DOI：10. 11830/ISSN. 1000-5013. 202404034

机器学习模型交易中的数据购买量与模型定价

林苗君[1]，羊梓敏[2]，陈斌[2]

（1. 华侨大学 财务处，福建 泉州 362021；
2. 华侨大学 数学科学学院，福建 泉州 362021）

摘要： 基于数据边界的 Shapley 值总和构建成本分配问题，采用截断蒙特卡洛的快速算法计算 Shapley 值，证明了数据边界最优解存在性。针对不同版本模型定价场景，定义模型经纪人收入最大化模型的定价问题，并将收入最大化模型的定价问题转换为等价整数线性规划问题。运用公共数据集数值验证文中方法的正确性，同时与已有的 4 种方法进行实验对比。实验结果表明：文中方法可以提高模型经纪人收入和模型买方的购买比例。

关键词： 数据定价；模型定价；Shapley 值；整数规划

中图分类号： TP 274；F 49　　**文献标志码：** A　　**文章编号：** 1000-5013(2025)01-0095-09

Data Purchase Volume and Model Pricing in Machine Learning Model Transactions

LIN Miaojun[1], YANG Zimin[2], CHEN Bin[2]

(1. Financial Department, Huaqiao University, Quanzhou 362021, China;
2. School of Mathematical Science, Huaqiao University, Quanzhou 362021, China)

Abstract: Cost allocation problem based on Shapley value summation on data boundaries is constructed. Using the fast algrithm of the Truncated Monte Carlo, the existence of the optimal solution of the data boundary is proved. Aiming at pricing scenarios of different versions of the models, the pricing problem of the income maximizing model of the model broker is defined, and the pricing problem of the income maximizing model is transformed into an equivalent integer linear programming problem. By public datasets , our proposed method is validated its correctness, and the experiment is compared with four existing methods, simultaneously. The experimental results show that the proposed method can increase the income of the model broker and the purchase ratio of the model buyer.

Keywords: data pricing; model pricing; Shapley value; integer programming

在大数据时代，数据的流通和共享已是大势所趋，数据定价与交易方法亦受到了广泛关注[1]。随着数据市场的发展，交易的范畴超越了单纯的原始数据买卖，涵盖了机器学习模型的交易。这一拓展使得据分析能力有限的小微企业和个人用户能够便捷地获取模型服务成果，而不是直接购入并处理原始数

收稿日期： 2024-04-11

通信作者： 陈斌(1984-)，男，副教授，博士，主要从事运筹学与控制论的研究。E-mail：chenbinmath@163. com。

基金项目： 国家自然科学基金资助项目(12071165)；福建省自然科学基金资助项目(2023J01124)；中央高校基本科研业务费专项资金资助(ZQN-1102)

据。因此,模型定价成为数据定价研究的重要组成部分[2]。

在数据市场的研究中,将数据视为商品并对其进行交易与定价的策略展现出丰富多样的面貌。基于数据定价的数据集(如Dawex,Twitter,Bloomberg,Iota和SafeGraph等)允许买家直接访问数据,允许买家根据数据的特征(信息熵水平[3-4]、数据质量[5-6]、新鲜度和实时性[7-8]及数据集的大小[9])进行估值。此外,竞争性数据交易模型[10]巧妙融合了数据供应商的售卖意向与模型需求方的支付意愿的纳什均衡,从而确保各参与者的利润最优化。尽管如此,这些研究仍具有一定局限性,主要体现在数据价值评估效率的低维度依赖。基于此,本文对机器学习模型交易中的数据购买量与模型定价进行研究。

1 市场交易框架

三方交易系统中,数据提供者将数据记录提交给专业的模型经纪人。模型经纪人扮演着中介与增值的角色,运用先进的机器学习算法对数据进行深度训练,从而打造高性能的预测或分析模型。随后,精心训练出来的模型以服务的形式被精准对接至有特定需求的模型买方,实现数据价值转化与商业化。三方模型市场交易框架,如图1所示。

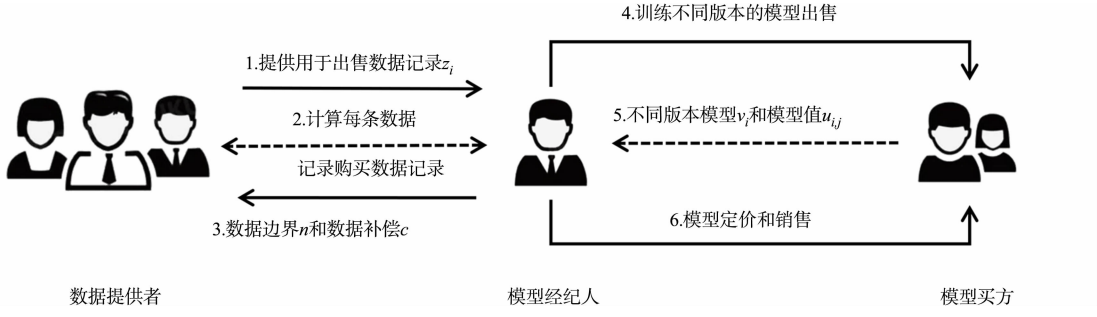

图1 三方模型市场交易框架

Fig. 1 Three-parties model market transaction framework

1) 数据提供者。在数据供应链的初始环节,数据提供者提供经过预处理(如数据脱敏)的高质量数据集D,确保所有数据均源自合法途径,不含侵犯隐私内容,并且满足市场流通的法定标准。这一环节的合法性与合规性对于维护整个数据市场的健康秩序至关重要。

2) 模型经纪人。模型经纪人扮演着核心中介角色,通过收购数据集D,利用先进的技术手段对其进行深度挖掘,训练出多样化精度层次的机器学习模型。模型经纪人的成功取决于精准平衡成本控制与市场需求,通过高效数据利用和差异化模型服务最大化经济回报。

3) 模型买方。模型买方作为数据交易生态系统的终端用户,根据业务需求和预算限制,选择购买或订阅满足精度需求的模型服务。模型买方的选择偏好和反馈信息是推动模型迭代升级和市场动态调整的关键驱动力。

模型经纪人在市场初期面临的核心挑战可归结为成本效益的最大化,具体体现两方面的策略制定:一是高效筛选并购买最具性价比的数据集;二是精准定位市场,推出多样化精度层次的模型,并设定合理价格体系,以覆盖更广泛的用户需求。

2 最优数据边界的求解

当前数据市场的定价机制往往比较粗放,用户在交易中不得不承担购买完整数据集的成本,实际上,买家仅需利用其中一部分数据实现特定目标。这种定价机制忽略数据使用效用与购买成本之间的精细平衡,未能充分体现数据价值的差异化和用户的实际需求。在数据交易的决策过程中,过度累积数据并不是最优解,在数据规模与实际效用之间寻求最佳平衡点才是最优解[11]。只有当边际收益高于边际成本时,才有必要购买数据;而相应的购买数据量称为数据边界n。在这个场景中,模型经纪人的首要任务是深入分析数据集的特性,识别最具预测价值或能显著提升模型性能的数据子集。

数据提供者搜集并整理数据记录,通过一系列预处理步骤(如数据清洗、脱敏处理等)确保数据质量

与合规性。完成这些准备后，数据被整合成具有特定主题或应用场景的数据集 D，通过计算 Shapley 值，模型经纪人选择一个最优的数据边界 n。在确定了所需数据后，模型经纪人利用选购的数据集训练多种机器学习模型，这些模型覆盖不同的复杂度和精度，旨在满足模型买方多样化的业务需求和预算限制。进入市场交易环节，模型经纪人将这些预训练好的模型推向市场，向模型买方展示。模型买方根据自身应用场景的具体要求（如预测精度、模型响应速度、成本预算等），选择最合适的模型进行购买。交易完成后，模型买方向模型经纪人支付模型费用 p_{m}，作为模型使用权的交换。

2.1　Fast-TMC Shapley 算法

在机器学习中，衡量每个数据点对模型性能的贡献是一个关键问题。Shapley 值法是合作博弈论中的一种经典方法[12]。在机器学习背景下，Shapley 值是考虑数据点在所有可能的数据子集中的边际贡献的平均值。在实际应用中，通常使用近似方法估计 Shapley 值，如蒙特卡罗截断采样 Shapley（TMC-Shapley）算法[13]。TMC-Shapley 算法引入截断技巧和蒙特卡洛采样策略，缓解了传统 Shapley 值估计的计算复杂性问题。然而，TMC-Shapley 算法存在如下 3 个问题。

1）计算成本高。尽管 TMC-Shapley 算法比 Shapley 值法有所优化，但计算量依然庞大，尤其是在处理大规模数据集时，所需的计算资源和时间可能依旧超出许多实际应用场景的承受范围。

2）参数敏感度高。TMC-Shapley 算法的性能严重依赖截断阈值的选择和蒙特卡洛采样的次数，这两个参数的设置需要仔细调优。不当的参数设置可能导致评估结果偏差，影响数据价值判断的准确性。

3）高维数据处理难。随着数据维度的增加，数据点之间的相互作用关系变得更加复杂，这直接加剧了 TMC-Shapley 算法的计算复杂度和内存需求，使其在高维数据集上的应用变得尤为困难。

Fast-TMC Shapley 算法，如图 2 所示。

Fast-TMC Shapley 算法着重解决了 Shapley 值法在大规模数据集应用中遇到的效率瓶颈。Fast-TMC Shapley 算法有 4 点核心改进。

1）优化采样策略。Fast-TMC Shapley 算法通过更加智能化的采样方法，减少了对数据子集的盲目探索，确保每次采样都能有效增加信息量，避免无效或重复的工作，从而在保持评估精度的同时，大幅降低所需的样本数量。

算法 1　快速蒙特卡洛采样方式计算 Shapley 值

输入：训练数据集 $D=(x_i,y_i)_1^N$，机器学习模型 A，抽样次数 N，评分指标 V

输出：训练数据 z_i 的 Shapley 值 sv

1: 初始化 $sv_i=0$ for $i=1,\dots,N$
2: 处理数据集和标签 L，将 i 对应的数据点和标签移到数组的末尾
3: 对于 $n=1$ 到 N 做：
　3.1: 初始化类别数 $C=1$
　3.2: 当 $C<2$ 时，重复以下步骤：
　　3.2.1: 随机选择一个子集长度 S
　　3.2.2: 生成一个随机排列 P
　　3.2.3: 根据随机排列选取子集 X_{R} 和 Y_{R}
　　3.2.4: 计算子集中的唯一类别数 C
　3.3: 使用子集训练模型并预测，计算评价指标 V_{C}
　3.4: 将 i 对应的数据点和标签添加到子集中，形成新的集合 X_{A} 和 Y_{A}
　3.5: 使用新的集合训练模型并预测，计算评价指标 V_{T}
　3.6: 计算并累加 Shapley 值的贡献，$sv+=|V_{\mathrm{T}}-V_{\mathrm{C}}|$
4: 计算 sv 除以 N 得到最终的 Shapley 值
5: 返回 sv

图 2　Fast-TMC Shapley 算法

Fig. 2　Fast-TMC Shapley algorithm

2）减少模型重训。通过高效的缓存和复用之前计算结果，即便是面对数据集的微小变化，也能够迅速定位并仅对必要部分进行重新评估，大大减少了计算资源的浪费。

3）高效存储与计算。利用高级的数据结构和算法优化，Fast-TMC Shapley 算法能够有效管理计算过程中的中间结果，确保在处理大规模数据时的内存使用效率和计算速度。

4）保证 Shapley 值估计的准确性。通过精心设计的近似策略和误差控制机制，即便在减少计算量的情况下，也能保持评估结果的可靠性和准确性。

2.2　数据边界建模

数据的 Shapley 值从大到小排序，从第一条数据开始购买，当购买第 n 条时，累计获得的总收益大于总成本，再根据模型买方的支付意愿（w_{TP}）构造收益的优化函数，从而找到数据边界 n。

$S(\mathrm{sv}_n)$ 表示前 n 个数据的 Shapley 总值，用来衡量第 n 个数据记录后模型可达到的准确度水平，$S(\mathrm{sv}_n)$ 为

$$S(\mathrm{sv}_n)=\sum_{i=1}^{n}\mathrm{sv}_i\ 。\tag{1}$$

式（1）中：sv_i 为购买数据记录。

假设 $w_{\mathrm{TP}}=\eta\cdot S(\mathrm{sv}_n)$，$\eta\geqslant 0$，效用函数满足 $\dfrac{\mathrm{d}S(\mathrm{sv}_n)}{\mathrm{d}n}>0$（递增函数），$\dfrac{\mathrm{d}^2S(\mathrm{sv}_n)}{\mathrm{d}^2n}\leqslant 0$（边际效用递减）。

由文献[4-5]中的式(2),拟合 $S(\mathrm{sv}_n)$和前 n 个数据记录之间的关联,显然方程满足效用函数的假设,即

$$S(\mathrm{sv}_n)=\sum_{i=1}^{n}\mathrm{sv}_i=\beta_1-\beta_2\cdot\exp(-\beta_3\cdot n)\ 。\tag{2}$$

在估计不同数据集参数时,计算原始数据集的 sv_i,并得到实验点$(n_1,S(\mathrm{sv}_1))$,…,$(n_N,S(\mathrm{sv}_N))$。使用非线性最小二乘算法,通过最小化平方误差之和优化 $\beta=(\beta_1,\beta_2,\beta_3)$,即

$$\min_{\beta_1,\beta_2,\beta_3}\sum_{i=1}^{N}(S(\mathrm{sv}_i)-(\beta_1-\beta_2\cdot\exp(-\beta_3\cdot n_i))^2\ 。\tag{3}$$

2.3 最优数据边界求解

假设 w 为模型买方的实际支付意愿,w'为模型买方名义支付意愿。实际支付意愿取决于 sv_n,并且有 $w=w'\cdot\mathrm{sv}_n$。其中,w'是一个正的随机变量,反映了模型买方的异质性。假设概率密度函数为 $f(w')$,针对不同现实情况,提出了 w'的两种分布。

1) 均匀分布。假设 W'为模型买方的最大名义支付意愿,W 为模型买方的最大的实际支付意愿,则 $W=W'\cdot\mathrm{sv}_n$。对于一组模型买方,其概率密度 $f(w')$遵循$[0,W']$的均匀分布。累积分布函数 $F(p_\mathrm{m})$表示 $W'\leqslant p_\mathrm{m}$ 的概率。因此,购买概率 p_r 为

$$p_\mathrm{r}=1-F(p_\mathrm{m})=W-p_\mathrm{m}。\tag{4}$$

模型经纪人的利润 $\Pi(p_\mathrm{m},n)$可以表示为

$$\begin{aligned}\Pi(p_\mathrm{m},n)&=p_\mathrm{m}\cdot m\cdot p_r-c\cdot n=p_\mathrm{m}\cdot m\cdot(W-p_\mathrm{m})-c\cdot n=\\&p_\mathrm{m}\cdot m\cdot(W'\cdot\mathrm{sv}_n-p_\mathrm{m})-c\cdot n。\end{aligned}\tag{5}$$

式(5)中:$\mathrm{sv}_n=\sum_{i=n}^{s}\mathrm{sv}_i=\beta_1-\beta_2\cdot\exp(-\beta_3\cdot n)$,$\beta_1,\beta_2,\beta_3\geqslant 0$;$p_\mathrm{m}\cdot m\cdot p_\mathrm{r}-c\cdot n=p_\mathrm{m}\cdot m\cdot(W-p_\mathrm{m})$为愿意支付不低于 p_m 价格的模型买方带来的收入;$c\cdot n$ 为支付给数据提供者的费用。

为获得最优的模型费用 p_m^* 和数据边界 n,有

$$\left.\begin{aligned}&\max_{p_\mathrm{m},n}\Pi(p_\mathrm{m},n)=p_\mathrm{m}\cdot m\cdot(W'\cdot S(\mathrm{sv}_n)-p_\mathrm{m})-c\cdot n,\\&\text{s.t.}\quad p_\mathrm{m}\geqslant 0,\quad n\geqslant 0。\end{aligned}\right\}\tag{6}$$

方程(6)的约束条件确保了 p_m 和 n 的非负解,有

$$\left.\begin{aligned}\frac{\Pi(p_\mathrm{m},n)}{\partial p_\mathrm{m}}&=m\cdot(W'\cdot S(\mathrm{sv}_n)-2p_\mathrm{m}),\\\frac{\Pi(p_\mathrm{m},n)}{\partial n}&=p_\mathrm{m}\cdot m\cdot W'\cdot\beta_2\cdot\beta_3\cdot\exp(-\beta_3\cdot n)-c。\end{aligned}\right\}\tag{7}$$

当 n 或 p_m 固定时,$\Pi(p_\mathrm{m},n)$的二阶导数为

$$\left.\begin{aligned}\frac{\partial^2\Pi(p_\mathrm{m},n)}{\partial^2 p_\mathrm{m}}&=-2m<0,\\\frac{\partial^2\Pi(p_\mathrm{m},n)}{\partial^2 n}&=-\beta_2\cdot\beta_3^2\cdot(p_\mathrm{m}\cdot m\cdot W')\cdot\exp(-\beta_3\cdot n)<0。\end{aligned}\right\}\tag{8}$$

因此,均匀分布的解是全局最优的。由 $\frac{\partial\Pi s(p_\mathrm{m},n)}{\partial p_\mathrm{m}}=0$ 和$\frac{\partial\Pi(p_\mathrm{m},n)}{\partial n}=0$,可以得到 n^* 和 p_m^* 的解。

2) 由于许多自然现象遵循对数正态分布(LND)[14],因此,使用对数正态分布模拟买方支付意愿比较符合实际。定义 w'遵循 LND 分布,其概率密度函数为 $f(w')$。累积分布函数 $F(p_\mathrm{m})$表示 $W'\leqslant p_\mathrm{m}$ 的概率。因此,购买概率为

$$p_r=1-F(p_\mathrm{m})=\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left(\frac{\ln p_\mathrm{m}-\mu}{\sqrt{2\sigma}}\right)S(\mathrm{sv}_n)。\tag{9}$$

式(9)中:erf(·)为度量理论中使用的误差函数。

模型经纪人的利润 $\Pi(p_\mathrm{m},n)$为

$$\Pi(p_\mathrm{m},n)=p_\mathrm{m}\cdot m\cdot p_\mathrm{r}-c\cdot n=p_\mathrm{m}\cdot m\cdot\left(\frac{1}{2}-\frac{1}{2}\mathrm{erf}\left(\frac{\ln p_\mathrm{m}-\mu}{\sqrt{2\sigma}}\right)\right)\cdot S(\mathrm{sv}_n)-c\cdot n=$$

$$\frac{p_{\mathrm{m}} \cdot m}{2}\left(1-\operatorname{erf}\left(\frac{\ln p_{\mathrm{m}}-\mu}{\sqrt{2\sigma}}\right)\right) \cdot S(\mathrm{sv}_n)-c \cdot n。 \tag{10}$$

式(10)中：$S(\mathrm{sv}_n)=\sum_{i=n}^{s}\mathrm{sv}_i=\beta_1-\beta_2 \cdot \exp(-\beta_3 \cdot n)$，$\beta_1,\beta_2,\beta_3 \geqslant 0$；$p_{\mathrm{m}} \cdot m \cdot p_r$ 为从支付意愿不低于 p_{m} 的模型买方处所得到的收入。

通过调整 μ 和 σ，可以灵活使用 LND 分布，以适应不同情况，假设 $\mu=0$ 和 $\sigma=1$，优化问题为

$$\left.\begin{aligned}&\max_{p_{\mathrm{m}},n}\Pi(p_{\mathrm{m}},n)=p_{\mathrm{m}} \cdot m \cdot \frac{1}{2} \cdot \left(1-\operatorname{erf}\left(\frac{\ln p_{\mathrm{m}}}{\sqrt{2}}\right)\right) \cdot S(\mathrm{sv}_n)-c \cdot n,\\&\text{s.t.}\quad p_{\mathrm{m}} \geqslant 0,\quad n \geqslant 0。\end{aligned}\right\} \tag{11}$$

式(11)中：约束条件为 p_{m} 和 n 的解非负。

通过对 $\Pi(p_{\mathrm{m}},n)$ 关于 p_{m} 和 n 进行微分，可以得到关于 $\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial p_{\mathrm{m}}}=0$ 和 $\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial n}=0$ 的具体形式。

$$\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial p_{\mathrm{m}}}=\frac{m \cdot S(\mathrm{sv}_n)}{2} \cdot \left(1-\operatorname{erf}\left(\frac{\ln p_{\mathrm{m}}}{\sqrt{2}}\right)-\sqrt{\frac{2}{\pi}} \cdot \exp\left(-\frac{\ln^2 p_{\mathrm{m}}}{2}\right)\right), \tag{12}$$

$$\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial n}=\left(\frac{p_{\mathrm{m}} \cdot m}{2} \cdot \left(1-\operatorname{erf}\left(\frac{\ln p_{\mathrm{m}}}{\sqrt{2}}\right)\right)\right) \cdot (\beta_2 \cdot \beta_3 \cdot \exp(-\beta_3 \cdot n))-c。 \tag{13}$$

$\Pi(p_{\mathrm{m}},n)$ 关于 p_{m} 或 n 的二阶导数是固定的且均为非正，即

$$\frac{\partial^2 \Pi(p_{\mathrm{m}},n)}{\partial^2 p_{\mathrm{m}}}=-\frac{m \cdot S(\mathrm{sv}_n)}{\sqrt{2\pi} \cdot p_{\mathrm{m}}} \cdot (1+\ln p_{\mathrm{m}}) \cdot \exp\left(-\frac{\ln^2 p_{\mathrm{m}}}{2}\right)<0, \tag{14}$$

$$\frac{\partial^2 \Pi(p_{\mathrm{m}},n)}{\partial^2 n}=-(\beta_2\beta_3^2)\left(\frac{p_{\mathrm{m}} \cdot m}{2} \cdot \left(1-\operatorname{erf}\left(\frac{\ln p_{\mathrm{m}}}{\sqrt{2}}\right)\right)\right) \cdot \exp(-\beta_3 \cdot n)<0。 \tag{15}$$

因此，LND 分布的解是全局最优的。通过解 $\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial n}=0$ 和 $\frac{\partial \Pi(p_{\mathrm{m}},n)}{\partial p_{\mathrm{m}}}=0$，得到 n^* 和 p_{m}^* 的解。

3　模型最优定价求解

3.1　模型定价模型的建模

为了实现促进双方更高效、互利的交易，需要考虑到的 3 个关键定价因素。

1) 基于调研，经纪人获得模型买方对各版本模型的具体购买意向，包括他们对不同性能指标组合的偏好排序和每个版本的最高支付意愿。

2) 根据收集到的信息，经纪人需制定精细的版本控制策略，决定推出哪些精度或功能级别的模型。

3) 在构建模型的定价模型(RM)时，经纪人还需考虑无套利原则，即确保市场中不存在利用价格差异进行无风险获利的机会。这意味着同一模型或相近性能模型之间的定价必须合理衔接，避免买方通过购买低价版本转售为高价版本，从而赚取差价，维护市场的稳定性和公平性。

模型经纪人使用数据集 D，训练不同版本的模型 $(v_1,v_2,\cdots,v_N)$ 进行销售，这些模型按准确度等级从低到高排序。假设模型买方对模型 v_i 感兴趣，且其模型价值为 $u_{i,j}$，模型买方仅在 $p(v_i) \leqslant u_{i,j}$ 时购买模型，模型经纪人则获得收入 $p(v_i)$。RM 可以写为

$$\left.\begin{aligned}&\max_{\langle p(v_1),p(v_2),\cdots,p(v_N)\rangle}=\sum_{i=1}^{N}\sum\nolimits_{j=1}^{M}p(v_i) \cdot I(p(v_i) \leqslant u_{i,j}),\\&\text{s.t.}\quad p(v_i)>p(v_j) \geqslant 0,\quad v_i \geqslant v_j,\\&\qquad p(v_i)+p(v_j) \geqslant p(v_i+v_j),\quad \forall v_i,v_j \geqslant 0。\end{aligned}\right\} \tag{16}$$

式(16)中：$p(v_i)$ 为整数变量；$I(p(v_i) \leqslant u_{i,j})$ 是指示变量；当 $p(v_i) \leqslant u_{i,j}$ 时，RM 为 1，否则为 0；约束条件确保了无套利属性。

3.2　模型最优定价求解

在求解式(16)时，因存在指示函数 $I(p(v_i) \leqslant u_{i,j})$ 而难以求解，需要将最大收益问题转换为等效的整数线性规划问题。首先，使用 0～1 整数变量 $x_{i,j,1}$，$x_{i,j,2}$ 和新增约束条件把非线性目标函数转化为了

一个二次整数线性规划问题。即

$$\left.\begin{aligned}&\max_{\langle p(v_i),x_{i,j,1},x_{i,j,2}\rangle}=\sum_{i=1}^{N}\sum_{j=1}^{M}p(v_i)\cdot x_{i,j,1},\\&\text{s.t.}\quad p(v_i)>p(v_j)\geqslant 0,\qquad v_i\geqslant v_j,\\&\qquad p(v_i)+p(v_j)\geqslant p(v_i+v_j),\qquad \forall v_i,v_j\geqslant 0,\\&\qquad x_{i,j,2}\cdot u_{i,j}\leqslant p(v_i)\leqslant x_{i,j,1}\cdot u_{i,j}+x_{i,j,2}\cdot L,\\&\qquad x_{i,j,1}+x_{i,j,2}=1,\\&\qquad 0\leqslant x_{i,j,1},x_{i,j,2}\leqslant 1。\end{aligned}\right\}\tag{17}$$

式(17)中:L 为无穷大;$p(v_i)$,$x_{i,j,1}$,$x_{i,j,2}$ 都是整数变量;第 1,2 个约束条件确保了市场的无套利特性,其余约束条件保证指示函数的转换;$u_{i,j}$ 为模型价值。

当 $x_{i,j,1}=1$,$x_{i,j,2}=0$ 时,目标函数变为 $p(v_i)$,第 3 个约束条件变为 $0\leqslant p(v_i)\leqslant u_{i,j}$,这意味着模型 v_i 的价格低于其模型价值,模型买方会购买此模型,模型经纪人获得收入 $p(v_i)$。

当 $x_{i,j,1}=0$,$x_{i,j,2}=1$ 时,目标函数变为 0,第 3 个约束条件变为 $u_{i,j}\leqslant p(v_i)$,这意味着模型 v_i 的价格高于其模型价值,模型买方不会购买。

由于目标函数是 0~1 整数变量的特性,将二次线性规划转换为一次线性规划问题,即

$$\left.\begin{aligned}&\max_{\langle p(v_i),x_{i,j,1},x_{i,j,2},y_{i,j}\rangle}=\sum_{i=1}^{N}\sum_{j=1}^{M}y_{i,j},\\&\text{s.t.}\quad p(v_i)>p(v_j)\geqslant 0,\qquad v_i\geqslant v_j,\\&\qquad p(v_i)+p(v_j)\geqslant p(v_i+v_j),\qquad \forall v_i,v_j\geqslant 0,\\&\qquad x_{i,j,2}\cdot u_{i,j}\leqslant p(v_i)\leqslant x_{i,j,1}\cdot u_{i,j}+x_{i,j,2}\cdot L,\\&\qquad x_{i,j,1}+x_{i,j,2}=1,\\&\qquad 0\leqslant x_{i,j,1},x_{i,j,2}\leqslant 1,\\&\qquad y_{i,j}\leqslant x_{i,j,1}\cdot L,\\&\qquad y_{i,j}\leqslant p(v_i),\\&\qquad y_{i,j}\geqslant p(v_i)-L\cdot(1-x_{i,j,1}),\\&\qquad y_{i,j}\geqslant 0。\end{aligned}\right\}\tag{18}$$

式(18)中:$p(v_i)$,$x_{i,j,1}$,$x_{i,j,2}$,$y_{i,j}$ 都是整数变量;前两个约束条件保障了市场无套利特性,第 3~5 个约束条件保证指示函数的转换,而其他约束条件项则负责 ILP 的转换。

当 $x_{i,j,1}=1$ 时,约束条件 $y_{i,j}\leqslant x_{i,j,1}\cdot L$ 变为 $y_{i,j}\leqslant L$,约束条件 $y_{i,j}\geqslant p(v_i)-L\cdot(1-x_{i,j,1})$ 变为 $y_{i,j}\geqslant p(v_i)$。同时,约束条件确保 $y_{i,j}\leqslant p(v_i)$。因此,3 个约束条件的结合确保了 $y_{i,j}=p(v_i)$。目标函数变为 $p(v_i)$,模型买方会购买此模型。

当 $x_{i,j,1}=0$ 时,约束条件 $y_{i,j}\leqslant x_{i,j,1}\cdot L$ 变为 $y_{i,j}\leqslant 0$,约束条件 $y_{i,j}\geqslant p(v_i)-L\cdot(1-x_{i,j,1})$ 变为 $y_{i,j}\geqslant -L$。同时,约束条件确保 $y_{i,j}\geqslant 0$。因此,3 个约束条件的结合确保了 $y_{i,j}=0$。目标函数变为 0,模型买方不会购买此模型。

以上过程将收入最大化问题转换为 ILP 问题(简称为 RM-ILP),可以获得提高模型经纪人收入的定价结果。

4 实验结果与分析

4.1 数据边界实验

数据边界实验采用糖尿病数据集和台湾省台北市房地产数据集两个公共数据集[15]。糖尿病数据集中共有 768 条数据,每条数据有 8 个特征和 1 个对应的标签。糖尿病数据是标签数据,首先,使用 Fast-TMC Shapley 算法计算 sv_i,即第 i 条数据的 Shapley 值,然后归一化 sv_i;接着,从大到小排序 sv_i 并累加,获得 $S(\mathrm{sv}_i)$。用最小均方误差的非线性最小二乘法进行拟合,拟合系数 $\beta_1=1.072$,$\beta_2=1.025$ 和 $\beta_3=0.034$。糖尿病数据集 Shapley 值,如图 3 所示。

房地产估值的市场历史数据集来自台湾省新北市新店区，一共有 414 条数据，每条数据有 6 个特征和 1 个对应的单位面积房价。评价函数是 R^2，R^2 是衡量模型解释的变异量的比例，R^2 在 0～1 之间，R^2 越接近 1，表示模型的拟合效果越好。最小均方误差下，非线性最小二乘法拟合系数 $\beta_1=1.041$，$\beta_2=0.925$ 和 $\beta_3=0.073$。台北市房地产数据集 Shaple 值，如图 4 所示。

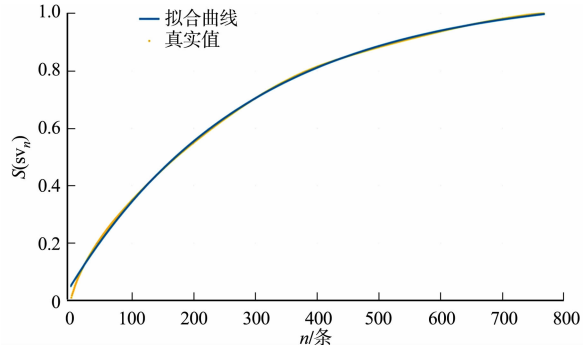

图 3　糖尿病数据集 Shapley 总值

Fig. 3　Gross Shapley value of diabetes data

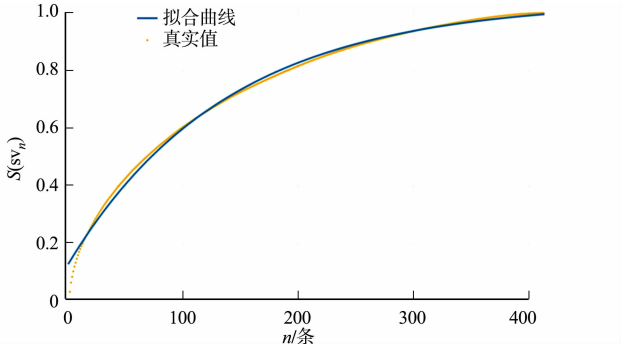

图 4　台北市房地产数据集 Shapley 总值

Fig. 4　Gross Shapley value of Taipei City real estate data

4.2　模型定价实验

针对不同的情境和目标，选择下列 4 种不同方法进行对比分析[16]。

1）Lin 方法。该方法较为简单直观，通过确定模型价值的最低点和最高点，构建线性函数，估算价格。该方法适用于价值与价格呈线性关系、且市场对价格敏感度相对稳定的场景。

2）Average 方法。该方法通过市场调研收集的数据，计算所有调查样本中模型价值的平均数作为定价基准。该方法试图平衡高估与低估的风险，适用于追求稳定和市场接受度优先的场景。

3）DPP 方法。该方法是一种更灵活和动态的方法，依据市场反馈、库存情况、竞争对手价格等多种因素，通过动态规划算法为不同模型版本设定价格，旨在实现利润最大化[17]。

4）DPP＋方法。作为 DPP 方法的升级版，DPP＋方法在动态规划的基础上，进一步扩展了解决方案空间，考虑了更多的变量和组合，以期找到更精准的最优定价策略[18]。

凹性预期价格和正态分布需求下的 $p(v_i)$，如图 5 所示。图 5 中：u 为正态分布需求。凸性预期价格和正态分布需求下的 $p(v_i)$，如图 6 所示。

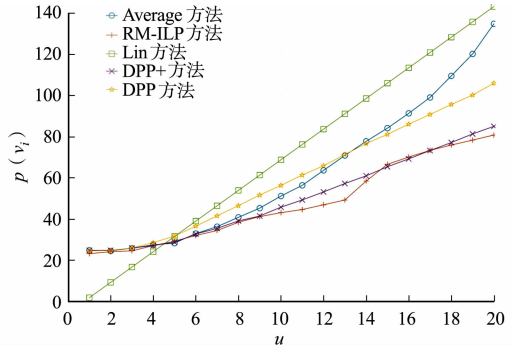

图 5　凹性预期价格和正态分布需求下的 $p(v_i)$

Fig. 5　$p(v_i)$ of concave expected price and normally distributed demand

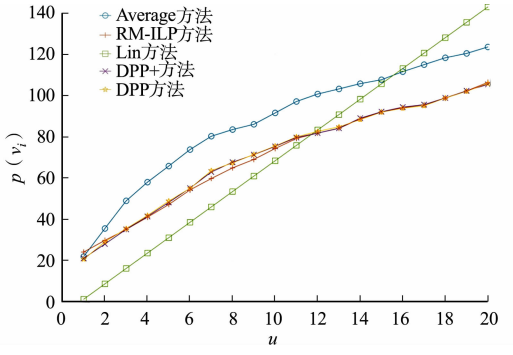

图 6　凸性预期价格和正态分布需求下的 $p(v_i)$

Fig. 6　$p(v_i)$ of convex expected price and normally distributed demand

凹性预期价格和正态分布需求的收入总值，如图 7 所示。图 7 中：p_t 为收入总值。凸性预期价格和正态分布需求的收入总值，如图 8 所示。由图 7，8 可知：相比于其他 4 种方法，RM-ILP 方法在收入增益方面的表现最为显著，最高可达 6.67 倍；RM-ILP 方法在凹性预期价格上相比于 DPP 方法和 DPP＋方法有更优的表现，模型经纪人收入分别增长了 23％和 3％；在凸性预期价格上相比于 DPP 方法和 DPP＋方法，模型经纪人收入分别增长了 1％，3％。

综上所述，RM-ILP 方法在数据市场中的模型定价和策略优化提供更优化的解决方案。其优势主要体现在以下 5 个方面。

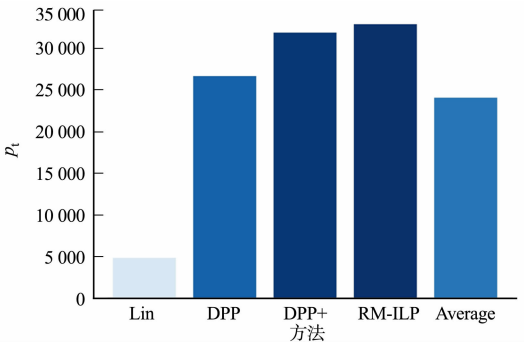

图7 凹性预期价格和正态分布需求的收入总值

Fig. 7 Concave expected price and total revenue of normally distributed demand

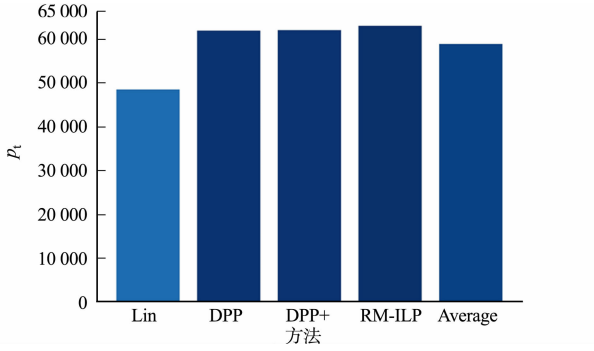

图8 凸性预期价格和正态分布需求的收入总值

Fig. 8 Convexity expected price and total revenue of normally distributed demand

1) 精细化价格设定。RM-ILP方法能够基于复杂的市场需求模型,精确调整不同模型版本的价格,确保每个版本都能在满足市场需求的同时,使利润最大化。这种精细化定价策略优于简单的线性或平均值定价法,因为它能够更准确地反映市场细分和买方偏好。

2) 优化收入增益通过优化函数。RM-ILP方法能够综合考虑各种定价和销售策略对总收入的影响,包括版本控制策略和市场需求的动态变化,从而实现收入的最大化。这种方法确保了模型经纪人在成本控制和市场需求之间找到最佳平衡点。

3) 确保可负担性比率。在追求利润最大化的同时,RM-ILP方法还能够通过优化过程考虑模型买方的支付能力和意愿,确保定价策略不会因为过高而排斥大量潜在买家,维持一个健康的市场可接受度和客户基础。

4) 适应动态市场环境。面对快速变化的市场环境和买方偏好,RM-ILP方法的动态调整能力至关重要。它能够及时反应市场信息,调整模型版本的供应、价格,甚至开发新的模型版本,以适应市场需求的变化,展现出了极强的灵活性和适应性。

5) 确保遵守市场规则。在构建优化模型时,RM-ILP方法充分考虑了数据市场的特性和规则(如版本控制和无套利原则),确保了定价策略的合法性和市场公平性,防止了市场操纵和不正当竞争行为。

5 结论

1) 精细化数据价值评估。通过先进的Fast-TMC Shapley算法,实现了对数据点贡献的精细化评估,为数据的定价与交易提供了科学依据。

2) 高效资源分配。数据边界的确定帮助模型经纪人实现资源的高效配置,避免了不必要的数据购置成本,提高了数据利用效率。

3) 动态定价策略。基于RM-ILP的模型定价优化,不仅考虑了买方的偏好和支付能力,还确保了市场规则的遵守,提高了模型产品的市场竞争力和接受度。对不同真实数据集的数据边界进行了数值研究,验证了最优数据边界的存在性。与其他4种方法相比,RM-ILP方法在模型经纪人收入最大化问题的求解中展现了明显优势。

通过应用RM-ILP方法,模型经纪人不仅能够更深入地理解模型买方的购买偏好和支付意愿,还能制定出更贴合市场需求、更符合买方行为习惯的定价策略,促进了数据市场的健康发展,通过提高数据和模型交易的透明度、效率和公平性,吸引更多参与者,推动数字经济的繁荣。

参考文献:

[1] SONG Jie, HE Guannan, WANG Jianxiao, *et al*. Shaping futurelow-carbon energy and transportation systems digital technologies and applications[J]. Energy, 2022, 1(3): 285-305. DOI: 10.23919/IEN.2022.0040.

[2] PEI Jian. A survey on data pricing: From economics to data science[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(10): 4586-4608. DOI: 10.1109/TKDE.2020.3045927.

[3] LI Xijun, YAO Jianguo, LIU Xue, *et al*. A first look at information entropy-based data pricing[C]// Proceedings of the 37th International Conference on Distributed Computing Systems. Atlanta: IEEE Press, 2017: 2053-2060. DOI: 10.1109/ICDCS.2017.45.

[4] SHEN Yuncheng, GUO Bing, SHEN Yan, *et al*. A pricing model for big personal data[J]. Tsinghua Science and Technology, 2016, 21(5): 482-490. DOI: 10.1109/TST.2016.7590317.

[5] YU Haifei, ZHANG Mengxiao. Data pricing strategy based on data quality[J]. Computers & Industrial Engineering, 2017, 112: 1-10. DOI: 10.1016/j.cie.2017.08.008.

[6] YANG Jian, ZHAO Chongchong, XING Chunxiao, *et al*. Big data market optimization pricing model based on data quality[J]. Complexity, 2019(2): 1-10. DOI: 10.1155/2019/5964068.

[7] ZHANG Meng, ARAFA A, HUANG Jianwei, *et al*. Pricing fresh data[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(5): 1211-1225. DOI: 10.1109/JSAC.2021.3065088.

[8] ZHANG Meng, ARAFA A, HUANG Jianwei, *et al*. How to price fresh data[C]// Proceedings of the 2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. Avignon: IEEE Press, 2019: 1-8. DOI: 10.23919/WiOPT47501.2019.9144091.

[9] NIYATO D, ALSHEIKH M A, WANG Ping, *et al*. Market model and optimal pricing scheme of big data and internet of things (IoT)[C]// Proceedings of the 2016 IEEE International Conference on Communications. Malaysia: IEEE Press, 2016: 1-6. DOI: 10.1109/ICC.2016.7510922.

[10] OH H, PARK S, LEE G, *et al*. Competitive data trading model with privacy valuation for multiple stakeholders in IoT data markets[J]. IEEE Internet of Things Journal, 2020, 7(4): 3623-3639. DOI: 10.1109/JIOT.2020.2973662.

[11] TIAN Yingjie, DING Yurong, FU Saiji, *et al*. Data boundary and data pricing based on the shapley value[J]. IEEE Access, 2022, 10: 14288-14300. DOI: 10.1109/ACCESS.2022.3147799.

[12] GHORBANI A, ZOU J. Data shapley: Equitable valuation of data for machine learning[C]// Proceedings of the 36th International Conference on Machine Learning. California: PMLR Press, 2019: 2242-2251. DOI: 10.48550/arXiv.1904.02868.

[13] CHEN Lingjiao, KOUTRIS P, KUMAR A. Towards model-based pricing for machine learningin a data marketplace [C]// Proceedings of the 2019 International Conference on Management of Data. Amsterdam: ACM Press, 2019: 1535-1552. DOI: 10.1145/3299869.3300078.

[14] ALBERINI A. Efficiency vsbias of willingness-to-pay estimates: Bivariate and interval-data models[J]. Journal of Environmental Economics and Management, 1995, 29(2): 169-180. DOI: 10.1006/jeem.1995.1039.

[15] YEH I, HSU T. Building real estate valuation models with comparative approach through case-based reasoning[J]. Applied Soft Computing, 2018, 65: 260-271. DOI: 10.1016/j.asoc.2018.01.029.

[16] DONG Xin, SAHA B, SRIVASTAVA D. Less is more: Selecting sources wisely for integration[J]. VLDB Endowment, 2012, 6(2): 37-48. DOI: 10.14778/2535568.2448938.

[17] LIU Jinfei, LOU Jian, LIU Junxu, *et al*. Dealer: An end-to-end model marketplace with differential privacy[J]. VLDB Endowment, 2021, 14(6): 957-969. DOI: 10.14778/ 3447689.3447700.

[18] JIA Ruoxi, DAO D, WANG Boxin, *et al*. Efficient task-specific data valuation for nearest neighbor algorithms[J]. VLDB Endowment, 2019, 12(11): 1610-1623. DOI: 10.14778/3342263.3342637.

（**责任编辑**：陈志贤　　**英文审校**：黄心中）