控制与决策
Control and Decision

带尺寸约束的二机流水车间生产运输协调博弈调度问题

宫华, 许可, 孙文娟

引用本文:
宫华,许可,孙文娟. 带尺寸约束的二机流水车间生产运输协调博弈调度问题[J]. 控制与决策, 2023, 38(7): 1942−1950.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2207

带尺寸约束的二机流水车间生产运输协调博弈调度问题

宫 华†, 许 可, 孙文娟

(沈阳理工大学 理学院，沈阳 110159)

摘 要: 研究二机流水车间生产运输协调调度问题,当工件在第1台机器加工完成后,由1台带有容量限制的运输车分批次运输到第2台机器加工,运输过程考虑工件尺寸约束,目标函数为最小化最大完工时间. 考虑到源于不同客户的工件对机器及运输设备的竞争,以工件为博弈方,工件在生产运输过程中等待时间为策略,各工件完工时间为收益,建立非合作博弈模型. 通过将问题转化为马尔可夫决策过程,设计线性逼近值函数的Q-learning算法求解纳什均衡调度. 实验结果表明Q-learning算法求得的纳什均衡调度具有较好的全局最优性,从而能够在满足客户的利益下,提高企业的生产效率,实现客户与企业的双赢.

关键词: 生产运输协调调度；非合作博弈；Q-learning算法；纳什均衡；二机流水车间

中图分类号: TP278 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2207 **开放科学(资源服务)标识码(OSID):**

引用格式: 宫华,许可,孙文娟. 带尺寸约束的二机流水车间生产运输协调博弈调度问题[J]. 控制与决策, 2023, 38(7): 1942-1950.

Coordinated two-machine flow-shop and transportation scheduling under size constraint and game theory consideration

GONG Hua†, XU Ke, SUN Wen-juan

(School of Science，Shenyang Ligong University，Shenyang 110159，China)

Abstract: In this paper, a coordinated production and transportation scheduling problem on two-machine flow-shop is studied. After being processed on the first machine, the jobs are transported to the second machine in batches by a transporter with capacity limitation. Each job has different size requirement during transportation. The objective is to minimize the maximum makespan. Since different jobs belonged to different customers have the competition for the machines and the transporter, a non-cooperative game model is established. In this game model, the jobs are viewed as players, the waiting time of each job is viewed as the strategy, and the completion time of each job is viewed as its profit. By transforming the problem into a Markov decision process, a Q-learning algorithm with linear approximate value function is designed to solve Nash equilibrium scheduling. The experimental results show that the approximate Nash equilibrium scheduling obtained by the Q-learning algorithm has a better global optimality. The algorithm can improve the production efficiency and achieve a win-win situation for customers and enterprises in satisfying with the interests of customers.

Keywords: production and transportation coordinated scheduling；non-cooperative game；Q-learning algorithm；Nash equilibrium；two-machine flow-shop

0 引 言

智能制造下生产运输协调优化是微观视角下的产品设备、运输与资源多维度的协同优化与博弈,传统调度方法难以协调均衡多主体博弈决策[1-2]. 以流程工业中带有高温热链的高能耗钢铁企业为例,各生产工序与物料运输紧密衔接,多工序与运输耦合是产品质量保障与高效生产的前提[3]. 高炉到炼钢过程中需要通过鱼雷车运输铁水,炼钢-连铸生产中天车和台车衔接运输钢水,经过连铸机生产的板坯要由辊道或汽车运输到下游加热炉进行热轧. 由于运输工具的能力限制,物料运输对温度、送达时间等要求苛刻,因此考虑带有在制品工序间运输的流水车间生产调

收稿日期: 2021-12-21；**录用日期:** 2022-03-15.

基金项目: 辽宁省“兴辽英才计划”项目(XLYC2006017)；辽宁省教育厅科学研究经费项目(LG202025)；沈阳理工大学科研创新团队建设计划项目(SYLUTD202102).

责任编委: 王凌.

†通讯作者. E-mail: gonghua@sylu.edu.cn.

度问题研究具有一定的实际意义. 科学合理地制定生产与运输协调调度方案,有利于提高生产效率,降低能源消耗,对企业降本增效具有重要意义.

针对二机流水车间生产间运输的协调调度问题大多集中在传统的调度理论与方法,在综合考虑加工能力、运输能力及交货期等其他约束指标的条件下,设计调度方案使得总体目标达到最优. 文献[4]研究了二机流水车间生产与工序间运输的协调调度,模型中考虑运输工具的数量、运输能力和运输时间,对问题的不同目标函数给出多项式算法或者复杂性分析. 文献[5]研究运输工具能力无限的二机流水车间的调度问题并设计近似算法. 针对考虑运输过程工件具有尺寸约束的二机流水车间生产运输协调调度问题,文献[6]提出基于bin-packing的启发式算法并证明近似最优解性能比为7/3. 文献[7]在文献[6]的基础上,提出了改进的bin-packing算法. 文献[8]研究了以最小化最大完工时间为目标函数的两阶段流水车间带中间运输的调度问题,提出了近似性能比为2的启发式算法. 文献[9]针对考虑组间准备时间和运输时间的二机流水车间生产间物流协调调度问题,建立混合整数线性规划模型,并提出采用协同进化离散差分进化算法进行求解. 综上,二机流水车间生产间运输协调调度问题集中在从生产角度出发考虑企业整体的利益,忽略了源于不同客户的工件之间存在的竞争,而这恰恰是影响流水车间生产运输协调调度的一个关键因素. 因此,以隶属于不同客户的加工工件为主体,从客户个体自身利益最大化出发,考虑工件对于加工设备及运输设备的竞争,利用非合作博弈理论来研究生产运输协调调度问题有重要的实际意义.

基于非合作博弈理论的生产调度问题研究强调个体性,博弈方通过竞争得到各自满意的调度方案. 一些学者将非合作博弈理论应用于并行机[10-11]、柔性流水车间[12]、作业车间[13]等调度问题中,通过建立静态博弈模型,设计相应算法求解纳什均衡. 文献[14]针对带有组件更改时间的柔性流水车间调度问题利用博弈理论设计机器分配方案,建立完全信息重复博弈模型,并设计紧致遗传算法求出纳什均衡机器分配方案. 文献[15]针对不确定性混合流水车间调度问题,建立了讨价还价博弈模型,并设计改进的遗传算法对模型进行求解. 文献[16]针对实时多目标柔性车间调度问题,提出了一种基于动态博弈的双层调度方法,通过子博弈完美纳什均衡求得问题的最优解. 文献[17]以最小化完工时间、总机器负荷及临界机器负荷为目标,将其映射为3个博弈方,提出了基于三方博弈的改进遗传算法求解多目标柔性作业车间调度问题,并求得子博弈完美纳什均衡. 综上可知,非合作博弈理论在调度研究的应用大多集中在生产阶段调度目标的最优,鲜少考虑运输与生产衔接与协调.

求解流水车间调度问题的传统方法主要包括精确算法、启发式算法以及智能优化算法. 针对大规模调度问题,精确算法应用比较受限,而启发式算法及智能优化算法不能有效利用历史数据进行学习,较难适应复杂多变的实际生产环境. 强化学习方法可以生成适应实际生产的调度策略,在调度问题中的应用也越来越广泛. 文献[18]将强化学习应用于以最小化平均加权拖期为目标函数的不相关并行机调度问题中,利用带有函数逼近的在线R-learning算法求解. 文献[19]将强化学习用于动态车间调度问题中,考虑工件随机到达和机器故障的发生,利用带有Q-因子的强化学习算法优化变邻域搜索算法的参数. 文献[20]利用Q-learning算法求解置换流水车间调度问题,通过标准数据集验证了算法的有效性. 文献[21]利用深度神经网络模拟状态值函数,提出了一种基于时序差分法的深度强化学习算法,将其应用于非置换流水车间调度问题中. 文献[22]以最小化最大完工时间为目标,提出求解流水车间调度问题的一种基于深度强化学习与迭代贪婪算法的框架,利用强化学习训练模型以获取优良输出结果.

本文研究工件带有尺寸约束的二机流水车间生产运输协调调度问题,应用非合作博弈理论,将源于不同客户的工件映射为博弈方,通过竞争生产设备及运输设备,得到各博弈方满意的调度方案. 同时考虑到基于个体理性出发的非合作博弈的均衡解,可能会导致集体利益受损,即能够使得博弈方个体满意的方案,并不一定是总体最优的调度方案. 本文加入引导机制,设计强化学习Q-learning算法求解生产间运输协调调度问题的博弈模型,使博弈方能够合理竞争,从而求得具有全局最优的纳什均衡调度,以提高生产效率,实现工件所属客户与企业的双赢.

1　问题描述

本文所研究的带有尺寸约束的二机流水车间生产运输协调非合作博弈调度问题描述如下:n个工件隶属于不同客户,需要经过两道工序的加工,每道工序上各有一台机器,各工件的两道工序加工顺序相同. 同一时刻,每台机器只能加工一个工件,每个工件也只能由一台机器加工. 有一辆运输车负责在两道工序之间的衔接运输,工件具有不同的尺寸,同一批运输的工件尺寸之和不能超过运输车的容量. 运输车完成一批运输后返回,等待下一批次的运

输. 并且假设: 1) 所有工件在零时刻等待第一台机器加工; 2) 生产运输过程中有无限缓冲区; 3) 工件在两台机器上的加工时间已知; 4) 运输车运输每一批次工件的时间相同,且运输车返回时间小于运输车运输时间; 5) 机器生产及运输一旦开始就不中断.

考虑到工件所属的客户会通过竞争生产机器及运输车来实现自己的利益最大化,因此,本文所考虑的生产间运输协调调度问题的决策是确定各工件在每台机器上的加工顺序,以及在运输车上的分批与运输顺序,以最小化完工时间为各客户目标,在整体最大完工时间最小的基础上,确定使得各客户满意的调度方案.

相关符号及说明如下.

$M=\{M_1,M_2\}$: 机器集合;

$N=\{J_j|j=1,2,\ldots,n\}$: 工件集合;

V: 运输车;

q_j: 工件J_j的尺寸;

cp: 运输车的容量;

t_1: 运输车从机器M_1运输工件到机器M_2的时间;

t_2: 运输车从机器M_2空车返回机器M_1的时间;

p_{ij}: 工件J_j在机器M_i上的加工时间;

$\mathrm{st}_{kj}(k=1,2,3)$: 分别表示工件$J_j$在机器$M_1$、运输车$V$和机器$M_2$上的开始时间;

$\mathrm{ct}_{kj}(k=1,2,3)$: 分别表示工件$J_j$在机器$M_1$、运输车$V$和机器$M_2$上的完工时间;

$w_{kj}(k=1,2,3)$: 分别表示工件J_j在机器M_1加工前、在M_1加工完成后到运输车V运输前和运输完成后在机器M_2加工前的等待时间;

w_j: 工件J_j在机器M_1、运输车V和机器M_2前的等待时间之和;

$\mathrm{cb}_{kl}(k=1,2)$: 分别表示第l批次运输的工件中最后一个工件在机器M_1、运输车V上的完成时间;

$C_{\max}$: 所有工件的最大完工时间;

Π_N: 所有工件的所有整体调度方案的集合.

2 非合作博弈模型

针对二机流水车间生产间运输协调调度问题,建立非合作博弈模型为三元组$G=\{N,S,U\}$. 其中: $N=\{J_j|j=1,2,\ldots,n\}$表示博弈方集, $S=(S_1,S_2,\ldots,S_n)$表示n个博弈方的策略集, $U=(U_1,U_2,\ldots,U_n)$表示n个博弈方收益函数集.

2.1 博 弈 方

由于n个源于不同客户的工件之间存在对加工机器及运输车资源的竞争,且其行为相互影响,因此将各工件作为博弈方.

2.2 策 略

策略表示博弈方的选择行为. 工件竞争的是加工及运输的先后顺序,所以工件J_j的每个调度方案(包括在两台机器上的加工顺序及运输车上的运输批次)构成博弈方j的一个策略. 但若以此调度方案作为博弈模型中博弈方的策略,则无法直观反映各博弈方策略的变化对总体调度方案的影响,故本文将工件的调度方案映射为工件的等待时间. 由于各工件调度方案的每个可行组合(可行组合是指能够形成一个整体调度方案的组合),对应一个整体调度方案$\pi(\pi\in\Pi_N)$. 而每个整体调度方案π,又与在该方案下工件J_j在机器及运输车前的等待时间w_j对应. 因此,将博弈方j的策略集记为$S_j(S_j=\{s_j|s_j=w_j(\pi),\pi\in\Pi_N\})$.

工件J_j在调度方案π下的等待时间为

$$w_j(\pi)=\sum_{k=1}^{3}w_{kj}(\pi). \tag{1}$$

其中

$$w_{1j}(\pi)=\mathrm{st}_{1j}, \tag{2}$$

$$w_{2j}(\pi)=\mathrm{st}_{2j}-\mathrm{ct}_{1j}=\max(\mathrm{cb}_{1l},\mathrm{cb}_{2,l-1}+t_2)-\mathrm{ct}_{1j}, \tag{3}$$

$$w_{3j}(\pi)=\mathrm{st}_{3j}-\mathrm{ct}_{2j}=\max(\mathrm{cb}_{2l},\mathrm{ct}_{3,j-1})-\mathrm{ct}_{2j}. \tag{4}$$

由于工件均在零时刻到达机器M_1,式(2)表示工件J_j在机器M_1前的等待时间等于其在M_1的开始加工时间;式(3)表示工件J_j在运输车V前的等待时间等于工件在V上的开始运输时间与工件在M_1的完工时间的差值;式(4)表示工件J_j在M_2上加工前的等待时间等于工件在M_2上的开始加工时间与运输完成时间的差值.

2.3 收益函数

在非合作博弈模型中,博弈方的收益函数是对博弈方策略的度量. 本文以各工件在机器M_2的完工时间的相反数作为博弈方的收益,因此工件的完工时间越小,收益值越大. 收益函数集合为$U=(U_1,U_2,\ldots,U_n)$,其中

$$U_j=-\mathrm{ct}_{3j}=-(p_{1j}+p_{2j}+t_1+w_j(\pi)),\quad j=1,2,\ldots,n. \tag{5}$$

2.4 纳什均衡

基于非合作博弈模型,将流水车间生产间运输协调调度问题转化为纳什均衡的求解,即满足: 对于每

个博弈方j,有

$$U_j(s_j^*, s_{-j}^*) \geqslant U_j(s_j, s_{-j}^*),\ \forall s_j \in S_j. \tag{6}$$

其中: s_j^*表示博弈方j的纳什均衡策略, s_{-j}^*表示除了博弈方j以外其他人的纳什均衡策略.

事实上,二机流水车间运输协调调度问题的博弈模型中,可能存在唯一或多个纳什均衡,也可能不存在纳什均衡. 若存在纳什均衡,则最小化所有工件的等待时间之和的最优调度一定是纳什均衡调度.

定理1　对于二机流水车间生产间运输协调调度问题,目标为$\min W(\pi) = \sum_{j=1}^{n} w_j(\pi)$的最优调度$\pi^*$为其博弈模型的纳什均衡调度.

证明　设最小化所有工件等待时间之和的最优调度π^*对应的各工件等待时间为$\omega^* = (w_1^*, w_2^*, \ldots, w_n^*)$. 若$\pi^*$不是纳什均衡调度,则一定存在一个工件$J_j$,其策略$s_j^* = w_j^*$不满足式(6). 即至少存在一个策略$\bar{s}_j = \bar{w}_j$,满足$U_j(s_j^*, s_{-j}^*) < U_j(\bar{s}_j, s_{-j}^*)$. 此时,策略组合$(\bar{s}_j, s_{-j}^*)$为$\omega = (w_1^*, \ldots, w_{j-1}^*, \bar{w}_j, w_{j+1}^*, \ldots, w_n^*)$,与其对应的调度方案设为$\bar{\pi}$.

由于$U_j = -(p_{1j} + p_{2j} + t_1 + w_j(\pi))$,其中$p_{1j}, p_{2j}, t_1$均为常数,当$U_j(s_j^*, s_{-j}^*) < U_j(\bar{s}_j, s_{-j}^*)$时,$w_j^* > \bar{w}_j$,从而$W(\pi^*) > W(\bar{\pi})$,与$\pi^*$是最优调度矛盾. 因此,定理得证. □

在二机流水车间生产运输协调调度问题的博弈模型中,当不存在纯策略纳什均衡或者存在多个纳什均衡时,博弈方无法选择各自策略. 为考虑博弈方利益,一方面要使各客户工件加工的等待时间最短,另一方面也应当使所有工件尽早加工完成. 因此,当博弈模型不存在纳什均衡解时,以最小化所有工件等待时间之和为目标,求解近似纳什均衡解. 由于带有尺寸约束的二机流水车间生产运输调度问题为NP-hard问题,本文利用强化学习Q-learning算法求解二机流水车间生产间运输协调调度问题的博弈模型,将博弈方的收益转化为奖励,通过智能体不断与环境交互进行自我学习,求得近似纳什均衡解.

3　强化学习Q-learning算法

强化学习是人工智能领域中应用马尔可夫决策过程解决序列决策问题的关键技术,已广泛应用于控制、调度等诸多领域. 本文利用基于值函数逼近的Q-learning算法求解二机流水车间生产运输协调调度问题的博弈模型,寻找所有工件等待时间之和最小的调度方案,从而找到博弈模型的近似纳什均衡解.

3.1　问题的转换

应用Q-learning算法求解生产运输调度的关键问题是将调度问题转化为马尔可夫序贯决策问题. 关键问题是构建二机流水车间运输协调调度系统各个时刻的状态特征、行为特征和奖励函数. 状态特征用来描述系统整体环境的特点和变化;行为特征表示智能体执行的动作;奖励函数是关于动作的函数,反映动作的即时奖励,累积奖励用来控制Q-learning算法的长远目标.

3.1.1　状态特征

状态特征主要描述二机流水车间环境的主要特点和变化,通过机器、运输工具和工件的状态变化来反映. 定义的状态特征一般是归一化的数值表征,可应用于不同问题规模且易于计算. 用$f_{i,k}$表示机器$M_i(i = 1, 2)$的第k个状态,其中机器M_1共有5个状态特征,机器M_2共有4个状态特征. $f_{3,1}, f_{3,2}$表示运输车V的2个状态特征, $f_{j+3,1}(1 \leqslant j \leqslant n)$表示工件$J_j$的状态. 因此,本文研究的二机流水车间生产间运输协调调度问题共定义$n + 11$种状态,各状态特征定义如表1所示.

状态特征1描述机器繁忙还是空闲的状态;状态特征2描述了各机器前等待加工的工件数量;状态特征3描述了当前各机器的负载;状态特征4描述了加工时间最小的工件是否在队列中等待;状态特征5根据Johnson法则定义;状态特征6描述了运输车的工作情况,是处于空闲、正在运输、还是空车返回阶段;状态特征7描述了等待运输的工件数量分布;状态特征8描述了工件在整个车间环境中的状态. 所有状态特征提供了某状态下机器、运输车及工件的信息.

3.1.2　行　为

在每个状态下,可供智能体选择的行为决定了机器上工件的加工顺序以及运输车上的运输批次. 针对生产间运输协调调度问题,本文将智能体的行为分为3类: 机器M_1的行为、运输车V的行为和机器M_2的行为. 定义行为空间的实质是缩小了博弈模型中策略全枚举的空间,使得智能体能够在有限区域内搜索到纳什均衡解. 各类行为定义如下:

1) 机器M_1的行为.

行为1 (Johnson规则)　将工件分成两个集合: N_1包含满足$p_{1j} \leqslant p_{2j}$的所有工件,集合N_2包含满足$p_{1j} > p_{2j}$的所有工件. N_1中的工件按照p_{1j}的非减序排列, N_2中的工件按照p_{2j}的非增序排列, 按此顺序选择工件.

行为2 (SPT规则)　按照p_{1j}排序优先选择加工时间最短的工件.

行为3 (LPT规则)　按照p_{1j}排序优先选择加工时间最长的工件.

行为4 (select no job)　等待, 不选择任何工件加

表1 状态特征信息表

类别	序号	状态特征函数	条件	描述
机器	1	$f_{i,1}=\begin{cases}0,\text{机器空闲};\\1,\text{机器繁忙}\end{cases}$	$i=1,2$	机器M_i是否处于工作状态
	2	$f_{i,2}=\dfrac{\eta(Q_i)}{n}$	$i=1,2$	机器M_i前等待加工的工件队列Q_i的工件个数$\eta(Q_i)$与工件总数之比
	3	$f_{i,3}=\Big(\sum\limits_{j\in Q_i}p_{ij}\Big/\eta(Q_i)\Big)\Big(n\Big/\sum\limits_{j=1}^{n}p_{ij}\Big)$	$i=1,2$	当前机器负载等于队列Q_i里工件的平均加工时间与该机器上加工工件的平均加工时间之比
	4	$f_{i,4}=\begin{cases}0,\ J_k=\arg\min\limits_{J_j\in N}\{p_{ij}\}\notin Q_i;\\1,\ \text{否则}\end{cases}$	$i=1,2$	机器M_i上加工时间最小的工件是否在队列Q_i中
	5	$f_{1,5}=\dfrac{\eta(JQ_1)}{\eta(Q_1)}$ $JQ_1=\{J_j\|p_{1j}>p_{2,j},J_j\in Q_1\}$	$Q_1\neq\varnothing$	队列Q_1中在机器M_1的加工时间大于在机器M_2的加工时间的工件数量与Q_1中工件总数之比
运输车	6	$f_{3,1}=\begin{cases}0,\ \text{运输车空闲};\\1,\ \text{运输车正在运输};\\-1,\ \text{运输车空车返回}\end{cases}$		运输车的3种状态: 运输、空闲及空车返回
	7	$f_{3,2}=\dfrac{\eta(Q_V)}{n}$		运输车前等待加工的工件数量$\eta(Q_V)$与工件总数之比
工件	8	$f_{j+3,1}=\begin{cases}0,\ \text{等待第一台机器加工};\\1,\ \text{正在第一台机器加工};\\-1,\ \text{在第一台机器和运输车之间};\\1/2,\ \text{正在被运输};\\-1/2,\ \text{在运输车与第二台机器之间};\\1/3,\ \text{正在第二台机器加工};\\-1/3,\ \text{完成在第二台机器加工}\end{cases}$	$j=1,2,\ldots,n$	工件的7种状态

工. 对于机器M_1,出现此种行为的情形是没有工件等待机器M_1加工或机器M_1繁忙.

2) 机器M_2的行为.

行为1 (FCFS规则) 按照运输到M_2上的工件的到达顺序先到先加工.

行为2 (SPT规则) 按照p_{2j}排序优先选择加工时间最短的工件.

行为3 (LPT规则) 按照p_{2j}排序优先选择加工时间最长的工件.

行为4 (select no job) 对于机器M_2,出现此种行为的情形是没有工件等待机器M_2的加工或机器M_2繁忙.

3) 运输车V的行为.

行为1 (A1规则) 按照工件在机器M_1上完成加工的顺序进行运输,以运输车的容量为限制对工件进行分批,同一批次的工件一起运输.

行为2 (A2规则) 对等待运输的工件按照工件尺寸的非增序排列,以运输车的容量限制对等待运输的工件进行分批次的运输.

行为3 (select no job) 对于运输车出现此种行为的情形是正在运输、空车返回或没有工件等待运输.

状态s下机器M_1、机器M_2及运输车V的行为构成一个行为组合$a(s)$,将所有的行为组合构成的集合记为$\Lambda(s)$.

3.1.3 奖励函数

奖励函数表示动作的即时奖励,累积的奖励表示目标函数. 应用强化学习求解博弈模型中的纳什均衡解的目标是最小化各工件的完工时间,即

$$F=(\min \mathrm{ct}_{31},\min \mathrm{ct}_{32},\ldots,\min \mathrm{ct}_{3n}). \tag{7}$$

由于每个工件的完工时间与工件在整个系统的等待时间相关,当工件不是正在M_1上加工,也不是正在运输车上运输,并且还没有被M_2加工时,工件处在等待状态. 定义示性函数

$$\delta_j(t)=\begin{cases}0,\ \text{工件}J_j\text{在时刻}t\text{正在}M_1\text{上加工},\\\quad\text{或正在运输车上运输},\\\quad\text{或已开始在}M_2\text{上加工};\\-1,\ \text{otherwise}.\end{cases} \tag{8}$$

奖励函数定义为

$$r_{jk}=\int_{t_k}^{t_{k+1}}\delta_j(\tau)\mathrm{d}\tau, \tag{9}$$

其中r_{jk}表示智能体从决策时刻t_k执行行为后转移到时刻t_{k+1}时关于第$j(j=1,2,\ldots,n)$个分量所获得的奖励. 对于第j个分量,最小化目标函数ct_{3j}即为最大化累积奖励R_j,即

$$R_j=\sum_{k=0}^{K-1}r_{jk}=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}\delta_j(\tau)\mathrm{d}\tau=$$

$$\int_0^{C_{\max}} \delta_j(\tau)\mathrm{d}\tau = -w_j, \tag{10}$$

其中K为一次迭代时间内决策时刻的数量. 又因为

$$w_j = \mathrm{ct}_{3j} - p_{1j} - p_{2j} - t_1, \tag{11}$$

从而有

$$R_j = -\mathrm{ct}_{3j} + p_{1j} + p_{2j} + t_1. \tag{12}$$

由式(12)可知,工件的完工时间越小,获得的奖励越大. 在强化学习算法中考虑所有工件生成奖励的平均值作为智能体所获得的累积奖励,故累积奖励越大,所有工件的等待时间和越小,从而据此来寻找近似纳什均衡解.

3.2　基于值函数逼近的Q-learning算法

本文采用异策略的Q-learning方法,利用线性值函数逼近构建强化学习算法,求解二机流水车间生产运输协调调度问题. 本文所采用的值函数逼近为参数化逼近,通过更新基函数权重来更新状态值函数,计算公式如下:

$$Q(s,a) = \sum_{k=1}^{n+11} \theta_k^a \phi_k(s). \tag{13}$$

其中: $n+11$表示状态向量中分量的个数, $\phi_k(s)(1 \leqslant k \leqslant n+11)$表示定义在状态空间中的基函数, θ_k^a表示在状态s_k下选择行为$a \in \Lambda(s)$的权重. 正规化的基函数如下所示:

$$\phi_k(s) = \begin{cases} f_{k,1},\ 1 \leqslant k \leqslant 2; \\ f_{k-2,2},\ 3 \leqslant k \leqslant 4; \\ f_{k-4,3},\ 5 \leqslant k \leqslant 6; \\ f_{k-6,4},\ 7 \leqslant k \leqslant 8; \\ f_{1,5},\ k = 9; \\ f_{3,1},\ k = 10; \\ f_{3,2},\ k = 11; \\ f_{k-8,1},\ 12 \leqslant k \leqslant n+11. \end{cases} \tag{14}$$

基于值函数逼近的Q-learning算法框架如下.

输入: 初始化问题和设置参数.

1) 输入调度问题参数: 工件数量n,运输车容量cp,运输车运输一批工件的时间t_1、返回的时间t_2,工件在各机器上的加工时间及尺寸;

2) 输入Q-learning算法参数: 学习率α,折扣因子γ,贪婪因子ε,衰减率λ,基函数的权重$\theta_a = (1,1,\ldots,1)_{n+11}$,行为组合$a$的资格迹$E(a) = (0,0,\ldots,0)_{n+11}$.

过程:

for $t = 0$: MI do

设置初始时刻t_0及初始状态s_0,初始化基函数.

for num $= 0$: n do

1) 根据ε贪婪策略选择行为并执行: 以ε的概率随机选择候选行为$a_k(a_k \in \Lambda(s_k))$,以$1-\varepsilon$的概率选择最佳行为$a_k^*$,即$a_k^* = \arg\max\limits_{a_k \in \Lambda(s_k)} Q(s_k, a_k)$,并执行选择的行为.

2) 确定状态转移时刻并更新状态: 工件完成加工,工件完成运输、运输车空车返回都是促使状态发生转移的事件,计算智能体从状态s_k采取行为a_k到状态s_{k+1}所获得的即时奖励$r(s_k, a_k, s_{k+1})$,并更新基函数的权重θ_k^a,从而更新状态值函数,其中

$$\theta_k^a = \theta_k^a + \alpha\delta(a_k)E(a_k),$$

$$\delta(a_k) = r(s_k, a_k, s_{k+1}) - Q(s_k, a_k) + \gamma \max_{a_{k+1} \in \Lambda(s_{k+1})} Q(s_{k+1}, a_{k+1}),$$

$$E(a_k) = \lambda E(a_k) + \nabla_{\theta_k^a} Q(s_k, a_k)$$

end for

end for

输出: 所有工件的等待时间及其完工时间.

算法框架主要包括两层循环,内层循环是一次迭代过程中每一步的状态转移,外层循环将在上一次迭代结束后所得到状态s_k下采用行为a时的参数θ_k^a传给下一次迭代相同状态相同行为时的参数,算法流程如图1所示. 通过每次迭代的传递参数,基于线性值

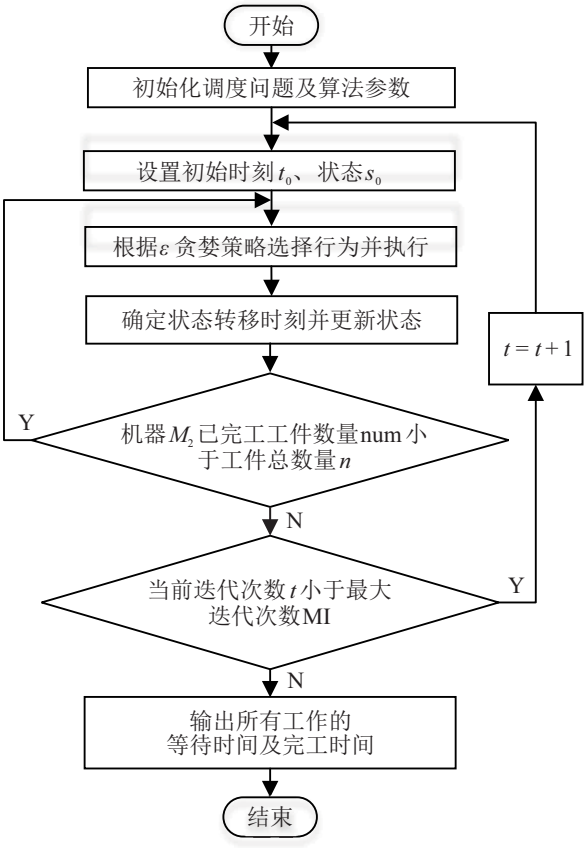


图 1　Q-learning算法流程

逼函数近的Q-Learning算法通过自我学习尝试,学习到博弈问题中的近似纳什均衡策略.

4 实验仿真

4.1 实验环境及参数设置

本节通过实验验证强化学习Q-learning算法求解生产运输协调调度问题博弈模型的有效性. 实验采用Intel(R) Xeon(R) Silver 4110 CPU @2.10 GHZ 处理器, 16 GB安装内存, JetBrains PyCharm Community Edition 2017.3.4软件编程实现.

参数设置如下:假设工件加工时间p_{ij}、工件尺寸q_j、运输时间t_1和t_2均服从均匀分布,具体为:$p_{ij} \sim U[1,50], q_j \sim U[1,25], t_1 \sim U[10,50], t_2 \sim U[10,30]$;运输车容量$\mathrm{cp} = \mathrm{random}[25,50]$.

基于线性值函数逼近的Q-learning算法中,参数$\alpha, \lambda, \gamma, \varepsilon$的取值通过四因素三水平的正交试验法得到. 算法中$\alpha$是学习率,表示对TD偏差的学习,控制整体向目标函数的收敛速度,一般设$\alpha \in (0,1]$;$\lambda(\lambda \in (0,1))$表示轨迹衰减率;$\gamma(\gamma \in [0,1])$表示折扣因子,用来计算累积回报;$\varepsilon$表示贪婪因子,通常取值不超过0.1. 设置参数$\alpha, \lambda, \gamma, \varepsilon$初始水平分别为:(0.001, 0.05, 0.01, 0.02); (0.002, 0.1, 0.02, 0.05); (0.005, 0.2, 0.002, 0.1). 通过多次实验选取3个水平的参数值,根据$L_9(3^4)$规则对各个参数进行交换,代入Q-learning算法中进行实验,最终得到参数取值为:$\alpha = 0.001$, $\lambda = 0.05$, $\gamma = 0.01$, $\varepsilon = 0.1$.

4.2 实验结果及分析

4.2.1 Q-learning算法的POA

为衡量博弈机制的协调性以及Q-learning算法的有效性,引入指标无秩序代价(POA),通过计算Q-learning算法得到的纳什均衡调度的最大完工时间与最优调度(目标为最小化最大完工时间)的最大完工时间的比值来验证博弈机制的偏差效果. POA计算公式如下:

$$\mathrm{POA} = \frac{C_{\max}^{\mathrm{NE}}}{C_{\max}^{\mathrm{OPT}}}. \tag{15}$$

其中:$C_{\max}^{\mathrm{NE}}$表示Q-learning算法得到的近似纳什均衡调度的最大完工时间,$C_{\max}^{\mathrm{OPT}}$表示最优调度的最大完工时间. 最优调度依然通过Q-learning算法求得,与求解近似纳什均衡调度的区别在奖励函数的设置上,即时奖励定义为机器与运输车空闲时间和的相反数,显然累积奖励越大,调度方案的最大完工时间越小.

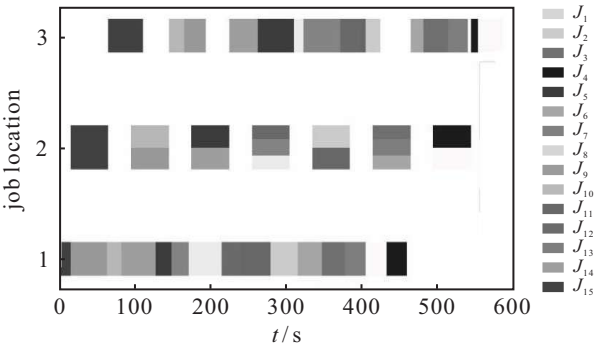
以工件个数$n = 15$为例进行实验,最大迭代次数$\mathrm{MI} = 500$,利用Q-learning算法得到的近似纳什均衡解如表2所示,表示近似纳什均衡调度方案映射出的每个工件的等待时间以及完工时间. 近似纳什均衡调度及最优调度如表3所示,其中近似纳什均衡调度最大完工时间为585,最优调度最大完工时间为579, POA值为1.01. 对应的近似纳什均衡调度及通过Q-learning算法得到最优调度方案如图2所示. 表3中,运输车运输的同一批次工件加方括号表示. 因此,利用Q-learning算法求出的近似纳什均衡调度不仅有利于各工件所属客户,而且从企业角度来说,也能较好地达到整体最优.

表2 近似纳什均衡解

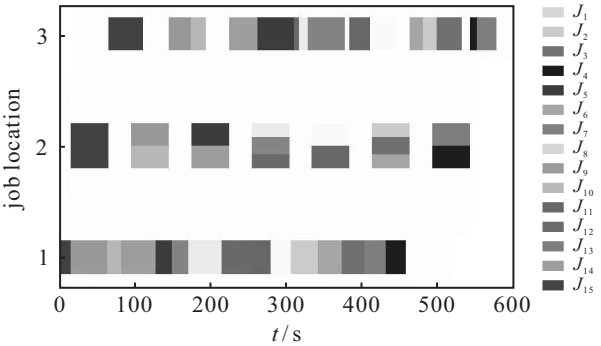
等待时间(策略组合)	完工时间(收益函数)
(476,319,401,467,191,382,259, 216,66,75,309,295,436,128,0)	(585,424,514,553,310,481,371,322, 193,164,405,398,540,262,110)

表3 近似纳什均衡调度及最优调度

工件位置	工件加工及运输顺序	
	近似纳什均衡调度	最优调度
机器M_1	$J_{15}, J_9, J_{10}, J_{14}, J_5, J_7, J_8, J_{12}, J_{11}, J_2, J_6, J_3, J_{13}, J_1, J_4$	$J_{15}, J_9, J_{10}, J_{14}, J_5, J_7, J_8, J_{11}, J_{12}, J_1, J_2, J_6, J_3, J_{13}, J_4$
运输车V	$[J_{15}], [J_9, J_{10}], [J_{14}, J_5], [J_7, J_8, J_{12}], [J_{11}, J_2], [J_6, J_3, J_{13}], [J_1, J_4]$	$[J_{15}], [J_9, J_{10}], [J_{14}, J_5], [J_7, J_8, J_{11}], [J_{12}, J_1], [J_6, J_3, J_2], [J_{13}, J_4]$
机器M_2	$J_{15}, J_{10}, J_9, J_{14}, J_5, J_8, J_7, J_{12}, J_{11}, J_2, J_6, J_3, J_{13}, J_4, J_1$	$J_{15}, J_9, J_{10}, J_{14}, J_5, J_{11}, J_8, J_7, J_{12}, J_1, J_6, J_2, J_3, J_4, J_{13}$



(a) 近似纳什均衡调度



(b) 最优调度

图2 近似纳什均衡调度及最优调度甘特图

4.2.2 不同规模问题的Q-learning算法实验结果

为验证算法稳定性,对不同规模的二机流水车间生产间运输协调调度问题进行实验. 分别取$n = 15, 30, 50, 80, 100, 150, 200$,将在不同启发式规则下求得的$C_{\max}$与$Q$-Learning算法得到的近似纳什均衡调度的$C_{\max}$做比较,如表4所示. 表4中的规则由机器$M_1$规则-运输车$V$规则-机器$M_2$规则3部分构成. 结果表明,与其他启发式规则相比,Q-learning算法得到的近似纳什均衡调度的$C_{\max}$更小,利用Q-learning算法得到的近似纳什均衡调度不仅对各客户有利,而且从整体(生产企业)角度来看更优. 选取$n = 200$,输出不同启发式规则及Q-learning算法得到的$C_{\max}$对比结果如图3所示.

Q-learning算法在每个状态下根据ε贪婪策略选择行为,在尝试中学习到纳什均衡调度方案. 以$n = 100$为例,基于Q-learning算法得到的最大完工时间随迭代次数的变化趋势如图4所示. 说明基于线性值函数逼近的Q-learning算法随着迭代次数的增加不断学习,最大完工时间呈下降趋势,从而获得最大完工时间较小的近似纳什均衡调度.

表 4　不同启发式规则及Q-learning算法得到的最大完工时间

符号	规则	$C_{\max}$						
		$n=15$	$n=30$	$n=50$	$n=80$	$n=100$	$n=150$	$n=200$
H_1	SPT-A1-SPT	860	1 474	2 531	3 968	5 102	7 830	9 159
H_2	SPT-A1-LPT	796	1 349	2 084	3 330	4 988	5 926	9 159
H_3	SPT-A1-FCFS	796	1 349	2 084	3 330	4 988	5 926	9 159
H_4	SPT-A2-SPT	860	1 474	2 562	4 107	5 102	7 749	9 303
H_5	SPT-A2-LPT	796	1 349	2 404	2 850	4 988	4 966	8 759
H_6	SPT-A2-FCFS	796	1 349	2 404	3 028	4 988	5 690	8 759
H_7	LPT-A1-SPT	926	1 851	2 745	4 463	5 835	8 690	11 205
H_8	LPT-A1-LPT	744	1 236	1 971	3 297	5 783	7 015	8 981
H_9	LPT-A1-FCFS	768	1 368	2 039	3 523	5 783	7 015	9 077
H_{10}	LPT-A2-SPT	846	1 712	2 611	4 281	5 435	8 854	11 097
H_{11}	LPT-A2-LPT	796	1 349	2 404	2 850	4 988	4 966	8 759
H_{12}	LPT-A2-FCFS	796	1 349	2 404	3 028	4 988	5 690	8 759
H_{13}	Johnson-A1-SPT	956	1 669	2 564	4 050	5 224	8 001	9 956
H_{14}	Johnson-A1-LPT	956	1 669	2 564	4 050	5 148	7 846	9 879
H_{15}	Johnson-A1-FCFS	956	1 669	2 564	4 050	5 148	7 846	9 879
H_{16}	Johnson-A2-SPT	956	1 669	2 644	4 112	5 224	8 124	10 103
H_{17}	Johnson-A2-LPT	956	1 669	2 644	4 050	5 148	7 766	9 879
H_{18}	Johnson-A2-FCFS	956	1 669	2 644	4 050	5 148	7 766	9 879
Q	Q-learning	**585**	**1 060**	**1 539**	**2 421**	**3 197**	**4 658**	**6 073**

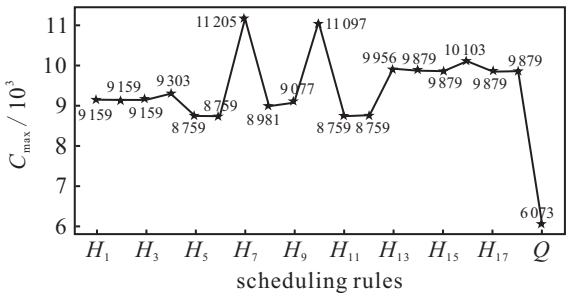


图 3　$n = 200$最大完工时间对比曲线

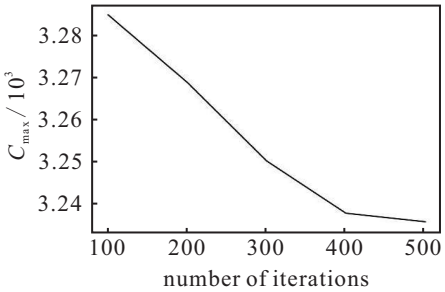


图 4　$n = 100$最大完工时间变化曲线

5 结　论

本文针对二机流水车间生产与生产间运输协调调度问题进行研究,考虑了运输过程工件带有不同尺寸约束. 以不同客户的工件为博弈方,将工件在生产运输过程中等待时间和映射为博弈方策略,各工件的完工时间为其收益,建立非合作博弈模型. 利用强化学习方法对博弈模型进行求解,通过定义系统状态、行为及奖励函数,将协调调度问题转化为马尔可夫决策过程,设计基于线性值函数逼近的Q-learning算法求解纳什均衡调度. 实验结果表明,Q-learning算法能针对不同的系统状态灵活选择行为,求得的近似纳什均衡调度具有较好的全局最优性,均优于其他启发式规则得到的调度方案.

参考文献(References)

[1] 柴天佑, 丁进良. 流程工业智能优化制造[J]. 中国工程科学, 2018, 20(4): 51-58.
(Chai T Y, Ding J L. Smart and optimal manufacturing

for process industry[J]. Engineering Science, 2018, 20(4): 51-58.)

[2] 席裕庚, 王长军. 控制、规划和调度问题中的博弈论应用[J]. 中国计量学院学报, 2005, 16(1): 8-12.
(Xi Y G, Wang C J. The game theory applications in control, planning and scheduling problems[J]. Journal of China Institute of Metrology, 2005, 16(1): 8-12.)

[3] 吴双平, 徐安军. 钢铁生产流程的物质流仿真研究[J]. 钢铁, 2021, 56(8): 73-85.
(Wu S P, Xu A J. A review of mass flow simulation in steel production process[J]. Iron & Steel, 2021, 56(8): 73-85.)

[4] Lee C Y, Chen Z L. Machine scheduling with transportation considerations[J]. Journal of Scheduling, 2001, 4(1): 3-24.

[5] Lee C Y, Strusevich V A. Two-machine shop scheduling with an uncapacitated interstage transporter[J]. IIE Transactions, 2005, 37(8): 725-736.

[6] Gong H, Tang L X. Two-machine flowshop scheduling with intermediate transportation under job physical space consideration[J]. Computers & Operations Research, 2011, 38(9): 1267-1274.

[7] Dong J M, Wang X S, Hu J L, et al. An improved two-machine flowshop scheduling with intermediate transportation[J]. Journal of Combinatorial Optimization, 2016, 31(3): 1316-1334.

[8] Zhong W Y, Chen Z L. Flowshop scheduling with interstage job transportation[J]. Journal of Scheduling, 2015, 18(4): 411-422.

[9] Yuan S P, Li T K, Wang B L. A discrete differential evolution algorithm for flow shop group scheduling problem with sequence-dependent setup and transportation times[J]. Journal of Intelligent Manufacturing, 2021, 32(2): 427-439.

[10] Li K L, Liu C B, Li K Q. An approximation algorithm based on game theory for scheduling simple linear deteriorating jobs[J]. Theoretical Computer Science, 2014, 543: 46-51.

[11] Cole R, Correa J R, Gkatzelis V, et al. Decentralized utilitarian mechanisms for scheduling games[J]. Games and Economic Behavior, 2015, 92: 306-326.

[12] Nie L, Wang X G, Pan F Y. A game-theory approach based on genetic algorithm for flexible job shop scheduling problem[J]. Journal of Physics: Conference Series, 2019, 1187(3): 032095.

[13] 周光辉, 王蕊, 江平宇, 等. 作业车间调度的非合作博弈模型与混合自适应遗传算法[J]. 西安交通大学学报, 2010, 44(5): 35-39.
(Zhou G H, Wang R, Jiang P Y, et al. Non-cooperation game model and hybrid adaptive genetic algorithm for job-shop scheduling[J]. Journal of Xi'an Jiaotong University, 2010, 44(5): 35-39.)

[14] Han Z H, Zhu Y H, Ma X F, et al. Multiple rules with game theoretic analysis for flexible flow shop scheduling problem with component altering times[J]. International Journal of Modelling, Identification and Control, 2016, 26(1): 1.

[15] Safari G, Hafezalkotob A, Khalilzadeh M. A Nash bargaining model for flow shop scheduling problem under uncertainty: A case study from tire manufacturing in Iran[J]. The International Journal of Advanced Manufacturing Technology, 2018, 96(1/2/3/4): 531-546.

[16] Zhang Y F, Wang J, Liu Y. Game theory based real-time multi-objective flexible job shop scheduling considering environmental impact[J]. Journal of Cleaner Production, 2017, 167: 665-679.

[17] 裴小兵, 李依臻. 基于三方博弈的改进遗传算法求解多目标柔性作业车间调度[J]. 工业工程与管理, 2020, 25(4): 59-68.
(Pei X B, Li Y Z. Improved genetic algorithm based on three-party game for MultiObjective flexible job shop scheduling[J]. Industrial Engineering and Management, 2020, 25(4): 59-68.)

[18] Zhang Z C, Zheng L, Li N, et al. Minimizing mean weighted tardiness in unrelated parallel machine scheduling with reinforcement learning[J]. Computers & Operations Research, 2012, 39(7): 1315-1324.

[19] Shahrabi J, Adibi M A, Mahootchi M. A reinforcement learning approach to parameter estimation in dynamic job shop scheduling[J]. Computers & Industrial Engineering, 2017, 110: 75-82.

[20] 张东阳, 叶春明. 应用强化学习算法求解置换流水车间调度问题[J]. 计算机系统应用, 2019, 28(12): 195-199.
(Zhang D Y, Ye C M. Reinforcement learning algorithm for permutation flow shop scheduling to minimize makespan[J]. Computer Systems & Applications, 2019, 28(12): 195-199.)

[21] 肖鹏飞, 张超勇, 孟磊磊, 等. 基于深度强化学习的非置换流水车间调度问题[J]. 计算机集成制造系统, 2021, 27(1): 192-205.
(Xiao P F, Zhang C Y, Meng L L, et al. Non-permutation flow shop scheduling problem based on deep reinforcement learning[J]. Computer Integrated Manufacturing Systems, 2021, 27(1): 192-205.)

[22] 王凌, 潘子肖. 基于深度强化学习与迭代贪婪的流水车间调度优化[J]. 控制与决策, 2021, 36(11): 2609-2617.
(Wang L, Pan Z X. Scheduling optimization for flow-shop based on deep reinforcement learning and iterative greedy method[J]. Control and Decision, 2021, 36(11): 2609-2617.)

作者简介

宫华(1976–), 女, 教授, 博士生导师, 从事生产调度与物流优化、深度学习与强化学习等研究, E-mail: gonghua@sylu.edu.cn;

许可(1982–), 女, 副教授, 博士生, 从事优化理论与算法、生产调度与物流优化等研究, E-mail: xuke@sylu.edu.cn;

孙文娟(1982–), 女, 副教授, 博士生, 从事生产调度与物流优化、优化理论与算法等研究, E-mail: sunwenjuan@sylu.edu.cn.