Journal of System Simulation

Volume 36 | Issue 7 Article 14

7-15-2024

Task Analysis Methods Based on Deep Reinforcement Learning

Xue Gong
Naval University Of Engineering, Wuhan 430033, China, gogxue@163.com

Pengfei Peng
Naval University Of Engineering, Wuhan 430033, China

Li Rong
Naval University Of Engineering, Wuhan 430033, China, 33574319@qq.com

Yalian Zheng
State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China

See next page for additional authors

Task Analysis Methods Based on Deep Reinforcement Learning

Abstract

Abstract: In response to the high coupling of task interaction and many influencing factors in task analysis, a task analysis method based on sequence decoupling and deep reinforcement learning (DRL) is proposed, which can achieve task decomposition and task sequence reconstruction under complex constraints. The method designs an environment for deep reinforcement learning based on task information interaction, while improving the SumTree algorithm based on the difference between the loss functions of the target network and the evaluation network, achieving the priority evaluation among tasks. The activation function operation mechanism is introduced into the deep reinforcement learning network, followed by extracting the task features, putting forward the greedy activation factor, optimizing the parameters of the deep neural network, and determining the optimal state of the intelligent agent, thus facilitating its state transition. The multi-objective task execution sequence diagram is generated through experience replay. The simulation experiment results show that the method can generate executable task diagrams under optimal scheduling; and it has better adaptivity to dynamic scenarios compared with static scenarios, showing a promising prospect of widespread application in domain task planning.

Keywords

task analysis, reinforcement learning, evaluation network, greedy factors, coupled tasks, activation functions

Authors

Xue Gong, Pengfei Peng, Li Rong, Yalian Zheng, and Jun Jiang

Recommended Citation

Gong Xue, Peng Pengfei, Rong Li, et al. Task Analysis Methods Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2024, 36(7): 1670-1681.

第 36 卷第 7 期 系统仿真学报© Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

基于深度强化学习的任务分析方法

龚雪[1]，彭鹏菲[1]，荣里[1*]，郑雅莲[2]，姜俊[1]
(1. 海军工程大学，湖北 武汉 430033；2. 武汉大学 水资源与水电工程科学国家重点实验室，湖北 武汉 430072)

摘要：针对任务分析中任务协同交互耦合度高、影响因素繁多等问题，提出了基于序列解耦与深度强化学习的任务分析方法，实现了复杂约束条件下的任务分解及任务序列重构。*设计了基于任务信息交互的深度强化学习环境，基于目标网络与评估网络损失函数间的差值改进SumTree算法，实现任务间的优先级评估；将激活函数运行机制引入深度强化学习网络，提取任务特征，提出贪婪激活因子，优化深度神经网络参数，确定智能体最优状态，从而进行智能体状态转换。通过经验回放生成多目标任务执行序列图。*仿真实验结果表明，该方法能生成最佳调度下的可执行任务图；且相对于静态情景，该方法对动态情景有较好的自适应性，在领域任务筹划中具有良好的推广应用前景。

关键词：任务分析；强化学习；评估网络；贪婪因子；耦合任务；激活函数
中图分类号：E917；TP391 文献标志码：A 文章编号：1004-731X(2024)07-1670-12
DOI: 10.16182/j.issn1004731x.joss.23-0443
引用格式: 龚雪, 彭鹏菲, 荣里, 等. 基于深度强化学习的任务分析方法[J]. 系统仿真学报, 2024, 36(7): 1670-1681.
Reference format: Gong Xue, Peng Pengfei, Rong Li, et al. Task Analysis Methods Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2024, 36(7): 1670-1681.

Task Analysis Methods Based on Deep Reinforcement Learning

Gong Xue[1], *Peng Pengfei*[1], *Rong Li*[1*], *Zheng Yalian*[2], *Jiang Jun*[1]
(1. Naval University Of Engineering, Wuhan 430033, China;
2. State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China)

Abstract: In response to the high coupling of task interaction and many influencing factors in task analysis, a task analysis method based on sequence decoupling and deep reinforcement learning (DRL) is proposed, which can achieve task decomposition and task sequence reconstruction under complex constraints. *The method designs an environment for deep reinforcement learning based on task information interaction, while improving the SumTree algorithm based on the difference between the loss functions of the target network and the evaluation network, achieving the priority evaluation among tasks. The activation function operation mechanism is introduced into the deep reinforcement learning network, followed by extracting the task features, putting forward the greedy activation factor, optimizing the parameters of the deep neural network, and determining the optimal state of the intelligent agent, thus facilitating its state transition. The multi-objective task execution sequence diagram is generated through experience replay.* The simulation experiment results show that the method can generate executable task diagrams under optimal scheduling; and it has better adaptivity to dynamic scenarios compared with static scenarios, showing a promising prospect of widespread application in domain task planning.

收稿日期：2023-04-14 修回日期：2023-06-01
基金项目：国家重点研发计划(2017YFC1405205)；海军工程大学科研发展基金自主立项项目(425317S107)
第一作者：龚雪(1998-)，女，硕士生，研究方向为人工智能与大数据。Email：gogxue@163.com
通讯作者：荣里(1980-)，男，讲师，硕士，研究方向为军事教育训练、舰艇综合试验训练。Email：33574319@qq.com

Keywords: task analysis; reinforcement learning; evaluation network; greedy factors; coupled tasks; activation functions

0 引言

目前，国内外学者对任务分析方法[1]的研究主要是从3个角度出发：①传统的任务解析模型[2-4]；②数学模型解析方法[5-6]；③智能分析方法[7-11]。在传统任务分析算法研究中，文献[12]从任务的细粒度出发，为降低对紧急任务的响应效率、减少等待时间和任务完成时间，在面向多机器人环境中，从动态异构任务的细粒度出发，研究了动态任务分配与调度方法。文献[13]引入任务与体系的关联映射规则，提出了一种基于网络信息体系“两维四网”模型构建及特性分析方法，并以空中突击作战为例，结合任务执行精度等资源冗余度指标，研究了面向任务的网络信息体系方法。文献[14]提出了一种基于深度强化学习(deep reinforcement learning, DRL)的非线性动力学系统自适应轨迹规划方法，在高保真模拟环境中的无人机上进行训练和评估，结果体现了基于DRL的自适应方案的学习曲线、样本复杂度和稳定性。

战术弹道导弹[15-17] (tactical ballistic missile, TBM)的特性是跨度大、精度高、高空高速，为了实现对TBM的拦截，要求在极短的窗口时间内进行分析并规划拦截。然而实际情况中，反TBM规划方案变化大、影响因素复杂，且反TBM在预警探测、目标截获、跟踪识别、火力拦截、杀伤效果评估等作战任务环节中。每一个环节都环环相扣，任意一个环节都不能单独进行，给反TBM的任务分析带来巨大挑战。任务分析是任务规划的关键步骤，因此，对反TBM的耦合任务分析和分解是十分必要的。

现有研究成果从任务细粒度出发，并通过任务协同关系的定量分析得出任务执行序列。上述算法在一定程度上能解决任务分析问题，但仍存在许多缺陷，如针对数学解析模型，该模型由于难以考虑多方面的任务交互信息，因而局部最优[18-19]；智能分析方法[20]存在依赖于初始解、参数复杂、迭代时间长、底层存储机能和收敛过早等问题；强化学习算法[21-22]虽可有效解决参数复杂及初始解依赖问题，但算法处理连续性问题的能力较弱，存在实时性较差、时间延迟等问题。

本文针对任务分析中任务信息交互因素复杂和任务间耦合度高的问题，运用深度强化学习算法思路[23-27]，提出基于序列解耦与深度强化学习的任务分析方法：先引入激活函数，动态设计深度强化学习中贪婪激活因子，并结合任务间信息交互的特征改进SumTree算法；通过SumTree算法评估的任务优先级，不断更换智能体执行的初始任务，开展深度网络的训练，并将学习经验存放至经验池中；通过经验回放策略对任务序列进行重构。

1 基于序列解耦的任务环境设计

本文针对任务分析中任务信息交互因素复杂的问题，提出可处理实际应用场景的基于深度强化学习[28]的序列解耦任务分析环境设计方法。任务环境是指针对任务间的信息交互，设计agent的可达路径，即agent可执行的任务序列。为解决任务间耦合度过高以及任务信息交互因素复杂的问题，设计agent的序列解耦运行环境，以期降低任务耦合度及任务可执行的困难程度。基于深度强化学习多目标任务分析算法在DQN(deep Q-Learning)算法[29]的基础上引入了任务序列池来更新经验池中的数据，使得DQN算法[29-31]在任务分析领域得到应用。本文算法流程如图1所示。

DOI: 10.16182/j.issn1004731x.joss.23-0443

第 36 卷第 7 期 系统仿真学报 Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

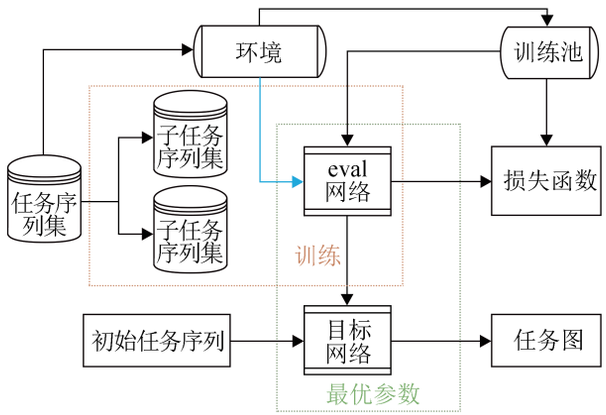


图1 基于深度强化学习多目标任务分析算法流程图
Fig. 1 Flow of multi-objective task analysis algorithm based on DRL

1.1 状态–动作空间定义

本文将一个任务视为一个状态，一个任务向另一个任务进行转接时，需要将转接的下一个任务视为当前任务的动作。因此根据任务信息矩阵，在知晓目标的前提下，寻找最优的任务执行序列图即为该智能体的目标。通过设计状态–动作矩阵，让智能体的进行策略性的选择，适当对智能体进行奖励和惩罚，以激励智能体不断和任务环境交互，从而在探索出的任务图中寻得最优。

(1) 状态–动作矩阵设计

状态–动作矩阵中每一行表示智能体的每一个状态，每一列表示智能体的每一个动作，则(状态，动作)值表示任务与任务直接的信息交互程度，例如，状态动作矩阵$\boldsymbol{I}(1,2)=0.7$表示该智能体在第一个任务，且智能体的状态是1的时候采取的动作2也即智能体将要执行任务2的时候，任务1和任务2的关联程度为0.7。

(2) 状态的选择

智能体通过深度强化学习进行动作选择，执行状态转移函数，状态转移函数关联任务信息矩阵和状态–动作矩阵，进行实时变换状态。

(3) 任务信息交互矩阵

任务信息矩阵表示任务间的无效时间耗用关系，若是任务间有耦合性则无效时间耗用就长，反之则短。该矩阵是一个动态变化的矩阵，在智能体不断进行参数更新的过程，随着奖励矩阵的更新，信息交互矩阵也随之更新，可以说，信息交互矩阵为奖励矩阵的扩展。

1.2 动作选择策略

针对ε-*greedy*算法中贪婪因子ε设计，提出了一种基于tanh函数机制的自适应动态调整策略：最初时，将ε设置较低值，随着迭代次数的增加，智能体对任务环境的认知能力变大，此时，对ε数值进行逐步增加，最终得到贪婪激活因子的取值。

贪婪激活因子ε的定义为

$$\varepsilon=\begin{cases}\varepsilon_0 & \\ \varepsilon+\dfrac{1}{(1+\mathrm{e}^{-episode})}, & \varepsilon<\varepsilon_{\max}\ \text{and}\ episode<episode_{\max}\\ \varepsilon_{\max}, & \varepsilon>\varepsilon_{\min}\end{cases} \tag{1}$$

式中：$\varepsilon_{\max}$为贪婪激活因子ε的最大值，$0<\varepsilon_{\max}<1$；ε_0为贪婪激活因子ε初始值，$0<\varepsilon_0<\varepsilon_{\max}$；$eqisode$为整数，代指算法当前迭代次数，$[0,episode_{\max}]$。设置$\varepsilon_{\max}=0.99$，$\varepsilon_0=0.1$，$eqisode_{\max}=30\ 000\ 000$。

1.3 网络结构设计

如图1所示，每隔一定时间将eval网络的参数传递给目标网络，最终通过训练得到最优的参数值。设置eval网络的损失函数为

$$L(w)=E\left[\left(r+\gamma\max Q'(s',a'_\lambda|w')-Q(s,a_\lambda|w)\right)^2\right] \tag{2}$$

式中：$Q(s',a'_\lambda)$为目标网络；w为网络的权值；s为当前状态；s'为下一个状态；a为状态s下的动作；a'为状态s'的动作值；γ为网络折损率；r为智能体与环境交互获得的奖励。

1.4 序列解耦环境设计

在设计深度强化学习的算法运行环境时，假设agent需要对n个任务进行处理，且最终的多个目标任务已知。在初始化时，agent对任务t_k进行判别，若任务t_k与任务t_{k+1}有信息交互，则记t_k为

智能体可达，此时，智能体可由任务t_k到达任务t_{k+1}或由任务t_{k+1}到达任务t_k，即针对任务间的信息交互，智能体能采用策略选择机制寻得最优的任务序列图。若任务t_k与任务t_{k+1}间没有信息交互，则记为智能体无法从任务t_k到达任务t_{k+1}。另外，设计环境奖励矩阵的依据为通过专家评价，若任务t_{k+1}完全依赖于任务t_k的信息输出，则记$r(s,a)$=1；若任务t_{k+1}不完全依赖于任务t_k的信息输出，则记$r(s,a)$处于(0,1)的区间；若任务t_{k+1}完全不依赖于任务t_k的信息输出，则记$r(s,a)$为−1；若是以某个任务为目标，则记$r(s,a)$处于(100,150)的区间，这个阶段是强化学习初期生成智能体执行序列元组的关键阶段。

在上述构建的环境中，因本文智能体执行的动作表示下一个任务，又因智能体当前的状态表示的是当前任务，故图2中A_s为智能体在状态s下所执行的动作。通过任务信息交互矩阵可知两个任务的实际耦合程度。通过设计耦合阈值过滤耦合任务集，根据智能体奖励机制，将智能体执行每一个任务的打分和任务信息交互矩阵进行比对、处理并存入经验池，将经验池的数据标准化后，更新任务信息交互矩阵。

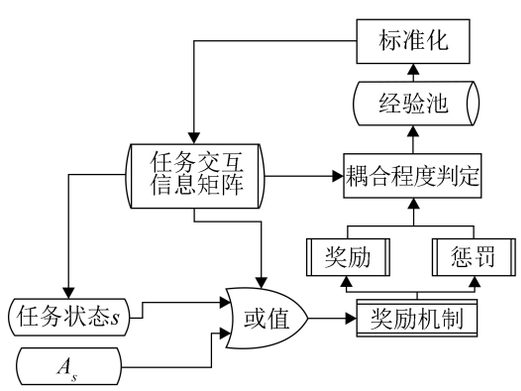


图2 解耦环境设计
Fig. 2 Decoupled environment design

在没有完成任务的情况下，做出如下界定：从最近的任务开始，尽量减少与目标之间的距离，并与最近的任务互动获得相应的行动奖励C。具体表达式为

$$r_t=\begin{cases}C, & \text{执行到最终任务}\\ -C, & \text{遇上死胡同}\\ r+\gamma C, & \text{其他}\end{cases} \tag{3}$$

式中：$\gamma=\alpha\times 1/(1+e^{r_{t-1}})$，$r_{t+1}$为上一个任务的奖励累积；$\alpha$为扰动系数，取值为[0,1]。

2 基于深度强化学习的任务分析方法

2.1 深度强化学习的任务分析方法设计

本文针对强化学习中的经验折损策略，引入激活函数机制，使经验折损率随着训练回合的增加而增加，而非单一的折损系数。经验折损率的动态设置能提升深度神经网络寻优的性能。通过动态设置贪婪因子，使网络搜寻的范围更加宽广，促使智能体的动作选择策略精度得到提升，基于改进高斯变异SumTree算法，动态评估任务优先级，并将任务优先级与网络参数结合，从而得到一个参数训练良好的目标网络。本文的任务分析方法步骤如下。

step 1：获取智能体正在执行的任务序号。

step 2：根据动作效用函数的计算方法，除当前任务外，计算其他任务的可能Q值。

step 3：利用贪婪策略挑选下一个可执行的任务。

step 4：智能体执行该任务。

step 5：将智能体执行的任务序列进行记忆存储。

step 6：进行样本批量训练。

step 7：将训练好的网络参数进行应用，通过经验回放，得到可执行任务序列图。

2.2 基于高斯变异的任务优先级的SumTree算法构建

高斯变异的SumTree算法的结构如图3所示，图中所示为二叉树结构，二叉树算法结构能提升子任务序列的重复利用率。在该算法中，任务序列样本代表叶子节点，且通过高斯变异处理叶子

第 36 卷第 7 期 系统仿真学报 Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

节点样本以增加抽取的随机性，叶子节点的优先级一般通过*TD-error*[32]确定。本文基于指数函数放大机制，将*Q_target*网络和*Q_eval*网络的*Q*值之差替代*TD-error*，以此确定叶子节点的优先级。*Q_target*网络和*Q_eval*网络的*Q*值相关性越不明显，则模型的预测精度越差，则被训练的价值就越大，因此节点优先级越高。高斯变异的SumTree算法优先级表示为

$$P=N\times\alpha+(b-b_{\mathrm{f}})\times d(b,b_{\mathrm{f}})\times \exp(-(b_{\mathrm{best},n-1}-b)/\varsigma^2(t)) \tag{4}$$

式中：n为当前叶子节点的序号；N为所有叶子节点之和；b_{f}为b的父节点；$d(b,b_{\mathrm{f}})$为叶子节点到父节点的欧式距离；$b_{\mathrm{best},n-1}$为上次训练最优的叶子节点；α为多样性指标，需根据求解模型自行设置；$\varsigma^2(n)$为高斯变异幅度，$\varsigma^2(n)=\varsigma_0^2\exp(-(b/N))$。

3 基于序列解耦与深度强化学习的任务分析方法

在训练过程中，目标网络的参数通过batch训练不断地更新优化，并通过固定时间节点给目标网络赋参。本文目标任务状态设置为多个，目标网络根据下一个状态的输入给出下一个状态的*Q*值(*Q_next*)，再将*Q_next*代入策略函数得到*Q_target*，将得到的目标任务状态和设置的目标任务状态群进行异化操作，若有一个目标状态与该*Q_target*吻合，则给予agent奖励，并且计算*loss*(*Q_value*,*Q_target*)，将损失反向传播并优化网络参数，即可得到更好的神经网络。经过反复的训练优化后，智能体能基本掌握各种任务状态下的动作决策，训练的主要流程如图4所示。

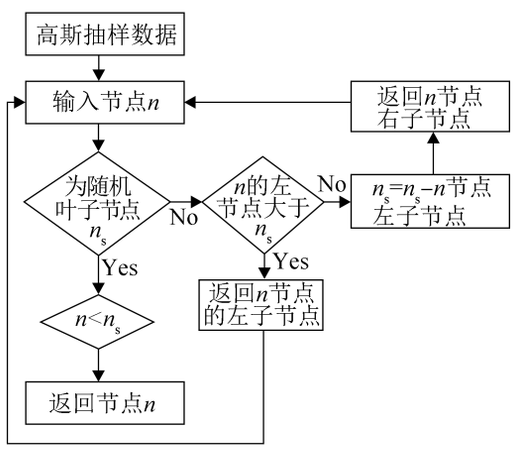


图3 高斯变异SumTree算法流程

Fig. 3 Flow of Gaussian mutation SumTree algorithm

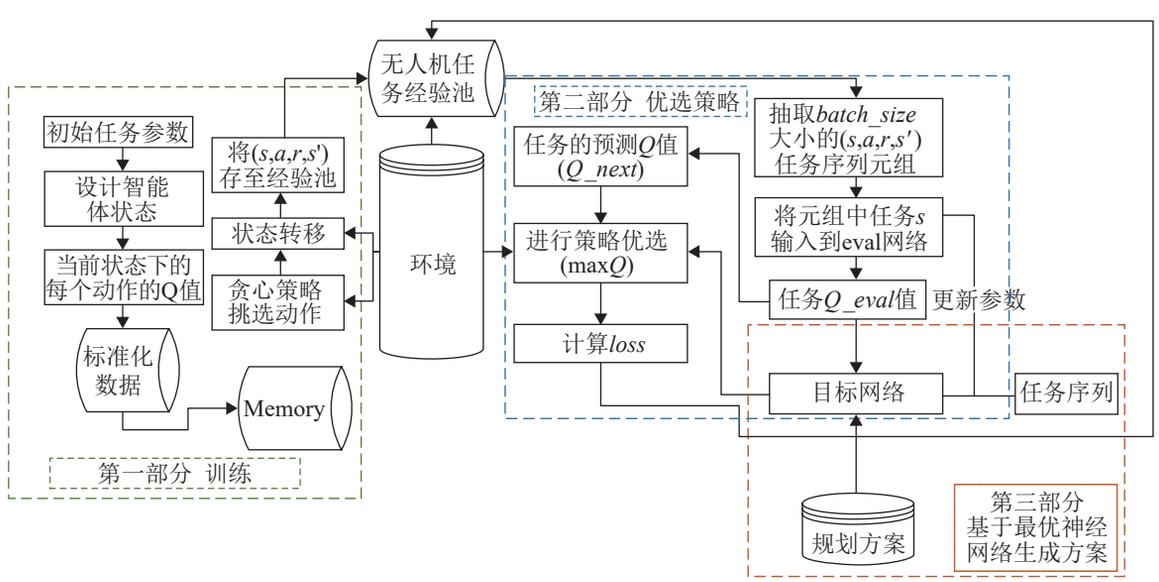


图4 基于改进强化学习的多目标任务分析训练流程

Fig. 4 Training process for multi-objective task analysis based on improved reinforcement learning

基于序列解耦与深度强化学习的任务分析方法步骤如下：

(1) 收集元组

step 1：获取随机任务状态。

step 2：根据 eval 网络，计算除当前任务外，计算其他任务的可能 Q 值。

step 3：基于高斯变异的任务优先级的 SumTree 算法计算当前任务优先级。

step 4：在和环境进行交互的同时，利用贪婪策略挑选下一个可执行的任务。

step 5：对当前任务的优先级进行记忆。

step 6：获得下一个需执行的任务即智能体的动作。

step 7：将任务当前状态的元组进行记忆存储。

(2) 训练网络

step 1：从经验池中抽取任务状态元组样本。

step 2：计算 eval 网络参数，并判断是否已经有一个周期 T，若是则执行 step 3，否则执行 step 4。

step 3：更新目标网络参数。

step 4：更新 Q_eval 值。

step 5：预测 Q 值。

step 6：和环境进行交互，策略优选 Q 值。

step 7：计算 $loss$。

(3) 基于最优神经网络生成任务图

step 1：输入任务序列。

step 2：最优的网络参数进行应用，将任务序列输入目标网络。

step 3：可执行任务序列图。

基于开关的机制，本文设计了 2 个策略函数更新的方式：①智能体在奖励映射很小时的动作空间的选择，②反馈奖励极大时智能体的状态空间的设置。其策略公式为

$$y=\begin{cases} r-\mathrm{e}^{-r}+\gamma\times\max Q, & r<0 \\ r+\gamma\times\max Q, & r>0 \end{cases} \tag{5}$$

$$\gamma=\gamma+1/(1+\mathrm{e}^{T})$$

$$Q=Q_s^{a_s}\times(1-\lambda)\times X+\lambda\times X\times\max A_s+ Q_s^{a_s}\times\lambda\times X+(1-\lambda)\times X\times\max A_s$$

式中：$Q_s^{a_s}$ 为当前任务状态下执行动作 a_s 的 Q 网络的值；X 为二进制函数，若当前总的奖励值小于 0 时取 0，反之取 1；λ=0.98 为折损值；maxA_s 为当前状态下所有动作的概率值中取最大；y 为策略函数；T 为回合数；γ 为每次训练后的折损值，γ 随着回合数的增加而不断降低；r 为每个回合智能体所得的奖励值。

4 实验分析

TBM 具有跨度大、高空高速、RCS 小的特点，所以 TBM 被探测和拦截的时间窗口很短。因此，要求拦截速度快，且拦截成功率高，针对多批次，多方向和多角度的临近饱和攻击的 TBM，本文以 60 枚 TBM 来袭任务事件模拟开展实验分析，将反 TBM 作战任务抽象为 60 个任务[19]，采用基于改进深度强化学习的任务分析算法完成任务序列重组及任务图重构。

4.1 改进深度强化学习的任务分析参数设置

基于深度强化学习的任务分析改进算法的模型参数，具体如表 1 所示。

表 1 仿真实验模型参数表

Table 1 Parameters for simulation experimental model

参数	值	含义
$\gamma_{\max}$	0.98	γ 经验折损率
EXPLORE	30 000,60 000	epsilon 衰减的总步数
BATCH	70	小批量训练样本数
memory_size	5 000	记忆上限
*neuro_layer*1	20	第一层隐藏层
*neuro_layer*2	64	第二层隐藏层
INITIAL_EPSILON	0.01	epsilon 的初始值

DOI: 10.16182/j.issn1004731x.joss.23-0443

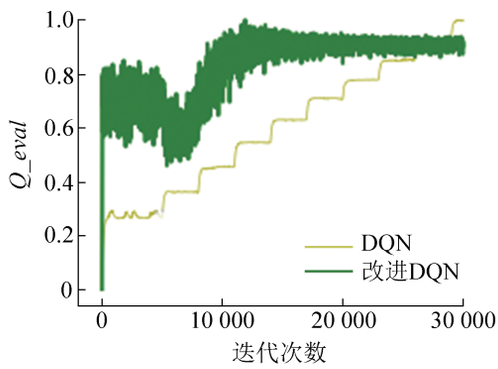
第 36 卷第 7 期 系统仿真学报 Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

4.2 改进深度强化学习的任务仿真实验分析

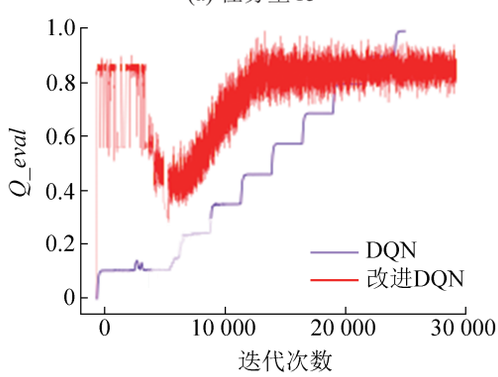
(1) 算法有效性分析

图 5 展示了改进深度强化学习算法和标准深度强化学习算法在不同任务量下的 Q 值分析曲线。可以看出：随着迭代次数增加，任务数量为 15 和 30 时，改进算法在 12 000 多步时达到收敛效果，且在 20 000 步时，*Q_eval* 已经没有很明显的变化，表示智能体成功跳出了局部最优并达到了全局最优，此时网络的训练参数已达最优，说明智能体在环境中已经寻得最优任务执行方案。当任务数量为 60 和 90 时，随着任务数量的增加，任务执行的环境随机性增加，则改进算法收敛速度变慢，都在 50 000 步左右开始震荡减小，并逐渐达到收敛，说明此时智能体已经掌握了环境的变化。而图 5 中，随着迭代和任务数量的增加，标准算法没能得到收敛，说明该算法下的智能体还未掌握环境的变化，还在适应环境，并对任务方案进行探索。因此证明在本文设计的解耦环境中，本文改进深度强化学习算法效率更高，实时性更强，解耦能力更高，且在影响因素繁多的情况下，本文改进深度强化学习算法收敛速度更快，智能体适应环境的速度更好，并最终寻得最优的任务执行方案。

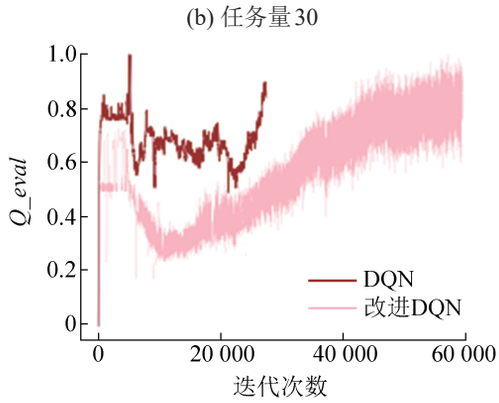
图 6 展示了基于改进深度强化学习在超参数不同情况下的观测值和真实值之间的距离变化。从图中可以看出，当训练样本为 90、经验折损率初始值为 0.85、任务数量为 15 和 30 时，其收敛速度更快，平稳度更高，震荡值较小，收敛效果更优；当任务数量为 60 和 90 时，其平稳性相对较高，但任务数量为 60 时 *Q_eval* 值几乎重叠，但任务数量为 90 时，其 *Q_eval* 值明显变小收敛效果明显更优。因此认为当训练样本为 90，经验折损率初始值为 0.85 时，其效果更优，其 loss 变化值如图 7 所示，随着任务数量的增加，loss 的收敛效果越来越小，相比样本数为 70、经验折损率为 0.9 时更小，因此从 loss 的收敛效果来看，当训练样本为 90，经验折损率初始值为 0.85 时，其效果更优。因此，本文选择的样本规模为 90。



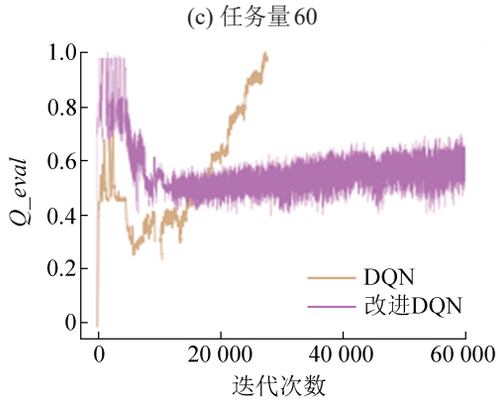
(a) 任务量 15



(b) 任务量 30



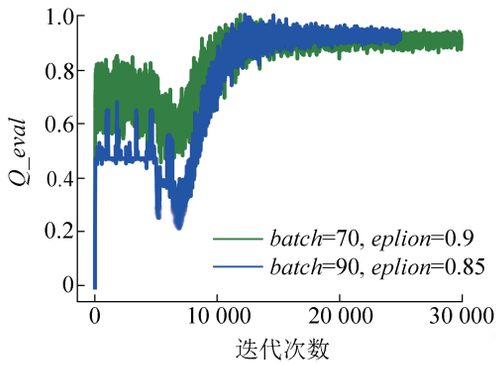
(c) 任务量 60



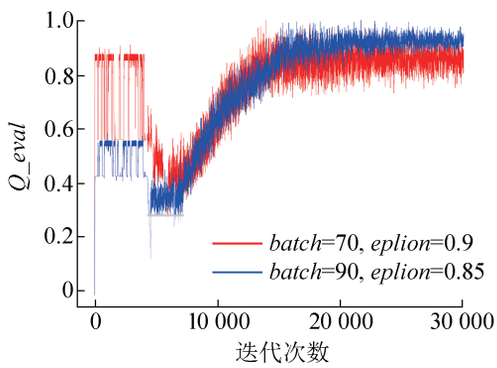
(d) 任务量 90

图 5 两种算法的最终 Q 值分析曲线

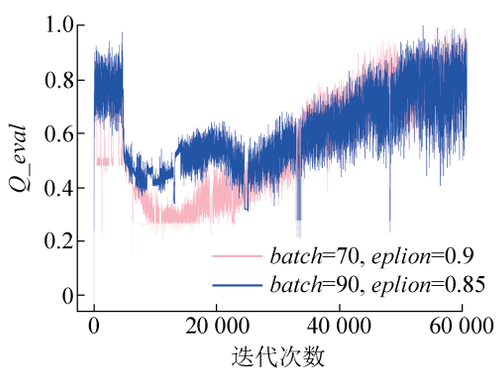
Fig. 5 Final Q-value analysis curves for two algorithms



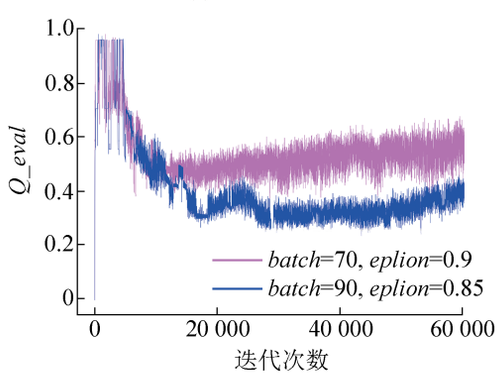
(a) 任务量15



(b) 任务量30



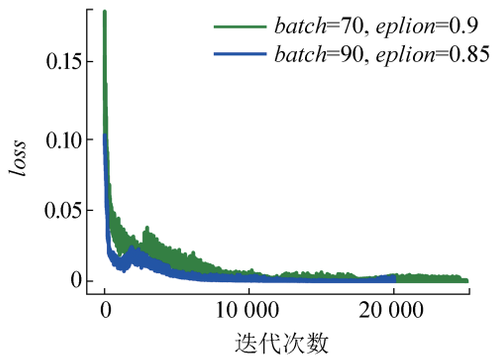
(c) 任务量60



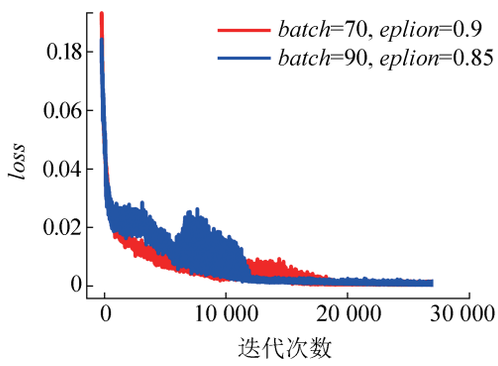
(d) 任务量90

图6　不同超参数的Q值对比

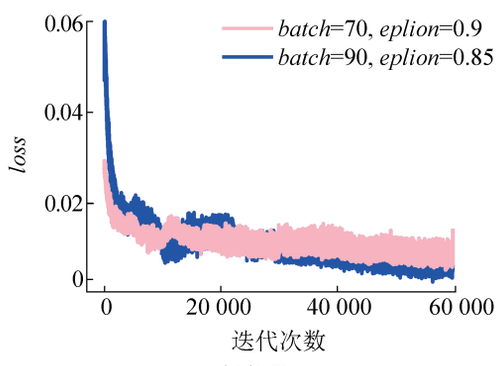
Fig. 6　Comparison of Q values with different byperparameters



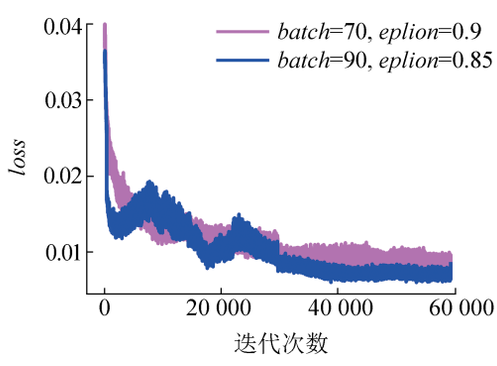
(a) 任务量15



(b) 任务量30



(c) 任务量60



(b) 任务量90

图7　不同超参数的loss值对比

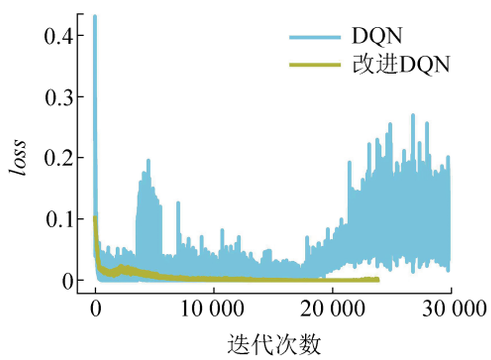
Fig. 7　Comparison of loss values with different byperparameters

https://dc-china-simulation.researchcommons.org/journal/vol36/iss7/14
DOI: 10.16182/j.issn1004731x.joss.23-0443

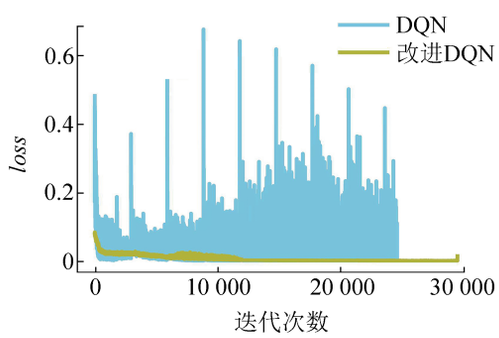
第 36 卷第 7 期 系统仿真学报 Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

图 8 展示了改进深度强化学习和标准深度强化学习算法的 loss 函数随迭代次数的变化。本研究设置的衰减值为 30 000 步，由图 8 可知，改进深度强化学习任务分析算法在 10 000 步时已有收敛趋势，而标准深度强化学习算法在 20 000 步以后才稍显收敛趋势，因而标准深度强化学习算法在进行任务分析时的计算效率较低。分析原因可知，改进深度强化学习任务分析算法能考虑多样化输入，故其收敛速度更快，运行效果更佳。

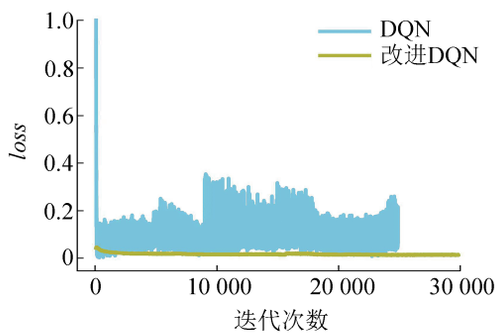
图 9 和 10 分别展示了基于序列解耦与改进深度强化学习算法所得的任务可执性系数图和任务分析图。任务可执行系数由模糊矩阵方法[33]的拓展应用可得，随机从 60 个任务中抽取 30 个任务进行计算，若任务可执行性系数大于 0.5 的数量总和超半数，则判定该任务集可执行。



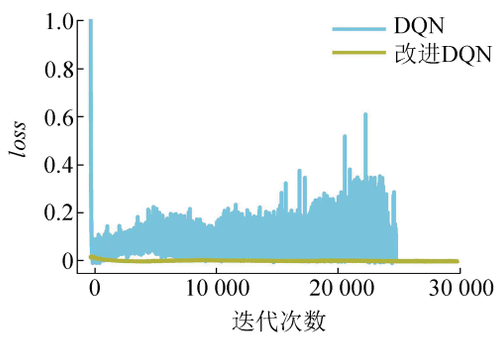
(a) 任务量 15



(b) 任务量 30



(c) 任务量 60



(d) 任务量 90

图 8 两种算法的 loss 迭代曲线

Fig. 8 Loss iteration curves for two algorithms

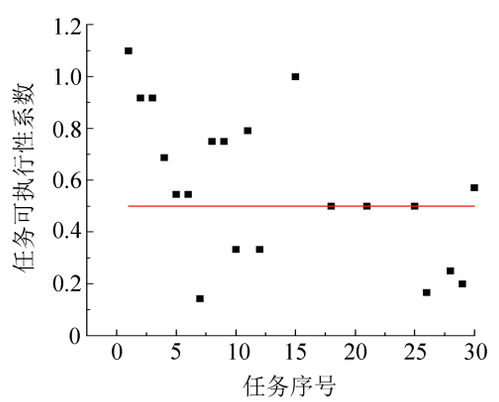


图 9 基于序列解耦与改进深度强化学习算法的任务可执行性系数

Fig. 9 30-task executability factors based on sequence decoupling and improved DRL algerithms

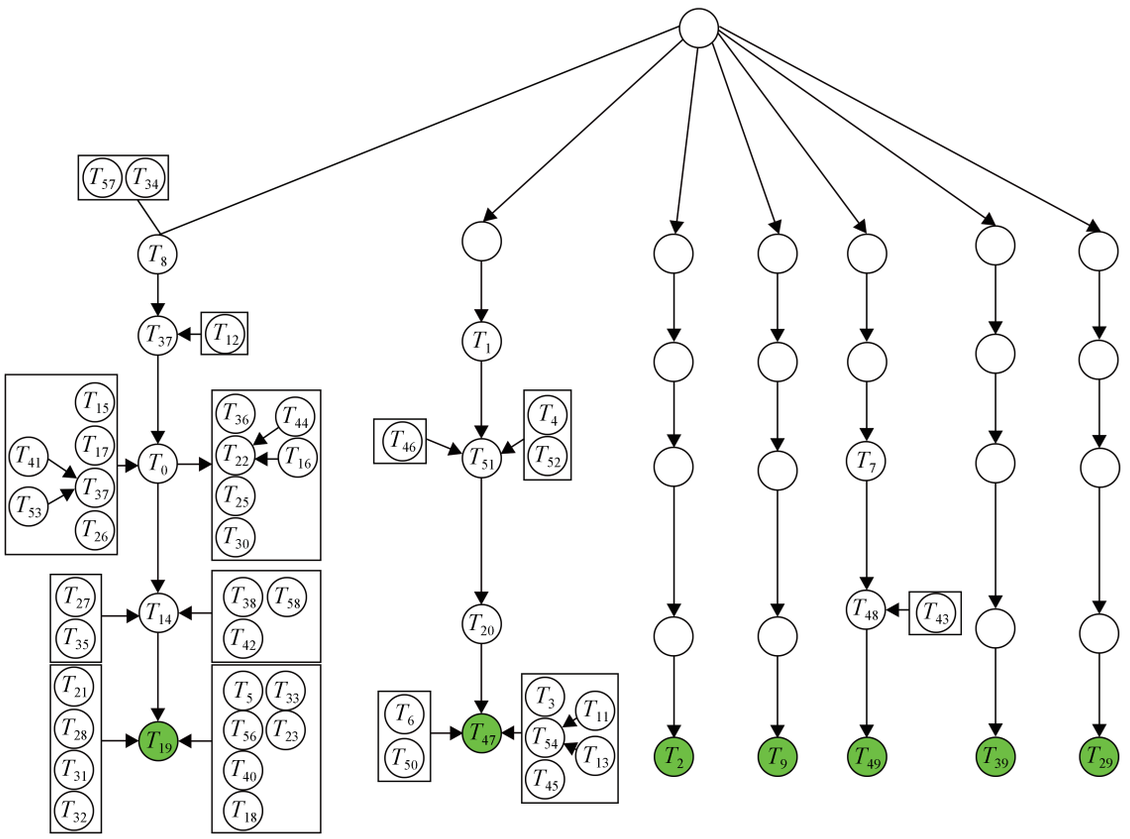


图 10　基于序列解耦与改进深度强化学习算法序列图

Fig. 10　Sequence diagram based on sequence decoupling and improved DRL algerithms

从图 10 可以看出，本文设置的任务目标为绿色填充的圆圈，若将任务执行路径看作为大树，则主干部分是 $T_8 \to T_{37} \to \cdots \to T_{19}$ 和 $T_1 \to T_{51} \to \cdots \to T_{47}$ 等，则支流部分的执行路径为 $T_{21} \to T_{19}$ 和 $T_{34} \to T_8 \to \cdots T_{19}$ 等，因此经过解耦后的任务可从支流到主干再到任务目标，也可从支流直接到任务目标；也即从开始的一个任务目标，分解为多个任务目标，实现支流并行，从而达到解耦合的目的。

5　结论

本文提出基于改进深度强化学习任务分析算法，通过不断地网络调参将训练值存入经验池，而后由经验池再进行网络取值训练，产生针对各个回合的最优网络，最终实现基于改进 SumTree 深度强化学习的回合网络，得到符合不同情景的任务分析图。

仿真实验结果表明，该算法比标准深度强化学习算法更具操作性，且计算效率更高、实用性更强，能快速在一个相对较小的回合数实现网络收敛，产生较好的任务分析图，具备解决任务分析问题的人工智能算法基础。

将深度强化学习和改进 SumTree 算法结合，能一定程度上解决任务分析问题。但算法也存在缺点，如无法克服 Q-learning 的过高估计问题，因此，可考虑将深度强化学习算法和粒子群算法结合，开展进一步的任务规划问题研究。

参考文献：

[1] 马悦, 吴琳, 刘昀, 等. 作战任务优选建模及求解方法研究[J]. 系统仿真学报, 2023, 35(3): 470-483.
Ma Yue, Wu Lin, Liu Yun, et al. Research on Modeling and Solution Method of Operational Tasks Optimization [J]. Journal of System Simulation, 2023, 35(3): 470-483.

[2] 贾正荣, 卢发兴, 王航宇. 基于解耦优化和环流APF的多平台协同攻击任务规划[J]. 北京航空航天大学学报,

DOI: 10.16182/j.issn1004731x.joss.23-0443

第 36 卷第 7 期 系统仿真学报 Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

2020, 46(6): 1142-1150.
Jia Zhengrong, Lu Faxing, Wang Hangyu. Multi-platform Cooperative Task Planning with Decoupling Optimization and Circulating APF[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(6): 1142-1150.

[3] 王晨旭, 王晓晨, 余敦辉, 等. 基于动态解耦的软件众包任务分解算法[J]. 计算机工程, 2019, 45(8): 120-124, 134.
Wang Chenxu, Wang Xiaochen, Yu Dunhui, et al. Software Crowdsourcing Task Decomposition Algorithm Based on Dynamic Decoupling[J]. Computer Engineering, 2019, 45(8): 120-124, 134.

[4] 杨伟刚, 张永永. 2020年以来美国国民警卫队遂行任务解析[J]. 中国军转民, 2021(15): 49-50.

[5] 吴红芳, 任南, 马梦园. 基于FDSM模型的WBS任务耦合问题的研究[J]. 上海管理科学, 2016, 38(6): 76-79.
Wu Hongfang, Ren Nan, Ma Mengyuan. Research on the Coupling Problem of WBS Tasks Based on FDSM Model [J]. Shanghai Management Science, 2016, 38(6): 76-79.

[6] 李永波. 基于解耦子任务的多目标跟踪方法研究[D]. 重庆: 重庆理工大学, 2022.
Li Yongbo. Research of Multi-object Tracking Method Based on Subtask Decoupling[D]. Chongqing: Chongqing University of Technology, 2022.

[7] 邵太华, 陈洪辉, 舒振, 等. 面向无人作战指挥控制的任务智能解析技术[J]. 指挥与控制学报, 2021, 7(2): 146-152.
Shao Taihua, Chen Honghui, Shu Zhen, et al. Mission Intelligent Parsing for Unmanned Combat Command and Control[J]. Journal of Command and Control, 2021, 7(2): 146-152.

[8] 胡云鹏, 彭祺擘, 武新峰, 等. 面向MBSE的航天任务风险分析方法[J]. 网信军民融合, 2022(增2): 23-29.

[9] 罗海龙, 赵得智, 王皓. 面向服务的跨域协同作战任务效费分析[J]. 军事运筹与评估, 2022, 37(3): 57-63.
Luo Hailong, Zhao Dezhi, Wang Hao. Efficiency-cost Analysis of Cross-domain Coordinated Operations Based on Service-oriented Architecture[J]. Military Operations Research and Assessments, 2022, 37(3): 57-63.

[10] 彭鹏菲, 龚雪, 郑雅莲, 等. 基于模拟退火与强化学习机制的任务分析方法[J]. 兵器装备工程学报, 2022, 43(9): 315-322.
Peng Pengfei, Gong Xue, Zheng Yalian, et al. Task Analysis Approach Based on Simulated Annealing and Reinforcement Learning Mechanisms[J]. Journal of Ordnance Equipment Engineering, 2022, 43(9): 315-322.

[11] Ren Jing, Huang Xishi, Huang R N. Efficient Deep Reinforcement Learning for Optimal Path Planning[J]. Electronics, 2022, 11(21): 3628.

[12] 王积旺, 沈立炜. 面向多机器人环境中动态异构任务的细粒度动作分配与调度方法[J]. 计算机科学, 2023, 50(2): 244-253.
Wang Jiwang, Shen Liwei. Fine-grained Action Allocation and Scheduling Method for Dynamic Heterogeneous Tasks in Multi-robot Environments[J]. Computer Science, 2023, 50(2): 244-253.

[13] 朱涛, 梁维泰, 黄松华, 等. 面向任务的网络信息体系建模分析方法研究[J]. 系统仿真学报, 2020, 32(4): 727-737.
Zhu Tao, Liang Weitai, Huang Songhua, et al. Research on Modeling and Analyzing Method of Task-oriented Network Information System of Systems[J]. Journal of System Simulation, 2020, 32(4): 727-737.

[14] Al Younes Y, Barczyk M. Adaptive Nonlinear Model Predictive Horizon Using Deep Reinforcement Learning for Optimal Trajectory Planning[J]. Drones, 2022, 6 (11): 323.

[15] 李龙跃, 刘付显, 赵慧珍. 弹道导弹防御M/M/N排队系统建模与仿真[J]. 系统仿真学报, 2018, 30(4): 1260-1271.
Li Longyue, Liu Fuxian, Zhao Huizhen. Modeling and Simulation of Missile Defense M/M/N Queueing System [J]. Journal of System Simulation, 2018, 30(4): 1260-1271.

[16] 李佳炜, 江晶, 刘重阳, 等. 弹道导弹目标群轨迹建模与仿真[J]. 系统仿真学报, 2020, 32(8): 1515-1523.
Li Jiawei, Jiang Jing, Liu Chongyang, et al. Modeling and Simulation for Target Complex Trajectory of Ballistic Missile[J]. Journal of System Simulation, 2020, 32(8): 1515-1523.

[17] 吴帅, 周晓华, 汪莉莉, 等. 基于实际采样的导弹弹道建模与仿真[J]. 系统仿真学报, 2019, 31(4): 811-817.
Wu Shuai, Zhou Xiaohua, Wang Lili, et al. Modeling and Simulation of Missile Trajectory Based on Practical Sampling[J]. Journal of System Simulation, 2019, 31(4): 811-817.

[18] 王伟, 刘付显. 基于任务关系矩阵的作战任务分解优化[J]. 军事运筹与系统工程, 2017, 31(4): 9-14.

[19] 董涛, 刘付显, 杜菲菲, 等. 基于矩阵的作战任务建模及重组[J]. 工程数学学报, 2013, 30(5): 633-641.
Dong Tao, Liu Fuxian, Du Feifei, et al. Modeling and Reengineering for Anti-TBM Operational Task Based on Matrix[J]. Chinese Journal of Engineering Mathematics, 2013, 30(5): 633-641.

[20] 马悦, 吴琳, 许霄, 等. 智能化作战任务规划需求分析[J]. 指挥控制与仿真, 2021, 43(4): 61-67.
Ma Yue, Wu Lin, Xu Xiao, et al. Requirement Analysis

of Intelligent Operation Task Planning[J]. Command Control & Simulation, 2021, 43(4): 61-67.

[21] 王小康, 冀杰, 刘洋, 等. 基于改进Q学习算法的无人物流配送车路径规划[J]. 系统仿真学报, 2024, 36(5): 1211-1221.
Wang Xiaokang, Ji Jie, Liu Yang, et al. Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm[J]. Journal of System Simulation, 2024, 36(5): 1211-1221.

[22] 胡鹤轩, 钱泽宇, 胡强, 等. 离散四水库问题基准下基于n步Q-learning的水库群优化调度[J]. 中国水利水电科学研究院学报(中英文), 2023, 21(2): 138-147.
Hu Hexuan, Qian Zeyu, Hu Qiang, et al. Optimal Scheduling of Multi-reservoir System Based on N-step Q-learning Under Discrete Four-reservoir Problem Benchmark[J]. Journal of China Institute of Water Resources and Hydropower Research, 2023, 21(2): 138-147.

[23] 唐斯琪, 潘志松, 胡谷雨, 等. 深度强化学习在天基信息网络中的应用–现状与前景[J]. 系统工程与电子技术, 2023, 45(3): 886-901.
Tang Siqi, Pan Zhisong, Hu Guyu, et al. Application of Deep Reinforcement Learning in Space Information Network-status Quo and Prospects[J]. Systems Engineering and Electronics, 2023, 45(3): 886-901.

[24] 宋健, 王子磊. 基于值分解的多目标多智能体深度强化学习方法[J]. 计算机工程, 2023, 49(1): 31-40.
Song Jian, Wang Zilei. Multi-goal Multi-agent Deep Reinforcement Learning Method Based on Value Decomposition[J]. Computer Engineering, 2023, 49(1): 31-40.

[25] Zhou Zhiqian, Zhu Pengming, Zeng Zhiwen, et al. Robot Navigation in a Crowd by Integrating Deep Reinforcement Learning and Online Planning[J]. Applied Intelligence, 2022, 52(13): 15600-15616.

[26] 倪郑鸿远. 强化学习的内在奖励优化方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2021.
Ni Zhenghongyuan. Research on Intrinsic Reward Optimization Method of Reinforcement Learning[D]. Harbin: Harbin Institute of Technology, 2021.

[27] 于航. 基于深度强化学习的多智能体协作学习算法研究[D]. 哈尔滨: 哈尔滨工业大学, 2021.
Yu Hang. Research on Multi-agent Cooperative Learning Based on Deep Reinforcement Learning[D]. Harbin: Harbin Institute of Technology, 2021.

[28] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 3083-3102.
Yan Chao, Xiang Xiaojia, Xu Xin, et al. A Survey on Scalability and Transferability of Multi-agent Deep Reinforcement Learning[J]. Control and Decision, 2022, 37(12): 3083-3102.

[29] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540): 529-533.

[30] 王锦, 张新有. 基于DQN的无人驾驶任务卸载策略[J]. 计算机应用研究, 2022, 39(9): 2738-2744.
Wang Jin, Zhang Xinyou. DQN-based Driverless Task Offloading Policy[J]. Application Research of Computers, 2022, 39(9): 2738-2744.

[31] 刘森, 李玺, 黄运. 基于改进DQN算法的NPC行进路线规划研究[J]. 无线电工程, 2022, 52(8): 1441-1446.
Liu Sen, Li Xi, Huang Yun. Research on Marching Route Planning of NPC Based on Improved DQN Algorithm[J]. Radio Engineering, 2022, 52(8): 1441-1446.

[32] 白辰甲, 刘鹏, 赵巍, 等. 基于TD-error自适应校正的深度Q学习主动采样方法[J]. 计算机研究与发展, 2019, 56(2): 262-280.
Bai Chenjia, Liu Peng, Zhao Wei, et al. Active Sampling for Deep Q-Learning Based on TD-error Adaptive Correction[J]. Journal of Computer Research and Development, 2019, 56(2): 262-280.

[33] 吴雨桐. 产品协同设计任务的排序与调度问题研究[D]. 太原: 太原科技大学, 2017.
Wu Yutong. Study on Task Scheduling and Dispatch in Collaborative Product Development[D]. Taiyuan: Taiyuan University of Science and Technology, 2017.

DOI: 10.16182/j.issn1004731x.joss.23-0443