Journal of System Simulation

Volume 36 | Issue 5 Article 15

5-15-2024

Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm

Xiaokang Wang
College of Engineering and Technology, Southwest University, Chongqing 400715, China,
1536340368@qq.com

Jie Ji
College of Engineering and Technology, Southwest University, Chongqing 400715, China,
jijiess@swu.edu.cn

Yang Liu
College of Engineering and Technology, Southwest University, Chongqing 400715, China

Qing He
College of Engineering and Technology, Southwest University, Chongqing 400715, China

Follow this and additional works at: https://dc-china-simulation.researchcommons.org/journal

 Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation. For more information, please contact xtfzxb@126.com.

Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm

Abstract

Abstract: To solve the traditional Q-learning algorithm for unmanned vehicle path planning suffers from the problems of low planning efficiency and slow convergence speed, for this reason, a path planning algorithm for unmanned delivery vehicles based on the improved Q-learning algorithm is proposed. Learning from the energy iteration principle of the simulated annealing algorithm, adjusts the greedy factor ε to make it change dynamically during the training process, so as to balance the relationship between exploration and utilization, and thus improve the planning efficiency. The reward value in the reward mechanism is changed from a discrete value to a continuous value, and it increases as the European distance between the unmanned delivery vehicle and the target point decreases, so that the target point can pull the unmanned delivery vehicle to move and accelerate the convergence speed of the algorithm. The improved Q-learning algorithm is simulated in two different environments, the simulation results show that the improved Q-learning algorithm can efficiently plan a path from the starting point to the target point with 34 steps, which is better path quality than comparison algorithms. The adaptability of the improved Q-learning algorithm to different environments is verified by changing the road environment, and the planning efficiency and convergence speed are still better than the traditional Qlearning algorithm.

Keywords

Q-learning, path planning, convergence speed, planning efficiency, path quality

Recommended Citation

Wang Xiaokang, Ji Jie, Liu Yang, et al. Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm[J]. Journal of System Simulation, 2024, 36(5): 1211-1221.

This paper is available in Journal of System Simulation: https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15

第 36 卷第 5 期 系统仿真学报© Vol. 36 No. 5
2024 年 5 月 Journal of System Simulation May 2024

基于改进Q学习算法的无人物流配送车路径规划

王小康，冀杰*，刘洋，贺庆
(西南大学 工程技术学院，重庆 400715)

摘要：为解决传统的Q学习算法用于无人车路径规划时，存在规划效率低和收敛速度慢等问题，为此，提出一种基于改进Q学习算法的无人物流配送车路径规划算法。借鉴模拟退火算法的能量迭代原理，对贪婪因子ε进行调整，使其在训练过程中动态变化，以平衡探索与利用之间的关系，提高规划效率。将奖励机制中的奖励值由离散值变为连续值，并使其随着无人物流配送车与目标点的欧式距离减小而增大，让目标点牵引无人物流配送车移动以加快算法收敛速度。在两种不同的环境下对改进的Q学习算法进行仿真实验，结果表明：改进后的Q学习算法可以高效地规划出一条从起始点至目标点的路径，步数为34步，优于对比算法的路径质量。通过改变道路环境，验证了改进Q学习算法对不同环境的适应性，规划效率和收敛速度依然优于传统Q学习算法。

关键词：Q学习；路径规划；收敛速度；规划效率；路径质量
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2024)05-1211-11
DOI: 10.16182/j.issn1004731x.joss.23-0051
引用格式: 王小康, 冀杰, 刘洋, 等. 基于改进Q学习算法的无人物流配送车路径规划[J]. 系统仿真学报, 2024, 36(5): 1211-1221.
Reference format: Wang Xiaokang, Ji Jie, Liu Yang, et al. Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm[J]. Journal of System Simulation, 2024, 36(5): 1211-1221.

Path Planning of Unmanned Delivery Vehicle Based on Improved Q-learning Algorithm

Wang Xiaokang, *Ji Jie**, *Liu Yang*, *He Qing*
(College of Engineering and Technology, Southwest University, Chongqing 400715, China)

Abstract: To solve the traditional Q-learning algorithm for unmanned vehicle path planning suffers from the problems of low planning efficiency and slow convergence speed, for this reason, a path planning algorithm for unmanned delivery vehicles based on the improved Q-learning algorithm is proposed. *Learning from the energy iteration principle of the simulated annealing algorithm, adjusts the greedy factor ε to make it change dynamically during the training process, so as to balance the relationship between exploration and utilization, and thus improve the planning efficiency. The reward value in the reward mechanism is changed from a discrete value to a continuous value, and it increases as the European distance between the unmanned delivery vehicle and the target point decreases, so that the target point can pull the unmanned delivery vehicle to move and accelerate the convergence speed of the algorithm.* The improved Q-learning algorithm is simulated in two different environments, the simulation results show that the improved Q-learning algorithm can efficiently plan a path from the starting point to the target point with 34 steps, which is better path quality than comparison algorithms. The adaptability of

收稿日期：2023-01-14 修回日期：2023-04-03
基金项目：重庆市科学技术局农业农村领域重点研发计划(cstc2021jscx-gksbX0003)；重庆市教育委员会科学技术研究项目(KJZD-M202201302)；重庆市博士后研究项目(2021XM3070)
第一作者：王小康(1999-)，男，硕士生，研究方向为车辆智能导航。E-mail:1536340368@qq.com
通讯作者：冀杰(1982-)，男，副教授，博士，研究方向为智能车辆与智能农机的道路环境感知、行为决策及底盘控制等。
E-mail： jijiess@swu.edu.cn

Journal of System Simulation, Vol. 36 [2024], Iss. 5, Art. 15

the improved Q-learning algorithm to different environments is verified by changing the road environment, and the planning efficiency and convergence speed are still better than the traditional Q-learning algorithm.

Keywords: Q-learning; path planning; convergence speed; planning efficiency; path quality

0 引言

传统的物流末端有多种配送方式，但依然存在配送效率低、人力资源浪费和成本高等情况，因此，无人物流配送车作为物流末端配送中一种新的交付方式，得到大力发展[1-3]。无人物流配送车的作业流程为在载货点装载货物后，自行规划路径，依靠机身携带的传感器感知并躲避障碍物，将货物送达目标点，提高了配送效率、弥补了人力不足、降低了作业成本[4-5]。

无人物流配送车包含感知、决策、规划和控制模块，其中，规划模块主要负责生成一条从起始点到目标点的最优可行驶路径[6-7]。按照搜索方式的不同，路径规划可以分为基于采样、基于图搜索及基于人工智能等方法[8]。Q学习算法作为一种强化学习算法，旨在解决智能体通过学习策略以达成回报最大化或实现特定目标的问题[9]，被逐渐应用于智能车辆的决策与路径规划，但该算法存在探索与利用的平衡问题，具有收敛速度慢等缺点[10]。传统Q学习算法的搜索策略中的贪婪系数为常数，使得整个训练过程中探索与利用的概率不变，导致易陷入局部最优和收敛速度过慢。基于该问题，文献[11]基于环境规模先对探索的步长进行阶段性调整，以减少搜索的重复度，再将每个阶段的路径进行拼接形成全局路径，该方法有效提升了收敛速度，但在不同的环境下，合适的阶段数量与每个阶段的步长不易获得。文献[12]引入反正弦函数对贪婪因子进行动态调整，并将改进的Q学习算法应用于AGV的路径规划中，对算法的收敛速度进行了改善。文献[13]设计了以softmax函数为主体的新搜索策略，使改进后的Q学习算法在前期选择不同动作的概率更加平等，且在训练后期不会选择次优动作，该方法有效提高了算法的收敛速度，但在路径规划中可能导致最优路径丢失。智能体与目标点之间的距离是影响训练时长的关键因素之一，文献[14]在Q学习算法中加入智能体与目标点的距离尺度来引导智能体向目标点移动，且在局部引入了虚拟目标点来绕开障碍物，提升了算法收敛速度的同时，增强了智能体避障能力。Q学习算法依靠Q表格的更新寻找最优策略，但Q表格的存储能力有限，当状态数量或者动作数量过多时，会产生维度灾难，文献[15]利用Q学习的4个派生性质来对Q表格只进行一次更新，相比传统Q学习算法中的反复更新，节省了大量存储空间和时间，降低了无人车硬件要求且规划路径的效率更高。文献[16]采用EM表记录距离信息，与Q表相结合提高了智能体对道路环境的应变能力，同时采用了静态和动态双重奖励机制，提高了智能体路径规划效率。

针对探索与利用的平衡和收敛速度两个问题，本文提出了一种改进的Q学习算法并用于无人物流配送车的路径规划。参考模拟退火算法的原理，设置动态贪婪系数，在保证不丢失最优路径的情况下加快收敛速度。同时，对奖励值进行改进，相比于传统Q学习算法中的离散式奖励值，基于欧式距离的连续奖励值能够有效减少无人物流配送车的盲目搜索。

1 基于改进Q学习算法的无人配送车训练过程

1.1 传统Q学习算法

1.1.1 马尔可夫决策过程

强化学习(RL)是机器学习中的一个重要领域，强调如何在环境中行动以取得最大收益，可描述为智能体与环境之间的相互学习过程[17]。Q学习

https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15
DOI: 10.16182/j.issn1004731x.joss.23-0051

Wang et al.: Path Planning of Unmanned Delivery Vehicle Based on Improved Q-le

算法是强化学习中一种典型的基于值的算法[18]，采用马尔可夫决策过程(MDP)的形式。在Q学习算法中，智能体需要和环境一直产生互动，在智能体和环境的交互过程中会产生一个序列：$S_0,A_0,R_1,S_1,A_1,R_2,S_2\cdots$，其中，$S_0,S_1,S_2$代表智能体的状态；$A_0,A_1$代表智能体的动作；$R_1,R_2$代表环境反馈。

该序列为一个序列决策过程，而MDP就是序列决策过程的公式化。MDP通常定义为一个四元组$M=\{S,A,P,R\}$，其中，S代表状态的集合，A代表动作空间的集合，P代表状态转移概率的集合，R代表奖励函数。在MDP形式下的无人物流配送车路径规划中，无人物流配送车从一个路径点移动到另一个路径点视为一种状态变化，且下一时刻的状态s'只和当前状态s有关。在状态$s\in S$下，对于某一特定动作$a\in A$，无人物流配送车从s转移到s'的转移概率为

$$P(s,a,s')=p(s'|s,a)\doteq \Pr(S_{t+1}=s'|S_t=s,A_t=a) \tag{1}$$

使用改进的Q学习算法对无人物流配送车进行路径规划，其目标是无人物流配送车在规定次数的训练下，寻找到最优策略π，使累计的奖励值最大，即找到从起始点到目标点的最优路径。由于无人物流配送车在未来状态选择动作得到的奖励值对当前状态的影响并不直观，所以加入了折扣率，加入折扣率后的累计奖励值为

$$G_t\doteq R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \tag{2}$$

式中：R_t为t时刻的回报值；$\gamma\in(0,1)$为折扣率，它决定了未来回报值占总回报值的比重。

状态价值函数表示无人物流配送车在某个状态s下价值的期望：

$$v_\pi(s)=E_\pi(G_t|S_t=s)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t=s\right] \tag{3}$$

式(3)并未考虑动作a所带来的影响，因此引入动作价值函数，即

$$q_\pi(s,a)=E_\pi(G_t|S_t=s,A_t=a)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t=s,A_t=a\right] \tag{4}$$

该式也称Q函数，用于计算Q学习算法中的Q值，表示无人物流配送车在某个状态下采取某个动作带来价值的数学期望。

1.1.2 Q学习算法

Q学习算法的主要思想是建立一个状态和动作组成的Q表格，通过Q表格的不断迭代变化来更新最优策略。Q值为在某一时刻的状态下，采取某个动作能够获得收益的期望。Q学习算法的主要优势是使用了时间差分法TD(融合了蒙特卡罗和动态规划)，能够进行离线学习[19]。

基于Q学习算法的无人物流配送车路径规划系统主要由无人物流配送车、状态、动作和道路环境组成。在某个时刻，无人物流配送车的状态用s_t表示，执行的动作用a_t表示。首先Q学习算法初始化$Q(s,a)$和初始状态s，无人物流配送车根据$\varepsilon-$贪婪策略选择动作a，获得反馈r和下一状态s'，更新Q值，如果无人物流配送车离目标点的距离变小，则状态–动作对得到的道路环境反馈值为正值，对应的Q值增大，如果无人物流配送车离目标点的距离变大或者碰到障碍物，则状态–动作对得到的道路环境反馈值为负值，对应的Q值减小，当无人物流配送车移动到目标点时，此轮结束，无人物流配送车返回起始点进行下一次训练。基于贝尔曼公式的Q学习算法更新为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_{t+1}+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)\right] \tag{5}$$

式中，$\gamma\in(0,1)$为折扣率；$\alpha\in(0,1)$为学习率。Q表格迭代更新完成后，无人物流配送车根据Q表格在每个状态下选择最优动作完成货物配送任务。传统的Q学习算法步骤如下所示。

步骤1：初始化$Q(s,a)=0,\forall s\in S,a\in A$；

步骤2：观察当前状态s，使用$\varepsilon-$贪婪策略选择一个动作a执行；

步骤3：执行完动作后，观察反馈r和新的

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 5, Art. 15

状态s'；

步骤4：根据式(5)更新Q值；

步骤5：将当前状态更新为下一个状态；

步骤6：重复第2~5步，直到达到预设的终止条件，例如达到最大迭代次数或者达到收敛条件。

图1为基于改进Q学习算法的无人物流配送车的路径规划框架，无人物流配送车与环境的交互学习过程为Q表格更新提供数据，在Q表格收敛时，选择Q值最大的状态动作对生成到达目标点的最优路径。

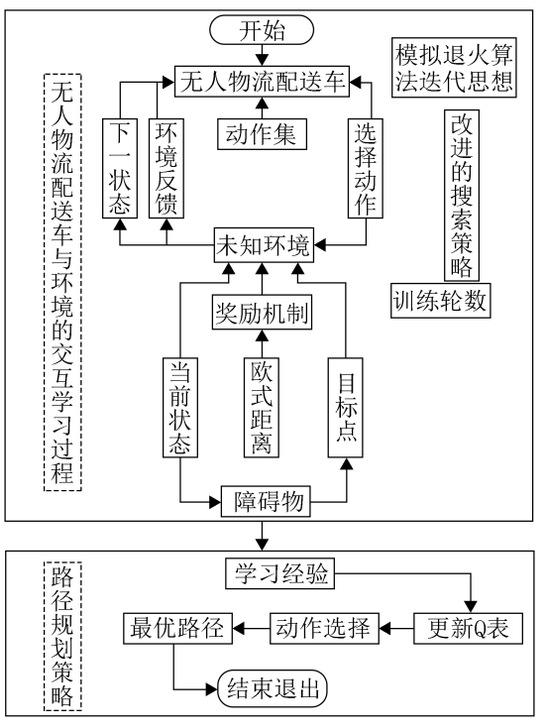


图1 无人物流配送车路径规划框架

Fig. 1 Path planning framework for unmanned delivery vehicle

1.2 Q学习算法改进

1.2.1 探索率动态调整

无人物流配送车需要对道路环境进行探索，同时也需要学习经验来最大化获得的奖励值。传统Q学习算法中无人物流配送车在选择动作时采用$\varepsilon-$贪婪策略，ε代表探索因子，ε的值越接近1无人物流配送车越倾向于探索环境，ε的值越接近0无人物流配送车越倾向于利用环境选择Q值最大的动作。$\varepsilon-$贪婪策略的原理是设置一个贪婪系数$\varepsilon\in(0,1)$，在选择动作时，有ε的概率从所有的动作中随机选择，有$1-\varepsilon$的概率选择具有最大奖励值的动作

$$\pi(a|s)=\begin{cases}1-\varepsilon+\varepsilon/m, a=a^*\\ \varepsilon/m, a\neq a^*\end{cases} \tag{6}$$

式中：$\pi a|s$为在状态s下选择动作a的概率；m为在状态s下动作集合A中动作的个数，动作$a\in A$；a^*为状态s下的最优动作。在该策略下，值函数的收敛效率较低且易陷入局部最优，故对贪婪因子进行动态调整。

模拟退火算法(SA)是一种常用的优化算法，它包含Metropolis算法和退火过程2个部分，其中，Metropolis算法主要是解决如何跳出局部最优解。如图2所示，固体在某一温度T下寻找能量最低值，迭代η次，迭代过程中固体温度不发生变化，能量发生变化，假设前一状态$x(n)$下，系统的能量为$E(n)$，系统根据某一指标状态变为$x(n+1)$，系统的能量变为$E(n+1)$，e为自然底数，则由状态$x(n)$变为$x(n+1)$的可接受概率为

$$P=\begin{cases}1, E(n+1)<E(n)\\ \mathrm{e}^{-\frac{E(n+1)-E(n)}{T}}, E(n+1\geqslant E(n))\end{cases} \tag{7}$$

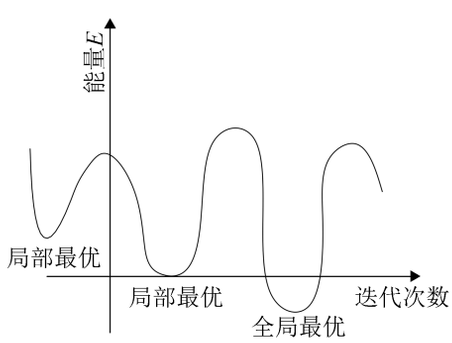


图2 迭代示意图

Fig. 2 Iteration diagram

结合模拟退火算法的思想，对Q学习算法中的$\varepsilon-$贪婪策略进行改进。在训练初期，无人物流配送车对环境进行广泛探索，寻找最优路径；在训练后期，无人物流配送车利用学习经验选择动作的概率增大，以加快算法收敛速度，同时仍保留一定的概率进行探索，防止陷入局部最优。为

https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15
DOI: 10.16182/j.issn1004731x.joss.23-0051

Wang et al.: Path Planning of Unmanned Delivery Vehicle Based on Improved Q-le

了达到这种结果，对$\varepsilon-$贪婪策略进行改进：

$$\varepsilon_k=\varepsilon_{\mathrm{f}}+\left(\varepsilon_1-\varepsilon_{\mathrm{f}}\right)\left(\mu_1+\mathrm{e}^{-\mu_2(k-N)}\right) \tag{8}$$

式中：ε_k、ε_1和ε_{f}分别为第k轮的探索率、初始探索率和最终探索率；μ_1、μ_2为比例因子；N为训练的总轮数。改进探索率后的$\varepsilon-$贪婪策略执行步骤如下。

步骤1：随机产生1个数$\omega\in(0,1)$；

步骤2：如果$\omega<\varepsilon_k$，随机选择动作$a\in A$，否则选择Q值最大的动作$a\in A$。

1.2.2 奖励机制调整

启发式搜索利用问题的启发信息来引导搜索，从而降低问题的复杂性，在人工智能中常用于空间规划。在每个状态下，启发式搜索需要考虑其延伸，启发信息越强，其延伸的无效节点就越少[20-21]。

奖励函数对Q学习算法的规划效率和收敛速度起着关键作用。在以往的研究中，奖励机制中大多采用的是到达目标点给予正奖励，碰到障碍物给予负奖励，中间状态奖励为0，导致无人物流配送车搜索盲目性大，探索效率低。将启发式思想与Q学习算法中的奖励机制相结合，以提高路径规划的效率。在实际应用中，无人物流配送车与目标点的距离对收敛速度有较大影响，因此，本文将无人物流配送车与目标点的欧氏距离作为启发式信息，设置即时奖励和无人物流配送车与目标点的欧式距离呈负相关，即随着欧式距离的减小，无人物流配送车获得的奖励值逐渐增大。具体设置为

$$r(s,a)=\begin{cases}r_1, d=0\\ \mu_3\mathrm{e}^{-\mu_4 d}, d\neq 0\end{cases} \tag{9}$$

式中：r_1为正常数，是无人物流配送车到达目标点的即时奖励；μ_3、μ_4分别为比例因子；d为当前状态下无人物流配送车与目标点的欧氏距离。

$$d=\left\| s-s_{\mathrm{goal}} \right\|_2 \tag{10}$$

改进后的Q学习算法奖励函数设计：当无人物流配送车到达目标点时，奖励值为+100，无人物流配送车碰到障碍物时，奖励值为−10，中间状态时，奖励值采用(9)式，具体的奖励函数为

$$r(s,a)=\begin{cases}100, \text{到达目标点}\\ -10, \text{与障碍物碰撞}\\ \mu_3\mathrm{e}^{-\mu_4 d}, \text{其他}\end{cases} \tag{11}$$

1.3 基于改进Q学习算法的无人物流配送车训练

将改进后的Q学习算法用于无人物流配送车的路径规划，具体训练过程如图3所示。

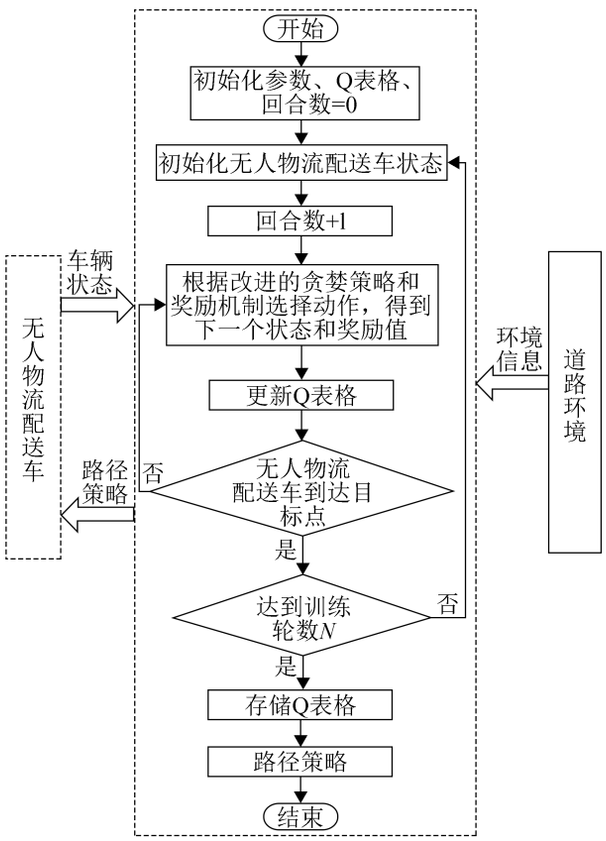


图3 改进后Q学习算法路径规划训练

Fig. 3 Path planning training of improved Q-learning algorithm

2 实验与分析

2.1 实验条件设定

该仿真实验针对无人物流配送车在水平道路上的路径规划问题，暂不考虑坡度及能耗。综合比较各种环境建模方法后，采用栅格法进行地图构建。栅格地图中不同颜色的栅格表示不同的环境信息，黑色栅格表示障碍物区域，白色栅格表

Published by Journal of System Simulation, 2024

Journal of System Simulation, Vol. 36 [2024], Iss. 5, Art. 15

示可通行区域。设定无人配送物流车为四轮模型，采用差速驱动实现转弯，其动作为沿上下左右四个方位移动，且一次移动一格。

2.2 参数设置

折扣率γ一般取0.9，能够充分考虑到未来奖励，γ的值过小会导致终点奖励值“辐射”范围过小，无人物流配送车可能无法到达目标点。学习率α用来权衡上一次学习结果和这一次学习结果的比重，一般取值为0.1，取值过大会导致在更新Q表时忽略了先前学习的经验。探索率ε表示无人物流配送车探索环境的概率，探索率过大会导致算法收敛速度过慢，过小会导致出现局部最优现象。

为了更好地对比3种算法的实验结果，3种算法的折扣率γ和学习率α由以往研究与经验进行设定。经过多次的预训练实验，确定合适的训练轮数N，初始探索率ε_1和最终探索率ε_f。传统Q学习算法与SARSA算法的探索率ε_2在训练过程中恒定不变，其值与初始探索率ε_1一致。比例因子μ_1、μ_2由式(8)在第1轮和第5 000轮时计算得出，比例因子μ_3、μ_4由式(11)在起始点和目标点计算得出。最终相关参数设置如表1所示。

表1 参数设置

Table 1 Parameter setting

参数	值	参数	值
折扣率γ	0.9	比例因子μ_2	0.000 1
学习率α	0.1	训练轮数N	5 000
探索率ε_2	0.4	比例因子μ_3	42.192 5
初始探索率ε_1	0.4	比例因子μ_4	–0.004 2
最终探索率ε_f	0.001	目标点即时奖励r_1	100
比例因子μ_1	–1		

2.3 仿真实验分析

时间差分方法(TD)分为两种类型：SARSA算法和Q学习算法，它们的本质区别在于更新Q值的方法不同。SARSA算法根据下一步的实际动作来更新Q表，对错误探索较为敏感，在动作选择上更为保守。Q学习算法根据下一步Q值最大的动作更新Q表，由于只在意Q值的最大化，因此，在动作选择上表现得更大胆。

在两种环境下进行仿真实验：环境I中，将SARSA算法、传统Q学习算法和改进的Q学习算法的仿真结果进行比较分析，以验证改进Q学习算法的可行性；在环境II中，对传统Q学习算法和改进Q学习算法的仿真结果进行对比分析，以验证改进Q学习算法对环境的适应性、收敛速度和规划效率。

2.3.1 建立仿真环境地图I

为了提高仿真的可信度和真实性，环境I选择重庆市主城区某路段，如图4所示。采用Tkinter创建栅格地图，地图大小为23×23，每一格为40像素，在该地图下，可以通过改变栅格的坐标位置对无人物流配送车和障碍物进行位置调整，如图5所示。

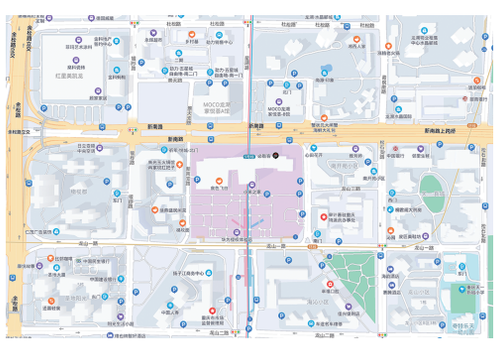


图4 重庆市某地段地图

Fig. 4 Map of a section in Chongqing

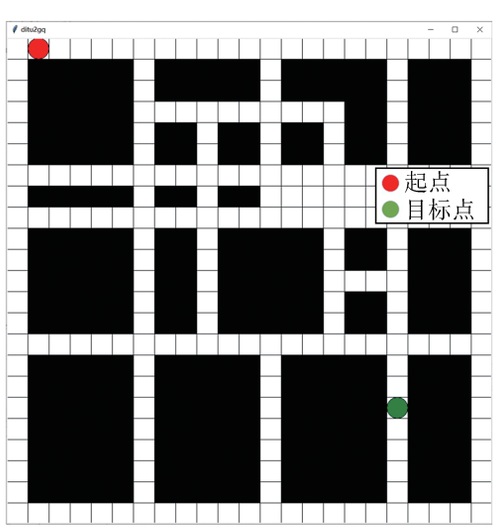


图5 仿真环境栅格地图

Fig. 5 Simulation environment grid map

https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15
DOI: 10.16182/j.issn1004731x.joss.23-0051

Wang et al.: Path Planning of Unmanned Delivery Vehicle Based on Improved Q-le

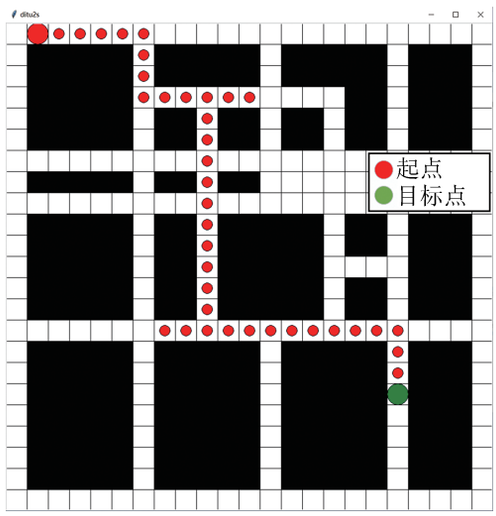
2.3.2 仿真实验I结果分析

图6是3种算法在地图I下训练结束后规划出的路径。使用SARSA算法规划的最短路径为42步，最长路径为2 130步，传统Q学习算法规划的最短路径为40步，最长路径为2 578步，改进后的Q学习算法规划的最短路径为34步，最长路径为2 288步。改进后的Q学习算法可以有效规划出一条从起始点至目标点的路径，且规划的路径更短、拐点更少。

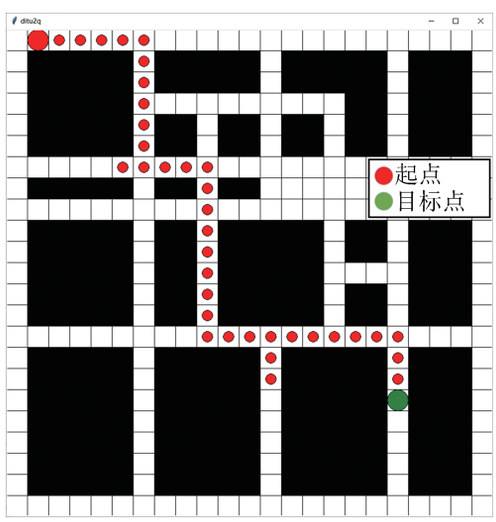
在训练前期，无人物流配送车以对环境探索为主，故在前期3种算法的步数都达到了2 000步以上。随着训练的进行，无人物流配送车对道路环境逐渐熟悉，利用的概率逐渐增大，步数变少，算法趋于收敛。如图7(a)所示，SARSA算法由于其过于保守，在4 000轮时才逐渐收敛。从图7(b)和图7(c)可以看出，传统Q学习算法在2 800轮左右开始收敛，改进的Q学习算法在2 500轮左右开始收敛。图7表明，SARSA算法探索时间长、收敛速度慢，传统Q学习算法由于贪婪系数常数，导致在前期的探索不够充分，易陷入局部最优，优化后的Q学习算法探索效率和收敛速度都得到了较大提升。

图8展示了3种算法训练中奖励值的变化。在训练初期，无人物流配送车随机选择动作的概率较大，易碰到障碍物，奖励值均为负值。SARSA算法和传统Q学习算法的初期奖励为–10 000左右，优化后的Q学习算法由于奖励值连续，初期奖励在–4 000左右。在训练后期，SARSA算法和传统Q学习算法的奖励值为800左右，改进后的Q学习算法奖励值为1 000左右，说明结合欧式距离的连续奖励值设置对无人物流配送小车起到了引导作用。

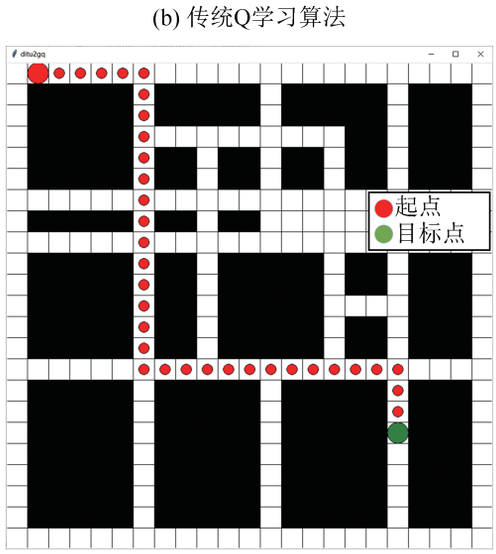
仿真实验I结果表明：3种算法都可以使无人物流配送小车规划出一条从起点到目标点的无碰撞可行驶路径，但改进后的Q学习算法相比其他两种算法在保证路径最优的情况下，用时更短，效率更高。



(a) SARSA算法



(b) 传统Q学习算法

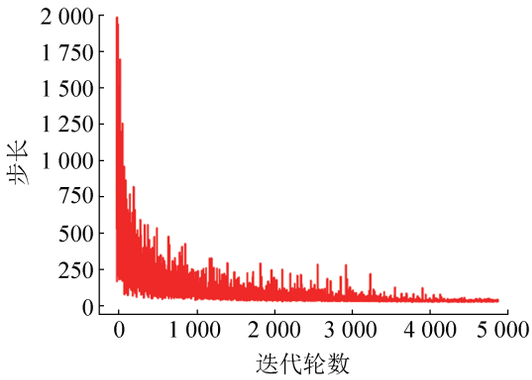


(c) 改进后的Q学习算法

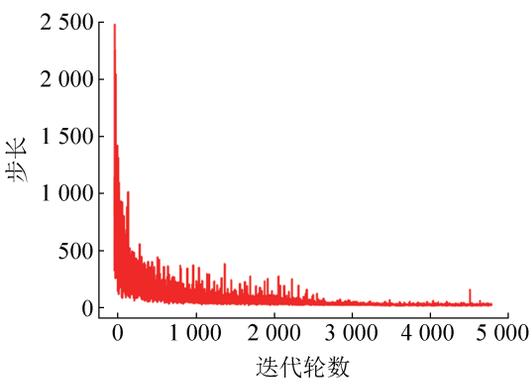
图6　3种算法规划的路径
Fig. 6　Paths planned by 3 algorithms

Published by Journal of System Simulation, 2024

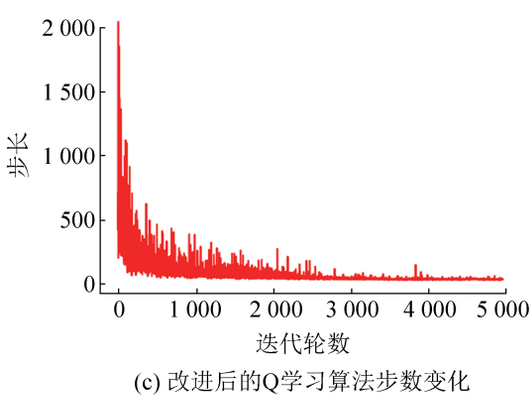
Journal of System Simulation, Vol. 36 [2024], Iss. 5, Art. 15



(a) SARSA算法步数变化



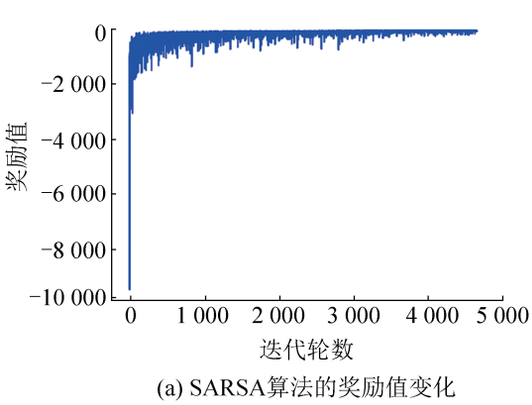
(b) 传统Q学习算法步数变化



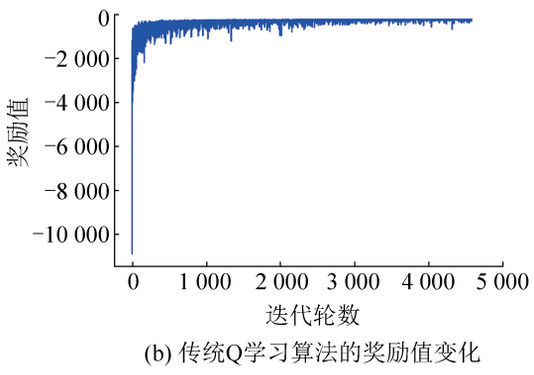
(c) 改进后的Q学习算法步数变化

图7 3种算法的步数变化

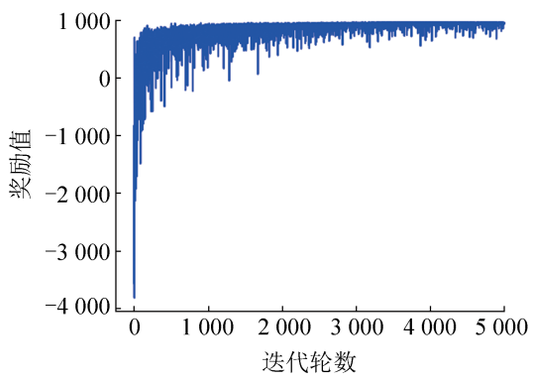
Fig. 7 Steps variation of 3 algorithms



(a) SARSA算法的奖励值变化



(b) 传统Q学习算法的奖励值变化



(c) 改进后的Q学习算法的奖励值变化

图8 3种算法的奖励值变化

Fig. 8 Rewards variation of 3 algorithms

2.3.3 建立仿真环境地图II

为了验证改进后的Q学习算法对环境的适应性，需在不同的环境下进行仿真实验。如图9所示，蓝色区域为正在施工的区域，禁止通行，无人物流配送车需要重新计算从起始点至目标点的路径。

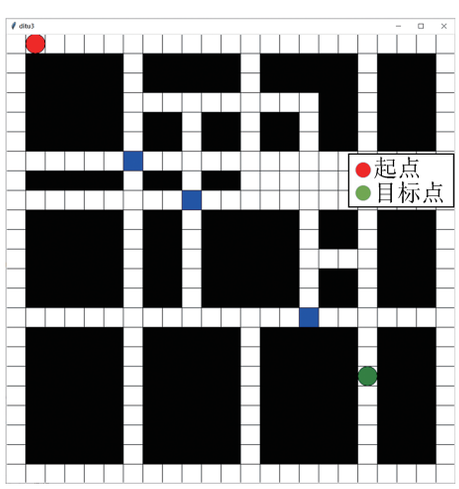


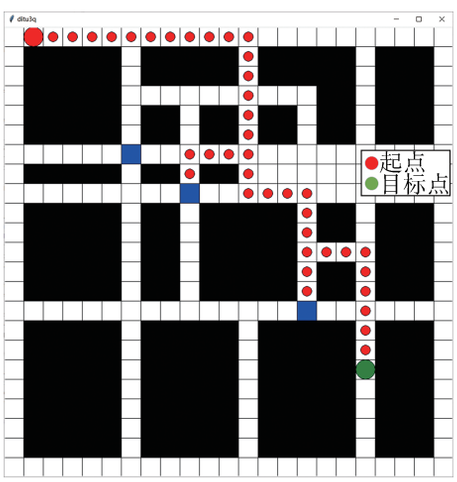
图9 施工环境下的栅格地图

Fig. 9 Grid map under construction environment

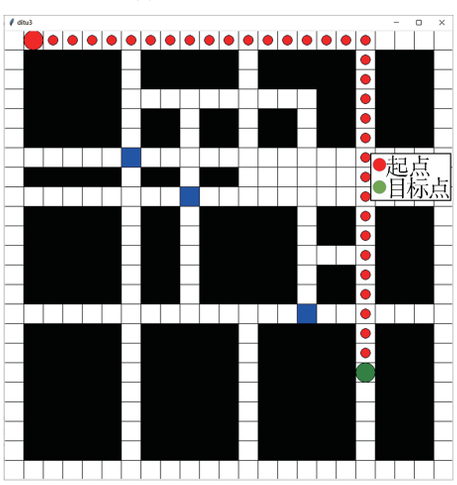
https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15
DOI: 10.16182/j.issn1004731x.joss.23-0051

2.3.4 仿真实验II结果分析

图10为传统Q学习算法和改进Q学习算法在在施工环境下规划的路径，传统Q学习算法规划的最短路径为46步，最长路径为3 662步，改进后的Q学习算法规划的最短路径为34步，最长路径为2 414步。在相同的训练次数下，改进后的Q学习算法规划的路径长度小于传统Q学习算法。



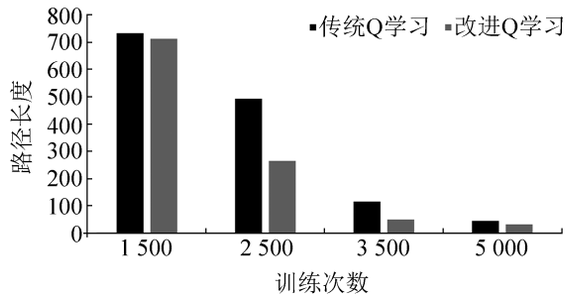
(a) 传统Q学习算法



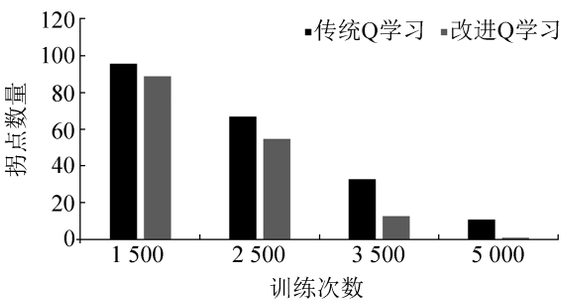
(b) 改进后的Q学习算法

图10　2种算法规划的路径
Fig. 10　Paths planned by 2 algorithms

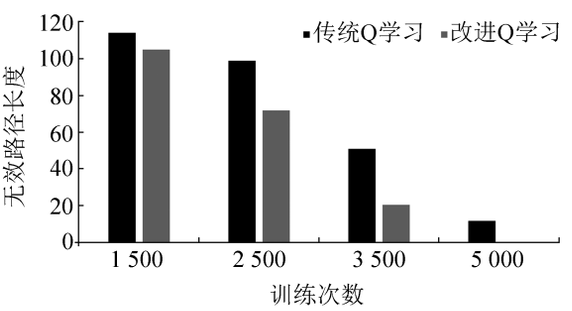
图11展示了训练过程中2种算法产生路径的评价指标，拐点数量越少，无人物流配送车行驶所消耗的能量越少。无效步数是指无人物流配送车在寻找最优路径时所产生的多余步长。在其他条件不变的情况下，3个指标越小，所产生的路径越优。



(a) 路径长度变化



(b) 拐点数量变化



(c) 无效路径长度变化

图11　训练过程中2种算法的评价指标
Fig. 11　Evaluation metrics for two algorithms during training

从以上数据可以看出，在相同的训练次数下，改进后的Q学习算法的无效步数比传统Q学习算法小，说明改进后的Q学习算法减少了不必要的探索，加快了收敛速度，而拐点数量和路径长度更小，表明了路径质量更优。

在训练前期2种算法主要以探索为主，步长均达到2 000以上，避免了局部最优的情况。如图12所示，改进后的Q学习算法由于探索因子ε的动态变化，在2 500轮以后开始逐渐收敛，而传统Q学习算法由于探索因子ε为固定值，在整个训练过程中并未出现明显收敛的情况。

Journal of System Simulation, Vol. 36 [2024], Iss. 5, Art. 15

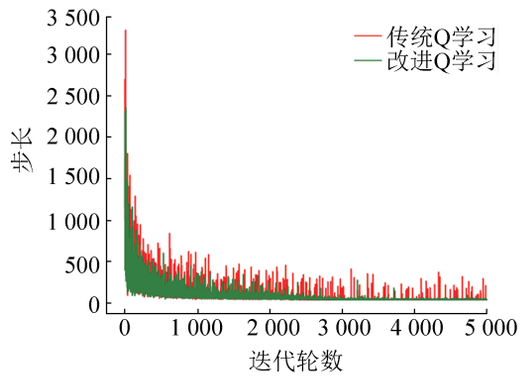


图12　2种算法的步数变化
Fig. 12　Steps variation of two algorithms

在前期探索环境过程中，无人物流配送车碰到障碍物的概率极大，2种算法的奖励值均达到了–6 000以下，如图13所示。与传统Q学习算法相比，改进后的Q学习算法奖励值由离散变为了连续导致波动幅度更大，但由于在奖励机制中加入了起始点与目标点之间欧氏距离的影响，使得即使在较大的波动幅度下，在3 000轮以后，依然能够快速收敛。

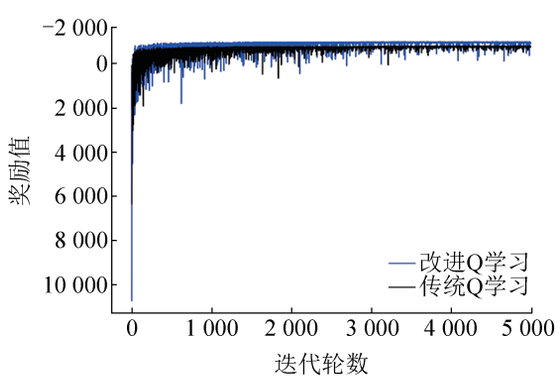


图13　2种算法的奖励值变化
Fig. 13　Rewards variation of two algorithms

仿真实验Ⅱ结果表明，改进后的Q学习算法可以适应不同的道路环境，且在规划效率和收敛速度方面，依然优于传统Q学习算法。

3　结论

本文针对无人物流配送车在城市道路下的路径规划问题展开研究，就传统Q学习算法易陷入局部最优和收敛速度慢2方面进行分析与改进。

(1) 为了平衡探索与利用问题，参考模拟退火算法的原理，提出一种贪婪系数ε动态变化的搜索策略，在保证无人物流配送车不陷入局部最优的情况下提高算法收敛速度。

(2) 传统的离散奖励值作用有限，新提出一种基于启发式思想的连续奖励值机制，令即时奖励和无人物流配送车与目标点之间的欧式距离呈负相关，使目标点牵引无人物流配送车移动，加快了算法收敛速度。

(3) 在2种不同的环境下进行了仿真实验，第一种环境下，将结果与SARSA算法和传统Q学习算法进行了对比分析，验证了改进后的Q学习算法的可行性。第2种环境下，将结果和传统Q学习算法进行了对比分析，表明改进后的Q学习算法可以适应不同的环境，且收敛速度和规划的路径质量依然优于传统Q学习算法，提高了无人物流配送车的作业效率。

参考文献：

[1] 张辉, 张瑞亮, 许小庆, 等. 基于关键节点的改进A*无人车路径规划算法[J]. 汽车技术, 2023(3): 10-18.
Zhang Hui, Zhang Ruiliang, Xu Xiaoqing, et al. Key Nodes-based Improved A* Algorithm for Path Planning of Unmanned Vehicle[J]. Automobile Technology, 2023 (3): 10-18.

[2] Li Xiaowei, Li Qing, Yin Chengqiang, et al. Autonomous Navigation Technology for Low-speed Small Unmanned Vehicle: An Overview[J]. World Electric Vehicle Journal, 2022, 13(9): 165.

[3] 罗洁, 王中训, 潘康路, 等. 基于改进人工势场法的无人车路径规划算法[J]. 电子设计工程, 2022, 30(17): 90-94, 99.
Luo Jie, Wang Zhongxun, Pan Kanglu, et al. Unmanned Vehicle Path Planning Algorithm Based on Improved Artificial Potential Field Method[J]. Electronic Design Engineering, 2022, 30(17): 90-94, 99.

[4] 黄凯文, 赵煜, 黄玲, 等. 基于机械视觉的Arduino智能物流配送车[J]. 河南科技, 2021, 40(22): 19-23.
Huang Kaiwen, Zhao Yu, Huang Ling, et al. Arduino Intelligent Logistics Delivery Vehicle Based on Mechanical Vision[J]. Henan Science and Technology, 2021, 40(22): 19-23.

[5] Wu Yuzhan, Ding Yuanhao, Ding Susheng, et al. Autonomous Last-mile Delivery Based on the

https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/15
DOI: 10.16182/j.issn1004731x.joss.23-0051

Wang et al.: Path Planning of Unmanned Delivery Vehicle Based on Improved Q-le

Cooperation of Multiple Heterogeneous Unmanned Ground Vehicles[J]. Mathematical Problems in Engineering, 2021, 2021: 5546581.

[6] 刘珂, 董洪昭, 张丽梅, 等. 基于改进人工势场法的物流无人配送车路径规划[J]. 计算机应用研究, 2022, 39(11): 3287-3291.
Liu Ke, Dong Hongzhao, Zhang Limei, et al. Path Planning for Logistics Unmanned Delivery Vehicles Based on Improved Artificial Potential Field Method[J]. Application Research of Computers, 2022, 39(11): 3287-3291.

[7] Li Jianqiang, Sun Tao, Huang Xiaopeng, et al. A Memetic Path Planning Algorithm for Unmanned Air/Ground Vehicle Cooperative Detection Systems[J]. IEEE Transactions on Automation Science and Engineering, 2022, 19(4): 2724-2737.

[8] 翟丽, 张雪莹, 张闲, 等. 基于势场法的无人车局部动态避障路径规划算法[J]. 北京理工大学学报, 2022, 42(7): 696-705.
Zhai Li, Zhang Xueying, Zhang Xian, et al. Local Dynamic Obstacle Avoidance Path Planning Algorithm for Unmanned Vehicles Based on Potential Field Method [J]. Transactions of Beijing Institute of Technology, 2022, 42(7): 696-705.

[9] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, et al. Q-learning Algorithms: A Comprehensive Classification and Applications[J]. IEEE Access, 2019, 7: 133653-133667.

[10] 李远哲, 胡纪滨. 强化学习在无人车领域的应用与展望[J]. 信息与控制, 2022, 51(2): 129-141.
Li Yuanzhe, Hu Jibin. Applications and Prospect of Reinforcement Learning in Unmanned Ground Vehicles [J]. Information and Control, 2022, 51(2): 129-141.

[11] 杨秀霞, 高恒杰, 刘伟, 等. 基于阶段Q学习算法的机器人路径规划[J]. 兵器装备工程学报, 2022, 43(5): 197-203.
Yang Xiuxia, Gao Hengjie, Liu Wei, et al. Robot Path Planning Based on Stage Q Learning Algorithm[J]. Journal of Ordnance Equipment Engineering, 2022, 43 (5): 197-203.

[12] 张祥来, 江尚容, 罗芹. 基于改进Q学习算法的"货到人"系统AGV路径规划[J]. 现代计算机, 2022, 28(2): 62-66, 72.
Zhang Xianglai, Jiang Shangrong, Luo Qin. Research on AGV Path Planning of "Goods-to-person" System Based on Q-learning[J]. Modern Computer, 2022, 28(2): 62-66, 72.

[13] 赵也践, 王艳红, 张俊, 等. 改进Q学习算法在作业车间调度问题中的应用[J]. 系统仿真学报, 2022, 34(6): 1247-1258.
Zhao Yejian, Wang Yanhong, Zhang Jun, et al. Application of Improved Q Learning Algorithm in Job Shop Scheduling Problem[J]. Journal of System Simulation, 2022, 34(6): 1247-1258.

[14] Ee Soong Low, Pauline Ong, Cheng Yee Low, et al. Modified Q-learning with Distance Metric and Virtual Target on Path Planning of Mobile Robot[J]. Expert Systems with Applications, 2022, 199: 117191.

[15] Amit Konar, Indrani Goswami Chakraborty, Sapam Jitu Singh, et al. A Deterministic Improved Q-learning for Path Planning of a Mobile Robot[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2013, 43 (5): 1141-1153.

[16] Zhao Meng, Lu Hui, Yang Siyi, et al. The Experience-memory Q-learning Algorithm for Robot Path Planning in Unknown Environment[J]. IEEE Access, 2020, 8: 47824-47844.

[17] Zhang Lieping, Tang Liu, Zhang Shenglan, et al. A Self-adaptive Reinforcement-exploration Q-learning Algorithm[J]. Symmetry, 2021, 13(6): 1057.

[18] Hu Yanming, Li Decai, He Yuqing, et al. Incremental Learning Framework for Autonomous Robots Based on Q-learning and the Adaptive Kernel Linear Model[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(1): 64-74.

[19] Ma Xin, Xu Ya, Sun Guoqiang, et al. State-chain Sequential Feedback Reinforcement Learning for Path Planning of Autonomous Mobile Robots[J]. Journal of Zhejiang University Science C, 2013, 14(3): 167-178.

[20] Shang Erke, Dai Bin, Nie Yiming, et al. An Improved A-star Based Path Planning Algorithm for Autonomous Land Vehicles[J]. International Journal of Advanced Robotic Systems, 2020, 17(5): 1729881420962263.

[21] Tang Gang, Tang Congqiang, Christophe Claramunt, et al. Geometric A-star Algorithm: An Improved A-star Algorithm for AGV Path Planning in a Port Environment [J]. IEEE Access, 2021, 9: 59196-59210.

Published by Journal of System Simulation, 2024