DOI:10.19651/j.cnki.emt.2210930

深度强化学习在机器人路径规划中的应用*

邓修朋[1] 崔建明[2] 李 敏[1] 张小军[1] 宋 戈[1]
(1.山东科技大学电子信息工程学院 青岛 266590; 2.山东科技大学计算机科学与工程学院 青岛 266590)

摘 要: 针对深度强化学习算法在路径规划的过程中出现与所处环境交互信息不精确、回馈稀疏、收敛不稳定等问题,在竞争网络结构的基础上,提出一种基于自调节贪婪策略与奖励设计的竞争深度 Q 网络算法。智能体在探索环境时,采用基于自调节贪婪因子的 ε-greedy 探索方法,由学习算法的收敛程度决定探索率 ε 的大小,从而合理分配探索与利用的概率。根据人工势场法物理理论塑造一种势场奖励函数,在目标处设置较大的引力势场奖励值,在障碍物附近设置斥力势场奖励值,使智能体能够更快的到达终点。在二维网格环境中进行仿真实验,仿真结果表明,该算法在不同规模地图下都取得了更高的平均奖赏值和更稳定的收敛效果,路径规划成功率提高了 48.04%,验证了算法在路径规划方面的有效性和鲁棒性。同时与 Q-learning 算法对比实验表明,所提算法路径规划成功率提高了 28.14%,具有更好的环境探索和路径规划能力。

关键词: 路径规划;强化学习;深度强化学习;ε-greedy 策略;人工势场
中图分类号: TP242 **文献标识码**: A **国家标准学科分类代码**: 520.2

Application of deep reinforcement learning in robot path planning

Deng Xiupeng[1] Cui Jianming[2] Li Min[1] Zhang Xiaojun[1] Song Ge[1]
(1. College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China;
2. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: To address the problems of inaccurate interaction information with the environment, sparse feedback, and unstable convergence of deep reinforcement learning algorithms in path planning, a dueling deep Q-network algorithm based on adaptive ε-greedy strategy and reward design is proposed. When exploring the environment, the agent uses the ε-greedy strategy with self-adjusting greedy factor, while the exploration rate ε is determined by the convergence degree of the learning algorithm, so that the probability of exploration and exploitation can be reasonably assigned. According to the physical theory of artificial potential field method, a potential field reward function is created which contains a larger gravitational potential field in the target, a repulsive potential field reward near the obstacle, making the agent to reach the end point faster. Simulation experiments are conducted in a 2D grid environment, the results show that the algorithm achieves higher average reward and more stable convergence under different scale maps, with an improvement of 48.04% in path planning success rate, which verifies the effectiveness and robustness of the algorithm in path planning. The method proposed in this paper is compared with the Q-learning, which has a 28.14% improvement in path planning success rate with better environment exploration and path planning capabilities.

Keywords: path planning;reinforcement learning;deep reinforcement learning;ε-greedy strategy;artificial potential field

0 引 言

由于新一代人工智能科技的快速发展,移动机器人得到普遍的应用[1],而人类对移动机器人需求的提高,使机器人的路径规划技术也成为了热点方向。早期经典路径规划算法有人工势场[2]、A* 算法[3]、快速扩展随机树法[4]、粒子群算法[5]、蚁群算法[6]等。随着研究的深入,对路径规划的速度与准确性要求性更高,但这些传统规划算法存在操作复杂、实时性低、易造成局部最优等缺陷,不具备快速响应复杂环境变化和灵活学习的能力,这使得传统算法遇到了关键难题。

面对传统的路径规划算法存在的问题,强化学习作为

收稿日期:2022-08-02
* 基金项目:山东省自然基金联合基金(ZR2019LZH001)项目资助

人工智能算法被应用到路径规划当中[7]，这种学习方法将序列决策问题转化为马尔可夫模型[8]，算法核心是通过智能体与环境的交互作用，建立环境状态与状态-动作值函数之间的映射，进而获得最优状态-动作值函数，最终得到最佳的动作序列[9]。但在更加复杂和不确定的高维环境下，强化学习中 Q-learning[10] 算法的 Q 值存储方式很难应用于连续的状态空间，其不仅耗时而且收敛速度慢，容易产生维数灾难导致规划能力降低。

面对这些状况，深度强化学习[11]结合了深度学习的感知能力与强化学习[12]的决策能力，有效的解决了维数灾难和收敛缓慢问题，其规划能力符合移动机器人发展的要求。Mnih 等[13]创造性的提出了深度 Q 网络(deep Q-network，DQN)，其存储记忆(experience replay)机制，有效的降低了样本数据的相关性，实现了一部分 Atari 游戏的操控，该算法能够计算给定状态下所有动作的 Q 值，采用其中最佳策略选取动作，从而能够完成任务。在 2015 年又引入了目标网络(target network)[14]，这使得 Q 值的更新方法得以改进，学习过程更加稳定。后续 DQN 的改进算法有很多，例如改良网络的深度双 Q 网络(double DQN，DDQN)算法[15]与竞争架构的 DQN(dueling DQN)算法[16]等，改良抽样方式的优先经验回放 DQN (prioritized replay DQN)算法[17]等。

强化学习是机器学习[18]中的一个领域，目标是让智能体了解要做什么(即如何将当前情况映射到行动中)，以最大限度使收益最大化。智能体动作选择过程并不是已知的，而是靠自己不断的尝试找出哪些动作会带来最大的好处。强化学习策略的养成和智能体与环境交互相互影响，所以算法学习过程中“探索”与“利用”的权衡十分关键。过度试探环境会使智能体规划效率降低，但是可能会带来更长久的收益回报，一味的利用可能会使智能体陷入局部最优，从而会失去长期更高的回报[19]。探索与利用是相互矛盾的，因此提高智能体的规划效率，获取高质量的反馈数据，可以改进学习的策略，进而提高深度强化学习算法的性能。

为了权衡探索与利用之间的关系，已经提出了许多算法。在基于值函数的 DQN 等算法中 ε-greedy 策略得以应用。但是通常将 ε 的设置是不够灵活的，它没有考虑在实际探索中对环境的掌握程度[20]，因此会使算法收敛效果差。再者智能体与环境交互中得到的奖励回报对策略好坏的养成十分重要，目前奖励稀疏问题也使深度强化学习变得更加困难。所以好的奖励函数设计具有重要意义[21]。

因此，本文在 Dueling DQN 算法基础上进行改进与验证，该算法通过优化网络结构使学习性能得到提升；设计一种自调节探索因子的 ε-greedy 策略，合理的调节探索与利用之间的关系。针对奖励稀疏问题，引入人工势场法设计连续的奖励函数，使智能体获得高质量的环境回报数据，从而达到探索的目的和需求。

1　相关工作

1.1　强化学习算法原理

强化学习是一个在当前环境下如何将状态(State)映射为动作(Action)并使奖励(Reward)最大化的过程。在这个设置中，智能体被赋予了对环境的观察，智能体可以通过一组定义的动作来操作。每执行一个动作后，智能体将会获得相应的奖励，目标是找到最优策略。图 1 给出了强化学习的学习过程[22]：

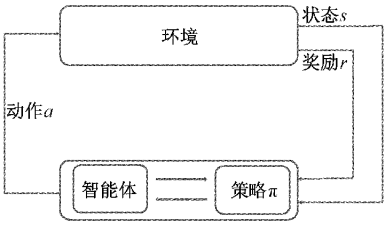


图 1　强化学习模型

马尔可夫决策过程(Markov decision process，MDP)是强化学习算法的关键，它通过状态空间 S，动作空间 A，转移概率分布 P，奖励回报 R 组成的一个四元组$<S,A,P,R>$表征。其中策略为任一个状态 s 下要选取动作 a 的映射，$\pi(a/s)$。在给定策略 π 时对应此样本的折扣累计奖励如式(1)：

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1}$$

$\gamma \in [0,1]$ 为折扣率，用来衡量即时和后续奖励的重要性。

智能体在环境中进行探索，目的是获得最优状态-动作价值函数 $q_*(s,a)$，并定义 $q_*(s,a)$ 为：

$$q_*(s,a) = \max_\pi q_\pi(s,a) \tag{2}$$

最优策略的学习方法可以将最大化 $q_*(s,a)$ 来决定：

$$\pi_*(a \mid s) = \begin{cases} 1, & a = \underset{a \in A(s)}{\operatorname{argmax}} q_*(s,a) \\ 0, & \text{其他} \end{cases} \tag{3}$$

智能体不断地与环境交互，得到最优状态-动作价值函数，从而得到最优策略。但在面临高维环境下，大的状态空间与动作空间使强化学习存储 Q 值得到限制，由此结合深度学习高维感知能力来完成规划任务。

贪婪算法为始终选取使值函数最大的策略。在强化学习中探索和利用的合理分配是一难题。在基于值的强化学习算法中，最普遍使用的是 ε-greedy 策略式(4)：

$$\pi(a \mid s) = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{|A(s)|}, & a = \operatorname{argmax}_a Q(s,a) \\ \dfrac{\varepsilon}{|A(s)|}, & a \neq \operatorname{argmax}_a Q(s,a) \end{cases} \tag{4}$$

该算法是一种随机性策略，每一个动作都有一定的概率被选中，但 ε 的取值一般为固定的，所以探索与利用的分配仍然不合理。

1.2 深度强化学习算法原理

DQN 算法由 Google DeepMind 团队提出,其框架[19]如图 2。深度 Q 网络在面对高维或连续的状态与动作空间时,它使用卷积神经网络来对动作值函数作非线性逼近。在神经网络的更新中,损失函数由目标网络值与主网络的预测值作均方误差进行计算如式(5)所示。

$$L(\theta)=E[(r+\gamma\max_{a'}Q(s',a',\theta)-Q(s,a,\theta))^2] \quad (5)$$

由于 DQN 算法中两个网络中的参数更新不同步,所以利用损失函数进行随机梯度下降来降低误差,同时通过反向传播来调试网络权重 θ,如式(6):

$$\theta_{t+1}=\theta_t+\alpha[r+\gamma\max_{a'}Q(s',a',\theta^-)-Q(s,a,\theta)]\nabla Q(s,a,\theta) \quad (6)$$

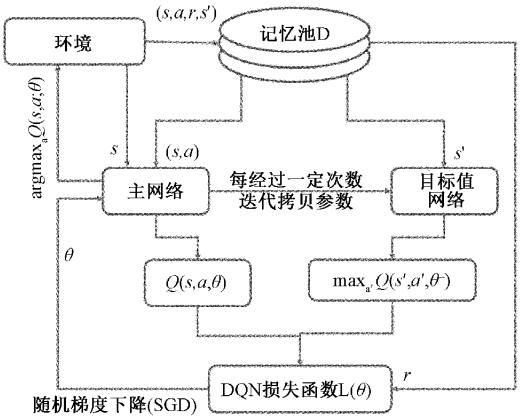


图 2 DQN 算法框架

Dueling DQN 算法属于 DQN 的改进算法,在 Dueling DQN 算法中将神经网络输出层的结构进行优化。Dueling DQN 算法是将 Q 网络划分,首先是仅与状态相关的状态值函数 $V(s;\theta,\beta)$,与动作选择无关。第二个是与状态和动作都相关的优势函数 $A(s,a;\theta,\alpha)$。其中 θ 表示公共部分的网络权重,α 与 β 表示在各自全连接层网络的参数,即分别在最终动作 Q 值中所占的比重。Dueling DQN 网络结构[21]如图 3 所示。

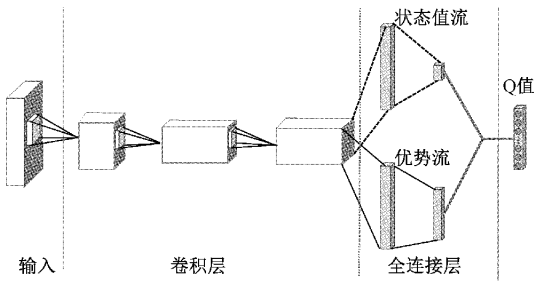


图 3 竞争网络结构

由图 3 结构可知最终价值函数可以表示为:

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha) \quad (7)$$

在这个式子中输出的价值函数遇到了 $V(s;\theta,\beta)$ 与 $A(s,a;\theta,\alpha)$ 可辨识性低的问题,在这里为了解决这个问题,其中最大化操作由平均 $1/|A|$ 替代,新的组合公式为:

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+(A(s,a;\theta,\alpha)-\frac{1}{|A|}\sum_{a'}A(s,a';\theta,\alpha)) \quad (8)$$

使用该方法在不改变 $V(s;\theta,\beta)$ 与 $A(s,a;\theta,\alpha)$ 的同时有效的提高了可辨别性。本文将在 Dueling DQN 算法的基础上进行改进。

1.3 人工势场法

人工势场法是一个完美融合了物理与计算机的算法,最早是解决机械臂移动规划的问题,后来也被应用到移动机器人的路径规划中。

在路径规划任务中算法的思想上假设给目标点与障碍物的周围加上虚拟势场,由产生的引力与斥力的合力促使机器人避开障碍物抵达目标,从而完成规划任务。如图 4 所示[23],在栅格地图中,右下方的圆圈代表目标位置,其对左上方的智能体产生引力,多个黑色的方块代表障碍物,其对智能体产生斥力。合力场的梯度方向为智能体的运动方向,所受引力或斥力的大小与智能体和目标或障碍物之间的距离有关。

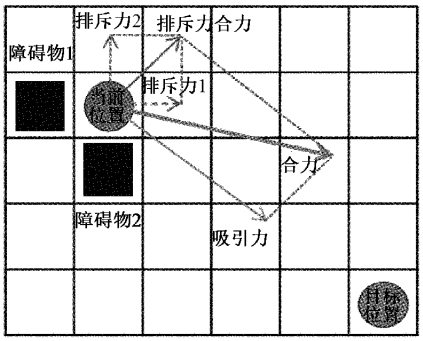


图 4 人工势场受力分析

由目的地产生的虚拟引力势场函数 $U_{att}(d)$ 表示为:

$$U_{att}(d)=\frac{1}{2}\alpha\rho^2(d-d_g) \quad (9)$$

其中,α 为引力势场增益系数,$\rho(d-d_g)$ 为智能体与目标位置之间的欧氏距离。当智能体抵达目标位置时,引力势场为 0。对 $U_{att}(d)$ 求其负梯度得到引力 $F_{att}(d)$,表示为:

$$F_{att}(d)=-\nabla U_{att}(d)=-\frac{1}{2}\alpha\nabla\rho^2(d-d_g)=-\alpha\rho(d-d_g) \quad (10)$$

与障碍物产生的虚拟斥力势场函数 $U_{rep}(d)$ 表示为:

$$U_{rep}(d)=\begin{cases}\frac{1}{2}\beta(\frac{1}{\rho(d-d_o)}-\frac{1}{d_*})^2, & 0<\rho(d-d_o)<d_*\\ 0, & \rho(d-d_o)\geqslant d_*\end{cases} \quad (11)$$

β 为斥力势场的增益系数,d_* 为障碍物的作用范围,$\rho(d-d_o)$ 为智能体与障碍物之间的欧氏距离。当智能体与障碍物的距离 $\rho(d-d_o)$ 超过的 d_* 时,该障碍物的斥力

势场不起作用。对 $U_{rep}(d)$ 求其负梯度得到斥力 $F_{rep}(d)$，表示为：

$$F_{rep}(d)=\begin{cases}\beta\left(\dfrac{1}{\rho(d-d_o)}-\dfrac{1}{d_*}\right)\dfrac{1}{\rho(d-d_o)^2}\nabla\rho(d-d_o), & 0<\rho(d-d_o)<d_*\\ 0, & \rho(d-d_o)\geqslant d_*\end{cases}\tag{12}$$

因此，智能体在环境中受到的合力为所受引力与斥力的矢量和(如图 4 所示)，该合力决定智能体在该位置的运动方向。

2　基于自调节贪婪策略与奖励设计的竞争深度 Q 网络算法

对于基于值的 DQN 及其改进算法等动作选择策略一般采用 ε-greedy 算法。在 Dueling DQN 算法的基础上对 ε-greedy 策略与奖励函数进行改进设计，提出一种基于自调节贪婪策略与奖励设计的竞争深度 Q 网络算法(adaptive ε-greedy and reward design dueling DQN，AGR-DDQN)。

2.1　自调节贪婪因子

在 ε-greedy 策略中，智能体在动作空间中进行每一次选择时，有 ε(ε<1)的概率任意做出一个动作，以 1-ε 的概率做出最大 Q 值对应的最优动作。但是需要智能体对环境进行大量的探索，其反馈信息利用率低，耗费大量的时间，导致智能体的规划效率大大降低。根据存在的问题，本文对 ε-greedy 策略进行改进，如式(13)所示。

$$\varepsilon=\begin{cases}\varepsilon_{lim}, & \dfrac{1}{1+\mathrm{e}^{\bar{R}_L\phi}}>\varepsilon_{lim}\\ \dfrac{1}{1+\mathrm{e}^{\bar{R}_L\phi}}, & \text{其他}\end{cases}\tag{13}$$

$$R_L=\frac{1}{L}\sum_{i=n-L}^{n-1}G_i\tag{14}$$

函数 $\dfrac{1}{1+\mathrm{e}^{\bar{R}_L\phi}}$ 的值域为(0,1)；式(14)中，$\bar{R}_L$ 为 L 个回合累计奖励的平均值，G_i 为第 i 回合累计奖励，$\bar{R}_L$ 的好坏能够反映智能体与环境交互的程度，$\bar{R}_L$ 越大说明当前的动作选择容易学习到最优策略，反之 $\bar{R}_L$ 越小说明当前策略需要增加智能体对环境的探索过程；ϕ 为收缩系数，ϕ 越大越趋向于利用，在不同环境中其设置不同；ε_{lim} 为贪婪因子的上限值，智能体刚进入环境中时以 ε_{lim} 的概率进行探索。

自调节的贪婪因子能够让智能体刚进入陌生环境时积极的探索环境，随着算法的学习，依据算法的收敛程度主动调整探索率的大小，以达到权衡探索和利用的目的。

2.2　人工势场法设计奖励函数

对于室内机器人的导航环境，智能体需要精确躲避室内的障碍物，还要在较短的时间内到达目标位置，所以需要制定一个合理的奖励规则。针对以往奖励函数的设置都存在奖励稀疏、不连续等问题，本文受人工势场法的思想引导，设置势场回报函数，利用人工势场值来缓解稀疏奖励问题。智能体越接近目标点引力势场越大，越接近障碍物斥力势场越大，总势场由引力势场与斥力势场构成。因此，利用势场值可以相对及时的评估智能体的决策。在智能体移动的每一步中，智能体观察状态，然后根据当前的策略选择动作，移动到一个新位置。其由引力势场函数构建目标引导奖励函数为：

$$R_{goal}=K_{att}\cdot d_g^2+r_g\tag{15}$$

其中，K_{att} 为引力奖励系数；d_g 为智能体与目标位置的距离；r_g 为常数，当距离为 0 时奖励为 r_g。这样使得目标引导奖励值由起始位置到终点位置逐渐变大，在目标处取得最大奖励。

依据斥力势场构建障碍躲避奖励函数如下：

$$R_{obsi}=\begin{cases}K_{rep}\cdot\left(\dfrac{1}{d_{obsi}}-\dfrac{1}{d_0}\right)^2, & d_{obsi}\leqslant d_0\\ 0, & d_{obsi}>d_0\end{cases}\tag{16}$$

$$R_{obs}=\sum_{i=0}^{n}R_{obsi}\tag{17}$$

其中，K_{rep} 为斥力奖励系数；d_{obsi} 为智能体与第 i 个障碍物之间的距离，距离越近奖励越小；d_0 为阈值，当智能体与障碍物的距离 d_{obsi} 大于 d_0 时，智能体将不受斥力势场的影响；R_{obs} 为智能体受多个障碍物的累计斥力奖励。依据人工势场法设计奖励函数：

$$Reward=R_{goal}+R_{obs}+R_s+R_c\tag{18}$$

其中奖励函数由 4 个部分组成，R_{goal} 与 R_{obs} 为引力与斥力势场生成的连续性奖励函数，该函数鼓励智能体向着目标前进，使智能体主动靠近目标点，绕开障碍物进行路径规划。R_s 是防止智能体局部振荡的前进奖励或后退惩罚。R_c 为智能体接触到障碍物的碰撞惩罚。

2.3　算法描述

本文针对 ε-greedy 策略进行了改进，提出了自调节贪婪算法，同时引入人工势场法设计了连续性的奖励函数。本文中在 Dueling DQN 算法的基础上进行改进。AGR-DDQN 算法伪代码如下：

算法 1 AGR-DDQN 算法

初始化：初始化 s_t、回放记忆池 D，设置容量 N、随机设置主网络权重 θ、目标网络权重 $\theta^-=\theta$，初始化贪婪因子 ε_{lim}，设置平均累计奖励长度 L，设置阈值 d_0。

for episode=1,M do

　for $t=0,T-1$ do

　　采用 ε-greedy 策略选择动作 a_t

　　环境 s_t 下执行动作 a_t，获得即时奖励 r_t、下一个状态 s_{t+1}，将 (s_t,a_t,r_t,s_{t+1}) 储存到记忆池 D 中，根据式(18)计算奖励

在记忆池中随机选取一小批量的历史元组 (s_t, a_t, r_t, s_{t-1}) 计算目标值:

$$y_i = \begin{cases} r_i, & s_{i\ 1} = \text{最终状态} \\ r_i + \gamma \max_{a'} Q(s_{i+1}, a', \theta), & \text{其他} \end{cases}$$

根据损失函数式(5)计算 $L(\theta)$,执行随机梯度下降(SGD)降低误差并更新 θ

经过 C 次迭代,将主网络参数赋值给目标网络:$\theta^{-} \leftarrow \theta$

end for

if episode % L=0:

根据式(14)计算 L 个 episode 累计奖励平均值

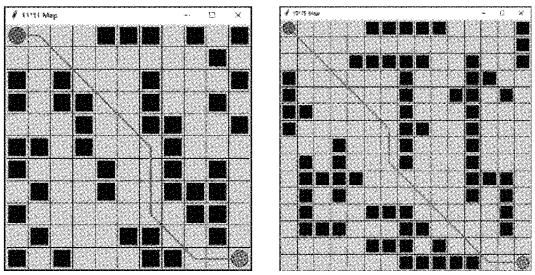
根据式(13)更新 ε

end for

3　实验与分析

3.1　实验环境与设置

所有实验均在配备 Intel(R) Core(TM) i7-7700HQ CPU@2.80 GHz、NVIDIA GeForce GTX 1050 GPU 的计算机上进行。使用 Tensorflow1.8 框架与 python3.6 模拟一个二维的路径研究环境,并将机器人的环境分解为小的网格,机器人的导航空间 S 为小正方形区域,每一个网格代表一种状态。网格地图的规模设置为两个,分别 440×440、600×600 像素,最小移动单元为 40,状态空间分别为 11×11、15×15。环境设置为有边界的状态空间,地图外围为智能体不可达位置,智能体由灰色圆圈表示,黑色正方形是障碍物。图 5(a)、(b)所示为 AGR-DDQN 算法在两种地图中规划的最优路径。



(a) 11×11仿真环境　　(b) 15×15仿真环境

图 5　仿真环境

同时,过少的动作会使路径变得不平滑,过多的动作会使算法难以收敛。我们设置动作的数量改进为 8 个,如图 6 所示。

使用 Tensorflow 框架建立算法模型,在不同规模地图中控制机器人进行模拟仿真,将改进的深度强化学习算法与传统的算法进行比较性实验。

为了验证本文提出算法的有效性,首先在 11×11 仿真环境中设置了 ORI_DDQN[16] 算法、AG_DDQN 算法、APF_DDQN 算法、AGR_DDQN 算法的对比实验以及 Q-learning 算法与 AGR_DDQN 算法的对比实验。然后在

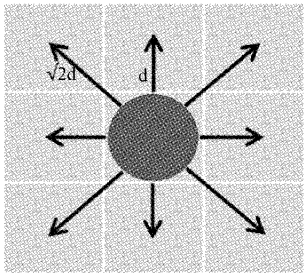


图 6　动作集与步长

15×15 仿真环境中也设置了验证实验。其中 ORI_DDQN 为原始的 Dueling DQN 算法,AG_DDQN 为只引入自调节贪婪因子的 Dueling DQN 算法,APF_DDQN 为只引入人工势场设计奖励函数的 Dueling 算法。实验参数设置如表 1 所示。

表 1　实验参数设置

参数	数值
动作空间 actions	8
训练回合数 episodes	10 000
学习率 learning_rate	0.01
折扣因子 γ	0.9
探索因子 ε	0.9
记忆池 memory_size	3 000
批大小 batch_size	32

APF_DDQN 算法与 AGR_DDQN 算法奖励函数采用式(18),R_{goal} 取值为(−5,5],R_s 为 0.5 或−5,R_c 为−10。其他算法的奖励函数设置为:

$$Reward = \begin{cases} 5, & \text{到达目标} \\ -10, & \text{碰撞惩罚} \\ 0, & \text{其他} \end{cases} \tag{19}$$

AG_DDQN 算法与 AGR_DDQN 算法引入的自调节贪婪因子参数如下:ε_{lim} =0.3,L=200。

通过以上的实验环境和表 1 的参数设置,使用上述不同的算法,本文设置了平均奖励积累、收敛平稳性和规划时间等方面的实验。

3.2　实验结果与分析

在图 7 中(a)～(d)依次表示在 11×11 仿真环境中不同算法的每 50 个回合的平均奖励收敛情况。比较图 7(a)、(b)可以发现,AG_DDQN 算法平均奖励高于原始 ORI_DDQN 算法,这是因为 AG_DDQN 算法引入自调节贪婪因子,使得智能体能够利用奖励反馈进行自适应权衡探索与利用,从而获得了更大的收益。图 7(c)APF_DDQN 算法中设计了连续性的奖励函数,初期智能体获得了额外的负向势能奖励,使得平均奖励较低。但随着训练的进行,智能体从环境中获得更精确的反馈数据,取得了较高的平均奖

励。图 7(d)中 AGR_DDQN 算法在训练后期平均奖励更高并且趋于收敛，有效的提高了智能体的规划能力。

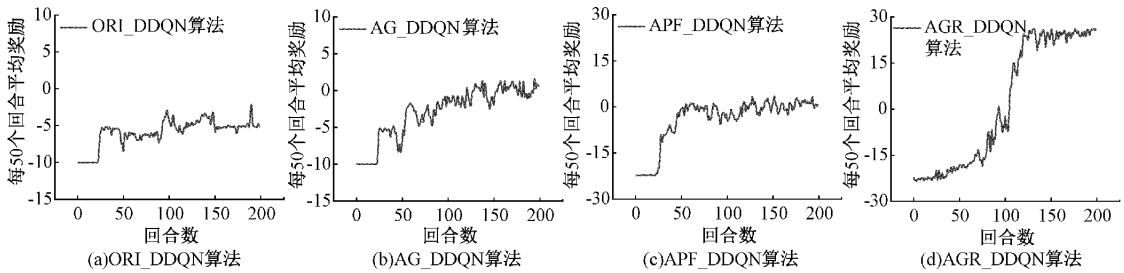


图 7　四种算法的每 50 个回合的平均奖励

图 8(a)～(d)为 4 种算法每个回合的步数变化。经过多次实验表明，在训练的前期 4 种算法都将对环境进行一定的探索。由图 8(a)～(c)可知 AG_DDQN 和 APF_DDPG 算法与 ORI_DDQN 算法相比，其步数震荡较小，具有较好的稳定性；AGR_DDQN 与其他 3 种算法相比，前期震荡较小并且随着训练的进行，其步数震荡更小，收敛速度更快，其稳定性更好。

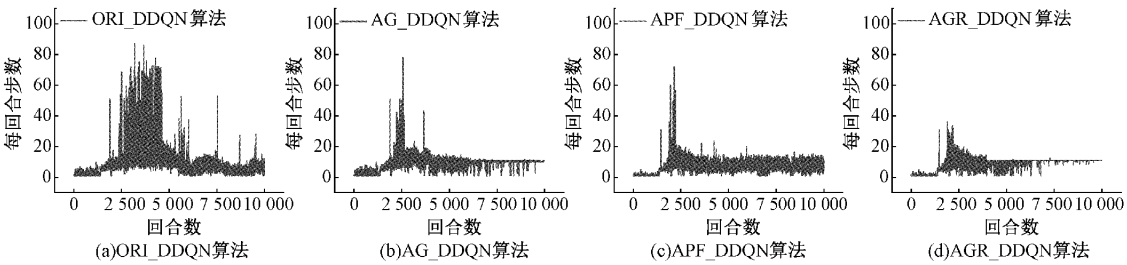


图 8　四种算法的每回合步数

表 2 为不发生碰撞情况下 4 种算法路径规划成功率对比。在训练的 10 000 个回合中进行成功率的计算，AGR_DDQN 算法成功率 54.83%明显高于其他算法，说明该算法让智能体在前期进行了更有益的环境探索，这使得智能体后期路径规划成功率更高。图 9(a)～(d)为 4 种算法到达目标位置的路径长度收敛图，横坐标为成功到达目标的次数。可以看到在刚开始到达目标时路径长度都存在一定的波动。随着迭代次数的增加，路径长度均呈现下降的趋势，这是因为在算法的初期智能体要尽可能多的对环境进行探索过程，从环境中得到更多的数据。由图 9(a)可知 ORI_DDQN 算法，在没有引入自调节贪婪因子的状态下，在后段仍然存在频繁的波动，而且在规定的 10 000 个训练回合内不发生碰撞到达目标点的次数不足 700，明显少于后 3 种算法。随着迭代次数的不断增加，本文提出的算法根据从环境中得到的反馈，自动调节贪婪因子，逐渐减小探索的概率，使迭代后期路径长度收敛速度更快且更稳定，证明了 AGR_DDQN 有更强的规划能力。

表 2　路径规划成功率对比

算法	回合数	到达目标次数	成功率/%
ORI_DDQN	10 000	679	6.79
AG_DDQN	10 000	3 728	37.28
APF_DDQN	10 000	2 731	27.31
AGR_DDQN	10 000	5 483	54.83

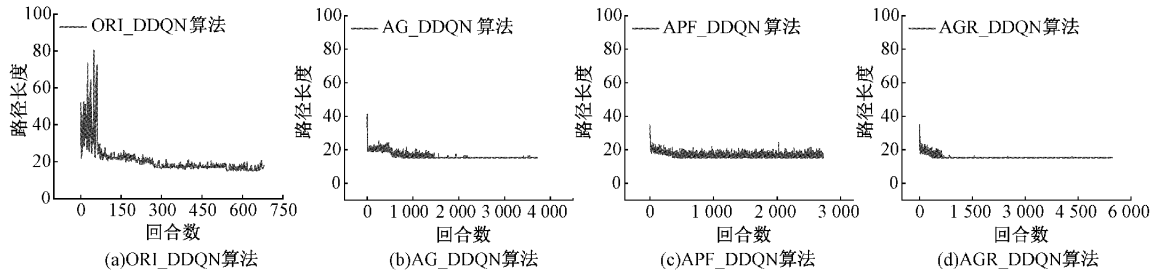
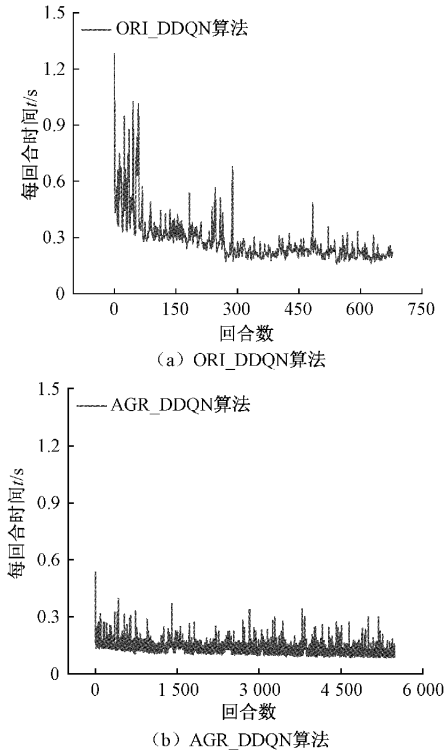


图 9　四种算法的路径规划长度

进一步对比算法成功规划路径的运行效率，图 10(a)为 ORI_DDQN 算法的运行时间图 10(b)为 AGR_DDQN 算法的运行时间，两种算法均有随着迭代次数的增加，运行时间都有下降的趋势，但图 10(b)运行时间少于图 10(a)，体现了改进的算法对环境具有更强的适应性，学习效率更高。



（a）ORI_DDQN算法

（b）AGR_DDQN算法

图 10　两种算法的规划时间

图 11 为 AGR_DDQN 算法与 Q-learning 算法平均每 50 个回合路径规划成功率的比较实验。随着迭代次数的增加，AGR_DDQN 算法的平均成功率呈上升趋势且明显高于 Q-learning 算法，整体提高 28.14%，并且表现出良好的收敛性稳定性。说明所提算法具有更好的自主学习能力和更强的路径规划能力。

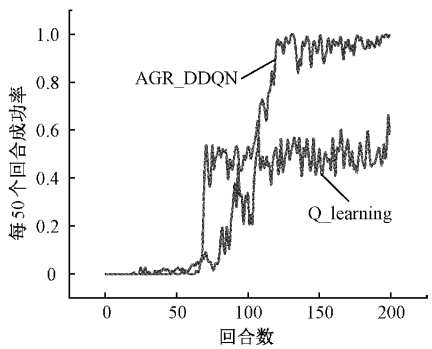
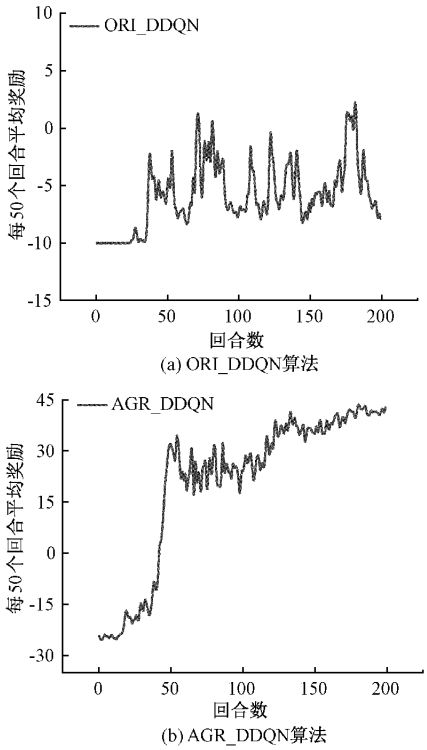


图 11　AGR_DDQN 和 Q-learning 算法对比

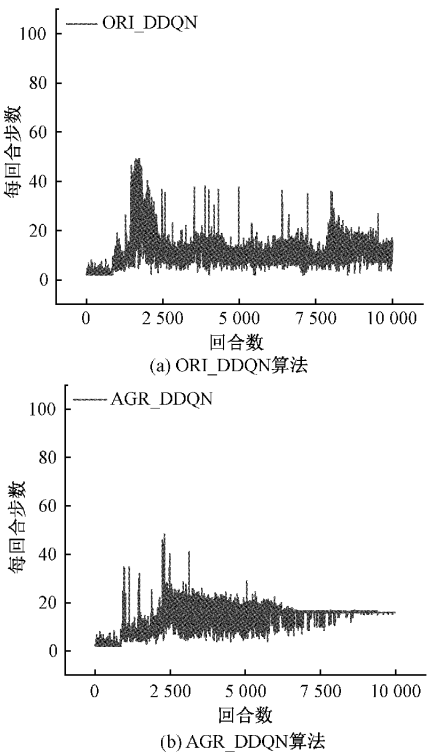
最后，在 15×15 的地图中 AGR_DDQN 和 ORI_DDQN 算法进行了奖励收敛和每回合运动步数的对比实验。观察图 12(a)、(b)，ORI_DDQN 算法一直存在强烈的震荡从而收敛效果较差，而 AGR_DDQN 算法，前期波动用于环境的探索，后期趋于稳定使奖励明显高于 ORI_DDQN 算法。在图 13(a)、(b)进行每回合步数比较，图 13(a)收敛趋势较小而图 13(b)算法后期步数趋于收敛。实验证明，在不同规模的地图环境中，本文的算法的奖励与每一回合运动步数均比 ORI_DDQN 算法收敛速度更快更稳定，路径规划能力更强，适应性更强。



(a) ORI_DDQN算法

(b) AGR_DDQN算法

图 12　不同算法的每 50 个回合的平均奖励



(a) ORI_DDQN算法

(b) AGR_DDQN算法

图 13　不同算法的每回合步数

4　结　　论

为了引导智能体规划出最优路径，提高其规划能力，本文在竞争网络的基础上提出了 AGR_DDQN 算法，使用自调节贪婪因子改进了动作选择策略并利用人工势场法优化了奖励函数，并将提出的改进算法应用到不同规模的环境中进行实验，同时将所提算法与 Q-learning 算法进行对比。实验结果表明该算法改善了探索与利用的关系并为智能体提供了连续的奖励函数，在算法效率与收敛稳定性上取得了一定的成效，从而提高了智能体的规划能力。下一步工作将会移植到其他的强化学习算法中验证算法的有效性。

参考文献

[1] ZHANG J, XIA Y, SHEN G. A novel learning-based global path planning algorithm for planetary rov-ers[J]. Neurocomputing, 2019,361:69-76.

[2] 朱颖，李元鹏，张亚婉，等. 基于改进人工势场法的搬运机器人路径规划[J]. 电子测量技术，2020,43(17):101-104.

[3] 张建光，张方，陈良港，等. 基于改进 A* 算法的自动引导车的路径规划[J]. 国外电子测量技术，2022,41(1):123-128.

[4] 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.

[5] 谢勇宏，孔月萍. 基于改进粒子群算法的三维路径规划[J]. 计算机测量与控制，2022,30(3):179-182,191.

[6] 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.

[7] 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]. 计算机工程，2021,47(10):16-25.

[8] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[9] LV L, ZHUANG S, DING D, et al. Path Planning via an Improved DQN-Based Learning Policy[J]. IEEE Access, 2019, 7: 67319-67330.

[10] 王科银，石振，杨正才，等. 改进强化学习算法应用于移动机器人路径规划[J]. 计算机工程与应用，2021,57(18):270-274.

[11] 张荣霞，武长旭，孙同超，等. 深度强化学习及在路径规划中的研究进展[J]. 计算机工程与应用，2021,57(19):44-56.

[12] POLYDOROS A S. Survey for model-based reinforcement learning: applications on robotics[J]. Journal of Intelligent and Robotic Systems, 2017, 86(2):153-173.

[13] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. Computer Science, 2013,15(13):123-131.

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

[15] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. Computer Science,2015,18(7):234-245.

[16] WANG Z, SCHUAL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]. Proceedings of International Conference on Machine Learning, 2016: 1995-2003, DOI: 10.48550/arXiv.1511.06581.

[17] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[C]. Proceedings of International Conference on Learning Representations, 2016:1-21, DOI:10.48550/arXiv.1511.05952.

[18] MURPHY K P. Machine learning: a probabilistic perspective[J]. Chance,2012,27(2):62-63.

[19] 尹旷，王红斌，方健，等. 基于强化学习的移动机器人路径规划优化[J]. 电子测量技术，2021,44(10):91-95.

[20] 杨彤，秦进. 基于平均序列累计奖赏的自适应 ε-greedy 策略[J]. 计算机工程与应用，2021,57(11):148-155.

[21] LI J, LIU Y, Deep reinforcement learning based adaptive real-time path planning for UAV[C]. 2021 8th Internati-onal Conference on Dependable Systems and Their App-lications(DSA), 2021:522-530, DOI: 10.1109/DSA52907.2021.00077.

[22] 毛国君，顾世民. 改进的 Q-Learning 算法及其在路径规划中的应用[J]. 太原理工大学学报，2021,52(1):91-97.

[23] 陈满，李茂军，李宜伟，等. 基于深度强化学习和人工势场法的移动机器人导航[J]. 云南大学学报(自然科学版)，2021,43(6):1125-1133.

作者简介

邓修朋，硕士研究生，主要研究方向为深度强化学习，机器人路径规划，人工智能。

E-mail:skdglendxp@163.com

崔建明，博士，副教授，主要研究方向为集成电路设计、物联网技术、嵌入式系统。

E-mail:cuijm399@163.com

宋戈(通信作者)，硕士，讲师，研究方向为深度强化学习、路径规划、动作识别、物联网技术。

E-mail: songge@sdust.edu.cn