基于 RPkNN-Sarsa(λ)强化学习的机器人路径规划方法*

王军红[1], 江 虹[1], 黄玉清[1], 伍晓利[2]
(1. 西南科技大学 信息工程学院, 四川 绵阳 621010; 2. 中国工程物理研究院 核物理与化学研究所, 四川 绵阳 621010)

摘 要: 基于 kNN-Sarsa(λ)强化学习的机器人路径规划方法虽然收敛速度快,但该算法容易陷入局部最优值,且未考虑环境信息的不完全可观测性。为此,设计了一种随机扰动(random perturbation)kNN-Sarsa(λ)强化学习算法,利用 Bayesian 规则对传感器探测信息的不确定性进行了处理,建立了基于栅格地图的仿真环境模型。仿真实验结果表明,该方法不仅收敛性好,能有效缓解 kNN-Sarsa(λ)算法易陷入局部最优的现象,且在传感器探测信息不确定的情况下仍能搜索到最优路径。

关键词: 路径规划; 强化学习; 随机扰动; 传感器探测信息不确定性
中图分类号: TP242 **文献标志码**: A **文章编号**: 1001-3695(2013)01-0199-03
doi:10.3969/j.issn.1001-3695.2013.01.051

Method of RPkNN-Sarsa(λ) reinforcement learning for robot path planning

WANG Jun-hong[1], JIANG Hong[1], HUANG Yu-qing[1], WU Xiao-li[2]
(1. *School of Information Engineering, Southwest University of Science & Technology, Mianyang Sichuan* 621010, *China*; 2. *Institute of Nuclear Physics & Chemistry, China Academy of Engineer Physics, Mianyang Sichuan* 621010, *China*)

Abstract: The method of robot path planning based on kNN-Sarsa(λ) reinforcement learning has fast convergence speed, but the algorithm is easy to fall into local optimal value and does not consider incomplete observability of environmental information. With regards to this, this paper designed a method of random perturbation kNN-Sarsa(λ) reinforcement learning algorithm. Also, it processed sensors detection data uncertainty using Bayesian theory. In addition, it used grid map to establish simulation environment model. The simulation experimental results show that the method not only has rapid convergence speed by alleviating the local optimal problem of kNN-Sarsa(λ) algorithm, but also can find a shortest path in the case of sensors detection data uncertainty.

Key words: path planning; reinforcement learning; random perturbation; sensors detection data uncertainty

0 引言

强化学习算法用于移动机器人路径规划,其优势在于该算法是一种非监督在线学习方法,且不需要环境的精确模型,因此在动态未知环境下的移动机器人路径规划应用中受到了广泛青睐。利用强化学习算法的机器人路径规划方法有 Q-learning[1,2]、Sarsa、Q(λ)、Sarsa(λ)和 Dyna-Q 等,其中 Dyna-Q 算法收敛速度最快[3]。文献[4]结合 Dyna-Q 强化学习算法提出了 Dyna-H,文献[5]提出了可扩展 Dyna-Q 算法,进一步提高了算法的收敛速度。但这些方法都基于环境信息是完全可观测的,且应用于复杂环境中的机器人路径规划时,存在路径搜索时间长、学习收敛性差的问题。文献[6]提出的 kNN-TD(λ)强化学习算法解决了机器人路径规划时存在的路径搜索时间长、学习收敛性差的问题,但该算法容易陷入局部最优值,且未考虑环境信息探测的不确定性。

本文对 kNN-Sarsa(λ)强化学习算法进行了改进,引入了随机扰动策略[7],使算法能随机调整动作选择策略,提高了算法的全局寻优能力。利用 Bayesian 规则对探测信息的不确定性进行处理,在保证 kNN-Sarsa(λ)算法收敛性的前提下搜索最优路径。

1 RPkNN-Sarsa(λ)强化学习算法

1.1 kNN-Sarsa(λ)强化学习算法

基于 kNN-Sarsa(λ)强化学习算法中值函数 Q 的估计值为[6]

$$\tilde{Q}(a)=\sum_{i=1}^{knn}Q(i,a)p(i)\quad i=1,2,\cdots,knn \tag{1}$$

式中:knn 是当前状态 s_t 的 k 个相邻状态集合;$Q(i,a)$ 为在状态 i 执行动作 a 的最大折算累积回报值;$p(i)$ 为 $P(Q(i,a)=\tilde{Q}(a)|s)$ 的概率,计算式为

$$p(i)=w_i/\sum_{i=1}^{knn}w_i,w_i=\frac{1}{1+l_i^2}\quad i=1,2,\cdots,knn$$

收稿日期: 2012-05-08; **修回日期**: 2012-06-21 **基金项目**: 国防基础科研计划资助项目(B3120110005); 四川省科技厅基金资助项目(2010JZ0020,09ZA136)

作者简介: 王军红(1986-),女,硕士研究生,主要研究方向为通信、智能控制等(junhong-4713@163.com);江虹(1969-),男,教授,博士,主要研究方向为认知技术、智能学习算法等;黄玉清(1962-),女,教授,主要研究方向为机器人技术等;伍晓利(1976-),男,工程师,主要研究方向为检测技术等.

其中，l_i 为 s_t 与其 k 个相邻状态的欧式距离。动作选择策略：

$$\pi^* = \arg\max_a \tilde{Q}(a) \tag{2}$$

Agent 按式(2)所述动作选择策略选择从 s'_t 到 a'的映射，强化学习中 agent 在 s_t 执行 a 后除得到立即奖赏外，还要对后续状态 s到动作 a映射的延迟奖赏进行估计。由式(3)计算：

$$\tilde{Q}'(a) = \sum_{i=1}^{knn'} Q(i,a)p'(i) \quad i = 1,2,\cdots,knn' \tag{3}$$

式中：knn'是 s'对应的 k 个相邻状态集合，$p'(i)$是集合 knn'中每个状态对应的概率。可得到 knn 集合中 k 近邻状态 Q 值函数更新式：

$$Q(i,a)_{t+1} = Q(i,a)_t + \alpha[r_t + \gamma \tilde{Q}'(a')_t - \tilde{Q}(a)_t]e_t \quad \forall i \in knn \tag{4}$$

其中：参数 α 和 γ 分别为(0,1)间的学习率与折扣因子；r_t 为立即回报值；资格迹 e 定义为

$$e(knn,j) = \begin{cases} p(knn) & j = a \\ 0 & j \neq a \end{cases} \tag{5}$$

式中：$p(knn)$是观测状态 s 的 k 近邻集合中各个状态对应的概率。资格迹 e 依据式(6)衰减[3]：

$$e_{t+1} = \gamma\lambda e_t \tag{6}$$

1.2 RPkNN-Sarsa(λ)路径规划算法

式(2)表明，在 kNN-Sarsa(λ)强化学习算法中，agent 选择某条路径上具有累积回报值 Q 最大且与 agent 当前位置距离最短的动作。若 agent 在路径探索过程中一直按式(2)选择动作，将可能导致路径探索出现局部最优。针对该问题，设计中引入了随机扰动策略[7]，在 kNN-Sarsa(λ)动作选择策略中随机增加一定程度的扰动。具有扰动特性的动作选择策略如下：

$$\begin{cases} \tilde{Q}'(a) = \sum_{i=1}^{knn} Q(i,a)p(i) & U \leqslant p_m \\ \tilde{Q}'(a) = \sum_{i=1}^{knn} Q(i,a)(p(i) + \delta) & U > p_m \end{cases} \tag{7}$$

$$\pi^* = \arg\max_a \tilde{Q}'(a) \tag{8}$$

式中：δ 为扰动因子；P_m 为随机转移率；U 为(0,1)间均匀分布的随机数。

式(7)中引入扰动因子 δ 后，将 kNN-Sarsa(λ)算法中动作选择策略分为确定性选择与随机选择两部分。确定性选择按照 kNN-Sarsa(λ)算法中动作选择策略引导 agent 不断探索最优路径；随机选择导致 agent 在选择动作时具有一定的随机性，进一步探索新路径，从而有效地避免了探索陷入局部最优。

2 传感器探测信息不确定性处理

2.1 栅格法建立环境模型

利用栅格法建立如图 1 所示的 2-D 笛卡尔矩形栅格地图表示环境信息，栅格地图将机器人仿真环境分解为 m 行 n 列的小栅格，栅格大小为 5 cm × 5 cm 的正方形，用 0/1 值表示该栅格是否有障碍物。

栅格环境中，每个栅格单元的状态都用一组概率值(x,y)表示。其中 x 表示该栅格单元有障碍物(obstacle)的概率，y 表示该栅格单元空闲(empty)的概率。实际中由于机器人实验环境是未知的，对该环境无任何先验信息，因此，所有栅格的初始状态都为(0.5, 0.5)。根据传感器对环境信息的不断探测来更新每个栅格状态，在更新过程中，若 x 大于初始状态值，则表示该栅格有障碍物。对无障碍物的栅格，由于传感器不会返回测量数据，其状态值保持初始化值，机器人确定其为空闲栅格。

2.2 激光测距传感器探测模型

激光测距传感器基本模型如图 2 所示(图中半圆形区域为激光测距传感器探测范围，黑色小块为障碍物)，其扫描范围为 0° ~ 180°，扫描最大距离为 R。以机器人当前运行方向为中心轴线扫描 180°，在该半圆范围内激光测距传感器发射激光束，当激光束在探测区域内碰到障碍物时，返回障碍物到机器人中心点的距离 d_i 与角度 α_i。

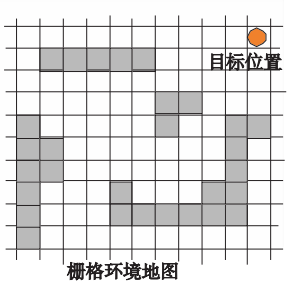


图1 机器人路径规划栅格环境模型

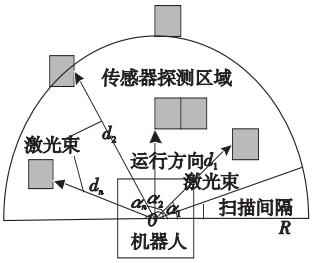


图2 激光测距传感器基本探测模型

2.3 利用传感器读数更新栅格状态

1)计算条件概率

将栅格 grid(i,j)值为 1 的区域视为有障碍物，计算该障碍物栅格到机器人中心点的距离 d_i 与角度 α_i，作为激光测距传感器探测到的障碍物信息，再转换为概率表示[8]：

$$P(S_{\text{sensors}}|O) = \frac{(R - d_i)/R + (180 - \alpha_i)/180}{2} \times \max_{\text{occupied}} \tag{9}$$

$$P(S_{\text{sensors}}|E) = 1 - P(S_{\text{sensors}}|O) \tag{10}$$

其中：$P(S_{\text{sensors}}|O)$表示栅格有障碍物的情况下传感器探测到该栅格有障碍物的概率；$P(S_{\text{sensors}}|E)$表示传感器探测到该栅格空闲的概率；$(R - d_i)/R$ 项表示栅格单元距激光传感器越近可信度就越高；$(180 - \alpha_i)/180$ 项表示栅格单元越靠近激光传感器中心轴线可信度就越高；$\max_{\text{occupied}}$表示传感器探测到该栅格有障碍物的可能性(依据传感器的分辨率等探测性能)。

2)计算后验概率

路径规划中机器人根据 $P(O|S_{\text{sensors}})$值进行避障，$P(O|S_{\text{sensors}})$为由传感器读数计算出的激光束扫描范围内有障碍物栅格的概率。贝叶斯规则揭示了这两种概率的关系[9]：

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{\sum_{i=1}^{n} P(E|H_i) \cdot P(H_i)}$$

其中：H_i 用有障碍物(O)、空闲(E)来代替，从而得到 $P(O|S_{\text{sensors}})$为

$$P(O|S_{\text{sensors}}) = P(S_{\text{sensors}}|O)P(O)/[P(S_{\text{sensors}}|O)P(O) + P(S_{\text{sensors}}|E)P(E)] \tag{11}$$

$$P(E|S_{\text{sensors}}) = 1 - P(O|S_{\text{sensors}}) \tag{12}$$

3)栅格状态更新

对传感器探测到的状态值未更新栅格，用栅格初始状态概率值计算当前状态的概率值，确定当前状态下该栅格有障碍物的概率。利用贝叶斯规则计算如下：

$$P(O|S_{\text{sensors}_t}) = P(S_{\text{sensors}_t}|O)P|S_{\text{sensors}_{t-1}})/[P(S_{\text{sensors}_t}|O)P(O|S_{\text{sensors}_{t-1}}) + P(S_{\text{sensors}_t}|E)P(E|S_{\text{sensors}_{t-1}})]$$

$$P(E|S_{\text{sensors}_t}) = 1 - P(O|S_{\text{sensors}_t})$$

由此可计算出栅格 grid(i,j)的状态值为$(P(O|S_{\text{sensors}_t}),$

$P(E \mid S_{\text{sensors\_}t})$),作为该栅格单元是否有障碍物信息的判断依据。

3　实验结果及分析

仿真实验中,建立如图 3 所示的迷宫为环境模型,大小为 39×36 个栅格,设置机器人起始点 S 与目标点 G 位置(该位置可随意设置),将机器人看成一圆形质点,机器人行走步长为一个栅格单元,基本动作:上(↑)、下(↓)、左(←)、右(→)。仿真参数取值:$\alpha = 0.98$,$\gamma = 0.95$,步长因子 $\lambda = 0.95$,$k = 4$,epsilon = 0.001(每个学习周期 epsilon = epsilon × 0.99),$P_m = 0.01$,δ 为(0,1)间均匀分布的随机数。

3.1　探测信息确定性与不确定性情况下两算法路径规划结果

在图 3(a)所示的仿真环境下,表 1 为在环境信息确定性观测与不确定性观测两种情况下,利用 kNN-Sarsa(λ)与 RPkNN-Sarsa(λ)强化学习算法路径规划结果。其中,学习周期数(Episodes)为强化学习中的学习次数;步长数(steps)为该算法收敛到近似最短路径时所用总步数;仿真时间(simulation times)为利用该算法完成路径规划所用时间。

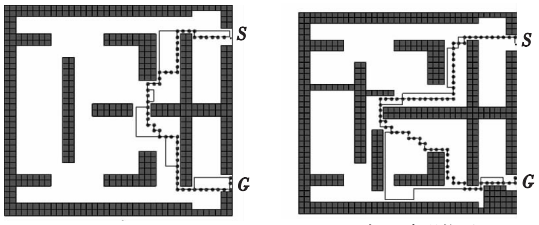
表 1　不同情况下两种算法路径规划结果对比

不同情况下	算法	学习周期数	步长数	仿真时间/s
确定性	kNN-Sarsa(λ)	50	64	43.08
	RPkNN-Sarsa(λ)	50	57	39.61
不确定性	kNN-Sarsa(λ)	50	64	91.21
	RPkNN-Sarsa(λ)	50	57	84.70

由表 1 可以看出:传感器探测信息确定性与不确定性两种情况下,kNN-Sarsa(λ)算法路径规划步长数均为 64,RPkNN-Sarsa(λ)算法路径规划步长数均为 57,说明前面所述传感器探测信息不确定性处理的合理性。传感器探测信息确定时,利用 RPkNN-Sarsa(λ)路径规划算法仿真时间比 kNN-Sarsa(λ)算法少约 8%;传感器探测信息不确定时,利用 RPkNN-Sarsa(λ)路径规划算法仿真时间比 kNN-Sarsa(λ)算法少约 7%,说明两种情况下 RPkNN-Sarsa(λ)算法都能比 kNN-Sarsa(λ)算法搜索到更短路径。

3.2　探测信息不确定性情况下两种算法路径规划仿真结果

在考虑传感器探测信息不确定性情况下,分别运用 kNN-Sarsa(λ)与 RPkNN-Sarsa(λ)算法的路径规划仿真结果如图 3 所示。

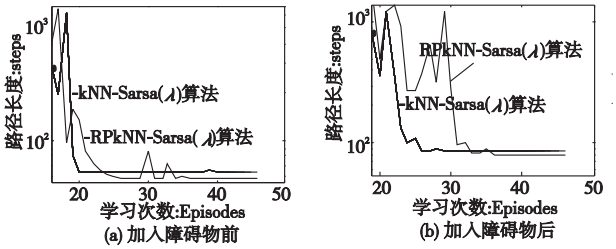


(a) 加入障碍物前　　(b) 加入障碍物后

图3　随机加入障碍物前后两种算法路径规划仿真结果

图 3 中实线为利用 kNN-Sarsa(λ)算法路径规划结果,带点虚线为利用 RPkNN-Sarsa(λ)算法路径规划结果。由图 3 可以看出:仿真环境中加入障碍物前后,利用 RPkNN-Sarsa(λ)算法搜索的路径长度均比 kNN-Sarsa(λ)算法短约 11.4%。

图 4 为仿真环境中加入障碍前后,kNN-Sarsa(λ)与 RPkNN-Sarsa(λ)算法的收敛性曲线。



(a) 加入障碍物前　　(b) 加入障碍物后

图4　两种算法收敛性曲线

由图 4 可以看出:(a)中 kNN-Sarsa(λ)算法经过 20 次学习后收敛,RPkNN-Sarsa(λ)算法在 36 次学习后收敛;(b)中 kNN-Sarsa(λ)算法经过 26 次学习后收敛,而 RPkNN-Sarsa(λ)算法在 36 次学习后收敛。由此可见,两种不同仿真场景中,kNN-Sarsa(λ)算法经过 20 多次学习后均陷入局部最优解;而 RPkNN-Sarsa(λ)算法虽然均在 36 次学习后才收敛,但仍能搜索到更短路径。因此,RPkNN-Sarsa(λ)算法的扰动策略能有效缓局部最优解的发生,避免了局部收敛。

4　结束语

本文在 kNN-Sarsa(λ)强化学习算法动作选择策略中加入随机扰动,将确定性选择与随机选择相结合,设计了基于 RPkNN-Sarsa(λ)强化学习算法的机器人路径规划方法。介绍了激光测距传感器模型,并利用 Bayesian 规则对传感器探测到的障碍物信息转换为概率表示,以确定栅格存在障碍物的可能性。最后将该方法应用于两种不同仿真场景,并与 kNN-Sarsa(λ)强化学习算法进行了对比。仿真实验结果表明:该方法不仅能够解决探测信息不确定性条件下的路径规划,还能够有效提高全局寻优能力,为最终实现该算法在真实机器人平台上的实际应用奠定了基础。

参考文献:

[1] CHAKRABORTY I G, DAS P K, KONAR A, *et al*. Extended Q-learning algorithm for path-planning of a mobile robot[C]//Proc of the 8th International Conference on Simulated Evolution and Learning. Berlin: Springer-Verlag, 2010: 379-383.

[2] MOHAMMAD A K J, MOHAMMAD A R, LARA Q. Reinforcement based mobile robot navigation in dynamic environment[J]. Robotics and Computer-Integrated Manufacturing, 2011, 27(1): 135-149.

[3] VIET H H, KYAW P H, CHUNG T C. Simulation-based evaluations of reinforcement learning algorithms for autonomous mobile robot path planning[C]//Proc of the 3rd FTRA International Conference on Information Technology Convergence and Services. 2012: 467-476.

[4] SANTOS M, MARTIN H J A, LOPEZ V, *et al*. Dyna-H: a heuristic planning reinforcement learning algorithm applied to role-playing game strategy decision systems[J]. Knowledge-Based Systems, 2012, 32: 28-36.

[5] VIET H H, AN S H, CHUNG T C. Extended Dyna-Q algorithm for path planning of mobile robots[J]. Journal of Measurement Science and Instrumentation, 2011, 2(3): 283-287.

[6] MARTIN H J A, LOPE J D, MARAVALL D. The kNN-TD reinforcement learning algorithm[C]//Proc of the 3rd International Work-Conference on the Interplay Between Natural and Artificial Computation. Berlin: Springer-Verlag, 2009: 305-312.

[7] 郝晋. 蚁群优化算法及其在电力系统短期发电计划中的应用研究[D]. 重庆: 重庆大学, 2002: 12-14.

[8] 王健, 张汝波. 基于 POMDP 模型的机器人导航控制方法[J]. 华中科技大学学报: 自然科学版, 2008, 36(增刊 I): 14.

[9] 吴峰. 基于决策理论的多智能体系统规划问题研究[D]. 合肥: 中国科学技术大学, 2011: 34.