

基于深度 Q 网络算法的空天地边缘 计算网络资源分配方法

李新春^{1,2}, 孙鹤源¹, 许 驰³

(1. 辽宁工程技术大学 电子与信息工程学院, 辽宁 葫芦岛 125105; 2. 辽宁理工学院 电气与智能控制学院, 辽宁 锦州 121013; 3. 中国科学院 沈阳自动化研究所, 沈阳 110016)

摘 要: 由于卫星、无人机和地面站位置不断变化, 导致空天地边缘计算网络链路不固定, 且网络需要快速响应用户请求, 对吞吐量与实时性的要求较高, 增加了网络资源分配的难度。对此, 本文提出基于深度 Q 网络算法的空天地边缘计算网络资源分配方法。首先, 考虑网络拓扑的动态性和资源异构性, 建立资源间的通信模型, 为资源分配提供基础框架; 然后, 基于最大吞吐量设计资源分配目标函数, 并利用马尔科夫决策模型表述目标函数, 将资源分配问题转化为序列决策问题, 便于在动态变化的网络环境中作出决策; 最后, 基于深度 Q 网络算法求解目标函数, 通过强化学习的方式, 使算法能够通过与环境的交互学习到最优的资源分配策略, 适应网络的实时性和动态性。实验结果表明: 应用该方法后, 网络累计回报较高, 资源任务平均能耗降低, 说明该方法实际可行。

关键词: 空天地一体化网络; 深度 Q 网络算法; 边缘计算; 资源分配; 马尔科夫决策模型

中图分类号: TP273 **文献标志码:** A **文章编号:** 1671-5497(2025)07-2418-07

DOI: 10.13229/j.cnki.jdxbgxb.20240743

Network resource allocation method of aerospace edge computing based on deep Q network algorithm

LI Xin-chun^{1,2}, SUN He-yuan¹, XU Chi³

(1. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China;

2. School of Electrical and Intelligent Control, Liaoning Institute of Science and Engineering, Jinzhou 121013, China;

3. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: Due to the constantly changing positions of satellites, UAVs and ground stations, the space space edge computing network link is not fixed, and the network needs to respond quickly to user requests, which requires high throughput and real-time, increasing the difficulty of network resource allocation. In this regard, this research proposes a network resource allocation method for space edge computing based on deep Q network algorithm. Firstly, considering the dynamic nature of network topology and resource

收稿日期: 2024-07-05.

基金项目: 国家自然科学基金项目(92267108, 62173322); 辽宁省科学技术计划项目(2023JH3/10200004, 2022JH25/10100005).

作者简介: 李新春(1963-), 男, 博士, 高级工程师. 研究方向: 无线传感器网络, 图像处理.

E-mail: s3268247250@outlook.com.

heterogeneity, establish a communication model between resources to provide a basic framework for resource allocation; Then, based on the maximum throughput, a resource allocation objective function is designed, and a Markov decision model is used to express the objective function. The resource allocation problem is transformed into a sequential decision problem, which facilitates decision-making in a dynamically changing network environment; Finally, based on the deep Q-network algorithm, the objective function is solved, and through reinforcement learning, the algorithm can learn the optimal resource allocation strategy through interaction with the environment, adapting to the real-time and dynamic nature of the network. The experimental results show that after applying this method, the cumulative return of the network is higher and the average energy consumption of resource tasks is reduced, indicating that this method is more suitable for practical applications.

Key words: integrated network of air, space, and space; deep Q-network algorithm; edge computing; resource allocation; markov decision model

0 引言

作为未来网络架构的重要组成部分,空天地一体化网络逐渐成为相关领域的研究热点^[1,2]。然而,面对网络资源的有限性和应用需求的多样性,如何合理分配空天地边缘计算网络中的资源,以最大化网络性能和用户体验,成为一个亟待解决的问题。

因此,在网络资源分配方面,张航等^[3]首先在无线供能的环境下,建立了关于匹配决策和资源分配的优化模型,并基于交替优化和匈牙利算法配合贪婪算法求解该模型。最后,通过改进后的求解算法实现网络资源分配。但是,该方法缺乏对网络动态变化的鲁棒性,在实际应用中,如果网络环境和用户需求频繁变化,会影响该方法的应用性能;李云等^[4]建立了一个去中心化的移动边缘计算服务编排和计算卸载模型,在算力、存储、带宽等多维资源约束下,以最小化任务处理时延为目标,将服务缓存和算网资源分配联合优化问题抽象为部分可观测马尔可夫决策过程,并引入长短时记忆网络(LSTM)来捕获时间相关的网络状态信息。最后,基于循环多智能体深度强化学习实现资源分配。但是,该方法在考虑服务请求的时间依赖性和与服务缓存间的耦合关系时,引入LSTM可能增加计算的复杂性。文献[5]在建立多层资源分配模型后,以最大限度地降低网络延迟为目标设计函数,以此实现资源分配。然而,当网络规模扩大或场景变化时,该方法的性能和效率可能受到影响。叶迎晖等^[6]在节点能量因果约束下,构建了以计算时延之和最小化为目标的

多维资源分配模型,之后引入松弛变量和辅助变量解耦优化变量,并基于二分法的迭代算法来求解该模型,实现资源分配。但是,该方法在节点能量因果约束下实行优化可能限制方法的灵活性和可扩展性。

深度Q网络(Deep network, DQN)作为深度学习与强化学习结合的产物,近年来在多个领域展现了强大的学习和决策能力。这种能力使得DQN在解决复杂、动态环境下的决策问题方面具有独特的优势,故展开基于深度Q网络的空天地网络资源分配。因此,为了解决上述方法中存在的问题,本研究提出了基于深度Q网络算法的空天地边缘计算网络资源分配方法。

1 空天地边缘计算网络资源分配方法

1.1 资源分配的目标函数设计

考虑到空天地边缘计算网络中卫星、无人机和地面站位置的动态变化,以及网络资源的异构性,为了给后续的资源分配提供一个稳定的基础框架,建立一个能够准确描述这些变化和资源特性的通信架构。以该架构为基础,将最大吞吐量作为资源分配的目标函数,确保网络能够高效地处理用户请求。

空天地一体化网络是一种融合空基、天基和地基网络资源来满足用户通信需求的新型通信网络^[7,8]。在空天地一体化网络中,每个域内均设有基站,用于实现通信、信号传输等功能^[9]。

在空天地一体化网络架构中,边缘计算网络可以作为一个分布式计算平台,部署在网络的边

缘,靠近数据源和用户,以提供低延迟和高效率的数据处理能力。这个边缘计算网络通过收集来自各个域(天域、空域、地域)的基站和用户之间的网络环境数据,并实行实时的资源分配策略,实现空中、地面和卫星基站资源的最佳分配。

参考信号接收功率是衡量用户在特定频率上接收到的基站信号功率的一个参变量。在空天地一体化网络中,该参数不仅涉及地面基站与用户之间的通信,还涉及卫星与用户之间的通信。

当卫星向用户提供服务时,用户端的参考信号接收功率取决于多个因素,包括卫星的有效各向辐射功率 E 、路径损耗 P 、大气损耗 A 以及天线增益与噪声温度之比($\frac{G}{T}$),这些因素共同决定了用户从卫星接收到的信号强度,此时,用户端的参考信号接收功率 P_U 为:

$$P_U = E - P - A + \frac{G}{T} \quad (1)$$

对于地面基站和无人机基站,用户端的参考信号接收功率 $P_{U'}$ 为:

$$P_{U'} = W + G - F - P \quad (2)$$

式中: W 为副载波功率; F 为馈线损耗。

在网络密集环境中,同频干扰是一个不可忽视的问题。因此,在设计资源分配目标函数前,需要计算考虑同频干扰时的信道容量,计算过程如下:

$$V = N \log_2(1 + S) \quad (3)$$

式中: V 为下行系统所有资源分配基本单位所占的总宽带; S 为信号干扰比。由此可以得到下行用户的最大吞吐量 $Y_{\max}$ 如下:

$$Y_{\max} = \chi N \log_2 \left(1 + \frac{V \times 10^{\frac{P_U + P_{U'}}{10}}}{\chi \sum_{l=1}^L 10^{\frac{P_0}{10}}} \right) \quad (4)$$

式中: χ 为宽带比例因子; P_0 为同频干扰基站对用户的参考信号接收功率。针对式(4),其应满足 $Y_{\max} \leq V$ 的条件。

以最大吞吐量 $Y_{\max}$ 为目标,设计空天地边缘计算网络资源分配的目标函数如下:

$$f = \max \sum_{u=1}^U \sum_{k=1}^K \frac{T_{u,k} \times Y_{\max}}{u} \quad (5)$$

式中: $u \in U$ 为用户端的数量; $k \in K$ 为基站的数量。该目标函数应满足 s. t. $\sum_{k=1}^K T_{u,k} \leq n_U$ 的约束,这里 $T_{u,k}$ 为 U 个用户收到的资源块; n_U 为满足用

户端的资源块数量。

1.2 基于马尔科夫决策模型描述目标函数

将资源分配问题转化为一个序列决策问题,可以在动态变化的网络环境中作出合理的决策,提高资源分配的适应性和灵活性^[10,11]。为此,本研究通过马尔科夫决策模型来表述资源分配的目标函数,将网络环境的变化、资源状态和用户需求等因素建模为MDP中的状态,以资源分配决策建模为动作,便于在复杂和动态的网络环境中进行决策。

(1) 状态空间

在空天地边缘计算网络资源分配中,利用状态 s^t 描述当前网络环境和资源需求的情况,其包括计算任务数据量 x_n^t 和信道增益幅值的平方 j_n^t ,这些信息对于智能体作出合理的资源分配决策至关重要。

时隙 t 的状态集合为:

$$s^t = \left\{ \frac{j_1^t}{f}, \frac{j_2^t}{f}, \dots, \frac{j_n^t}{f}, \frac{x_1^t}{f}, \frac{x_2^t}{f}, \dots, \frac{x_n^t}{f} \right\} \quad (6)$$

(2) 动态空间

智能体的动作 b^t (即资源分配决策)是系统性能优化的关键。通过精心选择这些动作,可以实现计算资源 φ_n^t 和通信资源 ξ_n^t 的有效利用,提高系统整体性能和任务处理成功率。

时隙 t 的动作集合为:

$$b^t = \left\{ \frac{\varphi_1^t}{f}, \frac{\varphi_2^t}{f}, \dots, \frac{\varphi_n^t}{f}, \frac{\xi_1^t}{f}, \frac{\xi_2^t}{f}, \dots, \frac{\xi_n^t}{f} \right\} \quad (7)$$

(3) 奖励函数

奖励 R^{t+1} 是智能体优化行为的关键反馈。在空天地边缘计算网络资源分配中,将奖励定义为处理成功的任务数量,这直接关联到优化目标——最大化任务处理成功概率。通过这种方式,智能体能够学习到哪些动作能够导致更多的任务被成功处理。

奖励函数 R^{t+1} 的公式表达形式如下:

$$R^{t+1} = \sum_{m=1}^M O_m^t \quad (8)$$

式中: O_m^t 为任务 m 是否成功的指示函数。

在每个时隙,智能体观察状态、选择动作、接收奖励并进入新状态,这个过程不断重复,形成了一系列的状态-动作-奖励过程^[12,13]。通过这些交互数据,智能体不断学习如何更好地分配资源以最大化累积奖励,即任务处理成功的数量。

1.3 资源分配的实现

针对目标函数的状态-动作-奖励过程,采用深度Q网络算法来求解资源分配的目标函数。

与传统的强化学习方法相比,DQN算法能够处理更复杂的网络环境和更大量的数据,具有更强的泛化能力和鲁棒性。在没有先验知识的情况下,DQN算法也可以通过试错学习到最优策略。为此,在确保空天地边缘计算网络中各业务队列长度维持在稳定可控范围内的基础上,通过最大化未来的累积奖励来实现最大化网络的总吞吐量,从而完成资源分配。

假设 η 表示折扣因子,在状态 s^t 下执行动作 b^t 的最优期望累计奖励,即最优Q值函数 $Q_{s,b}$:

$$Q_{s,b} = \max \left(\sum_{t=0}^T \eta^t R_{s,b}^{t+1} \right) \quad (9)$$

式中: T 为时间步的总数。

在DQN算法训练的过程中,通过深度强化学习的方式,可以使算法能够通过与环境的交互学习到最优的资源分配策略^[14,15],从而适应网络的实时性和动态性,提高资源分配的效率和效果。

由于DQN不是直接使用最新的经验来更新网络,而是从经验回放缓冲区中随机抽取一小批样本来实行训练。因此,在实际应用中,不能直接通过式(9)获取最优Q值函数,而是通常设计一个神经网络来表示Q值函数,并通过自适应学习率的优化算法来更新神经网络的参数。

基于均方误差的损失函数可用来衡量目标Q值和当前估计网络预测的Q值之间的偏差。这个偏差被用来通过反向传播和梯度下降来更新估值网络的参数,基于自适应学习率的优化算法的梯度的一阶矩和二阶矩的指数衰减平均值的修正值,来调整网络的最优权重参数。

梯度的一阶矩和二阶矩的指数衰减平均值 D_1 、 D_2 的公式表达如下:

$$\begin{cases} D_1 = \gamma_1(1 - \gamma_1) \times \mu \\ D_2 = \gamma_2 + (1 - \gamma_2) \times \mu^2 \end{cases} \quad (10)$$

式中: μ 为算法梯度; γ_1 、 γ_2 为一阶矩和二阶矩的学习率。对一阶矩和二阶矩的指数衰减平均值展开修正,修正后的值如下:

$$\begin{cases} D_1^* = \frac{D_1}{1 - \gamma_1^t} \\ D_2^* = \frac{D_2}{1 - \gamma_2^t} \end{cases} \quad (11)$$

更新后神经网络的最优权重参数 ω 公式表达如下:

$$\omega = 1 - \frac{D_1^*}{D_2^* + \delta} \quad (12)$$

式中: δ 为常数。

通过深度Q网络算法实现空天地边缘计算网络资源分配的步骤如下:

步骤1 利用深度Q网络算法框架,设计一个神经网络结构作为Q值的近似器,初始化一个用于存储历史经验的经验回放池及模型所需其他参数。

步骤2 在每个时间步,根据 ϵ -greedy策略选择动作 b 并执行,根据式(8)获得的奖励,并观察下一个状态 s^{t+1} 。

步骤3 将最优Q值函数 $Q_{s,b}$ 存储到经验回放池中。

步骤4 从经验回放池中随机抽取一批样本,用于训练DQN算法。通过获取的一阶矩和二阶矩的指数衰减平均值的修正值 D_1^* 、 D_2^* ,利用式(12)更新神经网络的权重参数 ω 。

步骤5 不断重复上述训练步骤,获取深度Q网络的最优权重参数。

步骤6 将训练完成的深度Q网络模型部署到空天地边缘计算网络中。对于给定的网络状态 s ,通过DQN模型计算所有可能动作的Q值。选择Q值最大的动作作为资源分配策略,并执行该策略以分配网络资源。

2 实验与结果分析

为了验证基于深度Q网络算法的空天地边缘计算网络资源分配方法的有效性,设计如下测试。

2.1 测试环境设计

在CPU为Intel Xeon E5-2690 v4的计算机上,选择可支持空天地一体化网络模拟的仿真平台NS-3搭建实验场景,以无人驾驶汽车业务为例,完成资源分配。

实验设备及参数设置如下:

2.1.1 网络设备

交换机为Cisco Catalyst 2960-X;路由器为Cisco 2901。

2.2.2 空天地边缘计算网络参数

网络拓扑:包括10颗卫星、20架无人机、50个地面基站。

数据流量:考虑到无人驾驶汽车需要实时传输高清视频、传感器数据等,假设其数据流量为10 Mbit/s。

时延要求:无人驾驶汽车对时延的要求非常高,因为任何延迟都可能影响车辆的安全性和驾驶决策。假设其端到端时延要求在5 ms以内。

信道参数:通信距离为100 km,频率为2.4 GHz,路径损耗大约为132 dB,噪声功率设置为 -100 (dB·m)/Hz。

2.2 深度Q网络算法的收敛性测试

设置DQN网络训练参数如下:学习率为0.001、经验回放缓冲区大小为1 000 000、折扣因子为0.99、目标训练步数或迭代次数为1 000步。

对比Q网络和深度Q网络算法在完成实现无人驾驶汽车业务的空天地边缘计算网络资源分配中平均奖励曲线来衡量算法的收敛性能,两者收敛曲线如图1所示。

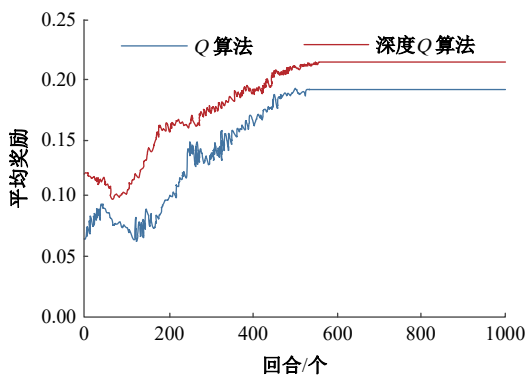


图1 Q网络和深度Q网络算法的收敛曲线对比

Fig. 1 Comparison of convergence curves between Q-network and deep Q-network algorithms

分析图1可知,深度Q网络算法的平均奖励在训练过程中呈现更快的上升趋势,且最终的平均奖励也高于Q算法。这表明深度Q网络算法能够有效学习和优化资源分配策略,以满足空天地边缘计算网络中的高效资源利用需求。

2.3 对比测试

为避免实验结果的单一性,采用本文方法、文献[3]方法和文献[4]方法展开对比测试,从网络资源累计回报、任务平均能耗两个角度来验证本文方法的应用效果。

2.3.1 网络资源累计回报测试

累计回报是一个用来量化边缘计算网络中资源分配效果的综合性指标。它反映了在一段时间内,系统通过合理的资源分配策略所获得的总体

收益或效益。引入累计回报衡量3种方法的空天地边缘计算网络资源分配效果,累计回报越高,表明方法的资源分配效果越好。

采用文献[3]方法和文献[4]方法和本文方法实现对空天地边缘计算网络的资源分配,记录3种方法的累计回报结果如图2所示。

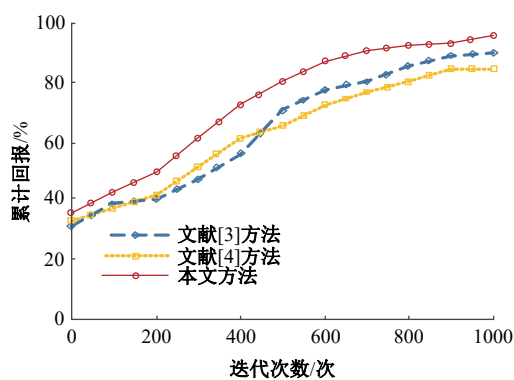


图2 3种方法的累计回报

Fig. 2 Cumulative returns of three methods

分析图2可知,随着迭代次数的增加,3种方法的累计回报都呈现出上升趋势,但本文方法的累计回报在相同的迭代点下明显高于文献[3]方法和文献[4]方法,表明本文方法在空天地边缘计算网络资源分配上能够更有效地利用资源,实现更高效益。

2.3.2 任务平均能耗

引入任务平均能耗衡量3种方法的空天地边缘计算网络的资源分配情况,有效的资源分配方法能够确保任务在执行过程中获得足够资源的同时避免资源的过度使用,从而降低任务的平均能耗。

采用文献[3]方法、文献[4]方法和本文方法完成空天地边缘计算网络资源分配,记录3种方法的任务平均能耗结果,如图3所示。

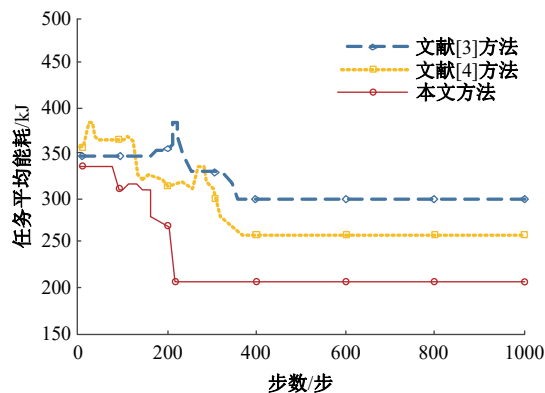


图3 3种方法的任务平均能耗

Fig. 3 Average task energy consumption of three methods

分析图3可知,随着步数的增加,本文方法的任务平均能耗均低于文献[3]方法、文献[4]方法,表明本文方法的空天地边缘计算网络的资源分配效果更好。

3 结束语

空天地边缘计算网络的资源分配是近年来通信领域研究的重点问题,本文提出了一种基于深度Q网络算法的空天地边缘计算网络资源分配方法。首先,建立资源分配目标优化函数;其次,描述该目标优化模型;最后,完成空天地边缘计算网络资源分配。该方法解决了传统空天地边缘计算网络的资源分配方法中存在的问题,具有重要的现实应用意义。

参考文献:

- [1] 徐晓斌,王琪,范存群,等. 面向空天地一体化信息网络的边缘计算资源融合管理方法[J]. 计算机学报, 2023, 46(4): 690-710.
Xu Xiao-bin, Wang Qi, Fan Cun-qun, et al. An aggregated edge computing resource management method for space-air-ground integrated information networks[J]. Chinese Journal of Computers, 2023, 46(4): 690-710.
- [2] 邓平科,张同须,施南翔,等. 星算网络——空天地一体化算力融合网络新发展[J]. 电信科学, 2022, 38(6): 71-81.
Deng Ping-ke, Zhang Tong-xu, Shi Nan-xiang, et al. Computing satellite networks—the novel development of computing-empowered space-air-ground integrated networks[J]. Telecommunications Science, 2022, 38(6): 71-81.
- [3] 张航,唐冬,黄高飞,等. 无线供能协作计算网络中的时延最小化资源分配方案[J]. 计算机应用研究, 2022, 39(1): 214-220.
Zhang Hang, Tang Dong, Huang Gao-fei, et al. Latency minimization resource allocation scheme in wireless powered cooperative computing network[J]. Application Research of Computers, 2022, 39(1): 214-220.
- [4] 李云,高倩,姚枝秀,等. 移动边缘计算中智能服务编排和算网资源分配联合优化方法[J]. 通信学报, 2023, 44(7): 51-63.
Li Yun, Gao Qian, Yao Zhi-xiu, et al. Joint optimization method of intelligent service arrangement and computing-networking resource allocation for MEC[J]. Journal on Communications, 2023, 44(7): 51-63.
- [5] Li Y H, Liu Z L, Tao Q. A resource allocation strategy for internet of vehicles using reinforcement learning in edge computing environment[J]. Soft Computing, 2022, 27(7): 3999-4009.
- [6] 叶迎晖,施丽琴,卢光跃. 无线供能移动边缘网络中计算时延最小化资源分配方法研究[J]. 电子与信息学报, 2022, 44(5): 1839-1846.
Ye Ying-hui, Shi Li-qin, Lu Guang-yue. Execution delay minimization in wireless powered mobile edge computing networks[J]. Journal of Electronics & Information Technology, 2022, 44(5): 1839-1846.
- [7] 杜丹冰. 一种优化空天地一体化网络吞吐量算法[J]. 火箭与制导学报, 2023, 43(5): 109-114.
Du Dan-bing. A throughput optimization in space-air-ground integrated networks[J]. Journal of Projectiles, Rockets, Missiles and Guidance, 2023, 43(5): 109-114.
- [8] 乔文欣,卢昱,刘益岑,等. 空天地协同的边缘云服务功能链动态编排方法[J]. 西安电子科技大学学报, 2022, 49(2): 79-88.
Qiao Wen-xin, Lu Yu, Liu Yi-cen, et al. Dynamic scheduling method for service function chains in space air terrestrial aided edge cloud networks[J]. Journal of Xidian University, 2022, 49(2): 79-88.
- [9] 郭淞岐,安康,孙艺夫,等. 基于深度学习的HARQ辅助空天地融合网络时延受限容量预测[J]. 电讯技术, 2023, 63(7): 963-971.
Guo Song-qi, An Kang, Sun Yi-fu, et al. Delay-limited throughput prediction of HARQ-assisted satellite-aerial-terrestrial integrated network: a deep learning approach[J]. Telecommunication Engineering, 2023, 63(7): 963-971.
- [10] 刘芳,朱天贺,苏卫星,等. 基于高斯隐马尔可夫模型的人机共享控制区域化决策算法[J]. 电子学报, 2022, 50(11): 2659-2667.
Liu Fang, Zhu Tian-he, Su Wei-xing, et al. Regionalized decision algorithm for human-machine shared control based on gaussian hidden markov model[J]. Acta Electronica Sinica, 2022, 50(11): 2659-2667.
- [11] 朱霸坤,朱卫纲,李伟,等. 基于马尔可夫的多功能雷达认知干扰决策建模研究[J]. 系统工程与电子技术, 2022, 44(8): 2488-2497.
Zhu Ba-kun, Zhu Wei-gang, Li Wei, et al. Research on decision-making modeling of cognitive jamming for multi-functional radar based on Markov[J].

- Systems Engineering and Electronics, 2022, 44(8): 2488-2497.
- [12] 王浩聪, 付主木, 孙昊琛, 等. 改进深度 Q 学习的燃料电池混合动力汽车能量管理[J]. 河南科技大学学报: 自然科学版, 2022, 43(4): 34-40.
- Wang Hao-cong, Fu Zhu-mu, Sun Hao-chen, et al. Energy management of fuel cell hybrid electric vehicle based on improved deep Q -learning[J]. Journal of Henan University of Science And Technology(Natural Science), 2022, 43(4): 34-40.
- [13] 夏天豪, 夏长清, 潘昊, 等. 基于强化学习的算力资源度量方法[J]. 燕山大学学报, 2023, 47(3): 246-254.
- Xia Tian-hao, Xia Chang-qing, Pan Hao, et al. Computational power resource measurement method based on reinforcement learning[J]. Journal of Yan-shan University, 2023, 47(3): 246-254.
- [14] 侯慧, 何梓姻, 陈跃, 等. 基于深度强化学习区间多目标优化的智能建筑低碳优化调度[J]. 电力系统自动化, 2023, 47(21): 47-57.
- Hou Hui, He Zi-yin, Chen Yue, et al. Low-carbon optimal dispatch of smart building based on interval multi-objective optimization with deep reinforcement learning[J]. Automation of Electric Power Systems, 2023, 47(21): 47-57.
- [15] 李文武, 周佳妮, 裴本林, 等. 梯级水库深度强化学习长期随机优化调度研究[J]. 水力发电学报, 2023, 42(11): 21-32.
- Li Wen-wu, Zhou Jia-ni, Pei Ben-lin, et al. Study on long-term stochastic optimal operation of cascade reservoirs by deep reinforcement learning[J]. Journal of Hydroelectric Engineering, 2023, 42(11): 21-32.