，即消费者的集合为$\phi = \{e_1, e_2, e_3, \dots, e_M\}$，产品的集合为$V = \{v_1, v_2, v_3, \dots, v_K\}$，将垄断者的策略公式化，也就是定价模型的第一层，如公式(3-11)

所示，相关约束条件如公式(3-12)、(3-13)以及(3-14)所示。

$$\max_{\{v_i,p_i\}} \varphi(v,p,x) = \sum_{j=1}^{M}\sum_{i=1}^{K}(p_i - c_i)x_{ij} \tag{3-11}$$

$$v_1 < v_2 < v_3 < \cdots < v_K \tag{3-12}$$

$$p_1 < p_2 < p_3 < \cdots < p_K \tag{3-13}$$

$$x_{ij} = 0 \; or \; x_{ij} = 1 \tag{3-14}$$

在公式(3-11)中，$\varphi$表示当前利润，$v_i$表示第$i$个版本的价值分数，$p_i$表示第$i$个版本的销售价格，$c_i$表示第$i$个版本的成本，$x_{ij}$表示消费者选择结果。

模型的最终的目的是实现利润最大化，产品的售价总是高于成本，每一件商品的价值分数及其对应的销售价格$\{v_i, p_i\}$是决策变量，公式(3-12)表示不同版本的产品，在价值分数之间会存在着差异，并且序号越大的版本价值分数越高。公式(3-13)表示随着产品价值分数的提升，该版本的售价也随之增加。$x_{ij}$表示消费者选择结果，取值只能是 0 或者 1，将在第二层进行详细介绍。

第二层考虑消费者的自由选择行为，在市场中的每一个消费者都有自己的偏好和兴趣，对于消费者来说，购买产品的动力是这件产品具有一定的效用，如公式(3-15)所示。

$$U(e_j, v_i, p_i) = W(e_j, v_i) - p_i \tag{3-15}$$

在公式(3-15)中，$U$表示产品对于消费者的效用，$e_j$表示第$j$个消费者对价值分数的敏感程度，$W$为支付意愿。

$W(e_j, v_i)$的计算方式如公式(3-10)所示，产品的效用为消费者的支付意愿与产品售价之间的差额，价值分数不同的产品对于消费者的效用也会不同，而消费者的目的为效用最大化，对于每一位消费者来说，可以用公式(3-16)来表示。

$$\max_{\{x_{ij}\}} \mu_j(x) = \sum_{i=1}^{K} x_{ij} U(e_j, v_i, p_i) \tag{3-16}$$

$$x_{ij} = 0 \; or \; x_{ij} = 1 \tag{3-17}$$

$$x_{i_1 j} x_{i_2 j} = 0, \; i_1 \neq i_2 \tag{3-18}$$

$$x_{ij} U(e_j, v_i, p_i) \geq 0 \tag{3-19}$$

在公式(3-16)之中，$\mu_j$表示消费者的总效用。

公式(3-17)表示$x_{ij}$的取值为 0 或 1，值为 0 表示消费者$j$不会购买产品$i$，相反，如果值为 1 的话，表示会购买。公式(3-18)表示每个顾客最多购买一个版本的产品，也就是对于确定的$j$，$i$取$[1, K]$任意值，$x_{ij}$的值要么都为零，要么只有一个值为 1。公式(3-19)表示消费者只会选择有效用的产品。

上述的模型不仅考虑了传统市场上的利润最大化的目标，还考虑了消费者的自我选择行为，具有很强的通用性。数据很大部分的价值在于使用上，因此，同样的数据

对于不同的消费者来说在效用上可能相差甚远，为了实现标准的定价模式，避免价格歧视，采用基于价值分数的定价模式是相对公平的。

由于在具有异构客户基于行为的估值函数的市场中，导出最优版本控制策略的解析解是不可行的，因此使用数值计算来评估版本控制策略的最优性以及多个版本的价值分数水平和价格的最佳位置。在上述模型中，消费者的选择起到了决定性的作用，商家根据选择结果来生产对应版本的产品并且制定最为合理的价格，以此来确保自己的产品在消费者群体中的高覆盖度。

# 第四章　模型的求解与结果分析

在上文中，构建了描述数据市场的模型，包括了数据市场的具体结构，参与角色以及各方利益分配的方案和定价策略，数据市场模型的关键在于数据定价策略的实现。

定价模型的第一层用来描述数据市场所有者的目的，即利润的最大化，而利润取决于消费者的购买选择，所以利润在一定程度上也代表了市场覆盖率。第二层模拟了消费者的自我选择行为，不同的消费者在多个版本的商品之间进行选择，由此来影响数据市场所有者的定价策略。由于在每一层中，都有着各自的决策变量以及约束条件，该模型属于典型的双层规划问题。在 Jeroslow 等人的研究中，认为线性双层规划问题是 NP-hard 的[59]，这种情况下采用标准的解析算法或者启发式的算法来求解此模型是非常困难的。在目前的研究中，主要有五种常用的方式用来解决此类的问题，分别是极值点搜索法、下降法、Karush-KushTucker 法、直接搜索法以及一些非数值优化方法，例如遗传算法、模拟退火算法等。在本文中，消费者的数量以及商品版本的数量的有限的，存在着最优解，由于模型涉及到了非线性的效用函数，因此采用了遗传算法对模型进行求解，并对传统的算法进行了改进，在下文中，将用具体的数值模拟定价的过程，并对结果进行分析。

## 第一节　子代择优遗传算法

遗传算法是一种基于启发式的方法，用于解决在短时间内无法解决的问题，例如一些经典的 NP-Hard 问题.遗传算法是在整体的过程上，目的是模拟在自然界中生物的演化过程。达尔文提出的生物进化理论可以解释自然界中每一种生物为什么会处于当前的形态以及为什么会具有一些功能性器官。在自然界中，随着环境的变化，只有那些更加适应当前的环境的个体，更加容易将自己的染色体遗传下去。我们通过计算机模拟的过程，就是从上帝的角度对结果进行自然选择的过程。

对于传统的遗传算法来说，整个思想是从生物自然选择的过程得到的启发，在大自然中，优势的基因总会遗传下去，这被称为优胜劣汰。对于具体的数值算法来讲，就是不合适的解会被淘汰，而更好的解将会生存下来并且将某些数值特征传承下去，这就相当于生物界中的繁衍。在我们的模型之中，首先需要确定数据产品的初始状态，以此作为初始的种群，整个过程可以划分为选择、交叉以及变异几个部分，流程如图 4-1 所示。

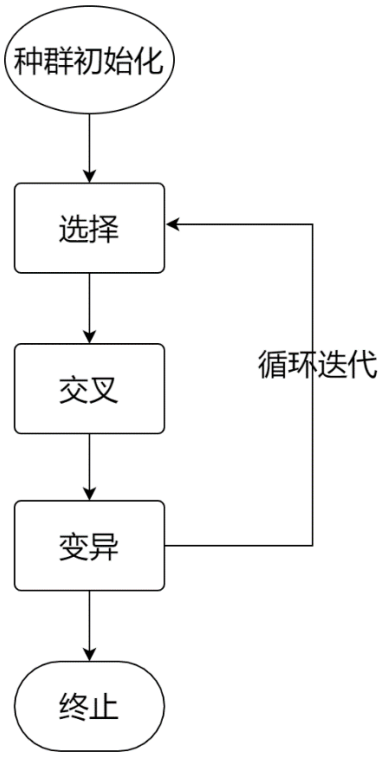


图 4-1 遗传算法的流程

遗传算法经常被用来解决复杂的数值问题，Alexouda 利用遗传算法解决了替代产品设计中的选择份额问题[60]，并且实验结果证明，进化方法在寻找更优的解方面具有优势。Jiao 等人在解决产品组合规划问题时[61]，也用到了遗传算法。本文建立了一个基于非线性函数的双层模型，在之间求解上比较困难，采用了遗传算法来逼近最优解。

由于在本文中，作为决策变量的价值分数与价格具有很强的相关性，这使得组成向量的值之间具有很强的上位性[62]，算法容易陷入局部最优解之中。为了提升算法的性能，本文对传统的遗传算法进行了改良，在交叉完成阶段产生了子代后，对每一个子代进行调整比较，选择最优的子代进入新种群，并将这种算法称为子代择优遗传算法。

假设通过交叉产生的子代$x=\{a_1,a_2,a_3,\dots,a_n\}$，适应度函数为$f(x)$，本文中的适应度函数为商家的利润最大化。为了找到更优秀的子代，对$x$中的每一个基因从左到右进行微调，$\sigma$表示调整的尺度，设置为合理的范围，经过第一次调整后，$x$产生了两个后代$x_1=\{a_1+\sigma,a_2,a_3,\dots,a_n\}$，$x_2=\{a_1-\sigma,a_2,a_3,\dots,a_n\}$，通过比较$f(x)$、$f(x_1)$以及 $f(x_2)$的大小，取最大的一个替换掉$x$，上述做法可以重复多次。在对每一个基因进行调整选择之后，留下的子代会更加的逼近最优解，提高了算法的优化能力，其过程如表 4-1 所示。

表 4-1 子代择优算法

**算法 1**：子代择优算法

**输入**：子代$x=\{a_1,a_2,a_3,\dots,a_n\}$;每次调整的最大限度$\sigma_{max}$;最大循环次数为$ln$

**输出**：择优调整后的子代$x$

设置循环次数记录$c=0$
***while*** $c<ln$ ***do***
  $i=1$
  ***for*** $i\le n$ ***do***
    设置调整的尺度$\sigma=random(0,\sigma_{max})$
    计算$x_1=\{a_1,a_i+\sigma,a_3,\dots,a_n\}$，$x_2=\{a_1,a_i-\sigma,a_3,\dots,a_n\}$
    计算$f(x),f(x_1),f(x_2)$
    取最大值$f(x_{max})=max\ (f(x),f(x_1),f(x_2))$
    $x=x_{max}$
    $i=i+1$
  $c=c+1$
***end while***
***return*** $x$

## 第二节　模型求解的关键步骤说明

### 一、种群初始化

由于模型无法采取传统的计算方式来得到结果，并且第二层的结果会影响第一层计算得到的值，因此采用的是寻优的思想，先确定一组初始解，然后对其进行优化，以此来逼近模型的最优解。

种群就是一组解，在最后的一次迭代之中，得到的种群里面最优秀的个体就是最终需要的结果。在本模型中，需要求解的为每一个版本的数据产品的价值分数以及其对应的价格，即$r=[(v_1,p_1),(v_2,p_2),\dots,(v_K,p_K)],K\in(1,2,3,4,5,6,7,8,9)$，种群中的每一个个体都是上述形式，当我们规定种群大小为$M$时，初始种群可以表示为$P=[r_1,r_2,\dots,r_M]$。我们以版本数量为 4 的情况下为例来进行说明。

当$K=4$的时候，$r=[(v_1,p_1),(v_2,p_2),(v_3,p_3),(v_4,p_4)]$，在算法的开始，需要确定价值分数以及价格的具体数值，两者都分布于 0 到 1 之间。在本文的算法中，采用了随机产生的方式，为了保持较高的市场覆盖率，价值分数一般会趋向于等区间的分布，并且在价值分数小于 0.1 的时候，消费者几乎不可能会选择该产品，由图 3-2 可知，当价值分数较小的时候，消费者只会为对应的产品出一个较低的价格。此外，即使有消费者购买了此商品，为商家带来的收益也是微乎其微的，因此，在产生随机的价值分数时，限制区间为 0.1-1，这样设置不会产生过低的价值分数，更为合理，而且

还能加快算法的寻优过程。

在设置好区间之后，在 0.1-1 之间随机产生$K$（版本数）个数，由公式(3-10)可知，价值分数有一定的大小关系，所以对其从小到大排序，以此对应数据产品的价值分数$v_1, v_2, v_3, \dots, v_K$。同样的，对于数据的价格来讲，只需要在 0-1 之间随机产生$K$个数，排序之后对应$p_1, p_2, p_3, \dots, p_K$，将价值分数与对应的价格合并起来组成一个初始种群中的个体$r = [(v_1, p_1), (v_2, p_2), \dots, (v_K, p_K)]$。确定初始种群的数量之后，重复的进行上述过程即可产生对应规模的初始种群,产生初始个体过程如表 4-2 所示。

表 4-2 产生初始个体

**算法 2**：产生初始个体

**输入**：版本数量$version_num$

**输出**：初始个体$individual$

```
while i < version_num do
    从(0,1)之间产生价值分数v
    for v ≥ 0.1 do
        将v加入价值分数的集合version中
end while
while i < version_num do
    从(0,1)之间产生价格p
    将p加入价格的集合price中
end while
将价值分数与价格分别排序
individual = [version[i], price[i]] for i in range(version_num)
return individual
```

## 二、选择

对于生物来讲，大自然会对每一个个体进行选择，影响选择结果的关键因素有气候、天敌以及自然灾害等等，顺利存活下来的个体才能继续繁衍下去。在天气炎热时，种群之中的耐高温个体可以存活下来，此时，进行选择的衡量标准为耐高温的程度。在本文中，我们的目标函数为利润最大化，计算每一个个体产生的利润之后，只有利润相对较大的会被留下来继续进行后续的步骤。

为了使种群之中优秀的特征能够在一次迭代之后得以保留，在进行选择时，会直接保留一部分个体，把种群中不会面临淘汰风险的个体称为“强者”，而剩下的全部个体都会进行选择的过程，并且有一定的几率会被淘汰掉，不会进入到下一次的迭代。因此在具体的实验之中需要设置种群中“强者”的比例，例如当我们定义利润排名前30%的个体为强者，假设种群的规模大小为 9，那么排名前三的个体将不会面临选择，直接存活下来，而剩下的 6 个个体将会被依次选择，如果将淘汰的几率设置为 0.5，

那么每一个个体都有 50%的可能性被直接淘汰掉。选择的过程是优化的关键步骤，剔除掉不合适的个体，保留强势的个体，在经历过每一次迭代之后，都会更加接近于最优解。

在本文涉及的实验中，选择过程所依据的标准为利润最大化，也就是遗传算法的适应度函数。而消费者的选择决定了商家的定价策略，因为商家不会生产没有消费者愿意购买的产品。消费者的支付意愿计算如表 4-3 所示。

表 4-3 计算消费者支付意愿

**算法 3**：计算消费者支付意愿

**输入**：价值分数$value$，价值分数敏感程度$v$，最高的价格$high_value$

**输出**：支付意愿$wtp$

设置$wtp$初始值为 0

消费者预期价值分数为$standard_value$

$standard\_value = (v/1) * high\_value$

***if*** $value < 0.5 * standard\_value$ ***do***

  $wtp = 0$

***if*** $value \geq 0.5 * standard\_value\ and\ value < standard\_value$ ***do***

  $temp = math.pow((standard\_value - value)/(0.5 * standard\_value),0.5)$

  $wtp = v * value * (1 - temp)$

***if*** $value \geq standard\_value$ ***do***

  $temp = math.pow((value - standard\_value)/(0.1 * standard\_value),0.5)$

  $wtp = v * standard\_value * (1 + 0.1 * temp)$

***return*** $wtp$

支付意愿的计算方法对应公式 3-10。在实际应用中，支付意愿就是消费者愿意为商品支付的钱，并且由消费者价值分数的敏感度决定，不同的消费者有着不同的敏感度，因此对于同一商品，消费者的购买意愿也会有所区别。支付意愿减去相应的售价之后就是商品对于消费者的效用，消费者只会选择效用最大的商品，据此可以得到在一组商品之下，市场中所有消费者的购买选择，这一过程实际上就是双层模型的第二层所起到的作用，如表 4-4 所示。

表 4-4 计算消费者选择矩阵

**算法 4**：计算消费者选择矩阵

**输入**：消费者集合$customer$，群体$individual$, 版本数量$version_num$

**输出**：消费者选择矩阵$x$

设置$flag_i, flag_j$初始值为-1
最大效用$max_utility$为 0
***for*** $i$ $in$ $range(customer\_num)$ ***do***
  ***for*** $j$ $in$ $range(version\_num)$ ***do***
    $x[i][j] = 1$
    计算效用$utility = wtp - price$
    ***if***效用大于目前的最大效用 ***do***
      记录当前的角标值$flag_i, flag_j$以及新的$max_utility$
  ***if*** 有效用最大值 ***do***
    $x[i][j] = 1$
将$flag_i, flag_j, max_utility$还原为初始值
***return*** $x$

在得到选择矩阵之后，就可以将结果代入模型的第一层之中，计算出此时的利润，后续就可以依据利润的大小在种群的个体之间进行选择，如表 4-5 所示，直接存活下来的“强者”以及幸存下来的一些个体组成了下一代种群的双亲。

表 4-5 选择算法

**算法 5**：选择算法

**输入**：消费者集合$customer$，个体$population$，存活率$retain_rate$

**输出**：下一代种群的双亲$parents$

计算个体利润$profits = total_profits(population, customer)$
按照利润进行排序得到双亲群体$origin_parents$
按照存活率$retain_rate$从$origin_parents$选择部分双亲
剩余的双亲从没被选择的个体中随机抽取
***return*** $parents$

## 三、交叉

生物学中，将两个染色体上的部分片段发生互换的过程称为交叉，在繁衍的过程中，基因会进行交叉，在交叉之后，后代可能会出现更优秀的个体，因为后代有几率继承双亲的优秀基因，这对寻优来说是很有帮助的。

进行选择之后，由于淘汰了一部分劣势个体，种群中个体的数量变少了，为了保持算法的稳定，每一次迭代产生的新种群都需要保持相同的规模，因此，交叉之后产生的新个体数量与选择过程中淘汰的个体的数量要相等。在交叉的方式上，传统的遗

传算法随机选择两个个体进行交叉的操作，在交叉片段的选择上，首先，对种群中的个体按照利润进行从大到小的排序，由于第一个个体的利润是最大的，为了稳定的保留此时种群中最为优秀的基因，此个体不参与交叉的过程，所以交叉片段从第二个个体到最后一个之间随机选择。此过程如表 4-6 所示。

表 4-6 交叉算法

**算法 6**：交叉算法

**输入**：双亲$parents$

**输出**：子代$children$

计算需要产生的子代个数$target_count$

初始化子代集合$children = [\,]$

***while*** $len(children) < target\_count$ ***do***

  从双亲群体中随机选取两个个体作为交叉对象$male, female$

  随机选取交叉片段并将双亲进行交叉操作产生子代$child_1, child_2$

  对子代进行择优

  将最优的子代添加到$children$

***end while***

***return*** $children$

确定好交叉的片段之后，将双亲的对应部分进行交换，形成两个子代。此外，在进行了交叉的过程之后，由于交换了部分数值，个体中可能不满足价值分数越大，价格越高的规定，因此每次交叉之后，需要对个体中的数据进行重新的排序。例如个体 a [(1.1,2),(2.4,2.5)]与个体 b [(1.5,1),(1.8,1.9)]进行交叉，交叉的片段为第二个基因，互换之后形成两个新个体 c[(1.1,2),(1.8,1.9)]以及 d[(1.5,1),(2.4,2.5)]，在个体 c 中，价值分数 1.8>1.1，但是价格上 1.9<2，这显然与前提假设矛盾，因此，需要进行一定的调整，将价格进行互换，调整之后的 c 变为[(1.1,1.9),(1.8,2)]。

## 四、变异

自然界中变异是指个体的染色体上基因发生突变，但是基因突变通常会带来不好的影响。在遗传算法之中，将变异这个过程引入的目的主要有两个：第一个作用是让遗传算法在寻优的过程中具有一定范围内的随机搜索能力。一次迭代在经过了选择、交叉两个过程之后，得到的种群已经接近了最优解，变异可以在交叉的基础上进行微调，如果恰好将种群中的一个弱势基因变成了优质的基因，可以加快寻优的过程。显然，变异同样有几率造成不好的影响，因此，变异的概率需要设置为一个较小的值。除此之外，变异还有一个目的是维持群体多样性，在迭代多次之后，种群容易陷入局部的最优解之中，为了防止此现象，用变异来进行适当的调整，使得最后的结果可以逼近全局的最优解。所以在遗传算法中，通常以交叉为主要步骤，变异作为辅助来相互配合。两者相互配合，但是又存在着一定的相互竞争关系。相互配合是指变异的局

部搜索能力可以在交叉的基础上，使得算法的结果更为接近于最优解。而竞争指的是经过交叉得到的优质解也可能被变异的操作所破坏。所以，需要把握好交叉与变异的方式，这样才能使算法的效率得到提升。

在本文的算法设计中，经过选择以及交叉过程产生的所有子代，在变异的过程中都有一定的概率会被调整，如表 4-7 所示。

**算法 7**：变异算法

**输入**：子代$children$，变异率$mutation_rate$

**输出**：变异后的子代

```
for i in range len(children) do
    if随机产生的值小于变异率mutation_rate do
        随机产生变异的位置，并调整该位置上对应的价值分数以及价格
        将该子代替换为变异后的子代
return children
```

在明确了遗传算法的整体架构以及一些关键步骤之后，在进行实验时，还有一些关键的参数，这些参数的具体数值如表 4-8 所示。

表 4-8 实验参数

| | |
|---|---|
| 消费者数量（customer_num） | 10000 |
| 种群规模（population_num） | 50 |
| 进化次数（iteration_time） | 1000 |
| “强者”比率（retain_rate） | 0.3 |
| 弱者存活概率（random_select_rate） | 0.5 |
| 变异率（mutation_rate） | 0.1 |

实验的操作中，数据产品的版本数量需要从 1 到 9 依次进行实验，版本数量为 1 和 2 的时候情况比较特殊，单独采取实验得到结果，实验的过程与上文介绍的过程一致，其余情况正常进行，代码部分用 python 实现。

双层模型面向一个消费者数量为 10000 的垄断市场，数据市场所有者提供完整的数据产品，并且将数据产品划分为一定的版本，消费者通过效用函数来确定购买选择，模型的目标函数为利润最大化。模型开始，构建个体数量为 50 的初始种群，每一次迭代之中，采用遗传算法进行优化得到下一代，经过 1000 次的迭代之后得到最终的种群，种群之中利润最大的个体就是模型的最优解。

将遗传算法运用于本文的模型之上，整个过程可以用表 4-9 来表示。

表 4-9 双层模型遗传算法主程序

**算法 8：**遗传算法主程序

**输入：**迭代次数$iteration_time$

**输出：**新种群$P$

设置$t=0$

初始化种群$P(t)$

***while*** $t < iteration\_time$ ***do***

  通过选择、交叉、变异产生后代$P_{offspring} = S.C.M(P(0))$

  将产生的后代与剩余的个体结合形成新的种群$P(t+1)$

  $t = t+1$

***end while***

***return*** $P(t)$

在上述过程中，计算利润首先需要计算消费者的购买选择，即通过消费者层的计算得到$x_{ij}$，对于种群中的每一个个体来讲，这一部分的步骤可以用表 4-10 来表示。

表 4-10 计算消费者购买选择矩阵

**算法 9：**计算消费者购买矩阵

**输入：**消费者总数$customer_num$,最大版本数$K$，个体$r$

**输出：**矩阵 $x_{ij}$

设置$i=1, j=1$

***while*** $i < customer_{num}$ ***do***

  ***while*** $j < K$:

    计算$WTP(e_i, v_j)$

    记录效用最大的时候的$j$值

    效用$U(e_j, v_i, p_i) = W(e_j, v_i) - p_i$

    记录$j = j+1$

  $x_{ij} = 1$

  $i = i+1$

***return*** $x_{ij}$

## 第三节 相关参数说明

商家层主要涉及到数据产品的版本数、每一个版本的价值分数、每个版本对应的价格以及成本。由于数据产品的边际成本与传统的商品存在着差异，为了更好的拟合现实中的情况，文中将把数据产品的成本分成三种情况进行考虑，分别是零边际成本、线性边际成本和非线性边际成本，具体的参数设置如表 4-11 所示。

表 4-11 商家层的主要参数

| | | | |
|---|---|---|---|
| 价值分数范围 | $v_k \in [0,1], q_H = 1$ | | |
| 价值分数 | $v_k, k = 1,2,\dots,K$ | | |
| 版本数量 | $K = 1,2,\dots,9$ | | |
| 价格 | $p_k, k = 1,2,\dots,K$ | | |
| 价格分布范围 | $p \in [0,1],\ p_H = 1$ | | |
| 产品成本分布 | 零边际成本 | 线性边际成本 | 非线性成本 |
| | $c_1 = c_2 = \cdots = c_K = 0;$ | $c_K = 0.25 * v_k, k = 1,2,\dots,K$ | $c_K = 0.25 * v_k + 0.25 * {v_k}^2, k = 1,2,\dots,K$ |

消费者层的关键参数为消费者的数量、价值分数敏感程度以及支付意愿函数的相关参数。消费者的数量设置为具体的数值，用来模拟市场中的所有潜在客户。质量敏感程度作为消费者之间的区分，所有消费者都分布在设置的敏感度区间之内，为了更好的讨论，文中将消费者的分布分成了三种情况进行讨论，分别是均匀分布、高斯分布以及指数分布，在结果上可以形成对比，增加说服力，具体的参数值如表 4-12 所示。

表 4-12 消费者层的主要参数

| | | | |
|---|---|---|---|
| 顾客分布范围 | $e \in [0,1], e_{max} = 1$ | | |
| 顾客分布类型 | $U(r_1, r_2)$ | $N(\mu, \sigma^2)$ | $Exp(\lambda)$ |
| | $r_1 = 0, r_2 = 1$ | $\mu = 0.5, \sigma = 0.1$ | $\lambda = 2$ |
| 顾客预期质量 | $v_e = \frac{e}{e_{max}} v_H$ | | |
| 支付意愿函数 | $\theta_1 = 0.1, \theta_2 = 0.5, \alpha = 0.5$ | | |
| 顾客总数 | $M = 10000$ | | |

此外，对于子代择优过程中的调整的最大限度$\sigma_{max}$取 0.1，最大循环次数$ln$取 10。

## 第四节　结果分析

目前数据市场并没有统一的定价标准，以国内的数据交易所为例，在定价模式上，最早的贵阳交易所为每一类数据制定了一个计价公式，用户上传数据之后由交易系统自动定价。上海交易所在数据定价规则的指定上考虑了三项指标：数据的成本、数据使用后的收益以及数据经过了多次交易后形成的稳定价格，以此来调整数据的价格。还有一些交易所以类似股票交易的形式来交易数据，由交易的双方自由协商制定数据的价格。因此，缺乏定价标准使得在定价策略的研究上，不能直接通过定价结果来验证价格的合理性，大部分的研究重点在于框架的制定或是计算出相对价格[63][64]。此外，由于个人以及小团队很难找到大量可供交易的数据，即使有数据资源也很难找到愿意购买数据的买家，通过建立起实际的数据市场来考察定价策略的表现是不可行的。在过往的研究中，为了验证定价策略的合理性，会对定价模型的特定指标进行检验，说明模型的合理性，这些指标包括模型的无套利、无折扣、利润最大化以及降低成本

等。利润最大化是一个常见的评估标准[65]，因此，本文在实验阶段选择通过利润的提升来说明模型的合理性。

上文中介绍的算法将被用来求解数据市场模型中最为关键的定价策略，算法对多个版本的质量水平以及对应的价格进行持续的优化，在经历过 1000 次的选择、交叉以及变异之后，输出模型的结果。实验中，讨论了消费者对价值分数敏感程度的三种分布，分别是均匀分布、高斯分布和指数分布，数据产品的成本也分为三种，分别是零边际成本、线性边际成本以及非线性边际成本，所以实验最后会得到九个结果，实验会进行多次，以确保实验结果准确。实验重点关注两个指标，一个是利润的最大化，另一项指标为市场覆盖率，通过这两项指标来分析定价策略的优劣性。下文将会对实验结果进行展示，并对结果进行分析讨论。

## 一、零边际成本下的价值分数与价格

在本小节中讨论数据产品的成本为零边际成本的情况，比较在消费者分布不同的情况下，数据产品的价值分数以及对应的价格的变化，并且可以得到在最大版本数确定的情况下，最优的定价策略。

利用遗传算法对模型进行求解，得到在最大版本数为$K = 1,2,\ldots,9$的情况下，最优的多版本定价策略，在各个版本数量之下的数据市场所有者的总利润如图 4-2 所示。

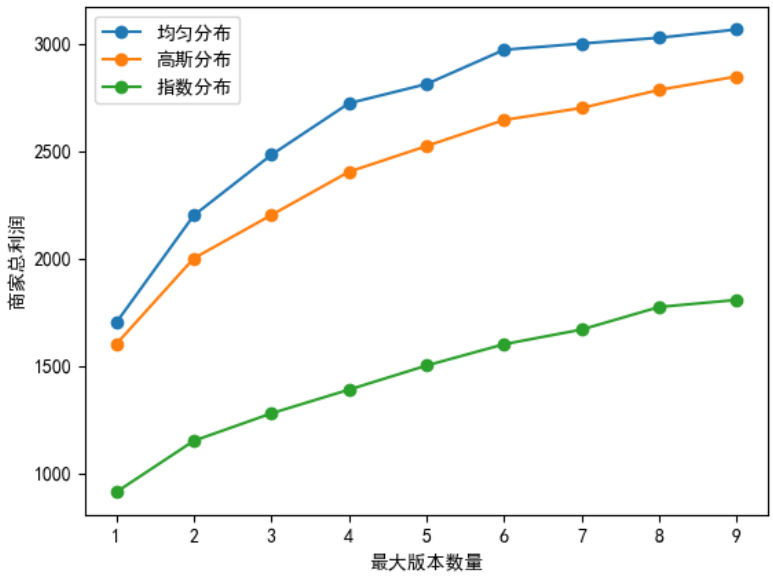


图 4-8 零边际成本下的总利润

从图 4-2 中可以看到，无论消费者对价值分数的敏感程度是何种分布，随着最大版本数量的增加，数据市场所有者的总利润也会增加，均匀分布、高斯分布以及指数分布在最大版本数量为 9 的时候，利润达到最大值 3066.4、2847.5 和 1806.6，对比利润最低的时候的值，利润分别增加了 80.4%、77.9%和 98.3%，这说明多版本策略在提高利润上起到了很大的作用。并且，可以观察到，随着最大版本数量的提升，利润增

加的速率逐渐减缓，这说明数据市场所有者添加低质量的版本带来的收益较小。

除了利润之外，在市场中，还有一个很重要的指标，即市场覆盖率。市场覆盖率越高，说明数据市场所有者采用的多版本策略被消费者认可的程度越高。图 4-3 显示了零边际成本下三种分布的市场覆盖率。

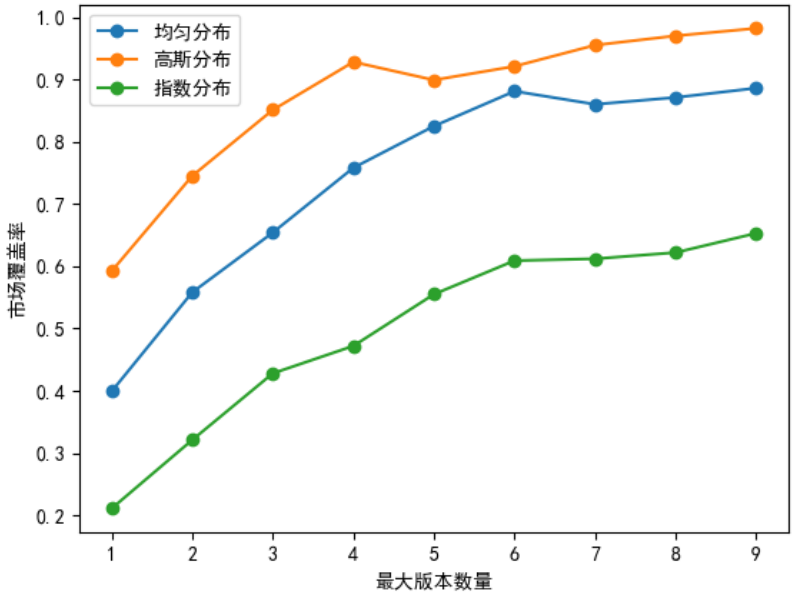


图 4-3 零边际成本下的市场覆盖率

当市场中只发行数据产品的一个版本的时候，三种分布下的市场覆盖率分别为 39.9%、59.2%和 21.1%（均匀分布、高斯分布、指数分布），而在最高的时候，市场覆盖率增长到了 88.6%、98.2%和 65.3%，可以看到，多版本相对于一个版本的时候，能为市场中更多的消费者提供服务，在消费者呈高斯分布时，数据产品的各个版本几乎覆盖到了全部的消费者。实行了多版本策略之后，市场覆盖率分别是一个版本时候的 2.22 倍、1.65 倍以及 3.09 倍，可以看到，通过双层模型找到最优解，并且将这种策略应用于市场之后，可以显著的提高数据市场所有者的利润，还可以让产品在市场中占据更高的份额。

在边际成本为零的情况下，最大版本数量不同，模型得到的最优解也会不同，图 4-4、4-5、4-6 列出了在不同的版本数量之下，每个版本的价值分数以及其对应的价格水平。

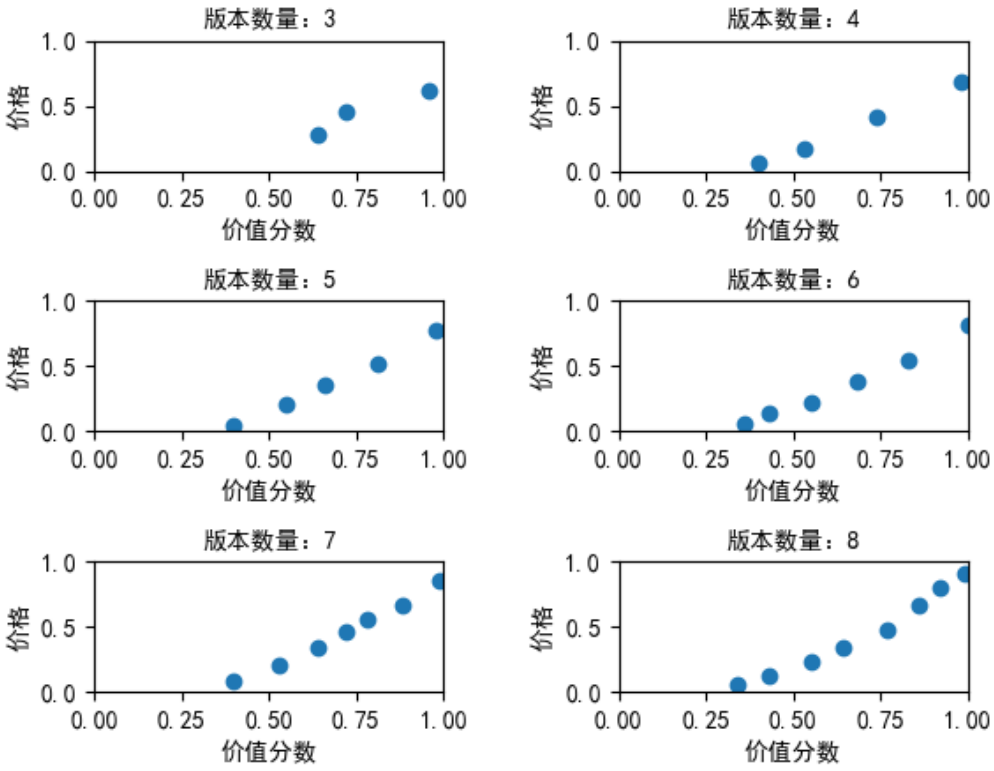


图 4-4 零边际成本下均匀分布的最优解

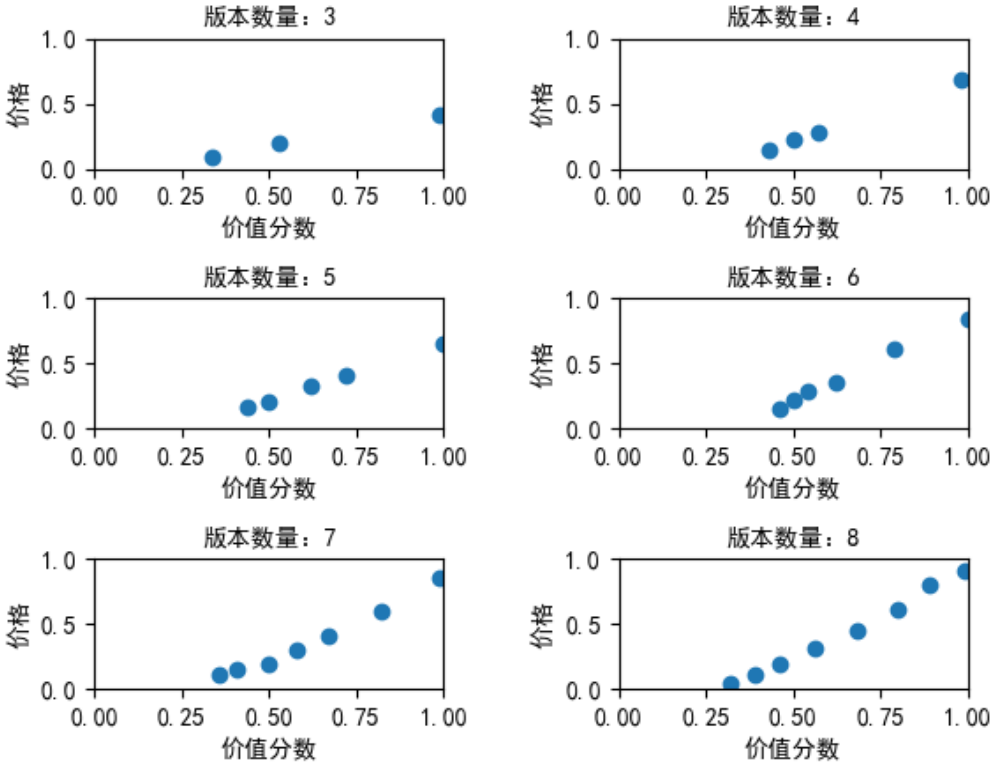


图 4-5 零边际成本下高斯分布的最优解

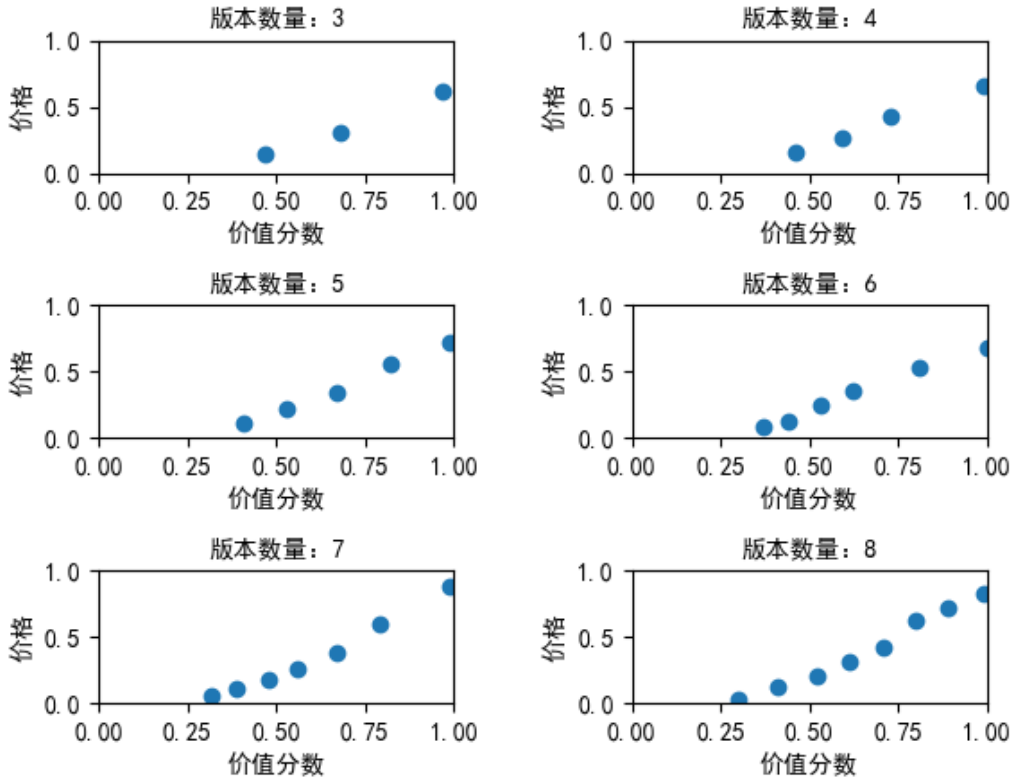


图 4-6 零边际成本下指数分布的最优解

从图 4-4、4-5 以及 4-6 中可以看到，在每个最优的方案之中，最高价值分数版本的价值分数总是接近于 1，由于消费者总是对数据产品的价值分数敏感，在提高版本数量的时候，会在一系列的产品之中添加低价值分数的版本，以迎合市场中对数据的价值不太敏感的消费者，扩大数据产品在市场的覆盖率。并且，当数据产品的新版本加入时，现有版本的价值分数水平和价格与之前的最优策略方案相比，都会有一定程度的提高，因此，当价值分数水平和价格处于最优位置时，多版本的定价方案对于数据市场所有者来说是非常有吸引力的。

## 二、线性与非线性成本下的价值分数与价格

下文将展示不同成本之下的最优解的情况，分别是零边际成本、线性边际成本以及非线性边际成本的情况，并对结果进行分析。

边际成本在实验中定义为价值分数水平的单调递增函数，价值分数越高的版本，其边际成本就越大，这必然会影响多版本方案中价值分数水平和价格的最优位置。图 4-7、4-8 及 4-9 展示了三种边际成本之下在消费者分布不同时的利润情况。

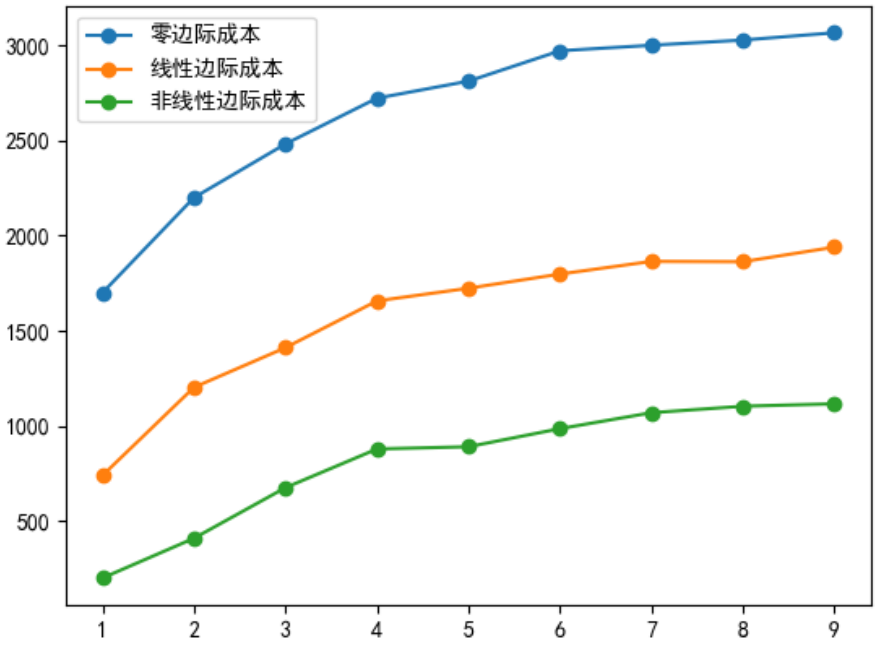


图 4-7 均匀分布下成本对利润的影响

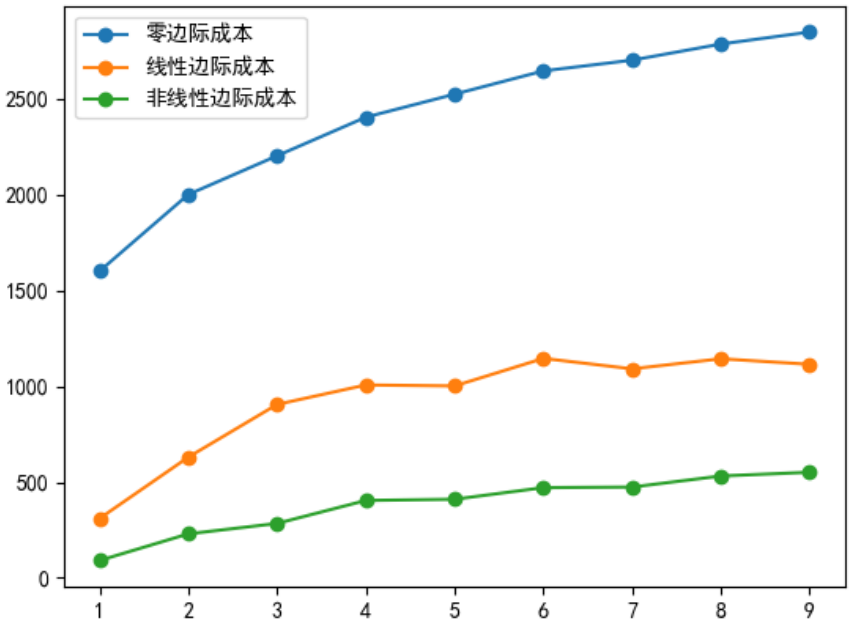


图 4-8 高斯分布下成本对利润的影响

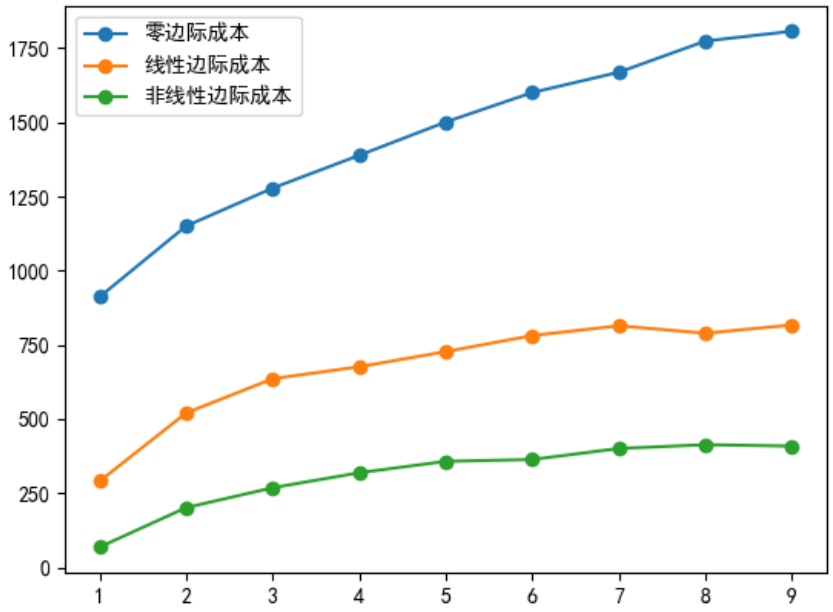


图 4-9 指数分布下成本对利润的影响

数据市场所有者的利润在总体上随着最大版本数量的增加而增加，并且，在线性成本和非线性成本下，利润相比于零边际成本会降低很多，这是由于成本的提高所造成的。图 4-10 表示消费者呈均匀分布时，最大版本数量为 4 的情况下，数据产品的价值分数以及价格水平。

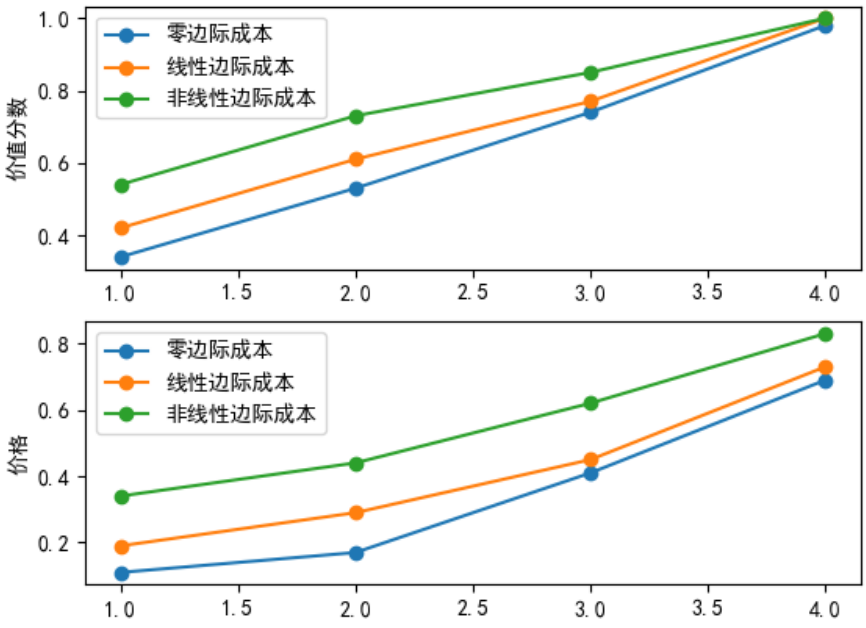


图 4-10 均匀分布下消费者分布不同时的价格水平与质量分数

从图 4-10 中可以看到，由于价值分数高的版本有更大的边际成本，数据产品的定价需要高于他们的固定边际成本，所以当边际成本函数随版本的价值分数线性或二次增加时，具有相同最大版本数的多版本方案的最优价值分数水平和最优价格均比边际成本为零的方案增加得多。此外，与零边际成本相比，线性或二次增加的边际成本导致最优多版本策略的市场覆盖率更低，如图 4-11 所示，无论是客户类型的高斯分布还是指数分布，都得到了类似的结果。

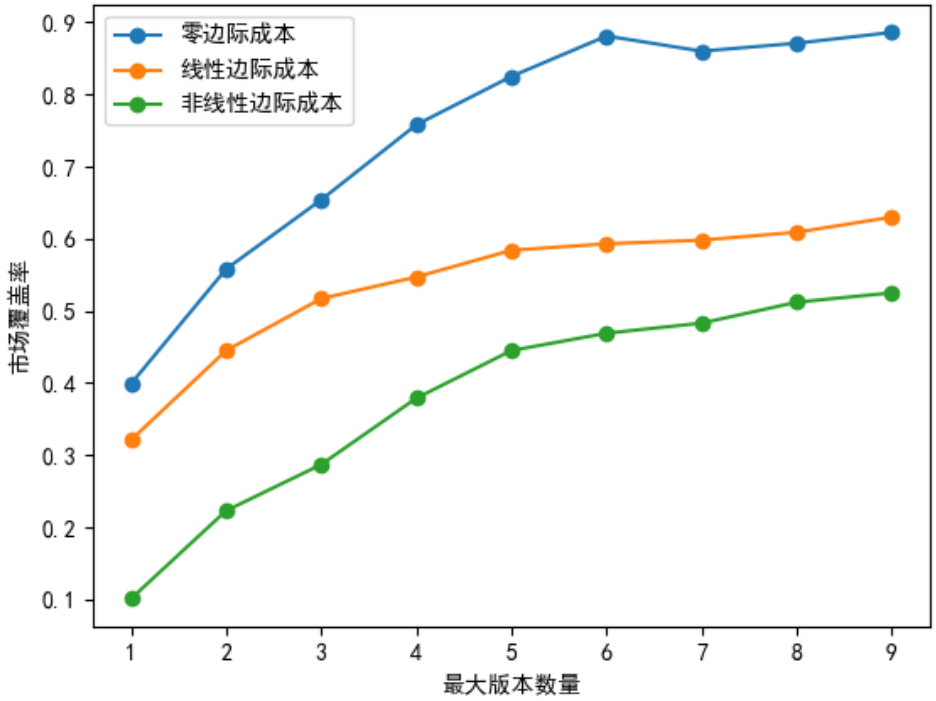


图 4-11 均匀分布下不同成本的市场覆盖率

上述的所有实验结果都表明，在垄断的数据市场中，数据市场所有者如果想要提高数据产品的利润，需要将产品进行更详细的划分，制定合理的定价策略。

## 三、实验结论

在考虑定价策略时，考虑了用户的自由选择行为，用对价值分数的敏度感来区分消费者个体，并且运用了非线性的效用函数来描述这一行为，符合本研究想要从数据的内在属性来制定价格的理论。实验的结果也说明为市场提供更多的版本可以为垄断者带来更大的回报，在现实生活中也有企业采用这样的模式，例如 Matlab，它的基础平台的价格是不变的，相当于价值分数最低的版本，当加上一些扩展工具箱（统计分析、金融工具等），相当于提供了更高价值分数的版本，同时也为这些版本制定了更高的价格，这一策略帮助 Matlab 获得了巨大的收益。

本文提出的定价模型提供了一种将数据平台垄断者和消费者的决策在两层上表示的通用方法，并通过自由选择行为连接起来。这个模型可以提供一个计算平台，以

优化数据各个版本的价值分数水平和多版本方案的价格，使垄断者的总利润最大化。例如对于上文提到的 glass 数据集，计算出价值分数后，如果想要提供多个版本进行售卖，可以通过此模型来计算相应的价值分数水平和对应的价格水平，通过对数据的处理改变价值分数，以此来面向市场上各类消费者的需求。并且，在实验中，基于具体数值的计算和结果，对于市场上不同的消费者分布，都可以得到不错的实验结果，由此，在面向现实生活中的市场时，只需要模拟出市场的实际分布，该模型也会有良好的表现。

# 第五章 总结与展望

## 第一节 总结

数据市场的出现可以促进数据在数据市场的参与多方之间的数据交易，然而，如今的市场中，现有的数据市场结构以及各项规章制度还存在着缺陷。数据交易作为一种新兴业务出现，其发展遇到了许多挑战，如数据隐私和安全和知识产权[66]。例如，一旦数据的交易过程完成，数据的所有权也可能转移给数据消费者，这些消费者可以通过转卖来赚取利益。如果没有严格的知识产权保护，数据平台所有者可能会失去对数据的控制，因为数据消费者随后可以转移、共享或出售数据。其次，允许数据消费者销售、提取或处理数据会在数据市场中产生竞争和冲突。在这种情况下，如果消费者可以直接以较低的价格出售数据副本，数据平台所有者就会失去其定价的权威与优势。如果消费者可以间接出售数据，通过提取、处理或设计将数据转化为新产品，数据平台所有者的收入将受到显著损害。因此，建立一个严格的数据市场是十分有必要的，并且在定价策略上，应该更加透明化，本文提出的基于信息熵以及数据质量的定价模型可以起到很好的借鉴作用，通过‘质’（数据质量）和‘量’（信息量）两方面对于数据价值的衡量，得出数据对应的价值分数，弥补了信息熵以及数据质量作为衡量数据价值的唯一因素时表现不良的缺陷，并以价值分数为基础来制定价格。

在大数据应用发展迅速的背景下，考虑机器学习对于数据的需求，为了使数据的流通更为便捷，提高数据资源的利用率，需要数据市场的推动作用。本文从数据收集阶段开始，将数据市场的参与者分为了三类角色，阐述了每一个角色的作用以及数据市场的运行过程。作为数据市场的所有者，负责制定相关策略，文中对最为重要的数据定价策略进行了设计。在以往的研究中，关注的重点在数据定价的具体形式，也有很多学者对这些研究做过综述[67][68]，但是数据定价的形式多种多样，没有统一的标准。对于数据定价策略，本文提出了一种新的数据市场定价方案，旨在为数据平台所有者和数据消费者提供一个有用的定价决策工具，该方案关注数据本身的价值，并考虑数据的信息熵以及多个质量维度属性，结果上不仅实现了商家利润的最大化，还考虑了数据产品对于消费者的效用问题，实现了效用的最大化，这种多版本策略可以帮助数据市场所有者在每个质量维度上细分市场，从而获得更多的利润。并且，文中提出的效用函数更准确地表征了消费者对数据产品的自我选择行为，考虑了每个消费者的标准需求，这符合商场中数据产品的内在属性。与线性效用函数导致版本策略的次最优性不同，这种新的效用函数保证了多版本策略是最优的，更多的版本可以为垄断者带来更大的回报。

本文在研究数据质量对数据价值的影响时，选取了三个最有代表性的维度，并且可以轻易的扩展至更多的维度，加入了信息熵作为衡量数据价值的因素，最终综合起

来使用价值分数作为数据定价的决定因素，使得数据定价模型更具有通用性，对于消费者更加友好，并且可以为垄断者优化多版本方案的提供了一个参考，以实现总利润的最大化，为建立起良好的数据市场做出贡献。

## 第二节　展望

本文以信息熵以及数据质量作为关键影响因素，确立了完整的数据市场结构，选取了主要的三类角色作为数据市场的相关参与者，随着数据市场理论研究的深入，未来可能会涉及到更多得参与方，数据市场结构也会因此而调整，数据市场拥有者所制定的策略以及规定也需要做同步的改进。在定价策略方面，效用函数揭示了版本控制策略的最优性取决于信息产品特定的客户评估函数，未来的研究应从网络外部性、外部需求变化、数据市场从低版本到高版本的动态升级等方面，着重研究多版本策略的特性，这些特征描述了现实生活中消费者对信息产品的估价和购买行为的更多方面，需要更加深入的研究。在信息熵方面，可以对熵的计算方式做出改良[69]，或者将熵的其他形式引入[70]。在数据质量的研究方面，数据在未来可能会进化出更多的特征，Batini 等人研究了时间维度上数据的特征[71]，将数据分为稳定、长期变化和频繁变化的数据类别，在这三种类别上数据定价策略肯定有所区别，这也是未来需要进行深入研究的一个重要课题。

# 参考文献

[1] 梁楠,李磊明.大数据技术在工业领域的应用综述[J].电子世界,2016(17):8-9.
[2] IDC 中国. 2021 年 V2 全球大数据支出指南[R].北京,2021.
[3] Van de Sandt S, Lavasa A, Dallmeier-Tiessen S, et al. submitter: The Definition of Reuse[J]. Data Science Journal, 2019,18(1):22.
[4] Allam Z, Dhunny Z A. On big data, artificial intelligence and smart cities[J]. Cities, 2019, 89: 80-91.
[5] Wang J, Xu C, Zhang J, et al. Big data analytics for intelligent manufacturing systems: A review[J]. Journal of Manufacturing Systems, 2021,62:738-752.
[6] Naeem M, Jamal T, Diaz-Martinez J, et al. Trends and future perspective challenges in big data[M]//Advances in Intelligent Data Analysis and Applications. Springer, Singapore, 2022: 309-325.
[7] 徐子伟,张陈斌,陈宗海. 大数据技术概述[C]//.系统仿真技术及其应用学术论文集（第 15 卷）.
[8] 康旗,韩勇,陈文静,刘亚琪.大数据资产化[J].信息通信技术,2015,9(06):29-35.
[9] 罗曼,田牧. 理想很丰满现实很骨感　贵阳大数据交易所这六年[N]. 证券时报,2021-07-12(A01).
[10] Akerlof G A. The market for “lemons”: Quality uncertainty and the market mechanism[M]//Uncertainty in economics. Academic Press, 1978: 235-251.
[11] Armstrong A A, Durfee E H. Mixing and memory: Emergent cooperation in an information marketplace[C]//Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160). IEEE, 1998: 34-41.
[12] Muschalle A, Stahl F, Löser A, et al. Pricing approaches for data markets[C]//International workshop on business intelligence for the real-time enterprise. Springer, Berlin, Heidelberg, 2012:129-144.
[13] Kantere V, Dash D, Gratsias G, et al. Predicting cost amortization for query services[C]//Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011: 325-336.
[14] Balazinska M, Howe B, Suciu D. Data markets in the cloud: An opportunity for the database community[J]. Proceedings of the VLDB Endowment, 2011, 4(12): 1482-1485.
[15] Koutris P, Upadhyaya P, Balazinska M, et al. Toward practical query pricing with querymarket[C]//proceedings of the 2013 ACM SIGMOD international conference on

management of data. 2013: 613-624.
[16] Koutris P, Upadhyaya P, Balazinska M, et al. Query-based data pricing[J]. Journal of the ACM (JACM), 2015, 62(5): 1-44.
[17] Zheng Z, Peng Y, Wu F, et al. An online pricing mechanism for mobile crowdsensing data markets[C]//Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing. 2017: 1-10.
[18] Henderson S, Peirson G, Herbohn K, et al. Issues in financial accounting[M]. Pearson Higher Education AU, 2015.
[19] 王婷婷. 基于拍卖理论的大数据交易定价策略研究[D].云南财经大学,2019.
[20] Yaïche H, Mazumdar R R, Rosenberg C. A game theoretic framework for bandwidth allocation and pricing in broadband networks[J]. IEEE/ACM transactions on networking, 2000, 8(5): 667-678.
[21] Niyato D, Hoang D T, Luong N C, et al. Smart data pricing models for the internet of things: a bundling strategy approach[J]. IEEE Network, 2016, 30(2): 18-25.
[22] Ballou D P, Pazer H L. Modeling data and process quality in multi-input, multi-output information systems[J]. Management science, 1985, 31(2): 150-162.
[23] Pipino L L, Lee Y W, Wang R Y. Data quality assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.
[24] Wang R Y, Strong D M. Beyond accuracy: What data quality means to data consumers[J]. Journal of management information systems, 1996, 12(4): 5-33.
[25] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement[J]. ACM computing surveys (CSUR), 2009, 41(3): 1-52.
[26] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2):6.
[27] Heckman J R, Boehmer E L, Peters E H, et al. A pricing model for data markets. iConference 2015 Proceedings, 2015.
[28] Yu H, Zhang M. Data pricing strategy based on data quality[J]. Computers & Industrial Engineering, 2017, 112: 1-10.
[29] Shannon, Claude E. A Mathematical Theory of Communication. Bell System Technical Journal. July 1948, 27 (3): 379–423.
[30] Holzinger A, Hörtenhuber M, Mayer C, et al. On entropy-based data mining[M]//Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer, Berlin, Heidelberg, 2014: 209-226.
[31] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization[J]. Advances in neural information processing systems, 2004, 17.
[32] Aurelio Y S, de Almeida G M, de Castro C L, et al. Learning from imbalanced data sets with weighted cross-entropy function[J]. Neural processing letters, 2019, 50(2):

1937-1949.
[33] 姚建国，李希君，管海兵.基于熵的数据价值衡量与定价方法[P].中国专利:106815743A,2017.06.09.
[34] Li X, Yao J, Liu X, et al. A first look at information entropy-based data pricing[C]//2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 2053-2060.
[35] 李希君. 基于信息熵的数据交易定价研究[D].上海交通大学,2018.
[36] Shen Y, Guo B, Shen Y, et al. Pricing personal data based on Information Entropy[C]//Proceedings of the 2nd International Conference on Software Engineering and Information Management. 2019: 143-146.
[37] Kim G H, Trimi S, Chung J H. Big-data applications in the government sector[J]. Communications of the ACM, 2014, 57(3): 78-85.
[38] J. Lin, W. Y u, N. Zhang, X. Yang, H. Zhang, and W. Zhao, ‘‘A survey on Internet of things: Architecture, enabling technologies, security and privacy, and applications,’’IEEE Internet Things J., vol. 4, no. 5,pp. 1125–1142, Oct. 2017.
[39] J. A. Stankovic, ‘‘Research directions for the Internet of Things,’’IEEE Internet Things J., vol. 1, no. 1, pp. 3–9, Feb. 2014.
[40] G. Xu, W. Y u, D. Griffith, N. Golmie, and P. Moulema, ‘‘Toward integrating distributed energy resources and storage devices in smart grid,’’IEEE
[41] J. Lin, W. Y u, and X. Yang, ‘‘Towards multistep electricity prices in smart grid electricity markets,’’IEEE Trans. Parallel Distrib. Syst., vol. 27,no. 1, pp. 286–302, Jan. 2016.
[42] J. Lin, W. Y u, X. Yang, Q. Yang, X. Fu, and W. Zhao, ‘‘A real-time enroute route guidance decision scheme for transportation-based cyberphysical systems,’’IEEE Trans. V eh. Technol., vol. 66, no. 3, pp. 2551–2566,Mar. 2017
[43] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction apis[C]//25th {USENIX} Security Symposium ({USENIX} Security 16). 2016: 601-618.
[44] Louis C (2020) Roundup of machine learning forecasts and market estimates, 2020. Forbes URL https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machinelearning-forecasts-and-market-estimates-2020, accessed: 2021-06-28
[45] Vomfell L, Stahl F, Schomm F, et al. A classification framework for data marketplaces[R]. ERCIS Working Thesis, 2015.
[46] ISO/IEC. Information Technology-Vocabulary, Online Browsing Platform. https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en:en.
[47] Pei J. A Survey on Data Pricing: from Economics to Data Science[J]. IEEE Transactions on Knowledge & Data Engineering, 2020 (01): 1-1.

[48] Akerlof G A. The market for “lemons”: Quality uncertainty and the market mechanism[M]//Uncertainty in economics. Academic Press, 1978: 235-251.

[49] Shapiro C , Varian H R . A strategic guide to the network economy. 1999.

[50] 王小鸥, 李琳. 计算机网络中 Markov 信息熵的证明[J]. 信息网络安全, 2012 (5): 7-9.

[51] Gao, S., Liu, Y., Wang, Y., Ma, X., 2013. Discovering spatial interaction communities from mobile phone data. Trans. GIS 17, 463–481

[52] 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,p p .1 – 6 ,Kuala Lumpur, Malaysia, 2016.

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

[54] F. Stahl and G. Vossen, “Fair knapsack pricing for data market-places,” in Advances in Databases and Information Systems, vol .9809, pp. 46–59, 2016.

[55] M. Mussa and S. Rosen, ‘‘Monopoly and product quality,’’J. Econ.Theory, vol. 18, no. 2, pp. 301–317, 1978.

[56] Chen Y J, Seshadri S. Product development and pricing strategy for information goods under heterogeneous outside opportunities[J]. Information Systems Research, 2007, 18(2): 150-172.

[57] Sundararajan A. Nonlinear pricing of information goods[J]. Management science, 2004, 50(12): 1660-1673.

[58] Krishnan V, Zhu W. Designing a family of development-intensive products[J]. Management science, 2006, 52(6): 813-825.

[59] Jeroslow R G . The polynomial hierarchy and a simple model for competitive analysis[J]. Mathematical Programming, 1985, 32(2):146-164.

[60] Alexouda G. An evolutionary algorithm approach to the share of choices problem in the product line design[J]. Computers & Operations Research, 2004, 31(13): 2215-2229.

[61] Jiao J R, Zhang Y, Wang Y. A heuristic genetic algorithm for product portfolio planning[J]. Computers & Operations Research, 2007, 34(6): 1777-1799.

[62] Li M, Kou J. Crowding with nearest neighbors replacement for multiple species niching and building blocks preservation in binary multimodal functions optimization[J]. Journal of Heuristics, 2008, 14(3): 243-270.

[63] Zhang M, Arafa A, Huang J, et al. Pricing fresh data[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(5): 1211-1225.

[64] Li C, Miklau G. Pricing Aggregate Queries in a Data Marketplace[C]//WebDB. 2012: 19-24.

[65] Gao L, Iosifidis G, Huang J, et al. Hybrid data pricing for network-assisted user-provided connectivity[C]//IEEE INFOCOM 2014-IEEE Conference on Computer Communications. IEEE, 2014: 682-690.

[66] BjrnLundqvist, BjrnLundqvist, BjrnLundqvist, et al. Big Data, Open Data, Privacy Regulations, Intellectual Property and Competition Law in an Internet-of-Things World: The Issue of Accessing Data. 2018.

[67] Zhang M, Beltrán F. A survey of data pricing methods[J]. Available at SSRN 3609120, 2020,30(6):21.

[68] Liang F, Yu W, An D, et al. A survey on big data market: Pricing, trading and protection[J]. Ieee Access, 2018, 6: 15132-15154.

[69] Kapur J N, Kesavan H K. Entropy optimization principles and their applications[M]//Entropy and energy dissipation in water resources. Springer, Dordrecht, 1992: 3-20.

[70] Mao J, Yao D, Wang C. A novel cross-entropy and entropy measures of IFSs and their applications[J]. Knowledge-Based Systems, 2013, 48: 37-45.

[71] Batini, C., & Scannapieco,M. (2016).Erratum to: Data and information quality : Dimensions, principles and techniques data and information quality. Springer, pp.E1–E1.