DOI: 10.13382/j.jemi.B2205932

# 基于 SPE-ICM 的移动机器人内在动机避障规划*

罗国攀 张国良 徐佳宝
(四川轻化工大学人工智能四川省重点实验室 宜宾 644000)

**摘 要**:针对动态环境下强化学习算法对移动障碍物的检测不理想,进而影响最优避障策略的问题。提出一种以状态预测误差为内在动机的奖励结构形式(state predict error - intrinsic curiosity module, SPE-ICM)来提高策略函数对 Agent 的环境探索能力。首先,引入内在奖励机制为 Agent 提供多重奖励(reward)结构;其次,依据内外的奖励结构优化提高 Agent 对环境信息的感知能力,改进对移动障碍物在数据结构上的采集检测方式,并且依赖新的检测方式对最优避障策略函数进行优化提升;最后,将该网络模型与深度确定性策略梯度算法(DDPG)结合,运用到以 ROS 搭建的路径规划仿真环境中进行对比实验,验证所提算法的可行性。结果表明,所提算法在检测能力、决策能力上效果明显更优。

**关键词**: 状态奖励误差;内在动机奖励;最优避障策略
**中图分类号**: TP242 **文献标识码**: A **国家标准学科分类代码**: 510.8050

# Obstacle avoidance strategy method for mobile robots based on SPE-ICM

Luo Guopan Zhang Guoliang Xu Jiabao
(Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644000, China)

**Abstract**: Aiming at the problem that the reinforcement learning algorithm is not ideal to detect moving obstacles in a dynamic environment, which affects the optimal obstacle avoidance strategy. A state predict error-intrinsic curiosity module (SPE-ICM) with state prediction error as intrinsic motivation is proposed to improve the ability of policy functions to explore the environment of Agents. First, the internal reward mechanism is introduced to provide multiple reward (reward) structure for Agent. Secondly, according to the internal and external reward structure optimization, the Agent's perception of environmental information is improved, the collection and detection method of moving obstacles on the data structure is improved, and the optimal obstacle avoidance strategy function is optimized and improved by relying on the new detection method. Finally, the network model is combined with the deep deterministic strategy gradient algorithm (DDPG), and the comparative experiment is carried out in the path planning simulation environment built by ROS to verify the feasibility of the proposed algorithm. The results show that the proposed algorithm has significantly better effects in detection ability and decision-making ability.

**Keywords**: status reward error; optimal obstacle avoidance strategy; intrinsic motivation rewards

## 0 引 言

面对疫情防控要求的提升,尽量减少面对面接触加强社区网格化管理,完善路径需求[1],其中对移动机器人的路径规划至关重要。路径规划算法旨在使移动机器人从初始位置到目标位置自主探索规划出一条平滑、无碰撞的路径轨迹[2]。其算法多种多样包括传统算法、智能仿生算法以及强化学习算法[3-4]等。例如传统算法中的 A* 算法[5]、智能仿生算法中的遗传算法[6]。

将强化学习算法与路径规划结合运用,规避了对复杂地图信息的构建,强调机器人本身的统筹规划能力,摆脱了传统算法依赖地图信息规避障碍物的不方便,尤其是拥有移动信息的物体[7-8]。但受其自身高运算成本的

收稿日期: 2022-10-21 Received Date: 2022-10-21
*基金项目:四川省应用基础研究项目(2019YJ00413)资助

局限性以及环境提供的即时奖励不足导致检测能力不全面,进而影响避障策略执行效果。

针对以上问题,有关学者提出相应改进措施。Minh 等[9]提出第 1 个深度强化学习模型,即深度 Q 网络(deep Q-network,DQN),该网络模型是将神经网络和 Q-learning 相结合,利用神经网络代替 Q 值表,解决 Q-learning 中的维数灾难运算问题。而为了解决环境多变检测能力不足,Pathak 等[10]提出内在奖励模块(intrinsic curiosity module,ICM)用以探索能力问题。谭庆等[11]为了解决在稀疏环境下的有效探索,提出用奖励预测误差为内在奖励模块优化修正状态预测误差的办法,提高 Agent 的探索能力。更有甚者,赵英等[12]结合新颖性和风险评估作为内在奖励,反馈给 Agent 学习更好的策略。而在路径规划中常常会需要移动机器人快速有效的检测环境信息,做到"即遇则避"的策略。

基于上述研究,提出一种基于状态预测误差为内在驱动奖励模块的移动机器人避障策略算法。首先,依据所获取的转移轨迹信息(transition)输入内在奖励模块,得到内在奖励值以及预测动作。其次,将内在奖励输入到 DDPG 算法中更新提升优化策略函数,预测动作作为辅助项循环迭代。最后,通过总的损失最小化函数求解最优网络参数。

## 1 涉及相关内容

### 1.1 基于马尔可夫决策过程的强化学习

在人工智能领域中,强化学习[13-14]是一类特定的机器学习问题。在一个强化学习系统中,决策者可以观察环境,根据观测做出行动,并且获得奖励,通过与环境的交互学习如何最大化回报奖励,其中决策的好坏是关键。

强化学习算法旨在利用最大化回报奖励来更新优化策略,完成指定任务。这种交互过程被视为一个马尔可夫决策过程[15-16](Markov decision process, MDP),通常表示为一个五元组 $(S,A,P,R,\gamma)$。具体如下:

1) $S=\{s_1,s_2,\cdots,s_t\}$ 是环境状态空间,Agent 所在环境的集合;

2) $A=\{a_1,a_2,\cdots,a_t\}$ 是动作集合,考虑连续性的动作空间,一个动作往往包含多个参数。

3) $R$ 是奖励值集合,旨在 Agent 采取相应的动作后所获得的即时奖励回报;

4) $\gamma$ 为折扣因子,$\gamma\in(0,1)$,一般取接近 1 的值,如 0.9。

在基于上述 1)~4)的基础上引入概率和 Markov 性,就可得到完整的 MDP 模型。

5) $P$ 为状态转移概率集合。定义在时间 $t$,从状态 $S_t=s$ 依据动作 $A_t=a$ 转移到 $S_{t+1}=s'$,所获得奖励的概率为:

$$\Pr[S_{t+1}=s',R_{t+1}=r\mid S_t=s,A_t=a] \tag{1}$$

所用具体模型的定义如图 1 所示。

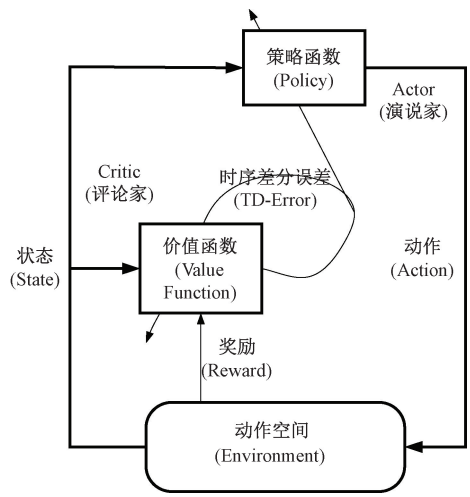



图 1 马尔可夫决策过程模型示意图

Fig. 1 MDP model graph

### 1.2 DDPG 算法

深度确定性策略梯度[17](deep deterministic policy gradient, DDPG)引入 Actor-Critic 框架,结合两套神经网络模型分别对策略网络模型 $Q(s,a;\theta^Q)$ 和价值网络模型 $\mu(s;\theta^\mu)$ 中状态动作价值函数和策略函数进行拟合,其中 $\theta^Q,\theta^\mu$ 为神经网络参数。在更新参数阶段,DDPG 延续了 DQN 算法中的经验复用,以状态序列 $(s_t,a_t,r_t,s_{t+1})$ 对网络进行策略训练,以最小化样本之间的相关性,同时采用 Q 目标网络在时间差异备份期间提供一致的目标,同时增加了批归一化防止梯度爆炸[18]。

首先,从经验复用池 D 中抽取若干经验,先通过目标网络得到目标回报值 $y_t$:

$$y_t=r_t+\gamma Q'(s_{t+1},\mu'(s_{t+1}\mid\theta^\mu)\mid\theta^{Q'}) \tag{2}$$

之后,更新在线价值网络,将 $s_t,a_t$ 共同输入得到实际值 $Q(s_t,a_t\mid\theta^Q)$,Critic 网络根据式(3)最小化损失函数 $L(\theta^Q)$ 更新参数:

$$L(\theta^Q)=\frac{1}{N}\sum_t(y_t-Q(s_t,a_t\mid\theta^Q))^2 \tag{3}$$

Actor 网络根据策略梯度更新策略参数:

$$\nabla_{\theta^\mu}J\approx\frac{1}{N}\sum_t\{\nabla_a Q(s,a\mid\theta^Q)\mid_{s=s_t,a=\mu(s_t)}\nabla_{\theta^\mu}\mu(s\mid\theta^\mu)\}\mid_{s_t} \tag{4}$$

## 2 基于好奇心内在奖励的避障策略算法

好奇心内在动机的运用起初是为了增加 Agent 的探

索能力,解决稀疏奖励问题所提出. 本文基于好奇心内在奖励的特性以及运用场景,将该模块作用于移动机器人探索躲避动态障碍物的功能模块使用,提高规划时的避障能力。

### 2.1　状态预测误差的内在奖励结构

状态预测误差法是以若干步长的状态预测误差值为内在奖励。该模块包含两个神经网络模型:前向模型(forward model,FM)和逆向模型(inverse model,IM),具体如图 3 所示。

首先,将 $s_t$ 进行激光数据处理得到 $f(s_t)$,将 $a_t$,$f(s_t)$ 作为前向网络模型输入,拟合预测新的状态 $\hat{s}_{t+1}$,并对 $\hat{s}_{t+1}$ 进行激光数据处理得到输出 $f(\hat{s}_{t+1})$,其中数据处理为求激光测距的几何距离。因为在二维平面上,所以激光采集的点云数据为二维。定义如式(5)所示:

$$f(S_t)=\begin{bmatrix}\sqrt{x_1^{\,2}+y_1^{\,2}}\\ \vdots\\ \sqrt{x_{10}^{\,2}+y_{10}^{\,2}}\end{bmatrix}_{10\times1} \tag{5}$$

$$\hat{S}_{t+1}=F(f(S_t),a_t;\theta^F)$$

其中,$f(S_t)$ 为当前时刻经过数据处理后得到的状态特征,$\theta^F$ 为前向网络参数。

神经网络[19]依靠其强大的计算拟合能力帮助解决数据复杂等问题。在利用好奇心奖励模块时用到了 3 个神经网络,在此对其设计原理进行说明。图 2 是神经网络以及输入输出数据的具体结构。

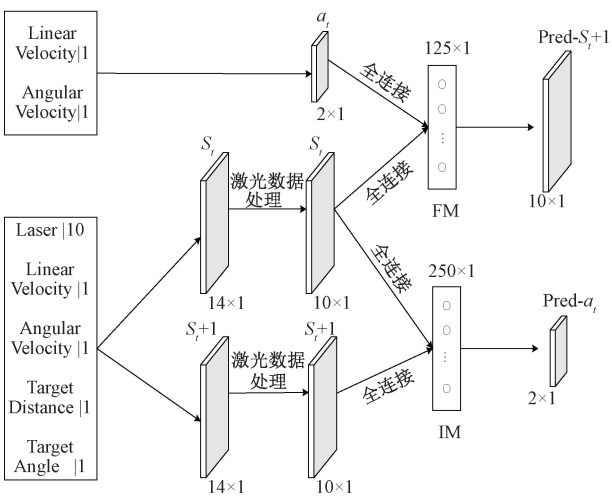



图 2　神经网络结构及输入输出数据示意图

Fig. 2　Schematic diagram of neural network structure and input and output data

在 DDPG 算法中采用的数据输入包括目标点的位置以及 Agent 的速度信息,在检测环境信息的误差时只需要移动障碍物的位置信息,所以在计算内在奖励时将状态 $s_t$,$s_{t+1}$ 进行数据处理只保留激光检测到的距离信息 10 个,之后的一系列操作不变更。

前向网络根据凹函数的定义希望预测状态和实际的状态尽可能一致,通过迭代更新优化网络参数 $\theta^F$,使前向最小化损失函数趋近于 0。

$$\min L(\theta^F)=\lim L_{Forward}(f(S_t),f(\hat{S}_{t+1}))=0 \tag{6}$$

$$L_{Forward}(f(S_t),f(\hat{S}_{t+1}))=\frac{1}{2}\|f(\hat{S}_{t+1})-f(S_{t+1})\|_2^2 \tag{7}$$

依据式(7)得到内在奖励的定义如下:

$$r_t^{in}(s_t,a_t)=\frac{\eta}{2}\|f(\hat{S}_{t+1})-f(S_{t+1})\|_2^2 \tag{8}$$

其中,$\eta$ 为大于 0 的比例因子。

同理,逆向网络拟合预测动作 $\hat{a}_t$ 使其与实际的动作 $a_t$ 的差值作为最小化损失函数,依据凹函数原理,尽可能优化修正动作决策误差使其趋近于 0,由于 $\hat{a}_t$ 可以通过网络参数更新迭代,从而进一步可以提升策略函数。

$$\hat{a}_t=g(s_t,s_{t+1};\theta^I) \tag{9}$$

$$L(\theta^I)=\frac{1}{2}\|\hat{a}_t-a_t\|_2^2 \tag{10}$$

$$\min L(\hat{a}_t,a_t) \tag{11}$$

### 2.2　改进环境感知能力

无地图探索一直是基于强化学习路径规划算法的特色,而对环境的感知以及信息采集则主要是通过外部环境提供奖励。强化学习算法依据即时奖励做出相应的动作决策,其中依据的奖励是由环境提供的,称之为外在奖励。但这种奖励往往是不定的,会随着环境中动态的物体而改变,所以即使当前获取的奖励存储记录之后,下一次面对相同的场景,由于移动障碍物是不定的,环境所提供的即时奖励也是不定的,这样就会导致训练策略不佳,对移动障碍物的检测不精确。在此基础上添加基于好奇心模块的内在奖励,该模块是由 Agent 自发产生的奖励,不接受由于环境的可变性因素,致使外在奖励给定数据不准确,导致环境信息感知不充分的问题,进而达到规避环境的不确定性,提高检测能力。另外,检测能力的最优将取决于内在奖励的消失而定,当预测误差趋近于 0 时,内在奖励消失,说明当前策略或者检测效果最优。

### 2.3　设计最优化内外奖励结构的避障策略

在 2.1 节中根据设计的好奇心网络模块的输出得到了内在奖励 $r_t^{in}$ 和动作预测值 $\hat{a}_t$。基于上述结果,可以设计最优化内外奖励结构的避障策略。关于避障的具体方法在于对最优策略函数的迭代更新,通过不断的试错训练,迭代出最优或者次优的最优策略是完成动态避障的关键。采用 2.1 节中提到的两个最小化循环修正方法不断地完善最优策略函数。

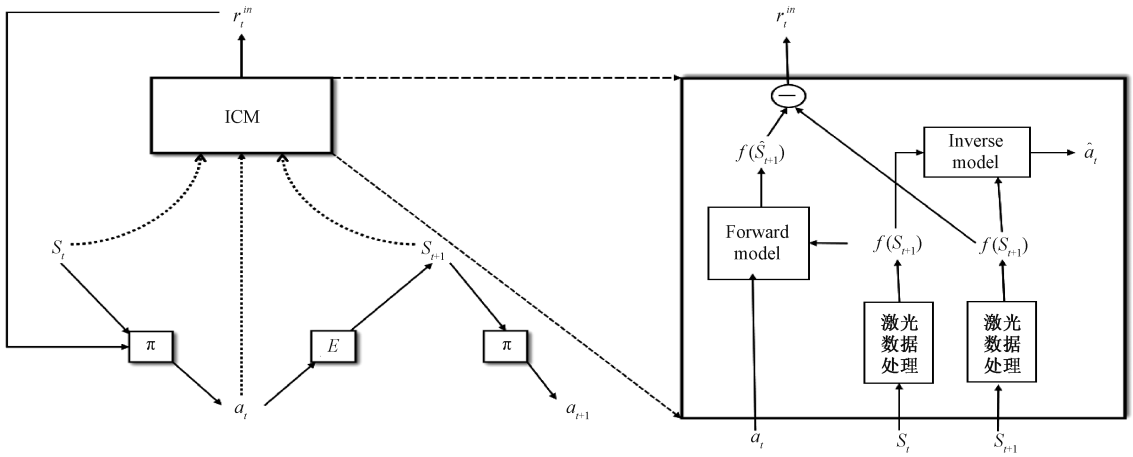



图 3 ICM 结构示意图

Fig. 3 Schematic diagram of the ICM structure

首先,最初的奖励结构只存在外在环境所提供,添加内在奖励后,奖励结构如下式:

$$R = r_{\eta}^{in}(s_t, a_t) + r^{ex} \tag{12}$$

其次,对未来时刻的回报奖励定义为:

$$U_t = R_t^{in+ex} + \sum_{i-t}^{\infty} \gamma^{i-t}(r_i^{ex} + r_{\eta}^{in}(s_t, a_t)) \tag{13}$$

进而依据 DDPG 算法的 Critic 网络评估动作价值函数 $Q$,最后通过 TD 误差反作用于 Actor 网络的策略函数做出修正优化后的动作,具体如下:

$$Q(s_t, a_t; \theta^Q) = E[U_t \mid S = s_t, A = a_t; \theta^Q] \tag{14}$$

$$\pi(a \mid s) = a^* = \underset{\pi^*}{\text{argmax}} Q(s, a; \theta^Q) \tag{15}$$

所以在此基础上通过内在奖励最小化损失函数不断通过梯度更新减小状态误差,内在奖励值会逐渐减小,当误差趋近于 0 时,说明检测能力趋于最优。其次,通过 ICM 逆向网络得到的预测动作也可以优化动作策略函数,具体为将式(10)和(11)改写为如下:

$$L(\theta^I) = \frac{1}{2} \| \hat{a}_t - a^* \|_2 \tag{16}$$

$$\min L(\hat{a}_t, a^*) \tag{17}$$

当预测动作与最优动作误差很小或者趋近于 0 时,代表当前策略函数为最优,加快网络参数收敛。

最后,将 Agent 所创建的循环迭代过程问题通过式(6)、(14)、(15)和(16)组合,可以写成:

$$\min_{\theta^Q, \theta^{\mu}, \theta^F, \theta^I} \begin{bmatrix} -\lambda E_{a=\pi^*(s;\theta^{\mu})}[U_t \mid s_t, a_t; \theta^Q] \\ +(1-\sigma) L(\theta^I) + \sigma L(\theta^F) \end{bmatrix} \tag{18}$$

其中,$0 \leqslant \sigma \leqslant 1$ 是一个标量,时衡量前向和逆向模型损失的标量,而 $\lambda > 0$ 是梯度损失和内在奖励信号的衡量标准,添加负号是为了求期望的最大值取反而来。通过 4 个神经网络参数相互迭代更新最终达到收敛时,训练出最优策略。

### 2.4 算法执行流程

基于整个构建思路,SPE-ICM-DDPG 算法的具体流程如算法 1 所示。

算法 1 SPE-ICM-DDPG Path Planning

1) 随机初始化 $Q(s, a \mid \theta^Q)$、$\mu(s \mid \theta^{\mu})$ 以及好奇心网络中的 $F(f(S_t), a_t; \theta^F)$、$\hat{a}_t = g(s_t, s_{t+1}; \theta^I)$ 的参数 $\theta^Q, \theta^{\mu}, \theta^F, \theta^I$
2) 初始化目标网络参数 $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^{\mu}$
3) 初始化经验回放池 $RM$
4) For episode = 1, M do
5) 初始化随机噪声动作探索过程 $N$
6) 获取初始观测状态 $s_1$
7) For t = 1, T do
8) 根据当前策略和探索噪声选择动作 $a_t = \mu(s_t \mid \theta^{\mu}) + N_t$
9) 执行 $a_t$ 并且观测奖励 $r_t$ 和下一时刻状态 $s_{t+1}$
10) 存储状态转移序列 $(s_t, a_t, r_t, s_{t+1})$ 到 $RM$
11) 随机从 $RM$ 中采样一组 minibatch,记为 $D$
12) For $D$ 中的每个状态转移序列 $(s_t, a_t, r_t, s_{t+1})$
13) 根据式(8)和(9)计算内在奖励 $r_t^{in}, \hat{a}_t$
14) 通过式(12)计算总的奖励 $R$
15) End for
16) 在样本 $D$ 上通过式(18)最小化损失函数更新前向和逆向网络
17) 通过式(3)更新价值网络
18) 通过式(4)更新策略网络
19) 更新目标网络
20) End for
21) End for

## 3 仿真实验与结果验证

### 3.1 实验平台介绍和设备说明

为了验证所提算法在动态环境中的规划避障效果,基于 ROS gazebo 设计了一个动态连续规划任务作为实

验环境。采用的设备为 Ubuntu 版本为 18.04,ROS 版本为 Melodic。实验测试采用的环境 RTX 2060,Python3.7,TensorFlow 1.5.0 电脑的主频为 3.59 GHz,运行内存为 16 GB。图 4 为实验环境。其中黑色物体为移动机器人,白色圆柱体为移动障碍物。在位姿保持不变的情况下,白色圆柱体环绕移动机器人所在初始位置进行速度大小不变的圆周运动。

相关网络以及强化学习中的超参数设置,具体如表 1 所示。

**表 1　实验参数表**

**Table 1　Experimental parameter table**

| 参数名称 | 度量值 | 参数名称 | 度量值 |
|---|---|---|---|
| 折扣因子 $\gamma$ | 0.9 | 前向学习率 | $10^{-2}$ |
| Critic 学习率 $\alpha$ | $10^{-2}$ | 逆向学习率 | $10^{-2}$ |
| Actor 学习率 $\beta$ | $10^{-2}$ | $\theta^F$,$\theta^I$ | 均匀分布 |
| $\theta^Q$,$\theta^\mu$ | 均匀分布 | 训练次数 | 2 000 |

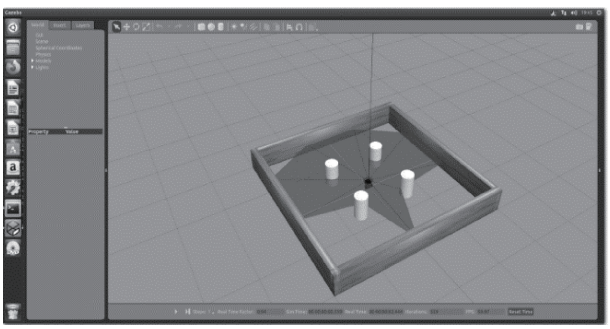


图 4　仿真环境

Fig. 4　Simulation environment

### 3.2　实验结果分析

为了验证所提算法在检测能力和决策策略函数效果更优,进行了如下实验数据对比分析。

首先,移动障碍物的检测能力主要体现在在既定路线的前提条件下及时规避即将到来的障碍物以及对障碍物的运行轨迹做出及时判断,具体到而言就是在于对奖励获取和利用的能力。表 2 通过对比与移动障碍物碰撞次数以及图 5、6 对状态动作价值函数 $Q(s_t,a_t;\theta^Q)$ 的预测值和奖励利用情况进行分析。

**表 2　检测能力对比分析表**

**Table 2　Comparative analysis table of detection capabilities**

| | 训练次数 | 单次总步长/步 | 动态碰撞/次 |
|---|---|---|---|
| DDPG | 2 000 | 600 | 637 |
| ICM-DDPG | 2 000 | 600 | 454 |

从表 2 分析,加入内在奖励之后与动态障碍物的碰撞次数减小了 9.1%,成功避障次数提高了 9.2%,该数据说明避障策略函数明显提升,侧面说明对奖励的利用更加充分,奖励结构的优化带动了 Agent 对环境的感知,进一步说明对动态移动物体的检测能力得到提升。为了更加明显的对比分析性能的提升,分析价值函数的变化也能看出所提算法的有效性,具体如图 5 所示。

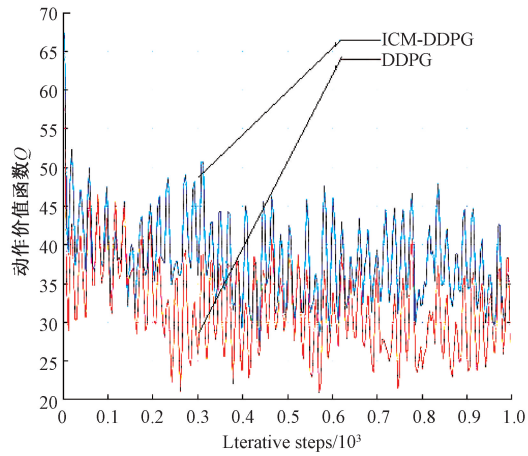



图 5　状态动作价值函数值对比

Fig. 5　State action value function value comparison plot

状态动作价值函数的定义是依据获取奖励的情况对当前状态下策略函数所要执行的动作 Action 进行计算评估,分数的高低说明执行该动作的好坏,决定是否要按照当前形势走向继续发展。经过 1 000 次的迭代实验过程,添加了好奇心内在奖励的动作价值函数值更高,证明对形势走向的判断力相比传统 DDPG 算法更精确。分析随着迭代次数的增加,局部会出现重叠现象,是因为随着式(18)不断完善,预测准确度提高,内在好奇心的奖励是依据状态误差,而误差越来越小,所以会出现相同的部分,当训练效果达到一定的收敛程度时,此时的策略函数接近最优或者次优。

奖励利用的情况如图 6 所示,通过随机采样收集了 200 份实验奖励数据进一步分析。对比说明 Agent 在奖励获取能力方面较原本算法更加容易,更有利于训练出最优或者次优的决策函数以完成避障规划。

其次,将训练结束后的策略函数 $\mu(s;\theta^\mu)$ 用于指导移动机器人多任务路径规划,观察单次规划成功次数以及单次终止步数对比验证算法的决策能力。在此,实验设置了 6 个目标位置点以供移动机器人完成多规划任务需求,并且按照依次顺序经过每个目标点,如图 7 所示。

将完成神经网络训练后的策略函数重新进行规划任务,总计 100 次对比了 DDPG 与 ICM-DDPG 的性能。具体实验过程为在指定的单次总步长内,完成从 1~6 目标点的过程为一次循环,当完成一次后未达到最大步长时重新刷新再次进行,如表 3 所示。

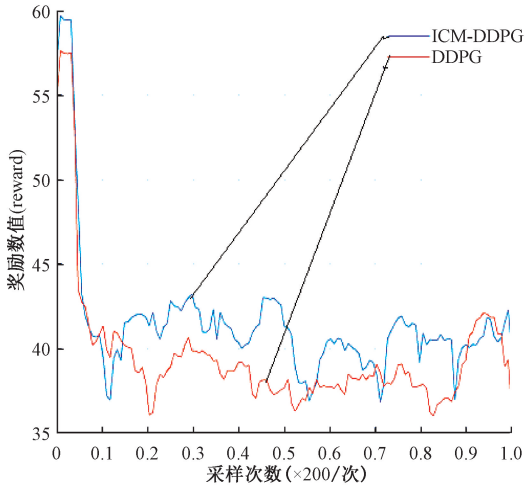



图 6 奖励利用对比图

Fig. 6 Reward utilization comparison chart

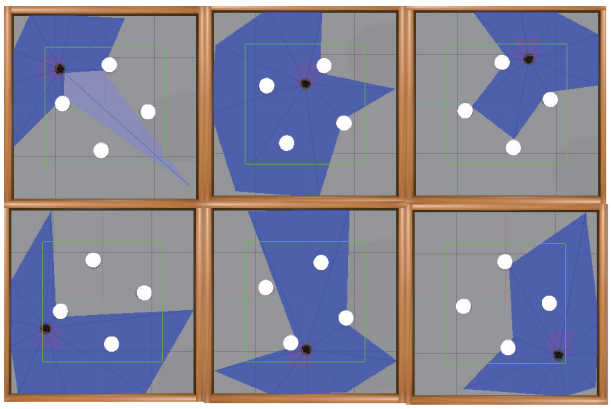


图 7 实验效果

Fig. 7 Experimental renderings

**表 3 决策能力对比分析表**

**Table 3 Comparative analysis table of decision-making capabilities**

| 规划次数 100 次 | 完成任务平均次数 | 完成一次平均步数/步 | 完成一次任务平均时间/s |
|---|---|---|---|
| DDPG | 3.8 | 269.8 | 124.2 |
| ICM-DDPG | 13.6 | 129.8 | 110.7 |

从表 3 分析,完成多次规划任务成功率提高了大约 4 倍,成功次数提升,说明避障策略函数更完善,间接验证了算法在决策执行能力上的提升;完成一次规划任务所用步数减少了 48.1%,充分说明了算法在收敛速度以及收敛性上有所提升。所用时间虽然减少较少,但相比规划效率提升明显。

## 4 结 论

为解决动态环境下强化学习算法对移动障碍物的检测不理想,进而影响最优避障策略的问题,本文将好奇心探索运用于移动机器人路径规划的避障策略,提高移动机器人在局部复杂动态环境中的探索能力。其具体的内容:1)设计内驱动模型解决单一奖励结构问题;2)通过构建的内外奖励结构优化奖励函数,进而改进最优避障策略函数;智能规划算法致力于让机器人自主的判断、规划、执行最终完成指定任务,而强化学习本身思路来源于心里学范畴,反映出“人类”纯粹行为,所以在算法的基础上加入自发探索和自给自足的奖励结构更能有效地规避外在因素带来的不确定性,提高算法控制效果的鲁棒性以及机器人自身的适应性。其次,原本的强化学习运用范围受自身算法的瓶颈局限很难扩展到连续控制问题上,引入神经网络结构并对其数据输入以及控制其输出结构,能有效地得到可靠且实用的数据样本,丰富了算法的应用范围。下一步研究将展开范围更广的全局规划。

## 参考文献

[1] 谢丹妮,张智文. 疫情防控常态化下,社区网格化管理的完善路径—以福建省漳州市为例[J]. 延边党校学报,2022,38(4):61-66.
XIE D N, ZHANG ZH W. The perfect path of community grid management under the normalization of epidemic prevention and control: A case study of Zhangzhou city, Fujian province[J]. Journal of Yanbian Party School, 2022,38(4):61-66.

[2] 霍凤财,迟金,黄梓健. 移动机器人路径规划算法综述[J]. 吉林大学学报(信息科学版),2018,36(6):639-647.
HUO F C,CHI J,HUANG Z J. Review of path planning for mobile robots [J]. Journal of Jinlin University (Information Science Edition),2018,36(6):639-647.

[3] SÁNCHEZ-IBÁÑEZ J R, PÉREZ-DEL-PULGAR C J, GARCÍA-CEREZO A. Path planning for autonomous mobile robots: A review [J]. Sensors, 2021, 21(23): 7898.

[4] ZAFAR M N, MOHANTA J C. Methodology for path planning and optimization of mobile robots: A review[J]. Procedia Computer Science, 2018, 133: 141-152.

[5] 李艳生,万勇,张毅,等. 基于人工蜂群-自适应遗传算法的仓储机器人路径规划[J]. 仪器仪表学报,2022,43(4):282-290.
LI Y SH, WAN Y, ZHANG Y, et al. Path planning for warehouse robot based on the artificial bee colony-adaptive genetic algorithm [J]. Chinese Journal of Scientific Instrument, 2022, 43(4): 282-290.

[6] 姜媛媛,张阳阳. 改进 8 邻域节点搜索策略 A*算法的路径规划[J]. 电子测量与仪器学报,2022,36(5):234-241.
JIANG Y Y, ZHANG Y Y. Improved path planning of A* algorithm of domain node search strategy 8[J]. Journal of Electronic Measurement and Instrumentation, 2022,36(5):234-241.

[7] QIANG L, NANXUN D, HUICAN L, et al. A model-free mapless navigation method for mobile robot using reinforcement learning[C]. 2018 Chinese Control and Decision Conference (CCDC). IEEE, 2018: 3410-3415.

[8] DOBREVSKI M, SKOCAJ D. Map-less goal-driven navigation based on reinforcement learning[C]. 23rd Computer Vision Winter Workshop, 2018.

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

[10] PATHAK D, AGRAWAL P, EFROS A, et al. Curiosity-driven exploration by self-supervised prediction[C]. International Conference on Machine Learning. Cambridge: JMLR, 2017: 2778- 2787.

[11] 谭庆,李辉,吴昊霖,等. 基于奖励预测误差的内在好奇心方法[J]. 计算机应用,2022,42(6):1822-1828.
TAN Q, LI H, WU H L, et al. Intrinsic curiosity method based on reward prediction error[J]. Journal of Computer Applications, 2022, 42(6): 1822-1828.

[12] 赵英,秦进,袁琳琳. 结合新颖性和风险评估的内在奖励方法[J]. 计算机工程与应用,2023,59(5):148-154.
ZHAO Y, QIN J, YUAN L L. Intrinsic reward method combining novelty and risk assessment[J]. Computer Engineering and Applications, 2023, 59(5): 148-154.

[13] NIU H, JI Z, ARVIN F, et al. Accelerated sim-to-real deep reinforcement learning: Learning collision avoidance from human player[C]. 2021 IEEE/SICE International Symposium on System Integration (SII). IEEE, 2021: 144-149.

[14] 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.
YAN J J, ZHANG Q SH, HU X P. Review of path planning techniques based on reinforcement learning[J]. Computer Engineering, 2021, 47(10): 16-25.

[15] 肖智清. 强化学习原理与 Python 实现[M]. 北京:机械工业出版社,2019.
XIAO ZH Q. Reinforcement Learning Principles and Python Implementation[M]. Beijing: China Machine Press, 2019.

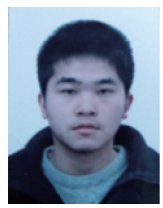
[16] 桂林,武小悦. 部分可观测马尔可夫决策过程算法综述[J]. 系统工程与电子技术,2008,345(6):1058-1064.
GUI L, WU X Y. Survey of algorithms for partially observable Markov decision processes[J]. Systems Engineering and Electronics, 2008, 345(6): 1058-1064.

[17] 柯丰恺,周唯倜,赵大兴. 优化深度确定性策略梯度算法[J]. 计算机工程与应用,2019,55(7):151-156,233.
KE F K, ZHOU W T, ZHAO D X. Optimized deep deterministic policy gradient algorithm[J]. Computer Engineering and Applications, 2019, 55(7): 151-156,233.

[18] 刘驰,王占健,戴子彭,等. 深度强化学习学术前沿与实战应用[M]. 北京:机械工业出版社,2020.
LIU CH, WANG ZH J, DAI Z P, et al. Academic Frontiers and Practical Applications of Deep Reinforcement Learning[M]. Beijing: China Machine Press, 2020.

[19] JESUS J C, BOTTEGA J A, CUADROS M A S L, et al. Deep deterministic policy gradient for navigation of mobile robots in simulated environments[C]. 2019 19th International Conference on Advanced Robotics (ICAR). IEEE, 2019: 362-367.

## 作者简介



**罗国攀**,就读于四川轻化工大学控制理论与控制工程专业,硕士研究生,主要研究方向为移动机器人路径规划。
E-mail: 1210977404@qq.com

**Luo Guopan** is now a M. Sc. candidate of control theory and control engineering at Sichuan University of Science and Engineering. His main research interest includes mobile robot path planning.

**徐佳宝**,就读于四川轻化工大学控制理论与控制工程专业,硕士研究生,主要研究方向为机器人控制技术。
E-mail:1248217710@qq.com

**Xu Jiabao** is now a M. Sc. candidate of control theory and control engineering at Sichuan University of Science and Engineering. His main research interest includes robot control technology.

**张国良**(通信作者),博士后,教授,硕士生导师,任教于四川轻化工大学,主要研究方向为先进控制理论、组合导航、机器人技术。
E-mail:zhgl@sohu.com

**Zhang Guoliang** (Corresponding author) is now a postdoctoral fellow, professor and master's supervisor at Sichuan University of Science and Engineering. His main research interests include advanced control theory, combined navigation and robotics