控制与决策
Control and Decision

# 基于蚁群信息素辅助的Q学习路径规划算法

田晓航, 霍鑫, 周典乐, 赵辉

引用本文:
田晓航,霍鑫,周典乐,赵辉. 基于蚁群信息素辅助的Q学习路径规划算法[J]. 控制与决策, 2023, 38(12): 3345−3353.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0476

---

# 基于蚁群信息素辅助的$Q$学习路径规划算法

田晓航[1], 霍 鑫[1†], 周典乐[2], 赵 辉[1]

(1. 哈尔滨工业大学 控制与仿真中心，哈尔滨 150080；2. 国防科技大学 前沿交叉学科学院，长沙 410073)

**摘 要:** 当$Q$学习应用于路径规划问题时,由于动作选择的随机性,以及$Q$表更新幅度的有限性,智能体会反复探索次优状态和路径,导致算法收敛速度减缓. 针对该问题,引入蚁群算法的信息素机制,提出一种寻优范围优化方法,减少智能体的无效探索次数. 此外,为提升算法初期迭代的目的性,结合当前栅格与终点位置关系的特点以及智能体动作选择的特性,设计$Q$表的初始化方法;为使算法在运行的前中后期有合适的探索概率,结合信息素浓度,设计动态调整探索因子的方法. 最后,在不同规格不同特点的多种环境中,通过仿真实验验证所提出算法的有效性和可行性.

**关键词:** $Q$学习；路径规划；$Q$表初始化；探索概率；蚁群算法；信息素

**中图分类号:** TP273 **文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2022.0476

**引用格式:** 田晓航,霍鑫,周典乐,等. 基于蚁群信息素辅助的$Q$学习路径规划算法[J]. 控制与决策, 2023, 38(12): 3345-3353.

# Ant colony pheromone aided $Q$-learning path planning algorithm

*TIAN Xiao-hang*[1], *HUO Xin*[1†], *ZHOU Dian-le*[2], *ZHAO Hui*[1]

(1. Control and Simulation Center，Harbin Institute of Technology，Harbin 150080，China；2. College of Advanced Interdisciplinary Studies，National University of Defense Technology，Changsha 410073，China)

**Abstract:** When $Q$-learning is applied to the path planning problem, due to the randomness of action selection and the limited update range of the $Q$ table, the agent will repeatedly explore sub-optimal states and paths, resulting in slower algorithm convergence. To address this problem, this paper introduces an ant colony pheromone aided $Q$-learning path planning algorithm, an optimization method for the optimization range is proposed to reduce the invalid exploration times of the agent. In addition, in order to improve the purpose of the initial iteration of the algorithm, according to the characteristics of the relationship between the current grid and the end point and the selection of the agent's action, an initialization method of the $Q$ table is designed. In order to make the algorithm have suitable exploration probability in the early, middle and late stages of operation, a method of dynamically adjusting the exploration factor is designed in combination with the concentration of pheromone. Finally, in a variety of environments with different specifications and different characteristics, the effectiveness and feasibility of the proposed algorithm are verified by simulation experiments.

**Keywords:** $Q$-learning; path planning; $Q$-table initialization; exploring probabilities; ant colony algorithm; pheromone

## 0 引 言

路径规划问题是机器人领域研究的热点,路径规划算法可应用于无人车(UGV)、 无人机(UAV)、水下移动机器人(AUV)等领域,具有很大的发展前景[1-4]. 路径规划算法可分为4类:传统路径规划算法、图形学方法、智能仿生学算法和其他算法. 蚁群算法和强化学习算法是当下热门的路径规划问题解决方案.

$Q$学习算法是路径规划问题中较为常用的强化学习方法[5-6]. 普通的$Q$学习算法在迭代过程中容易出现智能体反复探索次优路径,陷入局部最优,导致算法收敛速度缓慢的情况. 文献[7-8]利用人工势场法初始化$Q$表,为智能体提供了先验知识,大幅减少了智能体前期的随机探索,但这种方法在算法迭代后期并不能起到作用,在某些特定的地形中会失效,适用范围不广. 文献[9-10]基于智能体与目标的相对距

**收稿日期:** 2022-03-25；**录用日期:** 2022-07-17.

**基金项目:** 黑龙江省自然科学基金项目(LH2021F025);中央高校基本科研业务费专项资金项目(HIT.NSRIF202242);黑龙江省教改项目(SJGY20200185)；哈尔滨工业大学研究生教改核心项目(21HX0401).

**责任编委:** 易建强.

†通讯作者. E-mail: huoxin@hit.edu.cn.

离提出了两种启发式搜索策略,用以在算法迭代过程中指引智能体向目标靠近,但是启发函数与地形的相关度很高,不同的地形要设计不同的启发函数. 文献[11]提出了一种基于拉普拉斯映射的启发式$Q$学习,解决了欧氏距离作为启发函数在特定地形中失效的问题,但该方法只能用于状态空间的连接关系为无向图的任务,适用范围较小. 文献[12]通过锁定$Q$表中符合条件的元素以减少对$Q$表的更新次数,虽然该方法有效减少了算法的迭代次数,但不能保证被锁定$Q$表元素的精度,如果将该方法用于8方向路径规划问题中,算法不能保证收敛到最优路径.

蚁群算法是一种仿生学算法,它通过模拟蚂蚁觅食的过程,寻找由起点到终点符合条件约束的一条路径[13]. 针对蚁群算法易出现的局部最优和在复杂环境中收敛性差的问题,文献[14]提出了一种改进势场蚁群算法,同时改进了信息素更新策略,将信息素分为局部和全局两个更新阶段,并对一次迭代中的最优和最差路径设计了不同的信息素迭代策略;文献[15]通过改进的烟花算法快速生成一条参考路径,并以此来初始化信息素分布;文献[16]将蚁群划分为引导层蚁群和普通层蚁群,通过蚂蚁的分层分工提高了算法的路径规划能力. 关于蚁群算法的改进方法非常多,但是蚁群算法本身存在的容易陷入局部最优和死锁这两个问题并没有因为这些改进方法而得到完全的解决. $Q$学习算法不存在死锁问题,局部最优问题相比蚁群算法较轻,故本文选取$Q$学习作为研究对象.

针对$Q$学习在迭代过程中出现的智能体反复探索次优路径的问题,本文基于蚁群算法的信息素机制提出一种寻优范围优化方法,利用信息素浓度反映状态价值的特点,从信息素的浓度以及$Q$表记录的行为值两个角度去综合衡量状态的价值,一定程度上能减少智能体反复探索次优路径的情况,提高算法的迭代效率. 此外,针对$Q$表初始化方法和利用探索策略两方面提出分段线性的$Q$表初始化方法和基于轮盘赌的动作选择策略,结合有效栅格数量设计探索概率动态调整策略. 仿真实验表明,基于蚁群算法信息素机制的寻优范围优化能够进一步加速$Q$学习的迭代速度,加快算法的收敛. 最后,在多种环境下对比文献[7]有明显的优势,验证了本文算法的优越性.

## 1 环境建模及路径约束

本文采用栅格法对环境进行建模. 环境中分布着静态障碍物,如果栅格位置对应的实际环境中存在障碍物,则不管该栅格是否被障碍物完全覆盖,该栅格都将被标记为“1”,在仿真图中会以黑色栅格展现出来;完全没有障碍物的栅格则会被标记为“0”,在仿真图中会以白色栅格展现出来,如图1所示.

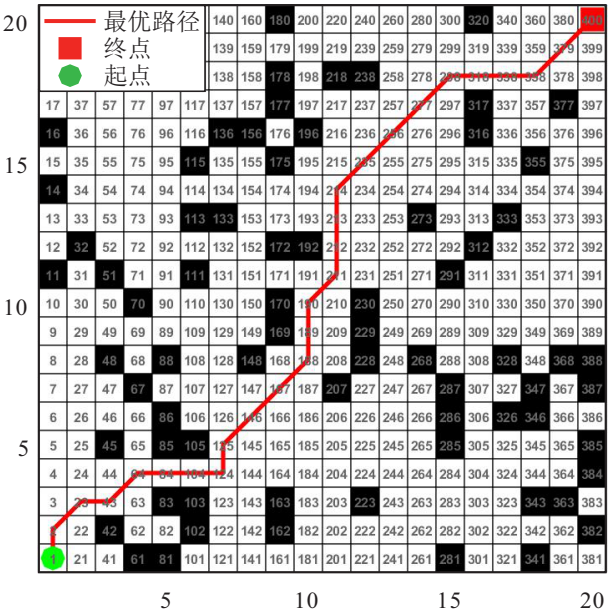


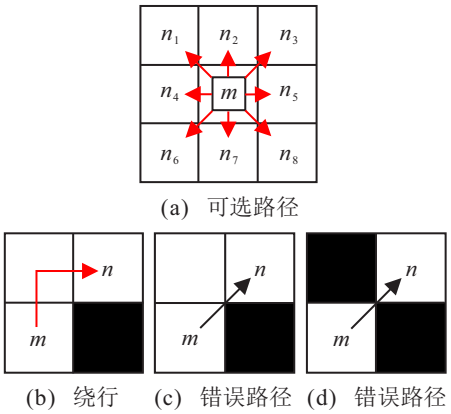

**图1 栅格编号**

以图1为例说明栅格编号与横纵坐标的关系. 利用直角坐标法对栅格进行编号,栅格横向序号为栅格横坐标,栅格纵向序号为栅格纵坐标. 图1中,栅格的横向序号最大值$r$为20,纵向序号最大值$c$为20,所以共有400个栅格,每个栅格按照从下到上,再从左到右依次编号为$1,2,\dots,r\times c$. 栅格编号与栅格的横纵坐标关系如下所示:

$$\begin{cases} x_m = \text{ceil}(m/r). \\ y_m = \begin{cases} \text{mod}(m,r),\ \text{mod}(m,r)\neq 0; \\ r,\ \text{mod}(m,r)=0. \end{cases} \end{cases} \tag{1}$$

其中:$x_m$表示栅格横坐标,$y_m$表示栅格纵坐标,$\text{ceil}(x)$表示取不小于$x$的整数运算,$\text{mod}(x,y)$表示求取$x/y$的余数,$m$表示栅格当前编号.

图1是一张栅格环境建模示范图,图中还展示了符合本文路径约束条件的最优路径. 本文的路径规划算法是一种避障式规划算法,规划得到的路径不能穿过障碍物,也不与障碍物发生任何擦碰. 关于擦碰的描述如图2所示,智能体从栅格$m$移动到栅格$n$的过程中,图2(a)中的智能体周围不存在障碍物,所以智能体能够选择周围8个栅格中的任意一个作为下一步所要到达的位置;图2(b)中智能体的右下侧存在障碍物,智能体要到达栅格$n$只能选择绕行,而不能以图2(c)的方式擦过障碍物;图2(d)中的智能体穿过了两个斜向分布的障碍物,这种方式也是不可行的.




(a) 可选路径　(b) 绕行　(c) 错误路径　(d) 错误路径

**图 2　路径约束**

## 2　数学基础

### 2.1　基本$Q$学习算法

$Q$学习可分为3个部分: $Q$表的初始化、动作选择策略和$Q$表更新. 基本$Q$学习在$Q$表初始化环节将$Q$表初始化为一个常值(通常是0), 动作选择策略通常选择$\varepsilon$-贪心策略, 在算法迭代期间探索概率始终保持固定. $Q$学习的行为值函数更新公式如下:

$$Q(s,a)=Q(s,a)+\alpha\cdot(r(s,a)+\gamma\cdot\max_{a'}Q(s',a')-Q(s,a)). \tag{2}$$

其中: $s$为当前智能体的状态, $s'$为下一个状态; $a$为智能体选择的动作(本文的智能体有8个动作可选: 上、下、左、右、左上、右上、左下、右下, 分别对应着$a=1,2,\ldots,8$), $a'$为下一个动作; $\alpha$为学习率; $r$为奖励函数; $\gamma$为衰减系数.

除了基本的3个部分之外, 在路径规划问题中还需要定义其回报函数和碰撞处理: 智能体每走一步都会受到惩罚, 惩罚大小是上一步所经过路径的长度; 当智能体碰到障碍物后, 智能体会停留在原地, 然后寻找动作避开障碍物, 直到达到终点为止. 回报函数定义如下:

$$r(s,a)=-\sqrt{(x_s-x_{s''})^2+(y_s-y_{s''})^2}. \tag{3}$$

其中: $(x_s,y_s)$是当前状态对应栅格的坐标, $s''$是上一个状态, $(x_{s''},y_{s''})$是上一个状态对应栅格的坐标.

### 2.2　蚁群算法

当蚂蚁在寻找最优路径时, 蚂蚁会在其走过的路径上留下信息素, 蚂蚁根据下一条路径上信息素浓度的大小选择动作. 在$t$时刻, 蚂蚁$k$由第$i$个节点选择下一个节点$j$是根据当前信息素$\tau_{ij}(t)$和启发信息$\eta_{ij}(t)$的大小判定的, 蚂蚁有$q_0(q_0\in[0,1])$的概率选择使$[\tau_{ij}(t)]^{\varsigma}[\eta_{ij}(t)]^{\beta}$最大的下一个节点$j$, 有$1-q_0$的概率按照轮盘赌的方式选择下一个节点.

$$P_{ij}^k(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\varsigma}[\eta_{ij}(t)]^{\beta}}{\sum\limits_{l\in\text{allowed}_k}[\tau_{il}(t)]^{\varsigma}[\eta_{il}(t)]^{\beta}}, & j\in\text{allowed}_k;\\ 0, & \text{otherwise.}\end{cases} \tag{4}$$

$$\eta_{ij}(t)=\frac{1}{d_{jg}}. \tag{5}$$

其中: $\varsigma$是信息素启发因子; $\text{allowed}_k$是第$k$只蚂蚁当前可以到达的下一节点的集合; $\beta$是期望启发式因子; $d_{jg}$是节点$j$到目标节点$g$的欧氏距离.

基本蚁群算法的信息素更新策略是: 在本次循环中所有蚂蚁完成自己的迭代过程之后, 进行一次全局信息素更新, 多次迭代后, 信息素的分布越来越靠近最优路径. 全局信息素更新公式为

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\Delta\tau_{ij}(t). \tag{6}$$

$$\Delta\tau_{ij}(t)=\sum_{k=1}^{M}\Delta\tau_{ij}^k(t). \tag{7}$$

$$\Delta\tau_{ij}^k(t)=\begin{cases}\dfrac{Q}{L_k}, & \text{蚂蚁}k\text{经过的路径}(i,j);\\ 0, & \text{otherwise.}\end{cases} \tag{8}$$

其中: $\rho$是信息素蒸发系数, 且$\rho\in[0,1]$; $\Delta\tau_{ij}(t)$是$t$时刻所有蚂蚁经过一次迭代后留下的信息素量; $M$是蚂蚁总数; $\Delta\tau_{ij}^k(t)$是$t$时刻第$k$只蚂蚁在迭代过程中留下的信息素量; $Q$是信息素强度; $L_k$是第$k$只蚂蚁在本轮迭代中遍历完成后的路径总长.

## 3　基于蚁群信息素机制的$Q$学习算法

### 3.1　基于蚁群信息素机制的寻优范围优化

在$Q$学习的迭代过程中, 由于动作选择具有一定的随机性以及$Q$表元素更新幅度的有限性, 智能体会反复探索次优状态, 导致算法的效率大大降低, 蚁群算法的信息素机制可以有效解决这一问题.

#### 3.1.1　$Q$学习智能体的信息素机制

$Q$表中存储了智能体在特定栅格选择特定动作期望获得的回报, 存储所有这样的信息需要$r\times c\times A$个存储单元($A$是所有动作的数量, $A=8$), 也就是说, 在图1所示的栅格环境下, $Q$表存储了3 200个数据. 图3(a)展示了$Q$表的结构, 以位于底部的$Q(1,1),Q(1,2),\ldots,Q(1,8)$为例, 这8个单元存储着智能体在编号为1的栅格中, 当$a=1,2,\ldots,8$时的期望回报. 按照$Q$表的结构生成一个信息素表, 如图3(b)所示, 在迭代过程中让$Q$学习的智能体在信息素表中留下信息素(如果智能体位于编号为1的栅格上并且选择了动作2, 那么智能体便会在$t(1,2)$这个存储单元中留下一定浓度的信息素). 将$Q$学习的迭代次数按种群划分: 每个种群有$M$个智能体, 该种群共

迭代$K$次,智能体共迭代$M \times K$次.

栅格编号$(r\times c)$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $Q(3,1)$ | $Q(3,2)$ | $Q(3,3)$ | $Q(3,4)$ | $Q(3,5)$ | $Q(3,6)$ | $Q(3,7)$ | $Q(3,8)$ |
| $Q(2,1)$ | $Q(2,2)$ | $Q(2,3)$ | $Q(2,4)$ | $Q(2,5)$ | $Q(2,6)$ | $Q(2,7)$ | $Q(2,8)$ |
| $Q(1,1)$ | $Q(1,2)$ | $Q(1,3)$ | $Q(1,4)$ | $Q(1,5)$ | $Q(1,6)$ | $Q(1,7)$ | $Q(1,8)$ |

(a) $Q$表的结构

栅格编号$(r\times c)$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $t(3,1)$ | $t(3,2)$ | $t(3,3)$ | $t(3,4)$ | $t(3,5)$ | $t(3,6)$ | $t(3,7)$ | $t(3,8)$ |
| $t(2,1)$ | $t(2,2)$ | $t(2,3)$ | $t(2,4)$ | $t(2,5)$ | $t(2,6)$ | $t(2,7)$ | $t(2,8)$ |
| $t(1,1)$ | $t(1,2)$ | $t(1,3)$ | $t(1,4)$ | $t(1,5)$ | $t(1,6)$ | $t(1,7)$ | $t(1,8)$ |

(b) 信息素表的结构

**图3 $Q$表和信息素表的结构**

种群中的蚂蚁在迭代过程中会在路径上留下一定量的信息素,对蚂蚁留下的信息素量进行调整,即对式(8)进行调整,有

$$\Delta\tau_{ij}^{k}(t)=\begin{cases}\tau_1,\ \text{蚂蚁}k\text{经过的路径}(i,j);\\ 0,\ \text{otherwise}.\end{cases}\tag{9}$$

其中$\tau_1$是蚂蚁在所经过路径上留下的信息素量.

在种群迭代一次之后,对本次迭代过程中最优路径上的信息素进行加强,该过程称为最优路径信息素强化,强化公式如下:

$$\tau_{ij}(t)=\tau_{ij}(t)+\tau_2,\tag{10}$$

其中$\tau_2$是强化信息素量.

在所有蚂蚁完成了一次迭代之后,按照式(6)进行全局信息素更新.

### 3.1.2 寻优范围优化

蚁群算法中的蚂蚁会在经过的路径上留下信息素,信息素浓度体现了该路径的价值,在蚂蚁不断迭代的过程中,信息素在栅格地图中的分布范围由小变大,当达到一定的探索度之后,由于蒸发作用,信息素分布范围会由大变小,最终留下一条或者少数几条信息素浓度大的路径,即可找到最优路径. 让$Q$学习的智能体在信息素表中留下信息素,信息素分布也会出现与蚁群算法相同的现象.

为了更直观地展示这一现象,以图1所示环境为例,让基本$Q$学习的智能体携带信息素. 将信息素表中处于同一行的8个单元所存储的8个信息素数据相加为一个数据,再将该数据按照式(1)映射到与栅格地图同样大小的$20\times 20$表格中,以颜色明暗程度体现该数据大小,如图4所示. 可以清晰地看到图4(a)中呈现高亮的栅格数量少于图4(b),图4(c)中高亮的栅格数量相比于图4(b)有所减少,图4(d)则进一步减少.

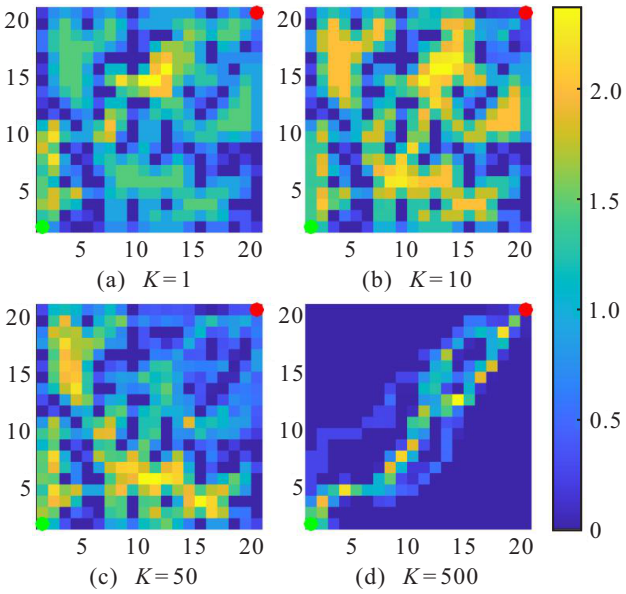
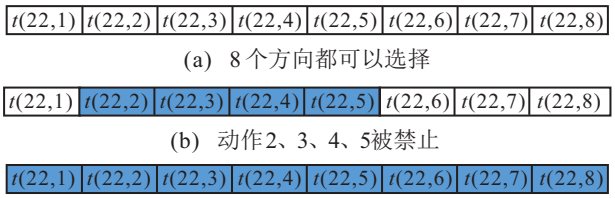



**图4 信息素分布变化图**

将信息素表中信息素浓度低于Kt的单元视为无效状态,大于(或等于)该阈值的单元视为有效状态. 有效状态数量由少变多的过程称为探索过程,有效状态数量由多变少的过程称为寻优过程. 在寻优过程中,部分单元连续数个种群都不会被智能体探索. 定义这样一条规则:连续St个寻优过程之后,标记所有的无效状态,禁止智能体对其进行探索. 比如:假设$t(2,3)$中信息素浓度低于Kt,并被标记为无效状态,那么当智能体处于编号为2的栅格中时,智能体将会被禁止选择动作3. 这样缩减探索范围,可以使$Q$学习保持一定全局寻优能力的同时减小由于智能体探索的随机性和$Q$表更新幅度的有限性带来的不必要的探索,加快迭代速度.

### 3.1.3 陷阱处理

在缩减寻优范围的过程中可能出现“陷阱”,在其中的智能体会被困住. 假设智能体在编号为22的栅格中可以选择8个方向到达它的任意相邻栅格,且8个存储22号栅格信息素浓度的信息素存储单元也没有被标记为无效状态,如图5(a)所示. 在后续的迭代过程中,信息素存储单元$t(22,2),t(22,3),t(22,4),t(22,5)$的信息素浓度低于Kt,并被标记为无效状态,当智能体再次到达第22号栅格时,智能体将不能选择动作2、3、4、5,如图5(b)所示. 在后续迭代过程中,智能体很少再探索第22号栅格,信息素表中关于22号栅格的8个存储单元都被标记为无效状态,如图5(c)所示. 此时在栅格表中,第22号栅格并不是障碍物,所以对于智能体来说,该栅格可达,但是一旦智能体到达第22号栅格,它不能选择任何动作逃离第22号栅格.




(a) 8 个方向都可以选择

(b) 动作2、3、4、5被禁止

(c) 8 个动作都被禁止

**图 5　陷阱的形成**

陷阱逃离策略: 智能体每到达一个状态都会检测其可执行动作的个数, 如果无动作可选, 则将该栅格在地图中标记为障碍物, 并将智能体放回起点.

### 3.2　$Q$表初始化

在基本$Q$学习算法中, $Q$表通常被初始化为栅格数量以内的某一数值, 这样的初始化方法没有充分利用已知条件. 栅格中起点和终点的位置是已知的, 一般情况下, 越靠近终点的状态, 其价值越大, $Q$表的真实值存在一定的规律. 文献[7]利用人工势场法初始化$Q$表, 极大提升了算法的迭代速度, 但文献[7]的初始化函数还有很大的改进空间, 本节在文献[7]的基础上设计一种$Q$表初始化方法.

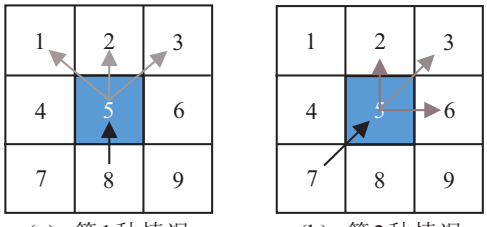
在没有障碍物的栅格地图中, 行为值与当前栅格到终点的距离大致成正比. 依靠这个规律可以设计一个初始化函数, 初始化公式如下:

$$Q(s,a) = -\varphi\sqrt{(x_s - x_g)^2 + (y_s - y_g)^2}. \tag{11}$$

其中: $(x_g, y_g)$是终点坐标, $\varphi$是系数.

本文设计的$Q$表初始化方法以当前栅格到终点栅格的有向线段为参考路径, 并以此对$Q$值进行大小设置. 智能体有8个寻优方向, 分别是前、后、左、右、左上、右上、左下、右下, 最优路径只用到其中的3个, 具体可分为两种情况.

第1种情况如图6(a)所示, 智能体位于灰色栅格, 栅格8是智能体上一步所在位置, 黑色箭头是上一步的最优路径, 那么下一条最优路径只可能是灰色箭头所示的3种情况. 原因是, 假设栅格7是下一个最优状态, 则黑色箭头所示最优路径不是最优路径, 与假设相悖; 假设栅格4是下一个最优状态, 则黑色箭头应该由栅格8指向栅格4, 与假设相悖; 由对称性可知栅格6和栅格9的情况同理. 图6(b)展示了第2种情况.



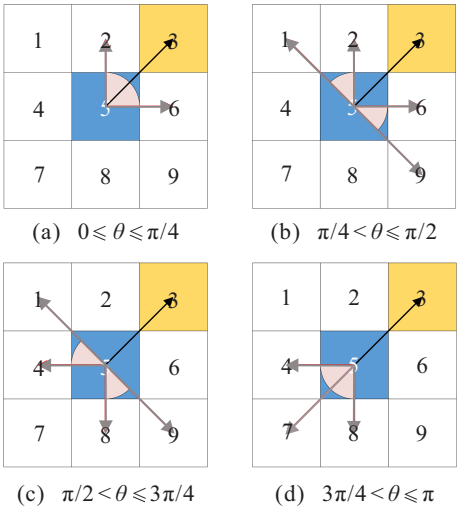

(a) 第1种情况　　(b) 第2种情况

**图 6　最优路径方向选择**

利用上述结论, 可以根据当前状态可选动作对应的下一段路径与参考路径的夹角, 对行为值初始化公式(11)给定不同大小的系数, 具体数值如下所示:

$$\varphi = \begin{cases} \varphi_1,\ 0 \leqslant \theta \leqslant \pi/4; \\ \varphi_2,\ \pi/4 < \theta \leqslant \pi/2; \\ \varphi_3,\ \pi/2 < \theta \leqslant 3\pi/4; \\ \varphi_4,\ 3\pi/4 < \theta \leqslant \pi. \end{cases} \tag{12}$$

其中: $\varphi_1$、$\varphi_2$、$\varphi_3$、$\varphi_4$是$\varphi$在不同情况下的具体取值, $\theta$是下一条路径与参考路径的夹角. 夹角示意图如图7所示, 图中由栅格5指向栅格3的黑色箭头是参考路径, 小扇形区域是角度范围.




(a) $0 \leqslant \theta \leqslant \pi/4$　　(b) $\pi/4 < \theta \leqslant \pi/2$

(c) $\pi/2 < \theta \leqslant 3\pi/4$　　(d) $3\pi/4 < \theta \leqslant \pi$

**图 7　与参考路径形成的夹角示意图**

### 3.3　动作选择策略

基本$Q$学习采用的$\varepsilon$-贪心策略有效地解决了贪婪算法容易陷入局部最优的缺陷, 使得算法一方面以大小为$1-\varepsilon$的概率选择当前$Q$值最大的动作, 另一方面又能以大小为$\varepsilon$的概率进行随机探索, 寻找全局最优. $\varepsilon$-贪心策略的公式描述如下:

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|A(s)|},\ a = \arg\max Q(s,a); \\ \dfrac{\varepsilon}{|A(s)|},\ a \neq \arg\max Q(s,a). \end{cases} \tag{13}$$

其中$A(s)$是智能体在状态$s$处所能采取的动作个数. 式(13)表示智能体在状态$s$下选择$Q$值最大动作的概率是$1-\varepsilon+\dfrac{\varepsilon}{|A(s)|}$, 选择其他动作的概率是$\dfrac{\varepsilon}{|A(s)|}$.

在智能体探索初期, 算法应保持较大的探索概率, 随着算法的运行, 探索概率可以适当降低. 结合本文的有效状态设计一种新的探索概率衰减策略: 当有效状态数量增多时, 算法通过探索获得的回报较大, 此时需要尽量保持当前的探索概率, 避免算法陷入局部最优; 当有效状态数量减小时, 算法已经大致寻找到最优路径, 正在局部范围内寻优, 此时应当减

小当前的探索概率. 基于有效栅格数量,提出一种动态调整探索概率的方法.

以本次规划中栅格地图所有的状态空间数为基准,连续St个寻优过程之后,依据有效状态衰减数量调整算法的探索因子,迭代公式如下:

$$\varepsilon = \varepsilon/(1+\mathrm{e}^{-\sigma\Delta s/S}). \tag{14}$$

其中:$\varepsilon$是探索概率,$\sigma$是衰减系数,$\Delta s$是有效状态衰减个数,$S$是状态总数量.

### 3.4 基于蚁群信息素机制的$Q$学习算法流程

本文算法步骤如图8所示.

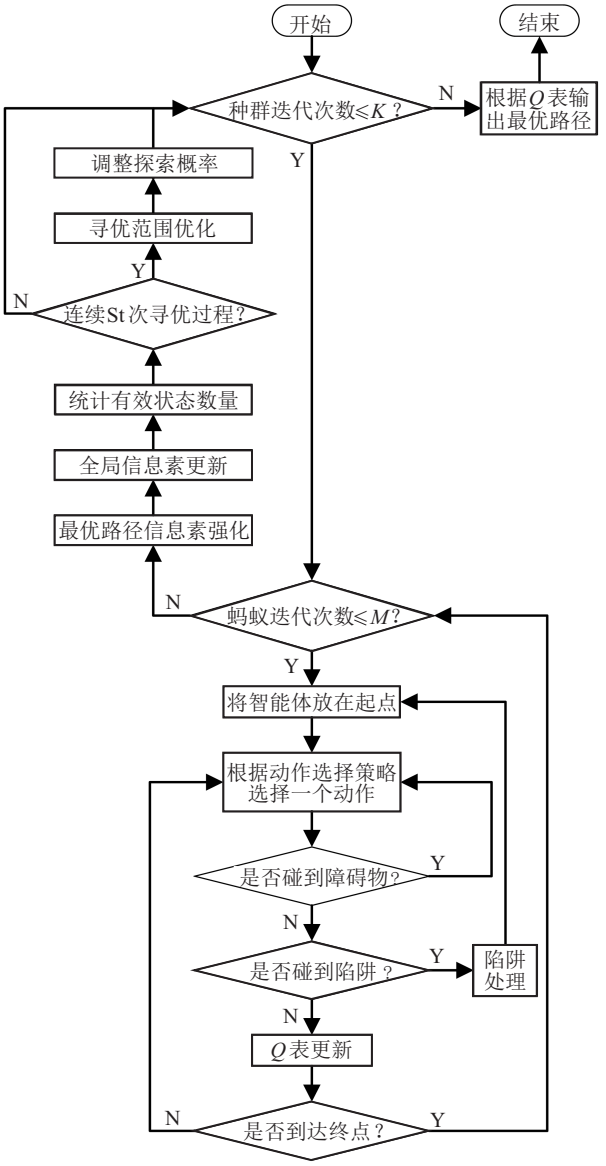



**图 8 本文算法流程**

step 1: 使用栅格法建立地图,确定起点和终点. 按照式(11)初始化$Q$表;初始化$Q$学习的各个参数,包括学习率$\alpha$、衰减系数$\gamma$、探索概率$\varepsilon$;初始化信息素相关参数,包括全局信息素蒸发系数$\rho$、蚂蚁的信息素残留量$\tau_1$、强化信息素量$\tau_2$、信息素阈值Kt、迭代阈值St.

step 2: 将智能体放置在起点.

step 3: 生成一个范围在0到1之间的随机数,根据该数与$\varepsilon$的大小关系选择智能体的动作以及下一步要到达的栅格.

step 4: 判断智能体是否陷入陷阱,如果是,则在完成陷阱处理后返回step 2,否则继续.

step 5: 更新$Q$表并判断是否同时满足到达终点及其需要的次数$M$,如果没到达终点,则返回step 3;如果到达了但是没满足次数要求,则返回step 2;如果到达了且满足次数要求,则继续.

step 6: 按式(9)使智能体在路径上留下信息素,种群迭代一次后按式(10)进行一次信息素强化,最后按式(6)进行全局信息素更新.

step 7: 统计有效状态数量并判断是否连续St次出现寻优过程,如果是,则进行寻优范围优化和探索概率调整,否则转入step 2让种群继续迭代,直到迭代次数达到$K$次.

step8: 根据$Q$表输出最优路径.

## 4 仿真实验

仿真实验设计两种栅格地图,分别为随机障碍物环境和特殊地形,在这两种地图中综合对比BAS-$Q$(Basic $Q$-learning)、文献[7]的算法、未开启信息素辅助(寻优范围优化)的IMP-$Q$(improved $Q$-learning)和开启了信息素辅助的PIMP-$Q$(pheromone aided IMP-$Q$),以验证算法的可行性和有效性(IMP-$Q$和PIMP-$Q$算法均使用本文设计的$Q$表初始化方法和动作选择策略,是否开启寻优范围优化是两者的唯一差别). 仿真实验运行环境:操作系统Windows10(64位),运行环境Matlab 2019b,处理器Intel (R) Core (TM) i7-10710U CPU 1.10 GHz. 本文方法中信息素相关的主要参数如表1所示,$Q$学习参数如表2所示.

**表 1 信息素相关参数具体数值**

| 信息素参数 | $\tau_1$ | $\tau_2$ | $\rho$ | St | Kt | $M$ |
|---|---|---|---|---|---|---|
| 具体数值 | 0.5 | 1 | 0.5 | 2 | 0.062 5 | 20 |

**表 2 $Q$学习相关参数具体数值**

| 算法 | $\varepsilon$ | $\alpha$ | $\gamma$ | $\varphi_1$ | $\varphi_2$ | $\varphi_3$ | $\varphi_4$ | $\sigma$ |
|---|---|---|---|---|---|---|---|---|
| BAS-$Q$ | 0.1 | 0.9 | 1 | \ | \ | \ | \ | \ |
| 文献[7] | 0.1 | 0.9 | 1 | \ | \ | \ | \ | \ |
| IMP-$Q$ | 0.1 | 0.9 | 1 | 1 | 1.1 | 1.3 | 1.4 | 1 000 |
| PIMP-$Q$ | 0.1 | 0.9 | 1 | 1 | 1.1 | 1.3 | 1.4 | 1 000 |

### 4.1 随机障碍物环境下的仿真实验

借助算法在Matlab中分别生成障碍物占比20%和30%(不考虑障碍物重叠)的两张40×40的栅格地图作为本节的仿真验证环境,设定算法迭代次数上限

为 20 000 次($K$ = 1 000). 每种算法执行 10 次, 在表格中统计其平均路径长度、平均迭代次数和平均寻优耗时, 取其中一次的迭代结果绘制迭代曲线对比图.

#### 4.1.1　障碍物占比 20 % 的 40×40 栅格地图

从图 9、图 10 和表 3 可以看到: 4 种算法找到的路径大体相同, 4 条路径在长度上相等, 都是最优解; IMP-$Q$ 和 PIMP-$Q$ 两种算法前期的曲线能够迅速地逼近最优解, 大幅优于 BAS-$Q$ 和文献 [7] 算法. 这是因为 IMP-$Q$ 和 PIMP-$Q$ 应用了本文的 $Q$ 表初始化方法, 大幅减少了智能体前期的探索次数和探索时间, 而采用了信息素辅助的 PIMP-$Q$ 算法能在迭代过程中逐步淘汰掉信息素浓度偏低状态, 减少智能体探索的范围, 其迭代曲线能够以非常快的速度逼近最优解, 并且在迭代后期保持平稳, 不发生波动. 从迭代次数来看, PIMP-$Q$ 的迭代次数只有 BAS-$Q$ 的 39 %, 文献 [7] 的 41 %, IMP-$Q$ 的 70 %, 在时间消耗上相比于没有使用信息素辅助的 IMP-$Q$ 也减少了 42 %.

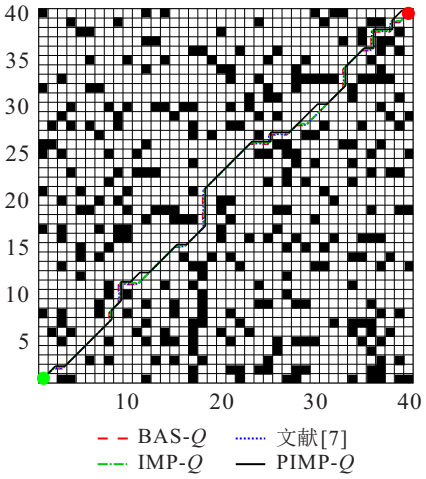



图 9　20 %随机障碍物环境中 4 种算法寻得的路径

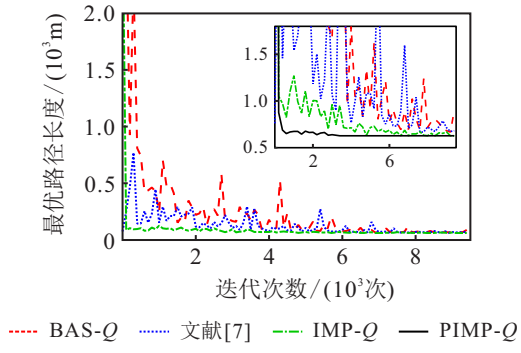



图 10　20 %随机障碍物环境中算法的迭代曲线对比

表 3　20 %随机障碍物环境中的算法性能对比

| 算法 | 平均路径长度/m | 平均迭代次数/次 | 平均寻优耗时/s |
|---|---|---|---|
| BAS-$Q$ | 62.769 6 | 7 322.1 | 36.340 3 |
| 文献 [7] | 62.769 6 | 6 923.8 | 11.901 6 |
| IMP-$Q$ | 62.769 6 | 4 068.8 | 5.019 5 |
| PIMP-$Q$ | 62.769 6 | **2 885.8** | **2.888 2** |

#### 4.1.2　障碍物占比 30 % 的 40×40 栅格地图

从图 11、图 12 和表 4 可以看到: 4 种算法的路径大体相同. 采用了本文信息素初始化策略的 IMP-$Q$ 算法, 其迭代曲线在前期相较于文献 [7] 有一定程度的优势. 从局部放大图可以看到, 采用了信息素辅助的 PIMP-$Q$ 其迭代曲线能在 IMP-$Q$ 的基础上更迅速地贴近最优解, 并且在约第 1 000 次迭代之后就保持了一定的平稳度; 算法平均迭代次数仅有 3 580.1 次, 迭代耗时也有较大幅度的减少.

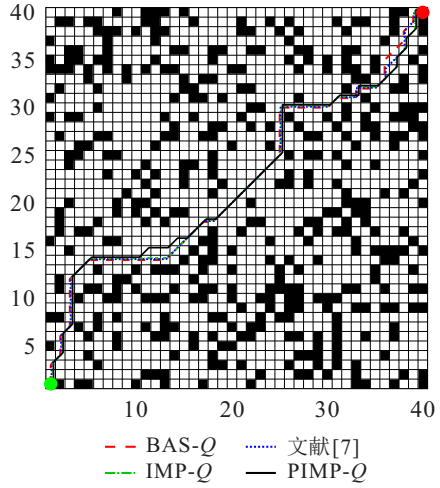



图 11　30 %随机障碍物环境中 4 种算法寻得的路径

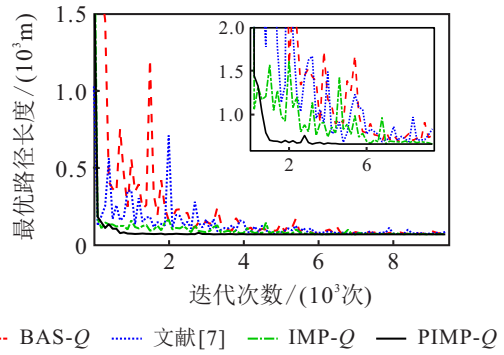



图 12　30 %随机障碍物环境中算法的迭代曲线对比

表 4　30 %随机障碍物环境中的算法性能对比

| 算法 | 平均路径长度/m | 平均迭代次数/次 | 平均寻优耗时/s |
|---|---|---|---|
| BAS-$Q$ | 66.284 3 | 7 494.6 | 23.346 5 |
| 文献 [7] | 66.284 3 | 7 066.3 | 11.629 8 |
| IMP-$Q$ | 66.284 3 | 5 792.5 | 7.016 6 |
| PIMP-$Q$ | 66.284 3 | **3 580.1** | **3.518 8** |

### 4.2　特殊地形下的仿真实验

本节参考文献 [14] 中的两张地图, 第 1 个 30 × 30 特殊地形是文献 [14] 中的原地图, 相较于随机障碍物, 该地图的障碍物体积更大, 且有一定数量的凹形障碍物, 这会给算法造成额外的负担; 第 2 个 30 × 30 回廊地图是文献 [14] 中 10 × 10 回廊地图的放大版, 这种回廊形地图会使人工势场法初始化 $Q$ 表在一定

程度上失效,使其不能像在随机障碍物环境中一样大幅提升算法性能,此外,更大的地图意味着更多的探索空间,这也会给算法带来负担.

#### 4.2.1 30×30特殊地形

从图13和图14可以看出:4种算法得到的路径大体相同;与在随机障碍物中类似,PIMP-$Q$算法的迭代曲线能迅速地逼近最优解并且在迭代后期保持平稳. 从表5可以看到,IMP-$Q$的迭代次数并没有像在随机障碍物环境中一样相较于文献[7]有大幅提升,但是算法的迭代时间仍然有所减少;PIMP-$Q$相较于IMP-$Q$,其迭代次数减半,迭代时间也减半,信息素辅助的效果明显.

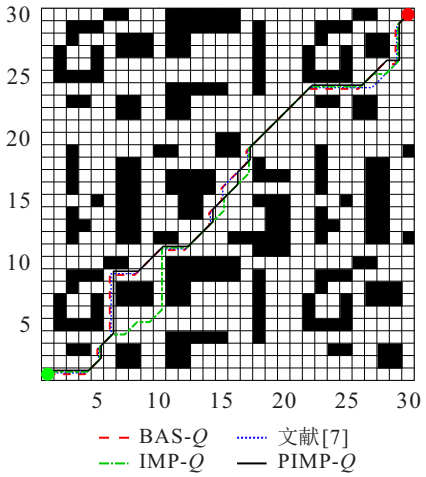



图13 30×30特殊地形中4种算法寻得的路径

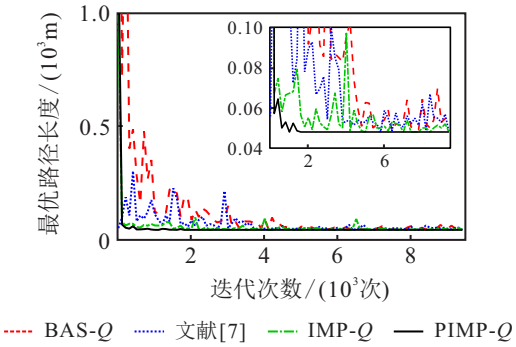



图14 30×30特殊地形中算法的迭代曲线对比

表5 30×30特殊地形中的算法性能对比

| 算法 | 平均路径长度/m | 平均迭代次数/次 | 平均寻优耗时/s |
|---|---|---|---|
| BAS-$Q$ | 48.041 6 | 4 841.2 | 11.566 9 |
| 文献[7] | 48.041 6 | 4 045.2 | 5.020 9 |
| IMP-$Q$ | 48.041 6 | 4 017.1 | 3.382 1 |
| PIMP-$Q$ | 48.041 6 | **1 808.7** | **1.561 4** |

#### 4.2.2 30×30回廊地形

回廊地形对算法造成了更大的阻碍. 从图15、图16和表6可以看到:虽然4种算法都迭代到了最优解,但文献[7]的算法在迭代时间上与BAS-$Q$几乎相同,且迭代次数甚至不降反升;同时可以看到,在这种回廊地形中,PIMP-$Q$仍然能够快速地逼近最优解并在迭代后期保持平稳,且减少智能体的迭代次数,使算法更快地找到最优路径.

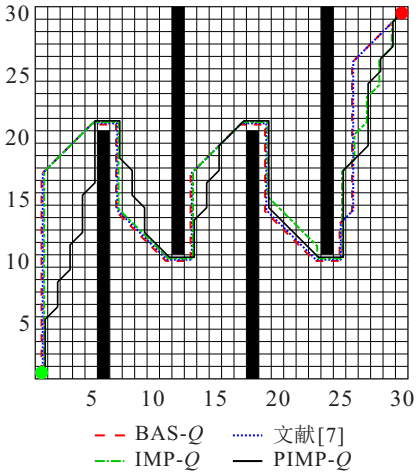



图15 30×30回廊地形中4种算法寻得的路径

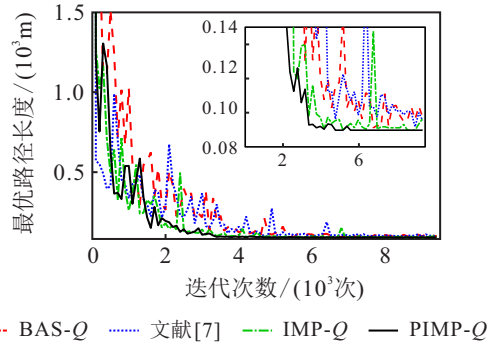



图16 30×30回廊地形中算法的迭代曲线对比

表6 30×30特殊地形中的算法性能对比

| 算法 | 平均路径长度/m | 平均迭代次数/次 | 平均寻优耗时/s |
|---|---|---|---|
| BAS-$Q$ | 89.698 5 | 7 766.8 | 34.634 8 |
| 文献[7] | 89.698 5 | 9 495.2 | 33.239 5 |
| IMP-$Q$ | 89.698 5 | 5 794.4 | 21.484 3 |
| PIMP-$Q$ | 89.698 5 | **4 942.3** | **19.262 3** |

## 5 结 论

本文基于蚁群算法的信息素机制为$Q$学习设计了一种寻优范围优化方法. 该方法适用于多种复杂环境,能在$Q$学习迭代过程中发挥作用,持续地优化智能体的探索范围,减少由于动作选择策略的随机性和$Q$表元素更新幅度的有限性导致的多余的探索次数,加快算法的运行速度. 本文利用路径规划问题的特点设计了$Q$学习的$Q$表初始化方法和动作选择策略;对于寻优范围优化造成的陷阱问题,文章也提供了相应的解决方法. 在仿真实验环节,在多个不同规模、不同特点的仿真环境中验证了方法的有效性和可行性. 本文的方法能有效减少算法的迭代次数

和时间,但不能保证寻得最优解,后续将研究本文方法的收敛性问题,使其兼顾解的最优性和运行的快速性.

## 参考文献(References)


[1] Josef S, Degani A. Deep reinforcement learning for safe local planning of a ground vehicle in unknown rough terrain[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6748-6755.

[2] Viswanathan S, Ravichandran K S, Tapas A M, et al. An intelligent gain based ant colony optimisation method for path planning of unmanned ground vehicles[J]. Defence Science Journal, 2019, 69(2): 167-172.

[3] 王思鹏, 杜昌平, 郑耀. 基于强化学习的扑翼飞行器路径规划算法[J]. 控制与决策, 2022, 37(4): 851-860.
(Wang S P, Du C P, Zheng Y. Local planner for flapping wing micro aerial vehicle based on deep reinforcement learning[J]. Control and Decision, 2022, 37(4): 851-860.)

[4] Ma Y N, Gong Y J, Xiao C F, et al. Path planning for autonomous underwater vehicles: An ant colony algorithm incorporating alarm pheromone[J]. IEEE Transactions on Vehicular Technology, 2019, 68(1): 141-154.

[5] Watkins C J C H, Dayan P. $Q$-learning[J]. Machine Learning, 1992, 8(3/4): 279-292.

[6] Tan B, Peng Y Y, Lin J G. A local path planning method based on $Q$-learning[C]. International Conference on Signal Processing and Machine Learning (CONF-SPML). Stanford, 2021: 80-84.

[7] 宋勇, 李贻斌, 李彩虹. 移动机器人路径规划强化学习的初始化[J]. 控制理论与应用, 2012, 29(12): 1623-1628.
(Song Y, Li Y B, Li C H. Initialization in reinforcement learning for mobile robots path planning[J]. Control Theory & Applications, 2012, 29(12): 1623-1628.)

[8] 徐晓苏, 袁杰. 基于改进强化学习的移动机器人路径规划方法[J]. 中国惯性技术学报, 2019, 27(3): 314-320.
(Xu X S, Yuan J. Path planning for mobile robot based on improved reinforcement learning algorithm[J]. Journal of Chinese Inertial Technology, 2019, 27(3): 314-320.)

[9] Yan C, Xiang X J. A path planning algorithm for UAV based on improved $Q$-learning[C]. The 2nd International Conference on Robotics and Automation Sciences (ICRAS). Wuhan, 2018: 1-5.

[10] Pei M, An H, Liu B, et al. An improved dyna-$Q$ algorithm for mobile robot path planning in unknown dynamic environment[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4415-4425.

[11] 朱美强, 李明, 程玉虎, 等. 基于拉普拉斯特征映射的启发式$Q$学习[J]. 控制与决策, 2014, 29(3): 425-430.
(Zhu M $Q$, Li M, Cheng Y H, et al. Heuristically accelerated $Q$-learning algorithm based on Laplacian Eigenmap[J]. Control and Decision, 2014, 29(3): 425-430.)

[12] Konar A, Goswami C I, Singh S J, 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.

[13] Dorigo M, Gambardella L M. Ant colony system: A cooperative learning approach to the traveling salesman problem[J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 53-66.

[14] 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.
(Wang X Y, Yang L, Zhang Y, et al. Robot path planning based on improved ant colony algorithm with potential field heuristic[J]. Control and Decision, 2018, 33(10): 1775-1781.)

[15] 张玮, 马焱, 赵捍东, 等. 基于改进烟花-蚁群混合算法的智能移动体避障路径规划[J]. 控制与决策, 2019, 34(2): 335-343.
(Zhang W, Ma Y, Zhao H D, et al. Obstacle avoidance path planning of intelligent mobile based on improved fireworks-ant colony hybrid algorithm[J]. Control and Decision, 2019, 34(2): 335-343.)

[16] 张恒, 何丽, 袁亮, 等. 基于改进双层蚁群算法的移动机器人路径规划[J]. 控制与决策, 2022, 37(2): 303-313.
(Zhang H, He L, Yuan L, et al. Mobile robot path planning using improved double-layer ant colony algorithm[J]. Control and Decision, 2022, 37(2): 303-313.)


## 作者简介


田晓航 (1998−), 男, 硕士生, 从事移动机器人路径规划的研究, E-mail: 1152895287@qq.com;

霍鑫 (1981−), 男, 教授, 博士, 从事模式识别与无人系统等研究, E-mail: huoxin@hit.edu.cn;

周典乐 (1983−), 男, 讲师, 博士, 从事军事智能与无人系统等研究, E-mail: laffiche@163.com;

赵辉 (1971−), 男, 教授, 博士, 从事电力电子与电机驱动等研究, E-mail: zhaohui@hit.edu.cn.