

面向可信数据共享的双层强化学习动态激励机制

段美姣^{1,2}, 黄日东¹, 李惠¹, 王熙¹

(1. 中央财经大学信息学院, 北京 100081; 2. 国家金融安全教育部工程研究中心, 北京 102206)

摘要: 数据驱动平台运营依赖用户持续贡献高质量数据, 但数据上传常伴随隐私风险与个体成本。现有激励机制多基于静态规则设计, 难以适应动态环境及用户异质性, 导致激励效率低下及平台收益与激励成本失衡。针对上述问题, 提出一种面向可信数据共享的双层强化学习动态激励机制。构建涵盖隐私效用、激励收益及近似 Shapley 值数据贡献评估的动态博弈模型; 设计双层强化学习框架, 上层平台智能体学习最优激励策略, 下层用户依据个体偏好决策数据上传行为, 平台通过边际激励与固定激励构成的二维向量激励用户差异化数据共享; 结合区块链与智能合约实现激励去中心化结算, 保障机制可信性。仿真实验表明, 与现有基线相比, 所提机制的训练总激励成本约压缩至基线均值的 1.9%, 生产力指标提升约 21%, 验证了其在异质用户行为响应下的有效性与适应性。

关键词: 双层强化学习; 数据共享; 动态激励机制; 隐私成本; Shapley 值; 区块链

中图分类号: TP309

文献标志码: A

doi: 10.11959/j.issn.1000

Bilevel Reinforcement Learning Dynamic Incentive Mechanism for Trustworthy Data Sharing

Duan Meijiao^{1,2}, Huang Ridong¹, Li Hui¹, Wang Xi¹

1. School of Information, Central University of Finance and Economics, Beijing 100081, China

2. Engineering Research Center of State Financial Security, Ministry of Education, Beijing 102206, China

Abstract: The operation of data-driven platforms relies on users' continuous contribution of high-quality data, while data uploading is often accompanied by privacy risks and individual costs. Existing incentive mechanisms are mostly based on static rules, rendering them inadequate for dynamic environments and user heterogeneity, thereby leading to low incentive efficiency and an imbalance between platform revenue and incentive costs. To address these challenges, a bilevel reinforcement learning-based dynamic incentive mechanism for trustworthy data sharing was proposed. A dynamic game-theoretic model incorporating privacy utility, incentive payoffs, and approximate Shapley value-based data pricing was constructed. A bilevel reinforcement learning framework was designed, in which the optimal incentive policy was learned by the upper-level platform agent, while data uploading behaviors were determined by lower-level users according to individual preferences. Differentiated data sharing was encouraged through a two-dimensional incentive vector consisting of marginal incentives and fixed incentives. Decentralized incentive settlement was implemented using blockchain and smart contracts to ensure the trustworthiness of the mechanism. Simulation experiments show that, compared with the baseline schemes, the proposed mechanism reduces the total training incentive cost to approximately 1.9% of the average baseline cost, and improves the productivity metric by approximately 21%. The results demonstrate the effectiveness and adaptability of the proposed mechanism under heterogeneous user responses.

收稿日期: XXXX-XX-XX; 修回日期: XXXX-XX-XX

通信作者: 王熙, xiwang@cufe.edu.cn

基金项目: 北京市社会科学基金资助项目 (No.24GLC056)

Foundation Items: Social Science Fund of Beijing (No.24GLC056)

Key words: bilevel reinforcement learning, data sharing, dynamic incentive mechanism, privacy cost, Shapley value, blockchain

0 引言

随着数字经济与人工智能的快速发展,数据已成为支撑智能服务优化和平台价值创造的关键生产要素^[1]。以推荐系统、个性化广告和用户行为分析等为代表的驱动平台,高度依赖用户持续贡献实时、高质量行为数据以优化模型和服务。例如,京东、淘宝等电商平台通过分析用户浏览、点击、收藏和购买等行为日志提升推荐准确性;健康管理服务则依据用户运动、饮食和睡眠等数据提供个性化建议。然而,用户数据贡献行为通常伴随显性成本和隐性风险。消费偏好、位置轨迹、社交关系等敏感信息一旦被不当使用,可能引发身份泄露、生活习惯及行为偏好推断等隐私风险^[2]。因此,用户不仅关注激励水平,也对长期隐私损失以及平台激励机制的可信性高度敏感。若平台无法提供稳定可信的激励兑现,理性用户将倾向于减少数据贡献,进而导致低信任与低参与相互强化的恶性循环。

平台侧同样面临激励预算有限、数据质量参差不齐、用户参与意愿动态波动等多重约束^[3-6]。传统激励机制,如固定积分、一次性补偿和静态博弈定价等,多基于同质化或静态的用户隐私偏好假设,而现实平台中用户偏好、隐私敏感度、上传数据量、数据质量、数据分布及数据对平台任务的边际价值均可能随时间变化。因此,平台需要根据用户响应和数据价值持续动态调整激励策略,而现有静态用户隐私模型及激励机制难以适应动态、异质的实际需求^[3,5]。此外,低质、无效或重复的数据会降低激励资源的使用效率,因此还需在贡献评估环节体现数据价值差异。

数据市场与众包领域近年来涌现出大量针对上述问题的激励机制研究^[3,7],主要从博弈建模和数据贡献评估两个方面展开:前者通过引入 Stackelberg 博弈,将平台建模为领导者、用户建模为跟随者,以刻画平台与用户之间的激励交互^[4-6];后者利用 Shapley 值等方法量化数据贡献以实现公平定价^[8-9]。然而,现有方法仍存在一定局限:其一,大多机制基于静态或单轮博弈,忽略长期多轮交互中的策略演化;其二,精确 Shapley 值计算复杂度高,在大规模、动态数据场景下存在计算效率瓶

颈;其三,对用户隐私成本的动态累积考虑不足;其四,通常仅采用货币或积分等单一激励形式,难以在长期交互框架中同时刻画动态用户响应与高质量数据的边际贡献,并实现多维个性化激励。

强化学习理论与算法的持续发展,为具有层级决策关系的复杂博弈问题提供了新的求解思路。Zheng 等^[10]提出 AI Economist 框架,将经济政策设计建模为社会规划者与异质经济代理人之间的两级共适应强化学习问题,并在税收政策场景中验证了该范式的可行性。在数据共享场景下,平台与用户之间的动态博弈同样具有类似的层级决策结构,可采用双层强化学习框架进行建模并自动寻优激励策略。此外,激励机制不仅需要关注策略效果,还需要关注执行过程的可信性。将区块链与智能合约引入数据交易场景,可有效缓解信息不对称与隐私风险问题^[11-12]。通过将激励规则编码为链上逻辑,可使计算过程可审计、激励发放过程可验证。

综上,如何刻画长期数据共享中的用户动态异质性,并在同一框架下协同实现层级动态策略优化、差异化激励与可信执行,以解决动态博弈策略寻优和激励透明可信执行等问题,是构建长期数据共享动态激励机制亟待解决的问题。

针对上述问题,本文主要的研究贡献如下:

(1) 构建面向动态异质用户的数据共享博弈模型。针对现有数据共享激励机制难以同时刻画用户数据动态增长、隐私成本累积以及数据边际贡献差异的问题,将隐私效用、上传成本、激励收益及基于近似 Shapley 值的数据贡献评估纳入动态 Stackelberg 博弈,刻画异质用户与平台在长期数据共享过程中的动态交互。

(2) 设计基于双层强化学习的动态激励机制。针对异质用户和多维激励耦合下动态 Stackelberg 博弈难以解析求解的问题,设计了基于双层强化学习的动态激励机制,通过上下层智能体交替冻结训练,实现平台激励策略与用户响应策略的动态高效寻优。并进一步设计了由边际激励与固定激励构成的二维激励向量,实现面向异质用户的自适应差异化激励。

(3) 构建基于联盟区块链与智能合约的可信激

励结算机制。针对现有激励机制在结算与执行过程中透明性与可验证性不足的问题,将激励结算规则编码为智能合约,通过链上执行实现激励结算过程的透明、可审计和可验证,保障机制可信性。

本文基于京东用户行为数据集构建异质用户动态数据共享环境,对所提机制开展仿真验证。实验结果表明,所提机制在异质用户响应环境下维持较高用户参与率的同时,训练总激励成本降至基线均值约 1.9%,平台生产力提升约 21%,并形成稳定的差异化激励策略,验证了其有效性与适应性。

1 相关工作

1.1 数据价值评估与定价

数据价值评估旨在衡量数据在特定场景下的潜在或实际价值,是数据定价与激励分配的重要基础^[14]。现有研究综合采用经济学、计算机科学和运筹优化等方法,从数据特征、任务效用以及参与者边际贡献等不同视角展开研究。根据价值评估依据与方法的不同,相关研究大致可分为以下三类^[13]:

(1) 基于数据特征的价值评估与定价

该类研究主要从数据资产自身固有特征出发^[14],通过对不同特征进行量化形成差异化价格体系。Yu 等^[15]从准确性、完整性、时效性等维度评估数据质量,构建双层规划定价模型,由平台根据数据质量制定不同版本及价格,消费者据此选择效用最优的数据产品。Li 等^[16]则将稀缺性纳入临床研究数据定价模型,实现基于数据特征的差异化价值评估。

(2) 基于任务效用的数据价值评估

该类研究侧重评估数据在支持特定任务或业务流程时产生的效用或贡献度,并据此确定数据价值^[14]。可通过数据加入前后任务性能或业务效用的变化衡量其增量价值,例如机器学习模型的性能提升。与基于数据自身特征的定价相比,任务效用方法能够更直接地反映数据与具体应用目标之间的关系,但其结果通常依赖特定任务和评价函数。同时,在多参与者场景下,如何公平分配联合产生的价值仍是需要解决的关键问题。

(3) 基于边际贡献的数据价值评估

该类研究依据数据样本或参与者对整体任务效用的边际贡献评估其价值,Shapley 值是其中具有

代表性的方法。Shapley 值源于合作博弈论,在数据市场中可用于量化边际贡献并公平分配联盟收益。Baghcheband 等^[8]面向机器学习数据市场提出增益数据 Shapley 值,通过量化数据对模型性能的贡献实现数据价值评估,并有助于提升数据补偿公平性和资源配置效率。Xia 等^[9]进一步针对时变数据提出动态 Shapley 计算方法。然而,精确 Shapley 计算的复杂度随参与者规模呈指数增长,在大规模、动态数据场景下难以直接应用,通常采用 Monte-Carlo 采样、回归调整 Monte-Carlo 与分层采样等近似方法^[17]。因此,如何在动态数据共享场景中兼顾贡献评估精度、计算效率和激励实时性,仍是数据共享激励机制设计中的关键问题。

此外,现有数据定价与价值评估研究更多关注数据本身及其任务价值,对数据供给方隐私成本的量化与补偿关注相对有限。Ghosh 等^[18]将差分隐私数据购买建模为采购拍卖,以隐私损失作为用户补偿依据。Hu 等^[19]在联邦学习场景下利用用户自主申报的隐私预算量化隐私损失,并基于 Stackelberg 博弈设计补偿机制。Kang 等^[20]基于公理化公平定义,研究联邦学习中不同隐私保护水平下的数据公平补偿机制。然而,现有涉及隐私因素的研究多聚焦于隐私损失度量、单轮补偿及任务级激励,对长期数据共享过程中隐私成本累积及用户响应的动态演化考虑不足,也较少将隐私成本与数据边际贡献、动态激励策略进行统一建模。

1.2 基于强化学习的动态激励机制

传统数据共享激励机制通常采用契约理论、拍卖机制或静态博弈模型求解均衡策略^[3]。然而,契约理论要求平台预先已知用户类型分布,拍卖机制难以适应用户行为随时间动态演化的场景,静态博弈方法则无法捕捉平台与用户之间的长期策略交互过程。强化学习(reinforcement learning, RL)通过智能体与环境的持续交互,以最大化长期累积奖励为目标进行策略学习,为动态环境下的激励策略优化提供了新的求解思路^[21]。

然而,当决策主体之间存在层级依赖关系时,例如平台作为领导者制定激励策略、用户作为跟随者动态响应,传统单层强化学习难以同时刻画不同层级主体的异质目标与主从决策结构。双层强化学习(bilevel reinforcement learning, BiRL)将具有层级关系的多主体决策问题形式化为上下层嵌套优化

结构, 上层领导者的目标函数依赖于下层跟随者的最优策略^[22], 两者之间形成 Stackelberg 博弈关系, 为数据共享激励、策略对齐等场景提供了可行的建模框架^[23]。

围绕具有层级决策关系的强化学习问题, 已有研究从双层优化和多智能体层次学习等角度进行了探索。Thoma 等^[24]提出上下文双层强化学习方法, 将下层决策过程建模为上下文马尔可夫决策过程, 并将整体问题表示为带随机上下文的 Stackelberg 博弈, 通过基于梯度的超策略优化实现领导者对多跟随者的激励诱导, 适用于激励对齐、奖励塑造、税收设计以及机制设计场景。Paolo 等^[25]从多智能体层次强化学习角度提出去中心化 TAG 框架, 通过 LevelEnv 抽象实现任意深度的层次结构与标准化的层级间信息交互, 支持异构智能体与松耦合信息流, 为复杂多智能体系统中的层次化策略学习提供了可扩展框架。

在动态激励机制研究方面, 强化学习已被用于众包、边缘学习等场景中的激励策略优化。Liu 等^[26]提出 Chiron 激励方案, 通过层次强化学习协调边缘学习中的参与者激励与鲁棒性目标。Jiang 等^[27]针对时空众包感知任务分配, 设计基于 Q-learning 的激励机制, 通过动态调整目标任务奖励以提高工作者参与度与任务完成率, 解决了静态分配导致的覆盖率低的问题。Luo 等^[28]提出面向移动众包场景的深度强化学习激励机制, 联合考虑感知与通信资源激励, 实现了多目标协同激励优化。上述研究表明, 强化学习能够根据参与者行为反馈动态调整激励策略, 为处理长期交互与环境变化提供了可行途径。

除利用强化学习优化动态激励策略外, 部分研究还通过 Stackelberg 博弈建模激励主体与参与者之间的主从决策关系。Han 等^[4]提出基于联盟区块链的联邦学习动态激励机制, 将任务发布者作为 Stackelberg 博弈领导者确定总奖励, 参与者作为跟随者决策参与程度, 并结合声誉机制实现参与方可靠性评估。Yuan 等^[5]提出 QI-DPFL 框架, 将中央服务器与参与客户端的交互建模为两阶段 Stackelberg 博弈, 服务器作为领导者设计时序奖励以最小化激励成本, 客户端作为跟随者决策隐私预算以最大化个人效用。He^[6]构建基于 Stackelberg 博弈的边缘联邦学习动态激励优化框架, 通过领导者—跟随者博

弈模型与动态奖惩策略调节异构边缘节点的参与行为。Wang 等^[29]进一步面向客户端数据分布不一致的差分隐私联邦学习场景, 将客户端的数据量、隐私预算决策与服务器支付策略建模为 Stackelberg 博弈, 并采用双层多智能体强化学习逼近博弈均衡, 为多维决策耦合下的动态激励求解提供了新的思路。

总体来看, 现有研究分别从强化学习的动态策略优化能力和 Stackelberg 博弈的主从决策建模能力出发, 形成了多种动态激励机制设计思路。Stackelberg 博弈能够较好地描述平台与参与者之间的层级交互关系, 但其策略求解通常依赖预设模型或特定响应关系。强化学习则无需显式求解复杂用户响应函数的闭式解, 可通过与环境的持续交互自适应地学习动态策略。而双层强化学习进一步将平台与用户的主从决策结构显式编码为上下层嵌套优化问题, 使得激励策略的学习过程与用户响应的动态演化能够协同优化, 更适合刻画异质用户环境下的长期激励均衡^[26, 29]。然而, 现有相关研究仍存在以下不足: 一是多数方法基于静态的参与者行为或上下文假设, 未充分考虑现实中用户私有数据的动态增长与隐私累积效应对长期激励策略的影响; 二是激励形式多为单维或通用设计, 难以同时兼顾数据质量与边际贡献、基础参与与隐私成本补偿等需求, 实现多维差异化激励; 三是实验验证多基于人工构造的马尔可夫决策过程仿真环境, 对利用真实用户行为日志构建动态交互环境并开展长期策略验证的研究仍较有限。

1.3 基于区块链的可信数据共享激励机制

在数据共享场景中, 平台与用户之间的激励结算长期依赖中心化执行, 其固有的不透明性与单点信任风险难以从根本上消除。区块链凭借去中心化账本的不可篡改性与智能合约的可编程自动执行特性, 为上述问题提供了可行的方案支撑, 近年来被广泛引入数据共享与众包激励机制的设计中。

在数据共享激励的机制构建方面, Xuan 等^[11]基于演化博弈理论构建了区块链与智能合约相结合的数据共享激励模型, 智能合约动态调控激励参数, 在无需可信第三方的前提下持续驱动用户参与数据共享。Alagha 等^[7]进一步在区块链数据共享平台上系统研究了多种激励合约形式, 通过智能合约实现数据所有者与请求方之间的自动化支付, 揭示

了不同激励结构对用户参与行为的差异化影响。

在贡献评估与公平激励方面,已有工作尝试将数据贡献评估与链上激励分发相结合。Li 等^[12]提出基于智能合约的隐私保护数据聚合与质量驱动激励机制,在链上实现数据质量评估与奖励分发的一体化执行,有效防止请求方对数据质量的虚假评价。Yang 等^[30]则将 Shapley 值引入区块链数据共享的收益分配,构建了面向多参与方的公平激励分配模型,从合作博弈视角为数据贡献的差异化定价提供了理论依据。

然而,现有方案中的激励规则通常以预定义逻辑固化部署于智能合约中,其执行过程缺乏对平台环境和用户行为动态变化的自适应调整能力。同时,如何在链上透明执行的约束下实现激励策略的持续寻优,尚未得到系统性探讨。本文将区块链智能合约与双层强化学习框架协同设计,智能合约承担激励结算的透明执行,双层强化学习实现激励策略的动态迭代优化,二者在功能分工上形成互补,共同构成完整的可信动态激励系统。

2 数据贡献评估与双层动态激励建模

2.1 数据贡献评估算法设计

在数据共享激励机制中,用户上传数据的边际贡献评估是边际激励计算的基础。本文采用基于近似 Shapley 值的数据贡献评估方法,将用户上传数

据对平台整体价值的增量贡献作为差异化激励的重要依据,同时综合考虑数据数量、数据质量、数据稀缺性、连续性和有效性等,更准确地评估用户数据对平台任务效用的贡献。具体参数如表 1 所示。

考虑一个有限的周期集 $\mathcal{T} = \{T_1, T_2, \dots, T_{\max}\}$ 。每个周期均包含一系列决策时间点 $\mathcal{D} = \{t_0, t_1, \dots, t_m\}$, 相邻时间点的间隔 $\Delta t = t_n - t_{n-1}$ 是固定的,以此模拟数据市场长期运行的多周期动态决策进程。

在每个时间步起始时,系统中已存在一定规模的用户数据池,平台根据当前的数据存量与质量进行本周期的初始激励预算规划。随着每个周期的时间推移,用户不断产生新的行为数据,平台需要实时响应这些动态涌现的数据,持续进行数据价值的评估与激励计算,进行激励资源的分配与调度,从而激励用户按照一定的策略持续上传或提高数据质量,参与到数据共享生态中,以促进平台的数据要素流转和价值增值。

在每个时间周期 T_k 内,用户 u_i 上传其隐私数据 $h_i^{T_k}$, 平台综合所有用户数据,计算本周期整体的数据稀缺性分数 $Y_U^{T_k}$ 、连续性分数 $T_U^{T_k}$ 和有效性分数 $M_U^{T_k}$, 并通过加权和得到总体数据质量 $Q_U^{T_k}$ 。由于精确 Shapley 值需对不同排列或所有联盟的边际贡献进行加权平均,计算复杂度随用户数量呈指数增长,在多用户长周期的动态决策场景中计算开销

表 1

近似 Shapley 值计算符号说明

符号	说明
U	所有参与者的集合
u_i	参与者 i
S	不包含参与者 i 的任意一个参与者子集
$B_S^{T_k}$	联盟 S 在第 k 个周期内创造的总数据价值
$B_{S \cup \{i\}}^{T_k}$	参与者 i 加入后,新联盟创造的总数据价值
$B_{S \cup \{i\}}^{T_k} - B_S^{T_k}$	参与者 i 对联盟 S 的边际贡献
$q_i^{T_k}(U)$	用户 u_i 在时间周期 T_k 的个体数据质量因子
$Q_U^{T_k}$	所有参与者在时间周期 T_k 的总体数据质量
$h_i^{T_k}$	用户 u_i 在时间周期 T_k 共享的隐私数据量
$Y_U^{T_k}$	数据稀缺性,用户数据加入后平台特征分布的信息熵增加量
$T_U^{T_k}$	数据连续性,刻画用户在本周期内是否持续产生可追踪数据流,例如活跃天数占比与行为时间间隔
$M_U^{T_k}$	数据有效性,刻画数据是否具备可被平台任务利用的信息,如有效值、完整字段

过大。本文借鉴 Shapley 值的边际贡献思想, 采用留一贡献近似方法, 衡量用户个体数据对整体质量的边际影响, 可在保持贡献评估精度的同时显著降低计算复杂度^[31]。具体而言, 将总体数据质量 Q_U^T 与剔除用户 u_i 后数据池质量 $Q_{U \setminus \{i\}}^T$ 的差值, 作为该用户的数据质量边际贡献因子 $q_i^T(U)$:

$$q_i^T(U) = Q_U^T - Q_{U \setminus \{i\}}^T \quad \#(1)$$

该因子量化了用户数据的不可替代性。在此基础上, 平台将 T_k 内的总数据价值定义为:

$$B_U^{T_k} = \sum_{i \in U} h_i^{T_k} \times q_i^T(U) \quad \#(2)$$

进行公平分配, 即最终价值等于数据数量或广度与数据质量相乘。平台通过计算用户 u_i 在所有可能用户子集联盟 $S \subseteq U \setminus \{i\}$ 中的边际贡献:

$$\text{Shapley} = [B_{S \cup \{i\}}^{T_k} - B_S^{T_k}] \quad \#(3)$$

得到用户应得的激励 $P_i^{T_k}(U)$, 即这个周期内基于用户 u_i 上传数据 $h_i^{T_k}$ 应给予的激励额。

2.2 双层动态模型构建

在 2.1 节数据贡献评估的基础上, 进一步构建平台与异质用户之间的双层动态激励模型。相关符号如表 2 所示。

表 2 激励机制符号说明

符号	说明
$g_i(t)$	用户 u_i 在时刻 t 拥有的数据总量
$h_i(t)$	用户 u_i 在时刻 t 选择共享的隐私数据量
$D_i(t)$	用户 u_i 在时刻 t 上传隐私数据后剩余数据量
π_i	用户 u_i 的响应策略
$L_i(t)$	用户 u_i 在时刻 t 获得的奖励
$\phi_i(t)$	留一法计算后的近似 Shapley 值
$P(g_i)$	用户拥有 g_i 单位隐私数据的隐私效用函数
R_i	用户 u_i 上传数据所获得的二维激励向量
Productivity(t)	平台生产力水平, 即总数据价值
$V(t)$	平台层目标, 即最大化数据价值
$C_i^{\text{priv}}(t)$	用户 u_i 在 t 轮次内上传数据的隐私成本

在动态数据贡献过程中, 用户 u_i 在时刻 t 的数据贡献量 $h_i(t)$, 由用户在特定系统状态下对隐私成本、激励收益与上传成本进行权衡后, 依据生成的决策策略 π_i 决定, 并进一步影响了个体数据存

量的动态演化与平台整体数据生态的更新。基于此, 本文构建了一种兼顾用户个体理性与平台决策目标的双层动态演化博弈框架。

考虑一个数据驱动平台与 N 个异质用户, 交互时间离散为 $t = \{1, 2, \dots, T\}$ 。每个用户 u_i 在第 t 轮拥有私有数据池 $g_i(t)$, 其中包含当轮新增的行为数据。用户从中选择上传量 $h_i(t) \in [0, g_i(t)]$, 当 $h_i(t) = 0$ 时表示不上传。上传后剩余的数据池记为 $D_i(t) = g_i(t) - h_i(t)$, 构成下一轮的起点。平台根据用户上传行为计算二维激励向量 $R_i(h_i(t)) = \{x_i(t), y_i(t)\}$, 其中:

$x_i(t)$ 表示边际激励, 平台将用户 u_i 本次上传数据的边际贡献, 即近似 Shapley 值 $\phi_i(t)$, 作为重要决策依据, 通过强化学习策略输出相应的边际激励强度, 以精准奖励高价值、高质量数据对全局模型性能的增量提升;

$y_i(t)$ 表示固定激励, 由全局基准项与平台策略输出的有界微调项构成: $y_i(t) = y_{\text{base}}(t) + \Delta y_i(t)$ 。其中基准项 y_{base} 主要依据全局用户的整体参与度动态调整, 用于覆盖低价值用户的上传成本、维持基础参与意愿, 并平滑系统整体稳定性。平台智能体在该基准上输出有界调整量 $\Delta y_i(t)$, 从而形成面向不同用户的个体化固定激励。

在机制设计中, 考虑两类影响用户上传成本的因素: 其一为每个用户对其持有的隐私数据存在固有效用 $P(g_i(t))$, 即用户 u_i 通过持有规模为 g 的隐私数据池而获得的隐私效用。当用户共享 $h_i(t)$ 单位数据时, 数据池变为 $D_i(t) = g_i(t) - h_i(t)$, 产生的隐私成本 $C_i^{\text{priv}}(t)$ 定义为效用函数 P 在共享前后的差值, 并额外附加一项与上传量成正比的固定风险溢价。其二为上传成本, 用户 u_i 在第 t 轮上传 $h_i(t)$ 单位数据所需的成本为 $C_i(t) = c_i \cdot h_i(t)$, c_i 为用户的单位上传成本系数。

2.2.1 用户层奖励函数

在用户层, 每个用户智能体被建模为一个基于 PPO 的策略网络。用户 u_i 在时刻 t 的奖励定义为当轮获得的总激励, 扣除因本轮上传而产生的隐私成本与上传成本, 即:

$$L_i(t) = F(h_i(t), q_i(t), \phi_i(t), y_i(t)) \quad \#(4) \\ - C_i^{\text{priv}}(t) - C_i(t)$$

其中:

F 为用户上传数据的激励收益;

$C_i(t)$ 为上传成本, $C_i(t) = c_i \cdot h_i(t)$, c_i 为用户的单位上传成本系数;

$C_i^{\text{priv}}(t)$ 为隐私成本, 定义为效用函数 P 在隐私数据共享前后的差值, 加上与上传量成正比的风险溢价, 用于刻画数据规模与历史上传行为带来的累积隐私代价:

$$C_i^{\text{priv}}(t) = P(g_i(t)) - P(D_i(t)) + \lambda h_i(t) \quad (5)$$

其中 λ 为单位数据的固定风险系数, $P(g_i(t))$ 为用户持有隐私数据的效用函数:

$$P(g_i(t)) = \alpha_i \log(1 + |g_i(t)|) \quad (6)$$

用户的目标是最大化长期折扣累积效用 $\sum_{t=1}^T \gamma^t L_i(t)$, 其中 $\gamma \in (0, 1)$ 为强化学习中的时间折扣因子, α_i 为用户隐私敏感度因子。

2.2.2 平台层奖励函数

平台层包含单一智能体, 其观测空间聚合了多用户信息和平台池状态。其目标函数定义为生产力、激励成本与参与度三者的加权组合:

$$V(t) = \alpha \cdot \text{Productivity}(t) - \rho \cdot \text{Cost}(t) + \delta \cdot I(t) \quad (7)$$

其中:

$\text{Productivity}(t)$ 为 Shapley 加权生产力:

$$\text{Productivity}(t) = \sum_i (\phi_i * h_i(t)) \quad (8)$$

$\text{Cost}(t)$ 为总激励成本:

$$\text{Cost}(t) = \sum_i (x_i(t) * h_i(t) + y_i(t)) \quad (9)$$

$I(t)$ 为参与度得分, 随上传用户比例与活跃用户数递增:

$$I(t) = \text{ParticipationRate} * \log(1 + \text{ActiveUsers}) / \log(1 + N) \quad (10)$$

$\alpha, \rho, \delta > 0$ 为平台在生产力、成本与参与度之间的权衡参数。综合考虑各组成项数值尺度及预实验结果, 设置 $\alpha = 1.0$, $\rho = 0.5$, $\delta = 2.0$ 。

该回报设计鼓励平台在保证足够多用户参与的前提下, 提高高质量数据的上传量, 同时控制激励成本, 从实质上优化平台的成本效率。考虑数据作为可持续生产要素的特殊性, 参与度奖励项的设计用于避免平台激励资源长期过度集中于少数高贡献用户, 以促进不同类型用户的均衡参与, 维护数据

共享生态的稳定性和可持续性。平台通过动态调整二维激励向量 $\{x_i(t), y_i(t)\}$, 将边际数据价值回报与基础参与奖励相结合, 在鼓励高价值数据贡献的同时兼顾整体参与水平, 引导用户群体形成长期稳定的差异化均衡状态。

2.3 动态 Stackelberg 博弈

整个交互过程构成一个无限时域的动态 Stackelberg 博弈。在每一轮 t , 平台作为领导者率先为每个用户发布二维激励向量 $\mathbf{R} = \{x_i(t), y_i(t)\}$, 用户作为跟随者在观测到 $\mathbf{R}$ 后, 选择最优上传策略。上传完成后, 平台接收各个用户数据摘要、更新全局策略, 并计算各用户本次贡献的近似 Shapley 值 $\phi_i(t)$, 这一新计算的 ϕ 值将影响平台智能体的下一轮边际激励决策 $x_i(t+1)$, 全局参与度则会影响固定激励决策 $y_i(t+1)$, 形成完整的闭环动态反馈, 如图 1 所示。

与传统静态 Stackelberg 博弈不同, 本文框架中的双层博弈具有以下显著动态特性:

(1) 激励向量 $\mathbf{R}$ 对每个用户是异质的, $x_i(t)$ 与 $\phi_i(t)$ 正相关, 体现了高质量数据高回报的定价原则; 而 $y_i(t)$ 受全局参与度 $I(t)$ 调控, 避免了对低贡献用户的过度激励或对整体参与的过度抑制。

(2) 用户私有数据池 $D(t)$ 随时间持续增长, 用户隐私成本 $C_i^{\text{priv}}(t)$ 呈现非线性累积, 导致效用函数随时间路径依赖; 平台收益 V 不仅取决于当期上传, 还受历史数据质量的长期影响。

(3) 数据池增长、模型性能提升以及近似 Shapley 采样过程均引入随机性, 使博弈呈现部分可观测随机博弈的特性。

理想的均衡概念为动态 Stackelberg 均衡。在折扣无限时域设定下, 平台的长期目标是选择激励策略 θ 以最大化自身预期折扣累积收益, 同时诱导用户采用最佳响应策略 π^* :

$$\theta^* = \operatorname{argmax}_{\{\theta_i\}} E \left[\sum_{t=0}^{\infty} \gamma^t L_{\text{platform}}(t) \mid \pi^* \right] \quad (11)$$

其约束条件为用户在给定平台策略下的最优响应:

$$\pi_i^* = \operatorname{argmax}_{\pi_i} E \left[\sum_{t=0}^{\infty} \gamma^t L_{\text{user}_i}(t) \mid \theta_i \right], \quad \forall i = 1, \dots, N \quad (12)$$

$L_{\text{platform}}(t)$ 和 $L_{\text{user}_i}(t)$ 分别为第 t 轮的即时效用, 定义见 3.2 节。期望 E 考虑随机性, 包括数据

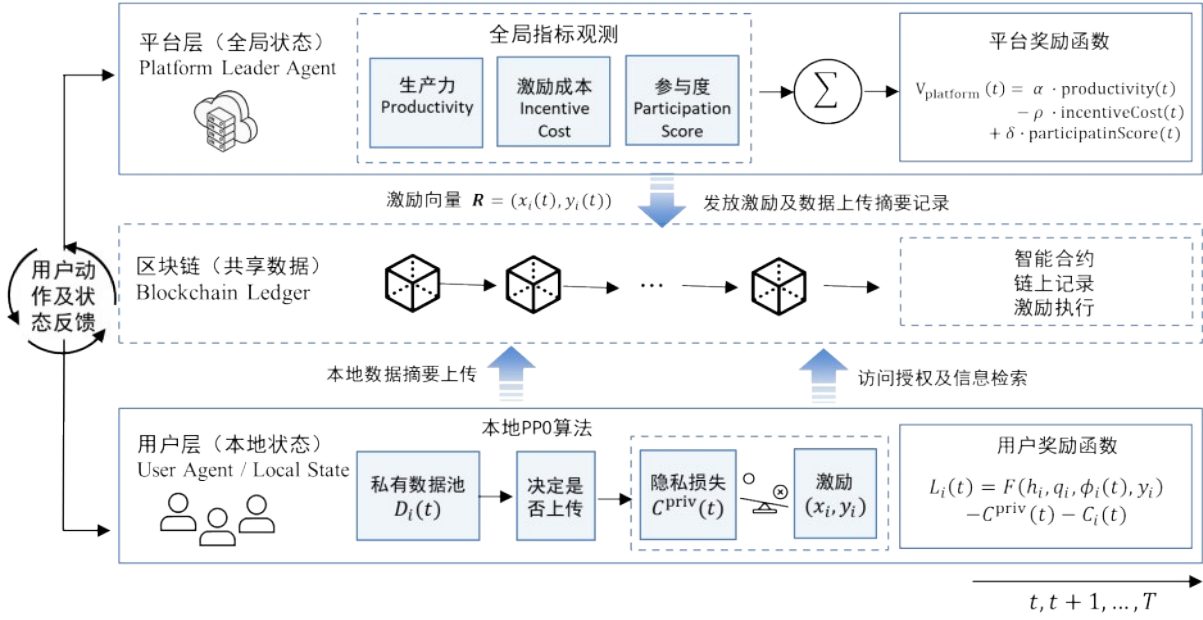


图1 双层动态博弈模型

池增长的随机过程、模型性能的随机波动以及近似 Shapley 计算的采样噪声。

上述定义将经典 Stackelberg 博弈扩展具有随机状态演化的动态环境，与反馈 Stackelberg 解或随机 Stackelberg 博弈^[24]建模思路类似。现有关于动态 Stackelberg 博弈的研究多局限于线性二次结构、可解析需求函数或有限状态马尔可夫决策过程，且往往假设完美信息或静态成本函数。但在常规的数据市场与平台激励场景中，由于用户行为的复杂异质性、隐私成本的时变累积以及个性化激励需求，这些简化假设难以成立。少数结合强化学习的工作虽尝试通过单层强化学习逼近领导者策略，但通常忽略用户的自主学习能力，导致在长期交互中无法有效适应策略的动态演化与用户异质响应。

3 双层强化学习框架

为求解第 2.3 节形式化的动态 Stackelberg 博弈问题，本文提出一个双层强化学习框架，将平台—用户长期交互过程建模为一个多智能体马尔可夫博弈，并采用层次化结构来处理领导者—跟随者的不对称决策关系。该框架无需预先精确的用户模型，通过数据驱动的持续交互自动逼近近似 Stackelberg 均衡。

3.1 多智能体马尔可夫决策过程建模

本文将数据驱动平台与 N 个异质用户的交互过

程建模为一个可观测的多智能体马尔可夫博弈。方案的双层强化学习框架如图 2 所示。该博弈为领导者—跟随者结构，由一个平台领导者智能体和 N 个用户跟随者智能体组成，定义如下：

状态空间 $\mathcal{S}$ ：全局状态 $\mathbf{s}(t) \in \mathcal{S}$ 包含平台侧可观测信息与用户侧私有信息摘要。具体包括：平台全局模型性能 $V(t)$ 、数据池状态、所有用户的近似 Shapley 边际贡献量 $\phi_i(t)$ 、全局参与度 $I(t)$ 、各用户的历史激励记录序列、用户聚类标签以及行为统计摘要。为降低维度灾难，用户私有数据池规模 $D_i(t)$ 、精确构成以及累积隐私损耗等高维信息仅由各用户智能体本地维护，平台仅通过摘要观测。

动作空间 A ：动作空间 $A = A_{\text{platform}} \times A_{\text{user}}$ 由两类智能体组成。平台智能体动作 $\mathbf{a}_{\text{platform}}(t) = (\mathbf{R}_i(t))_{i=1}^N$ ，其中 $\mathbf{R} = \{x_i(t), y_i(t)\}$ 为对用户 u_i 的个性化二维激励向量。边际激励 $x_i(t)$ 主要由用户 u_i 的边际贡献：近似 Shapley 值 $\phi_i(t)$ 动态决定，固定激励 $y_i(t)$ 则依据全局用户参与度 $I(t)$ 调控，用于维持基础参与率并覆盖低价值用户的上传成本。用户智能体动作 $\mathbf{a}_{\text{user}_i}(t) \in [0, g_i]$ ，表示用户 u_i 的上传数据量（或二元上传决策：上传与否、上传多少条记录）。用户可根据本地私有状态独立决策。状态转移方程为：

$$\mathbf{s}(t+1) \sim T(\mathbf{s}(t), \mathbf{a}_{\text{platform}}(t), \{\mathbf{a}_{\text{user}_i}(t)\}_{i=1}^N) \quad (13)$$

由以下因素共同决定:各用户私有数据池的随机增长过程(模拟真实行为日志每日新增);上传行为对用户私有池和平台全局模型的影响(上传后数据摘要融入平台模型,更新 $V(t)$ 与 $\phi_i(t)$);全局参与度 $I(t)$ 的演化;以及episode重置机制(长期训练中周期性重置部分状态以增强泛化)。转移概率体现了环境的不确定性与路径依赖特性。

奖励函数:同2.2.1节式(4)、式(5),平台回报关注生产力提升、激励成本控制与整体参与率的权衡。而用户回报 $L_i(t) = F(h_i, q_i, \phi_i(t), y_i) - C_i^{\text{priv}}(t) - C_i(t)$,目标为最大化个人隐私-收益权衡下的长期累积效用。

本文多智能体马尔可夫博弈建模自然地将动态Stackelberg博弈转化为可通过强化学习求解的形式,同时融入了个性化激励机制: x_i 依赖个体边际贡献, y_i 依赖全局参与度。

3.2 层次化交替学习方案

为降低多智能体同时训练的复杂性,且由3.1节可知,平台与用户本身即具有决策者与跟随者的层级关系,整体博弈分解为上层平台决策与下层用户响应的双层架构。故可通过交替冻结策略的训练机制,实现上下层策略的解耦与稳定收敛。算法伪代码见算法1。

下层用户层每个用户智能体运行仅基于自身状态及其观察向量(自身数据池、历史接收激励、累积隐私损耗、近期 ϕ)和平台下发的个性化激励 R 学习上传策略。该层目标是快速适应给定激励,实

算法1 双层交替训练算法

```

输入: 总训练步数 $T$ , 阶段长度 $K$ , 环境 $\mathcal{E}$ , 用户策略 $\pi^U$ , 平台策略 $\pi^P$ 
输出: 训练后的 $\pi^U, \pi^P$ 
1 构建双层层级: 平台层 $\mathcal{P} \rightarrow$  用户层 $\mathcal{U} \rightarrow \mathcal{E}$ 
2 初始化:  $\psi \leftarrow 0, \tau \leftarrow 0$  //  $\psi$ : 训练阶段编号
3 FREEZE( $\mathcal{P}$ )  $\leftarrow$  TRUE, FREEZE( $\mathcal{U}$ )  $\leftarrow$  FALSE // 初始: 训练用户层
4  $\mathcal{E}.$  RESET()
5 for  $t \leftarrow 1$  to  $T$  do
6   if  $\tau \geq K$  then // 阶段长度 $K=2048$ 
7      $\psi \leftarrow 1 - \psi, \tau \leftarrow 0$  // 每 $K$ 步切换阶段
8     if  $\psi = 0$  then
9       FREEZE( $\mathcal{P}$ )  $\leftarrow$  TRUE
10      FREEZE( $\mathcal{U}$ )  $\leftarrow$  FALSE // 阶段 0: 训练用户层
11    else
12      FREEZE( $\mathcal{U}$ )  $\leftarrow$  TRUE
13      FREEZE( $\mathcal{P}$ )  $\leftarrow$  FALSE // 阶段 1: 训练平台层
14    end
15  end
16  if FREEZE( $\mathcal{U}$ ) = FALSE then
17    对用户层执行 PPO 更新
18  end
19  if FREEZE( $\mathcal{P}$ ) = FALSE then
20    对平台层执行 PPO 更新
21  end
22  if done then
23     $\mathcal{E}.$  RESET() // 默认 $H_{\max} = 200$ 步/episode
24  end
25   $\tau \leftarrow \tau + 1$ 
26 end
27 return  $\pi^U, \pi^P$ 

```

现自身效用的最大化。上层平台层包含一个全局的激励决策智能体。该智能体观测全局状态(包括所

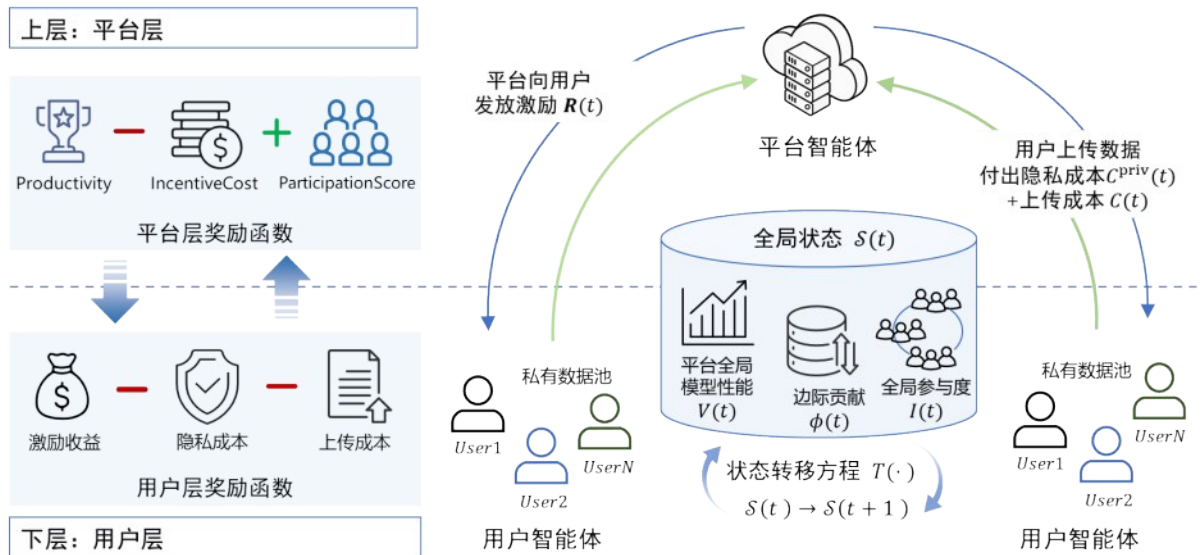


图2 双层强化学习框架

有用户的上传反应、用户参与度 $I(t)$ 以及各用户 Shapley 贡献), 输出个性化激励向量 R_i 。上层通过激励直接影响下层行为, 而下层的集体响应 (上传数据数量与质量) 又反过来塑造上层的长期奖励及效用, 形成闭环反馈。

上层与下层通过交替冻结训练策略进行优化: 在训练的某一阶段, 仅更新下层用户智能体的策略参数, 同时冻结平台策略, 用户学习在当前固定激励下的最佳响应。下一阶段, 仅更新上层平台智能体的策略参数, 冻结用户策略, 由平台学习如何最优地设置激励以诱导理想的下层用户行为。这种交替机制可有效缓解两层策略互相耦合带来的训练不稳定性, 同时加速了向近似 Stackelberg 均衡的收敛。训练过程中, 平台仅接收 Shapley 贡献摘要与模型更新梯度, 而非原始数据, 实现对用户原始数据的隐私保护。

3.3 基于联盟链的分布式可信执行架构

针对隐私数据共享场景, 本文采用基于联盟训练的分布式部署方案以降低集中式存储带来的隐私泄露风险。每个用户智能体仅运行于本地设备或边缘节点, 仅向区块链账本上传 Shapley 贡献摘要与模型更新梯度, 将激励计算、结算与存证的全流程纳入链上治理体系, 在不暴露用户原始数据的前提下实现激励机制的透明执行。

系统整体架构分为应用层、合约层、网络层与存储层四个层次, 如图 3 所示。(1) 应用层封装数据共享的业务逻辑并提供用户交互接口, 支持数据上传、访问授权及检索等核心操作; 用户首先在区块链系统中完成身份注册以获取链上账户, 作为后续交互的唯一可信凭证。(2) 合约层负责链上规则管理与激励结算, 包括用户注册、数据授权、访问控制和激励分发等智能合约, 合约条款公开透明且不可篡改, 参与方无法对结算结果进行单方面干预, 从而在制度层面保障了激励发放的公平性。其中激励结算上链执行算法见算法 2。(3) 网络层由联盟区块链网络构成, 底层采用 FISCO BCOS 联盟链架构, 该架构在隐私性与可扩展性之间具有良好的均衡特性, 原生支持国密算法 (SM2/SM3/SM4), 适用于多机构间隐私数据共享的监管合规场景; 共识机制采用 PBFT 算法, 在保证拜占庭容错能力的同时将确认延迟控制在可接受范围内, 以满足激励结算的时效性需求; 网络层同时集成 P2P

协议构建对等网络, 通过分布式消息协议支撑高并发条件下激励摘要的实时上链。(4) 存储层采用链上与链下相结合的混合存储结构^[32]: 用户原始行为数据存储于星际文件系统 (IPFS) 中, 经内容寻址与切片处理后分发至地理上分散的对等节点, 确保单一节点无法获取完整数据文件; 数据索引、内容哈希证明及访问控制策略等元信息则上链存储于分布式账本中, 确保数据完整性的可验证性与操作记录的可追溯性, 同时显著降低链上存储开销。

在上述架构基础上, 本文进一步将数据最小化

算法 2 激励结算上链执行算法

输入: 平台地址 Plat, 用户集合 $U=\{u_i\}_{i=1}^N$, 上传数据 $\{D_i\}$, 上传量 $\{h_i\}$, 质量分 $\{q_i\}$, 激励参数 $\{(x_i, y_i)\}$, 隐私损失 $\{P_i\}$, 上传成本 $\{C_i\}$, 定点缩放 SCALE=10⁶

输出: 结算轮次 R, 结算单集合 $\{Sid_i\}$, 批量哈希 H_{batch} , 交易回执集合 $\{Tx_i\}$

```

1  函数 ComputeReward(x, y, h, q, P, c) // 链上自动执行激励规则
2      linear  $\leftarrow (x \cdot q / \text{SCALE}) \cdot h / \text{SCALE}$ 
3      r  $\leftarrow \max(0, \text{linear} + y - P - c)$ 
4      return r
5  部署合约 C  $\leftarrow \text{Deploy}(\text{IncentivePlatform})$ ; C.AddPlatform(Plat)
   // 仅部署者可调用
6  for each  $u_i \in U$  do
7      C.AddUser( $u_{i.addr}$ , namei, cluster) // [身份登记]用户上链
8  end
9  R  $\leftarrow$  C.OpenRound() // [开轮]平台开启结算轮次
10 for i  $\leftarrow$  1 to N do
11     if 用户  $u_i$  未注册 then // [□ 用户身份校验]
12         reject 「未注册用户」; continue
13     end
14      $\delta_i \leftarrow \text{keccak256}(D_i)$  // [上传存证]链下→IPFS, 链上存
   哈希
15      $UID_i \leftarrow$  C.SubmitUpload( $\delta_i$ ,  $h_i \cdot \text{SCALE}$ ,  $q_i \cdot \text{SCALE}$ )
16     ( $\text{settled}, \_$ )  $\leftarrow$  C.IsSettled(R,  $u_i$ ) // [□ 重复结算检查]
17     if settled = TRUE then
18         reject 「本轮已结算」; continue
19     end
20      $\hat{r}_i \leftarrow \text{ComputeReward}(x_i, y_i, h_i, q_i, P_i, c_i)$ 
   // [□ 激励结果计算 (链下申报)]
21      $Sid_i \leftarrow$  C.SubmitIncentiveResult(R,  $u_i$ ,  $UID_i$ ,  $x_i, y_i, P_i, c_i$ ,  $\hat{r}_i \cdot \text{SCALE}$ )
   // [□ 提交+链上验算]
22     if Status( $Sid_i$ ) = REJECTED then // [异常处理]申报≠链上
   规则
23         记录异常; continue
24     end
25      $Tx_i \leftarrow$  C.ConfirmSettlement( $Sid_i$ )
   // [□ 结算结果上链确认]可审计可追溯
26 end
27  $H_{batch} \leftarrow \text{keccak256}(\{Sid_i, \hat{r}_i\}_{i=1}^N)$ 
28 C.CloseRound(R,  $H_{batch}$ ) // [封账]整轮批量哈希上链
29 return R,  $\{Sid_i\}$ ,  $H_{batch}$ ,  $\{Tx_i\}$ 

```

原则贯穿于双层强化学习训练的全流程以强化隐私保护。具体而言,每个用户智能体仅运行于本地设备或边缘节点,平台智能体仅接收经本地计算后的 Shapley 贡献摘要与模型更新梯度,而非用户的原始敏感数据;所有激励结算记录经智能合约链上公开执行后写入不可篡改的分布式账本,既保障了用户隐私,又使平台的激励行为处于多方可审计的透明环境之中。四层架构与隐私保护机制的协同设计,使本文框架在实现高效差异化激励的同时,满足实际数据驱动平台对隐私合规与可信执行的双重部署要求。

本文设计的链上数据索引与链下隐私数据离链存储相结合的混合结构,在保证数据完整性与可追溯性的同时,有效降低系统的链上存储成本。

4 实验验证

4.1 实验设置

4.1.1 用户聚类与异质性建模

本文实验数据来源于京东电商平台行为日志数据集。原始数据以用户 ID 与行为时间为主键进行排序,每条记录涵盖以下关键字段:行为时间(actionDate)、行为类型(type,包括点击、加入购物车、购买、评论等)、商品类目(category)、品牌(brand)、店铺标识(shopId),以及用户其他静态属性(年龄段、性别、注册时长、会员等级、城市等级等)与店铺属性(注册时间、综合评分、粉丝数量、VIP 用户数等)。

本文采用 K-Means 聚类与高斯混合模型(gaussian mixture model, GMM)的两阶段方法对原始数据集进行用户划分与数据生成。首先在数据集上对用户进行聚类分析与特征工程处理,构建用户异质性指标,所提取的核心指标如表 3 所示。

表 3 数据集核心指标说明

指标	描述
platformInterval	用户与平台交互的平均时间间隔
activeDaysRatio	观测期内用户活跃天数占比
type	用户不同的行为类型
categoryEntropy	用户行为所覆盖的商品类目
uniqueBrands	用户交互过的独立品牌数量

原始数据集中的 2,215 名用户实施 K-Means 聚类后,其划分为四个用户群体,具体如图 4 所示:簇 0 为高活跃度用户群,样本数量最多;簇 1 在各项特征指标上均处于最低水平,代表低价值用户群体;簇 2 在数据质量维度价值较高;簇 3 在综合指标上表现最优,为最高价值用户群。在此基础上,针对每一簇分别构建 GMM,对“行为时间—行为类型—商品类目”的联合分布进行建模,并依据用户所属簇的条件分布对行为时间与行为类型进行采样;商品及店铺等字段由对应簇内的真实记录填充,用户静态画像属性则保留该用户的真实观测值。上述流程在维持用户间活跃度差异的同时,确保合成数据的时间分布与行为模式均与各簇的统计

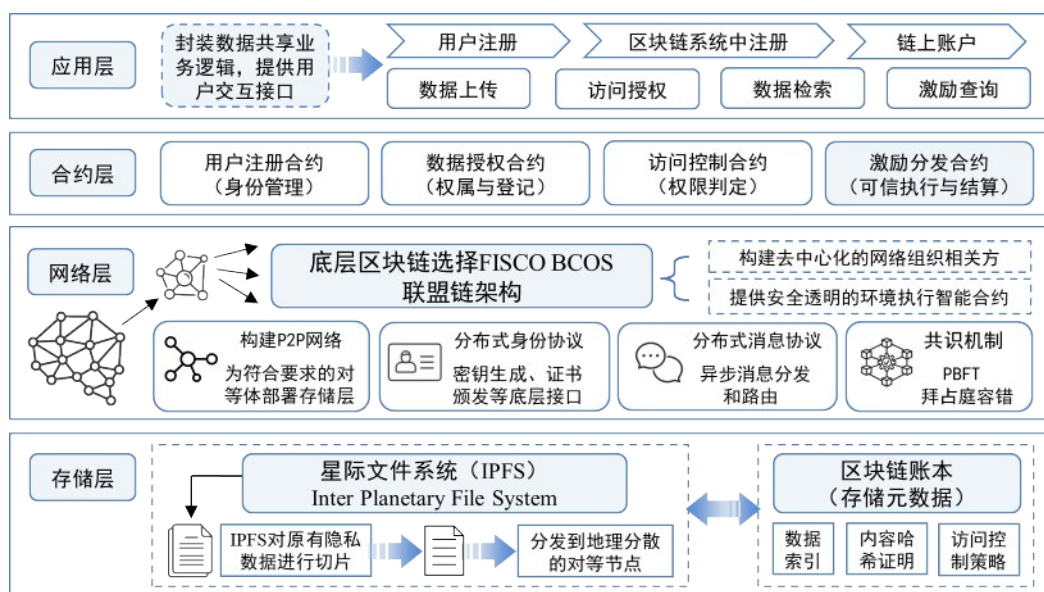


图 3 基于联盟链的可信数据共享系统分层架构

特性保持一致。

提取的聚类指标在实验中承担两方面的功能：一方面作为实验环境初始化的依据，为每位用户分配具有个体差异的私有数据池与行为参数；另一方面，现实世界中用户数据不是一成不变的，随着时间和与平台交互行为的增多，用户数据也在动态增长。故系统要对后续的数据增长过程与激励敏感性建模。具体而言，活跃天数更多及使用平台间隔越小的用户，数据连续性更高，用户决策更连贯数据价值更高；购买商品种类（categoryEntropy）较高的用户其兴趣分布更为分散，所贡献数据在类目覆盖维度上的多样性更强，对平台模型泛化能力的边际贡献相应更为显著；评论率（commentRatio）较高的用户所上传的行为记录包含更为丰富的语义信息，对基于 Shapley 值的贡献度量化结果具有重要影响。上述机制共同保障了实验中用户异质性的合理刻画，为后续激励策略的系统性验证提供了具有现实可信度的仿真基础。

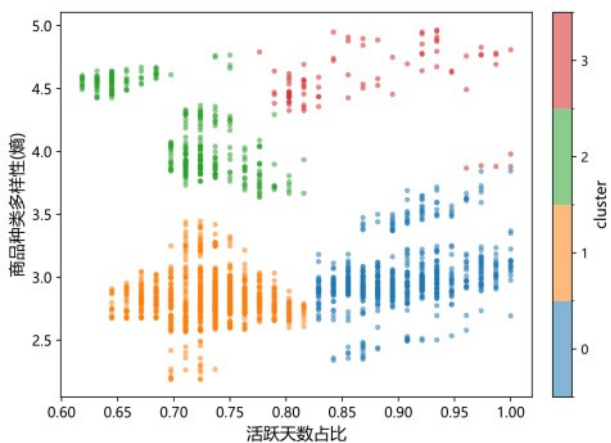


图4 各簇用户在关键特征平面上的对比

4.1.2 用户数据动态增长

现实场景中，用户可贡献的数据并非固定不变，而是随时间持续积累与演化。为准确模拟这一特性，本文设计了用户数据动态增长模块，将每位用户的可用数据建模为一个随时间动态扩张的私有数据池。

(1) 初始数据池。以该用户原始数据集中的历史行为记录作为数据池的初始状态，构成仿真起点。

(2) 用户偏好估计。从用户历史记录中统计提取商品类目、品牌及店铺的经验分布，以及日均行

为数等个体行为参数，作为后续数据生成的先验依据。

(3) 数据增长机制。在每一仿真时间步（对应现实中的一日或固定时间间隔），依据以下规则生成新行为记录：依据日均行为数参数进行泊松采样，确定该时间步的期望生成数量；类目、品牌与店铺字段采用混合采样策略—70%依据用户自身历史偏好分布采样以延续个体行为模式，30%依据全局分布采样以引入适度探索与多样性；具体行为记录则从匹配对应类目、品牌及店铺的历史记录池中随机抽取生成。

新生成的行为记录即时并入用户私有数据池，并用于实时更新该用户的偏好分布估计。在整个训练过程中，各用户数据池持续增长，在结构上保持个体行为偏好的延续性，同时允许一定程度的随机多样化。上述动态数据生成机制为激励策略的长期效果评估提供了更贴近真实场景的仿真环境，有效规避了现有研究中普遍采用静态数据集所引入的时序偏差。

4.1.3 平台数据池与数据价值评估

与用户私有数据池相对应，平台数据池汇聚用户自愿上传的数据记录，用于衡量激励框架所能激发的用户数据贡献总量及其对平台模型能力的提升效果。

在每一时间步，用户依据平台下发的个性化激励向量决策其上传数量，系统将所选定的私有数据记录汇入平台数据池。对用户每次上传的数据记录，依据第2.1节所述方法计算其 Shapley 边际贡献值，将用户在当前时间步的上传记录集合定义为上传集合，以此量化该批数据对平台模型的边际价值。

为防止数据池无限增长导致内存溢出，本文引入滑动窗口机制对平台数据池进行管理：在每个 episode 间隔，平台数据池选择性保留近期历史窗口内的数据，以控制内存占用并调节时序依赖强度。该设计既符合平台对数据的实际利用方式，亦为 Shapley 边际定价提供了计算可行性保障。

同2.2节定义，Shapley 边际贡献值作为边际激励分量，与固定激励共同构成平台对用户单次上传行为的完整激励向量。平台智能体的即时回报定义为模型性能增益与激励总成本之差，并附加参与率奖励项，以在短期收益最大化与长期用户留存之间

实现动态平衡;用户智能体的即时回报则基于个人效用函数,定义为激励收益扣除隐私累积成本与上传操作成本后的净收益。

4.2 习得的稳定策略

本文在原始数据集中各簇随机挑选8位用户共32位用户,主实验采用PPO算法实现双层交替训练,总训练步数1000k步(即50k轮),算法各超参数设置见表4。为验证双层结构及交替训练机制的作用,类标准对称MARL(上下层智能体同时持续更新)、单层RL(仅平台智能体更新)以及静态隐私模型作为对比基线,所有基线均在相同环境与数据集上复现,确保对比公平性。各训练agent与原始数据集user的映射关系见表5。具体而言,每个agent绑定对应的数据集customerId,并读取该用户的年龄段、会员等级等用户画像属性以及active-DaysRatio、platformLevel等行为特征,分别用于构造隐私敏感度参数 $\alpha_i \in [0.5, 1.5]$ 和单位上传成本 $c_i \in [0.1, 0.3]$ 。

训练指标如图5所示,从400k步开始平台习得较稳定策略,收益与成本曲线趋于收敛。此时簇3

用户(即高价值用户)的收益高于其他用户。前期平台激励用户上传数据,维持较高参与率的同时上传成本也在急剧增加,直到成本与收益达到一个稳态。平台数据池也趋于稳定。在相同训练环境和训练步数下,交替冻结在后段平台奖励、生产力与成本效率上优于同步更新MARL与单层RL基线。具体而言,单层RL长期处于低绩效平台,难以形成有效的“激励—行为”正反馈,验证了双层闭环与用户侧学习是必要的。训练稳定性以平台奖励的滑动标准差衡量,双层框架整体显著稳于单层RL;相对同时更新的MARL,本文交替冻结有利于缓解双层同更带来的非平稳抖动,奖励轨迹波动更小。

直到700k步后即训练后期,由图5中激励成本可看出激励用户上传相同数据的成本过高,由图6可以看出用户收益也在剧烈上升。平台转而去激励簇0和簇2的普通用户。以每个簇中第1位用户为例:簇0为活跃的普通用户,同时持有的数据较多,其上传行为贯穿训练始终;簇1大多是低价值数据,平台对该类低边际贡献用户分配的激励相对较低。直到簇0和簇2可进一步贡献的高价值数据

表4 算法各超参数设置表

参数	用户层	平台层
学习率	1×10^{-4}	5×10^{-5}
裁剪系数	0.2	0.1
Batch Size	2048	4096
折扣因子 γ	0.99	0.99
损失系数	0.7	1.5
熵系数	0.001	0.01
梯度裁剪	0.5	0.25
网络结构	MLP(64-64)+Tanh	MLP(64-64)+Tanh

表5 各Agent与数据集user映射关系

Cluster	Agent (Uid)			
0	user0 (153327)	user1 (1132533)	user2 (1332397)	user3 (1060064)
	user4 (708678)	user5 (715040)	user6 (155650)	user7 (1219573)
1	user8 (689717)	user9 (1243059)	user10 (795592)	user11 (580109)
	user12 (193797)	user13 (1314527)	user14 (774989)	user15 (282111)
2	user16 (1295561)	user17 (73745)	user18 (330172)	user19 (1265979)
	user20 (838551)	user21 (641653)	user22 (1352335)	user23 (99717)
3	user24 (539905)	user25 (829481)	user26 (1192419)	user27 (1568787)
	user28 (1302533)	user29 (1499427)	user30 (768483)	user31 (1045098)

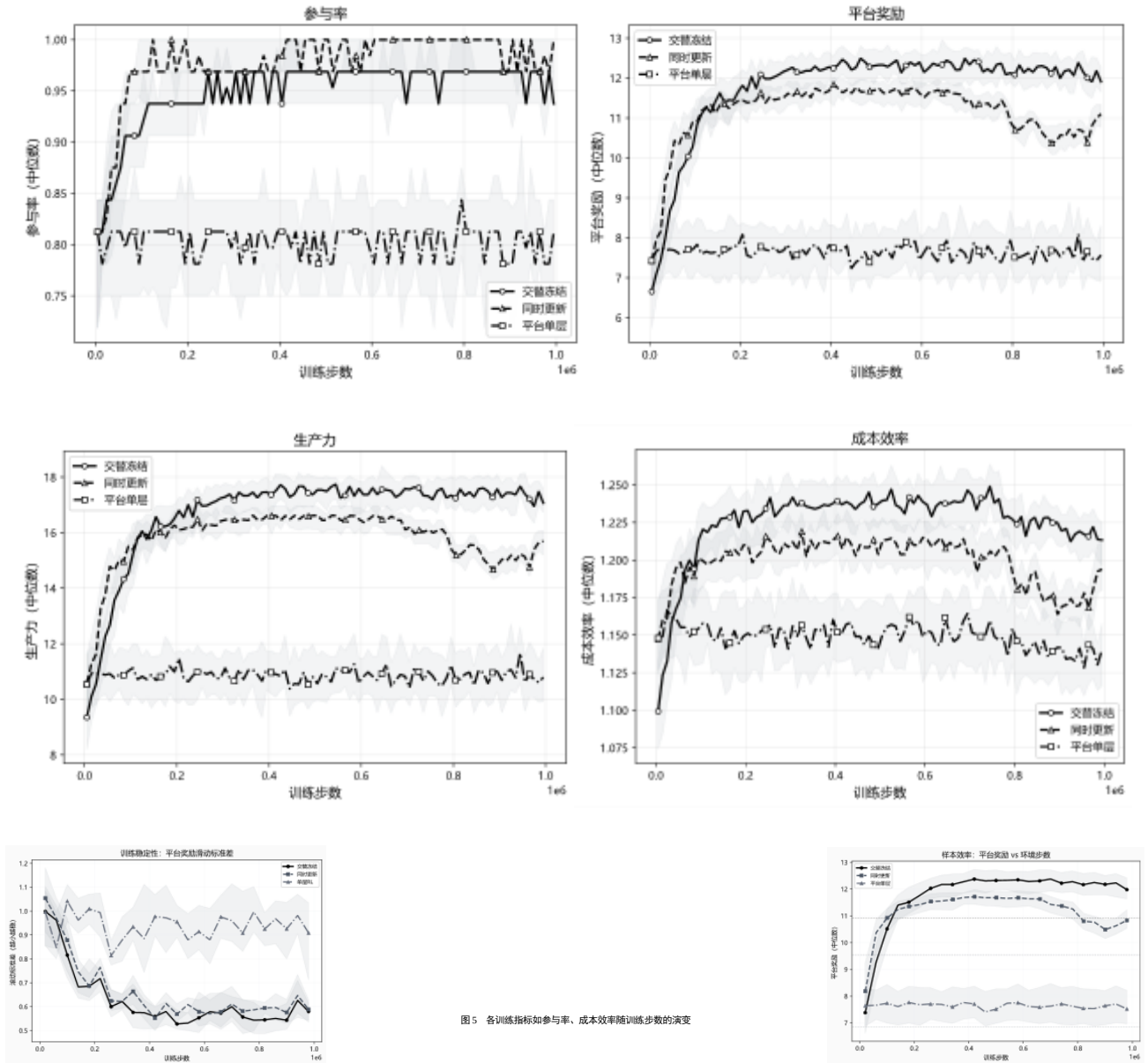


图5 各训练指标如参与率、成本效率随训练步数的演变

逐渐减少。这种各簇用户上传收益分段增加现象体现了框架的一大特性：对于不同类型的用户有很好的异质化激励。也表现了平台 agent 对于不同下层环境拥有良好的动态感知及适应性。

为刻画双层激励框架在训练过程中平台侧政策的阶段性变化，本文分别取训练序列的前 10% 与后 10% 记录点，统计其均值与取值范围，在二维动作空间分别对应平台输出的平均激励中对比训练初期与训练后期的政策分布，如图 7 所示。

在实验中平台政策呈现出明显的阶段性漂移： x 的均值由 0.375 上升至 0.618， y 的均值由 0.107 上升至 0.331。同时，后期在 y 维度的波动显著增大

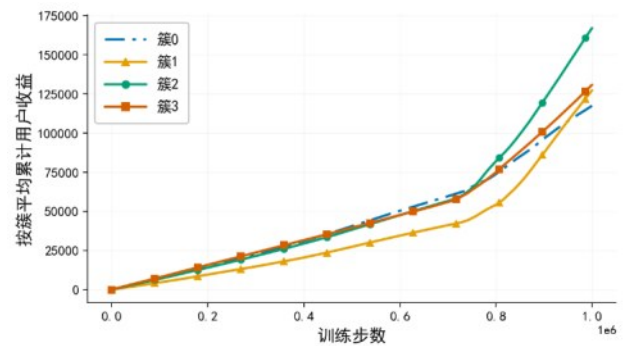


图6 各簇用户收益随训练步长的演化

(标准差由 0.011 增至 0.231)，表明平台后期会在更高的固定激励水平附近进行更大幅度的策略调整，

从而形成与训练初期显著不同的政策表现。平台政策初期整体更接近稳态:符合平台初期用户参与度较高,平台无需提高固定激励。训练过程中,平台侧政策表现为“早期即收敛的小幅微调”至“随训练推进发生显著漂移并在后期保持更高激励水平”的两类典型政策。该政策的转向体现了平台侧对于用户上传行为的适应性。

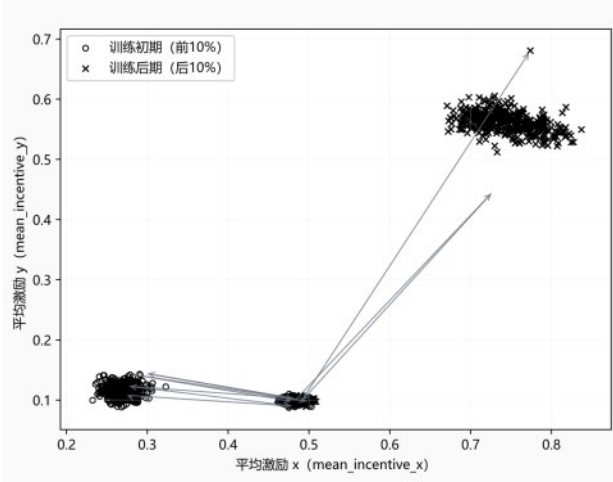


图7 平台初期至后期策略的阶段性变化

4.3 对比基线设计

为系统验证所提双层强化学习激励框架的有效性与机制设计的必要性,本文设计了两类共四种对比基线实验,从平台层奖励、激励成本与生产力激发三个维度对 BiRL 进行机制对比实验,分别针对现有激励机制研究中常使用的“静态定价”及“评分机制”。对比指标如平台奖励,激励成本,生产力定义同 2.2.2 节。为更好地模拟真实世界中平台初期,各个实验方案在训练初期至收敛阶段对平台生产力的激发与用户的激励效果,不同于 4.2 节用于验证长期策略演化、收敛趋势及稳定性的长期训练实验设置,本节进一步设置独立的固定训练预算对比实验,用户规模、数据集、环境参数及评价指标一致。本节方法及所有对比基线统一采用 100k 训练步长配置,4 簇 32 名用户参与。

4.3.1 Linear-Price:静态线性定价

Linear-Price 采用按上传条数线性定价,用户在每一时间步所获激励仅与其上传记录数量成正比。

$$F_i = p \times \text{records}_i \quad (14)$$

其中 p 为平台设定的统一单价参数。基准 p_0 设置为

BiRL 框架实验中的总激励成本/总上传条数,见 4.2 节。

该对照组 Shapley 贡献评估与数据质量等机制,以构造“按量付费基线”。为分析激励总量对实验结果的影响,本文在单价维度上设置三个档位— $p0.8$ 、 $p1.0$ 与 $p1.2$ (分别对应基准单价 p_0 的 0.8 倍、1.0 倍与 1.2 倍),从而回答如下问题:BiRL 框架的性能提升是否仅源于更高的等效激励总量,而非机制设计本身的有效性。

4.3.2 Budget-Proxy:“质量分”基线

平台仅对质量进行打分,依据不同数据质量分数在用户间进行分配激励。为排除总预算规模之间的差异对实验结果的干扰, Budget-Proxy 首先将每步总预算与 Linear-Price 在基准单价 p_0 下的等效支出进行对齐:

$$B = p_0 \times \sum_i \text{uploadRecords}_i \quad (15)$$

在此基础上,为每位用户计算代理质量分 q_i ,经 min-max 归一化至 $[0,1]$ 后,按如下线性形式分配预算:

$$F_i = f_{\text{fixed}} + m \cdot q_{\text{norm},i} \cdot \text{uploadRecords}_i \quad (16)$$

$$\sum_i F_i = B$$

该基线表示“不依赖 Shapley 值、以简单打分制指标替代精确贡献评估”的方案。通过预算对齐设计, Budget-Proxy 与 Linear-Price ($p1.0$ 档) 在总支出规模上保持可比,使实验能够独立评估贡献评估精度对激励效果的影响。若 BiRL 显著优于 Budget-Proxy,则表明基于 Shapley 值的精确边际贡献评估带来了不可替代的机制增益;反之则说明机制复杂性的边际收益有限。

4.4 实验结果与分析

BiRL 框架与四种对比基线在平台奖励、激励成本、生产力和成本效率四项核心指标上的后半程均值对比结果参见图 8~图 11。

(1) 平台奖励如图 8 所示, BiRL 实现正收益,线性定价基线全面亏损。在平台奖励指标上, BiRL 框架取得约 +5.6 的正值收益;而 Linear-Price 三个定价档位 ($p0.8$ 、 $p1.0$ 、 $p1.2$) 的平台奖励分别为 -146.3、-186.2、-224.9,均呈显著亏损,且亏损幅度随单价提升单调增大。该结果表明,单纯提高激励单价不仅无法改善平台收益,反而加剧了

激励成本的无节制累积。Budget-Proxy 的平台奖励为-187.6，与 Linear-Price (p1.0) 处于同一量级，说明以代理质量分替代近似 Shapley 值进行激励分配，不能有效解决激励效率问题。

(2) 后半程激励成本均值如图 9 所示，BiRL 实现极低成本运营，基线成本高出 45 到 70 倍。具体而言，BiRL 后半程逐步激励成本均值约为 6.62，而 Linear-Price 各档分别约为 307.4 (p0.8)、387.5 (p1.0) 和 465.0 (p1.2)，Budget-Proxy 约为 385.8；基线均值约为 386，即 BiRL 的约 58 倍。与之对应，训练全程总激励成本上 BiRL 累计支出约为 1.3×10^4 ，LP/Proxy 则约为 $5.46 \times 10^5 - 8.19 \times 10^5$ 。该结果从成本角度揭示了图 8 中基线方案持续亏损的根本原因——激励支出随训练步数线性累积且不受预算约束，最终大幅超出平台收入。BiRL 通过 Shapley 值精确识别高贡献用户并实施差异化激励，从根本上避免了对低贡献用户的无效支出。

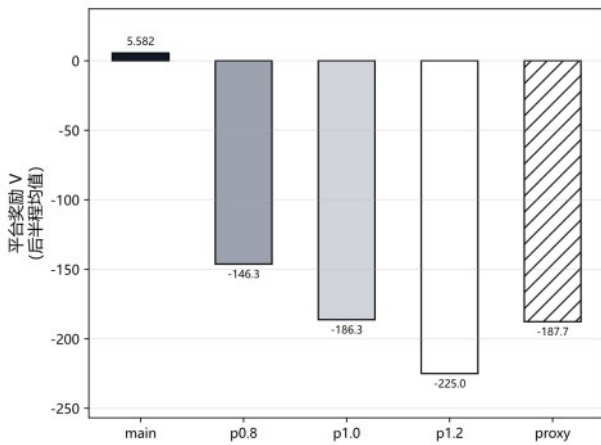


图8 各方案平台奖励对比

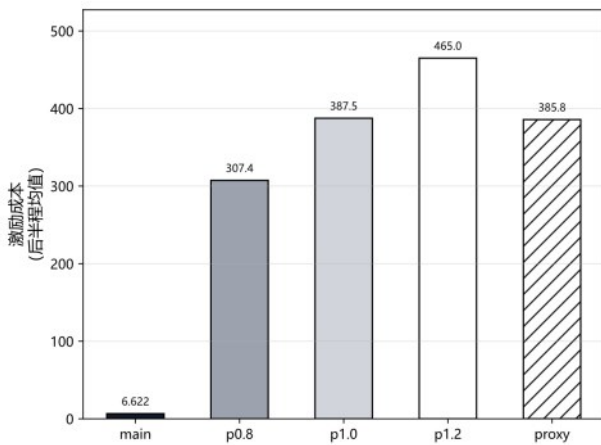


图9 训练后半程激励成本对比

(3) 各生产力对比如图 10 所示，BiRL 在最低成本下实现最优生产力。在统一生产力指标上，BiRL 框架的后半程均值约为 8.15，显著高于所有基线方法；Linear-Price 三个档位分别约为 6.74 (p0.8)、6.76 (p1.0) 和 6.77 (p1.2)，Budget-Proxy 约为 6.74，四者差异甚微、彼此接近。值得注意的是，从 p0.8 到 p1.2，激励预算成倍增长，而生产力提升幅度不足 0.4%，表明更高的激励支出并未带来相应的贡献质量提升。

(4) 成本效率如图 11 所示，四种基线方法的数据点则集中于右下区域，呈现高投入、低产出的低效格局。两组点之间存在显著的成本鸿沟与明显的生产力落差而 BiRL 在激励成本极低的条件下，生产力仍高出基线约 20%，印证了基于 Shapley 值的精准激励对用户贡献效率具有实质性驱动作用。

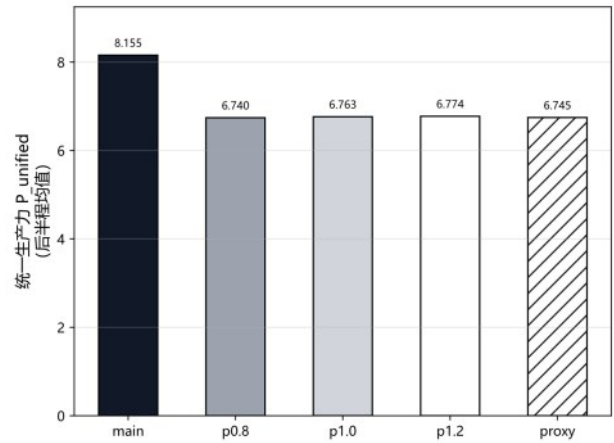


图10 各方案生产力对比

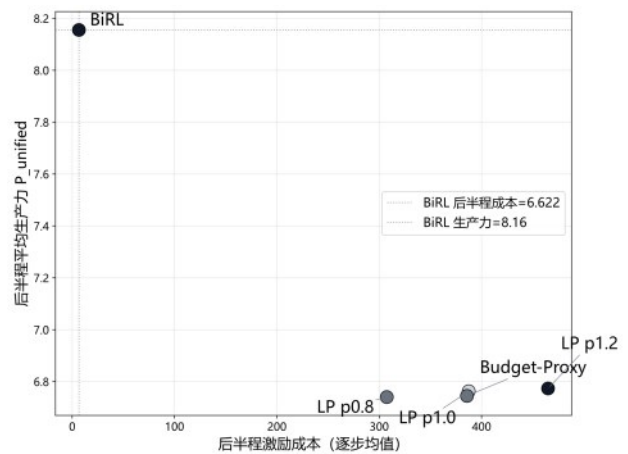


图11 成本效率散点图

综合以上指标, 本文 BiRL 框架是唯一在平台奖励为正、生产力接近最优、激励成本极低三个维度上同时表现优异的方案。Linear-Price 基线的对比证明, 性能提升并非源于更高的预算投入; Budget-Proxy 的对比证明, 以简单质量分代理指标替代 Shapley 值的精确贡献评估无法复现 BiRL 的成本效率。上述结果共同验证了所提双层强化学习激励框架在设计平台激励政策方面的有效性与可行性。

5 结束语

本文针对数据驱动平台中用户异质性、隐私成本动态累积及激励机制难以适应动态环境等问题, 提出了一种面向可信数据共享的双层强化学习动态激励机制。构建融合隐私效用、激励收益与近似 Shapley 值数据贡献评估的动态 Stackelberg 博弈模型; 设计双层强化学习策略优化框架, 通过交替冻结训练机制实现平台激励策略与用户响应策略的解耦优化, 在异质性用户间形成稳定的差异化共享均衡; 利用区块链智能合约实现激励执行的透明与可验证。实验结果表明, 相比于静态线性定价与预算代理分配基线, 本文所提 BiRL 框架在将训练全程总激励成本压缩至对比基线均值约 1.9% 的条件下, 生产力指标仍较各基线提升约 21%, 验证了其在激励效率与平台收益两个维度上的有效性与适应性。未来将进一步研究大规模并行训练及真实平台部署

与验证, 提升训练效率与可扩展性。



段美姣 (1985-), 女, 河北石家庄人, 博士, 中央财经大学副教授、硕士生导师, 主要研究方向为数据安全与隐私保护、区块链和人工智能安全。



黄日东 (2001-), 男, 江西抚州人, 中央财经大学研究生, 主要研究方向为强化学习和区块链。



王熙 (1987-), 女, 黑龙江哈尔滨人, 博士, 中央财经大学教授、博士生导师, 主要研究方向为数据挖掘与商务智能、智能信息系统及决策优化。



李惠 (1986-), 女, 湖北孝感, 博士, 中央财经大学信息学院副教授, 主要研究方向为决策与优化、物流与供应链管理。

参考文献:

- [1] 李凤华, 李晖, 牛犇, 等. 数据要素流通与安全的研究范畴与未来发展趋势[J]. 通信学报, 2024, 45(5): 1-11.
Li F H, Li H, Niu B, et al. Research category and future development trend of data elements circulation and security[J]. Journal on Communications, 2024, 45(05): 1-11.
- [2] 刘金钊, 汪寿阳. 数据要素市场化配置的困境与对策探究[J]. 中国科学院院刊, 2022, 37(10): 1435-1443.
Liu J Z, Wang S Y. Dilemmas and Suggestions on Market-based Data Allocation[J]. Bulletin of Chinese Academy of Sciences, 2022, 37(10): 1435-1443.
- [3] Zhan Y, Zhang J, Hong Z, et al. A Survey of Incentive Mechanism Design for Federated Learning[J]. IEEE Transactions on Emerging Topics in Computing, 2022, 10(2): 1035-1044.
- [4] Han B, Li B, Wolter K, et al. Dynamic Incentive Design for Federated Learning Based on Consortium Blockchain Using a Stackelberg Game [J]. IEEE Access, 2024, 12: 160267-160283.
- [5] Yuan W, Wang X. QI-DPFL: Quality-Aware and Incentive-Boosted Federated Learning with Differential Privacy[C]//Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2024: 1-8.
- [6] He L. The optimization of incentive mechanisms for edge federated learning based on Stackelberg game[J]. Computing, 2026, 108(6): 87.
- [7] Alagha A, Otrók H, Singh S, et al. Blockchain-based crowdsourced deep reinforcement learning as a service[J]. Information Sciences, 2024, 679: 121107.
- [8] Baghcheband H, Soares C, Reis L P. Shapley value-based data valuation for machine learning data markets[J]. Discover Applied Sciences, 2025, 7(12): 1431.
- [9] Xia H, Zhang J, Sun Q, et al. Computing Shapley Values for Dynamic Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(6): 3253-3271.
- [10] Zheng S, Trott A, Srinivasa S, et al. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning[J]. Science Advances, 2022, 8(18): eabk2607.
- [11] Xuan S, Zheng L, Chung I, et al. An incentive mechanism for data sharing based on blockchain with smart contracts[J]. Computers & Electrical Engineering, 2020, 83: 106587.
- [12] Li C X, Qu X D, Guo Y. TFCrowd: a blockchain-based crowdsourcing framework with enhanced trustworthiness and fairness[J]. EURASIP Journal on Wireless Communications and Networking, 2021(1): 1-20.
- [13] Schwabe D, Becker K, Seyferth M, et al. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review[J]. npj Digital Medicine, 7(1): 1-30.
- [14] Pei J. A Survey on Data Pricing: From Economics to Data Science[J]. IEEE Transactions on Automatic Control, 2022, 34(10): 4586-4608.
- [15] Yu H, Zhang M. Data pricing strategy based on data quality[J/OL]. Computers & Industrial Engineering, 2017, 112: 1-10.
- [16] Li J, Wang D, Qi G, et al. Alliance chain-based simulation on a new clinical research data pricing model[J]. Annals of Translational Medicine, 2022, 10(15): 836.
- [17] Beechey D, Simek Ö. Approximating Shapley Explanations in Reinforcement Learning[C]//Proceedings of the 39th Conference on Neural Information Processing Systems, 2025: 25320-25354.
- [18] Ghosh A, Roth A. Selling privacy at auction[C]//Proceedings of the 12th ACM conference on Electronic commerce. New York, NY, USA: Association for Computing Machinery, 2011: 199-208.
- [19] Hu R, Gong Y. Trading Data For Learning: Incentive Mechanism For On-Device Federated Learning[C]//GLOBECOM 2020, IEEE Global Communications Conference, 2020: 1-6.
- [20] Kang J, Pedarsani R, Ramchandran K. The fair value of data under heterogeneous privacy constraints[EB/OL]. arXiv preprint arXiv: 2301.13336, 2023.
- [21] Wang H N, Liu N, Zhang Y Y, et al. Deep Reinforcement Learning: A Survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1726-1744.
- [22] Beck Y, Ljubić I, Schmidt M. A survey on bilevel optimization under uncertainty[J]. European Journal of Operational Research, 2023, 311(2): 401-426.
- [23] Zeng R F, Zeng C, Wang X W, et al. A Comprehensive Survey of Incentive Mechanism for Federated Learning[EB/OL]. arXiv preprint arXiv: 2106.15406, 2021.
- [24] Thoma V, Pasztor B, Krause A, et al. Contextual Bilevel Reinforcement Learning for Incentive Alignment[C]//Proceedings of the 38th Conference on Neural Information Processing Systems, 2024: 127369 - 127435.
- [25] Paolo G, Benechehab A, Cherkaoui H, et al. TAG: A Decentralized Framework for Multi-Agent Hierarchical Reinforcement Learning[EB/OL]. arXiv preprint arXiv: 2502.15425, 2025.
- [26] Liu Y, Guo S, Zhan Y F, et al. Chiron: A Robustness-Aware Incentive Scheme for Edge Learning via Hierarchical Reinforcement Learning [J]. IEEE Transactions on Mobile Computing, 2024, 23(8): 8508-8524.
- [27] Jiang K, Wang Y, Wang H, et al. A Reinforcement Learning-Based Incentive Mechanism for Task Allocation Under Spatiotemporal Crowdsensing[J]. IEEE Transactions on Computational Social Systems, 2024, 11(2): 2179-2189.
- [28] Luo G, Zhan S, Liang C, et al. An incentive mechanism for joint sensing and communication Vehicular Crowdsensing by Deep Reinforcement Learning[J]. Computer Networks, 2025, 260: 111099.
- [29] Wang D, Pang X, Hu J, et al. Two-dimensional stackelberg game-based incentive mechanism for differential private federated learning with non-iid data[J]. IEEE Transactions on Mobile Computing, 2026, 25(8): 12688-12703.
- [30] Yang M, Feng H, Wang X, et al. Fair Incentive Allocation for Secure Data Sharing in Blockchain Systems: A Shapley Value Based Mechanism[C]//Proceedings of 3rd 2023 International Conference on Autonomous Unmanned Systems. Lecture Notes in Electrical Engineering, 1171. Singapore: Springer, 2024: 460-469.
- [31] 鲍世豪, 倪郑威. SWTA-Shapley: 一种高效的联邦学习贡献评估方法[J]. 电信科学, 2025, 41(12): 146-163.
Bao S H, Ni Z W. SWTA-Shapley: an efficient contribution evaluation method for federated learning[J]. Telecommunications Science, 2025, 41(12): 146-163.
- [32] Li M, Harish A R, Yu C, et al. Blockchain-Based Medical Data Asset Sharing Framework for Healthcare 4.0[J]. IEEE Transactions on Industrial Informatics, 2025, 21(4): 2779-2788.