DOI:10.19651/j.cnki.emt.2107490

基于改进深度强化学习的移动机器人路径规划*

王 军[1,2] 杨云霄[1,2] 李 莉[1,2]
(1.沈阳化工大学 计算机科学与技术学院 沈阳 110142; 2.辽宁省化工过程工业智能化技术重点实验室 沈阳 110142)

摘 要: 针对传统深度强化学习中移动机器人在稀疏奖励环境下只有在规定时间步内到达目标位置才能得到积极奖励,中间过程的每一步都是负面奖励的路径规划问题。提出了基于改进深度 Q 网络的路径规划方法,在移动机器人在探索过程中,对以真实目标为条件的轨迹进行采样,在经验回放过程中,把移动机器人已经到达的状态来代替真正的目标,这样移动机器人可以获得足够的积极奖励信号来开始学习。通过深度卷积神经网络模型,将原始 RGB 图像作为输入,通过端对端的方法训练,利用置信区间上界探索策略和小批量样本的方法训练神经网络参数,最后得到上、下、左、右 4 个动作的 Q 值。在相同的仿真环境中结果表明,该算法提升了采样效率,训练迭代更快,并且更容易收敛,避开障碍物到达终点的成功率增加 40%左右,一定程度上解决了稀疏奖励带来的问题。

关键词: 深度强化学习;路径规划;稀疏奖励;移动机器人;后见经验回放
中图分类号: TP242 **文献标识码:** A **国家标准学科分类代码:** 520.2

Mobile robot path planning based on improved deep reinforcement learning

Wang Jun[1,2] Yang Yunxiao[1,2] Li Li[1,2]
(1. School of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China;
2. Liaoning Key Laboratory of Intelligent Technology for Chemical Process Industry, Shenyang 110142, China)

Abstract: In the traditional deep reinforcement learning, the mobile robot can get a positive reward only when it reaches the target position within the specified time step in the sparse reward environment. Each step of the intermediate process is a path planning problem with negative reward. A path planning method based on improved depth Q network is proposed. In the process of exploration, the mobile robot samples the trajectory conditional on the real target, and replaces the real target with the state that the mobile robot has reached in the process of experience playback, so that the mobile robot can obtain enough positive reward signals to start learning. Through the deep convolution neural network model, the original RGB image is used as the input, trained through the end-to-end method, trained the neural network parameters by using the upper bound exploration strategy of confidence interval and the method of small batch samples, and finally obtained the Q values of up, down, left and right actions. In the same simulation environment, the results show that the algorithm improves the sampling efficiency, the training iteration is faster, and it is easier to converge. The success rate of avoiding obstacles to reach the end point increases by about 40%, which solves the problem caused by sparse reward to a certain extent.

Keywords: deep reinforcement learning; path planning; sparse reward; move robot; hindsight experience replay

0 引 言

移动机器人路径规划[1]是当前移动机器人研究的热点方向,随着移动机器人技术的快速发展和应用场景逐渐复杂化,对于移动机器人来说,在复杂未知的环境中规划路径是一个非常重要的问题。传统算法有 RRT *[2],遗传算法[3],蚁群算法[4]等。谷歌的 DeepMind 公司在《Playing Atari with Deep Reinforcement Learning》[5]提出 DQN (deep Q-Network, DQN)算法,将深度学习与强化学习相结合,使得深度强化学习在路径规划中显示出了巨大的潜力。DQN 使用神经网络模型逼近值函数,回放经验池用于打破样本顺序,以解决从强化学习中获得的经验与时序关

收稿日期:2021-08-04
*基金项目:辽宁省高校创新人才支持计划(LR2018057)、辽宁省"百万人才工程"资助项目(辽人社【2019】45 号)、辽宁省自然基金(2019-ZD-0068)、辽宁省教育厅项目(XXLJ2019010)资助

联的问题，它提高了深度神经网络稳定性并易收敛。文献[6]通过改进 Q-Learning 算法来解决移动机器人的路径规划问题，把花授粉算法(flower pollination algorithm，FPA)应用于 Q-Learning 的初始化。在具有不同障碍物的环境下实验的评估表明，当使用花授粉算法适当初始化值时，此方法可以加快 Q-Learning 的收敛。文献[7]提出了移动机器人在动态环境下的一种基于改进 Q-Learning 算法和启发式搜索策略的新方法，相对于传统的 Q-Learning 在路径规划所花费的时间更少，但是在接近实际情况的下，会有着很大的状态空间和连续的动作空间，出现维度灾难。文献[8]提出一种改进深度强化学习算法，加入了差值增长思想，解决了深度强化学习算法的过估计问题，并应用于移动机器人三维路径规划，但是奖励函数设计并未考虑到实际因素。文献[9]提出了一种基于深度强化学习的机器人多路径规划算法，可以实现在多种环境下进行路径规划，但在没有障碍物的环境下，机器人会出现不必要的移动，并且需要设计相对比较复杂的奖励函数。

在训练初期，移动机器人的探索是采用随机策略，需要与环境频繁交互，积极的奖励难以获得，在稀疏奖励环境中，有效的探索显得更为重要，奖励函数的设计需要考虑多方面因素，不恰当的奖励函数会导致算法无法收敛。在许多复杂的情况下，只有在满足特定条件的情况下才会给予奖励，因此稀疏奖励带来的负面影响，难以解决。因此本文提出稀疏环境下基于 DQN 改进的后见经验回放深度 Q 网络算法，其中置信区间上界[10]探索策略采用置信水平来实现对探索与利用之间的平衡；通过后见回放经验机制[11]的使用，很好地解决了在稀疏奖励[12]环境下算法难以收敛的问题，提升了样本利用率，加快收敛速度，一定程度上避免了强化学习在路径规划中需要设计复杂奖励函数。仿真结果表明所提出方法在稀疏奖励环境下移动机器人路径规划的有效性。

1　相关工作

强化学习是训练机器学习模型做出一系列决策，智能体需要不断学习，通过与不确定的、复杂的环境互动来实现目标，强化学习的目标就是最大化期望累计奖励。强化学习的关键要素有：环境、奖励、动作和状态。在强化学习应用到路径规划问题中，移动机器人处于复杂不确定的环境中，移动机器人从环境中获取状态，选择执行某个动作后，会使得环境转移到另一个状态，同时环境会给移动机器人反馈奖励。如果获得积极奖励，那么这个动作的概率增大；反之这个动作被采用的概率会降低，目的是通过学习行为策略来最大化奖励。

1.1　Q-Learning 算法

Q-Learning[13]是在 1989 年被 Watkins 等人提出的一种基于无模型、离线策略的强化学习算法，并结合时序差分法[14](temporal difference，TD)算法和蒙特卡洛[15](monte-carlo，MC)算法，利用表格存储 Q 值。行为策略和目标策略分别采用 ε-$greedy$ 算法(如式(1)所示)、$greedy$ 算法。行为策略通过探索得到的经验来优化目标策略，目标策略采用最优策略。Q-Learning 算法的目标就是找到最优策略的 Q 值。首先初始化 Q 值，通过 ε-$greedy$ 选择动作 a_t，通过观察得到奖励 r_t，进入 s_{t+1} 状态，通过软更新逼近目标值，如式(2)所示。

$$\pi(a\mid s)=\begin{cases}\varepsilon/\mid A\mid+1-\varepsilon, & a^*=\underset{a\in A}{\operatorname{argmax}}Q(s,a)\\ \varepsilon/\mid A\mid, & \text{其他}\end{cases}\tag{1}$$

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r_t+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)]\tag{2}$$

其中，α 为学习率，γ 为折扣率。但是在复杂的环境下，状态空间变得巨大，通过表格的方式储存状态容易产生维度灾难。

1.2　DQN 算法

Mnih 等[16]在 2015 年进一步提出了目标网络的概念。利用深度卷积神经网络逼近值函数和经验回放机制。神经网络是一个非线性函数 $f:R^n\rightarrow R^m$，n 和 m 分别表示状态空间和动作空间的维数。最优 Q 函数如式(3)所示。

$$Q^*(s,a)=E_{s'}[r(s,a,s')+\gamma\max_{a'}Q^*(s',a')\mid s,a]\tag{3}$$

利用神经网络估计函数 $Q(s,a;\theta)\approx Q^*(s,a)$。其中 θ 为神经网络的参数。状态 s 作为神经网络的输入，每个动作的 Q 值作为输出。首先通过采取 ε-$greedy$ 策略随机选取动作探索环境，把得到的样本 $(\phi_j,a_j,r_j,\phi_{j+1})$ 存储到回放经验池 D 中，然后从回放经验池 D 中的随机采样 $mini$-$batch$ 大小的样本，对损失函数执行梯度下降(stochastic gradient descent，SGD)，更新网络参数 θ，为了提高深度强化学习的稳定性和收敛性，每隔 C 步更新一次 TD 目标网络参数 θ^-，即令 $\theta^-=\theta$。其中行为值函数表示为 $Q(s,a,\theta)$，TD 目标值函数表示为 $Q(s',a';\theta^-)$，损失函数定义为式(4)。

$$L(\theta)=E[(r(s,a,s')+\gamma\max_{a'\in A}Q(s',a';\theta^-)-Q(s,a;\theta))^2]\tag{4}$$

2　改进 DQN 的算法

在稀疏奖励的环境下，移动机器人在环境中比较难能获得积极奖励，不易实现目标。本文根据 DQN 算法，提出了一种在稀疏奖励环境中改进的 DQN 算法，使移动机器人在二维环境中可以避开障碍物到达终点。

本文深度卷积神经网络模型包含 3 个卷积层和 2 个全连接层，通过端对端的方法训练，以原始 RGB 图像作为输入，经深度卷积神经网络处理最后得到上、下、左、右 4 个动作的 Q 值。如图 1 所示。

传统 DQN 算法中，回放经验池中大部分样本并未获

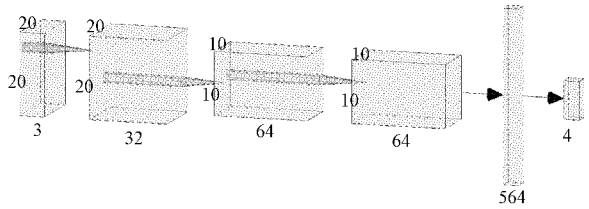


图 1　神经网络模型

得奖励,样本利用率较低。Andrychowicz 等在 2017 年提出了后见经验回放(hindsight experience replay, HER)算法,通过把已到达的状态作为目标,可以让移动机器人充分利用样本,此方法适用于离线策略的强化学习算法。因此本文提出其与 DQN 算法相结合的路径规划算法,在移动机器人的路径规划上取得了不错的效果。具体的算法流程如下。

算法 1 HERDQN 算法

初始化回放经验池 D,容量为 N,随机初始化估计网络参数 θ,初始化目标网络参 $\theta^{-}=\theta$,目标选择策略 $S(s_0,\cdots,s_T)=m(s_T)$

for $episode=1$, M do

采集目标 g,并初始化状态 s_0

for $t=0$, $T-1$ do

采用 UCB 策略选择动作 a_t

$a_t=\text{argmax}(Q_t(s,a,g;\theta)+U_t(a))$

移动机器人执行动作 a_t,得到下一个状态 s_{t+1}

end for

for $t=0$, $T-1$ do

$r_t:=r(s_t,a_t,g)$

将数据样本 $(s_t\|g,a_t,r_t,s_{t+1}\|g)$ 存储到回放经验池 D 中

从当前 *episode* 选取附加目标 $g'=S$

$r_t:=r(s_t,a_t,g')$

存储 $(s_t\|g',a_t,r'_t,s_{t+1}\|g')$ 到回放经验池 D

end for

for $t=1$, N do

从回放经验池 D 中采样 *mini-batch* 样本数据 B

执行梯度下降,更新网络参数 θ

每隔 C 步,更新目标网络权值 $\theta^{-}=\theta$

end for

end for

训练流程图如图 2 所示。此时 Q 函数不仅需要状态、动作,还需要目标。因此 Q 函数被定义为:$Q(s,a,g)$。奖励函数依赖于 $g\in G$,$r_g:S\times A\times G\to R$ 在每一个回合中采样一个目标 g,并在整个回合中保持固定,在稀疏奖励的

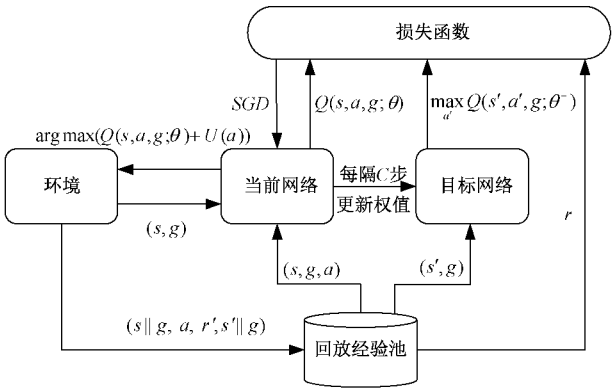


图 2　训练流程图

环境中采用相对简单的二分奖励函数,如果未实现目标奖励为−1,实现目标奖励为 0。奖励函数如式(5)所示。

$$r_t=r_g(s_t,a_t)=\begin{cases}0, & s_t=g\\ -1, & \text{其他}\end{cases}\tag{5}$$

样本附加存放对应的目标变为 $(s_t\|g',a_t,r'_t,s_{t+1}\|g')$,加上原目标产生的轨迹,即经验池里面存放 2 倍于真实采样到的样本。那么目标函数如式(6):

$$L(\theta_i)=E_{(s\|g,a,r,s'|g)\sim U(D)}[(r+\gamma\max_{a'\in A}Q(s',a',g,\theta'_i)-Q(s,a,g,\theta_i))^2]\tag{6}$$

其中,$(s\|g,a,r,s'\|g)$ 为经验数据,$U(D)$ 代表回放经验池的回放记忆单元。

深度强化学习的经典探索算法有 ε-*greed* 策略,玻尔兹曼策略,汤普森抽样。ε-*greed* 策略虽然每个动作都有被选择的概率,但是无引导性,这并不能有助于移动机器人很大概率的发现最优动作。本文采用与置信区间上界(upper-confidence-bound, UCB)采用置信水平来实现对探索与利用之前的平衡,置信区间越大,方差越大,采样的不确定性就越大。如式(7)所示。

$$a_t=\arg\max(Q_t(s,a,g;\theta)+U_t(a))\tag{7}$$

其中,$U_t(a)=c\sqrt{\dfrac{\ln t}{N_{t(a)}}}$,$N_{t(a)}$ 表示目前该动作 a 被选择的次数,c 为权值。开始训练时,所有动作均未执行,$U_t(a)$ 会趋于无穷大,移动机器人将会执行所有动作,随着训练时间的增长,当前动作被执行的次数很低时,$U_t(a)$ 值变大,不确定性越高,对应动作被执行的概率越大;反之亦然。随着训练次数的增加。*Int* 增长速度会越来越慢,$N_{t(a)}$ 增长速度基本保持不变,*Int* 值逐渐下降,每个动作的置信区间都变得收敛。

3　实验结果与分析

3.1　简单环境

为了验证 HERDQN 方法的有效性,首先在文献[11]中所提的 Bit-flipping 实验环境中,状态空间为 $S=\{0,1\}^n$,其中 $n=10$。对 DQN 算法,分别采用 ε-*greed* 探索策略和

UCB 探索策略的 HERDQN 算法进行比较，实验结果如图 3 所示。

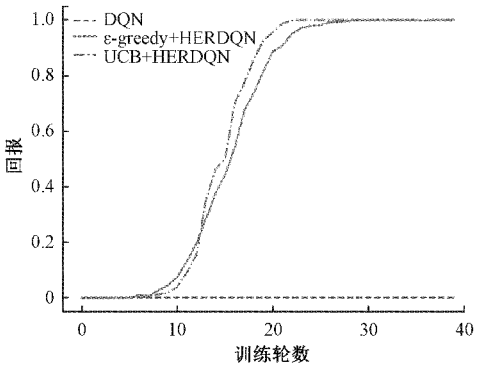


图 3　不同方法的成功率

在 Bit-flipping 实验环境中，如果智能体未能实现目标，体只能收到－1 的奖励，否则为 0。并且智能体需要探索巨大的动作状态空间，因此 DQN 算法在此环境下，只能获得负面奖励，无法成功。而在通过结合 HER 算法之后，重新采样新目标和计算奖励，并存入到回放经验池，智能体可以获得积极奖励，成功率最终收敛于最优。ε-*greed* 探索采用完全随机的策略，不能有助于智能体发现最优策略，而 UCB 策略可以更好的维持开发与探索之间的平衡，因为前期探索较多，所以前期学习速度相对较慢。从图 3 中表明，HERDQN 算法比 DQN 算法有着更好的适应性和学习能力，证明了 HERDQN 的可行性。

3.2　复杂环境

本文实验环境为 GPU RTX 2080ti，Python3.7，Torch 1.8.1，Cuda11.1。仿真环境是一个大小为 20×20 像素的二维栅格地图。其矩形代表障碍物，实心圆代表起点，五角星代表终点，白色区域代表可通行区域，如图 4 所示。

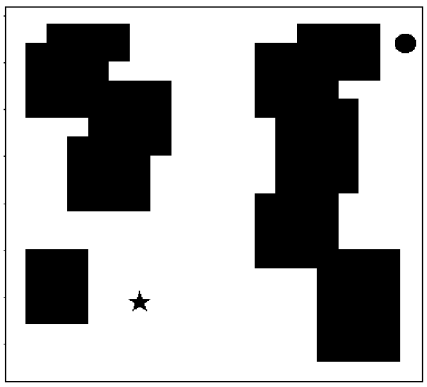


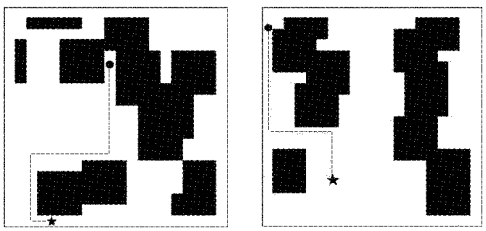
图 4　栅格仿真环境

移动机器人训练时，探索策略采用 UCB 方法，之后采集有附加目标的转移样本存储到回放经验池 D 中，对回放经验池中的样本，重新附加存放对应回合的新目标，然后利用回放经验池 D 中的样本训练网络。在每个回合的训练过程中，需要更新目标网络的参数 θ 。移动机器人有上、下、左、右 4 个方向维度的动作。参数设置如表 1 所示。

表 1　训练参数表

参数	值
C	2
Epochs	15 000
折扣率 γ	0.99
学习率 α	0.000 1
经验回放池大小	500 000
目标网络更新步数	3 000

在不同障碍物的环境中移动机器人路径规划如图 5 所示。



(a) 路径图一　　(b) 路径图二

图 5　移动机器人路径图

在相同的二维栅格仿真环境中，使用相同的二分奖励函数训练 15 000 次，分别对 DQN 算法和 HERDQN 算法进行比较，实验结果如下。

从图 6、7 中比较，横坐标代表训练回合次数，纵坐标代表移动机器人到达终点的成功率。DQN 在相同环境下经过约 15 000 次训练只能达到 30%左右的成功率，并且没有收敛，波动很大。训练初期，HERDQN 成功率上升更快，训练约 3 000 次时 HERDQN 算法就可以达到约 60%的成功率，逐步趋于收敛，最终可以达到 70%多的成功率。与 DQN 算法相比，HERDQN 算法下的移动机器人到达目标点的成功率提高约 40%。

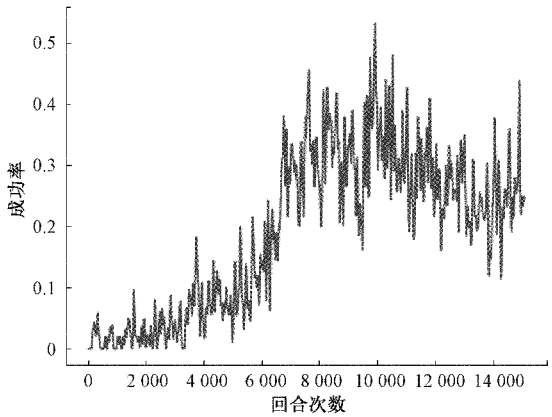


图 6　DQN 成功率

从图 8、9 中比较，横坐标代表训练回合次数，纵坐标代表移动机器人能得到的奖励值。训练时，获得的奖励值越

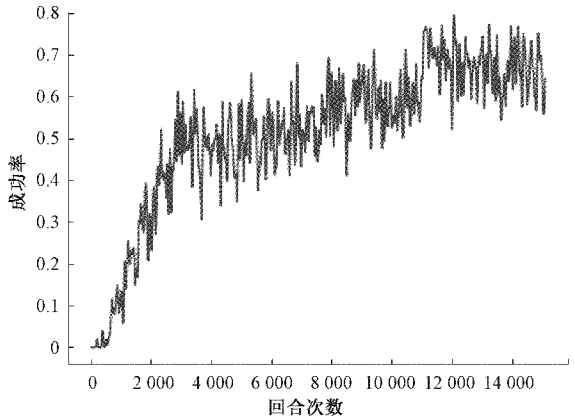


图 7　HERDQN 成功率

高,代表移动机器人在避开障碍物到达目标点时的路径越接近最优。DQN 算法在稀疏奖励的环境下,训练的前期负面奖励的样本过多,无法高效利用样本,约 7 000 次训练后奖励才逐渐收敛,而 HERDQN 算法,通过将已到达的状态映射为新的目标,并替换原目标,大部分回合都可以实现目标,因此奖励变得稠密,大大提高了样本使用效率,因此在约 1 000 次训练后奖励值就逐渐收敛。

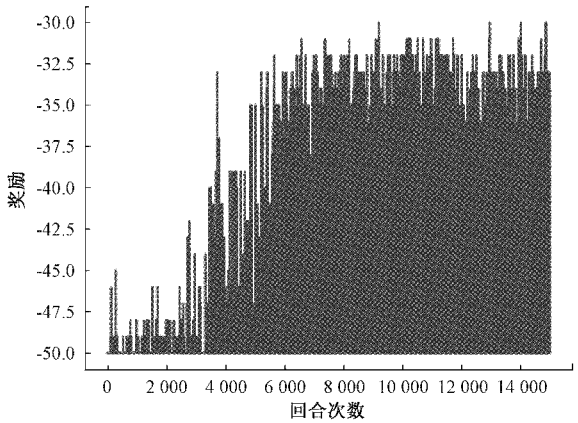


图 8　DQN 奖励

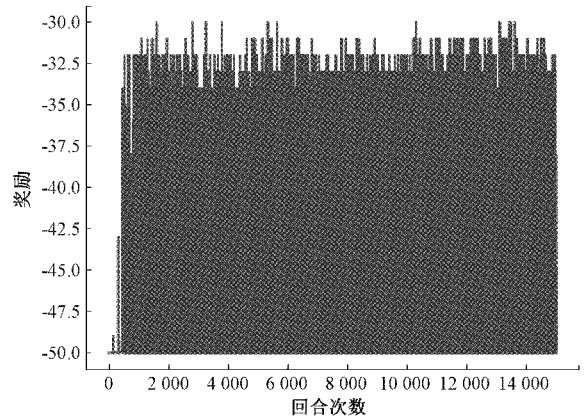


图 9　HERDQN 奖励

4　结　　论

稀疏奖励是强化学习应用在移动机器人路径规划中比较棘手的问题,在稀疏奖励环境下移动机器人与环境交互过程十分耗时,如果交互样本无法获得奖励,那么该样本对于算法训练的贡献将很小。针对这个问题本文提出了基于 DQN 改进的 HERDQN 算法,通过与 HER 算法和 UCB 探索策略的结合,不需要复杂的奖励函数设计,可以更加充分的利用样本,使算法迭代更快,更容易收敛,通过把移动机器人已到达的状态作为新的目标,这样成功的经验可以快速积累,移动机器人的规划路径水平也不断提高,从而克服稀疏奖励带来的问题。在相同环境的仿真实验表明,本文算法比较 DQN 算法在训练过程奖赏值收敛速度以及规划路径的成功率方面都有一定程度的提升和优化。总之,HERDQN 方法可以在较少的迭代中获得较大的奖励值,从而使移动机器人能够在最短的时间内获得最优路径。

参考文献

[1]　闫皎洁,张锲石,胡希平. 基于强化学习的路径规划技术综述[J/OL]. 计算机工程:1-14[2021-07-05].

[2]　林依凡,陈彦杰,何炳蔚,等. 无碰撞检测 RRT* 的移动机器人运动规划方法[J]. 仪器仪表学报,2020,41(10):257-267.

[3]　段建民,陈强龙. 基于改进人工势场-遗传算法的路径规划算法研究[J]. 国外电子测量技术,2019,38(3):19-24.

[4]　李志锟,黄宜庆,徐玉琼. 改进变步长蚁群算法的移动机器人路径规划[J]. 电子测量与仪器学报,2020,34(8):15-21.

[5]　MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. ArXiv Preprint, 2013, ArXiv:1312.5602, 2013.

[6]　LOW E S, ONG P, CHEAH K C. Solving the optimal path planning of a mobile robot using improved Q-learning[J]. Robotics and Autonomous Systems, 2019, 115: 143-161.

[7]　LI S, XU X, ZUO L. Dynamic path planning of a mobile robot with improved Q-learning algorithm[C]. 2015 IEEE international conference on information and automation, IEEE, 2015: 409-414.

[8]　封硕,舒红,谢步庆. 基于改进深度强化学习的三维环境路径规划[J]. 计算机应用与软件,2021,38(1):250-255.

[9]　BAE H, KIM G, KIM J, et al. Multi-robot path planning method using reinforcement learning[J]. Applied Sciences, 2019, 9(15): 3057.

[10]　ZHANG Y, CAI P, PAN C, et al. Multi-agent deep reinforcement learning-based cooperative spectrum

sensing with upper confidence bound exploration[J]. IEEE Access, 2019, 7: 118898-118906.

[11] ANDRYCHOWICZ M, WOLSKI F, RAY A, et al. Hindsight experience replay [J]. ArXiv Preprint, ArXiv:1707.01495, 2017.

[12] 杨惟轶,白辰甲,蔡超,等.深度强化学习中稀疏奖励问题研究综述[J].计算机科学,2020,47(3):182-191.

[13] GUO M, LIU Y, MALEC J. A new Q-learning algorithm based on the metropolis criterion[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2004, 34(5):2140.

[14] 陈学松,杨宜民.基于递推最小二乘法的多步时序差分学习算法[J].计算机工程与应用, 2010(8):52-55.

[15] 张孝军,程银宝,吴军,等.自适应蒙特卡洛法评定量块校准测量不确定度[J].电子测量技术,2020,43(20):84-88.

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

作者简介

王军,工学博士,教授,主要研究方向为工业物联网,工业无线网络,网络软件进化。

E-mail:wj_software@hotmail.com

杨云霄,硕士研究生,主要研究方向为工业物联网,工业移动机器人。

E-mail:1103836698@qq.com

李莉,工学博士,副教授,主要研究方向为水声换能器,水声传感器。

E-mail:779770934@qq.com