基于强化学习的多 Agent 路径规划方法研究

王毅然 经小川 田 涛 孙运乾 从帅军
(中国航天系统科学与工程研究院 北京 100048)

摘 要 以复杂任务下多个智能体路径规划问题为研究对象,提出一种基于强化学习的多 Agent 路径规划方法。该方法采用无模型的在线 Q 学习算法,多个 Agent 不断重复"探索-学习-利用"过程,积累历史经验评估动作策略并优化决策,完成未知环境下的多 Agent 的路径规划任务。仿真结果表明,与基于强化学习的单 Agent 路径规划方法相比,该方法在多 Agent 避免了相碰并成功躲避障碍物的前提下,减少了 17.4% 的总探索步数,形成了到达目标点的最短路径。

关键词 多智能体 强化学习 路径规划 Q 学习算法 未知环境

中图分类号 TP301.6 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.08.029

MULTI-AGENT PATH PLANNING BASED ON REINFORCEMENT LEARNING

Wang Yiran Jing Xiaochuan Tian Tao Sun Yunqian Cong Shuaijun
(*China Academy of Aerospace System Science and Engineering*, *Beijing 100048*, *China*)

Abstract Taking multiple agents path planning problems under complex tasks as the research object, we proposed a multi-agent path planning method based on reinforcement learning. The method adopted a model-free online Q learning algorithm. In this method, a model-free online Q-learning algorithm was adopted. Many agents repeated the process of "exploration-learning-utilization", accumulated historical experience, evaluated action strategies and optimized decision-making, and completed the task of multi-agent path planning in unknown environment. The simulation results show that compared with the single agent path planning method based on reinforcement learning, this method reduces the total exploration steps by 17.4% and forms the shortest path to the target point on the premise that multi-agent avoids collision and successfully avoids obstacles.

Keywords Multi-agent Reinforcement learning Path planning Q learing algorithm Unknown environment

0 引 言

随着科学技术的不断发展,路径规划技术的研究成果已经广泛应用人类生产和生活的各个方面。如在地震救灾中,无人机能够自主躲避障碍物,规划一组较优的路径到达指定灾区,完成灾情获取任务;在军事领域中,无人机和机器人在完成情报侦察以及作战打击任务过程中,要躲避敌方威胁和避免相撞,规划一条较优路径完成任务[1-4]。随着工作任务变得越来越复杂,往往需要多个智能体协同完成任务,每个智能体均是环境中的一部分,个体采取行动均会造成环境的改变,此时在动态环境中,单个智能体和其他智能体之间的协调与避障是多个智能体路径规划亟需解决的问题。路径规划的目标是寻找一条从给定的起始点到终止点的较优的运动路径。单智能体的路径规划在一个环境中的状态是有限的,目前解决的方法主要有 Dijkstra 算法[5]、粒子群算法[6]、A * 算法[7]、遗传算法、模拟退火算法、蚁群算法[8]等。多智能体系统与单个智能体相比,往往能够完成复杂艰巨任务,且通常能够付

收稿日期:2019-01-10。广东省科技厅应用型研发基金专项(2016B010127005)。**王毅然**,硕士生,主研领域:人工智能,计算机应用。**经小川**,研究员。**田涛**,博士。**孙运乾**,工程师。**从帅军**,工程师。

出更小的代价收获更大的整体效益,因此多个智能体的路径规划研究具有十分重要的意义。

多智能体系统是由具有一定自主性、能够在共同目标窗口内协作、竞争和通信的协作智能 Agent 组成的[9]。单个 Agent 解决问题的能力是有限的,复杂任务需要多个 Agent 协同合作,共同完成整体或局部目标。如果在同一环境中存在多个 Agent 同时移动,对其进行路径规划将会变得十分困难。目前解决多智能体的路径规划问题取得了一些进展,文献[10]提出了免疫协同进化算法并仿真实现静态障碍物环境中多个机器人避障、避碰的最短路径;文献[11]提出了一种主从结构的并行多水下机器人协同路径规划算法,子层结构应用粒子群并行算法,生成各个机器人当前的最优路径,同时主层结构应用微分进化算法实时给出当前考虑机器人与障碍物、机器人与机器人之间避碰情况下,总系统运行时间最短的路径组合方案;文献[12]提出了一种基于分层强化学习及人工势场的多 Agent 路径规划算法,首先将多 Agent 的运行环境虚拟为一个人工势能场,根据先验知识确定每点的势能值,它代表最优策略可获得的最大回报,其次利用分层强化学习方法的无环境模型学习进行策略更新;文献[13]提出了首先利用 A-Star 算法启发式地得到多个智能体到达目标点的临时最短路径,同时计算访问节点的时间,通过动态地对时间窗进行精确计算和加锁来重置路线以避免冲突。

为解决未知环境下多个 Agent 路径规划问题,上述算法随着 Agent 的数量以及环境规模变大时,算法的效率会变得很低。本文提出了一种基于强化学习的多 Agent 路径规划方法(Multi-agent path planning based on reinforcement learning,MAPP-RL),该方法中的多个 Agent 不断地与环境交互,当采取一个动作后,Agent 会从环境中得到一个反馈,用来评估该动作的好坏,然后把评估结果作为历史经验,不断地进行优化决策,最后找到一个可以得到最大奖励的动作序列,完成复杂未知环境下的多 Agent 路径规划任务。

1 整体框架

多智能体的路径规划整体框架主要包括四个层次:环境建模层、算法层、任务分配层、多 Agent 系统层,如图 1 所示。

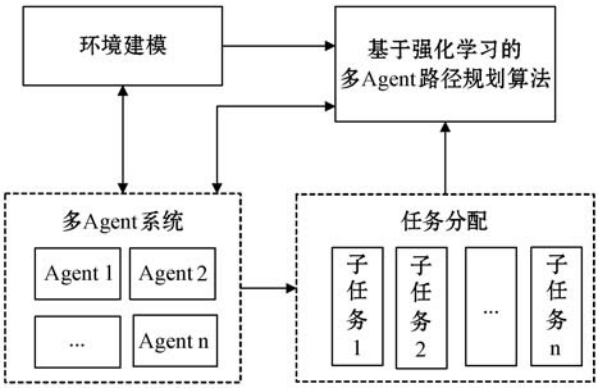


图 1 整体框架图

在图 1 中,首先对环境进行建模,包括对环境中障碍、目标点等信息设置,其次通过任务分配层主要根据实际任务划分多个子任务,然后算法层接收环境信息以及多个 Agent 信息和任务分配情况,并进行计算,将结果返回给 Agent。多 Agent 系统层与环境建模层、任务分配层、算法层进行交互,每个 Agent 均能执行动作与环境交互,同时也和任务分配模块的任务进行匹配,通过执行算法层,不断地更新策略,最后得到一组较优策略完成多个 Agent 的路径规划任务。

1.1 环境建模

对环境地图的建模常用的方法主要有三种:栅栏地图建模、拓扑地图建模和可视地图建模。本文采用的是栅栏建模法,如图 2 所示将环境分成 n^2 个面积相同的方格,每个方格均携带不同 0 ~ 3 的参数信息,当格子参数为 0 时表示该区域无障碍物,当格子参数为 1 时表示该区域含有障碍物,当格子参数为 2 时表示智能体的位置信息,当格子参数为 3 时表示目标点的位置信息。通过构建栅栏地图,能够很好地获取环境的信息。

0	0	0	1	1	1	1	0	0	0
0	2	1	1	1	1	1	0	0	0
0	0	0	1	0	0	1	3	1	0
0	0	0	1	1	0	0	0	1	0
0	0	2	0	1	1	0	0	1	0
0	1	0	0	0	0	1	0	1	0
0	1	0	0	0	1	0	0	1	0
0	1	1	1	1	1	1	3	0	0
0	1	1	0	1	1	1	1	1	0
0	0	1	0	1	1	1	1	0	0

图 2 栅栏环境图

1.2 任务分配

任务分配是多智能体协同合作中的一个重要研究内容。多 Agent 的路径规划的任务分配问题为:现假设系统环境中存在 m 个目标点,每个目标点至少一个 Agent 到达,所有目标点都有 Agent 到达时任务完成。

该任务分配的目标是将多个目标点分别分配给 Agent,以实现整体 Agent 到达目标点的路径总和最短。

1.3　多 Agent 路径规划算法

多 Agent 路径规划算法主要解决的问题是多个 Agent 的路径规划问题。本文采用的是基于强化学习的多 Agent 路径规划方法,多个 Agent 在同一环境中,不断与环境交互,根据环境的反馈进一步优化动作,完成整体的路径规划。对多个 Agent 进行路径规划主要有三个目标:一是对多个 Agent 进行路线规划时要考虑 Agent 间的路径冲突问题,避免多个 Agent 相撞;二是多个 Agent 进行路线选择时要避开障碍物;三是多个 Agent 到达目标点的路径总和尽可能的短。

2　基于强化学习的多 Agent 路径规划

2.1　强化学习相关理论

强化学习是一种无监督学习方法,Agent 通过与动态环境的反复交互,学会选择最优或近最优的行为以实现其长期目标[14]。Sutton 和 Barto 定义了强化学习方法的四个关键要素:策略、奖赏函数、价值函数、环境模型[15]。强化学习的基本模型主要包括环境和智能体两部分,如图 3 所示。

图 3　强化学习基本模型

在图 3 中,Agent 根据当前所处的环境状态,执行一个动作与环境交互,从环境中得到一个奖励,同时到达新的状态,进行学习更新策略,接着再执行一个动作作用于环境,不断重复此过程,优化策略完成任务。

很多强化学习问题可以形式化为马尔可夫决策过程(Markov decision process, MDP)。MDP 是由〈S, A, $\boldsymbol{P}$, R, γ〉构成的一个元组,其中:

S 是一个有限状态集;

A 是一个有限行为集;

$\boldsymbol{P}$ 是集合中基于行为的状态转移概率矩阵:

$\boldsymbol{P}^{a}_{ss'} = E[R_{t+1} \mid S_t = s, A_t = a]$;

R 是基于状态和行为的奖励函数:

$R^{a}_{s} = E[R_{t+1} \mid S_t = s, A_t = a]$;

γ 是一个衰减因子:$\gamma \in [0,1]$。

2.2　多 Agent 路径规划

在多 Agent 的强化学习过程中,每个 Agent 获得的奖励不仅仅取决于 Agent 自身的动作,同时还依赖于其他 Agent 的动作。因此本文将强化学习的 MDP 模型扩展为多马尔科夫决策过程(MDPs)。现假设有 n 个智能体,每个 Agent 可以选择的动作 m 个(即 $a_i, i = 1,2,\cdots,m$),每个 Agent 的状态个数为 k 个(即 $s_j, j = 1,2,\cdots,m$),则多个 Agent 采取的联合动作可以表示为 A_i,多个 Agent 的联合状态可以表示为 S_i。基于强化学习的基本模型,结合本文的任务目标,本文定义了多 Agent 的路径规划学习框架,具体情况如图 4 所示。

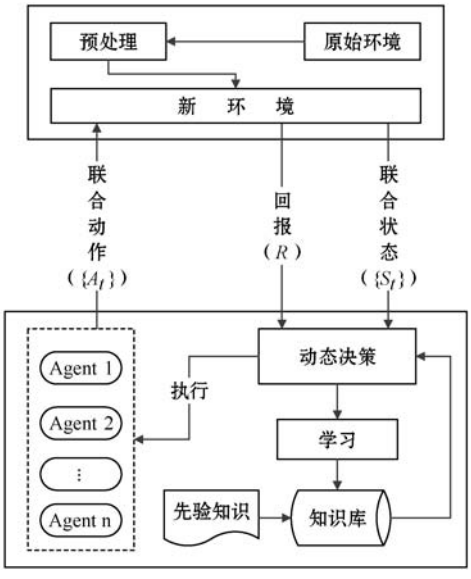


图 4　多 Agent 路径规划学习框架

在图 4 中,为了提高 Agent 的学习速度,本文中我们首先对多 Agent 所处的环境进行了预处理操作,剔除了一些无关的环境状态,同时将先验信息更新到知识库,提高了多 Agent 的学习效率。在该模型中,多个 Agent 基于当前所处的状态 S_t,每个 Agent 根据知识库的历史经验,按照一定的策略规则采取动作集中的一个动作 a_t,所有的 Agent 的动作组合成一次联合动作 A_t作用于环境。当联合动作 A_t执行完毕后,环境将转化为一个新的状态 S_{t+1},并且得到一个新的奖励值 R_{t+1}。然后进行学习,更新历史经验,进一步完善知识库。接着根据 Agen 所处的新状态 S_{t+1}和 R_{t+1}选择新的联合动作 A_{t+1}。多个 Agent 与环境进行周期性交互,不断重复“探索-学习-决策-利用”过程,从历史动作中进行学习更新自己的知识库,作为历史经验指导下次动作选择。

2.3　多 Agent 路径规划学习算法实现

2.3.1　联合状态设定准则

在多 Agent 的路径规划问题中,多个 Agent 所做的决策是基于环境的当前状态,因此对环境的状态定义十分重要。首先不考虑 Agent 的位置信息,将环境状态进行预处理操作,剔除一些如任务无关的环境状态,得到处理后的环境状态集合 $S = \{s_1, s_2, \cdots, s_n\}$。然而

要解决多 Agent 的路径规划问题,必须考虑多个 Agent 的位置信息,因此本文提出联合状态集合 S'。现假设存在 m 个 Agent,定义:C_s^1代表从集合 S 中随机选择一个状态,其中一个联合状态可以表示为 $S'_i = \underbrace{C_s^1, C_s^1, \cdots, C_s^1}_{m个}$。$m$ 个 Agent 的联合状态个数为 n^m个。

2.3.2 联合动作

动作空间表示了系统的解决方案空间。假设每个 Agent 的动作集合 $A = \{a_1, a_2, \cdots, a_k\}$,多个 Agent 在同一环境中要同时行动,因此需要引入联合动作集合 A',现假设该系统环境中存在 m 个 Agent,定义:C_A^1代表从集合 A 中随机选择一个动作,则其中一个联合动作可以表示为 $A'_i = \underbrace{C_A^1, C_A^1, \cdots, C_A^1}_{m个}$。$m$ 个 Agent 的联合动作个数为 k^m个。

2.3.3 奖励函数

奖励函数定义了 Agent 的学习目标,并确定了 Agent 基于环境的感知状态即时行动的价值。由于 Agent 试图最大限度地获得总报酬,因此奖励函数本质上是用来指导 Agent 实现其目标的。奖励函数的设置会决定强化学习算法的收敛速度和程度。常用的奖励函数设置方法有:稀疏奖励、形式化奖励、奖励系数变化奖励等。本文采用的是稀疏奖励的形式定义奖励函数,设置情况如下式所示:

$$R = \begin{cases} -a & 当\ Agent\ 碰到障碍物 \\ -b & 当\ Agent\ 相撞 \\ c & Agent\ 均到达目标点 \\ 0 & 其他情况 \end{cases} \tag{1}$$

式中:$a, b, c > 0$。

如式(1)所示,多 Agent 的路径规划目标是让多个 Agent 采取一组可以获得最大奖励的动作序列,到达指定的目标点。当 Agent 完成目标时,赋予一个正的奖励;当 Agent 碰到静态障碍物时,赋予一个负的奖励;当有两个或以上的 Agent 相互碰撞时,赋予一个负的奖励;其他情况的奖励值为 0。

2.3.4 价值更新函数

多 Agent 的路径规划采用的是 Q-learning 算法,在确定所有联合环境状态 S 和联合动作 A 后,要生成一个 $n^m \times k^m$ 维的矩阵 $\boldsymbol{Q}$,矩阵中的元素 $Q(S,A)$ 表示为多个 Agent 在环境状态 S_t下选择动作 A_t的价值。

更新的过程:当多个 Agent 在环境状态 S_t下,按照既定的动作选择策略,选择一个联合动作 A_t,执行完动作后 Agent 到达一个新的环境状态 S_{t+1},这时我们开始更新矩阵 $\boldsymbol{Q}$ 中的 $Q(S,A)$值。Agent 在状态 S_{t+1}时选择 $\boldsymbol{Q}$ 矩阵对应 Q 值最大的 $Q(S_{t+1}, A_{t+1})$,然后把 $Q(S_{t+1}, A_{t+1})$乘上一个衰减值 γ 并加上到达 S_{t+1}时所获取的奖励 R 作为现实中 $Q(S,A)$的值,然后减去之前的 $Q(S,A)$,接着乘以一个学习效率 α 累加上最初的 $Q(S,A)$的值则更新为新的 $Q(S,A)$。具体 $Q(S,A)$值的更新公式如下式所示:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R + \gamma \max_{A_{t+1}} Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \tag{2}$$

2.3.5 动作选择策略

在强化学习问题中,探索和利用是一对矛盾:探索意味着 Agent 必须尝试不同的行为继而收集更多的信息,利用则是 Agent 做出当前信息下的最佳决定[15]。探索可能会牺牲一些短期利益,通过搜集更多信息而获得较为长期准确的利益估计;利用则侧重于根据已掌握的信息而做到短期利益最大化。探索不能无止境地进行,否则就牺牲了太多的短期利益进而导致整体利益受损;同时也不能太看重短期利益而忽视一些未探索的可能会带来巨大利益的行为。

目前,常用的探索方法有:ε-贪婪探索、不确定优先探索以及利用信息价值进行探索等。本文采用的是 ε-贪婪探索,这里的 ε 是 Agent 随机选择的概率($0 \leqslant \varepsilon \leqslant 1$),在概率为 $1-\varepsilon$ 的情况下,Agent 使用贪婪的 Q 值方法选择 Q 值最大所对应的一个动作,当存在多个 Q 值相同的动作时随机选择一个;在概率为 ε 的情况下,Agent 从动作集合中随机选择动作。

2.3.6 多 Agent 路径规划算法步骤

在多 Agent 的路径规划中,多个 Agent 根据当前所处的环境状态,不断地与环境进行交互,在学习过程中对学习结果进行更新修正,用于指导 Agent 的动作选择,最终通过不断的学习,找到一组可以最大化奖励的动作序列,完成多 Agent 路径规划任务。该方法的伪代码如算法 1 所示。

算法 1 多 Agent 路径规划算法

```
Initialize: S_t, Q(s,a)
Repeat(for each episode):      Initialize S
    While S_t is not S_T
        If (Probability < ε)
            choose A_t = maxQ(S_t)
        Else
            Random choose A_t
        Take action A_t, return R 和 S'
        Update Q(s,a)
        S←S'
        If S_t is S_T
            Break
```

该算法的具体学习过程的形式化描述如下:

(1) 初始化设置:地图生成,设置 Agent 和目标点的数量及初始位置,奖励函数设置,Q 表初始化。

(2) 参数设置:终止学习周期 $T_{\max}$,学习效率 α、衰减度 γ 和探索度 ε。

(3) 根据 ε-贪婪策略选择动作 A_t。

(4) 执行 A_t,返回奖励值 R 和下一个状态 S_{t+1}。

(5) 按式(2)更新 Q 值。

(6) 判断是否满足终止条件:若满足终止条件,执行(7);否则,执行(3)。

(7) $T:=T+1$,判断 $T>T_{\max}$:若成立,则学习结束;否则转(3)。

3　实验仿真与分析

3.1　实验设置

为了验证该方法的有效性,本文多个 Agent 的路径规划设置了一个虚拟的环境。与文献[16]一样,本文创造了不同大小的栅栏地图环境,其中障碍和目标点是随机生成的。如图 5 所示,我们设置了包含 7 个障碍、两个智能体、两个目标点的 7×7 大小的原始环境地图。

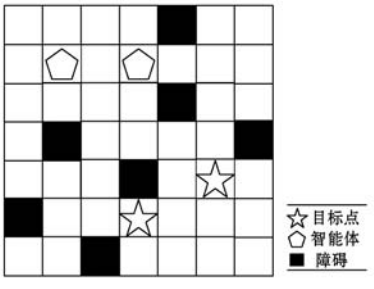


图 5　实验环境

针对同一任务目标,将文献[16]的方法与本文方法进行实验对比。其中文献[16]智能体的动作集合为$\{U,D,L,R,S\}$,其中 U 代表向上,D 代表向下,L 代表向左,R 代表向右,S 代表静止不动。本文方法的两个 Agent 的联合动作集为:

$$\boldsymbol{A}=\begin{bmatrix} UU & UD & UL & UR \\ DU & DD & DL & DR \\ LU & LD & LL & LR \\ RU & RD & RL & RR \\ SU & SD & SL & SR \\ US & DS & LS & RS \end{bmatrix}$$

其中文献[16]的奖励函数 R' 设置如式(3)所示,本文方法的奖励函数 R 具体设置如式(4)所示。

$$R'=\begin{cases} -1 & \text{碰到障碍} \\ 1 & \text{到达目标点} \\ 0 & \text{其他情况} \end{cases} \tag{3}$$

$$R=\begin{cases} -1 & \text{当 Agent 碰到障碍物} \\ -1 & \text{当 Agent 相撞} \\ 1 & \text{Agent 到达目标点} \\ 0 & \text{其他情况} \end{cases} \tag{4}$$

文献[16]和本文方法采用同一的学习更新函数的参数设置,如表 1 所示。

表 1　更新函数的参数设置

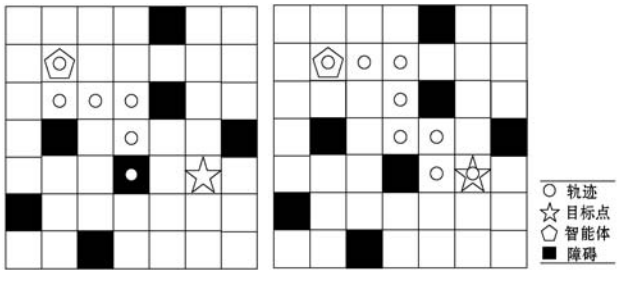
参数名	学习效率(α)	衰减度(γ)	探索度(ε)
值	0.01	0.9	0.1

本次实验假设两个智能体在环境中同时运动,不会出现故障情况,每次只能选择动作集合中的一个,环境是有边界的,当 Agent 选择超出边界的动作时,强制 Agent 留在环境内。任务目标是第 2 行第 2 列的 Agent1 到达第 5 行第 6 列的目标点,同时第 2 行第 4 列的 Agent2 到达第 6 行第 4 列的目标点,在 Agent 移动期间要避免相撞和避开障碍物。

3.2　实验结果与分析

为了验证本文方法的有效性,针对上述同一任务目标,进行两组实验,将本文方法与文献[16]方法进行对比,两组实验均训练 4 000 次。

本文运用文献[16]方法进行仿真实验,该方法分为两个阶段,首先分别对每个智能体进行路径规划,其次对发生碰撞的 Agent 进行动态调整。实验环境在图 5 基础上,分别进行单个智能体和目标点实验。首次实验时其中第 2 行第 2 列的 Agent1 运动轨迹如图 6(a)所示。在图 6(a)中,Agent1 在第 5 个步长时与静态障碍物发生碰撞,Agent1 的动作序列分别为:$\{D\rightarrow R\rightarrow R\rightarrow D\rightarrow D\}$,这是由于首次实验,Agent 并没有历史经验作为决策依据,而是随机的选择动作,不断“试错”。经过 Agent 不断与环境交互,更新 Q 表,进行动作选择,Agent 的最终路径规划路线结果如图 6(b)所示,Agent1 到达目标点的总步长为 7。

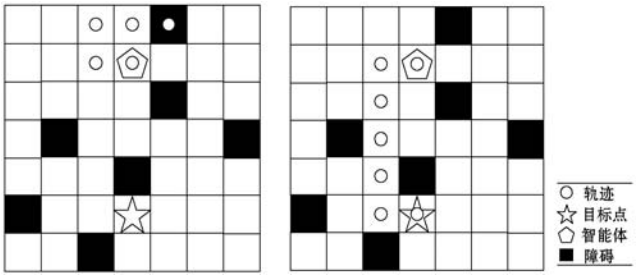


(a) 首次实验轨迹　　(b) 最终运动轨迹

图 6　Agent1 实验结果图

类似地,第 2 行第 4 列的 Agent2 运动轨迹如图 7(a)所示,Agent2 在第 4 个步长时与静态障碍物发生碰撞,Agent2 的动作序列为$\{L\rightarrow U\rightarrow R\rightarrow R\}$,经过

4 000 次学习,得到的最终路径规划结果如图 7(b)所示,Agent2 到达目标点的总步长为 6。



(a) 首次实验轨迹 (b) 最终运动轨迹

图 7 Agent2 首次实验运动轨迹

从图 6(b)和图 7(b)可以看出,当两个 Agent 在同一环境同时移动时,会在第 2 行第 3 列的位置相撞,运用动态规划思想对 Agent 的路径重新调整,最终的路径规划如图 8 所示。在图 8 中两个 Agent 在同一环境中同时移动,且能够躲避障碍物,两个 Agent 不会发生相撞,到达目标点路径最短。

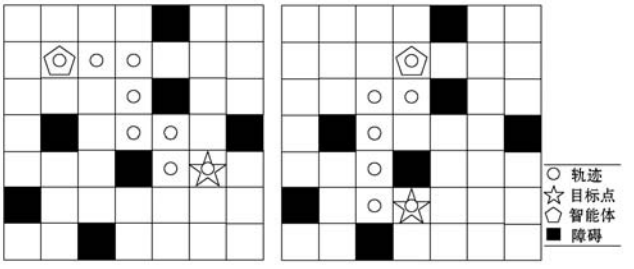


图 8 最终路径规划结果

运用本文的方法,在图 5 所示的环境中进行实验。首次实验时,两个 Agent 经过 18 个步长发生了相撞。这是由于本文的方法加入了先验信息,有历史经验作为决策支持,首次实验时避免了对障碍的学习,使 Agent 进行试错时避开了障碍。经过 499 次回合训练后,两个 Agent 第一次到达目标点,完成任务的总步长为 50。训练 4 000 次后最终的路径规划结果如图 9 所示,总步长为 14,其中联合动作序列为:

$$\{DL \rightarrow RS \rightarrow DD \rightarrow RD \rightarrow RD \rightarrow RD \rightarrow DR\}$$

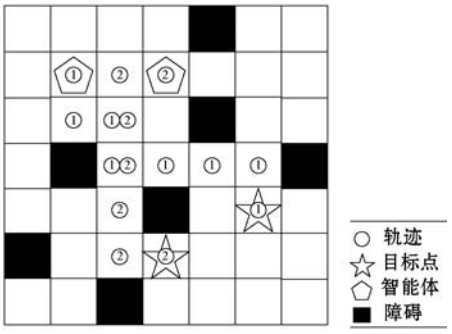


图 9 回合训练结果

为了验证本文的有效性,本文从总探索步数、完成任务的平均步数做了对比,具体情况如图 10、图 11 所示。在图 10 中,文献[16]的总探索步数是 65 810 步,本文方法的总探索步数是 54 375 步,由于本文方法两个 Agent 采取动作时要考虑双方的位置信息,引入联合动作,避免了对单个 Agent 相撞后的路径重新规划,减少了 17.4% 的总探索步数。从图 11 得出,本文完成任务的平均步数与文献[16]相比减少了 5 步。

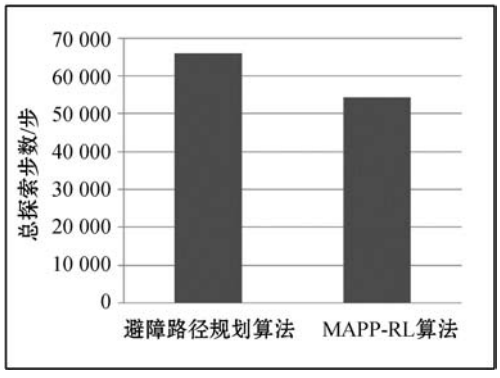


图 10 总探索步数

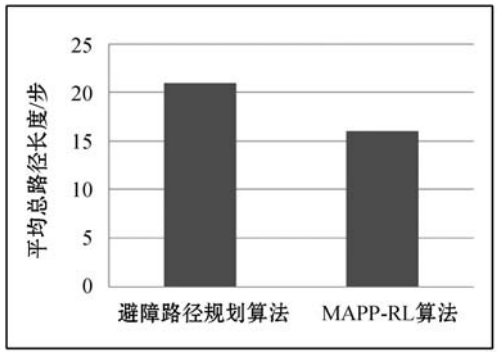


图 11 总路径平均步数

4 结 语

为解决复杂任务下多个 Agent 路径规划问题,本文提出一种基于强化学习的多 Agent 路径规划方法。首先建立了多 Agent 路径强化学习模型,并详细描述了各个基本要素,以及多个 Agent 如何从历史数据中积累经验优化决策。通过仿真实验表明,该方法是可行、有效的。为了提高该方法的学习效率,本文提出了 2 种解决方案:(1) 环境预处理,根据实际任务以及多 Agent 的信息,剔除一些无关的环境状态;(2) 加入先验信息的 Agent 决策 Q 表,基于先验信息更新 Q 表,作为历史经验提供给 Agent,大大提高了 Agent 的学习效率,与文献[16]方法相比,减少了 17.4% 的总探索步数。下一步将研究多 Agent 动态目标的路径规划问题,实现多 Agent 在复杂任务下的自主路径决策。

参 考 文 献

[1] Hildebrandt A C, Klischat M, Wahrmann D, et al. Real-Time Path Planning in Unknown Environments for Bipedal Robots[J]. IEEE Robotics and Automation Letters, 2017, 2

(4):1856 – 1863.
[2] Yu H, Meier K, Argyle M, et al. Cooperative Path Planning for Target Tracking in Urban Environments Using Unmanned Air and Ground Vehicles[J]. IEEE/ASME Transactions on Mechatronics, 2015, 20(2):541 – 552.
[3] Yang P, Tang K, Lozano J A, et al. Path Planning for Single Unmanned Aerial Vehicle by Separately Evolving Waypoints[J]. IEEE Transactions on Robotics, 2015, 31(5): 1130 – 1146.
[4] 熊超,解武杰,董文瀚. 基于碰撞锥改进人工势场的无人机避障路径规划[J]. 计算机工程, 2018, 44(9): 314 – 320.
[5] 姜涛,王建中,施家栋. 小型移动机器人自主返航路径规划方法[J]. 计算机工程, 2015, 41(1): 164 – 168.
[6] 刘洁,赵海芳,周德廉. 一种改进量子行为粒子群优化算法的移动机器人路径规划[J]. 计算机科学, 2017, 44(S2): 123 – 128.
[7] 赵晓,王铮,黄程侃,等. 基于改进 A * 算法的移动机器人路径规划[J]. 机器人, 2018, 40(6): 903 – 910.
[8] 游晓明,刘升,吕金秋. 一种动态搜索策略的蚁群算法及其在机器人路径规划中的应用[J]. 控制与决策, 2017, 32(3): 552 – 556.
[9] Lamini C, Fathi Y, Benhlima S. Collaborative Q-learning path planning for autonomous robots based on holonic multi-agent system[C]//2015 10th International Conference on Intelligent Systems: Theories and Applications(SITA). Rabat, 2015:1 – 6.
[10] 冯涛. 基于协同进化算法的多机器人路径规划研究[D]. 南京: 南京邮电大学,2015.
[11] 李东正,郝燕玲,张振兴. 基于主从结构的多水下机器人协同路径规划[J]. 计算机仿真, 2015, 32(1): 382 – 387.
[12] 郑延斌,李波,安德宇,等. 基于分层强化学习及人工势场的多 Agent 路径规划方法[J]. 计算机应用, 2015, 35(12): 3491 – 3496.
[13] 刘敬一. 自动化仓储调度系统中多 AGV 路径规划的研究与实现[D]. 沈阳: 中国科学院大学(中国科学院沈阳计算技术研究所),2018.
[14] Wang Z, Shi Z, Li Y, et al. The optimization of path planning for multi-robot system using Boltzmann Policy based Q-learning algorithm[C]//2013 IEEE International Conference on Robotics and Biomimetics(ROBIO). Shenzhen, 2013: 1199 – 1204.
[15] Sutton R S, Barto A G. Reinforcement Learning: An introduction[M]. Cambridge, Massachusetts: MIT Press, 1998.
[16] Kim J, Shin S, Wu J, et al. Obstacle Avoidance Path Planning for UAV Using Reinforcement Learning Under Simulated Environment[C]//IASER 3rd International Conference on Electronics, Electrical Engineering. Okinawa, 2017:34 – 36.

(上接第 92 页)

[11] Hangyo M, Nagashima T, Nashima S. Spectroscopy by pulsed terahertz radiation[J]. Measurement Science and Technology, 2002, 13(11):1727 – 1738.
[12] Withayachumnankul W, Fischer B M, Lin H, et al. Uncertainty in terahertz time-domain spectroscopy measurement[J]. Journal of the Optical Society of America B, 2008, 25(6):1059 – 1072.
[13] Duvillaret L, Garet F, Coutaz J L. Influence of noise on the characterization of materials by terahertz time-domain spectroscopy[J]. Journal of the Optical Society of America B, 2000, 17(3):452 – 461.
[14] Vartiainen E M, Ino Y, Shimano R, et al. Numerical phase correction method for terahertz time-domain reflection spectroscopy[J]. Journal of Applied Physics, 2004, 96(8): 4171 – 4176.
[15] Soltani A, Jahn D, Duschek L, et al. Attenuated Total Reflection Terahertz Time-Domain Spectroscopy: Uncertainty Analysis and Reduction Scheme[J]. IEEE Transactions on Terahertz Science & Technology, 2016, 6(1):32 – 39.
[16] Khazan M, Meissner R, Wilke I. Convertible transmission-re-flection time-domain terahertz spectroscopy[J]. Review of Scientific Instruments, 2001,72(8):3424 – 3430.
[17] Soltani A, Probst T, Busch S F, et al. Error from Delay Drift in Terahertz Attenuated Total Reflection Spectroscopy[J]. Journal of Infrared, Millimeter, and Terahertz Waves, 2014, 35(5):468 – 477.

(上接第 137 页)

[23] 王蓉蓉. 基于 GMM 语音谱包络表示的编码算法研究[D]. 南京:南京师范大学,2017.
[24] Zhang D, Entezami M, Stewart E, et al. Adaptive fault feature extraction from wayside acoustic signals from train bearings[J]. Journal of Sound and Vibration, 2018, 425:221 – 238.
[25] Wang C, Hu F, He Q, et al. De-noising of wayside acoustic signal from train bearings based on variable digital filtering[J]. Applied Acoustics, 2014, 83:127 – 140.
[26] Alemi A, Corman F, Lodewijks G. Condition monitoring approaches for the detection of railway wheel defects[J]. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit,2017,231(8):961 – 981.
[27] Li Z, He Q. Predicting failure times of railcar wheels and trucks by using wayside detector signals[C]//IEEE International Conference on Mechatronics & Automation. IEEE, 2014:1113 – 1118.
[28] Amini A, Entezami M, Huang Z, et al. Wayside detection of faults in railway axle bearings using time spectral kurtosis analysis on high-frequency acoustic emission signals[J]. Advances in Mechanical Engineering,2016,8(11):91 – 97.