2023年3月
第42卷 第3期

DOI:10.19652/j.cnki.femt.2204575

多用户蜂窝网络中基于深度强化学习的功率分配*

刘子怡[1] 李 君[2] 李正权[3,4]

(1.南京信息工程大学 南京 210044;2.无锡学院 无锡 214105;3.江南大学轻工过程先进控制教育部重点实验室 无锡 214122;4.北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要:在用户密集分布的蜂窝网络中,功率分配是决定系统性能和通信质量的重要因素之一。由于现有的功率分配算法往往达不到理想效果,而且泛化能力较差。在此基础上,提出一种基于D3QN(dueling double deep Q network)的功率分配算法来优化系统的传输速率。D3QN采用双神经网络和竞争网络优化神经网络的结构,通过解耦动作的选择和价值的评估,解决了DQN中出现的高估问题。仿真结果表明,该算法能够获得的平均速率比DQN高7.14%,在收敛速度和稳定性方面也有较好的表现,且泛化能力较强,可适用于不同实际场景。

关键词:功率分配;蜂窝网络;深度强化学习;D3QN算法

中图分类号:TN92 **文献标识码**:A **国家标准学科分类代码**:510

Power allocation based on deep reinforcement learning in multi-user cellular networks

Liu Ziyi[1] Li Jun[2] Li Zhengquan[3,4]

(1. Nanjing University of Information Science & Technology, Nanjing 210044, China; 2. Wuxi University, Wuxi 214105, China; 3. Key Laboratory of Advanced Control of Light Industry Process, Ministry of Education, Jiangnan University, Wuxi 214122, China; 4. State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Power allocation is one of the most important factors that determine the system performance and communication quality in the densely distributed cellular networks. Because the existing power allocation algorithms often can not achieve the desired results, and the generalization ability is poor. On this basis, a power allocation algorithm based on dueling double deep Q network(D3QN) is proposed to optimize the transmission rate of the system. D3QN uses dual neural networks and competitive networks to optimize the structure of neural networks. Through the selection and evaluation of decoupling actions, the overestimation problem in DQN is solved. The simulation results show that the average rate obtained by this algorithm is 7.14% higher than that of DQN algorithm, has better performance in convergence speed and stability, and has strong generalization ability, which can be applied to different actual scenarios.

Keywords: power allocation; cellular network; deep reinforcement learning; D3QN algorithm

0 引 言

随着无线通信网络的飞速发展,接入终端的数量逐渐增多,接入点的密度大幅度增加,给通信网络中数据的传输带来了严峻的挑战。在通信网络中密集部署微基站、家庭基站等小型基站,已成为解决问题的有效方法。由于基站数量更多、小区更小、接入点更密,导致整个无线通信网络中充斥着信号,使得小区间干扰和小区内干扰变得日益严重。无线资源分配和干扰管理变得尤为突出,不恰当的资源分配方案会降低网络频谱效率,同时也会造成资源的

收稿日期:2022-12-03

* **基金项目**:国家自然科学基金(61571108)、网络与交换技术国家重点实验室(北京邮电大学)开放课题(SKLNST-2020-1-13)项目资助

浪费。因此，有必要提出一种较好的功率分配算法来提高用户的服务质量。然而功率分配具有非凸性，文献[1]提出一种新的功率分配策略，根据比例公平方法将非凸目标函数转换成凸函数，用KKT最优约束条件求解。文献[2]提出一种改进的粒子群功率分配算法，通过对用户发射功率最优分配，降低用户间干扰，以提高蜂窝网络通信性能。文献[3]提出一种具有无线电池容量的低复杂度最优离线算法作为系统性能上限，进一步开发具有有限电池容量的启发式在线算法，来降低绿色通信中能耗。文献[4]设计了一种迭代优化算法解决资源分配问题。文献[5]利用分式规划理论(FP)，提高无线通信网络中的能量效率。文献[6]提出一种基于最优控制策略和最优值函数的分层博弈架构，用于解决不确定性干扰下的频谱资源分配问题。以上提出的算法通过理论分析和数值模拟均能优化网络的性能，但是传统算法依赖于大量的计算和完整的数学模型，无法满足通信实际场景。

机器学习算法是未来的发展方向，无线通信网络中的编码译码、信号处理、无人机、边缘计算[7-10]等均有应用。机器学习可以分为监督学习和强化学习。监督学习方法简单有效，通过监督学习训练神经网络逼近目标最优解。文献[11]提出基于强化学习及深度神经网络的框架以优化超密集网络中小基站的功率控制，仿真表明所提方案有很好的自适应能力。文献[12]利用无监督学习方法优化信道功率控制，仿真表明在保证低计算时延的同时获得更高的传输速率。

实际中通信网络的环境十分复杂且很难获得完整的数据集，监督学习存在很明显的局限。强化学习(reinforcement learning，RL)是面向目标的算法，相较于监督学习有一定的优势，是通过智能体与环境不断地交互学习，实现特定目标或获得奖励最大化[13]，对于不断变化的实际场景，RL是一个很好的解决方案。QL算法(Q-learning)是强化学习中最经典的算法，文献[14]提出一种协同QL算法，用于超密集异构网络中的联合无线资源管理，有效提高了用户的服务质量。文献[15]利用QL优化微蜂窝中的资源分配问题，仿真表明该方法在减少资源分配开销的同时可以大大增加用户数。

QL建立了状态－动作与该动作下最大奖励之间的映射关系，然而状态空间与动作空间巨大时，需要建立的Q表十分庞大，因此可以用神经网络来计算Q值。文献[16]将神经网络与QL相结合，提出DQN(deep Q network)。文献[17]提出一种基于全连接网络的DQN算法，用于多小区功率分配，收敛速度和稳定性方面有明显提高。文献[18]提出了用Double DQN优化用户关联和资源分配的联合问题。文献[19]研究了基于多智能体Dueling DQN的资源管理优化算法，经训练该算法可快速收敛到最优策略。本文研究的是蜂窝网络中多用户的功率分配问题，提出一种基于D3QN的功率分配算法，目标是在最大功率的约束下使该网络总体和速率最大化。

1 系统模型和问题描述

本文考虑的是下行链路中的蜂窝网络，蜂窝网络模型如图1所示。网络中有 N 个小区，在每个小区的中心部署一个基站，同时服务于 M 个用户。用户 m 在接收其所连接基站发送的信号时，会受到邻近基站和其所连接基站发送给其他用户的信号干扰。

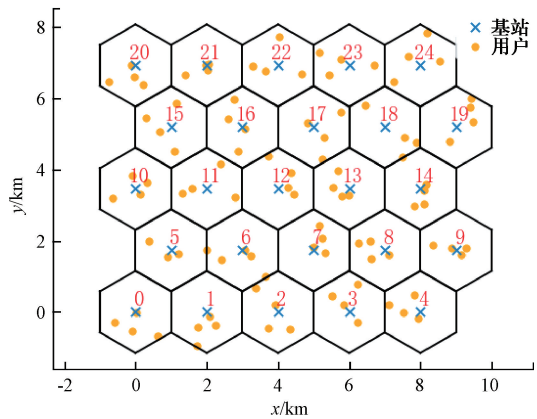

图1 蜂窝网络模型示例

在时隙 t，小区 j 中基站 n 到用户 m 的信道增益表示为：

$$g_{n,j,m}=|h^t_{n,j,m}|^2\beta_{n,j,m} \tag{1}$$

式中：$h^t_{n,j,m}$ 是小尺度衰落；$\beta_{n,j,m}$ 是大尺度衰落。在时隙 t，基站 n 到用户 m 的信干噪比(signal interference plus noise ratio，SINR)表示为：

$$SINR^t_{n,m}=\frac{g^t_{n,n,m}p^t_{n,m}}{\sum\limits_{m'\neq m}g^t_{n,n,m}p^t_{n,m'}+\sum\limits_{n'\in D_n}g^t_{n',n,m}\sum\limits_j p^t_{n',j}+\sigma^2} \tag{2}$$

式中：D_n 是邻近基站干扰集；p 是基站的发射功率；σ^2 是高斯白噪声功率。在时隙 t，用户 m 的干扰信息集合表示为：

$$I^t_{n,m}=\{g^t_{n,m}\mid n\in\{n,D_n\}\} \tag{3}$$

在带宽归一化的情况下，基站 n 到用户 m 的传输速率表示为：

$$C^t_{n,m}=\log_2(1+SINR^t_{n,m}) \tag{4}$$

本文的优化目标是在最大发射功率的约束下，使蜂窝网络中的和速率最大化，优化问题公式为：

$$\begin{aligned}&\max_{pt}\sum_n\sum_m C^t_{n,m}\\&0\leqslant p^t_{n,m}\leqslant P_{\max}\quad\forall n,m\end{aligned} \tag{5}$$

式中：$P^t=\{p^t_{n,m}\mid\forall n,m\}$；$P_{\max}$ 表示最大发射功率。

2 算法设计

2.1 深度强化学习

强化学习是机器学习的一个分支，其思想让智能体与

环境进行交互，通过学习最优策略，做出决策并获得最大回报。状态空间为 S，动作空间为 A，在时隙 t，智能体处于状态 $s_t \in S$，采用动作 $a_t \in A$ 与环境交互后，得到奖励 r_t，智能体到达下一个状态 s_{t+1}。本文将每个发射机视为智能体，将 DRL 的联合优化过程视为马尔可夫过程。本文的状态空间、动作空间及奖励函数具体设置如下。

1) 状态空间

状态的选取很重要，是智能体对环境探索的信息集合。为了降低输入数据的维度，在时隙 t，将用户 m 的干扰信息 $I_{n,m}^t$ 进行对数归一化处理，并按照从大到小的顺序排序，选取前 14 个元素作为神经网络的输入，用 $Z_{n,m}^t$ 表示。理论上最佳发射功率 p^t 与信道增益 g^t 有关，但实际上很难找到最优解。因此加入两个辅助信息，前一时隙的发射功率 p^{t-1} 及前一时隙的传输速率 C^{t-1}，最终状态表示为：

$$s_{n,m}^t = \{Z_{n,m}^t, p_{n,m}^{t-1}, C_{n,m}^{t-1}\} \tag{6}$$

2) 动作空间

动作是基站的发射功率。发射功率是一个连续变量且受到最大发射功率的约束，但 D3QN 中的动作是有限的，因此将动作离散为 A 个等级，A 中包括 1 个零值，和 $|A|-1$ 个介于最大发射功率和最小发射功率之间的均匀取值。动作表示为：

$$A = \{0, p_{\min}, \cdots, p_{\max}\} \tag{7}$$

式中：$p_{\min}$ 表示最小发射功率；$p_{\max}$ 表示最大发射功率。

3) 奖励函数

奖励是智能体根据状态进行下一步动作的函数，用于评估动作的好坏。奖励函数决定了优化方向，因此奖励函数的设定一般与优化目标有关。本文的奖励函数为当前时刻系统的和速率。

2.2 基于深度强化学习功率分配算法

QL 中使用状态－动作矩阵表示智能体每个回合结束后的累计奖励，Q 值更新公式表示为：

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)) \tag{8}$$

式中：$\gamma \in [0,1]$ 为折扣系数，权衡了当前和未来奖励的重要性，如果 $\gamma = 0$ 表示智能体只关心最大化当前收益；r 表示奖励。QL 的核心在于状态－动作矩阵，通过和查找矩阵中累计奖励为动作提供指引，但这只适用于状态和动作空间是离散的低维度数据，实际中的数据量十分庞大，这对于建立和查找 Q 表都是不现实的。文献[20]提出用函数来拟合 Q 值，深度学习在复杂特征提取中具有良好的表现，将神经网络与 QL 结合得到 DQN。在 DQN 中有两个结构完全相同但参数不同的网络分别为当前 Q 网络（参数 θ）和目标 Q 网络（参数 θ^-），当前 Q 网络参数实时更新，目标 Q 网络参数每隔固定步数将当前 Q 网络参数复制过来，目标网络 Q 值为：

$$y_t^{DQN} = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) \tag{9}$$

根据目标网络 Q 值与当前网络 Q 值计算损失函数，损失函数一般采用均方误差进行计算：

$$L(\theta) = E[(y_t^{DQN} - Q(s_t, a_t; \theta))^2] \tag{10}$$

研究结果表明，DQN 相较于 Q 学习效果更优，但是 Q 学习中的最大化操作导致 DQN 中的值函数比真实值函数偏大，为避免过估计问题 Deep Mind 团队提出了 Double DQN[21]，通过将动作选择和价值评估进行解耦来降低样本之间的相似性，从而避免 Q 值估计过高的问题。DDQN 中目标 Q 值的计算表达式为：

$$y_t^{DDQN} = r_t + \gamma Q(s_{t+1}, \mathrm{argmax}_{a'} Q(s_{t+1}, a'; \theta); \theta^-) \tag{11}$$

其中，动作选择基于参数 θ 使用 argmax 操作，而价值评估采用参数 θ^- 评估 Q 值。然而某些状态下的奖励值是相同的，而与智能体采取的动作无关，这就会导致智能体只关注状态的价值，而不关心不同动作导致的差异。此时引入了竞争网络（图 2），将二者分开建模使智能体更好的处理与动作关联较小的状态。

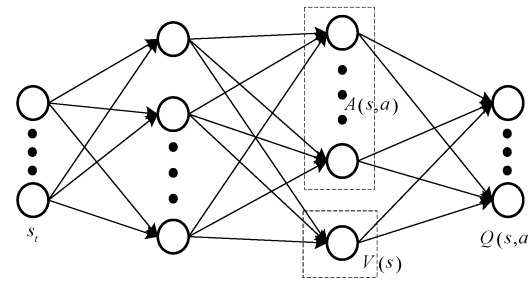

图 2 竞争网络结构

分别计算价值函数 $V(s)$ 和优势函数 $A(s,a)$ 来学习 Double DQN，其中 $A(s,a) = Q(s,a) - V(s)$，优势函数则是评估在状态 s_t 采取各个动作相对于平均回报的好坏，也就是这个动作的优势。为了提高算法的稳定性，通常将所采取动作的优势函数减去采取所有动作的平均优势值，即：

$$Q(s_t, a_t; \theta, \kappa, \beta) = V(s_t; \theta, \beta) + A(s_t, a_t; \theta, \kappa) - \frac{1}{|A|}\sum_{a_{t+1}} A(s_t, a_{t+1}; \theta, \kappa) \tag{12}$$

式中：θ 是 $V(s)$ 和 $A(s,a)$ 共享参数；κ 是 $A(s,a)$ 的参数；β 是 $V(s)$ 的参数；$|A|$ 表示所有可采取动作的数量。D3QN 是在 Double DQN 的基础上融入了 Dueling DQN 的思想，延续了 Dueling DQN 的神经网络框架，但目标 Q 值的计算采用了 Double DQN 的计算方法。基于 D3QN 功率分配算法的训练过程如下。

算法 1 基于 D3QN 的功率分配算法

1 初始化当前网络参数 θ、目标网络参数 θ^-、经验池
2 重复：
3 初始化状态 s_t
4 重复：

5　智能体在状态 s_t 下，根据 ε 一贪婪算法选择动作
6　基于环境中选择的动作，根据式(4)得到奖励 r_t 及下一个状态 s_{t+1}
7　将获得的经验 (s_t, a_t, r_t, s_{t+1}) 到经验池
8　if $episode > 100$
9　从经验池 D 中抽取小批量样本 B 用于训练神经网络
10　计算目标 Q 值，根据式(11)计算得到
11　计算损失函数 $L(\theta)$，反向传播更新当前网络参数 θ
12　更新目标网络参数 θ^-
13　结束
14 结束

3　仿真结果及结论

3.1　仿真环境设置

模拟场景中小区数为 $N = 25$ 个蜂窝网络，每个小区的中心配备一个基站，每个基站同时为 $M = 4$ 个用户服务。用户随机分布在小区 $r \in [R_{\min}, R_{\max}]$ 内，其中 $R_{\min} = 0.01$ km 为小区内基站到用户的最短距离，$R_{\max} = 1$ km 为小区内基站到用户的最长距离。小尺度衰落采用 Jakes 模型，服从瑞利分布，多普勒频率为 $fd = 10$ Hz，时间周期 $T = 20$ ms。路径损耗为 $\beta = 120.9 + 37.6\lg d + 10\lg z$ dB，其中 d 是基站到用户的距离，z 是对数正态随机变量，标准差为 8 dB。高斯白噪声功率 σ^2 为 -114 dBm。基站的最小发射功率 $P_{\min} = 5$ dBm，最大发射功率 $P_{\max} = 38$ dBm，功率等级 $|A| = 10$。邻近基站数为 $|D_n| = 16$。

D3QN 算法中的神经网络包括输入层、两个隐藏层和输出层，隐藏层的激活函数均为 ReLU，输出层采用线性激活函数，神经网络的输入和输出维度分别为 42 和 10。训练持续 $T_{\max}^{episode} = 10\,000$ 个回合，每个回合迭代 100 次，前 200 个回合只能随机选择动作，其余回合采用 ε- 贪婪算法。神经网络参数设置如表 1 所示。

表 1　神经网络参数设置

参数	值
折扣因子 γ	0.1
经验池大小 D	100 000
批量样本数量 B	512
学习率 λ	0.000 1
初始探索率 ε_{start}	0.3
最终探索率 ε_{end}	0.001
隐藏层神经元数量	128(1),64(2)

3.2　仿真结果与分析

将本文算法与 DQN、FP、WMMSE 及最大功率(Maximal power) 4 种功率分配算法进行对比，其中本文算法与 DQN 中设置相同的参数。图 3 所示为本文算法和 DQN 的训练结果经过平滑窗口处理后的收敛图。从图 3 可以看出，随着迭代回合数的增加，两种算法的平均速率都有所提高。但是在相同的训练回合内，本文算法相较于 DQN 整体平均速率更高，能收敛到的平均速率数值更大，故本文算法具有明显的优势。

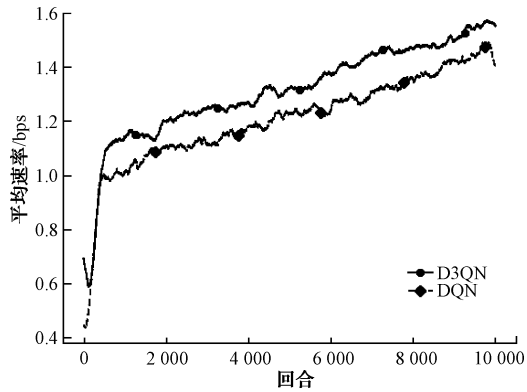

图 3　D3QN 与 DQN 收敛性能对比

在 1 000 个回合中 5 种算法的性能对比如图 4 所示。从图 4 可以看出，最大功率分配算法的平均速率明显低于其他算法。这种算法只是简单的按照设定好的方式进行功率分配，没有考虑到通信中实际环境的变化。WMMSE 和 FP 是功率分配中经常用到的两种算法，可以看出，这两种算法相比于最大功率分配算法的平均速率有明显的提高，但是这两种算法中涉及到复杂的数学计算，运行时会耗费很长的时间和占用过多的计算机资源。DQN 与以上 3 种算法相比，DQN 的平均速率更高，性能优势更加明显。本文算法与 DQN 大体上变化趋势一致，但本文算法的平均速率值更高，相对于 DQN 来说，曲线波动幅度更小，总体性能更加稳定。原因是本文算法在进行动作选择和价值评估时采用了不同的神经网络，远远降低了样本之间的相似性。由此可见，D3QN 的性能更优。

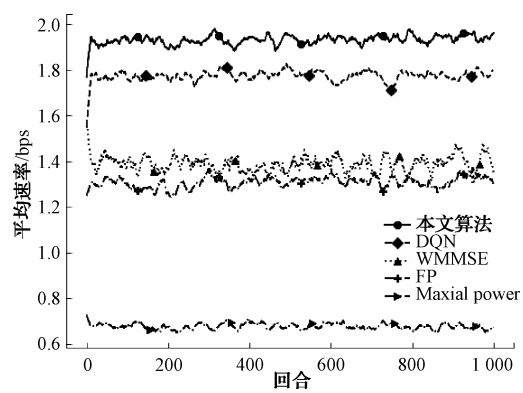

图 4　5 种算法的平均速率对比

在实际应用中，用户数不是固定的。用户从 1 增加到 8 时，5 种算法的性能对比如图 5 所示。从图 5 可以看出，当用户数量为 1 时，平均速率最高，随着用户数量的增加，通信环境的质量有所下降，5 种算法的性能皆有一定幅度

的下降，其中最大功率分配算法下降幅度最为明显，即稳定性最差。WMMSE 和 FP 的性能并无太大差异，均能达到较高水平，但平均速率均低于 DQN。RL 算法的平均速率和稳定性均优于其他普通算法，随着用户数的增加，RL 算法平均速率的下降幅度小于其他普通算法，故其稳定性更好。而本文算法在不同用户数时皆能达到最高的平均速率，稳定性相较于 DQN 也有一定优势，从而验证了本文算法在不同用户数时具有良好的泛化能力。

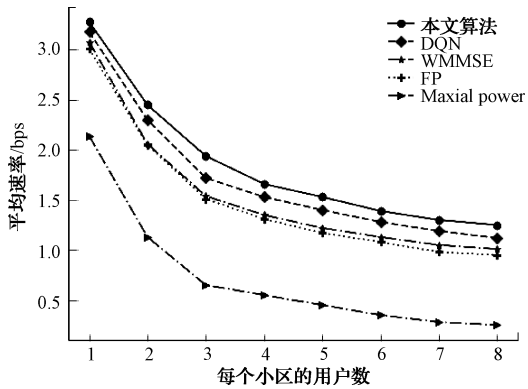

图 5　不同用户数时平均速率对比

不同小区数的 5 种算法性能对比如图 6 所示。从图 6 可以看到，5 种算法的平均速率随着小区数的增加而减小，这是因为当小区数量增加时，目标小区周围的干扰会增加，通信环境会变得复杂，以至于会降低目标小区的速率。但是，相比于其他 4 种算法，本文算法的平均速率均为最高，因此验证了本文算法在小区数不同时具有良好的泛化能力。

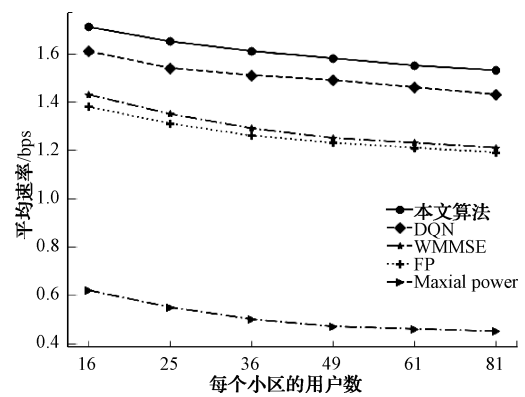

图 6　不同小区数时平均速率对比

4　结　论

本文研究了多用户蜂窝网络中的功率分配问题，提出了用 D3QN 算法来优化通信网络系统中功率分配问题，在最大发射功率的约束下，找到系统最佳发射功率，提高功率分配的合理性和高效性。仿真结果表明，本文算法可以显著提高通信网络的平均传输速率，且具有良好的泛化能力，可适用于不同小区数和用户数的环境。但是本文还存在一些局限性，D3QN 是基于价值的算法，只能适用于离散动作空间的情况，对于复杂的连续动作空间不再适用。未来，将考虑进一步优化蜂窝网络模型，如引入异构网络模型，同时采用 RL 中其他算法处理功率分配中连续动作空间的问题。

参 考 文 献

[1] 曹雍，杨震，冯友宏. 新的 NOMA 功率分配策略[J]. 通信学报，2017，38(10)：157-165.

[2] 张继荣，孟繁克，王晟寰. 改进粒子群算法的 D2D 功率分配[J]. 西安邮电大学学报，2021，26(2)：8-14.

[3] LIU D，CHEN Y，CHAI K K，et al. Two-dimensional optimization on user association and green energy allocation for HetNets with hybrid energy sources[J]. IEEE Transactions on Communications，2015，63(11)：4111-4124.

[4] ZHANG H，LIU H，CHENG J，et al. Downlink energy efficiency of power allocation and wireless backhaul bandwidth allocation in heterogeneous small cell networks [J]. IEEE Transactions on Communications，2018，66(4)：1705-1716.

[5] WU Q，TAO M，KWANG D W，et al. Energy-efficient resource allocation for wireless powered communication networks[J]. IEEE Transactions on Wireless Communications，2016，15(3)：2312-2327.

[6] 周燕. 基于最优控制策略和最优值函数的无线频谱资源分配[J]. 电子测量与仪器学报，2019，33(3)：44-50.

[7] 万飞，白宝明，朱敏. 多元 LDPC 编码调制系统 CNN 辅助迭代检测译码算法[J]. 无线电通信技术，2022，48(4)：673-679.

[8] 张玮，赵永虹，邱桃荣. 基于注意力机制和深度学习的运动想象脑电信号分类方法[J]. 南京大学学报(自然科学)，2022，58(1)：29-37.

[9] ZHONG R，LIU X，LIU Y，et al. Multi-agent reinforcement learning in NOMA-Aided UAV networks for cellular offloading [J]. IEEE Transactions on Wireless Communications，2022，21(3)：1498-1512.

[10] 杨东轩，吴叶兰，张刚刚. 基于边缘计算与深度学习的禽舍监测系统设计[J]. 江苏农业科学，2022，50(9)：219-225.

[11] 郑冰原，孙彦赞，吴雅婷，等. 基于深度强化学习的超密集网络资源分配[J]. 电子测量技术，2020，43(9)：133-138.

[12] 孙明，王淑梅，郭媛，等. 基于深度无监督学习的多小区蜂窝网资源分配方法[J]. 控制与决策，2022，37(9)：2333-2342.

[13] 杨思明,单征,丁煜,等. 深度强化学习研究综述[J]. 计算机工程,2021,47(12):19-29.

[14] IQBAL M U, ANSARI E A, AKHTAR S, et al. Improving the QoS in 5G HetNets through cooperative Q-learning[J]. IEEE Access, 2022(10): 19654-19676.

[15] AMIRI R, MEHRPOUYAN H, FRIDMAN L, et al. A machine learning approach for power allocation in HetNets considering QoS [C]. 2018 IEEE International Conference on Communications (ICC). IEEE, 2018.

[16] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[17] ZHANG Y, KANG C, MA T, et al. Power allocation in multi-cell networks using deep reinforcement learning [C]. 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018, 1-6.

[18] ZHAO N, LIANG Y C, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous networks[C]. 2018 IEEE Global Communications Conference (GLOBECOM). IEEE, 2018.

[19] YANG H, ZHAO J, LAM K Y, et al. Deep reinforcement learning based resource allocation for heterogeneous networks[C]. 2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2021: 253-258.

[20] LIU Q, KWONG C F, ZHOU S, et al. Autonomous mobility management for 5G ultra-dense HetNets via reinforcement learning with tile coding function approximation[J]. IEEE Access, 2021(9): 97942-97952.

[21] 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2018,41(1):1-27.

作 者 简 介

刘子怡,硕士研究生,主要研究方向为无线通信、深度强化学习方向、资源分配。

E-mail:20211249646@nuist.edu.cn

李君,副教授,主要研究方向为无线通信、资源分配、机器学习、编码译码等。

E-mail:07a0303105@cjlu.edu.cn

李正权(通信作者),教授,主要研究方向为无线通信、信号处理、信道编码译码。

E-mail:lzq722@jiangnan.edu.cn