许莉, 何开晟, 刘海容,等. 基于 MADDPG-R 的行人路径规划算法研究[J]. 智能计算机与应用,2025,15(12):107-112.
DOI:10.20169/j.issn.2095-2163.24031601

基于 MADDPG-R 的行人路径规划算法研究

许 莉[1], 何开晟[1], 刘海容[1], 向 进[1], 杨 婷[2], 陈思凡[3]
(1 福州大学 物理与信息工程学院, 福州 350108;2 同济大学 建筑与城市规划学院, 上海 200092;
3 福州大学至诚学院, 福州 350002)

摘 要: 随着城市化进程的加速,城市交通堵塞问题日益显著,尤其是在人口密集的城市中心区域,如何实现行人路径的有效规划,是一个亟待解决的问题。将强化学习算法应用于多智能体协同路径规划中,可以解决传统智能体路径规划方法在复杂环境场景下应用的局限性,本文提出了一种基于改进奖励机制下的多智能体确定性策略梯度算法(Multi-Agent Deep Deterministic Policy Gradient with Reward Enhancement,MADDPG-R),在多智能体深度确定性策略梯度算法的基础上,设计一个新的奖励机制,能够有效应对多智能体环境中的复杂情况,保障系统运行的实时性。同时,本文还设计了一个动态的仿真场景,并在二维环境中进行了仿真实验,验证了该算法的有效性。

关键词: 强化学习; MADDPG-R; 路径规划; 多智能体
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)12-0107-06

Research on pedestrian path planning algorithm based on MADDPG-R

XU Li[1], HE Kaisheng[1], LIU Hairong[1], XIANG Jin[1], YANG Ting[2], CHEN Sifan[3]
(1 College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China;
2 College of Architecture and Urban Planning, Tongji University, Shanghai 200092, China;
3 College of Zhicheng, Fuzhou University, Fuzhou 350002, China)

Abstract: With the acceleration of urbanization, the problem of urban traffic congestion is becoming increasingly prominent, especially in densely populated urban centers. How to achieve effective planning of pedestrian paths is an urgent issue that needs to be addressed. Applying reinforcement learning algorithms to multi-agent collaborative path planning can solve the limitations of traditional agent path planning methods in complex environmental scenarios. This paper proposes a multi-agent deep deterministic policy gradient with reward enhancement (MADDPG-R) based on an improved reward mechanism. On the basis of the multi-agent deep deterministic policy gradient algorithm, a new reward mechanism is designed to effectively cope with complex situations in multi-agent environments and ensure real-time system operation. Meanwhile, this article also designed a dynamic simulation scenario and conducted simulation experiments in a two-dimensional environment to verify the effectiveness of the algorithm.

Key words: reinforcement learning; MADDPG-R; path planning; multi-agent

0 引 言

根据 2023 年中华人民共和国国民经济和社会发展统计公报数据显示,中国人口城镇化率从 1978 年的 17.92%提升至 2023 年的 66.2%,城市化已成为 21 世纪不可逆且影响深远的趋势[1]。智能体作为人工智能领域的重要研究方向,其系统应用场景日趋复杂,这对智能体的路径规划能力提出了更高要求。

传统智能体路径规划方法过度依赖环境信息,不仅需要事先处理已知环境信息,且过程耗时显著,同时无法确保所得路径规划算法具备自适应性。在

基金项目: "十四五"国家重点研发计划课题(2022YFC3800205)。

作者简介: 许 莉(1998—),女,硕士,主要研究方向:多智能体路径规划,强化学习; 何开晟(1999—),男,硕士,主要研究方向:强化学习,加密算法; 刘海容(1999—),男,硕士,主要研究方向:联邦学习,智能体轨迹预测; 向 进(2000—),男,硕士,主要研究方向:路径规划,智能算法; 陈思凡(1997—),男,硕士,助教,主要研究方向:无人机智能规划,机器学习。
通信作者: 杨 婷(1983—),女,博士,副研究员,主要研究方向:智能仿真,城市建模。Email: 2002yangting@tongji.edu.cn。

收稿日期: 2024-03-16

此背景下,强化学习为环境未知情况下的路径规划问题提供了有效解决方案。强化学习算法是一种无需依赖周围环境信息和智能体自身存储知识的机器学习方法,该算法的出现为环境未知情况下的路径规划问题提供了有效解决方案[2-3]。

强化学习是早期人工智能领域的重要研究分支,强化学习算法无需预先获取环境信息和智能体自身的知识储备,为智能体路径规划问题提供了新的解决思路[4-5]。Q-learning 算法、深度 Q 网络(Deep Q-Network,DQN)算法、策略梯度(Policy Gradient,PG)算法均为目前应用较多的强化学习算法[6-8]。针对智能体路径规划问题,邓修朋等[9]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,解决了深度强化学习算法在环境交互中收敛不稳定的问题,但使用贪婪策略进行探索效率较低;史殿习等[10]在 DQN 算法的基础上,提出了基于 Dueling 的 Q 值计算优化机制,使得 Q 值的计算更加简单准确;司鹏搏等[11]提出基于网络剪枝的近端策略优化算法,提高了训练效率;Guo 等[12]提出了一种将深度确定性策略梯度算法与人工势场相结合的改进路径规划算法,该算法具有良好的收敛速度。Wang 等[13]提出了一种全局引导的强化学习方法,通过设计全新的奖励结构,以全分布的方式解决了移动机器人的路径规划问题;Chen 等[14]则采用深度强化学习算法中的软行为者批评(Soft Actor-Critic,SAC)方法,针对机械臂的动态避障路径规划进行了深入研究。然而,这类算法在多智能体环境中的应用面临较大挑战,一个主要的原因是智能体在训练的过程中其环境是动态变化的,智能体在这种环境中无法学习到有效的策略。多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)算法为复杂环境下的多智能体路径规划问题提供了新的解决方案[15-16]。多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient,MADDPG)算法是一种 MARL 算法,在多智能体协同工作中表现出良好的性能,并随着训练集数量的增加,该算法适应不断变化的复杂环境的能力也越来越强[17-18]。

本文针对复杂环境下的行人智能体动态路径规划问题,提出了一种改进奖励机制下的 MADDPG-R 算法。首先,将行人路径规划问题看成一个最短路径问题;其次,针对智能体运行过程中可能出现的碰撞问题建模;最后,通过设定奖励机制,为智能体规划出一条最优的无碰撞路径,提升算法的鲁棒性和适应性。

1 系统模型和问题公式

本文考虑了一种 N 个智能体在复杂环境中运行的二维场景,包含障碍物、目标点以及其他移动的智能体,这些智能体在复杂的环境中进行点对点路径规划,行人智能体的路径规划环境如图 1 所示。

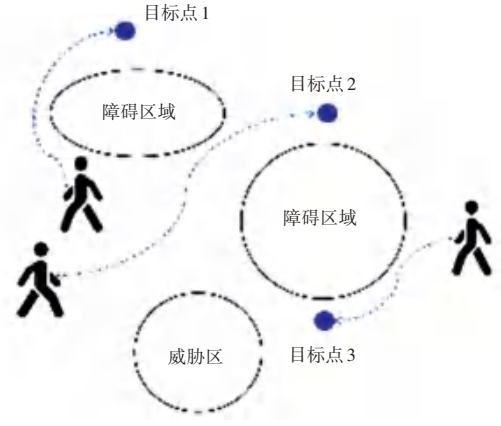


图 1 行人智能体路径规划环境

Fig. 1 Pedestrian agent path planning environment

将整个运行空间设置在笛卡尔坐标系下,空间中任意点 P 坐标表示为 (x,y),起始点 S 和目的地点 D 分别表示为 $S=(x_s,y_s)$ 和 $D=(x_d,y_d)$。在智能体训练过程中,障碍物的位置是随机生成的,并且障碍物会限制智能体的移动。将智能体在二维空间中的无碰撞轨迹路径定义为 T_{path},则智能体在二维空间的路径规划定义为从起始点、目的地点和中间任意点集合的函数,可以表示为:

$$\begin{cases} f(T_{\text{path}}) = f(S,p_1,p_2,\cdots,p_n,D), \\ p_i \in P \end{cases} \tag{1}$$

行人路径规划是一个复杂的优化问题,需对该优化问题展开系统性分析。在行人的路径规划过程中,环境信息并不全都是已知的,会遇到一个随机的复杂危险环境,如障碍物或者其他智能体的阻碍等。在行进过程中必须考虑这些因素,避免这些威胁区域对智能体路径规划的影响。因此,本文设计了优化条件。

1.1 最短行程设计

强化学习算法的优化目标是奖励最大化,与奖励密切相关的一个因素就是距离目标点的距离,因此,本文将行人路径规划问题定义为一个最小化路程问题,在保证安全前进的同时,通过减少相应的轨迹长度来降低总体的路径规划成本,用下式表示:

$$f_i = \sum_{i=1}^{n} D_i \tag{2}$$

其中，f_i 为总行程距离；n 为在该行程中行人的总步数；D_i 为行人每一步的行进距离，公式如下：

$$D_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \tag{3}$$

1.2　防撞系统设计

在复杂的路径规划问题中，碰撞是一个主要的问题。当行人与行人之间或行人与障碍物之间发生碰撞时，系统通常会给予其一个负向的奖励，以此作为对不良行为的惩罚。根据 Hall 等[19]提出的同心圆理论，本文设计了一种新的防撞系统，即在环境中为障碍物与行人添加关键区域，形成早期碰撞预警机制，防撞环境如图 2 所示。

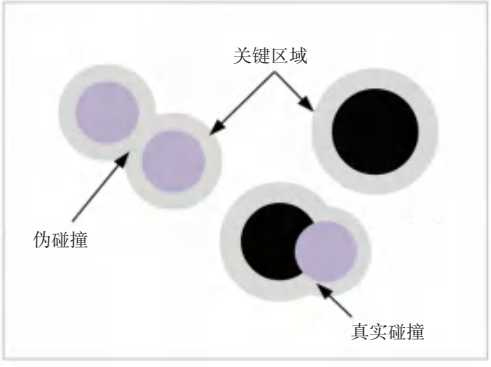


图 2　碰撞环境

Fig. 2　Collision environment

一般的碰撞检测标准是智能体与障碍物之间的距离小于二者半径之和，添加关键区域后，相当于设置了一个早期碰撞预警机制，当关键区域相互碰撞时，智能体会获得负奖励，这样就可以为智能体提供一定的避碰反应时间。

(1)伪碰撞

两个关键区域之间的碰撞为伪碰撞。定义 r_1 为智能体的半径，r_2 为静态障碍物的半径，α 表示关键区域的宽度，l_1 代表两个智能体之间的距离，d_{iw} 代表智能体与障碍物之间的距离，当 $r_1 + r_2 < d_{iw} \leqslant r_1 + r_2 + \alpha$ 时，称智能体与障碍物之间发生了伪碰撞，带伪碰撞区域的智能体与障碍物之间的位置关系如图 3 所示。

图 3　带伪碰撞区域智能体与障碍物位置关系

Fig. 3　Relationship between intelligent agents with pseudo collision zones and obstacle positions

相切与相交情况均认为在该环境中发生了伪碰撞，用 d_{ij} 表示智能体之间的实际距离，在智能体进行交互的环境中，当 $2r_1 < d_{ij} \leqslant 2r_1 + \alpha$ 时，称智能体之间发生了伪碰撞。

(2)真实碰撞

当智能体之间相互碰撞或者智能体与障碍物发生碰撞时，称为真实碰撞，智能体与障碍物位置关系如图 4 所示。当 $d_{iw} \leqslant r_1 + r_2$ 时，称智能体与障碍物之间发生了真实碰撞。

图 4　智能体与障碍物位置关系

Fig. 4　Position relationship between intelligent agents and obstacles

当 $d_{ij} \leqslant 2r_1$ 时，称智能体与智能体之间发生了真实碰撞。在训练过程中，通过惩罚伪碰撞的发生来减少真实碰撞的次数，使得智能体可以在避免碰撞的同时更加快速的到达目标点。

2　改进的奖励机制下的多智能体确定性策略梯度算法

2.1　MADDPG-R 算法

MADDPG-R 算法模型如图 5 所示。每个智能体包含两个网络：Actor 网络和 Critic 网络 。Actor 网络根据智能体输入的状态，输出将要执行的动作；Critic 网络对 Actor 网络计算出来的动作进行评价，计算动作的价值。在该模型中也存在经验回放缓冲区，用于存储一定数量的训练经验，Q 在更新网络时随机从中进行批量读取，打破数据之间的相关性，从而使训练的结果更加稳定。在训练阶段，Actor 网络只从自身获取观察信息，而 Critic 网络获取其他智能体的动作和观察等信息。在执行阶段，不涉及 Critic 网络，每个智能体只需要一个 Actor 网络，即执行是分散的。MADDPG-R 算法可以被认为是深度确定性策略梯度的多智能体版本，核心思想是集中训练和分散执行，不仅适用于合作，而且适用于竞争性或合作竞争混合的环境。

图 5 中的 π 为智能体策略，设定其参数为 θ，$J(\theta)$ 为第 i 个智能体的累计期望奖励 $E[R_i]$，o_i 表示第 i 个智能体的观测，$x = [o_0, \cdots, o_i]$ 为状态，Q_i^{π}

为价值网络。因此,第 i 个智能体的梯度可以表示为:

$$\nabla_{\theta_i} J(\theta_i) = E_{S\sim P^{\mu}, a_i\sim\pi_i} [\nabla_{\theta_i} \log\pi_i(a_i \mid o_i) Q_i^{\pi}(x, a_1, \cdots, a_N)] \quad (4)$$

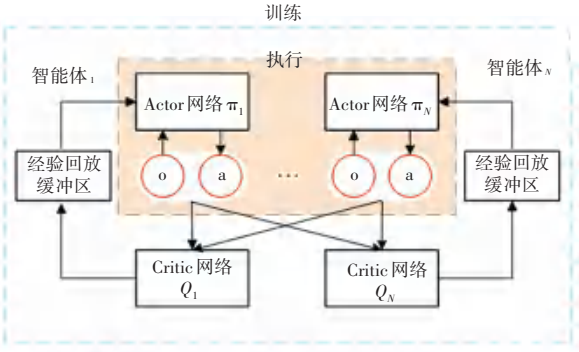


图 5 MADDPG-R 算法模型

Fig. 5 MADDPG-R algorithm model

2.2 构建多智能体系统

本文基于 MADDPG 算法框架构建多智能体系统,并对 MADDPG 算法中的奖励函数进行改进,使得该智能体系统可以在路径规划的同时实现目标分配,并应对训练中因环境变化而产生的动态环境。在该系统中,每一个行人都被抽象为一个智能体,该智能体与行人具有相同的动作模型。

本文考虑了一种 N 个智能体的场景,每个智能体使用的策略参数为 $\theta = \{\theta_1, \theta_2, \cdots, \theta_N\}$,并设所有智能体确定性策略的集合为 $\mu = \{\mu_{\theta_1}, \mu_{\theta_2}, \cdots, \mu_{\theta_N}\}$。因此,每个智能体的确定性策略为:

$$\nabla_{\theta_i} J(\mu_i) = E_{x,a\sim D} [\nabla_{\theta_i} \mu_i(a_i \mid o_i) \nabla_{a_i} Q_i^{\mu}(x, a_1, a_2, \cdots, a_N)\mid_{a_i=\mu_i(o_i)}] \quad (5)$$

其中,$Q_i^{\mu}(x, a_1, a_2, \cdots, a_N)\mid_{a_i=\mu_i(o_i)}$ 为价值函数,a_i 为每个智能体的行动,o_i 为智能体 i 的当前状态,包括智能体 i 与每个目标的距离、智能体 i 与其他智能体的距离、智能体 i 距离威胁区域的距离,D 是经验回放缓冲区,记录所有智能体的经验,包含 $(x, x', a_1, \cdots, a_N, r_1, \cdots, r_N)$,$x'$ 为多智能体系统执行动作之后的新动态,r_i 为第 i 个智能体的奖励。价值网络的梯度 ∇_{θ_i} 更新,公式如下:

$$\nabla_{\theta_i} \mathcal{L}(\theta_i) = E_{x,a,r,x'}[(Q_i^{\mu}(x, a_1, a_2, \cdots, a_N) - R)^2] \quad (6)$$

其中,γ 是折扣因子,R 为目标函数,公式如下:

$$R = r_i + \gamma Q_i^{\mu'}(x', a_1', a_2', \cdots, a_N')\mid_{a_i'=\mu_i'(o_i)} \quad (7)$$

为了获得其他智能体的策略,实现更加稳健的规划,本文为每个智能体 i 的策略设置一个经验缓冲区 D_i,由此,可以得到 Actor 网络的最小化梯度,如下式:

$$\nabla_{\theta i} J(\theta_i) = \frac{1}{S} E_{x,a\sim D_i} [\nabla_{\theta_i} \mu_i(o_i^j) \nabla_{a_i} Q_i^{\mu}(x^j, a_1^j, \cdots, a_N^j)\mid_{a_i=\mu_i(o_i^j)}] \quad (8)$$

其中,S 为样本的随机小批量大小,j 为样本的指数。

通过多智能体系统的构建,路径规划问题可以得到较好的解决。

2.3 奖励机制设定

在智能体系统中,奖励是衡量强化学习算法好坏的关键因素,强化学习算法优化目标就是达到奖励值的最大化。奖励函数的设置,会决定强化学习算法的收敛速度和收敛程度。本文第 i 个智能体的奖励函数 r_i 由 3 部分组成:

$$r_i = r_l + r_c + r_b \quad (9)$$

其中,r_l 表示最短行程奖励;r_c 表示伪碰撞奖励;r_b 表示真实碰撞奖励。

(1)最短行程奖励

在路径规划中,距离是首要考虑的问题。本文设计了一个奖励函数 r_l,用于实现智能体行程的最小化。在多智能体系统中,智能体通过遍历环境中的目标点确定距其最近的目标,并通过取该距离的相反数作为奖励值,距离越远,奖励值越小;距离越近,奖励值越大。通过这种取相反数的方式,可实现优化目标的统一:

$$r_l = -f_i \quad (10)$$

(2)伪碰撞奖励

智能体与智能体之间、智能体与障碍物之间的伪碰撞奖励旨在通过给予智能体在接近潜在碰撞时的负向奖励,从而有效减少真实碰撞的发生次数,进而提升整个系统的稳定性和安全性。伪碰撞奖励设置如下:

$$r_c = \begin{cases} -2, & \text{if } 2r_1 < d_{ij} \leqslant 2r_1 + \alpha \\ -2, & \text{if } r_1 + r_2 < d_{iw} \leqslant r_1 + r_2 + \alpha \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

其中,d_{ij} 表示两个智能体之间的实际距离;d_{iw} 表示智能体与障碍物之间的实际距离;α 代表临界区域宽度。

(3)真实碰撞奖励

设置碰撞奖励主要是为了减少在环境中出现碰撞的现象,主要包括智能体之间的碰撞、智能体与障碍物之间的碰撞,实现智能体安全迅速的到达目标点。真实碰撞奖励可表示为:

$$r_b = \begin{cases} -2, & \text{if } d_{ij} \leqslant 2r_1 \\ -2, & \text{if } l_2 \leqslant r_1 + r_2 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

3　仿真实验与分析

3.1　实验环境及评价指标

本文所有的实验都在 windows10/Ubuntu 16.04 双系统的台式计算机上运行，并且在该系统上创建智能体交互的环境，实现深度强化学习算法。实验的硬件环境为 Intel Xeon E5-2678V3，NVIDIA GeForce RTX 1050 Ti 处理器。为了提高训练速度，降低训练时间，本文实验采用多进程并行训练的方式。

实验基于 Gym 库来实现智能体探索的状态环境的设计，本文为行人多智能体系统设计了一个模拟训练环境，由代表行人的智能体、目标点和障碍物区域组成，以环境中心为坐标原点，环境中各象限的长度为 1 建立坐标系，其参数设置见表 1。

表 1　训练环境设置

Table 1　Training environment setting

参数	半径大小
智能体	0.005
目标点	0.075
障碍物	0.200

为了全面衡量训练过程的有效性，本文采用多维度指标对 MADDPG-R 算法的训练表现进行评价，分别为平均奖励值（Mean Reward，MR）、当前行人与其他行人之间的平均碰撞次数（Mean Collision with Agents，MCA）、行人与障碍物之间的平均碰撞次数（Mean Collision with Obstacles，MCO）以及行人到达目标点的平均步长（Mean Stepsize，MS）。基于上述指标进一步计算碰撞率（Collision Rate with Agents，CRA）和行人与障碍物之间的碰撞率（Collision Rate with Obstacles，CRO），公式如下：

$$\mathrm{CRA} = \frac{\mathrm{MCA}}{\mathrm{MS}} \tag{13}$$

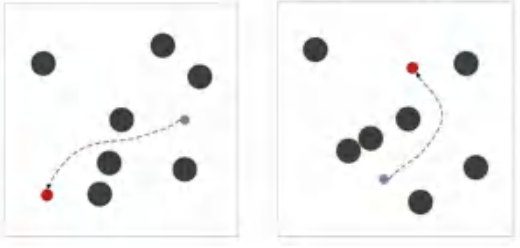
$$\mathrm{CRO} = \frac{\mathrm{MCO}}{\mathrm{MS}} \tag{14}$$

3.2　训练过程

所有智能体在每个训练周期内步长为 25 000，一个训练周期为 1 000 次 episodes，训练集包含 250 000 个样本。在运行过程中，智能体根据当前环境以及策略选择所要执行的动作，当所有智能体完成动作之后，系统会进入到下一个状态。当智能体在环境中完成任务或者步长达到 25 000 步，一个训练周期结束。

3.3　实验结果

对防撞系统优化设计中引入的伪碰撞区域进行有效性测试，分别在环境中引入关键区域 α 与不引入该区域进行对比实验，称未引入伪区域的实验为基线实验。在本次实验中，只考虑单个行人与单个目标点的环境，障碍物数量为 7 个，且实验中行人的起点位置、终点位置以及环境中障碍物位置均是随机生成。基线实验和含关键区域的对比实验结果如图 6 所示，紫色表示的是行人智能体，红色表示目标点，黑色表示障碍物。图 6(a)是在基线实验中行人从起点到终点的位置轨迹图，图 6(b)为带有关键区域且 $\alpha = 0.05$ 时行人从起点到终点的位置轨迹图，可见加入关键区域后，行人在经过障碍物时的轨迹路线明显比基线实验中的轨迹偏离障碍物。

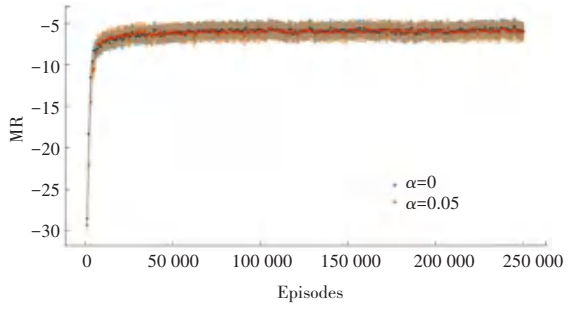


（a）基线实验　　（b）含关键区域实验

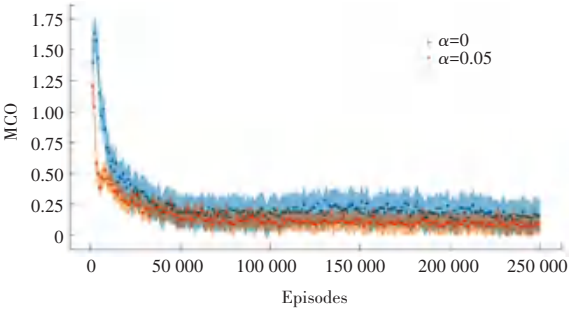
图 6　有无关键区域路径规划图对比

Fig. 6　Comparison of path planning diagrams with and without key regions

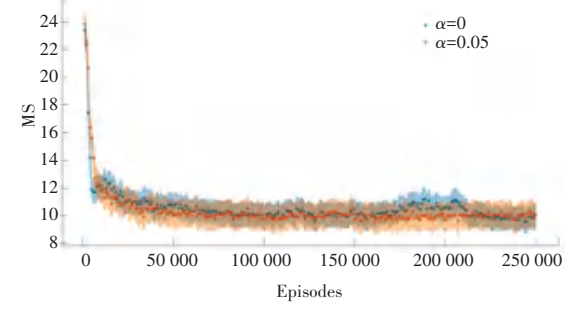
有无碰撞缓冲区域下的 MR、MCO、MS 对比实验结果如图 7 所示。图 7(a)为本次实验中所得到的平均奖励 MR 图，蓝色曲线表示基线实验的 MR 图像，橙色曲线表示关键区域 $\alpha = 0.05$ 时的 MR 收敛图像，可以看出随着训练次数的增加，奖励值逐渐趋于收敛，且在收敛时，带有关键区域的收敛速度和 MR 是优于基线实验的；图 7(b)为行人与障碍物之间平均碰撞次数 MCO 的对比图，蓝色曲线表示基线实验的平均碰撞次数，橙色曲线表示碰撞缓冲区域 $\alpha = 0.05$ 时的平均碰撞次数，可以明显看出相对于基线实验，收敛时加入碰撞缓冲区域的实验其路径规划过程中的碰撞次数平均减少了 61.9%，明显优于基线实验的 MCO；图 7(c)为两个实验的平均步长 MS 随着迭代次数的变化曲线图，可以看出加入碰撞缓冲区域的实验 MS 的变化明显比基线实验稳定。



(a) 有无碰撞缓冲区域的平均奖励 MR



(b) 有无碰撞缓冲区域的平均碰撞次数 MCO



(c) 有无碰撞缓冲区域的平均步长 MS

图 7 有无碰撞缓冲区域下的 MR、MCO、MS 对比实验结果

Fig. 7 Experimental comparison results of MR, MCO, and MS with and without a collision buffer zone

4 结束语

在多智能体路径规划问题中,传统的智能体路径规划算法过度依赖环境信息,在一些复杂多变的环境中,模型的收敛效果容易受到影响。本文提出利用强化学习算法来解决动态环境中的多智能体路径规划问题,并对算法中的奖励函数进行改进,提出 MADDPG-R 算法。实验证明,该算法在优化路径长度和降低碰撞率方面均表现出了良好的性能。未来将进一步对算法进行优化,尽可能减少训练时间,使其可以适应更加复杂多变的环境。

参考文献

[1] 国家统计局. 中华人民共和国 2023 年国民经济和社会发展统计公报[N]. 人民日报,2024-03-01. DOI:10.28655.n.cnki.nrmrb.

[2] 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.

[3] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2 (4): 314-326.

[4] PRUDENCIO R F, MAXIMO M R O A, COLOMBINI E L. A survey on offline reinforcement learning: Taxonomy, review, and open problems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(8): 10237-10257.

[5] NGUYEN T T, NGUYEN N D, NAHAVANDI S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications[J]. IEEE Transactions on Cybernetics, 2020, 50 (9): 3826-3839.

[6] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3/4): 279-292.

[7] KUMAR A, ZHOU A, TUCKER G, et al. Conservative q-learning for offline reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 1179-1191.

[8] CARTA S, FERREIRA A, PODDA A S, et al. Multi-DQN: An ensemble of deep Q-learning agents for stock market forecasting [J]. Expert Systems with Applications, 2021, 164: 113820.

[9] 邓修朋, 崔建明, 李敏, 等. 深度强化学习在机器人路径规划中的应用[J]. 电子测量技术, 2023, 46(6): 1-8.

[10] 史殿习, 彭滢璇, 杨焕焕,等. 基于 DQN 的多智能体深度强化学习运动规划方法[J]. 计算机科学, 2024, 51 (2): 268-277.

[11] 司鹏搏, 吴兵, 杨睿哲,等. 基于多智能体深度强化学习的无人机路径规划[J]. 北京工业大学学报, 2023, 49 (4): 449-458.

[12] GUO S, ZHANG X, ZHENG Y, et al. An autonomous path planning model for unmanned ships based on deep reinforcement learning[J]. Sensors, 2020, 20(2): 426.

[13] WANG B, LIU Z, LI Q, et al. Mobile robot path planning in dynamic environments through globally guided reinforcement learning [J]. IEEE Robotics and Automation Letters, 2020, 5(4):6932-6939.

[14] CHEN P, PEI J, LU W, et al. A deep reinforcement learning based method for real-time path planning and dynamic obstacle avoidance[J]. Neurocomputing, 2022, 497: 64-75.

[15] 陈人龙, 陈嘉礼, 李善琦,等. 多智能体强化学习方法综述[J]. 信息对抗技术, 2024, 3 (1): 18-32.

[16] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in Neural Information Processing Systems, 2017, 30: 6382-6393. DOI: 10.48550/arXiv,1706.02275.

[17] QIE H, SHI D, SHEN T, et al. Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning[J]. IEEE Access, 2019, 7: 146264-146272. DOI: 10.1109/ACCESS.2019.2943253.

[18] 劳天成, 刘义, 范文慧. 多智能体深度确定性策略梯度算法研究与改进[J]. 新疆大学学报(自然科学版), 2023, 40(6): 717-723.

[19] HALL E T. The Hidden Dimension: Man's Use of Space in Public and Private [M]. London: The Bodley Head, 1969: 121.