



基于多智能体合作博弈的液压支架调直策略研究

杨艺 王静怡 钱伟 王田

Study on Hydraulic Support Alignment Strategy Based on Multi-agent Cooperative Game

YANG Yi, WANG Jing-Yi, QIAN Wei, WANG Tian

在线阅读 View online: <https://doi.org/10.16383/j.aas.c250624>

您可能感兴趣的其他文章

基于多智能体强化学习的博弈综述

Multi-agent Reinforcement Learning Based Game: A Survey

自动化学报. 2025, 51(3): 540–558 <https://doi.org/10.16383/j.aas.c240478>

基于多智能体强化学习的乳腺癌致病基因预测

Prediction of Breast Cancer Pathogenic Genes Based on Multi-agent Reinforcement Learning

自动化学报. 2022, 48(5): 1246–1258 <https://doi.org/10.16383/j.aas.c210583>

多智能体强化学习控制与决策研究综述

Survey on Multi-agent Reinforcement Learning for Control and Decision-making

自动化学报. 2025, 51(3): 510–539 <https://doi.org/10.16383/j.aas.c240392>

多智能体博弈、学习与控制

An Interdisciplinary Survey of Multi-agent Games, Learning, and Control

自动化学报. 2023, 49(3): 580–613 <https://doi.org/10.16383/j.aas.c220680>

基于多智能体强化学习的流程工业多操作参数协同优化

Collaborative Optimization of Multiple Operating Parameters for Process Industries Based on Multi-Agent Reinforcement Learning

自动化学报. 2026, 52(1): 78–90 <https://doi.org/10.16383/j.aas.c250308>

多智能体协同研究进展综述: 博弈和控制交叉视角

Recent Advances on Multi-agent Collaboration: A Cross-perspective of Game and Control Theory

自动化学报. 2025, 51(3): 489–509 <https://doi.org/10.16383/j.aas.c240508>

基于多智能体合作博弈的液压支架调直策略研究

杨 艺^{1,2,3} 王静怡^{1,2} 钱 伟^{1,2} 王 田⁴

摘 要 综采工作面液压支架群调直过程中, 支架群间的强耦合关系与液压缸摩擦-滑移非线性构成典型的“耦合-非线性”双重复杂性, 使传统控制方法难以有效建模. 现有多智能体强化学习方法能实现并行决策, 但面临全局奖励无法精确归因至各支架动作、仅依赖观测状态难以捕获支架群姿态的时序演化规律等问题, 阻碍策略的有效收敛. 为此, 提出融合合作博弈与长短期记忆网络 (LSTM) 的多智能体强化学习算法 CG-LSTM-MATD3. 该算法基于去中心化部分可观测马尔科夫决策过程对液压支架调直过程建模, 引入 Shapley 值实现各液压支架边际贡献的合理归因, 并设计 Coalition 网络通过联盟生成降低计算复杂度. 其次, 在 Actor、Critic 以及 Coalition 网络中嵌入 LSTM 模块, 通过对状态序列等信息的记忆, 使模型能够捕获状态的时序依赖关系, 增强对环境动态特性和真实状态的感知能力. 实验结果表明, 该算法在七支架任务中直线度相对基线算法提升 80.59%, 消融实验进一步证明了合作博弈机制与 LSTM 模块的有效性.

关键词 液压支架; 多智能体强化学习; 合作博弈; MATD3 算法; 长短期记忆网络; Shapley 值

引用格式 杨艺, 王静怡, 钱伟, 王田. 基于多智能体合作博弈的液压支架调直策略研究. 自动化学报, 2026, 52(7): 1463–1476

DOI 10.16383/j.aas.c250624 **CSTR** 32138.14.j.aas.c250624

Study on Hydraulic Support Alignment Strategy Based on Multi-agent Cooperative Game

YANG Yi^{1,2,3} WANG Jing-Yi^{1,2} QIAN Wei^{1,2} WANG Tian⁴

Abstract During the alignment of hydraulic support clusters in fully mechanized mining faces, the strong coupling between support clusters and the friction-slip nonlinearity of hydraulic cylinders create a typical “coupling-nonlinearity” dual complexity, making it difficult for traditional control methods to model the system effectively. Existing multi-agent reinforcement learning methods can achieve parallel decision-making, while they face challenges such as the inability to precisely attribute global rewards to each support action, and the difficulty of capturing the temporal evolution of the support cluster’s posture based solely on observed states, which hinder effective policy convergence. To address these problems, this paper proposes the multi-agent reinforcement learning algorithm CG-LSTM-MATD3, which integrates cooperative games with long short-term memory networks (LSTM). This algorithm models the hydraulic support alignment process based on a decentralized partially observable Markov decision process, introduces Shapley values to reasonably attribute the marginal contributions of each hydraulic support, and designs a Coalition network to reduce computational complexity through coalition formation. Furthermore, LSTM modules are embedded in the Actor, Critic, and Coalition networks. By storing information such as state sequences, the model can capture temporal dependencies among states, thereby enhancing its ability to perceive the dynamic characteristics of the environment and the true state. Experimental results show that the algorithm achieves an 80.59% improvement in linearity compared to the baseline algorithm in the seven-support task. Ablation experiments further validate the effectiveness of the cooperative game mechanism and the LSTM modules.

Keywords hydraulic support; multi-agent reinforcement learning; cooperative game; MATD3 algorithm; long short-term memory; Shapley value

Citation Yang Yi, Wang Jing-Yi, Qian Wei, Wang Tian. Study on hydraulic support alignment strategy based on multi-agent cooperative game. *Acta Automatica Sinica*, 2026, 52(7): 1463–1476

收稿日期 2025-11-11 录用日期 2026-05-13

Manuscript received November 11, 2025; accepted May 13, 2026

国家自然科学基金 (92467108) 资助

Supported by National Natural Science Foundation of China (92467108)

本文责任编辑 徐昕

Recommended by Associate Editor XU Xin

1. 河南理工大学电气工程与自动化学院 焦作 454003 2. 河南省煤矿装备智能检测与控制重点实验室 焦作 454003 3. 郑州恒达智控科技股份有限公司 郑州 450000 4. 北京航空航天大学人工智能学院 北京 100191

综采工作面回采过程中, 煤壁、刮板输送机及液压支架的直线度直接决定工作面的安全性与开采效率^[1]. 液压支架作为顶板支撑的核心设备, 同时也是实现工作面直线度调控的主要执行机构. 受煤层

1. School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454003 2. Henan Key Laboratory of Intelligent Inspection and Control of Coal Mine Equipment, Jiaozuo 454003 3. Zhengzhou Hengda Intelligent Control Technology Co., Ltd., Zhengzhou 450000 4. School of Artificial Intelligence, Beihang University, Beijing 100191

起伏、断层、夹矸等复杂地质条件影响, 液压支架推溜与拉架距离难以达到理想状态, 易造成工作面直线度偏差持续累积, 当直线度偏差超出安全截割与高效回采允许范围时, 需要依据液压支架实时状态制定合理调直策略, 以保障综采工作面的安全、稳定、高效运行。

综采工作面通常包含上百台液压支架相互协同工作, 各支架需依据采煤机位置、顶底板状态、自身与邻架姿态等状态变量, 执行升柱、降柱、推移等动作。因此, 在进行工作面直线度调整时, 各液压支架的执行动作之间存在强耦合关系。传统的人工操控或基于简单逻辑判断的控制方式难以应对多支架间复杂的力学耦合与协同动作需求, 且煤层起伏、断层、夹矸等地质构造会造成局部支架承载突变, 进一步导致不同位置支架受力呈现非线性分布, 使支架移动轨迹难以精确预测。支架间存在动作耦合, 使得智能体状态演化具有较强不确定性。微小的操作误差或初始位姿偏差都可能会经非线性耦合关系放大, 最终造成调直效果偏离预期。这种“耦合-非线性”双重复杂性, 使得基于经典线性控制理论和经验式操作的方法难以实现高精度的直线度控制。

为解决上述问题, 国内外学者已开展多方面研究并取得阶段性进展, 主要形成了传统控制方法与智能控制方法两大路径。在智能控制方法不断演进的过程中, 强化学习^[2]凭借智能体与环境交互进行试错学习, 降低了对精确数学模型的依赖。基于此, 文献 [3-6] 将强化学习引入井下综采工作面领域, 并取得一系列前期成果。最新的研究综述指出, 离策强化学习具备高探索能力和样本效率, 在解决多智能体优化控制问题中备受关注^[7]。基于这一技术路径, 文献 [8] 探索了离策算法在液压支架调直中的应用, 提出基于分段经验池与多机制采样的 TD3 改进算法, 并实现了复杂地质条件下调直策略的有效收敛。该研究侧重于验证强化学习在液压支架调直中的可行性, 建模部分与协同决策机制仍需进一步完善。

在多智能体强化学习中, 多个智能体协同完成共同目标, 合理的信用分配能够避免某些智能体沦为“懒惰”智能体。合作博弈论通过收益分配机制, 为多智能体系统中的信用分配问题提供了坚实的数学基础。其中, Shapley 值作为同时满足效率性、对称性、虚拟性和可加性四大公理的唯一解概念, 能够公平地量化各智能体对团队收益的贡献。近年来, 相关研究在理论深化与应用拓展方面均取得了较大进展, 但在面对多智能体任务时, 保证任务完成度的同时减小计算开销以及提升算法在动态非平稳环境下的鲁棒性, 仍需深入探讨。

针对液压支架群调直过程中多支架强耦合交互建模困难、信用分配不合理、部分可观测井下环境下策略收敛效率低等问题, 本文提出一种基于 MATD3 算法^[9]的液压支架群调直决策框架, 将该过程建模为去中心化部分可观测马尔科夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP)^[10]。针对信用分配问题, 本文引入合作博弈论^[11]中的 Shapley 值方法^[12]实现全局奖励的合理归因, 创新性设计 Coalition 网络, 通过动态联盟生成近似计算各支架边际贡献, 缓解计算 Shapley 值带来的指数级复杂度问题。本文在 Actor、Critic 和 Coalition 网络中融入长短期记忆网络 (long short-term memory, LSTM)^[13], 利用时序记忆特性处理部分可观测环境下的状态历史信息, 提升策略收敛速度。综上所述, 本文的主要贡献与创新点如下:

1) 提出 Coalition 网络以降低计算复杂度。设计 Coalition 网络实现动态联盟生成, 将计算 Shapley 值时的指数级复杂度降低至线性复杂度, 在保证信用分配准确性的同时降低多支架调直的计算负担。

2) 构建时序感知的网络架构。在各网络模块中嵌入 LSTM, 通过历史状态信息增强智能体对环境非平稳性的感知能力, 提升算法的收敛速度。

3) 实现基于合作博弈、长短期记忆的双延迟深度确定性策略梯度算法 (CG-LSTM-MATD3) 的算法仿真验证。结合综采工作面实际, 基于 GYM 框架开发液压支架调直仿真平台, 通过仿真测试验证算法在模拟井下动态环境中的有效性与鲁棒性。

1 相关工作

针对综采工作面液压支架调直, 国内外学者围绕传统控制与智能控制两大路径开展了大量的研究, 形成了众多研究成果, 同时也存在相应的局限性。

1.1 基于传统方法的液压支架调直控制

传统控制方法基于经典控制理论和几何解析原理, 通过建立精确数学模型实现液压支架位姿调控, 目前在支架位姿测量、误差建模与控制优化方面积累了丰富的研究成果。在测量技术方面, 2019 年, 文献 [14] 通过架间行走机器人与多维传感器建立液压支架位置偏移与直线度的数学模型, 实现了四维偏移量的测量与误差分析; 同年, 文献 [15] 引入红外 LED 标识板与防爆相机的视觉测量算法, 进一步提升了支架位姿识别的精度与抗干扰能力。在误差建模与控制优化方面, 2020 年, 文献 [16] 基于 D-H 坐标系和多传感器融合技术提高了直线度推

导精度; 2021 年, 文献 [17] 通过建立推移公式与误差模型提出了可降低直线度误差的优化调直控制方法; 2022 年, 文献 [18] 引入最大熵卡尔曼滤波算法来抑制非高斯噪声影响, 进一步提升了控制精度。

然而, 传统方法对精确的数学模型有较强的依赖性, 综采工作面环境复杂多变, 多支架间的强耦合关系难以通过解析方法精准建模, 模型偏差易引发全局误差累积并通过耦合机制放大, 最终导致控制策略失效, 无法适配井下动态非平稳工况的实际需求, 因此推动了智能控制方法的发展。

1.2 液压支架调直智能控制方法

智能控制方法采用数据驱动和自学习机制, 无需建立精确数学模型, 可以从历史数据中自主学习运行规律, 对环境变化的适应性更强, 已成为近年来的研究热点。2014 年, 文献 [19] 基于支持向量机与遗传算法构建液压支架调直系统模型, 实现了在不同井下环境下的精准预测与自动调直。2021 年, 文献 [20] 针对液压支架跟机自动化过程中的丢架、推移不到位等问题, 提出基于遗传算法与反向传播神经网络组合模型的控制方法, 有效提高了推移精度。2022 年, 文献 [21] 通过引入先进的智能感知与自动控制技术, 攻克了关键设备的精确定位与调直瓶颈, 实现了综采工作面的智能化运行。2024 年, 文献 [22] 通过神经网络与控制理论相结合, 实现了对系统扰动和未知状态的智能感知与补偿, 进一步提升了控制精度。同年, 文献 [23] 将液压支架建模为具备自主感知与决策能力的智能体, 基于分布式协同控制与自适应通信拓扑实现了自主调直, 为多支架群协同调直提供了新思路, 但在解决支架间的信用分配等问题上有所欠缺。

1.3 多智能体合作博弈的信用分配方法

在多智能体强化学习中, 信用分配是实现多智能体协同的关键, 合作博弈论中的 Shapley 值方法为该问题提供了数学基础。近年来相关研究在理论与应用上均取得了显著进展。

在理论层面, 2019 年, 文献 [24] 通过引入扩展凸博弈框架并利用 Shapley Q-value 为智能体实现公平的信用分配, 有效解决了多智能体强化学习中全局奖励分配不均的问题。2021 年, 文献 [25] 将 Shapley 值推广至马尔科夫决策过程, 提出马尔科夫 Shapley 值 (MSV) 及 Shapley-Bellman 最优方程, 建立了可解释的价值分解理论, 完善了 Shapley 值在动态决策场景中的理论基础。同年, 文献 [26] 进一步针对集中式训练和分布式执行 (centralized training and distributed execution, CTDE)

范式提出基于 Shapley 值的显式信用分配方法, 利用蒙特卡洛采样有效降低了计算复杂度, 解决了传统 Shapley 值计算繁琐的问题。2022 年, 文献 [27] 严格证明了 MSV 满足效率、对称性等公平信用分配公理, 并将该理论拓展至部分可观测环境。2024 年, 文献 [28] 验证了 Shapley 值评估多智能体贡献的有效性, 进一步完善了相关理论体系。

在应用层面, 2024 年, 文献 [29] 面向快速公交系统提出八相位公交信号优先方法, 将合作博弈论与 Q 学习相结合, 在保障公交服务可靠性的同时优化了私家车通行稳定性。2025 年, 文献 [30] 针对无人机集群协同任务, 构建了基于公共品博弈的博弈论-强化学习 (GT-RL) 方法, 有效促进了动态复杂环境下的合作行为。此外, 文献 [31] 引入扩散概率模型来构建生成式策略, 通过增强对多峰分布的探索来弥补复杂环境下协同决策的滞后性, 从而在个体层面提升决策效率。然而, 如何在保证信用分配精度的前提下, 有效降低算法计算开销并增强在复杂场景下的实时决策能力, 仍是实现该类方法工程化应用的关键。

2 液压支架作业与调直过程建模

2.1 液压支架作业

综采工作面液压支架的协同调直控制是保障煤矿安全高效开采的关键技术环节。综采工作面设备的空间布局与协同作业场景如图 1(a) 所示, 采煤机沿刮板输送机往复行走, 通过滚筒截割煤层, 破碎后的煤炭经装载机构卸入输送机槽内。刮板输送机铺设于工作面全长, 链条驱动的刮板输送机将煤流连续输送至转载点。液压支架以一定间距成排布置于刮板输送机采空区侧后方, 通过立柱和顶梁构成支护结构, 承载上覆岩层压力, 保障采空区顶板稳定。液压支架的对齐精度直接影响顶板控制效果与设备移动安全性, 因此在采煤全周期内需要确保液压支架群保持良好的直线度。三类设备相互配合, 共同完成“割煤-运煤-移架-支护”的综采工艺流程。液压支架拉架调直过程如图 1(b) 所示, 由于地质条件复杂、操作误差积累等原因, 支架群往往偏离理想位置, 呈现出错列状态。调直过程中, 各支架根据自身位置偏差和与相邻支架的位置关系, 通过控制推移千斤顶的推移行程, 实现支架与刮板输送机的相对位移调节, 完成支架群的错位校正。基于强化学习的调直策略使每台支架能够根据局部观测信息自主决策推移量, 并通过多轮迭代逐步调整至理想直线排列状态。调直过程需兼顾调直效率与支护连续性, 单次调整量过大可能诱发支架失稳或导致顶

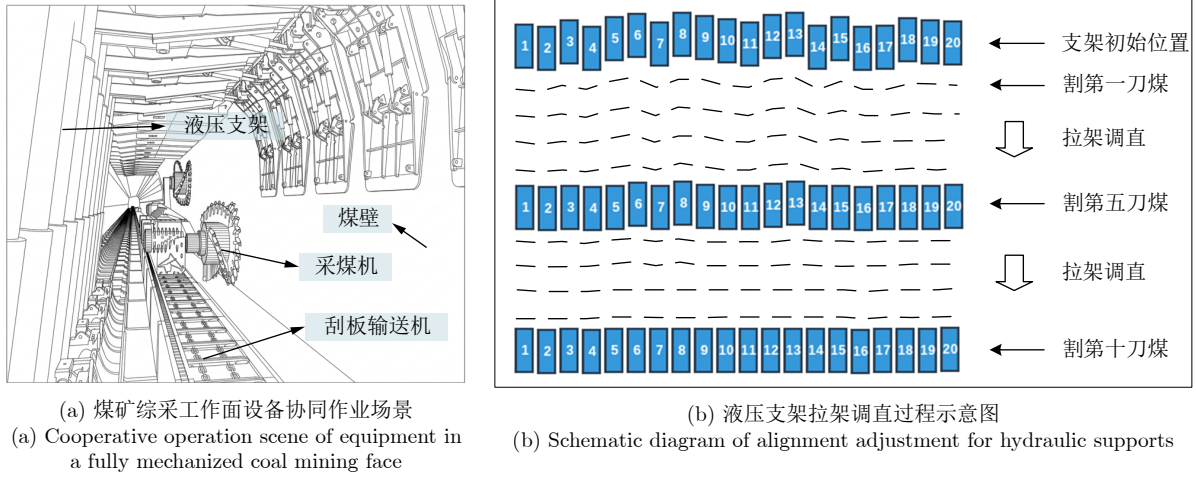


图 1 煤矿综采工作面液压支架协同作业

Fig. 1 Cooperative operation of hydraulic supports in fully mechanized coal mining face

板支护空白, 而调整量过小则会延长调直周期, 从而降低采煤效率.

2.2 调直过程建模

液压支架群协同调直的决策问题中决策变量为各支架的拉架距离, 该指令作为控制输入, 可直接驱动液压支架执行机构完成拉架调直动作. 综采工作面液压支架拉架调直过程具有典型的分布式决策特征, 各支架只能获取局部观测信息, 而整体调直效果取决于全局协同. 由于支架的位置变化满足马尔科夫性质, 即下一时刻位置仅由当前位置和拉架动作决定, 与历史轨迹无关, 因此该过程可被建模为 Dec-POMDP, 该模型由五元组 $B = \langle N, S, O, A, r \rangle$ 构成.

1) 支架集合 N . $N = \{1, 2, \dots, n\}$ 表示协同调直的 n 台液压支架集合, 支架编号按工作面布置顺序排列.

2) 全局状态空间 S . 全局状态 $s^t \in S$ 刻画 t 时刻所有支架的状态位置, s^t 定义为:

$$\begin{cases} s^t = [d_1^t, d_2^t, \dots, d_n^t]^T \\ d_i^t = p_{\max}^t - p_i^t, \quad i \in \{1, 2, \dots, n\} \\ p_{\max}^t = \max\{p_1^t, p_2^t, \dots, p_n^t\} \end{cases} \quad (1)$$

其中, i 代表第 i 台液压支架; d_i^t 表示第 i 台液压支架相对当前最远支架的滞后距离; p_i^t 是第 i 台液压支架的实际位置; $p_{\max}^t$ 是所有液压支架实际位置中的最大位置.

3) 观测空间 O . t 时刻液压支架的局部观测状态为 $o^t \in O$, 可表示为:

$$o^t = [d_1^t, d_2^t, \dots, d_m^t]^T \quad (2)$$

其中, $m < n$, n 为液压支架总数. 第 i 台液压支架能观测到自身和其他相连支架的状态信息.

4) 连续动作空间 A . 动作 $a^t \in A$ 表示为:

$$\begin{cases} a^t = [a_1^t, a_2^t, \dots, a_n^t]^T \\ a_i^t \in [a_{\min}, a_{\max}] \end{cases} \quad (3)$$

其中, a_i^t 表示第 i 台液压支架的拉架量; 考虑到工作面地质环境和机械设备的物理限制, $a_{\min}$ 和 $a_{\max}$ 分别限定了拉架动作的最小与最大距离.

5) 奖励函数 r . 奖励函数的设计需要同时兼顾调直效果与生产效率. 为实现这种平衡, 奖励函数采用加权复合形式进行设计:

$$r^t = \alpha r_{\text{linearity}}^t + (1 - \alpha) r_{\text{move}}^t \quad (4)$$

其中, $\alpha \in [0, 1]$ 为权衡系数; 直线度奖励 $r_{\text{linearity}}^t$ 采用指数衰减函数对支架给予奖励:

$$r_{\text{linearity}}^t = e^{-\frac{1}{n} \sum_{i=1}^n d_i^{t+1}} \quad (5)$$

当支架群趋近于共线状态时, 直线度奖励最大, 表示液压支架排列状态接近理想水平; 若存在显著支架错位, 奖励值将迅速衰减. 移动奖励 r_{move}^t 基于归一化平均拉架量定义:

$$r_{\text{move}}^t = \frac{1}{n} \sum_{i=1}^n \frac{a_i^t - a_{\min}}{a_{\max} - a_{\min}} \quad (6)$$

该奖励项激励支架在确保良好对齐的前提下最大化推进速度, 避免影响生产效率.

3 基于合作博弈与 LSTM 的多智能体强化学习算法

本节以 MATD3 算法为基础, 使用合作博弈量

网络参数, 这种平滑更新方式可缓解训练过程中的策略波动问题.

3.2 合作博弈与 MATD3 算法的融合

针对各液压支架对工作面直线度贡献存在差异的特点, 本节通过 Coalition 网络构建与 Shapley 值分配方法建立合理的贡献归因机制.

3.2.1 合作博弈 Shapley 值

1953 年, Shapley 基于合作博弈理论提出一种公平的信用分配方法, 本文将该方法引入 MATD3 算法, 用于量化单架液压支架在调直任务中的边际贡献. 第 i 台液压支架的 Shapley 值定义为:

$$\varphi_i(v) = \sum_{C \subset N} \frac{(k-1)!(n-k)!}{n!} (v(C) - v(C \setminus \{i\})) \quad (13)$$

其中, N 为液压支架集合; 联盟 C 为包含第 i 台液压支架的任意子集; $C \setminus \{i\}$ 表示不包含第 i 台液压支架的联盟; $v(\cdot)$ 为特征函数, 用于描述联盟的整体效能; $k = |C|$ 为联盟 C 内液压支架的数量. Shapley 值 $\varphi_i(v)$ 用于衡量支架 i 加入不同联盟时带来的边际贡献.

3.2.2 Coalition 网络的设计

传统 Shapley 值计算需遍历 2^N 种组合, 计算复杂度随支架数量呈指数级增长, 计算成本较高, 难以适配多支架调直场景.

本文设计一种基于神经网络的 Coalition 网络, 将复杂的联盟组合搜索转换为神经网络前向推理, 使计算复杂度从 $O(2^N)$ 降至 $O(N)$. 在时刻 t , 联盟向量 $\mathbf{c}^t$ 定义为:

$$\mathbf{c}^t = K_\psi(\mathbf{s}^t) \quad (14)$$

其中, K_ψ 表示参数为 ψ 的 Coalition 网络, 输入为全局状态 $\mathbf{s}^t$, 输出为二值化联盟成员标识; 联盟向量为 $\mathbf{c}^t = [c_1^t, c_2^t, \dots, c_n^t]^T$, $c_i^t = 1$ 表示第 i 台液压支架加入当前联盟, $c_i^t = 0$ 表示未加入. 由于同一联盟内的液压支架存在强耦合关系, 因此液压支架间拉架动作需要保持相互协同才能修正局部偏差.

Coalition 网络各层均包含 128 个神经元, 激活函数采用 ReLU 以构建非线性映射关系, 优化器选用 Adam. 为保证训练稳定, Coalition 网络与 Actor、Critic 网络采用一致的软更新机制, 通过平滑更新目标网络参数, 避免联盟划分结果出现剧烈波动.

Coalition 网络的损失函数定义为:

$$L(\psi) = E_{(\mathbf{s}^t, \mathbf{a}^t, \mathbf{s}^{t+1}, r^t) \sim D} [R_{in} - R_{out}] \quad (15)$$

其中, R_{in} 和 R_{out} 分别表示联盟成员与非成员在调

直任务中获得的奖励. 这一优化目标通过最大化联盟内外的奖励差值, 引导网络精准学习各支架状态间的关系, 最终生成包含关键支架的最优联盟, 从而保障液压支架拉架动作的一致性, 提升工作面整体直线度.

本节使用的 Shapley 值分配方法如图 3 所示, 该方法通过对比第 i 台液压支架参与联盟与不参与联盟两种场景下的 Q 值差异, 来近似计算液压支架的边际贡献. 在 t 时刻, 两种联盟动作向量定义如下.

联盟动作向量为:

$$\mathbf{a}_C^t = \mathbf{a}^t \odot \mathbf{c}^t \quad (16)$$

其中, $\odot$ 表示 Hadamard 积; 动作向量 $\mathbf{a}^t = [a_1^t, a_2^t, \dots, a_n^t]^T$ 表示液压支架群的联合动作向量; 联盟向量为 $\mathbf{c}^t = [c_1^t, c_2^t, \dots, c_n^t]^T$, $c_i^t = 1$ 表示第 i 台液压支架加入当前联盟 C ; $\mathbf{a}_C^t$ 表示联盟 C 内液压支架的联合动作向量, 保留了联盟成员 i 由 Actor 网络输出的拉架动作.

反事实联盟动作向量为:

$$\begin{cases} \mathbf{a}_{C \setminus \{i\}}^t = \mathbf{a}^t \odot \mathbf{c}^t \\ \mathbf{a}_i^t = 0 \end{cases} \quad (17)$$

其中, $\mathbf{a}_{C \setminus \{i\}}^t$ 表示联盟 C 中成员的联合动作向量, 第 i 台液压支架的动作置零, 模拟支架未参与联盟决策的假设场景.

Critic 网络通过计算状态向量分别与联盟动作向量 $\mathbf{a}_C^t$ 以及反事实联盟动作向量 $\mathbf{a}_{C \setminus \{i\}}^t$ 拼接后生成的 Q 值差异, 实现对第 i 台液压支架边际贡献的计算:

$$\begin{aligned} \varphi_i(Q_{\theta_{i,j}}(\mathbf{s}^t, \mathbf{a}^t)) &= Q_{\theta_{i,j}}(\mathbf{s}^t, \mathbf{a}_C^t) - \\ &Q_{\theta_{i,j}}(\mathbf{s}^t, \mathbf{a}_{C \setminus \{i\}}^t) \end{aligned} \quad (18)$$

该值表示第 i 台液压支架对工作面直线度的贡献程度. 若数值为正, 说明支架的动作对改善工作面直线度起到关键作用; 若数值为负, 说明动作对

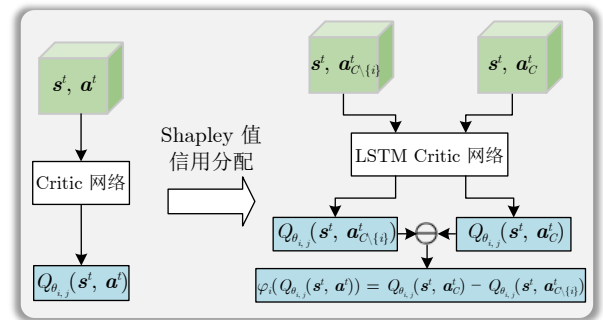


图 3 Shapley 值信用分配

Fig.3 Shapley value credit assignment

液压支架群的整体协同调直产生不利影响。

3.3 基于 LSTM 与经验回放的网络改进

液压支架调直过程中状态具有时序依赖性, 且传统经验回放易破坏时序信息. 本文通过将 LSTM 模块与经验回放机制相结合, 实现样本间独立性与样本内时序连贯性的平衡。

经验回放通过在经验池中随机抽取训练样本, 打破连续样本间的强相关性, 减少训练过程中的数据冗余, 提升状态空间探索效率. 但随机采样会忽略状态与动作的时间演化关系, 难以适应液压支架调直偏差具有累积效应的时序特性. LSTM 具备优异的长时序特征建模能力, 可对连续时间步内的状态变化、动作响应及偏差累积过程进行建模, 同时保留多步历史状态信息, 将历史决策特征融入当前状态信息, 从而挖掘时序依赖关系, 提升智能体在复杂动态环境下的决策准确性。

二者融合后, 在时序样本片段内部通过 LSTM 保证时序连贯性, 在不同样本片段之间通过经验回放保证独立性与随机性, 既缓解了样本相关问题, 又充分利用了时序信息. 本文将 LSTM 嵌入强化学习核心网络结构, 如图 4 所示, 其中 l 代表窗口大小。

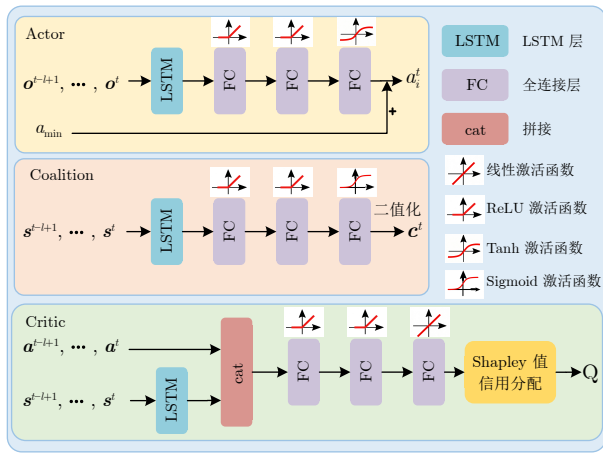


图 4 LSTM 融入各网络示意图

Fig. 4 Schematic diagram of LSTM integration into each network

在环境交互阶段, 智能体维护一个固定大小为 l 的滑动时间窗口以捕获时序信息. 针对训练初期历史数据不足的问题, 采用零填充策略补齐缺失数据, 保证模型输入维度统一. 当窗口数据填满后, 执行滑动更新, 剔除窗口中最早的历史观测值, 存入最新获取的环境观测值。

在决策阶段, Actor 网络以滑动窗口内的观测序列为输入, 通过 LSTM 层提取当前时刻状态信息

与历史时序特征, 特征经全连接层映射后, 输出适配当前环境的连续动作。

数据收集阶段, 智能体将完整的滑动时间窗口数据作为单条时序经验存储至经验池, 每条经验包含连续 l 步的观测序列、动作序列、下一时刻观测序列及下一时刻状态序列. 模型训练阶段, 网络从经验池中随机采样时序片段, 完成参数的梯度更新。

Actor、Critic 以及 Coalition 网络的运算过程如下: Actor 网络接收长度为 l 的观测序列 $o^{t-l+1}, \dots, o^t$, 利用 LSTM 挖掘环境动态变化的时序特征, 最终通过全连接层输出连续动作; Critic 网络以相同长度 l 的全局状态序列 $s^{t-l+1}, \dots, s^t$ 为输入, 经 LSTM 提取时序特征后, 与智能体联合动作序列 $a^{t-l+1}, \dots, a^t$ 进行拼接, 再传入全连接层采用 ReLU 激活函数增强非线性特征表达能力, 输出层采用线性激活函数得到价值评估结果; 同时, 引入 Shapley 值信用分配模块, 量化各智能体当前动作对整体收益的价值贡献; Coalition 网络接收长度为 l 的全局状态序列 $s^{t-l+1}, \dots, s^t$, 通过 LSTM 挖掘状态的时序信息, 再输入全连接层, 使用 ReLU 激活函数提取非线性特征, 输出层通过 Sigmoid 激活函数处理并经二值化处理后生成联盟向量。

4 实验与结果分析

4.1 实验环境与参数设置

为验证 CG-LSTM-MATD3 算法在液压支架群调直任务中的有效性, 本文基于 PyTorch 深度学习框架, 搭建了遵循 OpenAI Gym 接口规范的多智能体仿真实验平台. 平台用于模拟综采工作面中采煤机割煤后液压支架的拉架调直过程, 平台将每台液压支架抽象为独立智能体, 各智能体通过感知局部环境状态, 依据算法输出拉架动作, 环境根据液压支架群当前的整体直线度以及支架的移动距离计算奖励, 引导智能体通过协同决策将支架群调整至平直状态。

所有实验均在统一的硬件环境中进行, 服务器配置为 Intel Core i7-12700 处理器, 32 GB 内存及 NVIDIA GeForce RTX 4090 GPU, 操作系统为 Ubuntu 18.04. 训练过程采用 TensorBoard 进行可视化监控, 并基于日志中的实际运行时间统计算法训练耗时. 算法超参数设置如表 1 所示. 实验结果中奖励曲线与直线度曲线的横坐标均为训练回合, 纵坐标分别为奖励与直线度. 奖励综合工作面支架调直精度与移架推进效率, 奖励取值越高, 代表多支架协同调直综合性能越优. 直线度用来量化液压

表 1 算法超参数设置
Table 1 Algorithm hyperparameter setting

参数	Actor 学习率	Critic 学习率	联盟 学习率	折扣 因子	软更新 参数	步数	训练 总回合	经验池 容量	批容量	动作梯度 步数	扩散推理 时间步	动作 学习率	Actor、Critic 梯度裁剪范数	动作梯度 范数比例
MATD3	3×10^{-4}	3×10^{-4}	—	0.99	5×10^{-3}	10	2000	10000	256	—	—	—	—	—
CG-MATD3	3×10^{-4}	3×10^{-4}	3×10^{-4}	0.99	5×10^{-3}	10	2000	10000	256	—	—	—	—	—
CG-LSTM-MATD3	3×10^{-4}	3×10^{-4}	3×10^{-4}	0.99	5×10^{-3}	10	2000	10000	256	—	—	—	—	—
DIPO	3×10^{-4}	3×10^{-4}	—	0.99	5×10^{-3}	10	2000	10000	256	4	4	0.04	1.0	0.1

支架群整体偏移误差, 直线度数值越趋近于零, 支架排布直线度越好; 数值波动幅度越小, 调直控制策略的稳态性能越强.

4.2 不同智能体规模实验

本节通过在 5 ~ 10 个智能体规模下开展实验, 分析 CG-LSTM-MATD3 算法的可扩展能力. 图 5(a)、图 5(b) 分别给出算法在不同规模下的奖励与直线度曲线.

由图可知, 在 5 ~ 10 个智能体规模下, 算法均可稳定收敛, 并保持良好的调直效果. 智能体数量为 5 时, 收敛速度最快, 约 500 回合即可趋于稳定; 当数量增加至 6、7 个时, 收敛回合数逐渐增加, 分别在约 650 回合和 800 回合达到稳定; 当智能体数量从 8 个逐渐增加至 10 个时, 收敛回合逐渐增加, 策略稳定性出现小幅下降.

通常情况下, 综采工作面调直采用相邻 5 ~ 10 架作为一组, 通过成组拉架分段完成调直. 实验结果显示 7 个智能体在收敛速度与调直效果上取得了较好的平衡, 综合表现最优. 若将规模扩大至数十台甚至上百台时, 算法可能面临收敛困难、策略稳定性下降的挑战.

4.3 对比实验

通过将本文算法 CG-LSTM-MATD3 与多智

能体双延迟深度确定性策略梯度算法^[9] (MATD3)、合作博弈双延迟深度确定性策略梯度算法 (CG-MATD3)、Shapley Q 值深度确定性策略梯度算法^[24] (SQDDPG)、基于 Shapley 值的多智能体 Q 学习算法^[25] (SHAQ) 以及基于扩散模型的多智能体强化学习算法^[31] (DIPO) 进行对比, 验证各算法在液压支架调直任务中的表现.

4.3.1 五个智能体对比实验

如图 6 所示, 在五个智能体场景下, 各对比算法均能得到较好的直线度. CG-LSTM-MATD3 通过合作博弈机制与长短期记忆结构的融合, 使智能体捕捉到环境状态的时序信息和细致的信用分配, 在初期就实现了收敛速度与策略收益的平衡. CG-MATD3 凭借信用分配机制获得了较好的收敛效率. 相比之下, MATD3、DIPO 与 SQDDPG 等算法的收敛速度相对滞后, 但在后期均能达到最优直线度. SHAQ 在初期能获得较高的奖励, 但在直线度控制精度上仍有提升空间.

4.3.2 六个智能体对比实验

在六个智能体场景下, 由于任务复杂度增加, 各算法的收敛时间出现变化, 如图 7 所示. 在初期, CG-LSTM-MATD3 曲线受策略探索空间增大的影响出现一定波动, 但凭借长短期记忆结构对时序特征的建模能力, 以及合作博弈在信用分配上的优势,

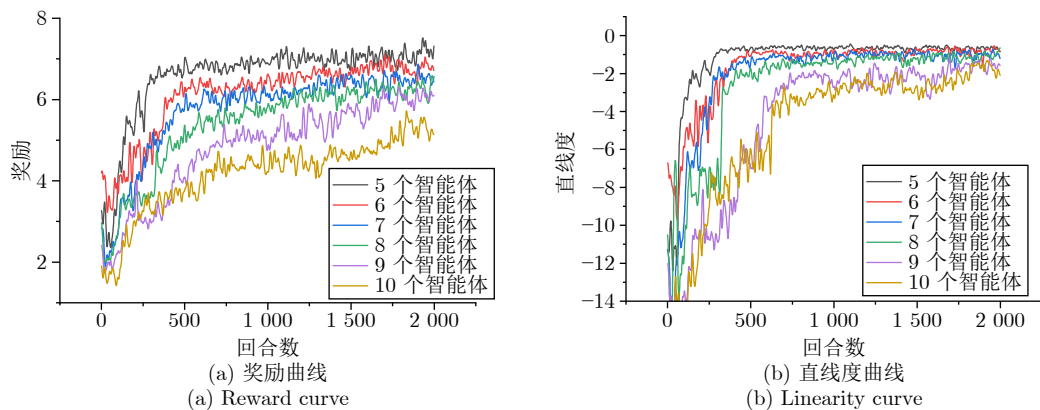


图 5 5 ~ 10 个智能体下的性能曲线
Fig.5 Performance curves with 5 ~ 10 agents

使得算法在性能曲线与收敛稳定性上均优于其他对比算法. CG-MATD3 与 MATD3 最终也可以获得较高的奖励以及直线度数值, 但在收敛效率上仍有待优化. 相较于五个智能体场景, SQDDPG 整体波

动表现有所改善, 最终仍然可以获得较高的直线度.

4.3.3 七个智能体对比实验

在七个智能体数量下, 环境复杂度显著增强, 各算法性能对比如图 8 所示. CG-LSTM-MATD3

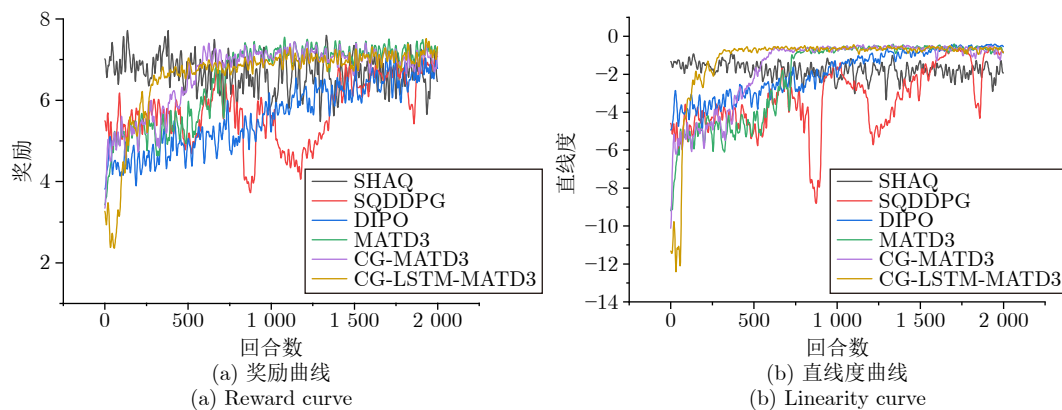


图 6 五个智能体下的奖励和直线度曲线

Fig.6 Curves of reward and linearity for 5 agents

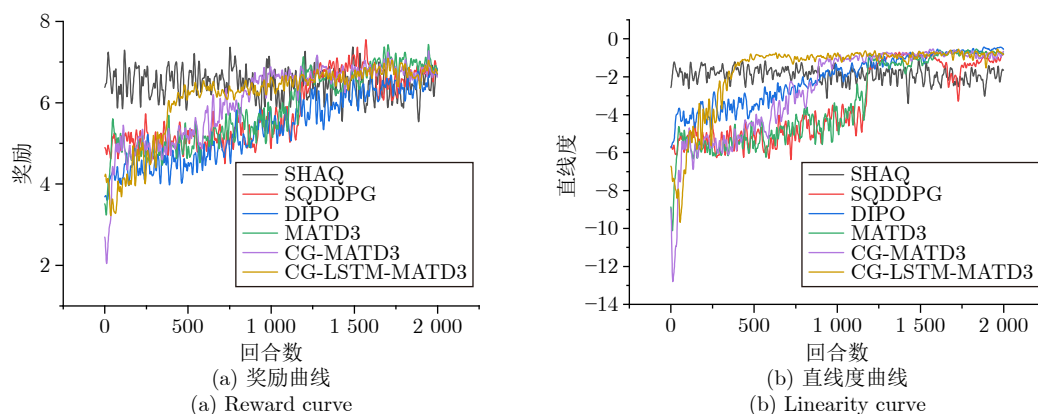


图 7 六个智能体下的奖励和直线度曲线

Fig.7 Curves of reward and linearity for 6 agents

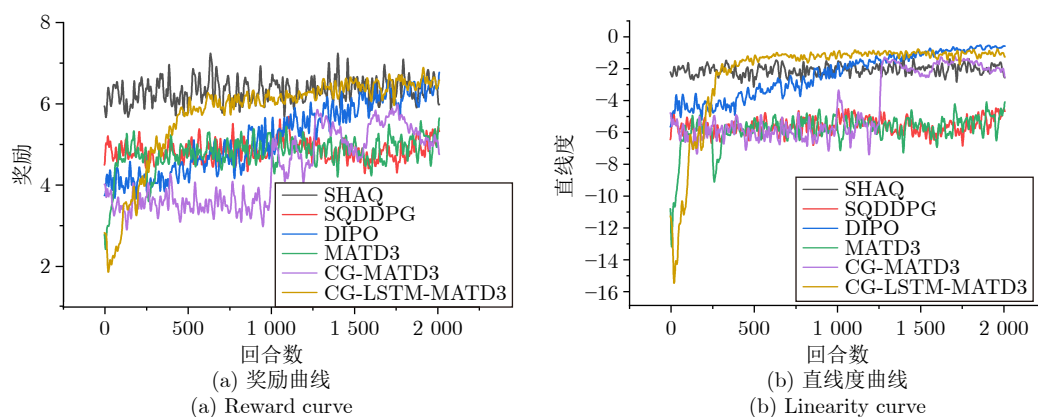


图 8 七个智能体下的奖励和直线度曲线

Fig.8 Curves of reward and linearity for 7 agents

凭借长短期记忆与合作博弈信用分配机制的协同作用, 在整体性能与收敛稳定性上有显著优势. 相比之下, CG-MATD3 由于缺乏对时序记忆信息的利用, 奖励曲线出现较大波动, 且直线度曲线的收敛效率明显降低. DIPO 算法收敛所需回合数较多, 但在最后阶段取得了略高于本文算法的直线度表现.

4.4 不同干扰下算法的鲁棒性

综采工作面井下环境极其复杂, 经常面临突发地质断层导致的地形突变, 以及因设备老化或损伤引发的支架故障等问题. 规模实验表明, 7 个智能体在收敛效率与直线度精度上达到最佳平衡. 本节以 7 个智能体为基准, 设置无干扰的原始场景 (Original)、液压支架偏斜 (Scene 1)、传感器噪声 (Scene 2)、煤层起伏 (Scene 3) 以及混合噪声 (Scene 4) 干扰下的实验. 为评估算法性能, 鲁棒性实验采取多智能体回报值收敛之后的数据进行对比分析.

4.4.1 干扰环境设置

为模拟综采工作面液压支架拉架过程所面临的极端条件, 本实验在原始场景 (Original) 的基础上, 设计以下四种干扰.

1) 液压支架偏斜 (Scene 1). 模拟液压支架偏斜等故障产生的动作执行效率衰减, 设定液压支架实际执行动作在指令值的 80% ~ 100% 之间随机波动, 即引入服从均匀分布的效率系数, 以此复现因设备老化导致的推移力不足现象.

2) 传感器噪声 (Scene 2). 模拟液压支架传感器因长期服役老化或井下强电磁干扰导致的读数偏差. 在智能体的观测状态上叠加均值为 0、标准差为 0.3 的高斯白噪声, 使智能体获取的拉架距离信息存在随机误差, 以此验证算法在感知信息不准确时的决策有效性.

3) 煤层起伏 (Scene 3). 模拟突发地质断层或煤层顶底板剧烈起伏导致的非平稳环境. 在智能体输出的动作上, 添加 0.4 的均匀分布随机漂移量, 模拟地质构造突变产生的额外动阻力或倾角重力分量, 导致支架实际位移与指令值产生偏差.

4) 混合噪声 (Scene 4). 上述三种噪声同时存在的最恶劣工况.

4.4.2 实验结果与分析

本节基于上述干扰设置对 CG-LSTM-MATD3 算法进行测试, 实验结果如表 2 与图 9 所示.

表 2 不同场景下的奖励与直线度数值

Table 2 Reward and linearity values under different scenes

场景	说明	奖励	直线度
Original	无干扰	6.476	-1.028
Scene 1	液压支架偏斜	6.180	-1.188
Scene 2	传感器噪声	5.930	-1.534
Scene 3	煤层起伏	6.416	-0.985
Scene 4	混合噪声干扰	5.895	-1.683

由实验数据可知, CG-LSTM-MATD3 算法在所有测试场景下均能收敛, 但是不同干扰模式下收敛速度、直线度精度存在差异. 在 Scene 1 中, 算法平均奖励为 6.180, 直线度为 -1.188, 相较于 Original 分别下降了 4.6% 和 15.6%, 液压支架偏斜对算法造成的影响较小. 在 Scene 2 中, 平均奖励为 5.930, 直线度降低至 -1.534, 相比 Original 分别下降了 8.4% 和 49.2%, 由直线度曲线可以明显看出, 算法收敛回合数延迟约 1 000 回合, 说明传感器噪声对算法的影响较大.

Scene 3 的表现优于预期, 平均奖励为 6.416, 与 Original 基本持平, 直线度达到 -0.985, 较 Ori-

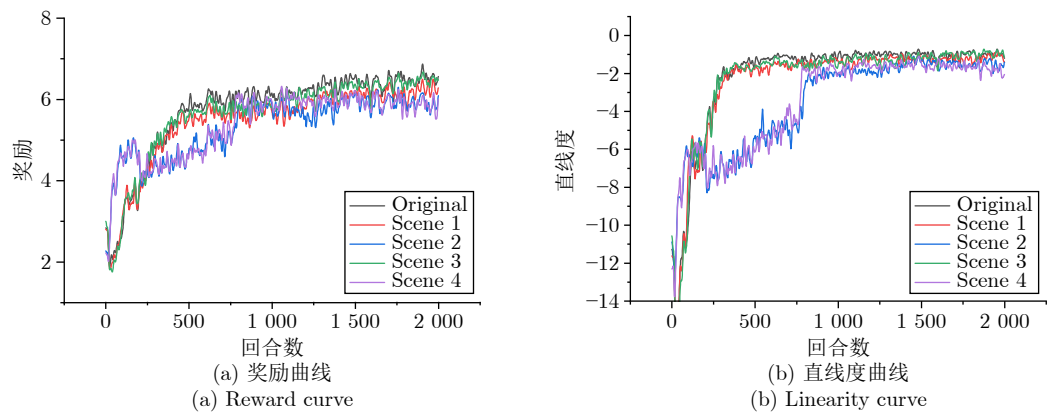


图 9 不同噪声场景下七个智能体的性能曲线

Fig.9 Performance curves of 7 agents under different noise scenes

ginal 提升了 4.2%, 说明煤层起伏这类环境干扰有益于算法训练. 在 Scene 4 的混合干扰场景中, 平均奖励为 5.895, 直线度为 -1.683, 相较于 Original 分别下降了 9.0% 和 63.7%, 且收敛所需回合数增加, 这一结果表明, 多种干扰因素的叠加对算法收敛速度、直线度精度产生了较为明显的负面影响.

综上所述, 随着环境干扰的随机性与复杂性逐步增加, CG-LSTM-MATD3 算法的奖励与直线度总体上呈现下降趋势, 收敛回合数也根据干扰类型受到不同程度的影响.

4.5 消融实验及工程部署分析

为进一步探究 CG-LSTM-MATD3 算法中核心机制的贡献, 明确模块间的协同效应, 本节围绕算法核心构成开展两部分消融实验, 同时结合工程实际需求测试相关技术指标, 验证算法的工程应用可行性. 具体实验分为两个: 一是经验回放机制与 LSTM 模块在样本处理与时序特征捕捉上的互补作用; 二是 LSTM 模块与合作博弈 Shapley 值对算法整体性能的独立贡献及协同增效效果.

由于在工程现场中液压支架多采用成组拉架的方法, 且经前期验证, 7 台液压支架在多智能体协同拉架策略的收敛速度、调直效果上达到整体最优, 因此本次消融实验在 7 台液压支架智能体上开展. 同时在消融实验中测试了运行时间、GPU 占用率、模型参数量三个核心工程化技术指标, 为算法的工程部署提供数据支撑.

为精准评估各模块在液压支架调直任务中的独立贡献与协同效应, 消融实验均使用多智能体回报值收敛之后的数据进行对比分析.

4.5.1 经验回放与 LSTM 的有效性验证

经验回放与 LSTM 分别从样本独立性和时序特征捕捉两个维度形成互补. 为量化验证二者互补作用带来的实际收益, 本文以基础 MATD3 算法为基准, 设计三组对照实验, 分别测试“仅保留经验回放”、“引入 LSTM+顺序采样”、“融合 LSTM+经验回放”三种配置下的算法表现. 实验选取算法收敛后的数据进行对比分析, 重点从收敛回合数、平均直线度、平均奖励三个维度验证各部分的有效性,

实验结果如表 3 所示.

实验表明, 仅采用经验回放的 MATD3 算法因无法捕捉状态时序关联, 最终未能收敛至最优策略, 平均直线度为 -5.301, 平均奖励为 5.001. 在经验回放机制上融合 LSTM 方法后, 平均直线度提升至 -1.497, 较单一经验回放提升了 71.76%; 平均奖励提升至 6.320, 涨幅达 26.37%. 这一结果表明, 经验回放与 LSTM 的融合有效提升了液压支架群的决策准确性, 使调直效果更接近理想的平直状态, 充分发挥了二者的互补优势.

通过对比“LSTM+顺序采样”与“LSTM+经验回放”可知, 融合 LSTM 的 MATD3 在采用经验回放时, 较顺序采样提前约 180 回合达到收敛; 收敛后的直线度额外提升了 5.62%, 奖励提升了 15.43%. 这表明, 经验回放的随机采样机制加快了液压支架调直决策算法的收敛速度, 并且能帮助智能体找到更优的协同拉架策略, 进一步验证了融合经验回放和 LSTM 的有效性.

4.5.2 消融实验及结果分析

在确立了时序采样机制优势的基础上, 本文进一步探究了 LSTM 模块与合作博弈机制的协同贡献. 在保持超参数一致的前提下, 本文使用 MATD3 算法、引入 LSTM 的 LSTM-MATD3、引入 Shapley 值的 CG-MATD3 以及融合两者的 CG-LSTM-MATD3 四种方案开展对比. 消融实验对比结果如表 4 所示. 表 4 中的 Avg 与 Std 分别表示各算法在收敛后数据的平均值与标准差.

MATD3 算法作为基准算法, 单步运行时间为 6.88 ms, GPU 占用为 7.71 MB, 模型参数量为 344.9 KB, 计算成本最低, 但算法的平均直线度为 -5.301, 直线度标准差较大, 表明智能体在液压支架调直决策过程中存在较大波动性; 平均奖励为 5.001, 奖励标准差为 0.89, 表明算法稳定性有待提高. 在引入时序记忆模块后, LSTM-MATD3 算法单步运行时间为 9.34 ms, GPU 占用为 17.42 MB, 模型参数量为 363.7 KB; 平均直线度为 -1.497, 与 MATD3 相比提升了 71.76%, 直线度标准差有较小改进; 平均奖励为 6.320, 奖励标准差由 0.89 降至 0.47, 这表明时序记忆机制能使智能体获取高回报,

表 3 LSTM 与经验回放模块互补性分析
Table 3 Complementary analysis of LSTM and experience replay modules

MATD3 算法配置	收敛回合数	平均直线度	直线度提升	平均奖励	奖励提升
经验回放	未收敛至最优	-5.301	—	5.001	—
LSTM+顺序采样	530	-1.795	66.14%	5.548	10.94%
LSTM+经验回放	350	-1.497	71.76%	6.320	26.37%

表 4 消融实验
Table 4 Ablation experiment

	CG	LSTM	直线度			奖励			运行时间 (ms)	GPU 占用 (MB)	模型参数量 (KB)
			Avg	Std	提升	Avg	Std	提升			
MATD3	×	×	-5.301	1.71	—	5.001	0.89	—	6.88	7.71	344.9
LSTM-MATD3	×	✓	-1.497	1.61	71.76%	6.320	0.47	26.37%	9.34	17.42	363.7
CG-MATD3	✓	×	-1.675	0.54	68.40%	5.428	0.53	8.54%	9.34	8.37	357.1
CG-LSTM-MATD3	✓	✓	-1.029	0.43	80.59%	6.477	0.44	29.51%	9.34	19.73	378.9

并且增强策略的稳定性。

CG-MATD3 单步运行时间为 9.34 ms, GPU 占用为 8.37 MB, 模型参数量为 357.1 KB; 算法直线度为 -1.675, 相比 MATD3 算法提升了 68.40%, 直线度标准差降为 0.54, 表明 Shapley 值机制通过合理量化各智能体的边际贡献, 提升了算法的稳定性; 平均奖励为 5.428, 相对 MATD3 算法提升了 8.54%, 奖励标准差为 0.53, 说明合作博弈机制能引导智能体学到更好的策略, 但在提升算法收益方面有限。

CG-LSTM-MATD3 算法融合了 LSTM 与 Shapley 值模块, 直线度为 -1.029, 相对于 MATD3 提升了 80.59%, 平均奖励为 6.477, 提升了 29.51%; 直线度标准差降至 0.43, 奖励标准差进一步减小, 这一结果优于任一模块单独作用的效果。与 LSTM-MATD3 相比, 直线度从 -1.497 提升至 -1.029, 体现了合作博弈机制的作用; 与 CG-MATD3 相比, 平均奖励从 5.428 提升至 6.477, 体现了时序记忆对策略质量的补充。实验表明 LSTM 赋予智能体利用历史信息的能力, 从而提升个体的最优决策能力; Shapley 值机制通过贡献分配促进多智能体间的协调配合。两个模块融合后提升了多智能体的收敛速度和策略收益。

在计算成本方面, 两个模块的引入使运行时间由 6.88 ms 增至 9.34 ms, GPU 占用从 7.71 MB 增至 19.73 MB, 模型参数量从 344.9 KB 增至 378.9 KB, 运行时间有所增加, 但调直决策的时间仍小于 0.2 s。通常情况下, 综采工作面成组调直的完成时间一般控制在几分钟范围内, 因此增加上述两个模块后, 调直决策时间能够达到工程应用的要求。

4.5.3 工程部署可行性及挑战

为评估算法的工程部署潜力, 本文首先在模拟边缘侧的硬件平台 (CPU: i7-12700, GPU: RTX 4090, RAM: 32 GB) 上进行性能测试, 测得单智能体决策时间为 9.34 ms。结果表明, 采用具备相似算力的嵌入式平台, 如 NVIDIA Jetson AGX Orin 系列, 即可满足实时性需求。

然而, 实际工程部署面临严峻的物理环境约束。根据煤矿井下设备的防爆规范, 所有电子设备必须置于密封的液压支架控制箱内。这种密封环境导致高性能计算板卡产生的热量无法及时排出, 散热问题成为当前算法落地的主要障碍。目前的解决方案依赖于在少数特定支架的防爆箱外部加装散热片, 但该方案普适性较差。因此, 解决密闭防爆环境下的高效散热问题, 是后续实现算法大规模工程应用的关键。

5 结束语

1) 在强化学习框架下, 针对多智能体液压支架拉架调直方法面临的信用分配与时序信息难以捕捉问题, 本文提出了 CG-LSTM-MATD3 算法, 基于 Dec-POMDP 建模, 引入 Shapley 值实现合理的信用分配, 设计 Coalition 网络生成联盟来近似计算 Shapley 值以降低计算复杂度。针对液压支架调直过程中的时序依赖特性, 本文将 LSTM 引入 Actor、Critic 和 Coalition 网络中, 使模型能够捕获状态的时序依赖关系, 增强对环境动态特性的感知能力。

2) 实验结果表明, 在智能体数量为 5、6、7 个的时候, 所提方法的直线度明显优于其他算法。在智能体数量为 7 个时, CG-LSTM-MATD3 算法的直线度数值为 -1.029, 相比于 MATD3 算法的直线度性能提升了 80.59%。

3) 在实际工程中, 液压支架多采用成组拉架方式实现调直, 这表明本文算法可满足综采工作面液压支架的实际运行需求, 具备工程实用价值。然而, 在包含数十或上百个液压支架的大规模决策场景中, 算法计算开销增加, 从而导致收敛速度降低甚至无法收敛的现象。因此, 后续工作将重点研究多智能体 DIPO, 深入探索算法框架在大规模液压支架拉架调直任务中的协同决策能力, 进而支撑更大规模液压支架群的协同决策任务。

参考文献

1 Zhang De-Sheng, Zhou Jie, Ren Huai-Wei, Feng Yin-Hui, Han Hui-Jun, Gong Shi-Xin. Key technologies for intelligent collaborative control of equipment group in super high mining face. *Coal*

- Technology*, 2025, **44**(10): 1–5
(张德生, 周杰, 任怀伟, 冯银辉, 韩会军, 巩师鑫. 超大采高工作面装备群智能协同控制关键技术. 煤炭技术, 2025, **44**(10): 1–5)
- 2 Murphy K. Reinforcement learning: An overview. arXiv preprint arXiv: 2412.05265, 2024.
- 3 Luo Kai-Cheng, Gao Yang, Yang Yi, Chang Ya-Jun, Yuan Rui-Fu. Control strategy of drawing opening based on mean deviation reward function. *Coal Engineering*, 2022, **54**(9): 105–111
(罗开成, 高阳, 杨艺, 常亚军, 袁瑞甫. 基于均值偏差奖励函数的放煤口控制策略研究. 煤炭工程, 2022, **54**(9): 105–111)
- 4 Yang Yi, Li Qing-Yuan, Li Hua-Min, Li Dong-Yin, Yang Yan-Lin, Fei Shu-Min. Research on intelligent decision-making of group drawing based on batch reinforcement learning. *Coal Science and Technology*, 2022, **50**(10): 188–197
(杨艺, 李庆元, 李化敏, 李东印, 杨延麟, 费树岷. 基于批量式强化学习的群组放煤智能决策研究. 煤炭科学技术, 2022, **50**(10): 188–197)
- 5 Yang Yi, Wang Sheng-Wen, Cui Ke-Fei, Fei Shu-Min. Intelligent drawing decision method based on fuzzy deep Q-network. *Industry and Mine Automation*, 2023, **49**(4): 78–85
(杨艺, 王圣文, 崔科飞, 费树岷. 基于模糊深度 Q 网络的放煤智能决策方法. 工矿自动化, 2023, **49**(4): 78–85)
- 6 Yang Yi, Sun Ying-Jie, Chang Ya-Jun, Liu Bin-Bin, Wang Ke-Ping. Shearer positioning method based on chain base station coordinate fusion. *Industry and Mine Automation*, 2025, **51**(5): 49–56
(杨艺, 孙英杰, 常亚军, 刘斌斌, 王科平. 基于链式基站坐标融合的采煤机定位方法. 工矿自动化, 2025, **51**(5): 49–56)
- 7 Luo B, Liu D, Wu H N, Huang T, Yang C, Gui W. Recent advances on off-policy reinforcement learning for optimization control. *IEEE Transactions on Cybernetics*, DOI: 10.1109/TCYB.2026.3683384
- 8 Yang Y, Dai Y, Wang T, Qian W. Hydraulic-supports alignment by TD3 with segmented experience pool. *Neural Processing Letters*, 2025, **57**: Article No. 35
- 9 Guo Lei, Liang Cheng-Qing. Multi-agent collision avoidance control based on MATD3 algorithm. *Computing Technology and Automation*, 2024, **43**(1): 9–15
(郭雷, 梁成庆. 基于 MATD3 算法的多智能体避碰控制. 计算技术与自动化, 2024, **43**(1): 9–15)
- 10 Shen G C, Wang Y. Review on Dec-POMDP model for MARL algorithms. In: *Proceedings of the Smart Communications, Intelligent Algorithms and Interactive Methods*. Singapore: Springer, 2022. 29–35
- 11 Alvarez-Mozos M, Macho-Stadler I, Perez-Castrillo D. Sequential creation of surplus and the Shapley value. *Games and Economic Behavior*, 2026, **155**: 149–166
- 12 Neuman E, Tuschmann S. Stochastic graphon games with interventions. arXiv preprint arXiv: 2507.00561, 2025.
- 13 Krichen M, Mihoub A. Long short-term memory networks: A comprehensive survey. *AI*, 2025, **6**(9): Article No. 215
- 14 Bai Jin-Ming, Wang Ran-Feng, Fu Xiang. Straightness measurement method of hydraulic support based on walking robot between supports. *Industry and Mine Automation*, 2019, **45**(1): 45–51
(白晋铭, 王然风, 付翔. 基于架间行走机器人的液压支架直线度测量方法. 工矿自动化, 2019, **45**(1): 45–51)
- 15 Zhang Xu-Hui, Wang Dong-Man, Yang Wen-Juan. Position and posture detection method of hydraulic support based on visual measurement. *Industry and Mine Automation*, 2019, **45**(3): 56–60
(张旭辉, 王冬曼, 杨文娟. 基于视觉测量的液压支架位姿检测方法. 工矿自动化, 2019, **45**(3): 56–60)
- 16 Zhang Shu-Nan, Cao Xian-Gang, Cui Ya-Zhong, Luo Xuan, Zhang Guo-Zhen. Research on straightness measurement method of hydraulic support based on multi-sensor. *Coal Mine Machinery*, 2020, **41**(4): 56–59
(张树楠, 曹现刚, 崔亚仲, 罗璇, 张国祯. 基于多传感器的液压支架直线度测量方法研究. 煤矿机械, 2020, **41**(4): 56–59)
- 17 Wang Yu-Zhuo, Chang Zong-Xu, Gao Fei, Lian Zi-Sheng. Research on alignment method of hydraulic support. *Journal of Mechanical & Electrical Engineering*, 2021, **38**(5): 645–649
(王宇卓, 常宗旭, 高飞, 廉自生. 液压支架的调直方法研究. 机电工程, 2021, **38**(5): 645–649)
- 18 Song Dan-Yang, Lu Chun-Gui, Tao Xin-Ya, Yang Jin-Heng, Wang Pei-En, Zheng Wen-Qiang. Hydraulic support alignment method based on maximum entropy Kalman filter algorithm. *Industry and Mine Automation*, 2022, **48**(11): 119–124
(宋单阳, 卢春贵, 陶心雅, 杨金衡, 王培恩, 郑文强. 基于最大熵卡尔曼滤波算法的液压支架调直方法. 工矿自动化, 2022, **48**(11): 119–124)
- 19 Hu Bo, Lian Zi-Sheng. Research on hydraulic support alignment system based on support vector machine and genetic algorithm. *Coal Mine Machinery*, 2014, **35**(10): 39–41
(胡波, 廉自生. 基于支持向量机和遗传算法的液压支架调直系统研究. 煤矿机械, 2014, **35**(10): 39–41)
- 20 Wang Hong, You Xiu-Song, Li Shou-Bin, Wei Wen-Yan. Research on support following automation based on genetic algorithm and BP neural network. *Coal Science and Technology*, 2021, **49**(1): 272–277
(王虹, 尤秀松, 李首滨, 魏文艳. 基于遗传算法与 BP 神经网络的支架跟机自动化研究. 煤炭科学技术, 2021, **49**(1): 272–277)
- 21 Li Wen-Jun, Zhou Zhan. Automatic alignment technology of fully mechanized working face based on inertial navigation system. *Shaanxi Coal*, 2022, **41**(4): 130–133
(李文俊, 周展. 基于惯导系统的综采工作面自动调直技术. 陕西煤炭, 2022, **41**(4): 130–133)
- 22 Wang Yun-Fei, Zhao Ji-Yun, Zhang He, Wang Hao, Zhang Yang. Straightness control method of hydraulic support group pushing system based on neural network compensation. *Coal Science and Technology*, 2024, **52**(11): 174–185
(王云飞, 赵继云, 张鹤, 王浩, 张阳. 基于神经网络补偿的液压支架群推移系统直线度控制方法. 煤炭科学技术, 2024, **52**(11): 174–185)
- 23 Sun Ming-Ze, Wang Yong-Qiang, Chang Ya-Jun, Zhu De-Sheng, Li Shi-Yan, Yang Ke-Hu. Distributed cooperative synchronization control of hydraulic support robot group for advancing. *Journal of China Coal Society*, 2024, **49**(S2): 1208–1222
(孙铭泽, 王永强, 常亚军, 朱德昇, 李石岩, 杨克虎. 液压支护机器人群组移架一致性分布式协同控制. 煤炭学报, 2024, **49**(S2): 1208–1222)
- 24 Wang J, Zhang Y, Kim T K, Gu Y. Shapley Q-value: A local reward approach to solve global reward games. arXiv preprint arXiv: 1907.05707, 2019.
- 25 Wang J, Zhang Y, Gu Y, Kim T K. SHAQ: Incorporating Shapley value theory into multi-agent Q-learning. arXiv preprint arXiv: 2105.15013, 2021.
- 26 Li J. Shapley counterfactual credits for multi-agent reinforcement learning. arXiv preprint arXiv: 2106.00285, 2021.
- 27 Heuillet A, Couthouis F, Diaz-Rodriguez N. Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with Shapley values. *IEEE Computational Intelligence Magazine*, 2022, **17**(1): 59–71
- 28 Wang J. Shapley value based multi-agent reinforcement learning: Theory, method and its application to energy network. arXiv preprint arXiv: 2402.15324, 2024.
- 29 Qin H, Zhang W, Tian R. Collaborative control method of transit signal priority based on cooperative game and reinforcement learning. In: *Proceedings of the 4th IEEE International*

Conference on Electronic Technology, Communication and Information (ICETCI). New York, USA: IEEE, 2024. 537–542

- 30 Tang C, Pan L, Chen J, Liu Y, Lai J. A game theory-reinforcement learning approach to cooperation for UAVs. *IEEE Transactions on Vehicular Technology*, 2025, **74**(6): 9864–9869
- 31 Yang L. Policy representation via diffusion probability model for reinforcement learning. arXiv preprint arXiv: 2305.13122, 2023.



杨 艺 河南理工大学电气工程与自动化学院副教授. 2007 年获得河南理工大学硕士学位, 2017 年获得北京航空航天大学博士学位. 主要研究方向为人工智能, 强化学习, 自适应动态规划及其在煤矿中的应用. 本文通信作者. E-mail: yangyi@hpu.edu.cn

(**YANG Yi** Associate professor at the School of Electrical Engineering and Automation, Henan Polytechnic University. He received his master degree from Henan Polytechnic University in 2007, and his Ph.D. degree from Beihang University in 2017. His research interests include artificial intelligence, reinforcement learning, adaptive dynamic programming, and their applications in coal mines. Corresponding author of this paper.)



王静怡 河南理工大学电气工程与自动化学院硕士研究生. 主要研究方向为多智能体强化学习.

E-mail: Anna_wangjy@outlook.com

(**WANG Jing-Yi** Master student at the School of Electrical Engineering and Automation, Henan Poly-

technic University. Her main research interest is multi-agent reinforcement learning.)



钱 伟 河南理工大学电气工程与自动化学院教授. 2005 年获得东南大学硕士学位, 2009 年获得浙江大学工业控制技术国家重点实验室博士学位. 主要研究方向为时滞系统, 随机系统, 网络控制系统和多智能体系统.

E-mail: qwei@hpu.edu.cn

(**QIAN Wei** Professor at the School of Electrical Engineering and Automation, Henan Polytechnic University. He received his master degree from Southeast University in 2005, and his Ph.D. degree from the State Key Laboratory of Industrial Control Technology, Zhejiang University in 2009. His research interests include time-delay systems, stochastic systems, networked control systems, and multi-agent systems.)



王 田 北京航空航天大学人工智能学院教授. 2010 年获得西安交通大学硕士学位, 2014 年获得法国特鲁瓦工程技术大学博士学位. 主要研究方向为人工智能, 计算机视觉和模式识别.

E-mail: wangtian@buaa.edu.cn

(**WANG Tian** Professor at the School of Artificial Intelligence, Beihang University. He received his master degree from Xi'an Jiaotong University in 2010, and his Ph.D. degree from Université de Technologie de Troyes, France, in 2014. His research interests include artificial intelligence, computer vision, and pattern recognition.)