DOI:10.15923/j.cnki.cn22-1382/t.2019.1.10

基于改进 SARSA(λ)移动机器人路径规划

宋　宇，　王志明*

（长春工业大学 计算机科学与工程学院，吉林 长春　130012）

摘　要：在机器人选择下一点坐标时，分别计算周围格子到达概率以及所受合力。记录机器人每轮到达终点所经路径总距离，将全局最优距离值与机器人到达终点所得奖励值相关，并进行加权更新 Q 值。仿真结果表明，采用该算法机器人到达目标点用时减少了 85%，路径总长度平均缩短 22%。

关键词：路径规划；强化学习；SARSA(λ)；人工势场法

中图分类号：TP 301.6　　**文献标志码：**A　　**文章编号：**1674-1374(2019)01-0055-05

Path planning based on improved SARSA (λ)

SONG Yu,　WANG Zhiming*

(School of Computer Science & Engineering, Changchun University of Technology, Changchun 130012, China)

Abstract: When robot comes to choose next point coordinates, we calculate the arrival probability and resultant force of the surrounding lattice, record robot's total path distance for each episode. The global optimal distance is related with the reward to update the weighted Q value. Simulation results show that with the algorithm, the time robot arrives the goal is reduced by 85% but the total path by 22%.

Key words: path planning; reinforcement learning; SARSA (λ); artificial potential field method.

0　引　言

由于具有广泛的应用场景，如机械臂运动规划、机器人运动规划等，近年来，路径规划得到国内外学者的关注，路径规划的相关算法被不断提出。从蚁群算法[1]、A* 算法、D* 算法、RRT 算法、PRM 算法、人工势场法[2]、优化算法[3]到模糊逻辑算法、神经网络算法、强化学习算法[4]。其中人工势场法在数学描述上简洁、美观，且规划出的路径较安全、平滑，但存在目标点不可达、障碍物附近抖动的问题。针对人工势场法的不足，国内外许多学者提出了改进方法，如文献[5]提出了扇区划分后增加虚拟障碍物的方法，文献[6]提出了预规划路径后增加虚拟质点的方法，文献[7]采用高斯组合隶属函数建立引力点函数，消除了极小值问题。近年来，基于强化学习算法[8-9]已经被初

收稿日期：2018-09-17
基金项目：吉林省青年科研基金资助项目(20160520020JH)
作者简介：宋　宇(1969－)，男，汉族，黑龙江呼兰人，长春工业大学教授，硕士，主要从事嵌入式系统及应用方向研究，E-mail:songyu@ccut.edu.cn. * 通讯作者：王志明(1991－)，男，汉族，山西神池人，长春工业大学硕士研究生，主要从事路径规划方向研究，E-mail:120354157@qq.com.

步用于路径规划问题中。SARSA(λ)算法是一种基于值函数的强化学习算法，如果直接将 SARSA(λ)应用于路径规划，会使初次探索随机性和撞墙概率较大，学习时间较长。

文中通过人工势场法的合力引导机制减少了路径搜索时间，首次探索时选择势场合力方向最大的概率最大，再次探索时，Q 值最大方向所占的比重增大。仿真实验表明，改进算法改善了人工势场法易陷入局部极值的现象。

1 Khatib 人工势场法

Khatib 人工势场法是通过计算合力机器人在一个虚拟势场环境中受到的合力来决定机器人下一步的方向，目标点在环境中任一点 x 产生的吸引力势场值为：

$$U_{x_d}(x)=\frac{1}{2}k_p(x-x_d)^2 \tag{1}$$

式中：x_d ——目标点坐标；

k_p ——引力增益系数。

如电场力等于电势的负梯度一样，引力为吸引力势场的负梯度，方向由机器人指向目标点。吸引力的大小为：

$$F_{\text{att}}(x)=-\text{grad}[U_{x_d}(x)]=-k_p(x-x_d) \tag{2}$$

单个障碍物在环境中任一点 x 产生的排斥力势场值为：

$$U_o(x)=\begin{cases}\frac{1}{2}\varepsilon\left(\frac{1}{\rho(x)}-\frac{1}{\rho_0}\right)^2, & \rho(x,x_{\text{obs}})\leqslant\rho_0\\ 0, & \rho(x,x_{\text{obs}})>\rho_0\end{cases} \tag{3}$$

式中：x_{obs}——目标点坐标；

ε ——斥力增益系数；

$\rho(x,x_{\text{obs}})$——机器人与障碍物之间的距离。

如电场力等于电势的负梯度一样，斥力为斥力势场的负梯度，方向由障碍物指向机器人。机器人排斥力的大小为：

$$F_{\text{rep}}(x)=-\text{grad}[U_o(x)]=\begin{cases}\varepsilon\left(\frac{1}{\rho(x)}-\frac{1}{\rho_0}\right)\frac{1}{\rho(x)^2}\frac{\partial\rho(x)}{\partial x}, & \rho(x,x_{\text{obs}})\leqslant\rho_0\\ 0, & \rho(x,x_{\text{obs}})>\rho_0\end{cases} \tag{4}$$

式中：$\frac{\partial\rho}{\partial x}$——距离函数的梯度。

2 SARSA(λ)算法

SARSA(λ)算法是一种基于值函数的强化学习算法，在强化学习中，状态行为值函数的定义为：

$$Q_\pi(s,a)=E_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\mid S_t=s,A_t=a\right] \tag{5}$$

式中：$\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$——累积回报；

γ ——折扣因子。

SARSA(λ)的 Q 值更新公式可以由式(5)得到，设状态行为对(s,a)的当前 Q 值估计值为 $Q(s,a)$，机器人在状态 s 做出动作 a 后得到奖励 r 以及到达下一个状态 $s\_$，若用 $r+Q(s_,a_)$去估计 $Q(s,a)$，分别给当前估计值 $Q(s,a)$与新的估计值 $r+Q(s_,a_)$一定的概率置信度 $1-aa$ 和 aa，aa 为一个 0 到 1 的数，则

$$Q(s,a)=(1-aa)*Q(s,a)+aa*(r+Q(s\_,a\_)) \tag{6}$$

展开得

$$Q(s,a)=Q(s,a)+aa*(r+Q(s\_,a\_)-\gamma*Q(s,a)) \tag{7}$$

为了记录获得奖励之前的所有状态动作对，当机器人获得奖励或惩罚时，所有已经记录的机器人经历过的状态动作对都一定程度更新 Q 值，引入资格迹矩阵 $\boldsymbol{E}(s,a)$，得到 SARSA(λ)的 Q 值更新公式为：

$$\begin{aligned}&Q(s,a)=Q(s,a)+aa*(r+Q(s\_,a\_)-\gamma*Q(s,a))*\boldsymbol{E}(s,a)\\&\boldsymbol{E}(s,a)=\gamma*\lambda*\boldsymbol{E}(s,a)\end{aligned} \tag{8}$$

机器人在一个回合内每走一步，之前步经历的资格迹衰减 $\gamma*\lambda$ 倍。

3 改进算法

若地图大小为 $n*n$，则初始化一个 $n*n$ 行，8 列的值全为 k 的 Q 表矩阵，k 为一个大于 0

的正数，机器人可移动方向为上、下、左、右、上左、上右、下左、下右 8 个方向，此矩阵每一行代表一个栅格的 Q 值信息。同时初始化一个 $n*n$ 行，8 列的值全为 0 的资格迹矩阵。

3.1　选择动作

机器人选择下一个动作时，分别计算周围 8 个邻居格子的 a 值与 q 值，式(9)为由合力与 Q 表共同确定的转移概率。其中的 q_i 值是查询 Q 表得到的，a_i 值的确定方法见 4.2。机器人最终转移概率由式(10)确定，即 50%的概率转移到由式(9)得到的 P 最大的栅格处，40%的概率转移到 q 值最大的动作所对应的栅格处，10%的概率向周围 8 个栅格方向随机移动一步。

$$P_i=\frac{a_i^{\alpha}*q_i^{\beta}}{\sum_{i=1}^{8}a_i^{\alpha}*q_i^{\beta}} \tag{9}$$

$$P=\begin{cases}P_{\max}, & \text{rand}<0.5\\ q_{\max}, & 0.5<\text{rand}<0.9\\ \text{random}, & \text{其他}\end{cases} \tag{10}$$

3.2　a 值

计算当前点(即当前位置)受到的合力大小与方向，将此合力在上、下、左、右、上左、上右、下左、下右 8 个方向上投影，得到 8 个分力，将这 8 个分力的大小(可正可负)从小到大排序，a_i 为一个大小为 1～8 序号值，a_i 值越大，表示此方向合力越大，即机器人向此方向移动的概率越大。

3.3　奖励函数

机器人第一次到达目标点时奖励为 R，记录此次机器人经过的总路径长度 mindistance，以后每轮比较路径长度与此路径长度，若出现更小的路径长度，则更新最优路径长度 mindistance。从第 2 轮开始的奖励规则由式(11)确定，其中的 distance 为本次从起点到终点的路径长度，其中的正负号由判断语句决定，若 distance 小于 mindistance 取正号，若 distance 大于 mindistance 取负号。dis 为执行 a 动作前机器人到目标点的距离，dis_为执行 a 动作后机器人到目标点的距离。

$$r=\begin{cases}R\pm(\text{distance}-\text{mindistance})^2, & \text{到达目标点}\\ -D, & \text{撞墙}\\ \text{dis}-\text{dis\_}, & \text{其他}\end{cases} \tag{11}$$

3.4　Q 值更新公式

如前所述，传统 SARSA(λ)的 $Q(s,a)$ 值是由当前的 $Q(s,a)$ 值与 $r+Q(s\_,a\_)$ 的加权平均得到的，其中 $Q(s\_,a\_)$ 是由机器人做出动作 a 后到达位置 $s\_$，再次根据 ε-greedy 策略选择动作 $a\_$，查询 Q 表得到的。为了充分利用之前探索过的信息，此处将 $Q(s\_,a\_)$ 改为

$$Q(s\_,a\_)=\frac{\sum_{i=1}^{8}Q(s\_,i)}{8} \tag{12}$$

式中：$Q(s\_,i)$——机器人在位置 $s\_$，选择动作 i (i 为 1 个 1～8 的整数，代表上、下、左、右、上左、上右、下左、下右)查询 Q 表得到的值。

即将式(8)改为：

$$\begin{aligned}&Q(s,a)=Q(s,a)+aa*\left\{r+\frac{\sum_{i=1}^{8}Q(s,i)}{8}-\gamma*Q(s,a)\right\}*\boldsymbol{E}(s,a)\\&\boldsymbol{E}(s,a)=\gamma*\lambda*\boldsymbol{E}(s,a)\end{aligned} \tag{13}$$

式中：n——格子序号。

分别计算 $s_$周围 8 个格子到目标点的距离加机器人从 $s_$到 8 个格子的移动距离的和，将 8 个格子的对应距离值从大到小排序，n 为格子的距离值的序号，即距离越小，n 值越大，n 为 1～8 的整数。

4　仿　真

文中在 PyCharm2017.1.2 环境下进行了仿真实验，学习率 aa 为 0.01，折扣因子 γ 为 0.9，资格迹衰减因子 λ 为 0.9，其余参数的设置见表 1。

表 1　参数设置

参数	值
n	20
α	1
β	1
ω_2	1
R	10 000
D	1 000

机器人横向或纵向移动距离值增加1，对角线移动距离值增加$\sqrt{2}$，主要对比了改进算法与原SARSA(λ)算法的平均成功次数，即每100回合中找到目标点的次数所占的比重，最短距离长度，首次成功时间，即第一次找到目标点所用的时间。

4.1 人工势场法目标点不可达地图环境

在合力为0的栅格附近人工势场法出现了在障碍物附近抖动的情形，如图1所示。

SARSA(λ)算法路径规划结果和改进算法路径规划结果分别如图2和图3所示。

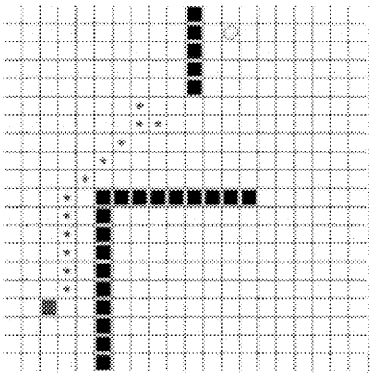

图1 人工势场法路径

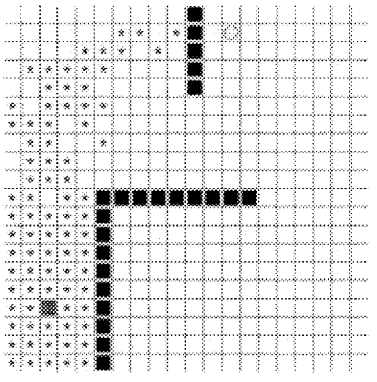

图2 SARSA(λ)路径

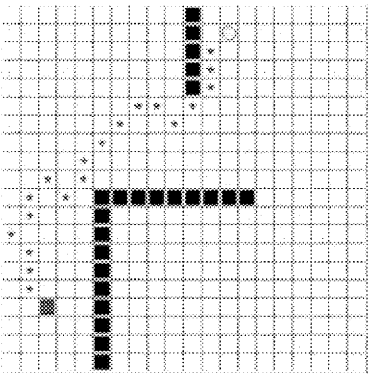

图3 改进算法路径

栅格地图大小为20×20，起点为左下角方块，终点为上方圆形，障碍物为黑色方块。图3中五角星为改进算法100次迭代输出的最优路径。传统SARSA(λ)容易撞墙而无法在100回合内找到目标点(见图2)，五角星为传统SARSA(λ)算法100次迭代中机器人所经过的所有路径点。

4.2 人工势场法目标点可达地图环境

在人工势场法与改进算法都能找到目标点的情况下，这里比较了直接应用SARSA(λ)算法与改进算法的从起点首次到达目标点用时，100次迭代内，成功到达目标点的路径的平均路径长度，100次迭代内，成功到达目标点的路径最短路径长度与从起点到达目标点的平均用时，相关指标对比见表2。

表2 算法结果对比

算法	改进前	改进后	结果对比/%
首次用时/s	45	7	减少85
平均长度	46.5	36	减少22
最短长度	39	32	减少18
平均用时/s	13	9	减少28

SARSA(λ)算法100次迭代输出最优路径如图4所示。

改进算法100次迭代输出最优路径如图5所示。

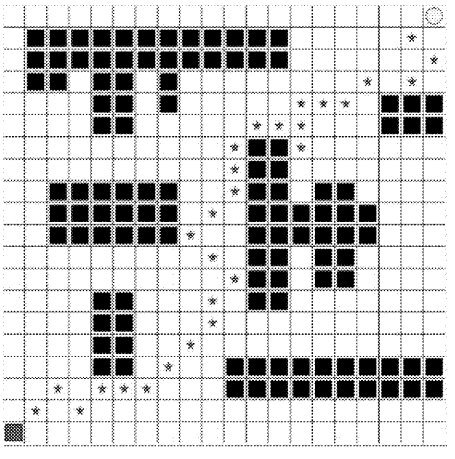

图4 SARSA(λ)算法路径图

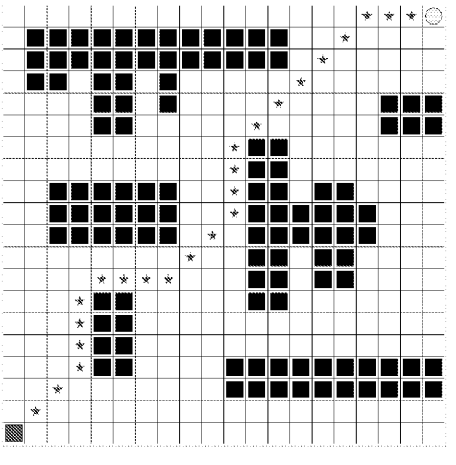

图5 改进算法路径图

5　结　语

由于 SARSA(λ)具有随机性与记忆性，寻路过程中随机挣脱极小值点的概率被增大。仿真实验表明，相比直接应用 SARSA(λ)于路径规划，势场力的引入大幅减少了路径规划所用的时间，由于 SARSA(λ)具有随机性与记忆性，一方面克服了传统人工势场法的目标点不可达的缺陷，另一方面增加了寻找到更优路径的概率。

参考文献：

[1] Dorigo M, Birattari M, Stutzle T. Ant colony optimization[J]. IEEE Computational Intelligence Magazine, 2007, 1(4): 28-39.

[2] Khatib O. Real-time obstacle avoidance for manipulators and mobile robots[C]//IEEE International Conference on Robotics and Automation. Proceedings. IEEE, 2003: 90-98.

[3] 宋宇，王志明.基于 WDO 的无人机全局路径规划方法[J].长春工业大学学报，2017，38(6)：555-559.

[4] 郭宪，方勇纯.深入浅出强化学习[M].北京：电子工业出版社，2018：76-83.

[5] 梁献霞，刘朝英，宋雪玲，等.改进人工势场法的移动机器人路径规划研[J].计算机仿真，2018，35(4)：291-294，361.

[6] 丁家如，杜昌平，赵耀，等.基于改进人工势场法的无人机路径规划算法[J].计算机应用，2016，36(1)：287-290.

[7] 修彩靖，陈慧.基于改进人工势场法的无人驾驶车辆局部路径规划的研究[J].汽车工程，2013，35(9)：808-811.

[8] Panov A I, Yakovlev K S, Suvorov R. Grid path planning with deep rei-nforcement learning: preliminary results[J]. Procedia Computer Science, 2018, 123: 347-353.

[9] 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.