基于改进深度强化学习的分布式车间调度研究

周文豪[1]，计效园[1]*，李海龙[1]，侯明君[1]，余朋[1]，涂先猛[1]，殷亚军[1]，周建新[1]

(1.华中科技大学材料学院，材料成形与模具技术全国重点实验室，湖北 武汉 430074)

*通讯作者：计效园，男，教授，博士。E-mail：mail@mail.com

摘　要：多车间分布式生产的主计划调度是调度中的主要挑战之一。对此，本文以最小化订单提前/延迟惩罚、最小化全局最大完工时间和最小化车间载荷分布为优化目标，提出了一种具有模糊加工时间的三目标分布式铸造车间优化模型，并通过改进深度强化学习算法对模型进行求解。在算法中，将车间及铸件作为整体环境建立了智能体，其状态由订单、车间和任务队列描述，动作空间则由任务分配和车间的笛卡尔积组成，并根据目标函数设置奖励机制。同时，采用双层 dueling 网络代替传统的神经网络。最后将所得结果与改进的粒子群算法的结果进行比较发现，提出的深度强化学习在整体性能上与改进的粒子群算法相当甚至优于改进的粒子群算法。

关键词：铸造生产；分布式车间；智能调度；强化学习；多目标优化

Research on Distributed Workshop Scheduling Based on Improved Deep Reinforcement Learning

Zhou Wen-hao[1], Ji Xiao-yuan[1], Li Hai-long[1], Hou Ming-jun[1], Chen Fa-yuan[1], Yin Ya-jun[1] and Zhou-Jianxin[1]

(1.School of Materials Science and Technology, Huazhong University of Science and Technology, National Key Laboratory of Materials Forming and Mold Technology, Wuhan 430074, China)

Abstract: The master scheduling of multi workshop distributed production is one of the main challenges in scheduling. In response to this, this article proposes a three objective distributed casting workshop optimization model with fuzzy processing time, with the optimization objectives of minimizing order advance/delay penalties, minimizing global maximum completion time, and minimizing workshop load distribution. The model is solved using an improved deep reinforcement learning algorithm. In the algorithm, an intelligent agent is established with the workshop and castings as the overall environment, and its state is described by orders, workshops, and task queues. The action space is composed of task allocation and the Cartesian product of the workshop, and a reward mechanism is set according to the objective function. Meanwhile, a double-layer dueling network is adopted to replace traditional neural networks. Finally, the results obtained were compared with those of the improved particle swarm algorithm, and it was found that the proposed deep reinforcement learning had overall performance comparable to or even better than the improved particle swarm algorithm.

Keywords: Casting production; Distributed workshop; Intelligent scheduling; Reinforcement learning; Multi-Objective Optimization

1 前言

铸造是一种工作流程基本固定的离散制造过程。随着生产力的提高，许多铸造企业铸件转向分布式车间生产，不同或者相同类型的产线分布在多个平行铸造车间，每个车间都能够执行铸件的整个生产过程。从本质上来说，这类似于一个并行机器调度问题，长期以来一直是学术界和工业界相当感兴趣的主题[1]。在生产时间可变，生产调度指标多以及分布式生产的情况下，目前常用的人工调度方法容易导致订单延迟、生产效率低、工作量不平衡等问题。对此，近年来采用智能算法进行车间调度的研究不断增多。

对于订单生产时间的不确定，并联铸造车间调度可视为具有模糊加工时间的并联机床调度模型，也称为模糊并联机床调度（FPMS）[2]。许多研究使用模糊数来表示真实环境中的不确定值，Yeh 等人[3-5]研究了具有学习效果和模糊处理时间的梯形并行机器调度问题，并推荐了提出的模拟退火算法。廖和苏[6]使用混合蚁群优化来解决具有梯形模糊处理时间、设置时间和发布时间的并行机器调度问题。何亚东[7]等提出了一种基于模糊偏好的模糊数排序方法，通过建立多个模糊数之间的偏好关系，提高信息的完整性和排序准确度。

帕累托支配规则目前广泛运用在多目标优化算法中，以处理无法通过设置权重或将冗余目标转化为约束的方法转化为单个目标的优化问题中。Sun 等人[8]提出了一种新的多目标两阶段进化算法，首先，使用非支配动态权重聚合方法找到帕累托最优解，然后利用这些解学习帕累托最优子空间进行收敛。Zhou 等人[9]使用多目标差分进化算法来最小化并行批处理机器调度问题的完工时间和总电力成本。然而，由于多目标优化的 NP-hard 特性，传统的基于帕累托支配规则的多目标算法在解决高维问题时仍存在较大的挑战。

近年来，随着人工智能技术的不断发展，深度强化学习（DRL）作为一种新兴算法，有望突破了传统强化学习在复杂高维问题上的限制[10]。经典的 DRL 算法深度 Q 网络（DQN）凭借其出色的无监督学习能力在生产调度领域得到了广泛的应用[11-13]。张[14]等人利用一种基于 DQN 网络改进的 DDQN 算法解决了芯片制造的双目标并行调度问题，Yuan[15]针对具有 AGV 的分布式异构柔性车间调度问题提出了一种基于 DQN 的实时调度方法，该方法具有较好的求解结果以及泛化性能。

综上所述，目前针对车间调度问题的研究大多聚焦于机械加工车间，针对铸造车间的分布式生产调度计划问题的研究较少。且开发的调度算法大多都基于帕累托规则与多目标元启发式算法，对高维问题的求解都存在一定的局限性。由此，对于实际需求，本文基于模糊时间理论建立了分布式车间主调度的数学模型，并开发了一种基于 DQN 算法的改进 DQN 算法，最后通过仿真数据对算法进行了验证，得到的求解结果与此前研究的改进粒子群算法效果相当甚至更加优秀。

2 问题描述与数学模型

2.1 模糊时间处理

我们研究的对象是拥有多个铸造车间的集团式铸造企业，每个车间都有固定的生产工艺，可以独立完成订单的加工。现在有一组 N 个订单需要分配给满足其工艺要求的 M 个平行铸造车间。此外，由于主客观因素的影响，每个订单的加工时间、交货时间并不固定。根据生产人员的经验，每个车间订单的估计生产时间 $\tilde{P}$以三元组的形式给出，分别为乐观完成时间 Opt、最有可能完成时间 Mos 和悲观完成时间 Pes。而订单的交货时间则由四个元素组成：订单预计最早完成时间(Eet)、订单交货期开始时间(Dst)、订单交货期结束时间(Det)、订单预计最晚完成时间(Elt)，我们使用质心去模糊方法对这两种模糊数进行处理，它们的隶属函数如式（1）~（4）所示。

$$\mu_{\tilde{P}}(x_1)=\begin{cases}\dfrac{x\text{-}Opt}{Mos\text{-}Opt} & Opt\le x\le Mos\\ \dfrac{Pes\text{-}x}{Pes\text{-}Mos} & Mos\le x\le Pes\\ 0 & else\end{cases} \tag{1}$$

$$P=\frac{\int x\cdot\mu_{\tilde{P}}(x_1)dx}{\int\mu_{\tilde{P}}(x_1)\,dx} \tag{2}$$

$$\mu_{\tilde{P}}(x_2)=\begin{cases}\dfrac{x-Eet}{Dst-Eet} & Eet\le x\le Dst\\ 1 & Dst\le x\le Det\\ \dfrac{Elt-x}{Elt-Det} & Mos\le x\le Pes\\ 0 & Dst\le x\le Det\end{cases} \tag{3}$$

$$T=\frac{\int_{Eet}^{Elt}x\cdot\mu_{\tilde{P}}(x_2)dx}{\int_{Eet}^{Elt}\mu_{\tilde{P}}(x_2)dx} \tag{4}$$

2.2 数学模型

为了方便表述，在本小节中所使用到的所有数

学符号及含义如表 2.1 所示。同时，简化平行铸造车间调度模型，消除实际情况下复杂约束的不必要影响，我们提出了以下假设。

1）车间的铸造工艺是固定的，不能改变。

2）订单不能拆分或合并。

3）订单一旦处理完毕，就不能被中断或抢占。

4）车间永远不会停止。

表 2.1 所用到的数学符号及含义

Tab. 2.1 The mathematical symbols and meanings

符号	含义
i	订单($i = 1, 2, ..., N$)
k	车间($k = 1, 2, ..., M$)
$\widetilde{P_{ik}}$	订单i在车间k的加工时间
$\widetilde{C_{ik}}$	订单i在车间k的完工时间
$\widetilde{C_k}$	车间k的最终完工时间
Ds_i，De_i	订单i的转运开始、结束时间
$\widetilde{L_k}$	车间k的剩余订单所需加工时间
Lc_i	订单i的单天延期惩罚成本
Ec_i	订单i的单天提前惩罚成本
T_{ik}	订单i在车间k的拖期天数
E_{ik}	订单i在车间k的提前天数
$\widetilde{U_k}$	车间k的总加工时间

基于上述假设与铸造企业的实际生产情况，我们从下面三个方便考虑目标函数：首先，要最大化满足铸造订单按时交货的需求保证企业信誉，同时避免订单过早的提前完成产生库存成本。其次，要最大化提高企业的生产效率，在一个排产周期内的所有的平行铸造车间应保证在尽可能短的时间内完成所有任务。最后，所有车间的加工载荷应尽量保持一致，从而保证员工利益与企业生产稳定。因此，本研究提出最小化订单延期/提前惩罚函数、最小化全局车间最大完工时间、最小化车间加工负载三个目标函数，如式（5）~（7）所示：

$$Min\ F_1 = \sum_{i=1}^{N}\sum_{k=1}^{M} Lc_i X_{ik} T_{ik} + \sum_{i=1}^{N}\sum_{k=1}^{M} Ec_i X_{ik} E_{ik} \quad (5)$$

$$Min\ F_2 = max\,(C_1, C_2, ..., C_M) \quad (6)$$

$$Min\ F_3 = \frac{1}{M}\sum_{k=1}^{M}\left(U_k - \frac{1}{M}\sum_{k=1}^{M} U_k\right)^2 \quad (7)$$

其中，（5）为最小化订单延期/提前惩罚函数，（6）为最小化全局车间最大完工时间，（7）为最小化车间加工负载。由于生产实际限制，模型具有如式（8）~式（14）所示的约束条件：

$$T_{ik} = \begin{cases} C_{ik} - Ds_i & x_{ik} = 1\ and\ C_{ik} > Ds_i \\ 0 & otherwise \end{cases} \quad (8)$$

$$E_{ik} = \begin{cases} De_i - C_{ik} & x_{ik} = 1\ and\ C_{ik} < De_i \\ 0 & otherwise \end{cases} \quad (9)$$

$$\sum_{k=1}^{M} X_{ik} \le 1; i = 1, 2..., N \quad (10)$$

$$\sum_{k=1}^{M} Y_{ijk} \le 1;\ i = 1, 2..., N; j = 1, 2, ..., N; i \neq j \quad (11)$$

$$\widetilde{C_{ik}} = \sum_{j=1}^{N} Y_{ijk}\widetilde{P_{jk}} + \widetilde{P_{ik}} + \widetilde{L_k};\ i = 1, 2..., N; k = 1, 2, ..., M \quad (12)$$

$$\widetilde{U_k} = \sum_{i=1}^{N} X_{ik}\widetilde{P_{ik}};\ i = 1, 2..., N; k = 1, 2, ..., M \quad (13)$$

$$\widetilde{C_k} = \widetilde{U_k} + \widetilde{L_k};\ k = 1,2, ..., M \quad (14)$$

其中，约束（8）和（9）分别表示车间 k 中订单 i 的延迟时间和提前时间。约束（10）确保订单只能分配给车间一次。约束、（11）用于确定车间内订单的顺序。式（12）确定车间 k 中作业 i 的具体完成时间。车间 k 中订单的总处理时间是所有订单处理时间的总和，如式（13）所示。车间 k 的完成时间等于其最后一个订单完成时间，也可以在式（14）中表示为总订单处理时间和初始剩余作业完成时间的总和。

3 改进深度强化学习算法

在强化学习框架中，智能体通过与环境持续交互实现策略优化。每个时刻智能体观测环境状态 S_t，基于当前策略$\pi(a|s)$选择动作 a_t 执行，随后环境返回即时奖励 R_{t+1} 并转移至新状态 S_{t+1}。这一过程形成的状态-动作-奖励序列构成马尔可夫决策过程（MDP），其核心目标是通过最大化累积奖励期望来学习最优策略。2013 年 DeepMind 将深度神经网络与 Q-learning 结合提出的深度 Q 网络（DQN），成功解决了高维状态空间下的学习难题，推动了深度强化学习的爆发式发展。而在深度强化学习架构中，通过引入深度学习的方式来估计智能体在不同状态下的价值（或策略），从而引导智能体选择最

优动作。其动作价值估计公式如式（15）所示：

$$q(s,a) \leftarrow E[R_{t+1} + \gamma max_{a'} q(s',a') | S_t = s, A_t = a] \quad (15)$$

其中，$q(s,a)$代表在当前状态s下执行动作a的动作价值，R_{t+1}为执行该动作得到的即时奖励，γ为折扣因子，$max_{a'} q(s',a')$表示在执行动作a后，状态由s转移到s'后，使$q(s',a')$值最大的动作的动作价值。在本研究中，具体的状态、动作以及奖励函数设计如下：

本研究只考虑主计划的调度，因此每个订单只有分配以及未分配两种状态，由此，先以一段二进制编码表示所有订单的分配状态，“0”代表未分配，“1”代表已分配，订单在二进制编码中的位置代表订单编号。

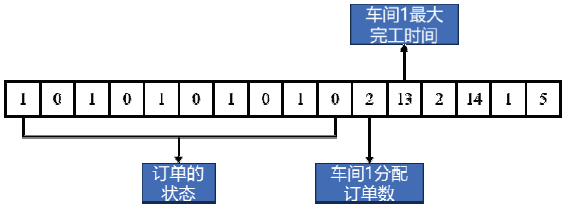

图 3.1 状态空间编码

Fig 3.1. State space encoding

车间有 2 个关键环境因素，一个是当前已被分配订单数，第二个是当前最大完工时间，被分配的具体订单可以通过完工时间以及订单编码耦合推断，因此可不考虑。由此，以位数为 2 的编码表示每个车间的状态，编码的第一位代表当前车间已被分配订单数量，第二位代表当前的最大完工时间，然后将每个车间的编码按车间编号顺序连接，组成所有车间的状态编码，最终的状态编码如图 3.1 所示。

在使用强化学习算法研究车间调度问题时，一般将启发规则作为动作进行调度，但由于本研究的研究对象一个调度周期内的订单数不太多，因此，直接采取订单数N_o及车间数N_W的笛卡尔积作为动作空间，每个订单占据连续的 3 个动作编号值，并且在分配完某一订单后，屏蔽掉其可能的 3 个动作，最终的动作空间如式（16）所示：

$$A = [0, 1, 2, \ldots, \ldots N_o * N_W] \quad (16)$$

针对本研究进行的多目标优化问题，将奖励函数划分为 3 部分：完工时间奖励 R_1、交货期奖励 R_2 以及车间负载奖励 R_3，计算公式如（17）-（19）：

$$R_1 = -\min\left(1, \frac{\Delta Ti}{Pi}\right) \quad (17)$$

$$R_2 = \begin{cases} -\min\left(0, \frac{T_f - T_{uf}}{T_{uf}} * P_l\right) & T_f > T_{uf} \\ -\min\left(0, \frac{T_{uf} - T_f}{T_{uf}} * P_e\right) & T_{uf} > T_f \end{cases} \quad (18)$$

$$R_3 = 2.5 * e^{-3*\frac{Std}{Mean}} - 0.5 \quad (19)$$

其中，ΔTi为分配完订单i后，全局最大完工时间的增加量；Pi是订单i的实际加工时间。T_f为订单的实际完成时间；T_{uf}为订单的模糊交货日期；P_l为拖期完成惩罚系数；P_e是提前完成惩罚系数。Std和$Mean$分别为三个车间的加工时间的标准差和平均值。

为增加算法的探索性，避免陷入局部最优导致算法过早收敛，每一次 Q 值更新都采用线性衰减 episilon-greedy 策略。同时，通过 experiment buffer 经验回放机制选取训练样本，并采用衰减学习率控制学习速度的大小。相关的超参数设置如表 2 所示。

表 2 参数设定

Tab. 3.1 hyperparameter setting

超参数名称	大小
初始学习率	0.0001
学习率下降速率	0.9/(100 epoch)
初始 epsilon	1
Epsilon 下降速率	0.995
Epsilon 最小值	0.001
Buffer 大小	64
Replay Buffer 容量	10000
奖励折扣率	0.9

最终的算法架构如图 3.1 所示。本研究采用改进的 Dueling DQN 算法，引入优势函数，当状态输入模型后，采用两个神经网络分别估计状态价值与动作价值，再通过聚合层，利用式（20）将动作价值与状态价值合并。

$$q(s,a) = V(s) + \left(A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a')\right) \quad (20)$$

其中，$V(s)$为状态价值估计值，$A(s,a)$为动作a的动作价值估计值，$\frac{1}{|A|}\sum_{a'} A(s,a')$为所有动作的价值估计值的平均值。

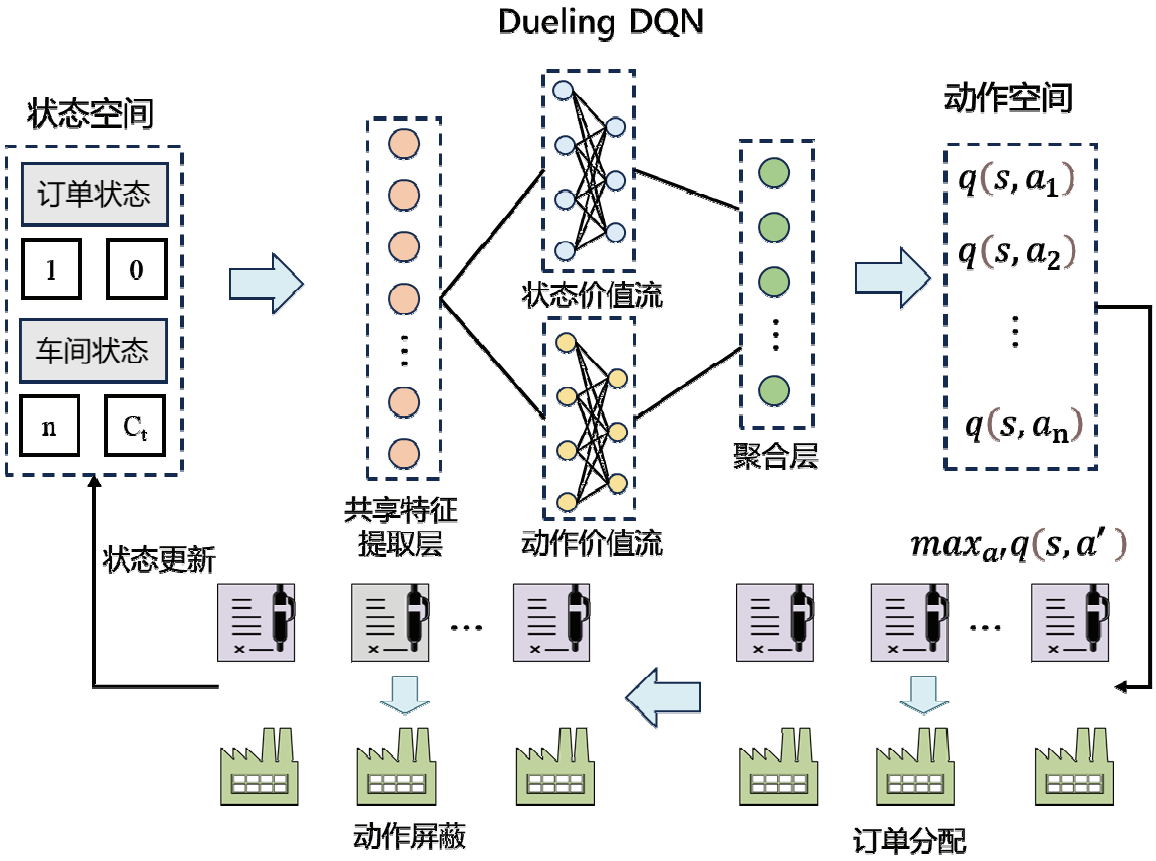

图 3.1 Dueling DQN 架构
Fig 3.1. Framework of Dueling DQN

4 仿真数据验证

为验证改进 Dueling DQN 算法得到有效性，根据某企业实际生产特点，设计了模拟仿真数据对算法进行验证。模拟仿真数据的生成按照如表 4.1 所示的规则进行。算法运行平台为 PC 机，CPU 为 AMD-7950x，GPU 为 Nvidia RTX 4060。

图 4.1 为训练过程中奖励函数的变化，可以发现在训练初期，奖励函数上升较快，随后呈波动趋势。由于改进的 Dueling DQN 算法具有较好的收敛性能，因此当 epoch 的代数超过 1700 次时，奖励已经趋于稳定，最终算法的奖励在 35 左右保持稳定。同时通过观察可以发现，最后稳定的奖励值会略低于训练过程中偶然出现的最高值，这是由于算法前期受 epsilon-greedy 策略的影响，算法主要处于探索阶段，对学习到的经验的利用率较小，因此可能会找到一个奖励值较高的方案。当迭代次数增大后，epsilon 的值减小，算法对经验的利用比较多，从而找到回报更加稳定的路径。由于强化学习的目标是找到一个稳定策略（如纳什均衡），而非追求单次最高奖励，最终稳定的策略会牺牲达到峰值的单次迭代，在长期平均回报上更优，因此最终稳定的奖励函数可能略低于训练过程中出现的最高奖励。

表 4.1 模拟数据生成指标表
Tab. 4.1 hyperparameter setting

仿真数据指标	levels
订单数	10
车间数	3
加工时间（T1，T2，T3）	随机数~(5, 20)
交货截止时间（T1，T2，T3，T4）	随机数~(5, 60)
拖期交货惩罚系数	随机数~(1, 5)
提前交货惩罚系数	随机数~(0, 2)

通过最后训练得到的模型形成的调度方案如图 4.2 所示，该方案的三个目标函数为[71.67，358.91，2.31]。为比较形成的调度方案的有效性，我们利用前期研究形成的改进离散粒子群算法 PDPSO[17]与

Dueling DQN 算法的结果进行了比较，表 4.2 为 PDPSO 算法得到的所有 pareto 解，由表可知，PDPSO 算法求得的 Pareto 解均未能支配 Dueling DQN 算法得到的解，而 Dueling DQN 算法得到的解支配了 PDPSO 算法得到的第 5 个 pareto 解，说明 Dueling DQN 算法效果与 PDPSO 算法相当，甚至优于 PDPSO 算法。

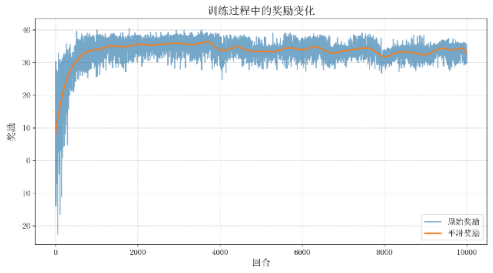

图 4.1 训练过程中奖励函数变化趋势图
Fig 4.1. Framework of Dueling DQN

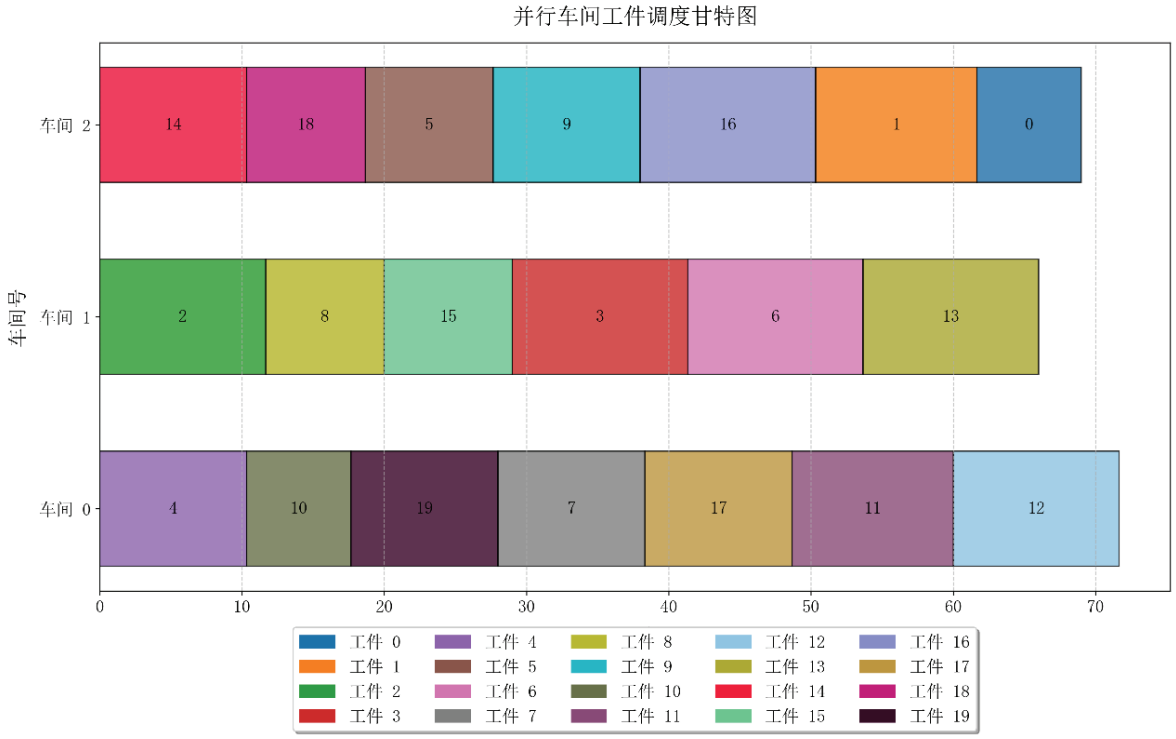

图 4.2 Dueling DQN 算法形成的调度方案
Fig 4.2. Framework of Dueling DQN

表 4.2 改进 PDPSO 算法求解结果
Tab. 4.2 hyperparameter setting

序号	F1	F2	F3
1	73.67	269.02	11.58
2	80.67	242.14	74.10
3	73.33	282.49	31.28
4	71.33	497.96	3.06
5	72.33	473.54	23.73

5 总结与展望

本文建立了一个具有模糊处理时间和截止日期的三目标并行车间调度数学模型，并开发了一种基于深度强化学习的 Dueling DQN 算法。生成随机实例来测试所提出的数学模型，并将所提出的 Dueling DQN 算法与前期研究形成的改进离散粒子群算法 PDPSO 进行比较。结果表明，所提出的 Dueling DQN 算法得到的解支配了 PDPSO 得到的一个 Pareto 解，而 PDPSO 所有的解均未能支配 Dueling DQN 算法的到的解，说明在本文研究的三目标并行车间调度数问题上，Dueling DQN 算法效果与 PDPSO 算法相当甚至更优。

未来的研究会增加模型的复杂性，将每个订单的所有工序的调度均考虑进来，而非目前建立的黑盒模型，尤其是部分涉及批处理的工序，需将组批和拆批策略考虑其中。更重要的是，考虑更多符合实际生产环境的目标，如工作优先级、可重入生成等等，最后，在未来的工作中将提高算法的效率和

精度。

参考文献:

[1] T.C.E. Cheng, C.C.S. Sin, A state-of-the-art review of parallel-machine scheduling research, EUR J OPER RES, (1990) 271-292.Ghosh S,Maiti J. Data mining driven DMAIC framework for improving foundry quality-a case study[J]. Production Planning & Control,2014,25(6).

[2] J. Behnamian, Survey on fuzzy shop scheduling[J], FUZZY OPTIM DECIS MA, 15 (2016) 331-366.

[3] P. Jin, B. Liu, Parallel machine scheduling models with fuzzy processing times[J], INFORM SCIENCES, 166 (2004) 49-66.

[4] S. Balin, Parallel machine scheduling with fuzzy processing times using a robust genetic algorithm and simulation[J], INFORM SCIENCES, (2011) 3551-3569.

[5] S. Balin, Non-identical parallel machine scheduling with fuzzy processing times using genetic algorithm and simulation, The International Journal of Advanced Manufacturing Technology, 61 (2012) 1115-1127.

[6] T.W. Liao, P. Su, Parallel machine scheduling in fuzzy environment with hybrid ant colony optimization including a comparison of fuzzy number ranking methods in consideration of spread of fuzziness, APPL SOFT COMPUT, 56 (2017) 65-81.

[7] 何亚东,邓超,招云芳,等.基于模糊偏好的多阶段多层级模糊调度问题优化[J].计算机集成制造系统, (2025),1-24.

[8] Y. Sun, B. Xue, M. Zhang, G.G. Yen, A New Two-stage Evolutionary Algorithm for Many-objective Optimization, IEEE T EVOLUT COMPUT, (2018).

[9] S. Zhou, X. Li, N. Du, et al., A multi-objective differential evolution algorithm for parallel batch processing machine scheduling considering electricity consumption cost, COMPUT OPER RES, 96,(2018),55-68.

[10] Wen Song, Xinyang Chen, Qiqiang Li, et al., Flexible job-shop schedulingvia graph neural network and deep reinforcement learning, lEEE Trans. Ind.Inform.19(2), (2023),1600-1610.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al., Human-level control through deep reinforcement learning[J], Nature, 518,(2015),529-533.

[12] Yu Du, Junqing Li, Chengdong Li, et al., A reinforcement learningapproach for flexible job shop scheduling problem with crane transportation andsetup times, IEEE Trans. Neural Netw, Learn. Syst,35 (4),(2024),5695-5709.

[13] Yu Du, Junging Li, A deep reinforcement learning based algorithm for adistributed precast conerete production scheduling, Int, J, Prod. Econ. 268,(2024),109102.

[14] Weijian Zhang, Min Kong, Yajing Zhang, et al.,A revised deep reinforcement learning algorithm for parallel machine scheduling problem under multi-scenario due date constraints, Swarm Evol Comput ,92,(2025),101808.

[15] Minghai Yuan, Songwei Lu, Liang Zheng, et al., Distributed heterogeneous flexible job-shop scheduling problem considering automated guided vehicle transportation via improved deep Q network, Swarm Evol Comput, 94, (2025), 101902.

[16] Wenhao Zhou, Fayuan Chen, Xiaoyuan Ji, et al., A Pareto-based discrete particle swarm optimization for parallel casting workshop scheduling problem with fuzzy processing time, Knowl-Based Syst,256,(2022),109872.