

基于强化学习与博弈优化的多工单协同排产研究

黄起豹¹, 沈建斌²

(1. 上饶师范学院 数字技术应用产业学院, 江西 上饶 340001;
2. 深圳市酷开智能系统科技有限公司, 广东 深圳 518000)

摘要: 针对智能制造多品种、小批量模式下多工单协同排产的资源冲突、多主体目标异构及动态扰动适应性差等问题, 提出融合强化学习与博弈优化的协同排产方法。该方法通过构建资源调度中心、生产单元、工单管理者三方主从—合作博弈模型来刻画利益交互, 同时引入深度Q网络(DQN)实现策略自适应优化。在汽车零部件制造、电子设备组装公开数据集上开展实验, 结果表明, 在汽车零部件制造场景中, 所提方法较GA、PSO、DQN算法的排产周期分别缩短18.5%、9.9%、7.2%, 资源利用率分别提升25.6%、18.3%、12.5%, 总成本分别降低16.2%、8.7%、6.3%; 在电子设备组装场景中, 所提方法较GA、PSO、DQN算法的排产周期分别缩短21.2%、11.9%、8.5%, 资源利用率分别提升27.9%、19.8%、13.8%, 总成本分别降低17.7%、10.7%、7.1%。此外, 鲁棒性测试表明, 在扰动概率为30%时, 所提方法在两类场景下的平均变异系数仅为0.045~0.047。该方法实现了多主体利益均衡与动态排产优化的统一, 为复杂制造场景下的排产提供了新思路。

关键词: 多工单协同排产; 强化学习; 博弈论; 生产调度; DQN

DOI: 10.11907/rjtk.261077

中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(2026)005-0036-07



扫描二维码阅读全文:

Research on Multi-Workorder Collaborative Scheduling Based on Reinforcement Learning and Game Optimization

HUANG Qibao¹, SHEN Jianbin²

(1. School of Digital Technology Application Industry, Shangrao Normal College, Shangrao 340001, China;
2. Shenzhen Coolview Intelligent System Technology Co., Ltd., Shenzhen 518000, China)

Abstract: To address the issues of resource conflicts, heterogeneous objectives among multiple entities, and poor adaptability to dynamic disturbances in collaborative scheduling for multiple work orders under the multi-variety and small-batch production mode in intelligent manufacturing, a collaborative scheduling method integrating reinforcement learning with game optimization is proposed. This method characterizes the interactions of interests by constructing a three-party leader-follower-cooperative game model involving a resource scheduling center, production units, and work order managers. Meanwhile, it introduces the Deep Q-Network (DQN) to achieve adaptive strategy optimization. Experiments conducted on publicly available datasets from automotive component manufacturing and electronic device assembly scenarios demonstrate that, in the automotive component manufacturing scenario, the proposed method reduces the scheduling cycle by 18.5%, 9.9%, and 7.2%, improves resource utilization by 25.6%, 18.3%, and 12.5%, and lowers total costs by 16.2%, 8.7%, and 6.3% compared to Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and DQN algorithms, respectively. In the electronic device assembly scenario, the proposed method reduces the scheduling cycle by 21.2%, 11.9%, and 8.5%, improves resource utilization by 27.9%, 19.8%, and 13.8%, and lowers total costs by 17.7%, 10.7%, and 7.1% compared to GA, PSO, and DQN algorithms, respectively. Furthermore, robustness tests indicate that, with a disturbance probability of 30%, the proposed method achieves an average coefficient of variation of only 0.045 - 0.047 in both scenarios. This method achieves the unification of multi-entity interest balancing and dynamic scheduling optimization, providing a novel approach for scheduling in complex manufacturing scenarios.

Key Words: multi-workorder collaborative scheduling; reinforcement learning; game theory; production scheduling; DQN

收稿日期: 2026-02-05

基金项目: 全国高等学校计算机教育研究会项目(CERACU2026R07); 中国机械工业教育协会项目(ZJJX25SY001)

作者简介: 黄起豹(1978-), 男, 博士, 上饶师范学院数字技术应用产业学院副教授、硕士生导师, 研究方向为人工智能与产业融合应用; 沈建斌(1975-), 男, 博士, 深圳市酷开智能系统科技有限公司总监, 研究方向为人工智能产品开发。本文通讯作者: 黄起豹。

0 引言

随着智能制造向多品种、小批量定制化模式深度转型,生产系统面临多工单并行、资源动态竞争与订单频繁扰动的复杂环境。多工单协同排产作为连接生产计划与现场执行的关键环节,其核心挑战在于在有限资源约束下协调加工顺序、分配生产单元、安排时间窗口,同时应对设备故障、紧急插单等动态事件。传统集中式排产方法往往假设静态确定环境,难以平衡资源调度中心、生产单元与工单管理者等多主体间的异构利益目标,导致全局优化与局部利益冲突、计划刚性与动态扰动适应性差等突出问题。因此,研究一种能够实现多主体利益均衡并具备动态自适应能力的协同排产方法,对于提升制造系统响应速度、资源利用效率及抗扰动能力具有重要的理论价值与工程意义。

1 相关研究

多工单协同排产是智能制造生产调度的核心问题,现有研究主要围绕排产优化算法、强化学习应用、博弈优化实践3条技术路线展开。在排产优化算法方面,刘建军等^[1]探讨了复杂制造系统的人机交互调度排产方法,强调了人在环境决策中的补充作用。宋庭新等^[2]提出一种基于多目标灰狼优化算法的柔性作业车间多目标排产方法,解决了制药车间生产工艺复杂、多品种小批量生产效率低及频繁清洗换型等问题。董绍康等^[3]对混合排产问题的求解与应用进行综述,指出多主体、多层次模型缺乏高效自适应算法。Zhang等^[4]系统梳理了从单智能体到多智能体强化学习的工业应用进展,指出动态环境下的策略鲁棒性与可扩展性是核心瓶颈。Fathollahi^[5]提出的深度学习模态优化方案解决了能源系统在动态扰动下模态失稳与功率波动的问题。Ju等^[6]构建了虚拟电厂集群两阶段调度优化与收益分配模型,解决了多主体利益冲突与收益分配不公问题。Kou等^[7]基于排队论和模糊建模进行可再生能源投资决策,解决了资源竞争与等待时间评估困难问题。

在强化学习应用方面,Dong等^[8]设计了大语言模型赋能的低代码工业机器人编程系统,为人机协同排产提供了新工具。Du等^[9]结合模仿学习与迭代贪心算法求解流水线调度问题,为强化学习与启发式规则的融合提供了新思路。杨博宇等^[10]基于启发式搜索实现极化雷达目标检测,其搜索策略可用于排产解空间剪枝。吴正佳等^[11]采用改进遗传算法(Genetic Algorithm, GA)求解多资源约束并行机分批调度问题,获得较好静态解。Qasim等^[12]提出结合MCDM的改进GA,用于云环境工业物联网作业调度,验证了GA在资源分配中的有效性。Xu等^[13]结合混合高斯

量子粒子群(Particle Swarm Optimization, PSO)与自适应GA求解柔性作业车间调度问题,在静态算例中表现较好。Thepphakorn等^[14]进行多目标混合优化排产,目标权衡思想可借鉴。Migliavacca^[15]对多能源静态建模方法进行综述,强调多主体互动建模的必要性,为排产博弈建模提供了理论参照。陆心屹等^[16]将深度强化学习引入柔性作业车间动态节能调度,验证了DQN(Deep Q-Network)在实时响应环境变化中的有效性。胡姜等^[17]探讨了数字技术赋能下新能源汽车企业的利益均衡决策,其均衡思想可用于生产排产。

在博弈优化实践方面,Wu等^[18]构建了独立共享储能系统的分层博弈优化模型,有效协调了多方利益主体。Tian等^[19]研究了有向图上异构高阶博弈器的分布式自适应纳什均衡求解算法,为生产排产中的多智能体协同提供了理论支撑。Liu等^[20]针对多联盟博弈提出分层预设性能算法,其层级架构对于构建生产单元与工单管理者的主从关系具有重要参考价值。都繁杰等^[21]提出基于在线容错的云计算资源调度方法,其容错机制可应对设备故障扰动博弈。魏金花等^[22]为多产消者构建两阶段四层主从一合作博弈鲁棒模型,其混合博弈结构直接启发了本文三方博弈框架的设计。Jiang等^[23]融合水力模型与智能算法进行大位移钻井多参数优化,体现了多目标博弈优化思想。Ventura等^[24]比较了合作博弈中沙普利指数与Choquet积分在多准则决策中的应用,为工单管理者间的利益分配提供了量化依据。Xiao等^[25]基于改进DQN算法优化AGV动态路径规划,进一步证明了强化学习在复杂序贯博弈决策中的应用潜力。

然而,现有研究仍存在一些不足:①排产优化算法多采用集中式静态框架,难以适配动态扰动,未兼顾多主体异构利益;②强化学习虽具备动态适应能力,但应用多聚焦单一工单或单元调度,未构建多主体协同学习机制;③博弈优化模型多停留在双主体静态博弈,三方及以上模型研究薄弱,策略缺乏自学习能力。

针对上述问题,本文提出融合强化学习与博弈优化的协同排产方法(Reinforcement Learning-Game Optimization, RL-Game)。与现有工作相比,其创新与改进在于:①构建三方主从一合作博弈模型,突破现有局限,贴合实际多主体利益交互关系;②将博弈策略优化转化为马尔可夫决策过程,引入DQN算法实现策略自适应迭代学习,弥补传统博弈模型缺陷;③通过消融实验与先进方法比较,定量证明三方博弈模型与强化学习深度融合在多方面的优越性,其实现了多主体利益均衡与动态排产优化统一。

2 多工单协同排产博弈模型构建

本文建立了一个三方主从一合作博弈模型,其中资源调度中心作为主导者,生产单元和工单管理者作为跟随

者。资源调度中心的目标为最大化全局生产效益,综合考虑排产周期、总成本和资源利用率等因素。表示为:

$$\max F_{center} = w_1 \frac{T_{base}}{T_{act}} + w_2 \frac{C_{base}}{C_{act}} + w_3 U_{avg} \quad (1)$$

式中: F_{center} 为资源调度中心的全局生产效益值,无量纲; w_1, w_2, w_3 为权重系数,分别表示排产周期、总成本、资源利用率的重要性占比,满足 $w_1 + w_2 + w_3 = 1$, 本文分别取值为 0.35、0.35、0.30; T_{base} 为基准排产周期(生产日); T_{act} 为实际排产周期(生产日); C_{base} 为基准总成本(万元); C_{act} 为实际总成本(万元); U_{avg} 为平均资源利用率(%),取值范围为 $[0, 100]$ 。

总成本 C_{act} 涵盖设备运行成本、人力成本与延期惩罚成本。计算公式为:

$$C_{act} = \sum_{i=1}^M \sum_{j=1}^{N_i} (Cost_{run}^{ij} + Cost_{lab}^{ij}) \times Time_{act}^{ij} + \sum_{k=1}^K Cost_{pen}^k \times \max(0, T_{act}^k - T_{due}^k) \quad (2)$$

式中: M 为生产单元数量; N_i 为第 i 个生产单元设备数量; $Cost_{run}^{ij}$ 为第 i 个生产单元第 j 台设备的单位运行成本(万元/h); $Cost_{lab}^{ij}$ 为单位人力成本(万元/h); $Time_{act}^{ij}$ 为实际运行时间(h); K 为工单数量; $Cost_{pen}^k$ 为第 k 个工单单位延期惩罚成本(万元/生产日); T_{act}^k 为第 k 个工单实际加工周期(生产日); T_{due}^k 为第 k 个工单的交货期(生产日)。

3 基于强化学习的博弈决策算法设计

3.1 强化学习与博弈的融合框架

为应对生产动态扰动,本文设计多工单协同排产动态博弈决策框架,结构见图 1。该框架包含环境层、博弈层、强化学习层与决策层 4 个模块,各模块协同实现博弈策略自适应优化。环境层模拟多工单生产过程,实时采集资源占用、工单进度、设备状态等信息;博弈层包含资源调度中心、生产单元和工单管理者等主体,各主体根据目标函数与环境状态生成初始策略;强化学习层以 DQN 算法为核心,将策略优化转化为序贯决策问题,通过与环境交互学习最优策略;决策层通过优化策略组合生成工单加工顺序、资源分配方案等排产结果。

3.2 马尔可夫决策过程建模

将多工单协同排产博弈过程转化为马尔可夫决策过程,明确状态空间、动作空间与奖励函数三大核心要素。状态空间 S 全面涵盖生产环境、博弈主体策略与工单加工状态,定义为 $S = \{V_{res}, V_{center}, V_{unit}, V_{order}, V_{proc}\}$, 其中 V_{res} 为资源状态向量; V_{center} 为资源调度中心策略状态; V_{unit} 与 V_{order} 分别为生产单元与工单管理者策略状态; V_{proc} 为工单加工状态向量,描述第 k 个工单剩余加工时间、已加工数量与质量合格率。动作空间 A 对应 3 类主体的策略调整行为,定义为 $A = \{A_{center}, A_{unit}, A_{order}\}$, 其中 A_{center} 包含资源调度中心的资源分配比例、排产周期、成本控制目标调整动作; A_{unit} 涵盖生产单元的加工作业优先级与设备运行参数调

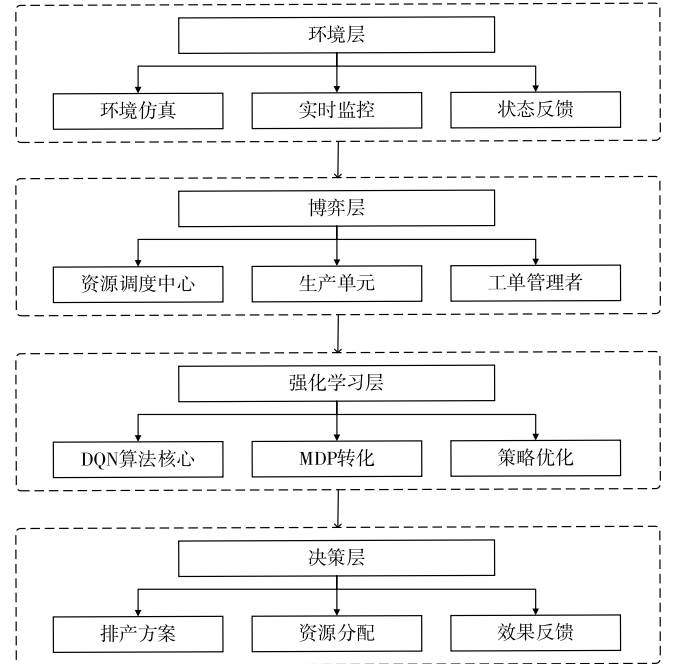


Fig. 1 Dynamic game decision framework for multi-order collaborative production scheduling

图1 多工单协同排产动态博弈决策框架

整动作; A_{order} 涉及工单管理者的订单优先级与质量要求调整动作。奖励函数 R 采用多目标加权形式,综合多主体利益,定义为 $R = \sum \lambda_i R_i$, 其中 λ_i 为奖励权重系数且和为 1, R_i 为 3 类主体的标准化奖励值。

3.3 DQN 算法优化设计

采用 DQN 算法逼近状态—动作价值函数 $Q(s, a)$, 通过经验回放与目标网络机制提升训练稳定性与收敛速度。DQN 网络为全连接神经网络,输入层维度为状态空间维度 $|S|$,接收状态向量 s ; 隐藏层设 3 层全连接层,激活函数为 ReLU,神经元数量分别为 128、64、32; 输出层维度为动作空间维度 $|A|$,输出每个动作的 Q 值。经验回放缓冲区 D 存储经验样本 (s, a, r, s') , 训练时随机采样批量样本打破相关性。缓冲区容量设为 $|D|$, 满时采用先进先出策略替换样本。引入目标网络 Q' 计算目标 Q 值,结构与主网络 Q 一致,每 C 步从主网络复制参数。目标 Q 值的计算公式为 $y = r + \gamma \max_a Q'(s', a')$, 其中 γ 为折扣因子。损失函数采用均方误差 (MSE): $L(\theta) = \frac{1}{B} \sum (y - Q(s, a; \theta))^2$, 式中 B 为批量样本大小, θ 为主网络参数。优化器采用 Adam, 学习率为 η 。

DQN 算法遵循“初始化—交互学习—收敛验证—方案生成—可行性验证”逻辑,具体流程为:在初始化阶段,设定博弈模型、DQN、算法训练参数,确定经验回放缓冲区和生产环境初始状态 s_0 。在迭代学习阶段,在最大时间步 T_{max} 范围内,每个时间步 t 根据 ϵ -贪婪策略选择动作 a_t ; 执行动作后更新主体策略,获得奖励 r_t 与下一状态 s_{t+1} , 样本存入缓冲区。当缓冲区样本量达到阈值时,随机抽取批量

样本计算目标 Q 值, 采用 Adam 优化器对主网络参数进行更新。每间隔 C 步, 将主网络的参数同步复制到目标网络中, 完成目标网络参数更新。在收敛验证阶段, 设置双重训练终止条件, 满足其一即停训: ①损失函数值持续小于收敛阈值 ϵ 且连续稳定步数 N_{stable} 达规定值, 表明算法收敛至最优策略; ②迭代步数达预设最大迭代次数 T_{max} 。最大迭代次数 T_{max} 依据算法收敛性预实验设置。汽车零部件制造和电子设备组装场景预实验结果显示, 4 000~5 000 步时损失函数趋于稳定, 继续增加迭代步数排产性能提升不明显, 故设 5 000 步为最大迭代次数, 以平衡求解效率与优化效果。最后, 对方案是否满足约束条件进行验证, 不满足则调整参数重新迭代。

4 验证实验

4.1 实验方法

4.1.1 实验场景设置

场景一为汽车零部件制造, 以 Brandimarte 标准算例集中的 MK02、MK05 和 MK08 算例为基础, 通过参数校准与扩展设计构建实验场景。该场景包含 3 个生产单元, 各配 8 台设备, 共 20 个并行工单(单批次包含工件数量为 50~150 件), 每个工单包含 4~6 道工序, 交货期为 8~15 个生产日, 最低质量合格率 $Q_{min}=95\%$ 。

场景二为电子设备组装, 基于电子制造调度基准算例 EM-05、EM-10 构建, 包含 4 个生产单元, 各配 6 台设备, 共 25 个并行工单(批量 30~120 件), 每个工单包含 3~5 道工序, 交货期为 6~12 个生产日, $Q_{min}=98\%$ 。两类场景均含有设备故障、订单插单等动态扰动, 采用固定概率随机触发方式模拟, 即每个生产时间步以 10%~35% 的概率随机触发设备故障(单台停机 1~2 个生产日)或订单插单(新增 1~2 个紧急工单, 要求交付周期压缩至 3~5 个生产日)。两类场景的底层参数均以公开基准算例为基准, 并依据典型制造企业的实际运行数据(如设备故障率分布、工序换型时间等)进行校准与扩展, 从而在保障学术可复现性的同时提升实验的工程真实性。

4.1.2 实验环境与参数设置

实验基于 Intel i7-12700H 平台, 采用 Python 3.8 及 PyTorch 1.12.0 实现。参数设置: 博弈模型中 $w_1 = 0.35$, $w_2 = 0.35$, $w_3 = 0.3$, T_{max_delay} 交货期为 10%; DQN 算法中 $\eta = 0.001$, $|D| = 10\ 000$, $B = 64$, $C = 100$, $\gamma = 0.95$, $T_{max} = 5\ 000$ 步。GA 与 PSO 算法种群规模为 100, 最大迭代次数为 500, GA 交叉概率为 0.8、变异概率为 0.05, PSO 惯性权重为 0.7、学习因子 $c_1 = c_2 = 2$ 。新增对比算法为 DQN 单算法、Game-GA, 参数与原设置一致。

4.1.3 消融实验设计

为验证所提三方博弈模型与 DQN 强化学习的协同优化效果, 设计 3 组消融实验: ①消融实验 1(仅强化学习)。

采用 DQN 单算法实现多工单协同排产, 去除三方博弈模型, 直接通过 DQN 优化全局排产指标; ②消融实验 2(博弈模型+传统优化算法)。构建三方博弈模型, 采用 GA 算法作为求解工具, 去除 DQN 强化学习模块; ③消融实验 3(简化版博弈模型)。将三方博弈模型简化为资源调度中心—生产单元双主体博弈模型, 保留 DQN 强化学习模块, 分析工单管理者主体对排产结果的影响。

消融实验与原实验采用相同的场景参数、评价指标, 确保对比的公平性。

4.2 实验结果与分析

在以上两类场景下对所提方法(RL-Game)、GA、PSO、DQN 单算法、Game-GA 进行 10 次重复实验, 取平均值分析; 消融实验结果单独分析, 以验证核心模块的有效性。

4.2.1 场景一实验结果与分析

由表 1 可知, RL-Game 方法在排产周期、资源利用率、总成本及工序衔接效率方面均优于对比算法, 较 GA 算法平均优化率分别高达 18.5%、25.6%、16.2% 和 18.2%。原因在于: ①区别于 GA、PSO、Game-GA 的静态优化逻辑, RL-Game 的动态博弈决策框架具备实时扰动感知与自适应调整能力, 能捕捉生产中的动态扰动事件, 通过实时反馈和更新策略保障排产连续性与稳定性, 解决传统方法不匹配动态生产环境的问题; ②三方主从—合作博弈模型可实现多主体利益全局均衡, 资源调度中心统筹分配资源, 平衡各生产单元工单负载, 避免极端情况; 生产单元目标函数绑定加工收益与资源利用率, 形成正向激励机制, 提升全局资源利用效率; ③基于 DQN 算法的强化学习模块将工序衔接优化转化为序贯决策问题, 通过经验回放与目标网络机制学习最优衔接策略, 减少工序等待时间与切换成本, 契合汽车零部件制造多工序协同需求, 实现了排产性能的综合优化。

Table 1 Experimental results of scenario 1
表 1 场景一实验结果

排产方法	排产周期 /生产日	资源利用 率/%	总成本 /万元	工序衔接 效率/%
GA	16.8±0.5	68.3±2.1	89.6±3.2	75.2±1.8
PSO	15.2±0.4	72.5±1.9	82.3±2.8	79.6±1.5
DQN	14.8±0.3	73.3±1.5	80.1±2.3	79.4±1.2
Game-GA	15.3±0.4	74.2±1.6	82.6±2.5	79.1±1.3
RL-Game	13.7±0.3	85.8±1.2	75.1±2.1	88.9±1.1
RL-Game 较 GA 优化率/%	18.5	25.6	16.2	18.2
RL-Game 较 PSO 优化率/%	9.9	18.3	8.7	11.7
RL-Game 较 DQN 优化率/%	7.2	12.5	6.3	9.5
RL-Game 较 Game-GA 优化率/%	10.5	14.2	9.2	10.8

4.2.2 场景二实验结果与分析

如表 2 所示, RL-Game 方法在排产周期、资源利用率、

总成本、订单按期交付率方面均优于对比算法,较GA算法的平均优化率分别高达21.2%、27.9%、17.7%和21.4%,说明RL-Game方法在不同制造场景下均具备优异的排产性能,是解决多工单协同排产中资源冲突、多主体目标异构、动态扰动适应性差三大核心问题的有效路径,且具备良好的场景适配性,可推广至各类多品种、小批量制造场景。

Table 2 Experimental results of scenario 2
表 2 场景二实验结果

排产方法	排产周期 /生产日	资源利用 率/%	总成本 /万元	订单按期 交付率/%
GA	11.3±0.4	65.7±2.3	68.9±2.5	78.4±2.2
PSO	10.1±0.3	70.2±2.0	63.5±2.1	83.6±1.8
DQN	9.7±0.2	70.3±1.4	61.0±1.8	85.0±1.1
Game-GA	10.0±0.3	69.6±1.5	62.9±2.0	83.7±1.2
RL-Game	8.9±0.2	84.1±1.3	56.7±1.7	95.2±1.0
RL-Game较GA优化率/%	21.2	27.9	17.7	21.4
RL-Game较PSO优化率/%	11.9	19.8	10.7	13.9
RL-Game较DQN优化率/%	8.5	13.8	7.1	10.2
RL-Game较Game-GA优化率/%	11.2	15.5	9.8	11.5

4.2.3 消融实验结果与分析

消融实验采用场景一的参数设置,聚焦三方博弈模型与DQN强化学习的协同有效性,实验结果取10次重复实验的平均值,结果如表3所示。可以看出,仅使用DQN单算法时,各项指标优化效果最差,因缺乏多主体利益协调机制,只能实现全局指标单一优化,无法平衡局部利益,证明博弈模型是多主体利益均衡的核心支撑。博弈模型结合传统GA算法的优化效果优于DQN单算法,但远不及所提方法,因GA静态优化逻辑无法应对生产动态扰动,验证了强化学习对博弈策略动态优化的必要性。双主体博弈模型保留DQN模块,但因缺失工单管理者主体,资源利用率与总成本优化率相较所提方法分别下降6.5%、4.5%,说明三方博弈模型的完整性是全局优化关键。RL-Game的多主体利益均衡度达92.3%,显著高于其他消融方案,充分证明了三方主从一合作博弈模型与DQN强化学习的深度融合是实现多主体利益均衡与动态排产优化统一的关键。

Table 3 Ablation experiment results
表 3 消融实验结果

排产方法	排产周期 /生产日	资源利用 率/%	总成本 /万元	多主体利益 均衡度/%
消融1(仅DQN单算法)	14.8±0.3	73.3±1.5	80.1±2.3	65.7
消融2(博弈+GA)	15.3±0.4	74.2±1.6	82.6±2.5	81.5
消融3(双主体博弈+DQN)	15.6±0.4	78.2±1.4	83.5±2.4	78.2
RL-Game(原方法)	13.7±0.3	85.8±1.2	75.1±2.1	92.3

4.2.4 与先进强化学习/博弈方法的比较

为进一步验证所提方法的优越性,新增2类当前领域的先进方法作为对比:①基于近端策略优化(Proximal Policy Optimization, PPO)的多主体强化学习(主流多智能体调度算法);②双主体博弈+PPO(经典多主体博弈求解方法),对比结果取两类场景平均值,结果见表4。可以看出,RL-Game在各项指标上均显著优于当前先进方法,具体来说:①相较于PPO多主体强化学习,RL-Game通过三方博弈模型解决了多主体目标异构问题,资源利用率提升11.05%,多主体利益协调能力更优;②相较于双主体博弈+PPO,RL-Game的三方博弈模型更贴合实际生产的主体构成,且DQN算法在序贯决策中的策略优化效率更高,排产周期缩短13.9%;③RL-Game的动态扰动适应度达94.5%,远高于对比方法,因其融合框架实现了博弈策略的实时自适应更新,而非先进方法中“博弈+强化学习”的简单组合。

Table 4 Comparison results with advanced methods
表 4 与先进方法的比较结果

排产方法	排产周期 /生产日	资源利用 率/%	总成本 /万元	动态扰动 适应度/%
PPO多主体强化学习	11.3±0.3	76.5±1.4	65.8±2.0	78.3
双主体博弈+PPO	10.8±0.2	78.2±1.3	63.5±1.8	82.6
RL-Game	9.9±0.2	83.5±1.1	58.2±1.5	94.5
RL-Game较PPO优化率/%	12.4	9.15	11.55	20.69
RL-Game较博弈+PPO优化率/%	8.33	6.78	8.35	14.41

4.2.5 鲁棒性测试结果

在两类场景上进行鲁棒性测试,分别设置10%、15%、20%、25%、30%五档动态扰动概率,运用变异系数($CV = \frac{\sigma}{\mu}$,其中 σ 为标准差, μ 为平均值)对所提方法的稳定性进行度量,变异系数越小表示鲁棒性越优。场景一与场景二的鲁棒性测试结果分别见表5、表6。可以看出,当扰动概率逐步提升时,3类方法的CV均呈现上升趋势,且RL-Game的CV始终显著低于GA和PSO,上升幅度更小,鲁棒性更优。具体来说,在场景一中,当扰动概率达到30%时,RL-Game的平均CV仅为0.045,而GA为0.139,PSO为0.116;在场景二中,RL-Game的平均CV为0.047,GA为0.147,PSO为0.123。RL-Game的鲁棒性优势来源于DQN算法强大的动态学习能力,这种能力使系统在运行过程中能实时感知环境中的扰动强度变化,并通过高效的自适应机制快速调整策略,从而有效抵消外界扰动对系统性能的影响。相比之下,GA和PSO采用基于固定规则的静态优化逻辑,缺乏应对动态变化的灵活性,因此在面对复杂多变的环境时难以及时调整自身行为模式,导致其整体性能容易出现剧烈波动。

Table 5 Robustness test results of scenario 1					
表 5 场景一鲁棒性测试结果					
扰动概率	排产方法	排产周期	资源利用率	总成本 CV	平均 CV
		CV	CV		
10%	GA	0.032	0.038	0.041	0.037
	PSO	0.028	0.031	0.035	0.031
	RL-Game	0.015	0.018	0.020	0.018
15%	GA	0.045	0.052	0.058	0.052
	PSO	0.039	0.043	0.049	0.044
	RL-Game	0.021	0.024	0.026	0.024
20%	GA	0.068	0.075	0.082	0.075
	PSO	0.057	0.062	0.069	0.063
	RL-Game	0.028	0.031	0.033	0.031
25%	GA	0.092	0.105	0.113	0.103
	PSO	0.078	0.086	0.094	0.086
	RL-Game	0.035	0.038	0.041	0.038
30%	GA	0.125	0.142	0.151	0.139
	PSO	0.103	0.118	0.127	0.116
	RL-Game	0.042	0.045	0.048	0.045

Table 6 Robustness test results of scenario 2					
表 6 场景二鲁棒性测试结果					
扰动概率	排产方法	排产周期	资源利用率	总成本 CV	平均 CV
		CV	CV		
10%	GA	0.035	0.041	0.044	0.040
	PSO	0.030	0.034	0.038	0.034
	RL-Game	0.016	0.019	0.022	0.019
15%	GA	0.048	0.055	0.061	0.055
	PSO	0.041	0.046	0.052	0.046
	RL-Game	0.022	0.025	0.028	0.025
20%	GA	0.072	0.080	0.087	0.080
	PSO	0.061	0.067	0.073	0.067
	RL-Game	0.029	0.032	0.035	0.032
25%	GA	0.098	0.112	0.120	0.110
	PSO	0.083	0.091	0.099	0.091
	RL-Game	0.037	0.040	0.043	0.040
30%	GA	0.132	0.150	0.159	0.147
	PSO	0.109	0.125	0.134	0.123
	RL-Game	0.044	0.047	0.050	0.047

5 结语

针对智能制造多品种、小批量模式下多工单协同排产存在的资源冲突、多主体目标异构、动态扰动适应性差问题,本文提出融合强化学习与博弈优化的协同排产方法。经理论建模、算法设计与实验验证,本文构建的资源调度中心—生产单元—工单管理者三方主从—合作博弈模型可精准刻画主体利益交互,以正向激励绑定局部与全局利益,使多主体利益均衡度达92.3%,解决了传统集中式排产利益平衡痛点;其融合框架将博弈策略优化转为马尔可夫决策过程,采用DQN算法实现策略自适应迭代优化,突破了传统博弈模型局限,适配动态扰动场景。在两类典型场景上的实验结果表明,RL-Game方法较传统算法核心指标优化率达8.7%~27.9%,较先进方法优化率达8.63%~20.7%;鲁棒性测试结果证实了其具有较强的动态适应性。

本研究虽然实现了多工单协同排产优化,但仍存在场景假设简化、算法适配性有限、博弈机制未充分贴合实际等局限。未来将从以下两个方面深化研究:①拓展场景假设,引入随机过程理论,适配加工时间随机、工艺路线可变的复杂场景;②优化算法性能,采用PPO、SAC等先进强化学习算法结合迁移学习,提升大规模场景下的求解效率。

参考文献:

[1] LIU J J, QIU J H, ZHANG Z M. Human-computer interactive scheduling of complex manufacturing systems and case study[J]. Journal of Mechanical Engineering, 2025, 61(15):351-371.
刘建军,邱俊豪,张忠明. 复杂制造系统人机交互调度与案例研究[J]. 机械工程学报,2025,61(15):351-371.

[2] SONG T X, LIU H Z. Scheduling optimization of pharmaceutical workshop considering equipment setup time and production balance rate[J]. Journal of Machine Design, 2024, 41(10):93-100.
宋庭新,刘慧中. 考虑设备准备时间和生产平衡率的制药车间排产优化[J]. 机械设计,2024,41(10):93-100.

[3] DONG S K, LI C, YANG G, et al. Review on solution and application of mixed game problems[J]. Journal of Software, 2025, 36(1):107-151.
董绍康,李超,杨光,等. 混合博弈问题的求解与应用综述[J]. 软件学报,2025,36(1):107-151.

[4] ZHANG D, YUAN Q, MENG L, et al. Reinforcement learning for single-agent to multi-agent systems: from basic theory to industrial application progress, a survey[J]. Artificial Intelligence Review, 2025, 59(2):46.

[5] FATHOLLAHI A. Intelligent control framework for improving energy system stability through deep learning-based modal optimization scheme[J]. Sustainability, 2024, 16(21):9392.

[6] JU L, LV S, LI Y, et al. Two-stage scheduling optimization model and benefit allocation strategy for virtual power plant clusters aggregated by multi-dimensional information indicators [J]. Renewable Energy, 2025, 240: 122245.

[7] KOU G, YÜKSEL S, DINÇER H, et al. Queuing-theory-driven renewable energy investment decisions through artificial expert choices with dynamic hybrid multi-facet fuzzy modeling[J]. iScience, 2025, 28(11):113711.

[8] DONG W, LI D, JI Y, et al. Towards a next-generation LLM empowered low-code programming industrial robotic system for human-centric smart manufacturing[J]. Journal of Manufacturing Systems, 2025, 83:675-686.

[9] DU S Z, LI Q J, SONG N H, et al. Solving the permutation flow shop scheduling problem with sequence-dependent setup time via iterative greedy algorithm and imitation learning[J]. Mathematics and Computers in Simulation, 2025, 234: 169-193.

[10] YANG B Y, YANG Y, LI Y Z, et al. UAV detection method for polarimetric radar based on heuristic search[J]. Journal of Signal Processing, 2025, 41(5):867-885.
杨博宇,杨勇,李永祯,等. 基于启发式搜索的极化雷达无人机检测方法[J]. 信号处理,2025,41(5):867-885.

[11] WU Z J, XI S L, MENG R H, et al. Research on batch scheduling problem of parallel machines considering multi-resource constraints[J]. Journal of China Three Gorges University (Natural Sciences), 2024, 46(6): 104-112.

- 吴正佳,席帅乐,孟荣华,等. 考虑多资源约束的并行机分批调度问题研究[J]. 三峡大学学报(自然科学版),2024,46(6):104-112.
- [12] QASIM M,SAJID M,LAPINA M, et al. COBGA: MCDM-assisted improved genetic algorithm for scheduling Industrial-Internet-of-Things jobs in cloud computing[J]. Cluster Computing,2025,28(15):944.
- [13] XU Y,ZHANG M,WANG D, et al. Hybrid Gaussian quantum particle swarm optimization and adaptive genetic algorithm for flexible job-shop scheduling problem [J]. Engineering Applications of Artificial Intelligence,2025,154:110882.
- [14] THEPPHAKORN T,PONGCHAROEN P,VITAYASAK S. Multi-objective hybrid optimizations for designing course schedules based on operating costs and resource utilization [J]. Annals of Operations Research, 2024(prepublish):1-68.
- [15] MIGLIAVACCA G. Multi-energy static modeling approaches: a critical overview[J]. Energies,2025,18(7):1826.
- [16] LU X Y, HAN X L. Research on dynamic energy-saving scheduling of flexible job shop based on deep reinforcement learning [J]. Operations Research and Management Science, 2025, 34(8):99-104.
- 陆心屹,韩晓龙. 基于深度强化学习的柔性作业车间动态节能调度问题研究[J]. 运筹与管理,2025,34(8):99-104.
- [17] HU J, WANG J Y, YIN L J, et al. Benefit equilibrium decision-making of new energy vehicle enterprises empowered by digital technology [J/OL]. Operations Research and Management Science, 2026-02-03. <https://link.cnki.net/urlid/34.1133.g3.20250915.1411.002>.
- 胡姜,王佳月,殷旅江,等. 数字技术赋能下新能源汽车企业利益均衡决策[J/OL]. 运筹与管理, 2026-02-03. <https://link.cnki.net/urlid/34.1133.g3.20250915.1411.002>.
- [18] WU N, WU Q, REN H, et al. Hierarchical game optimization of independent shared energy storage system considering multiple profits and penalty[J]. Energy,2025,321:135448.
- [19] TIAN C, MEI J, TIAN K, et al. Distributed adaptive Nash equilibrium seeking for games of heterogeneous high-order players over a directed graph[J]. Automatica,2025,179:112296.
- [20] LIU J W, YAO Y X, HUA M, et al. Distributed Nash equilibrium seeking for multi-coalition games with constrained NASVs: a hierarchical prescribed-performance algorithm[J]. Neurocomputing,2025,650:130821.
- [21] DU F J, LI J, WANG L, et al. MEF: a cloud computing resource scheduling method based on online fault tolerance[J]. Computer Applications and Software, 2024, 41(3):336-344.
- 都繁杰,李静,王亮,等. MEF:基于在线容错的云计算资源调度方法[J]. 计算机应用与软件,2024,41(3):336-344.
- [22] WEI J H, SHANG G G, LIU X F, et al. Robust model and solution of two-stage four-layer stackelberg-cooperative game for multiple prosumers[J]. Electric Power Automation Equipment, 2025, 45(1):208-216.
- 魏金花,尚广广,刘雄飞,等. 多产消者的两阶段四层主从—合作博弈鲁棒模型及求解[J]. 电力自动化设备,2025,45(1):208-216.
- [23] JIANG H,ZHANG T,XI Y, et al. Intelligent multiple parameters optimization methods integrating hydraulic model and SIAs with various constraints for extended reach drilling[J]. Geoenergy Science and Engineering,2025,251:213846.
- [24] VENTURA F, JAMES J, CARISTI G, et al. From cooperative play to multi-critical decision: the Shapley index and the Choquet integral in comparison[J]. Energy Economics,2025,150:108816.
- [25] XIAO Q,PAN T,WANG K, et al. Research on AGV path planning based on improved DQN algorithm[J]. Sensors,2025,25(15):4685.

(责任编辑:尹晨茹)