DOI:10.19344/j. cnki. issn1671-5276.2023.03.001

基于强化学习的柔性作业车间动态调度算法研究

王立群[1],唐敦兵[1],朱海华[1],马国财[2],曹志宏[2]

(1. 南京航空航天大学 机电学院,江苏 南京 210016;
2. 北京电子工程总体研究所 复杂产品智能制造系统技术国家重点实验室,北京 100854)

摘　要:随着制造模式往多品种、小批量的转变,车间生产过程变得复杂多变,传统的依赖于人工和静态式的调度方法已经无法适应实际的车间环境。为此,设计一种基于马尔可夫决策过程的柔性作业车间调度模型。以车间环境作为状态空间,以设备选择作为动作空间,以最小化完工时间作为调度目标,将柔性作业车间调度视为序列化决策问题,使用一种基于策略梯度下降的深度学习方法训练该模型,在仿真环境中对该算法进行验证。结果表明:本算法降低了总完工时间,均衡了机器负载,提高了生产效率和调度智能性。

关键词:柔性作业车间;强化学习;马尔可夫决策过程;车间调度

中图分类号:TH165　　**文献标志码**:A　　**文章编号**:1671-5276(2023)03-0001-03

Research on Dynamic Scheduling Algorithm of Flexible Job Shop Based on Reinforcement Learning

WNAG Liqun[1], TANG Dunbing[1], ZHU Haihua[1], MA Guocai[2], CAO Zhihong[2]

(1. College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China; 2. State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institute of Electronic System Engineering, Beijing 100854, China)

Abstract: In adaption to complex and changeable workshop production process owing to the shifting of manufacturing model from traditional manual and static scheduling method to multiple varieties and small batches, this paper proposes a flexible job shop scheduling model based on Markov decision process, in which the flexible job shop scheduling is regarded as a sequenced decision-making problem, with workshop environment as the state space, equipment selection as the action space, and minimized makespan as the scheduling objective. A deep learning method based on policy gradient descent is used to train the model, which is verified in a simulation environment. The results show that the proposed algorithm can reduce makespan, balance the load of machines, and improves production efficiency and scheduling intelligence.

Keywords: flexible workshop; reinforcement learning; Markov decision process; workshop scheduling

0　引言

工业 4.0 时代,新兴技术如物联网[1]、云计算[2]、人工智能[3]、网络物理系统[4-5]等蓬勃发展,企业的生产模式也从原有的流水线生产往多品种、小批量发展。传统的集中式、静态调度算法已经不能满足目前生产过程所需要的高灵活性、高可重构性。

目前,研究人员为了克服传统车间调度算法的缺点,对车间调度算法进行了许多研究,如运筹学方法、启发式算法、基于规则的算法等。董蓉等[6]为了解决柔性作业车间机器分配不均和工序混乱的问题,设计了一种混合遗传-蚁群算法;SHA D Y 等[7]重新设计了粒子群算法,可以对设置的多个车间调度目标进行优化;周刚[8]设计了一种基于路径搜索的 ABC 算法,并结合人工蜂群算法随机更新种群,提高了算法的全局优化能力。这类方法在车间调度问题是静态调度时,求解速度快,全局性能好,但是存在着无法利用历史学习经验的缺点。

许多新兴技术的出现也促进了柔性作业车间动态调度的研究,比如在线学习、强化学习等。AISSANI N 等[9]设计了一个基于多智能体的自适应调度模型,智能体通过强化学习训练,提高了对动态变化环境的响应能力。陈鸣[10]基于上下文赌博机设计了多 Agent 的自学习策略,Agent 通过学习能够根据不同的环境选择不同的调度策略。蒋静静[11]在每一个决策点,利用预设的调度规则给每个任务分配操作,同时基于多智能体的强化学习算法对模型进行训练。

尽管各研究人员已经在智能制造系统方面取得了一定的成就,但是目前的研究对象主要集中于普通作业车间的静态调度问题;智能体学习的结果多是人为预设的调度规则集,调度智能体只能在规则集中选,缺乏足够的智能性与柔性。

基于上面提出的问题,本文首先设计了一个应用于柔

基金项目:国家重点研发计划项目(2018YFE0117000);国防基础科研计划项目(JCKY2018204B007)

第一作者简介:王立群(1997—),男,江苏淮安人,硕士研究生,研究方向为物联环境下的智能制造系统。

性作业车间的强化学习模型,该模型基于马尔可夫决策过程(Markov decision process,MDP),以机床的选择为动作空间,以最短完工时间为奖励函数。其次,采用策略梯度方法(policy gradient,PG)训练模型,提高算法的计算效率。最后,通过在模拟环境中随机生成订单对算法进行验证。

1　结合 MDP 的车间调度模型

如图 1 所示,车间调度问题可以用 MDP 模型进行描述,在车间的每一个调度阶段,Agent 可观察自身所处的环境,评估每个动作的奖励,从动作集中选择一个动作执行,之后 Agent 和车间环境都会进入到下个状态。本节具体设计了柔性作业车间强化学习模型中状态空间、动作空间、奖励函数这几个关键因素。

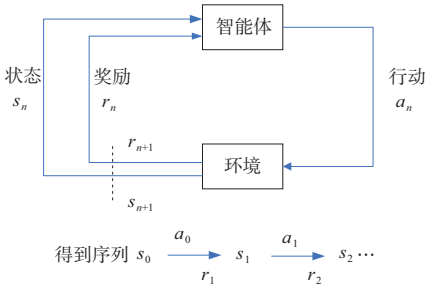


图 1　智能体与环境的交互

1)状态空间

状态空间指的是车间环境处于阶段 n 时包含的生产和调度信息,记为s_n。车间环境包括工件和机床,因此本文将状态空间也分为两类。

第一类状态空间指的是工件的调度信息,包括该工件已经加工完成的工序、等待加工的工序、工序类别等信息,记作s_{n1i}。其中 n 表示工件处于阶段 n,1 表示第一类状态空间,i 表示该工件的工件号。

第二类状态空间指的是机床状态信息,包括机床加工能力、工作状态、工作负载、预计完工时间等,记作s_{n2}。其中 n 表示机床处于状态 n,2 表示第二类状态空间。

综合工件和机床的状态信息,得到状态空间$s_n=(s_{n11},s_{n12},\cdots,s_{n1N},s_{n2})$。

2)动作空间

动作空间也同样分为两类:

第一类动作空间为机床选择动作,在柔性作业车间中,待加工工件的某一道特定工序,只有具备相应加工能力的机床才可以加工。该动作记作a_{1ij},其中 1 表示第一类动作,i 表示工件的工序号,j 表示具备加工能力的机床编号;第二类动作空间为机床加工动作,当工件到达机床开始加工时,修改工件的调度状态,记作a_2,2 表示第二类动作。

智能体位于调度的不同阶段时,可以选择的动作空间也不一样,记可选动作空间为A_{s_n}。其中s_n表示智能体处于的调度阶段,判断是否可以选择第一类动作空间和第二类动作空间见式(1)和式(2)

$$\begin{cases} a_{1ij}\in A_{s_n} & j\in A_{M_i} \\ a_{1ij}\notin A_{s_n} & j\notin A_{M_i} \end{cases} \tag{1}$$

如果机床 j 可以加工工件 i,则可以选择动作空间a_{1ij},否则,不可以选择。A_{M_i}表示对工件 i 有加工能力的设备集合。

$$\begin{cases} a_2\in A_{s_n} & B_M=\varnothing \\ a_2\notin A_{s_n} & B_M\neq\varnothing \end{cases} \tag{2}$$

如果有机床被分配到了加工任务,那么动作空间a_2是可以选择的,否则,不可以选择。B_M 表示所有被分配到了加工任务的设备集合,$\varnothing$ 表示空集。

3)奖惩函数

奖惩函数为智能体的行动提供了即时反馈,使智能体偏向于做出获得最大奖励值的行动。本文设置的调度目标为最小化完工时间,对应的奖惩函数如式(3)所示。

$$r_{n+1}=\begin{cases} 0 & a_n\in A_1 \\ F_n & a_n\in A_2 \end{cases} \tag{3}$$

式中:r_{n+1}表示智能体获得的奖励值;A_1表示第一类动作空间;A_2表示第二类动作空间;F_n表示车间环境从当前状态转移到下一个状态的过程中,车间内所有设备的空闲时间之和。

直接使用r_{n+1}训练模型的时候,存在着奖励值数量级过大和系统偏差的问题,因此本文将式(3)所示的奖励值乘以缩放系数 λ,加上偏置 b 后作为实际的奖励值。

2　智能体训练

1)智能体的训练方法

目前智能体常用的训练方法有深度强化学习(deep Q-learning,DQN)、策略梯度方法(policy gradient,PG)两种。在用这两种方法对智能体进行训练的过程中,发现同样的训练效果,PG 的训练速度比 DQN 要快得多,进行 3 000次训练,PG 只花费了 10 min 而 DQN 花费了 32 min。猜测原因是在车间调度环境比较复杂的时候,由于 DQN 算法每次都会更新经验回放池,每一次更新和采样花费了大量时间。考虑训练的时间和效果,本文选择了 PG 算法来训练上一小节提出的车间调度模型。具体训练方法如下。

输入:策略函数 $\pi(a|s,\theta)$。

输入:状态函数$\hat{v}(s|w)$。

算法超参数:策略函数和状态的更新步长α^θ,α^w。

初始化:策略神经网络变量 θ 和状态值神经网络变量 w。

循环:

对于每一个试验 $s_0,a_0,r_1,\cdots,s_{N-1},a_{N-1},r_N$, 依照策略 $\pi(\cdot|\cdot,\theta)$:

$$\begin{cases} G\leftarrow\sum_{k=n+1}^{N}\gamma^{k-n-1}\cdot r_k \\ \delta\leftarrow G-\hat{v}(s_n|w) \\ w\leftarrow w+\alpha^w\delta\dfrac{\partial\hat{v}(s_n|w)}{\partial w} \end{cases} \tag{4}$$

$$\theta\leftarrow\theta+\alpha^\theta\gamma^n\delta\frac{\partial\ln(\pi(a_n|s_n,\theta))}{\partial\theta} \tag{5}$$

本节设定的目标函数 $J(\theta)$ 如式(6)所示。

$$J(\theta)=-E_{\pi_\theta}\left(\sum_{k=0}^{N}\gamma^k\cdot r_{k+1}\right) \tag{6}$$

式中:奖励值 r 见式(3);γ^k表示第 k 步奖励的衰减率;π_θ

表示策略神经网络参数;s_0表示状态空间。

对目标函数 $J(\theta)$关于策略神经网络参数 θ 求偏导之后,可以得到式(7)。

$$\frac{\partial J(\theta)}{\partial \theta}=-G_n\cdot\frac{\partial \ln(\pi(a_n|s_n,\theta)}{\partial \theta} \tag{7}$$

式中:G_n 表示到阶段 n,智能体获得的总奖励值;$\frac{\partial \ln(\pi(a_n|s_n,\theta)}{\partial \theta}$是利用神经网络计算的结果。

为了降低方差,在梯度表达式(7)中增加了与动作a_n不相关的基准项 $v(s_n)$。

$$\frac{\partial J(\theta)}{\partial \theta}=-[G_n-v(s_n)]\cdot\frac{\partial \ln(\pi(a_n|s_n,\theta))}{\partial \theta} \tag{8}$$

神经网络参数 θ 的更新表达式见式(9)。

$$\theta\leftarrow\theta-\alpha^{\theta}\cdot\frac{\partial J(\theta)}{\partial \theta} \tag{9}$$

式中α^{θ}表示学习率。

将式(8)代入式(9),就可以得到式(5)表示的策略梯度更新表达式。

2)智能体的训练过程

基于上文提出的算法,图 2 展示了训练的完整流程。

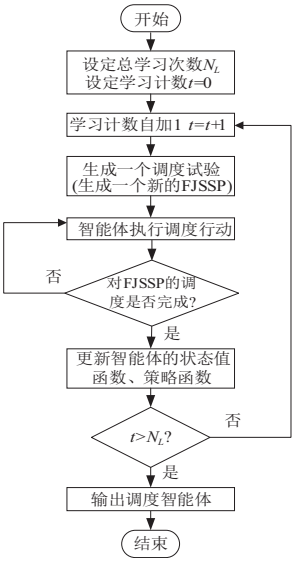


图 2　智能体的训练过程

Step1:初始化总训练次数和当前学习次数。

Step2:当前训练次数自增 1。

Step3:系统随机生成一个调度订单,由智能体进行调度。

Step4:如果当前试验完成,利用本次的经验更新状态和策略函数,提升智能体的智能性;如果没完成,智能体接着执行调度行动。

Step5:如果计数值 t 大于总的学习次数,结束训练,输出训练好的智能体;否则,智能体继续训练。

3　实验论证

1)模拟环境设置

本文在模拟车间中对本文提出的动态调度算法进行了验证。模拟车间环境如图 3 所示,包含 2 台车床、2 台铣床、2 台雕刻机,具体的加工能力如表 1 所示。为了验证本文算法的普适性,参照实际的生产环境生成了订单。每批订单包括 10 个工件,每个工件随机分配车削、铣削、雕刻这 3 个工艺中的 1 个到多个,同时随机设置每道工序的加工时间。

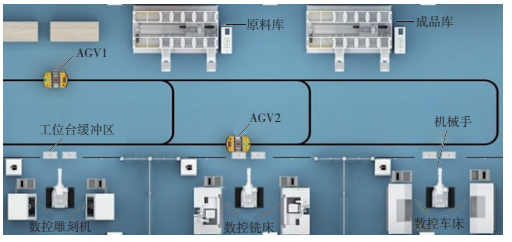


图 3　模拟车间环境

表 1　机床加工能力信息表

机床	工艺类型	缓冲区容量
M1	车	2
M2	车	2
M3	铣	2
M4	铣	2
M5	雕刻	2
M6	雕刻	2

2)调度结果分析

图 4 展示了本文设计的调度智能体的学习曲线。横轴表示 Agent 的学习次数,纵轴表示 Agent 获得的奖励值总和,奖励值总和越大,学习效果越好。从图 4 中可以看出,在 1 100 次试验之前,Agent 可以从每次训练中正向收益,学习效果较好;在 1 200~1 700 次试验之间,由于 PG 方法动作选择的随机性,Agent 在该处的动作选择不好,产生了负向收益;在 2 500 次试验之后,调度模型趋于稳定。

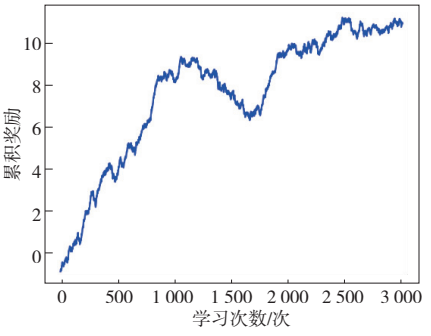


图 4　强化学习算法的学习曲线

图 5 展示了调度智能体的训练效果。图 5(a)是未经训练的智能体产生的调度甘特图;图 5(b)是经过本文提出算法训练之后智能体产生的调度甘特图。从本文设置的最小化完工时间的目标来看,训练后智能体的完工时间降低了近 400 s;从平衡多个机器负载的角度来看,训练前加工任务主要分配给了 M3 号机床,而 M5 号机床却没有被分配到加工任务。训练后,各机床的负载达到了一个总体均衡的状态。这表示经过训练后的智能体具有智能性。

(下转第 10 页)

1)当脉冲能量较小时,在较小的光斑重合度和线重合度下,刻蚀槽外轮廓及扫描轨迹间均会出现较多的材料残留,随着光斑重合度和线重合度的增大,材料去除率增大,逐渐形成底面及外轮廓光滑的凹槽;同时,通过提高脉冲能量也可以明显减缓这种现象,增大材料去除率。

2)线重合度为-33%、0%时,凹槽外径随脉冲能量的增大而增大,且均小于理论外径 430 μm。增大线重合度为 33%时,得到的凹槽外径与理论外径值相当,且随着脉冲能量的增大,其值没有明显变化。因此,对于 100 μm~200 μm 深度的表面刻蚀加工,采用较低光斑重合度及线重合度与较大脉冲能量配合加工对于提高加工效率有利。

参考文献:

[1] 邱海鹏,陈明伟,谢巍杰. SiC/SiC 陶瓷基复合材料研究及应用[J]. 航空制造技术,2015,58(14):94-97.

[2] 刘巧沐,黄顺洲,何爱杰. 碳化硅陶瓷基复合材料在航空发动机上的应用需求及挑战[J]. 材料工程,2019,47(2):1-10.

[3] 焦健,王宇,邱海鹏,等. 陶瓷基复合材料不同加工工艺的表面形貌分析研究[J]. 航空制造技术,2014,57(6):89-92.

[4] 张灿祥,张卫锋,刘致君,等. 旋转超声钻削加工的研究现状及发展趋势[J]. 机械制造与自动化,2021,50(1):1-5.

[5] 夏博. 飞秒激光高质量高深径比微孔加工机理及其在线观测[D]. 北京:北京理工大学,2016:14-42.

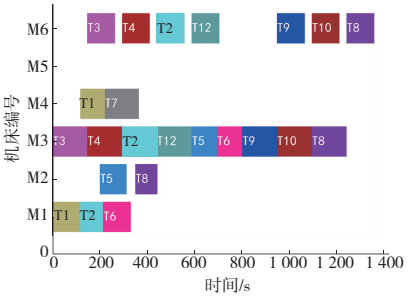
[6] 邱一,刘壮,李元成,等. 飞秒激光扫描去除 CFRP 复合材料的热累积分析[J]. 应用激光,2021,41(5):1004-1010.

[7] ZHAI Z Y, WEI C, ZHANG Y C, et al. Investigations on the oxidation phenomenon of SiC/SiC fabricated by high repetition frequency femtosecond laser[J]. Applied Surface Science, 2020, 502:144131.

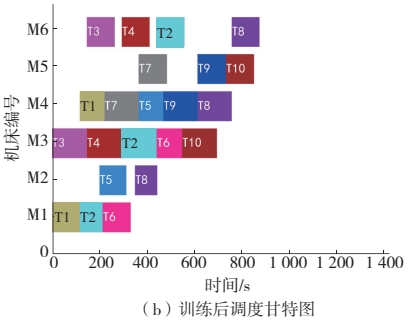
[8] LIU Y S, WANG C H, LI W N, et al. Effect of energy density and feeding speed on micro-hole drilling in C/SiC composites by picosecond laser[J]. Journal of Materials Processing Technology, 2014, 214(12):3131-3140.

收稿日期:2022-01-05

(上接第 3 页)



(a) 训练前调度甘特图



(b) 训练后调度甘特图

图 5　智能体训练效果图

4　结语

随着工业 4.0 时代的到来,多品种小批量的生产模式、快速变化的客户需求需要制造系统具有高灵活性和高可重构性。针对当前动态车间调度算法存在的不足,本文提出了一个应用于柔性作业车间调度模型,并设计了其关键要素;在该模型的基础上,又设计了其对应的训练方法;最后仿真结果表明,本文训练的调度智能体有效地提高了制造车间的生产效率。

由于时间的限制,所做的研究还存在着以下不足。1)本文提出的调度智能体模型是基于全局调度的,但是当制造系统的规模变大和扰动事件频发时,全局调度的性能很差,后续可以考虑以车间的每个设备作为一个自主的调度智能体,设计多智能体强化学习模型。2)本文设计的调度模型目标函数只考虑了最短机床空闲时间,后续可以考虑设备负载、加工率等其他优化目标。

参考文献:

[1] CHEN F, DENG P, WAN J F, et al. Data mining for the internet of things: literature review and challenges[J]. International Journal of Distributed Sensor Networks, 2015, 11(8):431047.

[2] XU X. From cloud computing to cloud manufacturing[J]. Robotics and Computer-Integrated Manufacturing, 2012, 28(1):75-86.

[3] TING D S W, PASQUALE L R, PENG L, et al. Artificial intelligence and deep learning in ophthalmology[J]. The British Journal of Ophthalmology, 2019, 103(2):167-175.

[4] ZHANG Y F, GUO Z G, LYU J X, et al. A framework for smart production-logistics systems based on CPS and industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2018, 14(9):4019-4032.

[5] ZEADALLY S, SANISLAV T, MOIS G D. Self-adaptation techniques in cyber-physical systems (CPSs)[J]. IEEE Access, 2019, 7:171126-171139.

[6] 董蓉,何卫平. 求解 FJSP 的混合遗传—蚁群算法[J]. 计算机集成制造系统,2012,18(11):2492-2501.

[7] SHA D Y, LIN H H. A multi-objective PSO for job-shop scheduling problems[J]. Expert Systems With Applications, 2010, 37(2):1065-1070.

[8] 周刚. 基于人工蜂群算法的柔性调度问题研究[D]. 北京:清华大学,2012.

[9] AISSANI N, BEKRAR A, TRENTESAUX D, et al. Dynamic scheduling for multi-site companies: a decisional approach based on reinforcement multi-agent learning[J]. Journal of Intelligent Manufacturing, 2012, 23(6):2513-2529.

[10] 陈鸣. 面向混线生产的多 Agent 智能调度方法研究[D]. 南京:南京航空航天大学,2020.

[11] 蒋静静. 基于深度强化学习的离散型制造企业车间动态调度研究[D]. 西安:西安理工大学,2020.

收稿日期:2021-12-03