DOI:万方数据 cnki robot 万方数据
第 28卷第 5期
2006年 9月
机器人 ROBOT
Vol 28 No 5
Sept, 2006

文章编号: 1002 0446(2006) 05 0544 04

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*

沈 晶, 顾国昌, 刘海波

(哈尔滨工程大学计算机科学与技术学院, 黑龙江 哈尔滨 150001)

摘 要: 提出了一种基于分层强化学习的移动机器人路径规划算法. 该算法利用强化学习方法的无环境模型学习能力以及分层强化学习方法的局部策略更新能力, 克服了路径规划方法对全局环境的静态信息或动态障碍物的运动信息的依赖性. 仿真实验结果表明了算法的可行性, 尽管在规划速度上没有明显的优势, 但其应对未知动态环境的学习能力是现有其它方法无法比拟的.

关键词: 移动机器人; 未知动态环境; 路径规划; 分层强化学习

中图分类号: TP24 文献标识码: B

Mobile Robot Path Planning Based on Hierarchical Reinforcement Learning in Unknown Dynamic Environment

SHEN Jing GU Guo-chang LIU Hai-bo

(*School of Computer Science and Technology, Harbin Engineering University, Harbin* 150001, *China*)

Abstract: A path-planning algorithm based on hierarchical reinforcement learning is presented. Since the reinforcement learning approach is introduced, the algorithm is provided with the capability of learning without environment model. The hierarchical reinforcement learning method is mainly employed for updating local strategies. So, this algorithm can eliminate its dependence on the static information of the global environment or the moving information of the dynamic obstacles. Simulation experiment shows the feasibility of the algorithm. Although there is no obvious advantage in planning speed, the learning ability of the algorithm in unknown dynamic environment is unique.

Keywords: mobile robot; unknown dynamic environment; path planning; hierarchical reinforcement learning

1 引言 (Introduction)

路径规划是移动机器人导航中最基本的环节之一, 是指在有障碍物的工作环境中寻找一条从给定起点到终点的较优的运动路径, 使机器人在运动过程中能安全、无碰撞地绕过所有的障碍物, 且所走路径最短, 这是一类 NP难问题.

静态环境中, 如果机器人具备全局环境信息, 可用全局路径规划方法, 这方面的研究已有广泛的报道, 例如: 可视顶点图[1]、人工势场法[2]、遗传算法[3]、神经网络[4]、随机树[5]等方法; 如果机器人不具备全局环境信息, 可采用强化学习[6]或基于滚动窗口的方法[7]. 但机器人的工作环境多数是动态不确定的, 机器人对有动态障碍物的环境很难具有先验知识, 这种情况下, 机器人只能根据实时探测到的环境信息进行避碰规划, 解决此类问题已有诸多方法: 滚动路径规划方法[8]有效地解决了动态障碍物环境下机器人运动过程中的安全避碰问题, 然而由于缺少全局信息, 加上机器人的视野有限, 滚动出的路径很难达到全局较优; 蚂蚁预测算法[9]可以在障碍物非常复杂的环境中迅速规划出安全路径, 但该算法需要先基于静态全局信息规划出全局最优路径; 采用基于链接图的遗传算法建立全局优化路径并采用基于行为的方法进行局部路径规划[10], 可以很好地解决全局与局部路径规划的关系, 但该方法必须假定静态障碍物的分布是已知的; 人工水流法[11]有效解决了多机器人共存环境中的路径规划问题, 使得每个机器人无需知道其他机器人的轨迹和

* 基金项目: 国防基础研究计划资助项目; 哈尔滨工程大学基础研究基金资助项目(HEUFT05068, HEUFT05021).
收稿日期: 2005 - 12 - 19

各机器人的运动优先权就可以实现避碰, 但该方法未将路径最短作为规划目标, 常会出现环路, 而且该方法需要根据环境信息事先生成虚拟地形图; 模糊神经网络[12]、遗传算法[13]和改进的人工势场法[14]也都表现出了良好的动态路径规划能力, 但这些方法都必须要求知道障碍物的运动信息 (如速度等); 通过建立交通规则[15]可以有效解决动态环境中多机器人避碰规划问题, 但在很多实际问题中动态障碍物未必会遵守交通规则.

上述动态环境中的路径规划方法均高度依赖于全局环境的静态信息或动态障碍物的运动信息, 但在未知动态环境中, 这些信息都是很难获得的, 因而限制了算法的应用. 本文提出一种基于分层强化学习 (HRL: hierarchical reinforcement learning)的移动机器人路径规划算法, 该算法利用强化学习方法的无环境模型学习能力以及 HRL的局部策略更新能力克服机器人在路径规划时对全局环境的静态信息或动态障碍物的运动信息的依赖性.

2 分层强化学习原理 (Principles of HRL)

强化学习[16]通过试错 (trial-and-error)与环境交互来改进策略, 其自学习和在线学习的特点使其成为机器学习研究的一个重要分支, 但强化学习一直被维数灾难所困扰, 为此, 研究人员又提出了分层强化学习[17]. 目前, 代表性的 HRL方法主要有 Sutton提出的 Option[18]、Parr提出的 HAM[19]和 Dietterich提出的 MAXQ[20]方法.

HRL的核心思想是引入抽象机制对学习任务进行分解, 抽象机制允许学习 Agent忽略与当前阶段任务有关的具体细节, 而把完成具体细节的基本动作组成的动作序列看作一个抽象, 并根据抽象内部策略执行相应的基本动作序列. 因此, Agent无需在每个时间步都对动作做出决策, 而是在每个抽象执行完成之后才进行下一次决策. 宏[21]是最简单的一种抽象形式, 每个宏算子包含一个动作序列, 可按名称调用其他宏或被其它宏调用. 从控制角度来说, 宏是一种开环控制策略, 并不满足控制系统 (尤其是随机控制系统)的要求, 但 HRL将宏思想拓展到了闭环控制策略中, 定义出基于状态子集的局部策略, 便形成了 Option、HAM 和 MAXQ等方法. 本文的路径规划算法是在 Option方法的基础上提出的, 下面介绍该方法.

Option方法中, 学习任务被抽象成若干 Option, 并将这些 Option作为一种特殊的“动作 ”加入到原来的动作集中. 一个 Option可以理解为为完成某子目标而定义在某状态子空间上的按一定策略执行的动作或 Option序列, 普通动作可以视为 Option的一种特例. 设 S 和 A 分别为 Agent的状态集和动作集, 最简单的一种 Option是直接定义在 MDP(Markov decision process)上的, 用三元组 〈φ, π, β〉表示. 其中, $\varphi \subseteq S$ 为入口状态集, 当且仅当 $s \in \varphi$ 时, Option 〈φ, π, β〉可依策略执行, 通常, φ包含且只包含该 Option经历的所有可能状态; $\pi: \varphi \times A_\varphi \to [0, 1]$ 为 Option内部策略, A_φ 为在状态集 φ上可执行的动作集; $\beta: S \to [0, 1]$ 为 Option终止条件, Option在某一状态 s' 依概率 $\beta(s')$ 终止, 通常, 将 Option要达到的子目标状态 s_ξ 定义为 $\beta(s_\xi)=1$. 如果将策略定义在 Option之上, 即 $\mu: \varphi \times O_\varphi \to [0, 1]$, O_φ 为状态集 φ上的可执行的 Option集, φ和 β定义不变, 则 〈φ, μ, β〉即形成分层 Option. Option 〈φ, π, β〉称为 Markov-Option, Option〈φ, μ, β〉称为 Semi-Markov-Option, 将 Semi-Markov-Option叠加在核心 MDP上, 便形成了 SMDP (Semi-MDP).

Option方法中, Option执行结束时才进行一次学习, 式 (1)为 Q学习更新公式.

$$Q_{k+1}(s, o) = (1-\alpha_k) Q_k(s, o) + \alpha_k \left[r + \gamma^{\tau} \max_{o' \in O_{s'}} Q_k(s', o') \right] \quad (1)$$

其中, k 为迭代次数, α_k 为学习率, r 是 Agent从环境状态 s 转移到 s' 后所接受的奖赏值 (强化信号), γ为折扣因子, τ为 Option o 持续的时间. Precup证明了在标准 Q学习收敛条件下, Option方法以概率 1收敛到最优策略[22].

Option可以由设计者根据专家知识事先确定, 也可以自动生成. 目前解决自动分层问题 (包括 Option、HAM 和 MAXQ)的研究工作多集中在状态空间的子目标发现上, 根据子目标即可对状态和动作进行抽象以形成分层子任务[23~25].

3 未知动态环境中的路径规划算法 (Path planning algorithm in unknown dynamic environment)

强化学习是一个 MDP, MDP模型本身未考虑动态环境问题, 作为 SMDP的分层强化学习同样不能直接处理动态环境中的路径规划问题. 但是, 强化学习任务经过分层之后, 学习任务或局限于机器人当前所处的规模较小的局部空间 (局部路径规划), 或局限于与底层细节无关的维数较低的高层空间 (全局

路径规划),这样,采用分层强化学习在未知动态环境中进行路径规划,机器人只需关注当前局部空间内的环境变化和分层任务子目标状态的变化,将策略更新过程限制在局部空间或高层空间上,从而加快学习(规划)速度,使算法在可以接受的时间内收敛.

设有静态和动态障碍物的二维栅格环境中有1个机器人执行给定起终点的路径规划任务,机器人对环境完全未知,但可在环境中执行"上"、"下"、"左"、"右"单步移动动作或符合入口条件的Option,在移动到下一栅格前,可对该栅格进行探测,以确定该栅格是否有障碍物.路径规划(学习)算法如下:

1)初始化空白状态转移结构图和随机Q表;

2)观察当前状态 s;

3)选择并执行一个动作 a;

4)观察下一个状态 s';

5)获得一个奖赏值 r;

6)按式(1)调整 Q 值;

7)修改状态结构图;

8)记录学习经验(s, a, s', r);

9)如果连续 T 个探测周期内没有探索到新状态则转10),否则转2);

10)计算子目标;

11)生成Option入口状态集;

12)各Option内部策略学习;

13)Semi-Markov-Option策略学习;

14)观察当前状态;

15)选择一个Option;

16)依当前Option内部策略执行动作,同时观察环境状态,若环境状态发生非受控变化,则转17),否则直到Option终止后转20);

17)若子目标环境发生变化,则转18),若当前Option内部环境发生变化,则转19),否则转16);

18)Semi-Markov-Option策略重新学习,转16);

19)当前Option内部策略重新学习,转16);

20)按式(1)更新 Q 值;

21)若不满足规划(学习)结束条件则转14),否则算法停止.

步骤1)~9)是在标准Q学习算法中插入了构造状态转移结构图的步骤,状态转移结构图定义为三元组 $\langle V, U, E\rangle$,其中 V 为探明状态,即第2)步中观察过的状态,U 为探明的不可达状态或未探测状态,E 为状态间的可达关系,状态转移结构图用于计算子目标和生成Option.步骤6)按式(1)调整 Q 值,前已述及,普通动作可以视为Option的一种特例,则式(1)中取 $\tau=1$ 并用 a 替换 o 即成适用于步骤6)的标准Q学习更新公式.步骤8)记录的学习经验可在步骤12)和19)中(采用经验回放[26])学习Option内部策略时使用.步骤2)到9)的一次循环称为一个探测周期,参数 T 用于控制探测程度,可根据经验确定,一般设 $T=2$ 即可满足要求.步骤10)中计算子目标可以采用Q-Cut[25]等方法计算.算法规划(学习)结束条件有多种设置方法,如达到预定学习周期或误差要求等.

算法执行期间,通过若干周期的反复试错和学习,同时进行全局和局部路径规划(包括探测和优化);算法结束后,从起点状态出发,按照Q表中记录的最优策略得到的就是机器人实际走过的到达终点的较优路径.

4 仿真实验与分析(Simulation and analysis)

仿真以移动机器人在有静态和动态两种障碍物的二维栅格空间中规划给定的起终点间的最短路径为任务背景(图1),黑色栅格为障碍空间,白色栅格为无障碍空间,标有字母的4类栅格:S、G 分别为起终点,R 为学习机器人,M 为移动障碍物,移动障碍物在每个时间步内可随机地向上、下、左、右任一方向移动一个栅格.

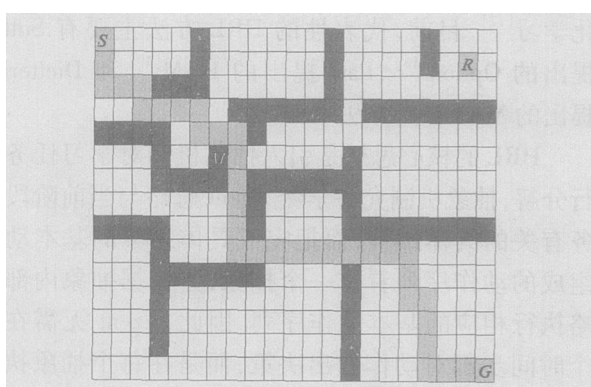

图1 实验环境

Fig 1 Experiment environment

执行规划前,机器人 R 对环境信息完全未知.规划过程中,R 以0.2的概率对环境进行探测(贪婪策略),与障碍物碰撞前可以感知障碍物的存在,并得到值为 -20 的惩罚信号,R 到达目标点即完成一个学习周期,并获得值为100的奖励信号,然后开始下一个周期的学习.设 $T=2$,即连续两个周期中未发现新状态,即开始计算子目标并生成Option,计算子目

标采用 Q-Cut算法[25], Q学习中学习率设为固定的 0.1, 折扣因子设为 0.9. 图 1中灰色标记一条规划好的路径. 图 2显示了规划(学习)算法的收敛情况(为清楚观察曲线变化的趋势, 纵坐标做了取对数处理).

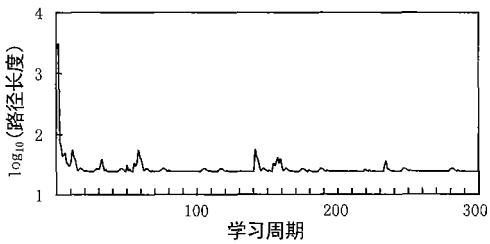

图 2　规划算法收敛情况

Fig. 2　Convergence of the path-planning algorithm

由图 2可见, 算法尽管不能严格收敛(由于动态障碍物的干扰是不可能严格收敛的), 但是平均路径长度(44步)已接近最短路径(34步), 这是因为算法学习过程中没有盲目地对环境的所有动态变化都做出响应, 只是关注影响子目标的状态和当前 Option的内部状态而忽略了其它状态的变化, 并根据不同类型状态的变化做出不同的反应, 将策略更新过程限制在局部空间(当前 Option的内部状态变化时)或高层空间(子目标状态变化时)上, 从而加快更新速度, 在受到移动障碍物干扰后能够快速收敛到最短路径, 从而体现出一种"几乎处处"收敛的趋势.

机器人在规划前不具备环境的全局静态信息, 在规划过程中也无从获取动态障碍物的运动信息, 文[1~7]中的规划算法均不适用于动态环境, 文[8~15]中的规划算法因缺乏全局环境的静态信息或动态障碍物的运动信息也不能正常运行, 因此实验中并未将本文的算法与引言中提及的上述算法直接进行比较. 但事实上, 在静态环境中本文的算法并不会规划出优于文[1~7]中算法规划出的路径, 而在全局环境的静态信息及动态障碍物的运动信息可知的情况下, 文[8~15]中算法的规划速度也要远比本文算法快, 因此说, 本文算法仅在未知动态环境中才能体现出其学习能力上的优越性.

5　结论(Conclusion)

本文以 Option分层强化学习方法为基础提出一种适用于未知动态环境的移动机器人路径规划(学习)算法, 该算法利用强化学习方法的无环境模型学习能力以及 HRL的局部策略更新能力克服机器人在路径规划时对全局环境的静态信息或动态障碍物的运动信息的依赖性. 仿真实验表明, 该算法在动态环境中能达到"几乎处处"收敛的效果, 从而在可容忍的时间内规划出近似最优路径, 因此可以说本文的算法是可行的, 尽管在规划速度上没有明显的优势, 但其应对未知动态环境的学习能力是现有其它方法无法比拟的. 如何加快规划速度、如何在统计动态障碍物运动特征的基础上提高规划结果的稳定性, 尚需进一步研究.

参考文献　(References)

[1] Lozano-Perez T, Wesley M A. An algorithm for planning collision-free paths among polyhedral obstacles[J]. Communications of the ACM, 1979, 22(10): 560-570.

[2] Khatib O. Real-time obstacle avoidance for robot manipulator and mobile robots[J]. The International Journal of Robotics Research, 1986, 5(1): 90-98.

[3] Wang C-M, Soh Y C, Wang H, *et al*. A hierarchical genetic algorithm for path planning in a static environment with obstacles[A]. Proceedings of the 2002 IEEE Canadian Conference on Electrical and Computer Engineering[C]. Piscataway, USA: IEEE, 2002. 1652-1657.

[4] D'Amico A, Ippoliti G, Longhi S A. Radial basis function networks approach for the tracking problem of mobile robots[A]. Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics[C]. Piscataway, USA: IEEE, 2001. 498-503.

[5] Bruce J, Veloso M. Real-time randomized path planning for robot navigation[A]. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems[C]. Piscataway, USA: IEEE, 2002. 2383-2388.

[6] 张汝波, 杨广铭, 顾国昌, 等. Q-学习及其在智能机器人局部路径规划中的应用研究[J]. 计算机研究与发展, 1999, 36(12): 1430-1436.

[7] 张纯刚, 席裕庚. 全局环境未知时基于滚动窗口的机器人路径规划[J]. 中国科学(E辑), 2001, 31(1): 51-58.

[8] 张纯刚, 席裕庚. 一类动态不确定环境下机器人的滚动路径规划[J]. 自动化学报, 2002, 28(2): 161-174.

[9] 朱庆保. 动态复杂环境下的机器人路径规划蚂蚁预测算法[J]. 计算机学报, 2005, 28(11): 1898-1906.

[10] 朴松昊, 洪炳镕. 一种动态环境下移动机器人的路径规划方法[J]. 机器人, 2003, 25(1): 18-21, 43.

[11] 韩学东, 洪炳镕, 孟伟. 一种新型的路径规划方法——人工水流法[J]. 高技术通讯, 2004, 14(4): 53-57.

[12] 谢宏斌, 刘国栋, 李春光. 动态环境中基于模糊神经网络的机器人路径规划的一种新方法[J]. 江南大学学报(自然科学版), 2003, 2(1): 20-23, 27.

[13] 刘国栋, 谢宏斌, 李春光. 动态环境中基于遗传算法的移动机器人路径规划的方法[J]. 机器人, 2003, 25(4): 327-330, 343.

(下转第 552页)

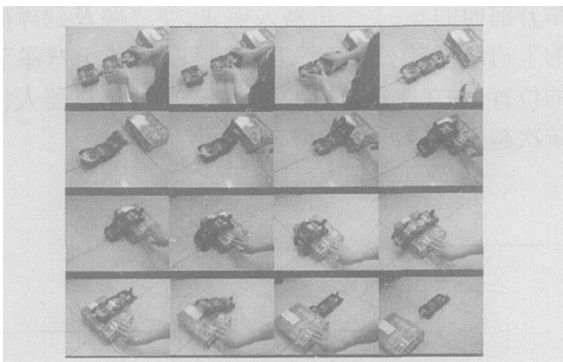

图 10 三节链形机器人攀爬 100mm 盒子
Fig. 10 Three-module chain-shaped robot climbing 100mm height box

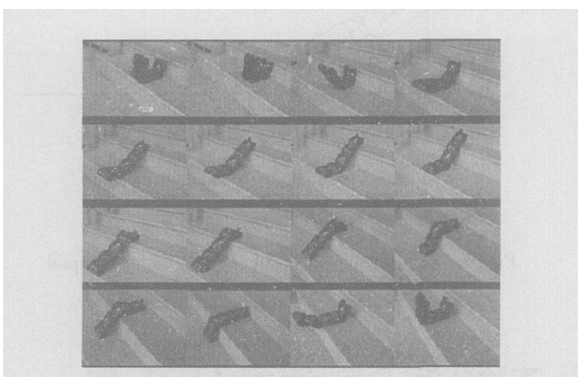

图 11 五节链形机器人爬越楼梯
Fig. 11 Five-module chain-shaped robot climbing stairs

7 结论(Conclusion)

本文研制一种模块化可重构履带移动式微小型机器人,单个机器人模块由推进机构、弯举机构、重构机构和躯体组成,可独立运动,多个微小型机器人模块可重构成链形机器人及环形机器人,链形机器人具有较强的越障能力,环形机器人具有高速及路面适应能力强的特点.对链形机器人越障步态进行了分析,应用该分析结果进行攀越箱体和楼梯的实验,取得了预期效果,验证了分析结果的有效性.

参考文献 (References)

[1] Fujita M, Kitano H, Kageyama K. A reconfigurable robot platform [J]. Robotics and Autonomous Systems, 1999, 29(2-3): 119-132.

[2] Goldenberg A, Kircanski N M, Kuzan P, *et al*. Modular expandable and reconfigurable robot[J]. Robotics and Computer-Integrated Manufacturing, 1997, 13(2): 173-174.

[3] 费燕琼,董庆雷,赵锡芳. 自重构模块化机器人的结构[J]. 上海交通大学学报, 2005, 39(6): 877-883.

[4] 刘国才. 模块化可重构履带式微小型机器人的研究[D]. 哈尔滨:哈尔滨工业大学, 2005.

作者简介:

李满天 (1975-),男,硕士,助理研究员. 研究领域:移动机器人.

黄 博 (1974-),男,硕士,讲师. 研究领域:移动机器人.

孙立宁 (1964-),男,博士,教授. 研究领域:移动机器人,微驱动 微操作,医疗机器人,MEMS.

(上接第 547页)

[14] 覃柯,孙茂相,孙昌志. 动态环境下基于改进人工势场法的机器人运动规划[J]. 沈阳工业大学学报, 2004, 26(5): 568-571, 582.

[15] 徐潼,唐振民.动态环境中的移动机器人避碰规划研究[J]. 机器人, 2003, 25(2): 117-122, 139.

[16] 张汝波. 强化学习理论及应用[M]. 哈尔滨:哈尔滨工程大学出版社, 2001.

[17] Barto A G, Mahadevan S. Recent advances in hierarchical reinforcement learning[J]. Discrete Event Dynamic Systems: Theory and Applications, 2003, 13(1-2): 41-77.

[18] Sutton R S, Precup D, Singh S P. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1): 181-211.

[19] Parr R. Hierarchical Control and Learning for Markov Decision Processes[D]. Berkeley: University of California, 1998.

[20] Dietterich T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Journal of Artificial Intelligence Research, 2000, 13(1): 227-303.

[21] Korf R E. Learning to Solve Problems by Searching for Macro-operators[M]. London: Pitman Publishing Ltd., 1985.

[22] Precup D. Temporal Abstraction in Reinforcement Learning[D]. Amherst: University of Massachusetts, 2000.

[23] Digney B L. Learning hierarchical control structures for multiple tasks and changing environments[A]. From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior[C]. Cambridge, USA: MIT Press, 1998. 321-330.

[24] McGovern A, Barto A. Automatic discovery of subgoals in reinforcement learning using diverse density[A]. Proceedings of the 8th International Conference on Machine Learning[C]. San Francisco: Morgan Kaufmann, 2001. 361-368.

[25] Menache I, Mannor S, Shimkin N, *et al*. Q-cut: dynamic discovery of sub-goals in reinforcement learning[A]. Proceedings of the 13th European Conference on Machine Learning[C]. Berlin, Germany: Springer-Verlag, 2002. 295-306.

[26] Lin L G. Self-improving reactive agents based on reinforcement learning, planning and teaching[J]. Machine Learning, 1992, 8(3): 293-321.

作者简介:

沈 晶 (1969-),女,博士生. 研究领域:分层强化学习,人工免疫系统.

顾国昌 (1946-),男,教授,博士生导师. 研究领域:智能机器人技术.

刘海波 (1976-),男,博士,IEEE计算机学会专业会员. 研究领域:智能机器人体系结构,多智能体系统.