中国科学: 信息科学 2024 年 第 54 卷 第 6 期: 1441–1457
SCIENTIA SINICA Informationis
论文

基于深度强化学习和知识迁移的飞机装配脉动生产线调度方法

钟金成[1,2†], 马浩宇[1,2†], 龙明盛[1,2*], 王建民[1,2*]

1. 清华大学软件学院, 北京 100084
2. 北京信息科学与技术国家研究中心, 北京 100084
* 通信作者. E-mail: mingsheng@tsinghua.edu.cn, jimwang@tsinghua.edu.cn
† 同等贡献

收稿日期: 2023–06–30; 修回日期: 2023–09–18; 接受日期: 2023–10–02; 网络出版日期: 2024–06–12

科技创新 2030—“新一代人工智能” 重大项目 (批准号: 2020AAA0109201)、国家自然科学基金 (批准号: 62021002, 62022050) 和北京市科技新星计划 (批准号: Z201100006820041) 资助项目

摘要 飞机装配是飞机制造中的关键环节, 如何对飞机装配脉动生产线进行合理调度, 实现降本增效, 是智能制造领域的重要科学问题. 然而, 飞机装配脉动生产线场景复杂, 装配单架飞机就包含上万道工序, 这为飞机装配调度问题的形式化建模和高效求解带来新的挑战, 因而当前生产实践中主要依靠人类专家经验进行手工调度. 本文聚焦降低人力负载的优化目标, 提出两种领域特定的技术以解决飞机装配调度问题. 首先, 将飞机装配脉动生产线调度问题建模为两个马尔可夫 (Markov) 决策过程, 通过双重强化学习智能体决策生成飞机装配的近似调度方案. 其次, 针对强化学习决策鲁棒性不足的缺陷, 提出领域知识迁移方法, 将强化学习的求解知识迁移到整数规划约束剪枝中, 最后利用整数规划求解器优化得到综合性能优异的调度方案. 在飞机装配生产线的真实数据上完成了实验验证, 结果表明本文提出的基于深度强化学习和知识迁移的调度方法能够成功扩展到年产量近百架次的飞机装配脉动生产线调度问题, 将组合优化方法难以求解的问题优化到分钟级求解, 相较于基线方法取得显著性能优势.

关键词 飞机装配, 智能调度, 组合优化, 强化学习, 知识迁移

1 引言

随着飞机制造工业的快速发展和市场竞争的日益激烈, 飞机制造工厂迫切需要优化产线效率, 降低生产成本, 保质保量地完成加工生产任务. 飞机装配是飞机制造中的关键环节, 对飞机装配生产流程进行调度和优化是产线优化的重要方向. 目前, 飞机装配脉动生产线 [1~3] 的临场调度决策主要依赖

引用格式: 钟金成, 马浩宇, 龙明盛, 等. 基于深度强化学习和知识迁移的飞机装配脉动生产线调度方法. 中国科学: 信息科学, 2024, 54: 1441–1457, doi: 10.1360/SSI-2023-0197
Zhong J C, Ma H Y, Long M S, et al. Scheduling approach for aircraft assembly pulsation production lines with deep reinforcement learning and knowledge transfer (in Chinese). Sci Sin Inform, 2024, 54: 1441–1457, doi: 10.1360/SSI-2023-0197

现场调度专家的经验和努力. 因此, 如何高效地对飞机装配脉动生产线进行合理的工序调度, 实现降本增效, 是智能制造领域中重要而紧迫的科学问题. 飞机装配调度的核心决策问题是确定各个装配工序的起始时间和所在的装配型架 (fixture), 旨在确保脉动生产线在满足拓扑约束、型架约束和物料约束等限制的前提下, 按照稳定的生产节拍 (takt time) 完成生产计划. 在脉动生产线中, 工期是固定的, 降低资源负载 (resource load) 是本文考虑的主要优化目标.

为了解决飞机装配调度问题, 本文将飞机装配脉动生产线的调度场景形式化建模为一个整数规划 (integer linear programming, ILP) 模型. 该模型用一组约束条件描述装配型架、生产节拍、装配工作空间等领域特定的需求, 以降低人力负载为优化目标. 可以使用现有的整数规划求解器软件 (例如 Gurobi[1)], CPLEX[2)] 和 OR-Tools[3)]等) 进行优化求解, 找到满足各项约束并且人力负载最低的调度方案. 但是求解整数规划是经典的 NP- 难问题 [4], 存在严重的扩展性缺陷. 单架飞机的装配工序已经达到数万量级, 不同飞机之间存在工序差异. 随着年产飞机数量的增加, 变量数量飞速膨胀, 朴素的整数规划方法已经很难产生可行的调度方案.

本文提出了一种基于深度强化学习和知识迁移 (knowledge transfer) 的调度方法来解决飞机装配调度问题. 相较于传统手工设计的启发式算法, 强化学习具备两大优势: 一是可以通过在模拟数据环境中进行探索和利用, 自动地学习归纳问题的结构特点, 减少对专家经验的要求; 二是调度问题的全局优化目标需要在调度方案完成后才能验证计算, 相较于传统手工设计的启发式算法只能考虑局部最优结构, 强化学习方法可以对全局目标进行优化. 强化学习方法的潜力已经在很多组合优化问题上被证明有效, 如旅行商问题 [5, 6]、最大独立集 [7, 8] 等. 但应用强化学习解决飞机装配调度问题仍然面临两个难题.

第一个难题是飞机装配脉动生产线的工序调度问题难以直接建模为马尔可夫决策过程 (Markov decision process, MDP). 本文创新性地提出将飞机装配脉动生产线的调度问题转化为两个子任务: 一是根据总体调度信息预测人力负载; 二是在预测的人员负载下, 生成符合约束的具体方案. 在此基础上, 本文提出一种双重强化学习框架 (bi-level reinforcement learning framework), 两个决策智能体分工协作, 分别完成两个子任务.

第二个难题是强化学习缺乏求解鲁棒性的保证. 考虑到待求解的调度问题是 NP- 难的, 无法保证强化学习输出的调度方案的质量. 本文提出一种两阶段知识迁移的方法以缓解这一难题. 即先使用深度强化学习对调度问题进行预求解, 生成初步调度方案; 再通过对初步方案中的知识进行迁移, 构建剪枝规则, 使用整数规划求解器进一步优化. 这一方案既可以提升求解质量, 也能够微调初始方案中不符合约束条件的工序安排, 确保最后输出方案满足所有约束条件. 更进一步, 本文还结合飞机装配脉动生产线的特点, 提出了工序节拍稳定假设, 在第一阶段中假设所有飞机工序均相同, 降低强化学习的训练难度, 在后续知识迁移和微调过程中, 再进一步考虑工序变动等复杂细节. 在实验中证明引入知识迁移技术是必要的, 能够显著提高求解效率.

综上所述, 本文主要贡献如下:

• 本文以降低人力负载为优化目标, 面向真实飞机装配脉动生产线的调度需求建立了整数规划模型. 通过对调度问题中求解最优工期的经典贪心框架进行分析, 将真实需求中的人力负载优化问题转化为最优负载预测和工序调度两个子问题, 提出了对应的双重强化学习优化算法, 用两个相互关联的决策智能体分别预测人力负载和在预测的人力负载下进行工序调度, 可以高效生成较高质量的调度

1) Gurobi Optimizer. https://www.gurobi.com.
2) CPLEX Optimizer. https://www.ibm.com/analytics/cplex-optimizer.
3) Google OR-Tools. https://github.com/google/or-tools.

方案.

• 为缓解强化学习方法缺乏求解鲁棒性的难题, 本文提出了一种两阶段的知识迁移求解方法, 可以将强化学习探索归纳的知识迁移到飞机脉动生产线多变的现场调度环境中, 对整数规划模型进行剪枝, 用现有组合优化求解器进行精细优化. 知识迁移方法充分利用强化学习方案中的求解知识, 为本文提出的调度方案增加了对场景变动的自适应感知, 大幅提升了运算效率.

• 本文实现了一个基于深度强化学习和知识迁移的飞机装配调度问题的求解方案, 并在飞机装配脉动生产线的真实数据上进行了实验验证. 结果证明了本文提出的方法能够成功扩展到年产量近百架次的大规模飞机装配脉动生产线调度场景, 生成满足全部约束条件并且人力负载指标优越的调度方案. 本文提出的方法不但求解效率高, 相较于对应的基线算法可以获得超过 7% 的负载降低.

2 飞机装配脉动生产线调度问题

飞机装配是将飞机的各个生产部段进行连接组装成型, 是飞机制造工艺中的重要步骤. 大型飞机装配生产线具有工序繁多、生产要素复杂、工期紧张、自动化程度低等特点. 飞机装配调度问题是在保障装配任务能够按照生产计划完成的前提下, 优化资源利用率, 降低人力负载, 控制管理成本, 指导实际生产的关键科学问题.

2.1 飞机装配脉动生产线

脉动式生产线和生产节拍. 现代飞机装配技术通常采用脉动生产来保障生产计划的有序完成. 传统飞机装配生产线围绕组装中的飞机展开, 工人在固定的场所进行搭积木式作业, 俗称 “人动机不动”. 脉动生产是 “机动人不动”, 是一种围绕组装型架进行作业的模式, 组装中的飞机像脉搏跳动一样按固定时间周期在不同工位的型架之间流转, 形成一种特殊的流水线模式 [1~3]. 从总体视角看, 稳定运行的脉动生产线以一个固定的脉动频率完成飞机装配任务, 每个脉动周期都有一架飞机开始装配, 有一架飞机完成出厂, 这个固定的脉动间隔被称为生产节拍. 脉动生产线运作的关键, 在于维持稳定的生产节拍, 只要预先确定生产节拍, 并跟踪每个生产节拍周期内的装配任务的完成情况, 就能按时完成生产计划. 由脉动生产线的运作模式可知, 生产节拍由计划产量和计划工时决定, 生产节拍 T 是规划工时和生产目标之间的比值, 如式 (1) 所示:

$$T = \frac{\text{计划工时}}{\#(\text{计划产量})}. \tag{1}$$

例如, 如果计划生产时间是 300 天, 每日 12 小时, 当飞机订单为 30 架时, 生产节拍根据式 (1) 计算得到: 每 10 天产出一架飞机, 即可满足生产计划, 此时生产节拍为 10 天, 共计 120 小时.

生产节拍和型架的生产周期 (working horizon). 飞机装配时, 型架是基本的加工场所. 飞机各个部段的装配任务都需要在对应型架上完成. 飞机的一个部段上架某个型架后, 会在型架上完成该部段的全部工序, 之后进入对接型架, 等待流转至后续型架开始下一阶段装配. 型架上完成的部段都是一个相对完整的装配环节, 在实践中, 为了确保脉动生产线的总体生产节拍的稳定, 会进一步跟踪每一类型架上的实际进度. 对于某一类型架, 每次脉动都会有一架飞机的对应部段上架, 有一架飞机的对应部段完成装配转出型架. 由于每套型架只能被一架飞机占有, 为了实现多架飞机并行装配, 现场会根据任务负荷配备复数套同类型架. 型架的生产周期指一个飞机部段实际可以在对应型架上的停留时间, 当保持节拍稳定时, 每个型架的生产周期 H_f 是生产节拍 T 的倍数, 如式 (2) 所示, 其中 $\#(f)$

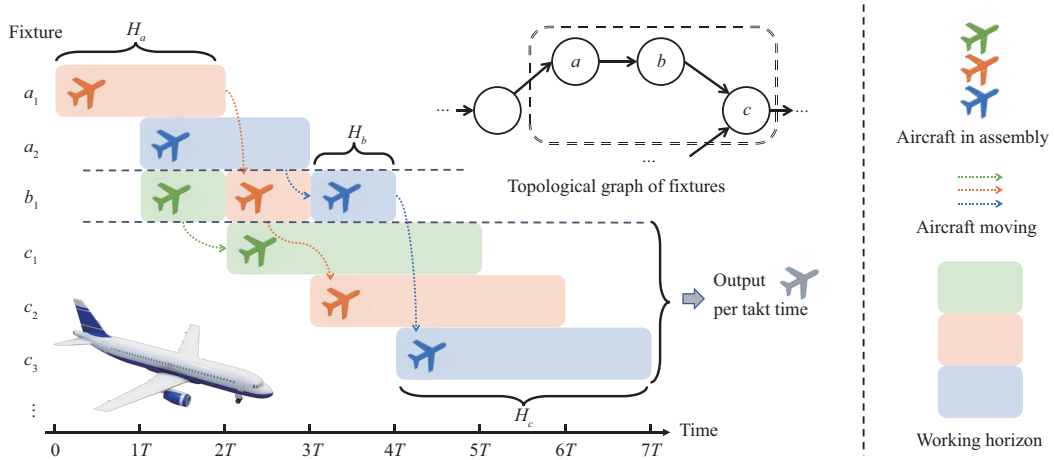

图 1 (网络版彩图) 飞机装配脉动生产线的型架和节拍模型

Figure 1 (Color online) Fixtures and takt-time model in aircraft assembly line

是对应型架 f 的配套数目.

$$H_f = T \times \#(f). \tag{2}$$

图 1 说明了型架的生产周期和生产节拍之间的关系. 飞机部段在型架 a, b, c 上顺序流转. 对于型架而言, 每个生产节拍都有一个飞机部段上架和转出. 型架 a 有 a_1, a_2 两套型架, 因此生产周期 H_a 为 $2T$, 类似地, 型架 c 的生产周期 H_c 为 $3T$. 每个型架上的任务都要在对应的生产周期内完成, 确保每个飞机部段都按照生产节拍完成装配, 从而保证飞机整体上的装配按照生产节拍进行.

2.2 飞机装配调度整数规划模型

本小节将详细阐述飞机装配脉动生产线的工序调度问题的各个约束细节, 并根据实际应用场景给出详细的整数规划建模.

整体而言, 飞机装配脉动生产线可以抽象为多个子图逐层叠加而成的有向无环图, 每一层子图都表示一架飞机, 每层子图中的节点表示这一架次飞机的装配工序, 图中的有向链路表示装配工序之间的先后序依赖关系, 子图各层之间受生产节拍和型架规则约束. 整个脉动装配生产线上还需要考虑工人、工作空间和型架之间的协同限制.

脉动生产线通过保持稳定的生产节拍和生产周期来保障生产活动按计划执行, 有利于资源调配和进度管理. 经典调度问题中的工期优化目标在脉动生产线的场景中被建模为时间约束. 整体优化目标是在满足节拍、资源等约束的前提下降低人员负载, 以达到降低工人需求, 控制资源成本, 提高换班容错率等目的. 下面首先给出本文数学描述中出现的关键符号含义.

- $\mathbb{J}$: 装配工序集合, 其中单个工序 $j \in \mathbb{J}$;
- $\mathbb{F}$: 型架集合, 其中单个型架 $f \in \mathbb{F}$;
- $\mathbb{K}$: 工种集合, 其中单个工种 $k \in \mathbb{K}$;
- $\mathbb{M}$: 作业地点集合, 其中每个作业地点 $m \in \mathbb{M}$;
- $\mathbb{J}_m$: 作业地点 m 所包含的工序集合, 其中 $\mathbb{J}_m \subset \mathbb{J}$;
- $\mathbb{A}$: 计划装配飞机集合, 其中 $a \in \mathbb{A}$;
- H_f: 型架 f 的生产周期限制;

- C_m: 作业地点 m 的装配空间限制;
- $t_j^{(a)}$: 飞机 a 的装配任务 j 的工作耗时;
- $n_{jk}^{(a)}$: 飞机 a 的装配任务 j 需要工种 k 的工人人数;
- $n_k^{\max}$: 同时工作的工种 k 的最大工人人数;
- $s_j^{(a)}$: 飞机 a 的装配任务 j 的起始时间;
- $s_f^{(a)}$: 飞机 a 在型架 f 上任务的起始时间;
- $P_j^{(a)}$: 飞机 a 的装配任务 j 的前置任务集合;
- $\mathrm{Loc}(j^{(a)}, f)$: 装配任务 $j^{(a)}$ 在型架 f 完成时为 1, 否则为 0;
- $\mathrm{NextTakt}(a)$: 装配飞机 a 后下一节拍对应的飞机.

优化目标. 本文优化目标是降低人员负载. 假设不同工种的工人不可互相替代, 该优化目标可表示为最小化同时工作的各类工人的最大人数; 同时, 传统生产线调度问题中常见的工期优化目标被建模为整数规划模型中的节拍和周期约束. 本文将飞机脉动生产线调度问题形式化建模为

$$\min \quad \sum_{k\in\mathbb{K}} n_k^{\max}, \tag{3}$$

$$\text{s.t.} \quad s_j^{(a)} + t_j^{(a)} \leqslant s_{j'}^{(a)}, \qquad \forall a \in \mathbb{A}, \forall j' \in \mathbb{J}, j \in P_{j'}^{(a)}, \tag{4}$$

$$s_f^{(a')} - T = s_f^{(a)}, \qquad \forall f \in \mathbb{F}, a' = \mathrm{NextTakt}(a), \tag{5}$$

$$s_j^{(a)} + t_j^{(a)} \leqslant s_f^{(a)} + H_f, \qquad \forall j^{(a)} \in \{j^{(a)} | \mathrm{Loc}(j^{(a)}, f)\}, \tag{6}$$

$$\sum_{j^{(a)\prime}\in\mathbb{J}_m} \mathbb{I}(s_j^{(a)} \leqslant s_{j'}^{(a)}) \mathbb{I}(s_{j'}^{(a)} < s_j^{(a)} + t_j^{(a)}) \cdot \sum_{k\in\mathbb{K}} n_{jk}^{(a)} \leqslant C_m, \qquad \forall m \in \mathbb{M}, j^{(a)} \in \mathbb{J}_m, \tag{7}$$

$$\sum_{j^{(a)\prime}\in\mathbb{J}} \mathbb{I}(s_j^{(a)} \leqslant s_{j'}^{(a)}) \mathbb{I}(s_{j'}^{(a)} < s_j^{(a)} + t_j^{(a)}) \cdot n_{jk}^{(a)} \leqslant n_k^{\max}, \qquad \forall k \in \mathbb{K}, j^{(a)} \in \mathbb{J}, \tag{8}$$

其中, 约束公式 (4)~(8) 的含义如下.

- 式 (4) 表示拓扑约束: 每个任务需要在其全部前置任务结束后才能开始执行. 先序依赖拓扑约束是调度问题的基本约束. 在实际应用中, 时间调度粒度是离散的, 例如在本文讨论的真实飞机装配案例中, 以半小时为一个单位工时.

- 式 (5) 表示型架节拍约束: 在脉动生产线中, 对于同一类型架, 每一个脉动节拍 T 都要稳定完成一架飞机的装配工作. 以当前飞机 a 的装配部段上架型架 f 的起始时间 $s_f^{(a)}$ 为锚点. 一个节拍单位时间 T 后, 相对应的下一架飞机 $a' = \mathrm{NextTakt}(a)$ 的相关部段也会上架同类型架 f. 两架相邻脉动节拍的飞机之间的时间间隔固定为一个节拍单位时间 T. 一个型架只能同时承载一架飞机, 当多架飞机需要同时装配时, 相邻飞机会分别在各自的型架上架装配. 飞机装配是围绕型架展开的, 只要保持飞机的各个部段之间以一个固定的节拍 T 在型架间流转, 就能在整体视角上保持稳定节拍 T. 飞机相关装配部段在型架内的具体装配工序的调度安排不作额外约束.

- 式 (6) 表示型架生产周期约束, 作为型架节拍约束的补充: 由于型架有实际的物理条件限制, 一个型架同时只能上架一架飞机, 为了保证飞机部段在型架上都能稳定保持生产节拍 T 执行装载和卸载操作, 要求每架飞机在型架上的停留时间有一个最长期限, 即型架节拍周期 H_f. 如前文讨论的式 (2) 所示, 型架节拍周期 H_f 由生产节拍和对应型架 f 的配套数目 $\#(f)$ 决定.

- 式 (7) 是装配工作空间限制: 工人在飞机部段上进行手工装配作业时, 受到空间和安全限制, 在同一个装配作业区域内同时承载的作业人数不能超过 C_m. $\mathbb{I}(\cdot)$ 是条件函数, 当输入的条件成立时返回

1, 否则返回 0. $\mathbb{I}(s_j^{(a)} \leqslant s_{j'}^{(a)})\mathbb{I}(s_{j'}^{(a)} < s_j^{(a)} + t_j^{(a)})$ 判断了 j 与 j' 是否在工作时间上存在交集, 存在则返回 1, 反之返回 0.

• 式 (8) 是最大同时工作人数的定义约束: $n_k^{\max}$ 表示种类 k 的工人的最多同时工作人数, 也是优化目标的组成部分. 人数限制与式 (7) 所表示的空间限制形式相似, 但是需要区别的是, 在实际的数据中工作空间 m 与型架绑定, 不同装配架次飞机 a 和 a' 在不同的型架上进行装配, 在工作空间上互不干扰, 而工人资源由所有工序全局共享, 在式 (8) 中从 j 和 j' 的枚举范围差异中体现.

上述整数规划模型根据真实的飞机脉动生产线得到, 也是本文方法和实验的基础. 在实践中还可能存在其他应用上的限制条件 (如物料限制、装备套件等), 也可以较为容易地整合进问题建模中, 本文不再额外讨论.

扩展性难题. 通过将飞机装配脉动生产线调度问题建模为整数规划问题, 可以使用现有的整数规划求解器软件如 Gurobi, CPLEX 和 OR-Tools 进行求解. 但是整数规划问题是 NP- 难的, 对求解规模增长的可扩展性差 [4]. 例如, 虽然使用现有软件求解包含 1 架飞机, 2 个型架, 总计 400 个装配任务的调度问题时, 可以在数分钟内完成, 但整数规划搜索方法的计算复杂度按照待求变量数的指数函数膨胀, 一架飞机的装配工序通常数以万计, 且随着先进飞机型号的研制, 工序规模还会持续增加. 求解年度脉动生产线调度问题时, 还需要考虑近百架飞机的协同调度. 在真实飞机装配脉动生产线的历史调度数据中, 待调度工序就有数十万, 现有求解器仅仅是找到一个可行调度方案就需要几天甚至几周. 更进一步地, 由于人工作业现场经常存在临时变动, 求解相应调度问题需要更强的时效性和适应性. 因此飞机装配脉动生产线的工序调度仍依赖于调度员的经验进行人工调度.

2.3 深度强化学习求解飞机装配调度问题

本文提出使用深度强化学习来构建数据驱动的启发式规则, 缓解随求解规模增长带来的扩展性难题. 强化学习是一种通过与环境交互学习如何进行序列决策的范式. 强化学习智能体通过观察环境, 做出一系列决策, 利用环境反馈优化自身, 最大化智能体从环境获得的奖励. 强化学习求解组合优化问题时能够针对全局目标, 自动地在问题验证环境中进行探索利用, 学习问题搜索空间中潜在的最优子结构, 而人类领域专家将局部决策和全局目标建立关系是非常困难的. 强化学习的一大优势是可以自动探索问题的解空间, 不像监督学习和模仿学习那样需要 “问题 – 解” 的标记数据. 然而, 将强化学习应用到飞机装配脉动生产线的调度问题时仍面临 3 个挑战.

挑战 1: 对飞机装配脉动生产线的调度问题进行 MDP 建模. 强化学习求解组合优化问题的场景中, 建立 MDP 的常见方法是依托经典贪心框架 (classic greedy framework) [5, 6], 强化学习智能体动态地选出优先级 (priority) 最高的任务, 由贪心框架确定任务的起始时间. 图 2(a) 展示了经典贪心框架的局限. 经典贪心方法基于局部工期最优, 尽可能提前任务起始时间, 在早期会有大量任务同时工作, 导致负载失衡, 与全局优化目标相悖. 本文创新性地提出一种双重强化学习框架 (bi-level reinforcement learning framework), 如图 2(b) 所示. 双重强化学习框架将脉动生产线调度优化问题拆分成预测人力负载峰值和工期优化调度两个子任务: 资源预测智能体 π_r (resource agent) 先预测人力负载, 在预测出的人力负载限制下再使用工序调度智能体 π_p (scheduling agent) 生成完整的调度方案.

挑战 2: 调度问题状态特征动态建模. 智能体在工序调度过程中会根据各个时刻的实时调度情况, 动态评估各个工序的优先级, 因此需要解决如何对调度问题的求解信息进行动态特征建模的问题. 所有求解调度问题所需的信息需要被编码成特征向量, 输入给神经网络生成调度动作, 有效的状态特征建模能够明显加速学习进程. 本文将每个任务工序视作节点, 节点之间的约束关系是有向边, 整个调度问题构成一个拓扑图 G, 其结构和节点信息随调度问题的求解进度而改变. 为了更好地建模调度问

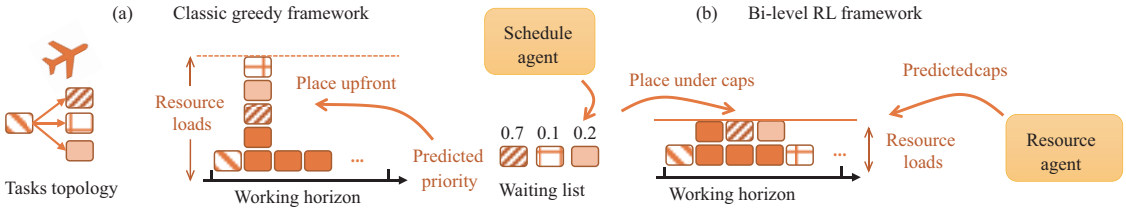

图 2 (网络版彩图) 双重强化学习框架与经典贪心框架的比较. 在没有显式资源限制时, 经典贪心框架会倾向于将任务尽早开始以满足工期要求, 导致资源负载失衡; 双重强化学习方法在经典贪心框架之外引入预测的资源约束, 能够得到优化资源负载后的调度方案

Figure 2 (Color online) Comparison between (a) classic greedy framework and (b) bi-level reinforcement learning framework. In the absence of explicit resource constraints, the classical greedy framework tends to start tasks as early as possible to meet project deadlines, resulting in an imbalance in resource loads. The bi-level reinforcement learning approach introduces predicted resource constraints beyond the classical greedy framework, enabling the generation of scheduling solutions with optimized resource loads

题, 本文引入图注意力网络 (graph attention networks, GATs) [9] 进行调度问题的表示学习. 图注意力网络是一类专门为处理图类型数据而设计的神经网络, 可以接受不同大小的图作为输入, 为图上的节点、边或者整张图生成对应的特征向量. 相较于强化学习常用的循环神经网络 [6, 10], 图注意力网络消除了对节点顺序的依赖, 引入了注意力机制自适应学习工序之间的关系, 便于捕捉不同工序之间的相关性.

挑战 3: 强化学习求解鲁棒性的保证. 基于强化学习的方法直接预测出完整调度方案是十分高效的. 但是考虑到飞机装配调度问题本质是 NP- 难问题, 同时模型还受到自身泛化性的限制, 样本分布外泛化问题一直都是深度学习方法的重要难题之一 [5, 11], 因此无法保证强化学习输出的调度方案的质量. 本文提出一种知识迁移方法以解决这一难题. 先使用深度强化学习对调度问题进行预求解, 输出中间方案; 再通过对中间方案中的调度知识进行迁移, 构建剪枝规则, 使用整数规划求解器进行最终优化. 这一方案将强化学习学到的调度知识迁移到整数规划剪枝规则中, 同时结合了强化学习快速推理的优势, 求解剪枝后的整数规划可以提升求解质量, 也能够微调强化学习方案中不符合约束条件的工序安排, 确保最后输出方案满足所有约束条件, 具备鲁棒性保证. 另外, 当一线生产出现工序意外变动等问题时, 也可以将已有调度方案中的知识迁移到新的剪枝规则构建中, 让整数规划求解器在已有调度方案上进行微调, 以达到现场调度的高效性和鲁棒性.

3 基于深度强化学习和知识迁移的调度问题求解方法

3.1 完整的调度工作流程

图 3 展示了本文提出的基于深度强化学习和知识迁移的调度方法的完整工作流程. 调度方法接受的输入包括: 计划产量或生产节拍、各个工序的加工信息、工序之间的拓扑关系图、整数规划模型、突发工序变动等信息. 强化学习智能体根据各个工序的加工信息和拓扑关系作出调度决策, 生成调度方案. 调度方案评估器作为外部环境, 接收强化学习智能体给出的调度方案, 检查是否满足工期、资源和节拍等整数规划模型中的约束条件, 评估计算出当前调度策略所对应的优化目标值. 强化学习奖励函数基于约束条件的满足与否和当前优化目标值, 为智能体提供反馈信号.

本文提出的完整调度方法可被分为两个阶段: 在第一阶段由强化学习直接输出初始的调度方案; 第二阶段将强化输出的调度知识迁移到整数规划剪枝中, 用现有整数规划求解器在强化学习生成的初

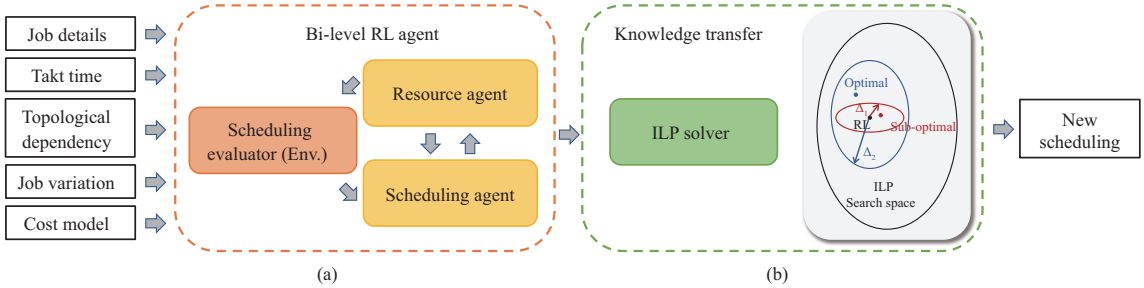

图 3 (网络版彩图) 本文方案的完整工作流: 整个求解工作流划分为双重强化学习求解和知识迁移两个阶段

Figure 3 (Color online) The complete workflow of this approach in the paper can be divided into two main stages: bi-level reinforcement learning and knowledge transfer

始调度方案附近进行微调. 在发生突发事件而需要重新调度规划时, 可以重新执行完整的调度工作流, 也可以直接将已有调度方案中的知识迁移到整数规划剪枝中, 用现有整数规划求解器微调已有方案, 快速获得适应新场景的调度方案.

3.2 双重强化学习框架

为了建立求解脉动生产线调度问题的 MDP, 本文提出了一种双重强化学习框架, 如图 2 所示, 将脉动生产线调度问题拆分成预测资源限制和工期优化调度两个相互协同的子问题. 资源预测智能体首先预测出各种资源的使用上限 (resource cap), 工序调度智能体在相应的资源限制下生成具体的调度方案. 本节将以人力资源为例, 详细阐述双重强化学习智能体的技术细节.

资源预测智能体. 资源预测智能体 π_r 根据输入的全局问题信息, 直接预测出当前工序调度智能体策略 π_p 下的最低人力负载. 其中可观测的状态包括所有可以提供的调度问题信息, 如生产节拍、工序信息和拓扑关系等. 单步动作输出各个不同工种对应的人力负载峰值的概率分布. 具体地, 资源预测智能体根据调度问题输入图 G, 输出高斯 (Gauss) 分布的均值 $\pi_r(\bar{\boldsymbol{n}}^{\max}|G) \in \mathbb{R}^{|\mathbb{K}|}$, 维度和人力资源种类数量一致. 预测的人力负载峰值单步动作从高斯分布 $\mathcal{N}(\pi_r(\bar{\boldsymbol{n}}^{\max}|G), 1)$ 上采样获得. 外部环境则将预测出的人力负载峰值作为限制约束, 由工序调度智能体生成可行的调度方案. 资源预测智能体奖励函数和所优化调度问题的优化目标保持一致, 是输出的调度方案所需的人数资源总数的相反数 $-\sum_k n_k^{\max}$. 当工序调度智能体无法在当前人力负载下找到符合节拍周期约束的解时, 会在奖励信号中额外加入一个非常大的惩罚项. 资源预测智能体的完整决策只有一步动作, 是单步 MDP, 完整的交互轨迹包括接受问题输入, 输出预测的人力负载, 及由外部环境评估预测的人力负载的准确性.

工序调度智能体. 工序调度智能体 π_p 在给定的资源限制下, 在节拍周期限制内完成调度工作, 生成满足所有约束的可行调度方案. 可观测的状态除了原调度问题信息外, 额外包括资源预测智能体给出的人力负载峰值限制和调度分配过程中工序的动态信息. 工序调度智能体的 MDP 基于经典贪心框架构建, 每次执行动作从输出的概率分布 $\pi_p(j'|G, j)$ 中, 采样选出一个未开工的工序, 将其分配到资源允许的最早时刻开始执行. 不断重复这一决策动作直到所有任务均被调度执行. 当智能体完整生成一个调度方案后, 外部环境将评估并返回整个决策轨迹的奖励信号. 奖励函数为完成全部工序的总时长的相反数 $-\max_j(s_j + t_j)$, 如果最终生成的调度方案无法满足节拍周期约束, 在奖励函数中额外加入一个惩罚项. 工序调度智能体在决策中间步骤的奖励信号都是 0, 只在完成整个序列决策后返回奖励信号, 也可以被认为是一个单步 MDP. 完整的交互轨迹包括接受问题输入和人力负载, 生成完整的调度方案, 以及外部环境评估生成的调度方案是否符合约束.

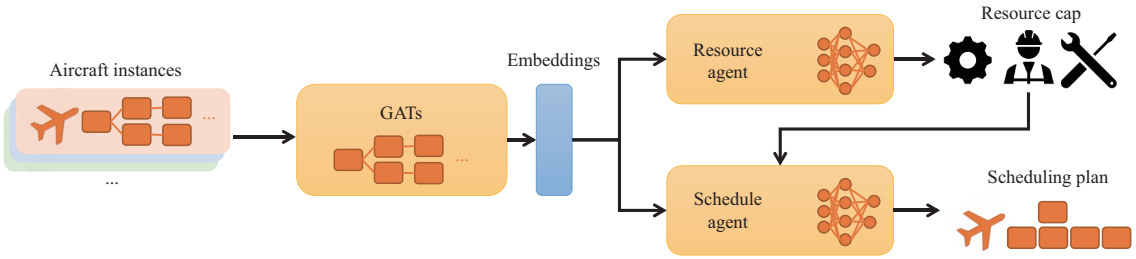

图 4 (网络版彩图) 双重强化学习框架. 将飞机装配任务构成的有向无环图作为输入, 图注意力网络把每个工序映射为对应的特征嵌入, 并生成全局特征. 资源预测智能体和工序调度智能体共享这些问题的特征嵌入. 资源预测智能体根据问题特征嵌入预测资源限制, 随后工序调度智能体结合问题特征嵌入和预测的资源限制逐步生成调度方案

Figure 4 (Color online) Bi-level reinforcement learning framework. The framework takes as input a directed acyclic graph representing aircraft assembly tasks. A graph attention network maps each task to its corresponding feature embedding and generates a global feature. These problem feature embeddings are shared between the resource agent and the scheduling agent. The resource agent utilizes the problem feature embeddings to predict resource constraints. Subsequently, the scheduling agent combines the problem feature embeddings and the predicted resource constraints to iteratively generate scheduling plans

网络架构和拓扑关系编码. 图 4 展示了双重强化学习智能体的网络架构. 本文使用两层图注意力网 (GATs) [9] 编码调度问题中工序之间的拓扑关系 G, 为每个工序和全局调度任务生成对应的特征向量. 图注意力网络允许节点与其邻居节点之间相互传递信息, 能够有效捕捉节点之间的关系, 如式 (9) 所示:

$$h_j^{(i)} = \sigma\left(\sum_{j'\in\mathcal{N}_j} a_{jj'}\boldsymbol{W}_{\text{value}}^{i} h_{j'}^{(i-1)}\right), \tag{9}$$

其中 $h_j^{(i)}\in\mathbb{R}^d$ 表示第 i 层工序 j 的特征向量, $\mathcal{N}_j$ 表示工序 j 的后继节点相邻节点, $\boldsymbol{W}_{\text{value}}^{i}$ 是共享的参数矩阵, σ 是非线性激活函数. 注意力系数 $a_{jj'}$ 的计算如下:

$$a_{jj'} = \text{softmax}_{j'\in\mathcal{N}_j}\left(\frac{(\boldsymbol{W}_{\text{query}}h_j)(\boldsymbol{W}_{\text{key}}h_{j'})^{\mathrm{T}}}{\sqrt{d}}\right), \tag{10}$$

其中 $\boldsymbol{W}_{\text{query}}$ 和 $\boldsymbol{W}_{\text{key}}$ 是共享的两个参数矩阵. 资源预测智能体 π_r 是两层感知机网络, 工序调度智能体 π_p 是由两层图注意力网络构成的指针网络 [10]. 两个智能体共享图注意力网络编码后的特征向量. 在飞机装配脉动生产线应用场景下, 拓扑图 G 中边的依赖关系由工序的前置要求构成, 其中每个工序结点的原始特征向量 $h_j^{(0)}$ 由工序元信息组成, 包括工序所需工种的人数、执行时间、所在型架站位等, 特征向量的每个维度都被归一化到区间 $[0,1]$ 后输入网络. 在原拓扑图基础上, 额外设置一个超源结点, 将该超源结点的特征作为问题的全局特征向量.

交替式训练算法. 资源预测智能体 π_r 和工序调度智能体 π_p 在整个飞机装配调度任务中分别为其负责的子任务工作, 逻辑上将彼此视作外部环境的一部分, 训练时采取相互交替的训练方式: 训练资源预测智能体时, 将工序调度智能体的参数固定; 训练工序调度智能体时, 将资源预测智能体的参数固定. 为了让训练过程更加稳定, 所有的奖励函数都被规范化到 $[-1,0)$ 的区间内, 奖励函数中的惩罚项被统一为在最终奖励中额外加上 -1. 两个智能体都采用经典的策略梯度强化学习算法 REINFORCE [12] 进行优化. 每步训练时, 针对当前调度问题拓扑图 G 采样生成一组探索解 $\{\tau^1,\tau^2,\ldots,\tau^B\}$. 为了最大化期望奖励 J, 使用式 (11) 进行梯度上升轮流对两个智能体进行优化:

$$\nabla_\theta J \approx \frac{1}{B}\sum_{i=1}^{B}\big(R(\tau^i)-b(G)\big)\nabla_\theta\log\pi_\theta(\tau^i|G), \tag{11}$$

其中 $\pi_\theta(\tau^i|G)$ 对于资源调度智能体 π_r 输出的是资源上限的概率 $\Pi_{k=1}^{|\mathbb{K}|}\pi_r(n_k^i|G)$, 对于工序调度智能体 π_p 输出的是整个工序执行优先级顺序的概率 $\Pi_{k=1}^{|\mathbb{J}|}\pi_r(j_k^i|G, j_{k-1}^i)$. $R(\tau^i)$ 表示求解轨迹 τ^i 的奖励函数, $b(G)$ 表示蒙特卡洛 (Monte Carlo) 贪心基线. REINFORCE 算法适合优化本文中资源预测智能体和工序调度智能体所对应的单步 MDP. 基于蒙特卡洛的 REINFORCE 算法具备训练收敛稳定的特点, 而主要缺陷是数据利用率较差, 但是在求解组合优化问题的场景下, 训练时探索轨迹的获取是比较廉价的.

有监督预训练. 研究中发现, 在训练双重强化学习智能体的早期阶段, 随机初始化的智能体会花费大部分的探索机会去寻觅一组可行的资源配比. 大量失败尝试在训练早期容易让智能体失去方向, 无法稳定训练. 针对该难题, 本文提出可以在训练早期对资源预测智能体引入有监督预训练实现强化学习热启动, 减少错误尝试. 由于飞机装配调度问题的 "问题 – 解" 标注数据无法获得, 本文提出使用任意可行的初始解作为弱标签进行有监督预训练. 一种可以快速获得可行的平凡解的方式是在假设资源无限的理想情况下, 对全部工序进行优先调度. 这种平凡解虽然距离最优解相去甚远, 但已经可以引导强化学习在训练早期就将优化重心落在惩罚项较少的可行解区域, 绕过大量失败尝试, 提高探索效率, 加速训练进程. 第 4.3 小节中的实验证明了使用有监督预训练的重要性.

优化导向探索. 资源预测智能体的决策动作是预测目标资源上限, 这和脉动生产线调度问题的优化目标一致. 对于资源预测智能体来说, 决策动作有着较好的可解释性: 当预测的资源能够生成可行调度方案时, 说明资源可能溢出, 减少某些资源是获取更多奖励的优化方向; 当预测的资源无法生成可行调度方案时, 增加某些资源是跳出非法解的优化方向. 因此本文在实现资源调度智能体探索时, 引入优化导向的探索机制, 即每次训练探索时, 先贪心尝试当前输出的动作 (将高斯分布均值 $\pi_r(n|G)$ 作为动作), 根据返回的奖励信号是否位于可行解区间, 分别决定是否向资源增大或减少的方向探索, 降低探索的随机性. 为了训练探索过程的稳定性, 每次探索时, 单个资源维度的随机范围限制在 ±1 的区间内. 第 4.3 小节中的实验证明了引入优化导向探索机制对于提高收敛速度和训练稳定性至关重要.

3.3 调度知识迁移方法

在 2.3 小节中已经讨论过, 强化学习求解 NP- 难问题缺乏鲁棒性保证, 本文提出使用知识迁移技术, 将强化学习学到的调度知识迁移到整数规划的剪枝规则中, 使用整数规划微调强化学习生成的调度方案. 该调度知识迁移方法融合了强化学习高效求解的特点和整数规划求解的鲁棒性优势.

如图 3 所示, 知识迁移技术在整个调度问题求解方法中可以视作强化学习的后处理阶段. 知识迁移基于强化学习方法生成的调度方案, 使用整数规划求解器进一步微调, 将微调的调度方案限制在强化学习生成的调度方案的邻域内. 本文引入一个可以由操作人员调整的迁移系数 Δ 来控制整数规划搜索的邻域半径, 具体约束描述如下:

$$s_j^0 - \Delta \leqslant s_j \leqslant s_j^0 + \Delta, \tag{12}$$

其中 s_j^0 表示任务 j 在现有调度方案的起始时间. 该约束要求调整后的 s_j 在现有调度方案的周围 Δ 邻域内. 微调阶段使用整数规划求解器对式 (3)~(8) 和 (12) 共同定义的整数规划模型进行求解. 图 3(b) 展示了不同 Δ 对整数规划搜索范围的影响: 当 Δ 取值较小时, 强调现有调度知识的重要性, 减少微调幅度, 提高微调速度, 但是可能引起负迁移, 导致无法找到最优调度方案; 当 Δ 取值较大时, 降低迁移知识限制, 增大微调幅度, 有更大概率找到最优调度方案, 但是会显著增加计算开销. 一线操作人员可以根据实际应用的实效性调整迁移系数 Δ, 获得负迁移和求解开销之间的权衡.

工序节拍稳定假设. 引入知识迁移和整数规划作为强化学习方法的鲁棒性保证后, 可以充分释放

整数规划的微调能力, 简化强化学习的学习环境, 将一部分优化难度留到知识迁移阶段. 因此本文提出工序节拍稳定假设. 如果将不同架次飞机在型架内的生产进度完全独立考虑 (如式 (5) 中只考虑上架型架时保持节拍), 会引入大量待求变量, 大幅增加工序调度智能体的执行轨迹长度, 不利于强化学习智能体的训练. 本文提出在第一阶段强化学习求解时, 将型架节拍稳定约束扩充到工序节拍稳定假设约束, 即不同飞机之间的所有对应任务都严格保持节拍稳定, 即将约束条件由式 (5) 加强为

$$s_j^{(a')} - T = s_j^{(a)}, \quad \text{其中 } \forall j \in \mathbb{J}, a' = \mathrm{NextTakt}(a). \tag{13}$$

式 (5) 仅对每架飞机上架各自型架的起始时间 $s_f^{(a)}$ 做出约束, 而式 (13) 要求任意两架相邻飞机 a 和 a' 的相同装配工序的起始时间都固定相差一个节拍单位 T. 在该假设下, 只需要安排第一架飞机的装配任务, 就可以依照约束限制 (13) 递推出所有飞机的调度方案, 从而极大减少工序调度智能体所需做出的决策数量.

节拍稳定假设的微调修正. 虽然工序节拍稳定假设能够大幅减少待求解变量的数量, 降低强化学习的训练难度, 但是加强约束势必会导致求解质量受限. 同时飞机订单中存在部分定制化需求, 手工作业现场也难免出现突发情况等问题. 因此即使是同样型号飞机, 不同架次之间也存在 $1\% \sim 5\%$ 的工序差异. 这些差异意味着无法将所有工序都通过式 (13) 推导出起始时间 s_j^0. 为了处理这些存在特别差异的装配任务, 本文将存在差异的任务 s_j 的约束范围从式 (12) 放宽到其所在型架 f 的对应工作区间 $[s_f, s_f + H_f)$ 内. 实验证明了工序节拍稳定假设能够显著降低问题的微调求解难度.

针对工序变动的适应性. 在生产实践中, 作业现场经常会出现突发变动, 现有的解可能已经不再符合全部约束. 这要求调度系统能够具备重规划或规划微调的功能. 此时可以重复整个求解流程, 也可以直接应用提出的知识迁移技术. 本文提出的知识迁移技术也可以将已有调度方案的知识迁移到整数规划剪枝约束中, 进行快速微调和重调度, 从而实现调度系统对于工序变动的适应性.

4 生产线实验评估与分析

为验证本文所提方法的有效性, 本文实现了一个求解飞机装配脉动生产线调度问题的原型方案, 以人力资源负载为优化目标, 在仿真模拟数据和在真实飞机装配产线上收集的历史生产数据上进行了实验验证. 真实案例的验证结果证明了本文提出的方法的优越性能.

本文选取了 5 个不同的评测案例, 包括 4 个生成的仿真调度问题 A, B, C, D 和产线真实历史数据 E. 具体地, 测试数据 E 包含实际飞机装配生产线某一年的完整生产工序信息, 模拟数据 A, B, C, D 通过从实际飞机装配生产线的完整装配生产工艺中采样一部分子图构造而成, 采样比例依次提高, 问题规模依次递增: A 中每架飞机包含约 200 道工序和两个顺序连接的装配型架, 大约需要数十人即可完成任务; B, C 和 D 中每架飞机分别包含约 1000, 2000 和 3000 道待调度工序; 原始测试数据 E 中每架飞机包含近万个工序, 需要数百人才能完成生产计划. 各个问题都以年产 30 架飞机计算节拍周期进行求解. 在所有实验评测中, 本文提出的强化学习方法只在真实案例数据 E 上进行一次完整训练, 可在所有测试案例上使用同一模型进行预测调度. 在不同问题上测试时不再对模型进行额外微调训练, 具有很高的实际运行效率. 本文提出的模型也可以在目标案例上进一步微调学习 [13], 通过牺牲一定的求解效率进一步提升求解质量, 相关扩展不在本文的讨论范围内.

本文在实验基线上主要选取了 ILP, ILP-2, SPT+ 和 WKR+. 其中 ILP 使用现有的整数规划求解器直接求解 3.3 小节中定义的整数规划模型. ILP-2 在 3.3 小节整数规划定义的基础上, 引入了装配工序完全节拍稳定假设和知识迁移微调技术. 所有 ILP 相关算法都提供了 72 小时运算时间. SPT

(shortest processing time) 和 WKR (work remaining) 是经典的工期优化启发式方法, 为了验证强化学习算法的有效性, 本文将双重强化学习中的工序调度智能体分别用 SPT 和 WKR 替换, 嵌入到双重强化学习框架中, 与资源预测智能体一同训练和预测, 保持与其他完整求解流程一致, 作为基线 SPT+ 和 WKR+. 图 5 中的纵坐标是优化目标数值, 即飞机装配脉动生产线的人力资源负载 (越小越好), 为了方便比较, 图中结果是将本文方法 (ours) 结果作为 1 后各方法的相对比值.

4.1 对不同规模问题的求解效果

本文提出的双重强化学习框架可以克服 ILP 类方法的扩展性难题. 图 5(a) 比较了 ILP, ILP-2, SPT+, WKR+ 和本文的方法在 A, B, C, D 和 E 五个问题规模依次递增的测试问题上的评测结果. 其中, ILP 方法只能解决较小的测试用例 A, 在更大的问题 B, C, D 和 E 上, 无法在合理的时间 (约 1 周) 内返回可行解. 作为对比, 两阶段求解的 ILP-2 能够在 B, C, D 和 E 上初步返回结果, 表明装配工序完全节拍稳定假设和知识迁移微调技术能够显著降低求解难度, 使 ILP-2 也能够在一段较长的运算时间内获得可行解. 实验结果表明本文提出的完整方案 (ours) 在各个不同问题上都展现出一致的优越性. 与本文提出的双重强化学习框架之内的基线 SPT+ 和 WKR+ 的比较同时证明了工序调度智能体在与环境的交互中学习到了比传统启发式算法更优质的调度策略. 图 5(b) 比较了双重强化学习方法与各个实验基线方法在第一阶段求得初始求解方案时的效果. 只考虑双重强化学习自身的预测能力, 也要强于所有基线方法. 由于本文方法在第一阶段求解初始方案时引入了工序完全节拍稳定假设, 在初始求解方案中假设各架飞机之间的工序一一对应 (满足式 (13)), 在第二阶段知识迁移微调过程中才会将飞机间的差异工序纳入考虑, 并微调初始方案中可能不满足约束条件的部分. 因此, 初始求解方案与实际可行方案存在差异, 其指标也无法与最终方案的指标直接比较.

4.2 对不同产量需求的求解效果

在飞机装配脉动生产线上, 增加计划产量会对应地缩短生产周期, 意味着每架飞机需要在更短的时间窗口内完成装配. 增加飞机产量会让问题求解规模增加, 同时减小可行解的范围, 增加调度方法得到可行解的难度. 图 5(c) 比较了各个方法在不同飞机产量时的结果, E-1.0 是飞机装配产线历史案例, E-1.1, E-1.2, E-1.4 和 E-1.5 是将订单数量增加一定比例后的结果, 数值越大则产量要求越高. 实验结果证明了本文方法在订单增加的情况下依旧具备明显的性能优势. 值得注意的是场景 E-1.5 已经达到了装配脉动产线在当前物料限制下的产能极限, 即使此时赋予无限资源, SPT+ 方法也无法得到可行解. 本文提出的强化学习智能调度方法在各个不同产能需求下, 依旧一致地表现出优异的性能. 图 5(d) 比较了不同产量需求下各个方法在第一阶段得到的初始调度方案的质量.

4.3 敏感性分析

为了更好地说明本文提出的各种技术对调度方法性能带来的影响, 进行了一系列敏感性分析. 在强化学习训练阶段, 实验验证了强化学习智能体的有监督预训练机制和优化导向探索机制的作用; 在推理阶段, 实验验证了强化学习的决策采样机制效果和不同的迁移系数 Δ 的影响.

有监督预训练和优化导向探索. 图 6(a) 展示了本文提出的相关训练策略的消融实验结果. 其中绿色曲线表示训练前将资源预测智能体随机初始化而其余实验条件保持一致的训练结果, 实验说明对资源预测智能体进行有监督预训练可以显著提高训练效率, 提高收敛速度. 橙色曲线表示使用标准的 ϵ-贪心探索策略的训练结果, 与本文最终方案的蓝色曲线相比较, 对资源预测智能体引入优化导向探索机制后, 显著减少了无效探索的次数, 整体提高了训练收敛速度和稳定性.

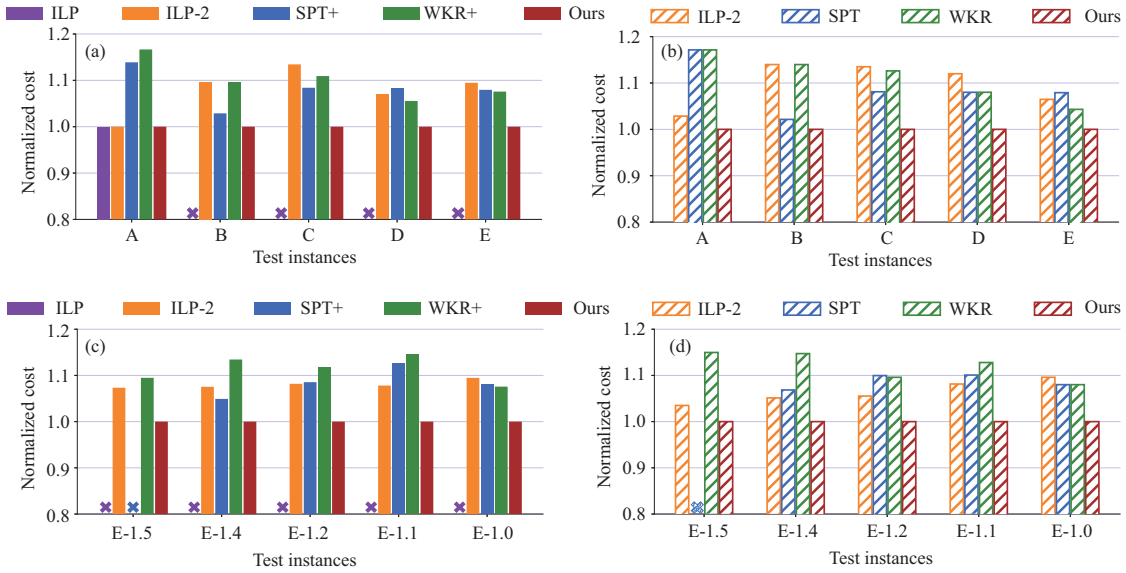

图 5 (网络版彩图) 在仿真数据和飞机装配产线历史数据上的人力资源负载调度结果. (a) 和 (b) 是在不同问题规模的测试案例下的评测结果, (c) 和 (d) 是真实案例 E 在不同产量要求下的评测结果; 其中, (a) 和 (c) 是两阶段求解的最终方案比较, (b) 和 (d) 是第一阶段的初始求解方案比较

Figure 5 (Color online) Results of worker resource loads on simulated data and historical data from aircraft assembly lines. (a) and (b) present evaluation results for different problem scales; (c) and (d) display evaluation results for real case E under different production demands. (a) and (c) compare final solutions of the two-stage solving process, while (b) and (d) compare initial solutions in the first stage

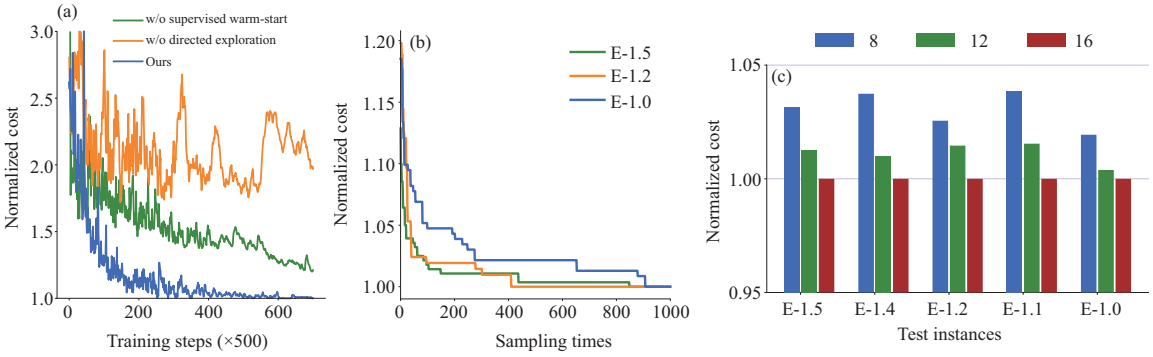

图 6 (网络版彩图) 敏感性分析. (a) 有监督预训练策略和优化导向探索对于整体强化学习优化过程的影响; (b) 强化学习求解的样本效率; (c) 不同迁移系数 Δ 对于二阶段迁移效果的影响

Figure 6 (Color online) Sensitive analysis. (a) Impact of supervised warm-start and directed exploration on the overall reinforcement learning optimization process; (b) sampling efficiency of reinforcement learning solutions; (c) influence of different transfer coefficients Δ on the two-stage transfer effects

强化学习决策采样. 对强化学习智能体进行多次采样是常见的强化学习后处理策略, 相较于整数规划阶段, 强化学习采样的计算开销基本可以忽略不计. 图 6(b) 展示了对强化学习智能体进行决策采样的性能和代价权衡曲线, 可以看出进行适当次数的采样尝试是一种具备高性价比的方法. 进行约 200 次采样可以获得最佳的性能代价平衡, 但是 400 次过后边际效益递减严重. 具体采样次数可以根

据应用现场的计算能力和求解时效要求进行选择.

迁移系数 Δ. 图 6(c) 展示了在第二阶段, 在不同的问题规模下 (A, B, C, D, E) 选取不同的 $\Delta \in \{8, 12, 16\}$ 的结果. 本文方法在各个测试用例上都显示出一致性, 即将 Δ 选取从 8 扩大到 16, 付出更多的运算代价, 在更大范围内能够搜索到更优的解 (约 4%). 这个参数可以由调度操作员在产线上进行调整, 也体现了负迁移和求解开销之间的权衡.

5 相关工作

资源受限调度和飞机装配脉动生产线优化. 飞机装配脉动生产线调度问题旨在保证完成年度生产计划的前提下, 优化资源配置和利用率, 对保障飞机有序生产, 实现降本增效具有重要意义 [14]. 飞机装配问题属于资源受限项目调度问题 (resource-constrained project scheduling, RCPSP) [15]. 经典 RCPSP 工期优化启发式算法已经受到广泛关注与研究 [16]. 针对飞机装配生产线, 近期的工作主要围绕不同业务需求建立整数规划模型, 使用遗传算法等演化算法进行局部搜索求解 [17~19]. 现有工作主要以工期优化为目标, 很少考虑实际脉动生产中的节拍稳定性, 主要在理想的小规模仿真数据中进行验证. 而本文紧紧围绕飞机装配脉动生产线的实际需求, 从节拍稳定性、负载优化和求解有效性等多个因素出发, 在稳定脉动生产节拍的前提下对人力资源负载进行全局优化, 并在真实飞机装配脉动生产线的历史数据上进行验证.

图表示学习和强化学习. 图神经网络是一类为图数据设计的高效表示学习算法 [9, 20~23]. 图表示学习算法已经在众多领域中被采用, 如旅行商问题 [24]、芯片布线 [25]、蛋白质结构预测 [26] 等. 本文在飞机装配工序拓扑图上使用图注意力网络来学习飞机装配调度问题的有效特征表示. 近年来强化学习技术针对不同学习环境的特点, 发展出多种高效算法 [27~30]. 本文根据飞机装配调度场景中模拟数据交互成本低、需要全局优化的特点, 将经典的 REINFORCE 算法 [12] 拓展到双重强化学习框架下.

强化学习求解组合优化问题. 强化学习技术在很多经典组合优化应用上取得了进展 [5~8], 如路径规划问题 [6, 24]、芯片布线 [25]、编译器优化 [31, 32] 和神经网络结构搜索 [33] 等. 本文将飞机装配脉动生产线调度问题转化成了两个子问题, 引入了双重强化学习框架进行求解. 本文还提出了基于调度知识迁移的两阶段混合求解方案, 同时结合强化学习和整数规划求解的优势. 近年来, 将机器学习方法引入整数规划问题的求解过程是又一研究热点 [34~38], 该类方法改进求解器内部运作机理, 不针对某个特定应用场景, 具备一定的通用性. 相关研究成果大多可以直接引入到本文提出的基于知识迁移的第二阶段求解, 取代现有的整数规划求解器.

6 结论与展望

本文提出了一种基于深度强化学习和知识迁移的方法解决飞机装配脉动生产线的调度问题. 通过设计一种双重强化学习机制, 将问题拆分成两个子任务协同求解. 同时, 针对强化学习鲁棒性不足的难题, 提出一种基于知识迁移的整数规划微调方法进一步完善求解流程, 提高算法鲁棒性和对全部约束的满足性. 本文实现了调度原型方案, 在真实飞机装配脉动生产线的历史调度数据上进行评测, 结果显示本文提出的方法在求解时效和求解质量上都具备显著优势. 本文探索了人工智能技术与传统装备制造之间的深度融合, 期待本文提出的求解框架能够在更多的工业场景中获得应用.

参考文献

1 Chang S M, Yang G J, Chen J. Research and application of intelligent manufacturing technology for aircraft final assembly pulsation production line. Aeron Manuf Technol, 2016, 59: 41–47 [苌书梅, 杨根军, 陈军. 飞机总装脉动生产线智能制造技术研究与应用. 航空制造技术, 2016, 59: 41–47]

2 Li X N, Zhi S W, Jiang B, et al. Digital pulsation production line for aircraft final assembly. Aeron Manuf Technol, 2016, 59: 48–51 [李西宁, 支劭伟, 蒋博, 等. 飞机总装数字化脉动生产线技术. 航空制造技术, 2016, 59: 48–51]

3 Li J L, Du B R, Wang B L, et al. Application and development of pulse assembly line. Aeron Manuf Technol, 2013, 56: 58–60 [李金龙, 杜宝瑞, 王碧玲, 等. 脉动装配生产线的应用与发展. 航空制造技术, 2013, 56: 58–60]

4 Karp R M. Reducibility among combinatorial problems. In: Complexity of Computer Computations. Boston: Springer, 1972. 85–103

5 Bengio Y, Lodi A, Prouvost A. Machine learning for combinatorial optimization: a methodological tour d'Horizon. Eur J Oper Res, 2021, 290: 405–421

6 Bello I, Pham H, Le Q V, et al. Neural combinatorial optimization with reinforcement learning. 2016. ArXiv:1611.09940

7 Chen X, Tian Y. Learning to perform local rewriting for combinatorial optimization. In: Proceedings of Advances in Neural Information Processing Systems, 2019

8 Mazyavkina N, Sviridov S, Ivanov S, et al. Reinforcement learning for combinatorial optimization: a survey. Comput Operations Res, 2021, 134: 105400

9 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. In: Proceedings of International Conference on Learning Representations, 2018

10 Vinyals O, Fortunato M, Jaitly N. Pointer networks. In: Proceedings of Advances in Neural Information Processing Systems, 2015

11 Jiang J, Shu Y, Wang J, et al. Transferability in deep learning: a survey. 2022. ArXiv:2201.05867

12 Sutton R S, Barto A G, Williams R J. Reinforcement learning is direct adaptive optimal control. IEEE Control Syst Mag, 1992, 12: 19–22

13 Hottung A, Kwon Y D, Tierney K. Efficient active search for combinatorial optimization problems. In: Proceedings of International Conference on Learning Representations, 2021

14 Li Y, Chang Q, Ni J, et al. Event-based supervisory control for energy efficient manufacturing systems. IEEE Trans Automat Sci Eng, 2018, 15: 92–103

15 Kelly J E. The critical path method: resource planning and scheduling. In: Industrial Scheduling. Upper Saddle River: Prentice Hall, 1963. 347–365

16 Kolisch R, Hartmann S. Heuristic algorithms for the resource-constrained project scheduling problem: classification and computational analysis. In: Project Scheduling. Boston: Springer, 1999

17 Shan S, Hu Z, Liu Z, et al. An adaptive genetic algorithm for demand-driven and resource-constrained project scheduling in aircraft assembly. Inf Technol Manag, 2015, 18: 41–53

18 Fang P, Yang J, Liao Q, et al. Flexible worker allocation in aircraft final assembly line using multiobjective evolutionary algorithms. IEEE Trans Ind Inf, 2021, 17: 7468–7478

19 Jiang C, Zhang J, Long T, et al. An optimization framework for worker allocation in aircraft final assembly lines based on simulation alternative modelling and historical data. Eng Optimization, 2023, 55: 1387–1402

20 Li Y, Tarlow D, Brockschmidt M, et al. Gated graph sequence neural networks. In: Proceedings of International Conference on Learning Representations, 2016

21 Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model. IEEE Trans Neural Netw, 2008, 20: 61–80

22 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2020, 32: 4–24

23 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations, 2018

24 Khalil E, Dai H, Zhang Y, et al. Learning combinatorial optimization algorithms over graphs. In: Proceedings of Advances in Neural Information Processing Systems, 2017

25 Mirhoseini A, Goldie A, Yazgan M, et al. A graph placement methodology for fast chip design. Nature, 2021, 594: 207–212

26 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589

27 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533

28 Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. In: Proceedings of International Conference on Learning Representations, 2016

29 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. 2017. ArXiv:1707.06347

30 Hafner D, Lillicrap T, Ba J, et al. Dream to control: learning behaviors by latent imagination. In: Proceedings of International Conference on Learning Representations, 2019

31 Haj-Ali A, Huang Q J, Xiang J, et al. AutoPhase: juggling HLS phase orderings in random forests with deep reinforcement learning. In: Proceedings of Machine Learning and Systems, 2020. 2: 70–81

32 Trofin M, Qian Y, Brevdo E, et al. MLGO: a machine learning guided compiler optimizations framework. 2021. ArXiv:2101.04808

33 Jia Z, Zaharia M, Aiken A. Beyond data and model parallelism for deep neural networks. In: Proceedings of Machine Learning and Systems, 2019. 1: 1–13

34 Tang Y, Agrawal S, Faenza Y. Reinforcement learning for integer programming: learning to cut. In: Proceedings of International Conference on Machine Learning, 2020. 9367–9376

35 Cappart Q, Moisan T, Rousseau L M, et al. Combining reinforcement learning and constraint programming for combinatorial optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 35: 3677–3687

36 Nair V, Bartunov S, Gimeno F, et al. Solving mixed integer programs using neural networks. 2021. ArXiv:2012.13349

37 Zhang J, Liu C, Li X, et al. A survey for solving mixed integer programming via machine learning. Neurocomputing, 2023, 519: 205–217

38 Guo W, Zhen H L, Li X, et al. Machine learning methods in solving the boolean satisfiability problem. Mach Intell Res, 2023, 20: 640–655

Scheduling approach for aircraft assembly pulsation production lines with deep reinforcement learning and knowledge transfer

Jincheng ZHONG[1,2†], Haoyu MA[1,2†], Mingsheng LONG[1,2*] & Jianmin WANG[1,2*]

1. *School of Software, Tsinghua University, Beijing* 100084*, China*;
2. *Beijing National Research Center for Information Science and Technology, Beijing* 100084*, China*
* Corresponding author. E-mail: mingsheng@tsinghua.edu.cn, jimwang@tsinghua.edu.cn
† Equal contribution

Abstract Aircraft assembly is a critical process in aircraft manufacturing. Scheduling the assembly pulsation production lines of aircraft assembly in a rational manner for cost reduction and efficiency improvement is an important scientific problem in the intelligent manufacturing field. However, the scenario of aircraft assembly lines is complex, with each assembly involving tens of thousands of operations, which poses new challenges for formally modeling and efficiently solving the aircraft assembly scheduling problem. Thereby, current industry practices heavily rely on manual scheduling through the expertise of human professionals. This paper aims to minimize human resource load and proposes two domain-specific techniques to address the scheduling problem of aircraft assembly pulsation lines. Firstly, the scheduling problem of aircraft assembly pulsation production lines is modeled as two Markov decision processes, and a bi-level reinforcement learning agent is used to make decisions on feasible scheduling solutions for aircraft assembly. Secondly, to tackle the problem of robustness deficiency in reinforcement learning decisions, a domain-knowledge transfer paradigm is proposed, whereas the problem-solving knowledge obtained via reinforcement learning is transferred to the constraint pruning process of the integer linear programming model, and the final scheduling solutions with excellent overall performance are attained through an integer programming solver. Experiments are conducted on real scheduling data from aircraft assembly pulsation production lines. Results demonstrate that the proposed scheduling method based on reinforcement learning and knowledge transfer can successfully scale up to scheduling the assembly pulsation production lines with a yield of nearly one hundred aircraft per year, a problem intractable for combinatorial optimization methods. The solving time of the proposed method is reduced to minutes, and the performance exhibits significant advantages compared to baseline methods.

Keywords aircraft assembly, intelligent scheduling, combinatorial optimization, reinforcement learning, knowledge transfer