基于探索-利用权衡优化的 Q 学习路径规划

彭云建,梁 进
(华南理工大学 自动化科学与工程学院,广东 广州 510640)

摘 要:针对移动智能体在未知环境下的路径规划问题,提出了基于探索-利用权衡优化的 Q 学习路径规划。对强化学习方法中固有的探索-利用权衡问题,提出了探索贪婪系数 ε 值随学习幕数平滑衰减的 εDBE(ε -decreasing based episodes)方法和根据 Q 表中的状态动作值判断到达状态的陌生/熟悉程度、做出探索或利用选择的 AεBS(adaptive ε based state)方法,这一改进确定了触发探索和触发利用的情况,避免探索过度和利用过度,能加快找到最优路径。在未知环境下对基于探索-利用权衡优化的 Q 学习路径规划与经典的 Q 学习路径规划进行仿真实验比较,结果表明该方法的智能体在未知障碍环境情况下具有快速学习适应的特性,最优路径步数收敛速度更快,能更高效实现路径规划,验证了该方法的可行性和高效性。

关键词:强化学习;Q 学习;探索-利用;路径规划;未知环境
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)04-0001-07
doi:10.3969/j.issn.1673-629X.2022.04.001

Q-learning Path Planning Based on Exploration/Exploitation Tradeoff Optimization

PENG Yun-jian, LIANG Jin
(School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China)

Abstract: Aiming at the path planning problem of mobile agent in unknown environment, a Q-learning path planning based on exploration/exploitation tradeoff optimization is proposed. For the inherent problem of exploration/exploitation tradeoff in reinforcement learning, the εDBE(ε -decreasing based episodes) method of exploring greedy coefficient ε value decreasing smoothly with the number of learning episodes and the AεBS (adaptiveεbased state) method of judging strangeness/familiarity of arriving state and making exploration or exploitation selection according to the state action value in Q table are proposed. This improvement determines the situation of triggering exploration or triggering exploitation, avoids over exploration and over exploitation, and can speed up finding the optimal path. In unknown environment, the Q-learning path planning based on exploration/exploitation tradeoff optimization is compared with the classical Q-learning path planning. The simulation results show that the agent with the proposed method has the characteristics of fast learning and adaptation in the unknown obstacle environment, the optimal path steps converge faster, and can realize the path planning more efficiently. The feasibility and efficiency of the proposed method are verified.

Key words: reinforcement learning; Q-learning; exploration/exploitation; path planning; unknown environment

0 引 言

随着人工智能的发展,能自主移动的智能体机器人在工业、军事以及医疗领域得到广泛使用[1],路径规划要求智能体避开障碍物,找到从出发点到目标点的最佳或次优路径[2],是移动智能体被广泛使用和发挥价值的基础。其中未知环境下的路径规划是研究的难点和热点,目前主要的方法有人工势场法[3]、神经网络、遗传算法、粒子群等智能算法[4]。

在利用强化学习解决未知情况下的路径规划方面,M. C. Su 等人提出在路径规划的理论中增加强化学习方法[5]。沈晶等人提出基于分层强化学习的路径规划的方法[6]。Y. Song 等人提出一种有效的移动机器人 Q 学习方法[7]。然而,在利用强化学习解决路径规划时,都会遇到强化学习本身固有的问题,即探索-利用问题[8]。为了解决探索-利用问题,目前提出的方法有 ε 贪婪方法和对其改进的 ε -first 方法[9]、ε -decreasing 方法[10],还有梯度算法[11]、value difference based exploration(VDBE)方法[12]等,但各有优点和不

收稿日期:2021-05-07 **修回日期**:2021-09-10
基金项目:国家自然科学基金(61573154)
作者简介:彭云建(1974-),男,副教授,研究方向为动态系统建模与应用、强化学习;梁 进(1996-),男,硕士研究生,研究方向为强化学习。

足,仍然有优化的空间。

该文根据优化 ε 值的改变方式和利用动作价值来动态选择采取的动作的思想,提出了基于探索-利用权衡优化的Q学习路径规划方法,解决移动智能体在未知环境下的路径规划问题。

1 探索-利用权衡优化的Q学习算法

为了实现智能体在未知环境下的路径规划,基于探索-利用权衡优化的Q学习路径规划可以分为两个部分,一是利用强化学习中Q学习不需要事先知道环境,智能体依然能与未知环境的互动中学习的特点,通过获得足够的幕数学习经验,不断更新Q表的动作价值,进而不断更新优化路径规划策略,实现路径规划;二是利用提出的 εDBE方法和AεBS,权衡强化学习中固有的探索-利用问题,提高未知环境下路径规划的快速性。

基于探索-利用权衡优化的Q学习路径规划如图1所示。提出改进探索-利用权衡问题的 εDBE方法和AεBS方法,着重优化 ε 值的改变方式和利用Q表中的动作价值来动态选择采取的动作,通过智能体与环境互动产生每幕学习经验来影响Q表动作价值的评估,进而获得更优动作行为、更新获得更优路径规划策略。

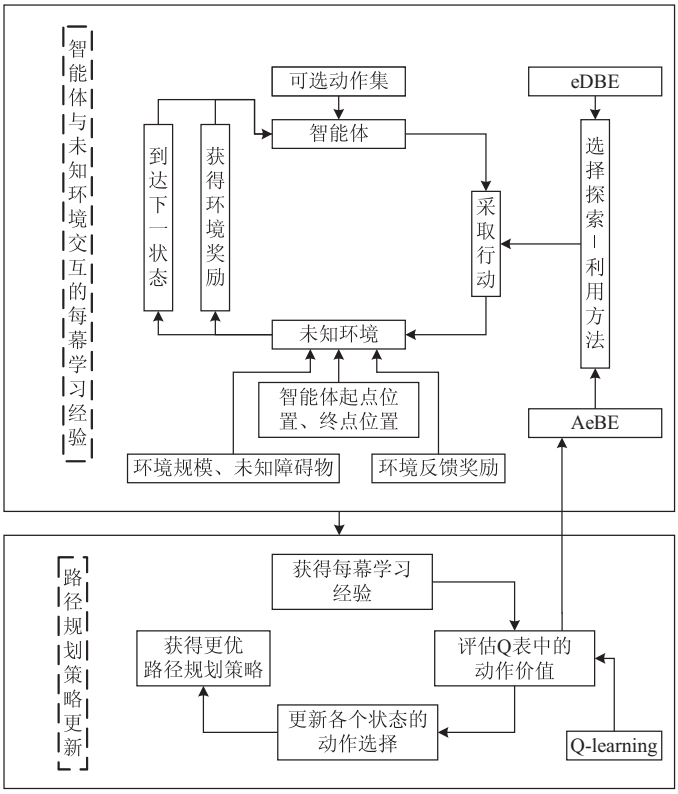


图1 探索-利用权衡优化的Q学习路径规划

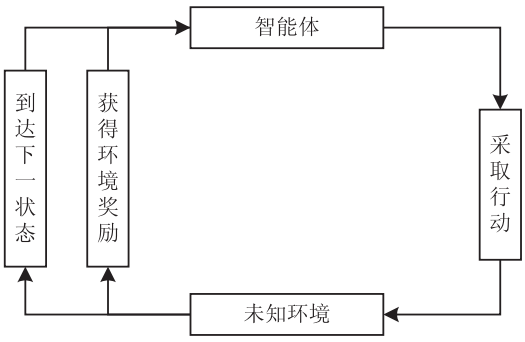


图2 智能体与环境交互图

智能体与环境交互如图2所示,每幕学习经验定义如下:在 t 时刻,智能体处于状态 s_t,采取动作 a_t,因此在 $t+1$ 时刻,智能体获得来自环境的奖励 r_{t+1},并在环境中发生了状态转移,到达了状态 s_{t+1}。在智能体与环境的不断交互过程中可获得一个状态、行动、奖励的序列:$s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3, \cdots, s_T$,其中 T 是终止时刻,这样有终止状态的一个序列也称为一幕(episode)学习经验。

2 改进探索-利用权衡问题的 εDBE 方法和 AεBS 方法

2.1 权衡探索-利用问题的基本方法

为了解决强化学习固有的探索-利用问题,经典的Q学习算法中采用了 ε-贪婪方法。之后有研究人员提出了改进 ε-贪婪方法的 ε-first 方法、ε-decreasing 方法,都是为了更好权衡探索-利用问题,提高Q学习算法解决问题的能力。

2.1.1　ε -贪婪方法

ε -贪婪方法的思想是设定一个小的贪婪探索系数,$0 < \varepsilon \leqslant 1$,在选择要采取哪个动作时,有 ε 的概率从所有可选的动作中随机选择,有 $1-\varepsilon$ 的概率选择目前能获得最大回报的动作。可用式(1)表示:

$$\pi(a \mid s)=\begin{cases}1-\varepsilon+\dfrac{\varepsilon}{m}, \text{if } a=a^* \\ \dfrac{\varepsilon}{m}, \text{ if } a \neq a^*\end{cases} \tag{1}$$

其中,$\pi(a \mid s)$ 为在状态 s 下选择动作 a 的概率,m 为状态 s 下动作集合 $A(s)$ 中动作 a 的总个数,$a \in A(s)$,a^* 为状态 s 下的最优动作。

2.1.2　ε -first 方法

ε -first 方法[9]的思想是一开始将 ε 的值设为 1,让智能体处于完全探索状态,一段训练幕数(episode)之后,将 ε 的值设为 0,让智能体处于完全利用环境状态。可用式(2)表示:

$$\varepsilon=\begin{cases}1, \text{if episode} \leqslant \text{preset\_episo} \\ 0, \text{if episode} > \text{preset\_episo}\end{cases} \tag{2}$$

其中,episode 为幕数变量, preset_episo 为预先设定的幕数值。

2.1.3　ε -decreasing 方法

改进的 ε -decreasing 方法[10]是 ε - 贪婪方法和 ε -first 方法的折中,思想是初始将 ε 设为一个较大的值,从训练幕数来看, ε 随着训练幕数增加不断减少;从单幕的步数来看, ε 随者步数增加而增大。可用式(3)表示:

$$\varepsilon=\varepsilon_0 * 0.1^{\frac{\text{episode}}{\text{step}}} \tag{3}$$

其中,ε_0 为贪婪系数的初始设定值,episode 为幕数变量,step 为每幕的步数变量。

2.2　εDBE 方法和 AεBS 方法

针对 Q 学习中固有的探索-利用问题,该文提出随幕数(episodes)平滑衰减 ε -值的 ε -decreasing based episodes(εDBE)方法,以及根据 Q 表中的状态动作值判断到达状态的陌生/熟悉程度、做出探索或利用选择的 adaptive ε based state (AεBS)方法。

2.2.1　εDBE 方法

随幕数(episodes)平滑衰减 ε 值的 εDBE 方法结合了 ε -decreasing 方法和 ε -贪婪方法的特点,即将初始 ε 设为一个较小的值,从训练幕数的角度来看,随着训练幕数增加而不断衰减;从单幕的步数角度来看 ε 保持不变,结合了 ε -decreasing 方法中 ε 衰减的特点,同时也具有 ε -贪婪方法在每一幕步数中 ε 保持不变的特点。在选择同时满足上述两个特点的 ε 衰减函数上,采用式(4)控制 ε 值的衰减。

$$\varepsilon=\frac{\varepsilon_0}{\sqrt{\text{episode}}} \tag{4}$$

其中,ε_0 为贪婪系数的初始设定值,$0 < \varepsilon_0 \leqslant 1$,episode 为幕数变量。

将式(4)与式(1)结合可得式(5)。

$$\pi(a \mid s)=\begin{cases}1-\dfrac{\varepsilon_0}{\sqrt{\text{episode}}}+\dfrac{\varepsilon_0}{m\sqrt{\text{episode}}}, \text{if } a=a^* \\ \dfrac{\varepsilon_0}{m\sqrt{\text{episode}}}, \text{if } a \neq a^*\end{cases} \tag{5}$$

规定了探索或利用的概率,即有 $\varepsilon_0/\sqrt{\text{episode}}$ 的概率从所有可选的动作中进行探索选择,有 $1-\varepsilon_0/\sqrt{\text{episode}}$ 的概率利用已学到的状态动作值,选择目前能获得最大回报的动作。在引入到下节的 Q 学习方法时,令从 Q 表中得到策略 π,通过 εDBE 方法进行策略评估和策略改进后得到的改进策略为 π',根据策略改进定理[13]可知,π' 相比于 π 更优,最终不断迭代后得到最优策略 π^*。

2.2.2　AεBS 方法

根据到达位置的陌生/熟悉程度和动作价值,从而做出探索/利用的动态动作选择 AεBS 方法。引入不断学习更新的 Q 表中动作价值作为陌生/熟悉程度的指标,当状态 s 下所对应的所有动作价值全为 0 时,认为该状态对于智能体来说是陌生的;当状态 s 下所对应的所有动作价值不全为 0 时,认为该状态对于智能体来说是熟悉的。在每幕学习的每一个步(step)中,遇到陌生的位置状态, ε 值变为 1,采取探索模式随机选择动作集中的任一动作;遇到熟悉的位置状态, ε 值变为 0,采取利用模式选择状态动作价值最大的动作。另外融合 ε -first 方法的思想,根据未知环境情况的不同,在幕数段中加入很小的 ε 值对 Q 表更新进行微调整。可用式(6)表示:

$$\varepsilon=\begin{cases}1, \text{ if } \forall Q(s,a)=0, a \in A(s) \\ 0, \text{ if } \exists Q(s,a) \neq 0, a \in A(s) \\ \varepsilon_0, \text{ if episode} \in [\text{episo1}, \text{episo2}]\end{cases} \tag{6}$$

其中, episode 为幕数变量, episo1 和 episo2 为设定的幕数值, ε_0 为贪婪系数的初始设定值, $0 < \varepsilon_0 \leqslant 1$, $A(s)$ 为状态 s 下的动作集合。

由于初始阶段中 Q 表的动作价值均初始化为零,因此采用 AεBS 方法的智能体可以充分探索环境,即每当遇到动作价值为零时智能体会判断出自身处于陌生环境,更倾向于随机选择不同的动作进行探索,更有可能不断遇到陌生情况,探索更为充分。同时在与环境的交互中不断更新 Q 表的动作价值,增加环境熟悉程度,从而利用 Q 表的动作价值的大小比较选择最优

动作,进而不断更新路径策略。

3 引入 εDBE 方法和 AεBS 方法的 Q 学习路径规划

在未知环境路径规划下,移动智能体在不同的状态 s 下通过策略 π 选择要采取的动作 a,与环境进行交互获得奖励 r,并到达下一状态 s'。重复上述过程不断迭代探索,更新 Q 表中的动作价值,找到更好的动作,直至找到最优策略 π^*,完成未知环境下的路径规划。时序差分方法是评估价值函数和寻找最优策略的实用方法。时序差分方法可以使智能体能直接与环境互动的经验中学习,不需要构建关于环境的动态特性。

Q 学习是 off-policy 下的时序差分控制方法,是强化学习的一个重要突破[14]。Q 学习更新的是动作价值函数,更新方法如式(7)所示:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma\max_a Q(s_{t+1},a) - Q(s_t,a_t)] \tag{7}$$

其中,α 为学习率,$0 < \alpha < 1$;γ 称为折扣因子,表示未来奖励对当前状态的影响程度[15],$0 \leqslant \gamma \leqslant 1$。

在 t 时刻智能体处于状态 s_t,动作状态价值为 $Q(s_t,a_t)$,当智能体采取动作 a_t 后在 $t+1$ 时刻到达状态 s_{t+1} 并获得奖励 r_{t+1},此时智能体将在 Q 表中找到能够使在状态 s_{t+1} 下动作价值最大的动作 a,以此来获得 $Q(s_{t+1},a)$,从而对 $Q(s_t,a_t)$ 进行更新。

可将式(7)改写成式(8)。

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha[r_{t+1} + \gamma\max_a Q(s_{t+1},a)] \tag{8}$$

假设 s_{t+1} 所对应的 $\max_a Q(s_{t+1},a)$ 恒定,通过式(8)可迭代求得稳定的 $Q(s_t,a_t)$。

一次迭代:

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha[r_{t+1} + \gamma\max_a Q(s_{t+1},a)] \tag{9}$$

二次迭代:

$$\begin{aligned} Q(s_t,a_t) &\leftarrow (1-\alpha)[(1-\alpha)Q(s_t,a_t) + \alpha[r_{t+1} + \gamma\max_a Q(s_{t+1},a)]] + \alpha[r_{t+1} + \gamma\max_a Q(s_{t+1},a)] \\ &\leftarrow (1-\alpha)^2 Q(s_t,a_t) + [1-(1-\alpha)^2][r_{t+1} + \gamma\max_a Q(s_{t+1},a)] \end{aligned} \tag{10}$$

以此类推,n 次迭代:

$$Q(s_t,a_t) \leftarrow (1-\alpha)^n Q(s_t,a_t) + [1-(1-\alpha)^n][r_{t+1} + \gamma\max_a Q(s_{t+1},a)] \tag{11}$$

因为 $0 < \alpha < 1$,所以 $0 < 1-\alpha < 1$,当 $n \to \infty$ 时,$Q(s_t,a_t)$ 将以概率 1 收敛到最优值,即:

$$Q(s_t,a_t) \leftarrow r_{t+1} + \gamma\max_a Q(s_{t+1},a) \tag{12}$$

当 Q 表更新后,根据式(13)即可选出状态下具有最大动作状态价值的动作,从而获得路径规划更优策略 π' 的更新。

$$\pi'(s) = \arg\max_a Q(s,a) \tag{13}$$

该文以稀疏奖励的形式定义奖励函数 r,如式(14)所示,将状态分为障碍状态、路径状态和目标终点状态,分别用状态集合 $O(s)$、$P(s)$、$G(s)$ 表示。其中到达障碍状态获得-1 奖励值,到达目标终点状态获得+1 奖励值,到达路径状态获得 0 奖励值,促使智能体避开障碍物快速到达目标终点。

$$r = \begin{cases} -1, \text{if } s \in O(s) \\ 0, \text{if } s \in P(s) \\ 1, \text{if } s \in G(s) \end{cases} \tag{14}$$

每个状态有上、下、左、右四个动作可选择,训练的过程为输入当前状态后,根据(εDBE)方法或根据(AεBS)方法从 Q 表中选出当前状态的相应动作,与未知环境交互后获得奖励,进入下一状态并判断是否撞到障碍物。

若判定会撞到障碍物,则根据式(8)更新 Q 表后结束本幕学习,开始下一个幕的学习;若判定不会撞到障碍物,则根据式(8)更新 Q 表后进入下一状态,本幕学习直至到达终点或判定会发生碰撞障碍物后结束。重复学习过程,不断更新 Q 表中各个状态的动作价值,直至找到最优策略,实现路径规划。

4 实验结果及分析

4.1 实验设计

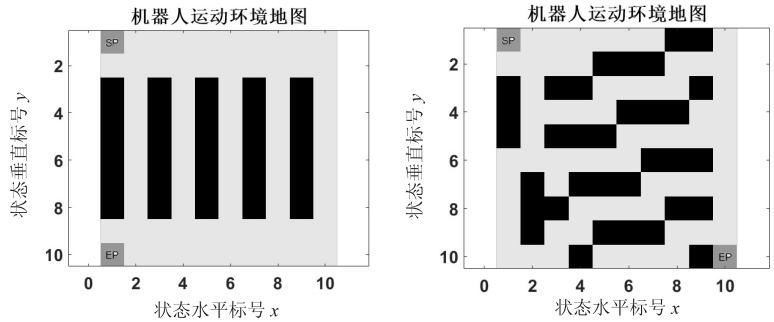
该文在 10 * 10 的地图上进行 Q 学习路径规划,设定了两种智能体未知的不同环境,对提出的基于探索-利用权衡优化的 Q 学习路径规划与基于经典的 ε-贪婪方法、ε-first 方法、ε-decreasing 方法的 Q 学习路径规划进行比较,验证提出方法的可行性和高效快速性。

其中每个网格对应一个状态,用不同的状态标号表示[16]。即在位置 (x,y) 处的网格对应的状态标号 stateno 可用式(15)表示。

$$\text{stateno} = 10(x-1) + y \tag{15}$$

图 3 所示为两种智能体未知的情况地图,状态 SS 为起点位置,GS 为终点位置,起始位置和路径均用深灰色表示,黑色为障碍物。智能体在每个无障碍物的浅灰色位置状态下,有上、下、左、右四个动作可以选择,碰到障碍物意味着一幕学习以失败结束,获得-1 奖励值,并返回起点位置;到达终点意味着一幕学习以成功结束,获得+1 奖励值,并返回起点位置;到达其余

状态均获得 0 奖励值。

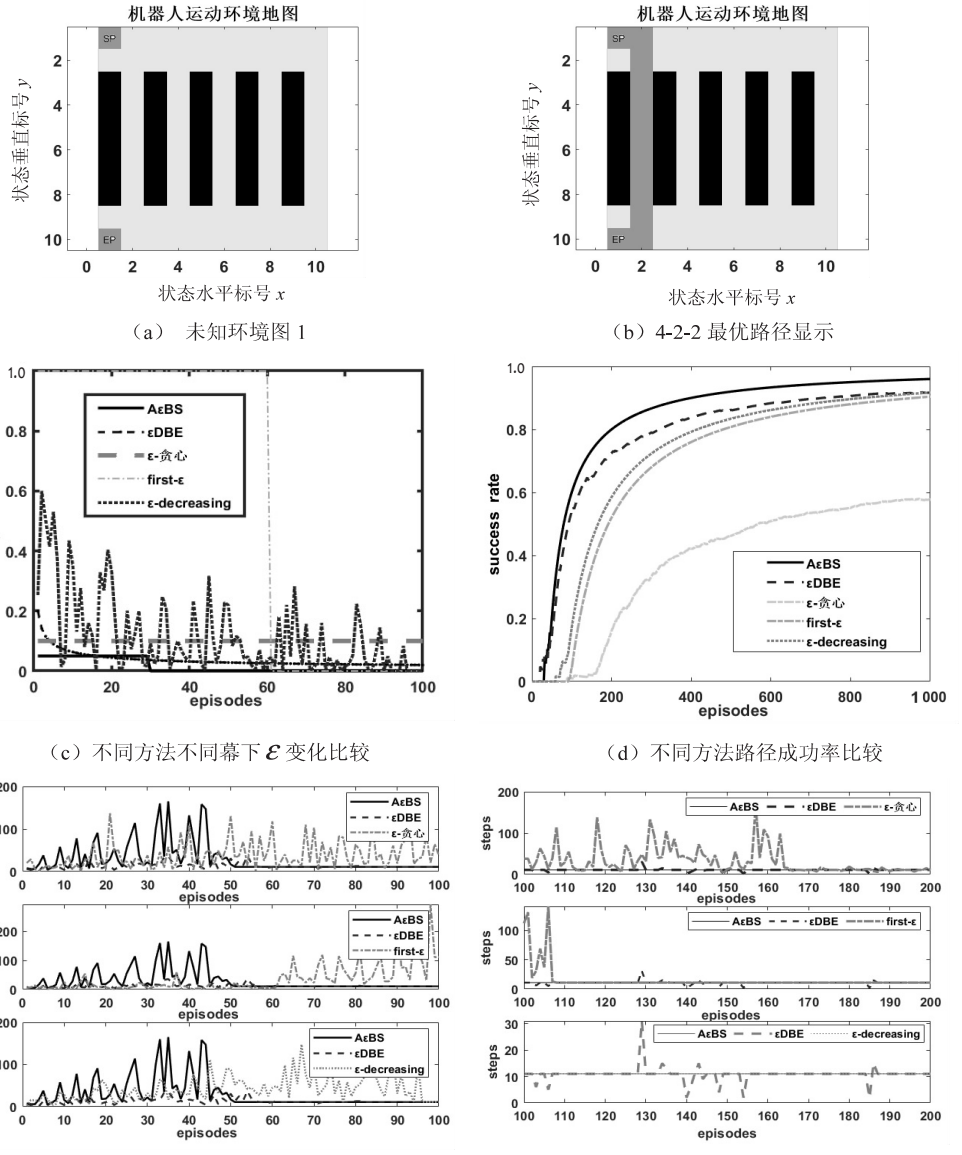


（a）未知环境地图 1　　（b）未知环境地图 2

图 3　两种智能体未知的环境地图

4.2　结果分析

通过 Q 学习路径规划可以得到以下仿真实验结果:图 4 所示为未知环境地图 1 下的仿真实验结果,其中折扣因子 $\gamma = 0.8$,学习率 $\alpha = 0.2$, ε -贪婪方法的 ε 值为 0.1, ε -decreasing 方法的 ε 初始值为 0.8, εDBE 的 ε 初始值为 0.2,AεBS 方法在 30 幕前的 ε 值为 0.05。从图 4(b)可以发现,Q 学习可以实现路径规划,找到从起点到终点的最优路径,状态转移步数为 11 步。



（a）　未知环境图 1　　（b）4-2-2 最优路径显示

（c）不同方法不同幕下 ε 变化比较　　（d）不同方法路径成功率比较

（e）不同方法路径步数收敛变化比较　　（f）不同方法路径步数收敛后特性比较

图 4　未知环境地图 1 中的仿真实验结果

从图 4(c)不同方法不同幕下 ε 变化比较中，ε -decreasing 方法中 ε 衰减过程较为波动，εDBE 方法中 ε 衰减过程较为平缓。

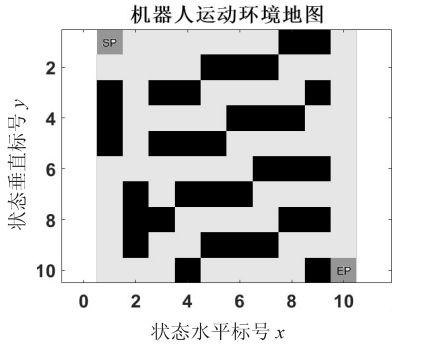
从图 4(d)不同方法路径成功率比较中可以看到，提出的(εDBE)方法和(AεBS)方法都比经典的 ε -贪婪方法、ε -first 方法、ε -decreasing 方法能更快找到最优路径，在相同的幕数经验学习下成功率更高。

从图 4(e)不同方法路径步数收敛变化比较中也可以看出，(εDBE)方法和(AεBS)方法最优路径收敛更快，ε -贪婪方法大约在 170 幕左右收敛、ε -first 方法大约在 110 幕左右收敛、ε -decreasing 方法大约在 100 幕左右收敛，(εDBE)方法和(AεBS)方法大约在 60 幕左右收敛。

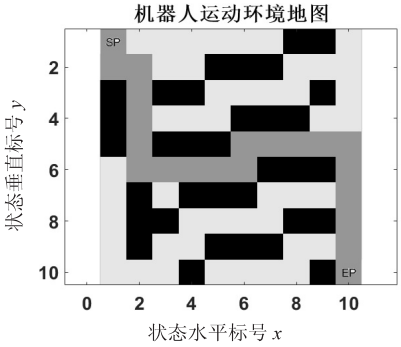
在图 4(f)不同方法路径步数收敛后方法特性比较中，由于(εDBE)方法和 ε -贪婪方法中 ε -值不为零，会出现一些细小的探索“尖刺”，但这些额外探索并不会妨碍智能体根据 Q 表中动作价值函数找到最优路径。

为了检验不同未知复杂环境下基于探索-利用权衡优化的 Q 学习路径规划方法的适应性，在未知环境地图 2 下继续进行仿真实验，其中折扣因子 $\gamma = 0.8$，学习率 $\alpha = 0.2$，ε -decreasing 方法的 ε 初始值为 0.8，εDBE 的 ε 初始值为 0.2，AεBS 方法在 100 幕到 300 间的 ε 值为 0.1。

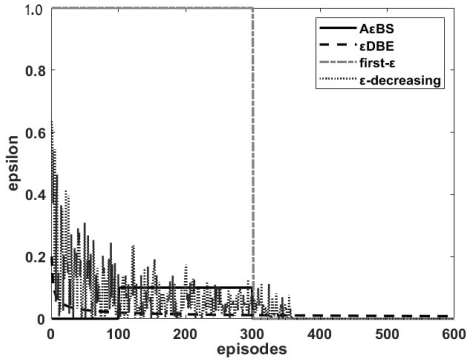
图 5 所示为未知环境地图 2 下的仿真实验结果。实验结果表明，提出的(εDBE)方法和(AεBS)方法较



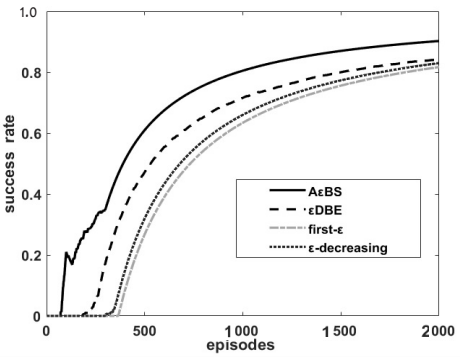
（a）未知环境图 2



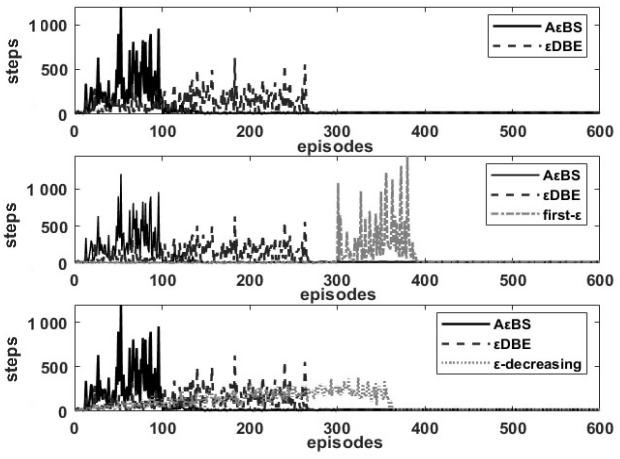
（b）最优路径显示



（c）不同方法不同幕下 ε 变化比较



（d） 不同方法路径成功率比较



（e）不同方法路径步数收敛变化比较

图 5　未知环境地图 2 中的仿真实验结果

ε -first 方法、ε -decreasing 方法,在 Q 学习路径规划中依然能更有效快速找到最优路径,状态转移步数为 20 步,最优路径收敛所需的学习幕数更少,找到路径的成功率更高。

5　结束语

针对移动智能体在未知环境下的路径规划问题,提出了基于探索-利用权衡优化的 Q 学习路径规划。实验结果表明,该方法具有可行性和高效性:提出方法能找到最优路径,实现路径规划;与现有的权衡探索-利用的 ε -贪婪方法、ε -first 方法、ε -decreasing 方法比较,提出的(εDBE)方法和(AεBS)方法能更好权衡探索-利用问题,在未知障碍环境情况下具有快速学习适应的特性,最优路径步数收敛速度更快,能高效可行地解决未知环境下的路径规划问题。

参考文献:

[1] 宋晓茹,任怡悦,高　嵩,等. 移动机器人路径规划综述[J]. 计算机测量与控制,2019,27(4):1-5.

[2] YAO Q,ZHENG Z,QI L,et al. Path planning method with improved artificial potential field—a reinforcement learning perspective[J]. IEEE Access,2020,8:135513-135523.

[3] 罗乾又,张　华,王　姮,等. 改进人工势场法在机器人路径规划中的应用[J]. 计算机工程与设计,2011,32(4):1411-1413.

[4] 王晓燕,杨　乐,张　宇,等. 基于改进势场蚁群算法的机器人路径规划[J]. 控制与决策,2018,33(10):1775-1781.

[5] SU M C,HUANG D Y,CHOU C H,et al. A reinforcement-learning approach to robot navigation[C]//Proc of the international conference on networking. Taiwan:IEEE,2004:665-669.

[6] 沈　晶,顾国昌,刘海波. 未知动态环境中基于分层强化学习的移动机器人路径规划[J]. 机器人,2006,28(5):544-547.

[7] SONG Yong,LI Yibin,LI Caihong,et al. An efficient initialization approach of Q-learning for mobile robots[J]. International Journal of Control,Automation and Systems,2012,10(1):166-172.

[8] 赵英男. 基于强化学习的路径规划问题研究[D]. 哈尔滨:哈尔滨工业大学,2017.

[9] CASTRONOVO M,MAES F,FONTENEAU R,et al. Learning exploration/exploitation strategies for single trajectory reinforcement learning[C]//European workshop on reinforcement learning. Edinburgh,UK:[s. n.],2013:1-10.

[10] CA E O,BONTEMPI G. Improving the exploration strategy in bandit algorithms[C]//Second international conference on learning and intelligent optimization. Trento,Italy:[s. n.],2007:56-68.

[11] 王建中,尹义龙. 基于动态信息模型的 LPN 路经规划方法[J]. 计算机工程,2006,32(17):66-68.

[12] TOKIC M. Adaptive ε-greedy exploration in reinforcement learning based on value differences[C]//Annual conference on artificial intelligence. AI,Karlsruhe,Germany:Springer Berlin,2010:203-210.

[13] DAYAN P. Motivated reinforcement learning[J]. Advances in Neural Information Processing Systems,2002,1:11-18.

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

[15] 刘志荣,姜树海,袁雯雯,等. 基于深度 Q 学习的移动机器人路径规划[J]. 测控技术,2019,38(7):24-28.

[16] KONAR A,CHAKRABORTY I G,SINGH S J,et al. A deterministic improved Q-learning for path planning of a mobile robot[J]. Systems,Man,and Cybernetics:Systems,2013,43(5):1141-1153.