Journal of System Simulation

Volume 36 | Issue 3 Article 20

3-15-2024

Intelligent Optimization of Coal Terminal Unloading Scheduling Based on Improved D3QN Algorithm

Baoxin Qin
Guoneng (Tianjin) Port Co. , Ltd, Tianjin 300450, China

Yuxiao Zhang
Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China

Sirui Wu
Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China

Weichong Cao
Guoneng (Tianjin) Port Co. , Ltd, Tianjin 300450, China

See next page for additional authors

Intelligent Optimization of Coal Terminal Unloading Scheduling Based on Improved D3QN Algorithm

Abstract

Abstract: Intelligent decision scheduling can improve the operation efficiency of large ports, which is one of the important research directions for the implementation of artificial intelligence technology in the smart port scenario. This article studies the intelligent unloading scheduling tasks of coal terminals and abstracts them as a Markov sequence decision problem. A deep reinforcement learning model for this problem is established, and an improved D3QN algorithm is proposed to realize intelligent optimization of unloading scheduling decisions by considering the characteristics of high action space dimension and sparse feasible action in the model. The simulation results show that for the same set of random task sequences, the optimized scheduling strategy obviously improves the efficiency compared with the random strategy. At the same time, the trained scheduling strategy is directly applied to the randomly generated new task sequence, and the scheduling efficiency is improved by 5%~7%, which indicates that the optimization method has good generalization ability. In addition, with the increasing complexity of decision models, traditional heuristic optimization algorithms are faced with prominent problems such as difficult modeling and low solving efficiency. This article provides a new idea for studying this kind of problem, which is expected to realize the wider application of deep reinforcement learning-based intelligent decision-making in port scheduling tasks.

Keywords

terminal unloading scheduling, scheduling strategy optimization, intelligent decision-making, deep reinforcement learning, Dueling Double DQN algorithm

Authors

Baoxin Qin, Yuxiao Zhang, Sirui Wu, Weichong Cao, and Zhan Li

Recommended Citation

Qin Baoxin, Zhang Yuxiao, Wu Sirui, et al. Intelligent Optimization of Coal Terminal Unloading Scheduling Based on Improved D3QN Algorithm[J]. Journal of System Simulation, 2024, 36(3): 770-781.

第 36 卷第 3 期 系统仿真学报© Vol. 36 No. 3
2024 年 3 月 Journal of System Simulation Mar. 2024

基于改进D3QN的煤炭码头卸车排产智能优化方法

秦保新[1]，张羽霄[2]，吴思锐[2]，曹卫冲[1]，李湛[2,3*]

(1. 国能(天津)港务有限责任公司，天津 300450；2. 哈尔滨工业大学 航天学院智能控制与系统研究所，黑龙江 哈尔滨 150001；
3. 鹏城实验室 数学与理论部，广东 深圳 518055)

摘要：采用智能化决策排产能够提高大型港口的运营效率，是人工智能技术在智慧港口场景落地的重要研究方向之一。针对煤炭码头卸车智能排产任务，将其抽象为马尔可夫序列决策问题。*建立了该问题的深度强化学习模型，并针对该模型中动作空间维度高且可行动作稀疏的特点，提出一种改进的D3QN算法，实现了卸车排产调度决策的智能优化。*仿真结果表明，对于同一组随机任务序列，优化后的排产策略相比随机策略实现了明显的效率提升。同时，将训练好的排产策略应用于随机生成的新任务序列，可实现5%~7%的排产效率提升，表明该优化方法具有较好的泛化能力。此外，随着决策模型复杂度的提升，传统启发式优化算法面临建模困难、求解效率低等突出问题。所提算法为该类问题的研究提供了一种新思路，有望实现深度强化学习智能决策在港口排产任务中的更广泛应用。

关键词：码头卸车排产；调度策略优化；智能决策；深度强化学习；Dueling Double DQN算法

中图分类号：TP391.9; TP278　　文献标志码：A　　文章编号：1004-731X(2024)03-0770-12

DOI: 10.16182/j.issn1004731x.joss.22-1320

引用格式: 秦保新, 张羽霄, 吴思锐, 等. 基于改进D3QN的煤炭码头卸车排产智能优化方法[J]. 系统仿真学报, 2024, 36(3): 770-781.

Reference format: Qin Baoxin, Zhang Yuxiao, Wu Sirui, et al. Intelligent Optimization of Coal Terminal Unloading Scheduling Based on Improved D3QN Algorithm[J]. Journal of System Simulation, 2024, 36(3): 770-781.

Intelligent Optimization of Coal Terminal Unloading Scheduling Based on Improved D3QN Algorithm

Qin Baoxin[1], *Zhang Yuxiao*[2], *Wu Sirui*[2], *Cao Weichong*[1], *Li Zhan*[2,3*]

(1. Guoneng (Tianjin) Port Co. , Ltd, Tianjin 300450, China; 2. Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin 150001, China; 3. Department of Mathematics and Theory, Peng Cheng Laboratory, Shenzhen 518055, China)

Abstract: Intelligent decision scheduling can improve the operation efficiency of large ports, which is one of the important research directions for the implementation of artificial intelligence technology in the smart port scenario. This article studies the intelligent unloading scheduling tasks of coal terminals and abstracts them as a Markov sequence decision problem. *A deep reinforcement learning model for this problem is established, and an improved D3QN algorithm is proposed to realize intelligent optimization of unloading scheduling decisions by considering the characteristics of high action space dimension and sparse feasible action in the model.* The simulation results show that for the same set of random task sequences, the optimized scheduling strategy obviously improves the efficiency compared with the random strategy. At the same time, the trained scheduling strategy is directly applied to the randomly

收稿日期：2022-11-05　　修回日期：2023-04-24

基金项目：国家自然科学基金(62273122)

作者简介：秦保新(1976-)，男，高工，硕士，研究方向为大型机械智能化，码头设备与生产管理。E-mail：11620065@chnenergy.com.cn

通讯作者：李湛(1987-)，男，副教授，博导，博士，研究方向为智能控制与优化、 机器人智能系统。E-mail：zhanli@hit.edu.cn

generated new task sequence, and the scheduling efficiency is improved by 5%~7%, which indicates that the optimization method has good generalization ability. In addition, with the increasing complexity of decision models, traditional heuristic optimization algorithms are faced with prominent problems such as difficult modeling and low solving efficiency. This article provides a new idea for studying this kind of problem, which is expected to realize the wider application of deep reinforcement learning-based intelligent decision-making in port scheduling tasks.

Keywords: terminal unloading scheduling; scheduling strategy optimization; intelligent decision-making; deep reinforcement learning; Dueling Double DQN algorithm

0 引言

由于经济全球化的程度日益加深，国与国之间的贸易往来增长迅速，对运输效率提出了更高的要求。海运凭借其运量庞大，长距离运输费用较低等优势，在全球贸易中占据着越来越重要的地位，全球90%的国际贸易都由海运来承担。早在2008年，美国的IBM公司就在报告中提出了“智慧城市”概念，通过以信息物理系统为结构框架，结合大数据、物联网以及人工智能等技术，实现全面感知、智能决策、信息共享的新型智能化、自动化港口模式。

目前，散货码头生产调度问题在大多数情况下仍被认为是一个有约束的优化问题，基于运筹学和启发式的智能优化算法目前已经较为成熟，主要方法包括启发式算法[1]、数学规划模型[2]、遗传算法[3]和蚁群算法[4]等。文献[5]针对码头自动堆垛机(automatic stacker crane, ASC)协同作业的过程，建立了总完成时间最小的混合整数规划模型，并利用模拟退火算法和遗传算法求解。针对场桥混合作业模式，文献[6-7]建立了包含时间同步约束的混合整数模型，引入遗传算法用以处理大规模调度问题。文献[8]在考虑联合调度情况下，将调度问题转化为旅行商问题(travelling salesman problem, TSP)，设计了自适应权重系数的邻域搜索算法进行快速求解。文献[9]设计了遗传蚁群算法，求解场桥调度和货位选择的结合问题，使任务完成的总时间最小。文献[10]改进了模拟退火算法，以场桥移动时间和车辆等待时间为目标，建立港口排产优化模型。

对于散货船舶生产调度问题的深度强化学习方法仍有待研究，在其他调度场景下的应用同样具有一定的参考价值。文献[11]研究了离散型制造企业车间的动态调度问题，基于DDPG(deep deterministic policy gradient)算法学习了复杂动态生产过程中的决策方法。对于流水车间调度问题，文献[12]选择每台机器上作业的处理时间作为状态，并提出了一种改进的指针网络学习策略。针对智能制造中的动态作业车间调度问题，文献[13]定义了3个矩阵：作业中操作的处理时间矩阵、作业处理状态矩阵和机器指定矩阵作为状态，并将这3个矩阵输入神经网络作为学习策略。通过将调度问题转化为图，并根据图中节点和边的情况定义状态，较好地考虑了问题的结构特征，并有效地表示了排产环境。通常采用图神经网络(graph neural network, GNN)、卷积神经网络(convolutional neural network, CNN)、图卷积网络(graph convolutional network, GCN)和其他网络来提取问题的有效特征。文献[14]采用析取图来建模车间调度问题，并提出了一种近似策略优化(proximal policy optimization, PPO)来优化GNN。对于自适应作业车间调度问题，文献[15]将生产信息表示为多通道图像，并使用CNN来近似状态动作值函数。对于柔性制造系统的动态调度，文献[16]使用Petri网对问题进行建模，并采用GCN来近似DQN(deep Q-network)中的状态作用值函数。

在港口排产决策领域，文献[17]研究了散货码头控制系统的智能优化方法，文献[18]研究了基于机器学习的散货码头泊位调度系统，文献[19]研究

DOI: 10.16182/j.issn1004731x.joss.22-1320

第 36 卷第 3 期 系统仿真学报 Vol. 36 No. 3
2024 年 3 月 Journal of System Simulation Mar. 2024

了煤炭码头卸车流程的启发式智能优化算法。这些研究工作主要针对泊位调度问题。在这些研究中，针对散货码头自动化滞后、信息化程度低的问题，对散货码头控制系统进行了优化，提高了控制系统的控制效率和可操作性。此外，目前也有一些针对集装箱码头的强化学习研究。例如，文献[20]提出了一种基于强化学习的堆场起重机调度模型，用于解决堆场起重机调度优化问题。文献[21]面向洋山港集装箱码头出口的配载问题，提出基于DQN的多分枝搜索算法以提高堆场设备的利用率。文献[22]研究了码头起重机调度和堆场起重机调度的问题，提出一种基于深度Q网络的方法解决船舶配载规划问题。但与集装箱码头调度相比，问题在于散货码头调度模型往往具有规模大、情况复杂的特点。文献[23]针对具有较多约束的码头装船排产问题，提出了一种基于深度强化学习的算法，建立了散货码头装船流程的强化学习模型，得到了包含取料机和装船机调度在内的智能决策算法。

本文在团队前期研究的基础上，进一步针对卸车排产问题展开研究，提出一种基于深度强化学习的煤炭码头卸车排产调度方法。通过对国能(天津)港务有限责任公司历史数据的分析，对煤炭码头排产调度中涉及的主要过程进行了总结和提取，建立了满足马尔可夫性质的强化学习模型，并确定了其状态空间和动作空间。通过改进D3QN算法进行学习和训练，得到了任务随机到达、堆场状态随机的快速调度方法。

本文的创新和贡献如下：

(1) 建立了港口卸车排产模型，设计了合适的奖励函数，满足长期奖励与优化目标完全等价的条件，保证了获得理论上最优训练结果的可能性。

(2) 改进了 *ε-greedy* 搜索策略，使其执行完全随机选择和可行的随机动作，使神经网络在训练过程中无需额外屏蔽就能满足收敛条件，加快了训练进度。

(3) 在多个随机任务和历史任务中验证了模型的迁移性，能够应对实际工况中不同的任务场景。

1 煤炭码头卸车排产数学模型

本节首先介绍国能(天津)港务煤炭码头堆场分布及煤炭转运流程，以及优化求解时所需的假设条件；码头排产模型所包含的状态信息、动作信息，以及状态转移的过程。然后说明奖励函数定义的方法，以及优化的目标。国能(天津)港务有限责任公司煤炭码头实景如图1所示。

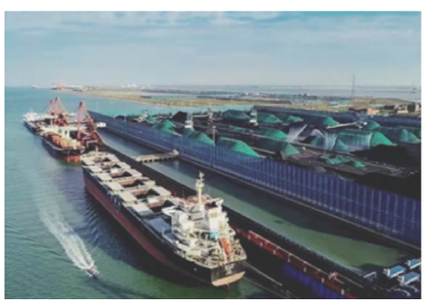


图1 国能(天津)港务有限责任公司煤炭码头实景
Fig. 1 Coal terminal of Guoneng (Tianjin) Port Co., Ltd

1.1 模型的假设

国能(天津)港务煤炭码头堆场分布及卸车煤炭转运流程如图2所示，堆场共分为6条，每条有7个煤堆，共42个煤堆呈点阵形式排列。码头共有4台翻车机和4条堆料线，其中2条堆料线位于2条堆场中间而另2条位于堆场边缘，翻车机负责将进港火车所载的煤炭传送到皮带机上，堆料线上的堆料机负责转运传送带上的煤炭到指定的煤堆。进港火车在缓冲区顺序排列，每个位置上按计划停靠相应的火车，每列火车的车厢数为30~50个，并且都载有相同种类的煤炭。图2为国能(天津)港务煤炭码头堆场分布及排产流程图。

基于卸车调度问题与卸车调度作业的流程，依据天津港务码头进车流程人工排产的历史信息，结合强化学习的特点以及问题求解难度作出如下假设：①每列火车进港时选择一台翻车机，且工作中不会进行切换；②火车进港接受翻车机服务后开始计算在岗时间，直至完成全部煤炭转运；③火车进港后的等待时间和翻车机工作速度为定

值，依旧历史数据设置；④堆料机的移动速度已知且相同，取自现实数据；⑤堆场煤堆的大小相同，排布间距取自实际数据；⑥翻车机与堆料机之间的传送装置默认为是自动化的。

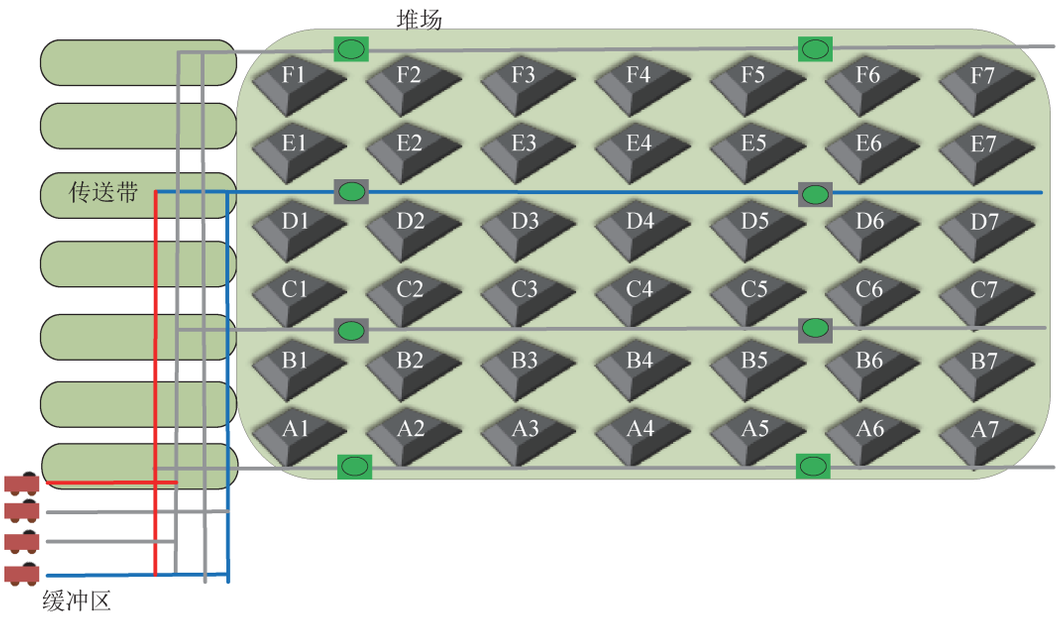


图2 煤炭码头卸车排产流程示意图
Fig. 2 Coal terminal unloading scheduling flow

在上述假设条件下，煤炭码头卸车排产任务是一个动态环境下的连续最优决策过程；在无突发情况的实际排产过程中，系统下一时刻的状态(设备使用状态、煤堆种类数量等)仅与当前时刻的状态和采取的动作有关，严格满足马尔可夫序列决策问题。同时还考虑了实际排产过程中的全部决策内容，在认为传送装置自动化的情况下，建模精度能够满足实际排产决策的要求。

1.2 港口排产模型的数学描述

在此港口模型中，设置满负荷的任务序列，让堆场的设备持续进行调度工作，优化目标是减少排产的总时间，提高设备的利用率。设进港火车有 n 列，记为 $T=\{T_1,T_2,\cdots,T_n\}$，在堆场的4台翻车机与4台堆料机上进行调度作业，所有火车在堆场堆满前都需进行调度安排。码头排产调度策略记作 $\pi=\left(\pi_{ij}^1,\pi_{ij}^2,\cdots,\pi_{ij}^n\right)$，其中，上标代表火车的标号，下标 i,j 分别为翻车机和堆料机的选择，需要配置设备的调度策略，使得所有火车的总完成时间 T_{all} 最小。为方便描述，定义符号如表1所示。

表1 港口卸车流程的状态信息
Table 1 Status information of port unloading process

符号	含义
n	火车总数
F	码头翻车机设备集
D	码头堆料机设备集
k	进港火车的序号
i,j	翻车机与堆料机的序号，i,j =1,2,3,4
π_{ij}^k	第 k 辆火车选择的翻车机和堆料机
A_{ij}^k	$\begin{cases}1, & \text{第}k\text{列火车使用第}i\text{台翻车机}\\0, & \text{存在设备占用情况动作不可行}\end{cases}$
t_i	翻车机 i 工作的剩余时间
t_j	堆料机 j 工作的剩余时间
T^k	第 k 辆火车的作业耗时
T_{all}	所有火车的总完成时间，$T_{\text{all}}=\sum_{k=1}^{n}t^k$

在 π_{ij}^k 策略下，设备占用时间约束条件为

$$t^k \leqslant t_i + t_j$$

第 k 辆进港火车可安排调度的判断条件为

$$\sum_{i=1}^{4}\sum_{j=1}^{4}A_{ij}^k \geqslant 1$$

表示火车至少存在一组可选设备进行调度作业。

系统调度结束的判断条件为

$$\sum_{i=1}^{4} t_i + \sum_{j=1}^{4} t_j \leqslant 0, \text{且} \sum_{k=1}^{n}\sum_{i=1}^{4}\sum_{j=1}^{4} A_{ij}^{k} = 0$$

1.3 港口排产强化学习模型建立

强化学习是通过设置训练目标，让智能体与环境不断交互进行探索和学习，最终可以得到全局最优解。在处理序贯决策类问题时有着非常理想的效果。它根据对象的马尔可夫模型，包含状态空间S、动作空间A、环境奖励函数R、环境状态转移函数P，智能体根据当前状态进行动作的决策，得到环境返回的奖励值，进而更新到下一个状态。

(1) 状态空间的设计：码头卸车任务的调度安排与进港的火车有关，也与当前卸煤流程各个设备占用的情况有关。卸车任务涉及的设备包括4台翻车机，传送带，以及4台堆料机。因此状态空间的信息包括：42个煤堆的种类和数量、堆料机的占用及故障情况、翻车机的占用及故障情况、4台堆料机所对应的煤堆、4条传送带对应的占用时间、4台翻车机对应的堆料机、以及翻车机对应的到港火车。如此设计卸车排产模型均衡了实际的工作情况和模型的可解性。表2为模型状态空间所包含的信息。

表2 港口卸车流程的状态信息
Table 2 Status information of port unloading process

状态空间	数据结构	维数	类型	含义
T_{wait}	数组	4	Int	火车的等待时间
M_{size}	数组	42	Int	42个煤堆的存量
M_{kind}	数组	42	Int	42个煤堆的种类
X_{size}	数组	4	Int	火车运煤量
X_{kind}	数组	4	Int	火车运煤种类
Q	数组	4	Bool	传送带占用情况
Z	数组	4	Bool	翻车机占用情况
Q_i	数组	4	Int	4条传送带对应的煤堆
C_i	数组	4	Int	4列火车对应的翻车机
T^i_{occupy}	数组	4	Int	4条传送带占用时间
Q_z	数组	4	Int	堆料机与翻车机的对应关系

(2) 动作空间的设计：在选择调度动作时，所做的选择包括42个煤堆对应4台堆料机、4台翻车机，共有672种动作组合。动作选取为这三者的组合关系，然而由于条件的限制，大部分的动作处于不可用状态。通过函数来判断动作是否合法：动作对应的车厢煤的种类、煤堆的煤种与车厢运送种类相对应、对应的煤堆存量数量上限、动作对应的翻车机和堆料机均空闲。智能体与堆场环境交互流程如图3所示。

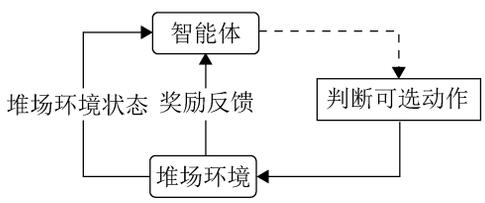


图3 排产调度交互过程
Fig. 3 Scheduling interaction process

(3) 回报函数的设计：对于该模型，将系统的总运行时间作为优化的目标，在每个动作执行时，会令系统的总等待时间增加，记为t，取其负值与一参数系数相乘，记为$-10\times t$，若该动作被占用或非法则设置一个惩罚，取固定值−1，并且保持当前状态不变。另外，考虑到运煤量多的动作占用的时间会更长，因此单步动作的回报函数设计为$r=(km-10\times t)/100$，此时的长期回报为$R=(kM-10\times T)/100$，其中，k为权重系数，m为单步动作运煤量，M为总的运煤量，T为系统的总运行时间。可以看出，当总体需求量相同时，运行的时间越短，则长期回报越大。

2 高维度动作空间稀疏动作下的D3QN算法设计

D3QN算法是一种基于DoubleDQN算法框架和Dueling Q Network网络结构的深度强化学习算法。为了方便将状态的价值与状态−动作值解耦，使得智能体更好地学习到状态与动作对其所获得汇报的影响，该算法在网络结构上做了改进，提出了决斗Q网络结构。

(1) Q-learning算法，是一种著名的基于时序差分的强化学习算法，是无模型强化学习方法的

一种。通过为每一个状态–动作对给予一个Q值，通过贝尔曼方程来实现状态–价值对Q值的更新，最终达到全局最优的结果，表示为

$$Q(s_t,a_t)\leftarrow(1-\alpha)Q(s_t,a_t)+(r_t+\gamma\max Q(s_{t+1},a_{t+1}))$$

式中：$Q(s_t,a_t)$为在状态s_t下采取动作a_t的长期回报，是一个估计的Q值；α为学习率，是更新新老Q值差异的幅度，实验过程中通常将α设置为较小的值；r_t为当前步奖励；γ为折扣因子；$\max Q(s_{t+1},a_{t+1})$为在状态s_{t+1}下能够取得的最大Q值。

(2) DQN算法，是将神经网络应用于强化学习Q-learning的深度强化学习算法，可以被看作是一个神经网络Q与权重函数θ逼近的过程。

DQN算法的损失函数为当前值函数和目标值函数的均方误差，其最小值是DQN算法的优化目标，表示为

$$L(\theta)=E((r_t+\gamma\max Q(s_{t+1},a_{t+1},\theta_{t+1})-Q(s_t,a_t,\theta_t))$$

对损失函数进行求导操作，得到梯度公式：

$$\frac{\partial L(\theta)}{\partial\theta}=E(r_t+\gamma\max Q(s_{t+1},a_{t+1},\theta'))-Q(s_t,a_t,\theta)\frac{\partial Q(s_t,a_t,\theta)}{\partial\theta}$$

(3) Double DQN算法：上述DQN算法只使用一个神经网络，既要用于估计Q值又要估计决策过程中的下一个动作，导致算法的收敛效果不稳定。另外，Q-learning算法包含一个最大化算子max的操作，Q值又会因为环境、近似函数或其他原因而引入噪声，这使得Q值往往会被过估计，其学习目标表示为

$$r_t+\gamma\arg\max_a Q(s_{t+1},a;\theta)$$

而Double DQN算法的核心思想是在估计Q值和预测动作阶段使用两个Q网络，用以去除二者噪声的相关性，其学习的目标为

$$r_t+\gamma Q'(s_{t+1}\arg\max_a Q(s_{t+1},a;\theta);\theta')$$

(4) Dueling DQN算法：将每个动作状态值函数拆分为状态值函数$V(s_t)$加上优势函数$A(s_t,a_t)$。

Dueling DQN：将每个动作状态值函数拆分为状态值函数$V(s)$加上优势函数$A(s,a)$。

$$Q(s_t,a_t;\theta,\alpha,\beta)=V(s_t;\alpha,\beta)+A(s_t,a_t;\theta,\alpha)-\frac{1}{|A|}\sum_{a_{t+1}}A(s_t,a_{t+1};\theta,\alpha)$$

使用该方法是因为：在某些状态s_t，无论做什么动作a_t，对下一个状态s_{t+1}都没过多影响，当前状态动作函数也与当前动作选择不太相关。在这种情况下，Dueling DQN更适用，比DQN学习更快，收敛效果更好。

针对排产调度问题，智能体在进行决策时面临着动作空间的大多数动作不可选的问题，这使得智能体在探索动作空间时难度较大，探索的效率较低，面对这样的动作空间，有2种主流的探索方式：① 设置动作合法性判断条件，强制智能体在合法动作集中进行决策；②为动作设计不同的奖励函数，当智能体选择到非法动作时给予一个较大的惩罚，使得智能体学习到当前状态下的非法动作。

2种方法均有一定的适用范围，但也各存在一些问题，方法1在非法动作较多的情况下探索效率过低，很难学习到比较好的策略；方法2的问题是设计较为困难，智能体学习到策略的好坏受到人为设置的reward的影响很大。鉴于上述存在的问题，本文改进了算法的探索策略，根据当前环境中合法动作与非法动作的比例，设置两个不同的ε参数：ε_1和ε_2，分别在合法动作集和全动作集上进行探索，提高学习策略的效率。

因此Q网络更新的方式为

$$Q^*(s_t;a_t)=\begin{cases}r(s_t,a_t)+\gamma\times\max Q(s_{t+1},a_{t+1}), & \text{合法动作}\\ \gamma\max Q(s_{t+1},a_{t+1}), & \text{非法动作}\end{cases}$$

在开始时，设定了较大的探索率，随着训练次数的增加，按照$\varepsilon=(1-/decay)\times\varepsilon$的速率衰减，其中，$\varepsilon_0=1,decay=15\,000$，并规定$\varepsilon$的最小值为0.02。使得在训练时始终保持一定的探索率。改进后的D3QN算法流程如下：

改进D3QN算法

输入：状态空间S，动作空间A(合法动作集A_{leagal}，全动作集A_{all})，折扣率γ，学习率α

1 初始化经验池D，容量为N

2 随机初始化Q网络的参数ω，当前网络Q_ω，目标网络$Q_{\omega'}$

3 随机初始化目标网络

4 repeat

5 初始化起始状态s

6 repeat

7 在状态s下，基于改进探索选择动作

8 $$a=\begin{cases} rand(A_{\text{leagal}}), & \varepsilon<\varepsilon_1 \\ rand(A_{\text{all}}), & \varepsilon_1<\varepsilon<\varepsilon_2 \\ \pi_\omega & \varepsilon>\varepsilon_2 \end{cases}$$

9 执行动作a_t，观测环境，得到奖励r_t和新的状态s_{t+1}

10 将s_t,a_t,r_t,s_{t+1}存入D中

11 从D中采样s_j,a_j,r_j,s_{j+1}

12 $$Q^*(s_t;a_t)=\begin{cases} r(s_t,a_t)+\gamma\max Q(s_{t+1},a_{t+1}), & \text{合法动作} \\ \gamma\max Q(s_{t+1},a_{t+1}), & \text{非法动作} \end{cases}$$

13 以$(Q^*(s_t,a_t)-Q(s_t,a_t))^2$为损失函数训练Q网络

14 $s_t\leftarrow s_{t+1}$

15 每隔C步，$\omega\leftarrow\omega'$

16 until s_t为终止状态

17 until $\forall s,a,Q(s,a)$收敛

输出：网络Q_ω

3 实验结果及分析

通过对码头卸车人工排产的流程进行分析，得到如下信息：火车运输的煤炭种类中神混48、神混52、神混55、外购55、外石5 000、神混45六种煤炭占据港口的煤炭需求量90%以上，并且实际的港口任务序列具有一定的数量规律，为使得实验更具有一般性，首先选择历史数据中主要的运输煤炭种类生成了随机任务订单进行训练，再通过实际的任务序列进行验证。

选择使用上述这6种煤炭生成随机任务订单，使在港口设备满负荷运转的同时能够覆盖大多数实际工作场景，设计500列火车的任务序列，以模拟近一个月的实际任务，火车运煤种类用随机函数生成，运煤的数量设置为4 000 t左右，则任务的总运煤量为200万t；在堆场的设计上，设置每个煤堆的最大容量为50 000 t，堆场的总容量为210万t，为保证堆场不出现某种煤炭率先装满导致仿真提前终止的情况，通过分析煤炭出港历史数据，在300列火车进港后，每当火车进港卸煤随即减少此种煤炭的堆场存量3 000 t，平均减少在6个相同种类的煤堆上。

本文利用DQN系列深度强化学习算法，设计合理的网络结构，针对此高纬度动作空间稀疏动作的强化学习问题，改进了探索策略，提高了算法的收敛速度和训练效果。采用Double DQN、D3QN算法进行了对比实验。图4~5分别为采用Double DQN算法，利用全连接网络和卷积网络的训练结果。

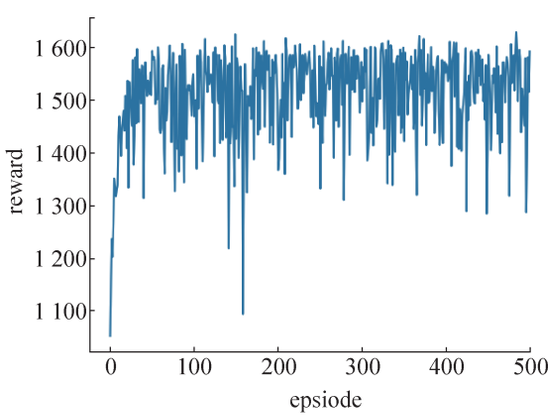


图4 卷积网络训练结果

Fig. 4 Training results of convolutional network

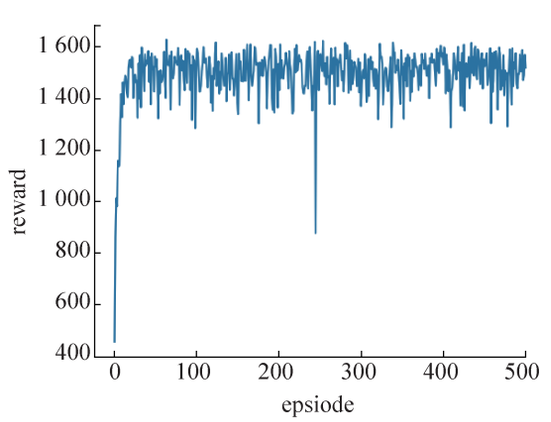


图5 全连接网络训练结果

Fig. 5 Training results of fully-connected network

可以看出，智能体学习到了一定的调度策略，但是排产结果的波动比较大，即在时间范围上的波动较大，效果不够理想。经分析主要原因有以下两点：

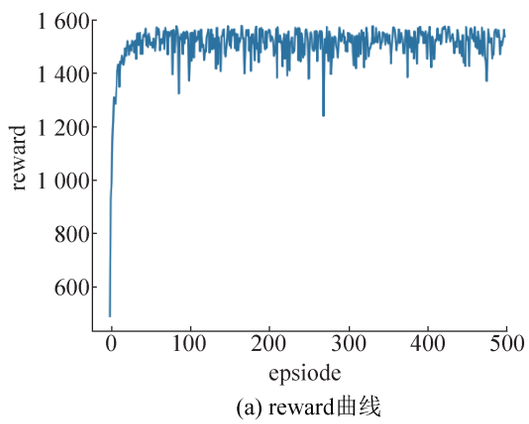
(1) Q网络学习到的是状态–动作价值$Q(s,a)$，然而在此排产问题中，一些状态下采取不同的动作得到的奖励比较接近，这说明此时影响Q值的主要是状态S，通过将与动作无关的状态值和Q值解耦，可以得到更具有鲁棒性的学习效果。

(2) 由于在强化学习的训练过程中，需要不断地使用更新参数后的模型进行下一步的预测，因此不宜使用过于复杂的网络结构，应使用较浅的网络来保证快速拟合。

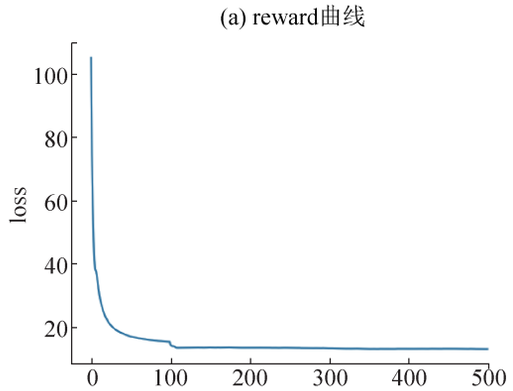
经过算法的改进和网络结构的设计，采用Dueling DQN算法，利用改进的ε-*greedy*策略选择动作。由于在动作选择时，大部分动作处于非法的状态，为提高训练的速度，以探索率ε_1和ε_2分别进行完全随机选择和可行的随机动作选择，提高网络收敛的速度和稳定程度，可以看到，在经过算法改进后的奖励函数曲线明显优于其前两组实验，结果拥有更高的均值以及更好的稳定性。其训练曲线如图6~7所示。

由图6~7可知，在前20个epsiode中reward曲线快速上升，对应到模型可视化图2中可以发现，在训练初期，4台堆料机的移动幅度很大，消耗了大量的时间，随着训练过程的增加，智能体基本掌握了不同煤堆需求的种类以及所处的位置，避免了长距离的移动，这也是reward曲线显著提升的主要原因。在智能体学习的初期，其动作选择近似在符合各冲突约束和状态约束的合法决策动作中进行随机选取，这与经验较少的排产工程师的决策过程类似，经过训练，智能体学习到了在当前随机任务序列中更好的排产顺序，以及设备调度方法，证明构建的模型能够有效学习到翻车机和堆料机调度的策略。

通过绘制训练前后排产过程的甘特图，可以更好地体现训练取得的效果，在前期堆场煤炭数量较少时二者的差别不明显，在经过一段时间的堆煤后，一些煤堆已经堆满，可选的动作进一步下降，在模型训练之前进行调度选择时设备工作的时间占比较低，大量时间花费在了移动上，对应到图8上显示为同一翻车机使用的堆料机频繁变更，设备调度之间等待的时间占比长。



(a) reward曲线



(b) loss曲线

图6 改进D3QN算法训练结果

Fig. 6 Training results of improved D3QN algorithm

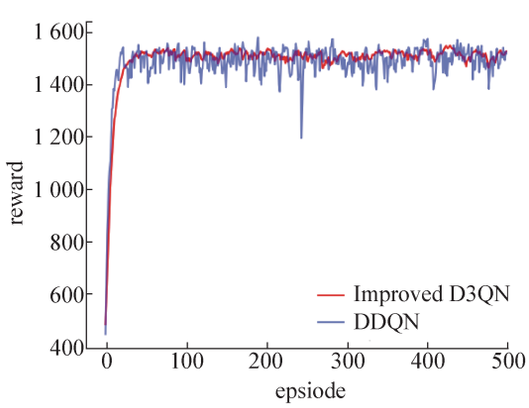


图7 两种算法训练结果均值对比

Fig. 7 Comparsion of mean training results between two algorithms

DOI: 10.16182/j.issn1004731x.joss.22-1320

第 36 卷第 3 期 系统仿真学报 Vol. 36 No. 3
2024 年 3 月 Journal of System Simulation Mar. 2024

若 4 台翻车机满负荷持续工作，则完成堆场的装填需要 260.5 个时间单位，在排产过程中大约有 50 个时间单位花费在等待和调度上，训练前后排产的总时长有了明显的下降。训练前后的排产甘特图如图 8~9 所示，具体优化指标如表 3 所示。

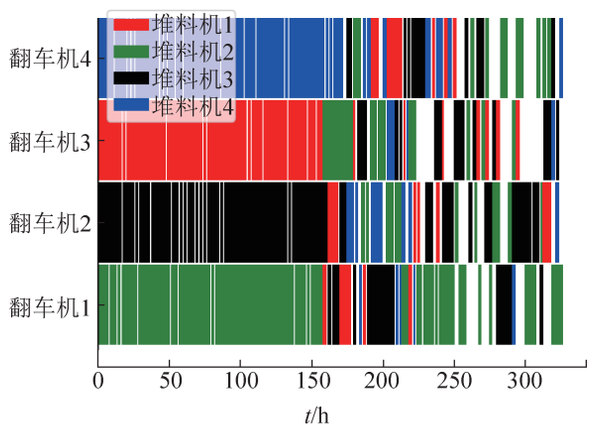


图 8 随机策略排产流程图

Fig. 8 Flow chart of random strategy scheduling

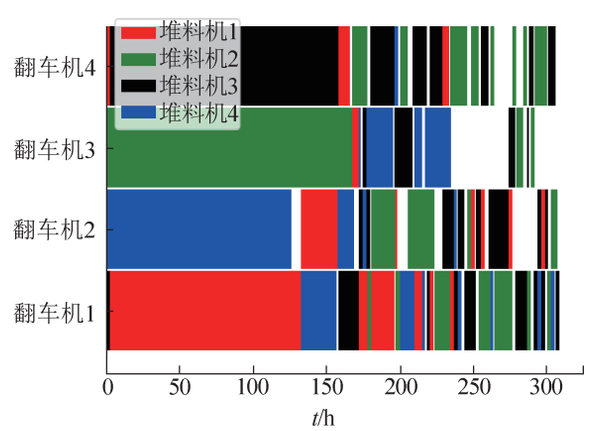


图 9 训练后策略排产流程图

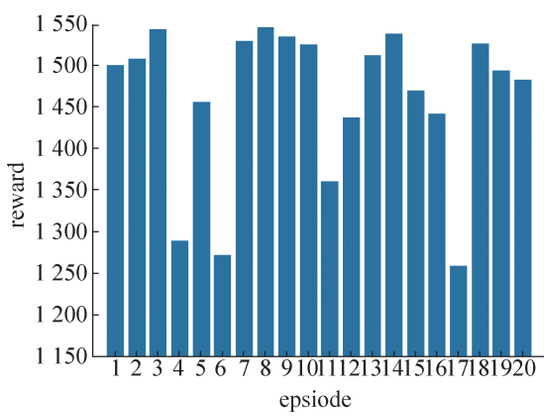
Fig. 9 Flow chart of strategy scheduling after training

生成另外 2 组随机任务序列验证所得的训练结果具有迁移性，在 2 组新遇到的任务中分别进行 20 次排产实验，二者的 reward 曲线如图 10 所示，在部署模型时留有 2% 的探索策略，为方便后续在新的任务进行训练。可以看到，在面对新任务进行排产调度时，总奖励值平均在 1 450 左右，远高于随机的调度策略，证明本文模型算法具有迁移性。训练策略在新任务中的优化指标如表 4 所示。

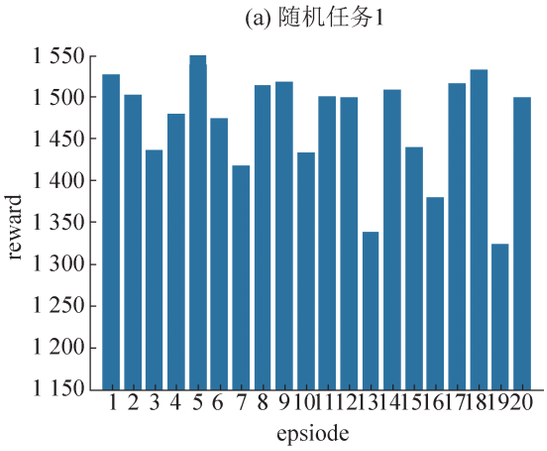
表 3 训练策略的优化指标

Table 3 Optimization index of training strategy

排产策略	非法动作次数	堆料机切换次数	调度总时长/h	优化比例/%
随机	约 1 700	[20,60,58,19]	324.18	基准
训练后	约 10	[16,35,31,16]	300.21	7.5



(a) 随机任务1



(b) 随机任务2

图 10 在随机任务上的排产奖励柱状图

Fig. 10 Bar chart of scheduling rewards on random tasks

表 4 训练策略在新任务中的优化指标

Table 4 Optimization index of training strategy in new task

任务	方法决策	非法动作次数	堆料机切换次数	调度总时长/h	优化比例/%
1	随机策略	基准	[20,60,58,19]	321.21	基准
	训练模型策略	约 10	[16,35,31,16]	301.10	6.7
2	随机策略	基准	[33,55,63,25]	332.45	基准
	训练模型策略	约 12	[21,38,30,18]	309.21	6.4

为进一步验证模型在实际任务序列上的排产效果，本文收集整理了2020年1月份15天中的火车进港任务序列进行实验验证，其内容包括人工排产选取的翻车机线路、火车运煤的种类和数量、以及在港作业时间、等待时间、故障信息等等，其主要信息如表5所示。

表5 港口历史排产信息
Table 5 Port history scheduling information

翻车机	车型	煤种	作业实绩	
			作业时间/h	大票吨
CD2	C80	外购55	111.80	4 320
CD1	C80	神混45	90.38	4 320
CD4	C80	外购55	77.50	4 320
CD3	C64	外石5 000	0	768
⋮				
CD4	C80	外购45	78.63	4 320
CD3	C64	外石5 000	92.78	4 224
CD2	C80	神混45	82.53	4 320

实际排产过程中，港口4台设备全部投入工作，完成全部任务序列的总调度时间为49 441.6 min (824.03 h)；在仿真实验中得到的总排产时间约为691.70 h，优化了16%的设备占用时间和等待时间，其具体的排产结果如图11所示。

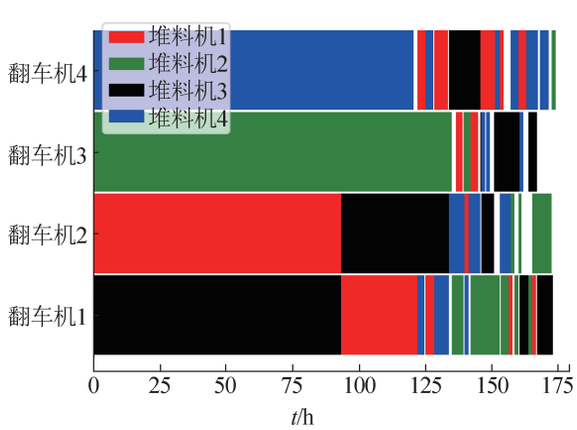


图11 历史数据排产流程图
Fig. 11 Flow chart of historical data scheduling

利用2020年1—6月的人工排产数据与智能排产方法得到的结果进行对比，如图12所示，可以发现，智能排产方法的优化比例在2%~20%之间，其幅度波动受到人工排产策略好坏的影响，具体对比信息如表6所示。

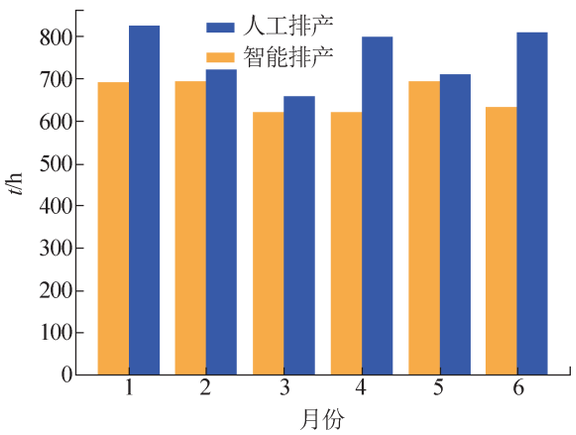


图12 人工排产与智能排产时间对比
Fig. 12 Comparison between manual scheduling time and intelligent scheduling time

表6 人工排产与智能排产时间对比
Table 6 Comparison between manual scheduling time and intelligent scheduling time

月份	t/h		优化比例/%
	人工排产	智能排产	
1	824.03	691.70	16.0
2	721.24	693.79	3.80
3	658.05	620.14	5.80
4	798.50	621.92	22.4
5	710.46	693.23	2.40
6	807.93	632.84	21.7

4 结论

随着海运市场竞争加剧，大型港口由传统人工调度的模式向智能化排产调度转型的需求与日俱增。本文针对大型煤炭码头卸车排产的任务特点和堆场结构，建立了满足马尔可夫性的数学模型，并且基于该模型的特点，改进了D3QN算法及其探索策略，并进行了仿真实验验证。通过与Double DQN算法在不同网络结构模型下的结果进行对比，证明了对算法改进可行有效，与可行动作空间内随机决策策略相比取得了明显效率提升。此外，通过将训练完成的模型在新的随机任务序列和历史数据上进行排产测试，表明该方法具有较好的泛化能力。

DOI: 10.16182/j.issn1004731x.joss.22-1320

参考文献：

[1] Daniela Ambrosino, Anna Sciomachen, Elena Tanfani. A Decomposition Heuristics for the Container Ship Stowage Problem[J]. Journal of Heuristics, 2006, 12(3): 211-233.

[2] Daniela Ambrosino, Anna Sciomachen, Elena Tanfani. Stowing a Containership: The Master Bay Plan Problem [J]. Transportation Research Part A: Policy and Practice, 2004, 38(2): 81-99.

[3] Todd D S, Sen P. A Multiple Criteria Genetic Algorithm for Containership Loading[C]///Proceedings of the 7th International Conference on Genetic Algorithms. [S. l.]: [s.n.], 1997: 674-681.

[4] 卫家骏. 集装箱船智能配载研究[D]. 大连: 大连海事大学, 2012.
Wei Jiajun. The Research on Container Ship's Intelligent Stowage[D]. Dalian: Dalian Maritime University, 2012.

[5] Dirk Briskorn, Simon Emde, Nils Boysen. Cooperative Twin-crane Scheduling[J]. Discrete Applied Mathematics, 2016, 211: 40-57.

[6] 魏晨, 胡志华, 高超锋, 等. 自动化集装箱码头堆场内双起重机调度模型与算法[J]. 大连海事大学学报, 2015, 41(4): 75-80, 89.
Wei Chen, Hu Zhihua, Gao Chaofeng, et al. Scheduling Model and Algorithm of Twin Synchronized Stacking Cranes in Stack Yard of Automated Container Terminal [J]. Journal of Dalian Maritime University, 2015, 41(4): 75-80, 89.

[7] 黄继伟, 韩晓龙. 基于遗传算法的自动化集装箱码头双轨道吊协同调度优化研究[J]. 计算机应用与软件, 2018, 35(9): 92-98, 143.
Huang Jiwei, Han Xiaolong. Collaborative Scheduling Optimization of Twin Automated Stacking Cranes in Automatic Container Terminals Based on Genetic Algorithm[J]. Computer Applications and Software, 2018, 35(9): 92-98, 143.

[8] Amir Hossein Gharehgozli, Laporte G, Yu Yugang, et al. Scheduling Twin Yard Cranes in a Container Block[J]. Transportation Science, 2015, 49(3): 686-705.

[9] 魏亚茹, 朱瑾. 自动化码头双场桥调度与集装箱存储选位建模[J]. 计算机应用, 2018, 38(4): 1189-1194, 1206.
Wei Yaru, Zhu Jin. Modeling of Twin Rail-mounted Gantry Scheduling and Container Slot Selection in Automated Terminal[J]. Journal of Computer Applications, 2018, 38(4): 1189-1194, 1206.

[10] 初良勇, 李淑娟, 阮志毅. 多箱区多场桥调度优化模型及算法实现[J]. 上海海事大学学报, 2017, 38(1): 37-42.
Chu Liangyong, Li Shujuan, Ruan Zhiyi. Scheduling Optimization Model and Algorithm Implementation of Multiple Container Blocks with Multiple Yard Cranes[J]. Journal of Shanghai Maritime University, 2017, 38(1): 37-42.

[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.
Jiang Jingjing. Research on Jobshop Dynamic Scheduling of Discrete Manufacturig Enterprises Based on Deep Reinforcement Learning[D]. Xi'an: Xi'an University of Technology, 2020.

[12] 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 Ling, Pan Zixiao. Scheduling Optimization for Flow-shop Based on Deep Reinforcement Learning and Iterative Greedy Method[J]. Control and Decision, 2021, 36(11): 2609-2617.

[13] Wang Libing, Hu Xin, Wang Yin, et al. Dynamic Job-shop Scheduling in Smart Manufacturing Using Deep Reinforcement Learning[J]. Computer Networks, 2021, 190: 107969.

[14] Luo Shu, Zhang Linxuan, Fan Yushun. Dynamic Multi-objective Scheduling for Flexible Job Shop by Deep Reinforcement Learning[J]. Computers and Industrial Engineering, 2021, 159: 107489.

[15] Han Baoan, Yang Jianjun. Research on Adaptive Job Shop Scheduling Problems Based on Dueling Double DQN[J]. IEEE Access, 2020, 8: 186474-186495.

[16] Hu Liang, Liu Zhenyu, Hu Weifei, et al. Petri-net-based Dynamic Scheduling of Flexible Manufacturing System Via Deep Reinforcement Learning with Graph Convolutional Network[J]. Journal of Manufacturing Systems, 2020, 55: 1-14.

[17] Wang Xuelin, Shi Hankun. Research on Intelligent Optimization of Bulk Cargo Terminal Control System[J]. Journal of Physics: Conference Series, 2020, 1601(5): 052044.

[18] Alan Dávila de León, Eduardo Lalla-Ruiz, Belén Melián-Batista, et al. A Machine Learning-based System for Berth Scheduling at Bulk Terminals[J]. Expert Systems with Applications, 2017, 87: 170-182.

[19] 高天佑. 输出型煤炭码头卸车生产调度优化模型和方法研究[D]. 武汉: 武汉理工大学, 2014.
Gao Tianyou. Optimization Models and Algorithms for Unloading Scheduling of the Export Coal Terminals[D]. Wuhan: Wuhan University of Technology, 2014.

[20] Fotuhi F, Huynh N, Vidal J M, et al. Modeling Yard Crane Operators as Reinforcement Learning Agents[J]. Research in Transportation Economics, 2013, 42(1): 3-12.

[21] 杨奔, 王炜晔, 赵婉婷, 等. 基于DQN的动态深度多分支

搜索自动配载算法[J]. 计算机工程, 2020, 46(8): 313-320.
Yang Ben, Wang Weiye, Zhao Wanting, et al. DQN-based Automatic Stowage Planning Algorithm Using Dynamic Depth Multi-branch Search[J]. Computer Engineering, 2020, 46(8): 313-320.

[22] Shen Yifan, Zhao Ning, Xia Mengjue, et al. A Deep Q-learning Network for Ship Stowage Planning Problem[J]. Polish Maritime Research, 2017, 24(S3): 102-109.

[23] Li Changan, Wu Sirui, Li Zhan, et al. Intelligent Scheduling Method for Bulk Cargo Terminal Loading Process Based on Deep Reinforcement Learning[J]. Electronics, 2022, 11(9): 1390.

DOI: 10.16182/j.issn1004731x.joss.22-1320