DOI:10.16652/j.issn.1004-373x.2019.17.028

基于仿生学内在动机的Q学习算法移动机器人路径规划研究

李福进，张俊琴，任红格
（华北理工大学 电气工程学院，河北 唐山 063200）

摘 要：针对移动机器人在未知环境中避障和路径规划自适应能力差的问题，受心理学方面内在动机启发，以加入引力势场的Q学习理论为基础，提出一种基于内在动机机制的引力场Q(IM-GPF-Q)学习算法。该算法以Q学习为理论框架，加入引力势场为算法提供先验知识，以内在动机作为内部奖励，与外部信号一起生成取向评价值，指引机器人学会自主选择最优路径。通过模拟客厅环境和两种具有陷阱的环境中进行的仿真实验，结果表明该算法能使机器人通过与外界未知环境进行交互获得认知，最终完成路径规划任务，与传统强化学习方法相比具有更快的收敛速度以及更好的自学习和自适应能力。

关键词：移动机器人；路径规划；内在动机；Q学习算法；引力势场；智能发育
中图分类号：TN99-34；TP183　　**文献标识码**：A　　**文章编号**：1004-373X(2019)17-0133-05

Research on mobile robot path planning by Q-learning algorithm based on bionics intrinsic motivation

LI Fujin, ZHANG Junqin, REN Hongge
(College of Electrical Engineering, North China University of Science and Technology, Tangshan 063200, China)

Abstract: In allusion to the problem of poor self-adaptive ability of mobile robots in obstacle avoidance and path planning in unknown environment, a kind of gravitational field Q-learning algorithm based on intrinsic motivation (IM-GPF-Q) is put forward, which is inspired by intrinsic motivation theory of psychology and is based on the Q learning theory of the gravitational potential field. The algorithm takes Q-learning as a theoretical framework, adds gravitational potential field to provide prior knowledge, introduces intrinsic motivation as an internal reward, generates orientation evaluation values together with external signals, and guides robots to learn to choose the optimal path. The simulation experiment was performed in the simulated living room environment and two trapped environments. The results show that the algorithm can enable robots to gain cognition by interacting with the unknown environment, and ultimately finish the path planning tasks. In comparison with the traditional method of reinforcement learning, IM-GPF-Q algorithm has faster convergence speed, and better self-learning and self-adaptability capabilities.

Keywords: mobile robot; path planning; intrinsic motivation; Q-learning algorithm; gravitational potential field; intelligent development

0 引 言

移动机器人在未知环境中通过不断交互增加经验完善行为选择从而具备高度自治的能力将是移动机器人研究的最终目标[1]。移动机器人研究中的重要课题就是避障和路径规划，依据机器人自身传感器对周围环境的感知情况，它可分成两方面：全局路径规划和局部路径规划[2]。由于后者随着环境的变化复杂性不断增加，因此需要准确的环境模型，但其效率就会大大降低。文献[3]提出拓扑法在路径规划中的应用，利用拓扑特征极

收稿日期：2018-09-28　　**修回日期**：2018-11-07
基金项目：国家自然科学基金(61203343)；河北省自然科学基金(F2018209289)；河北省高等学校科学技术研究青年基金项目(QN2016102，QN2016105)
Project Supported by National Natural Science Foundation of China (61203343), Natural Science Foundation of Hebei Province (F2018209289), Hebei Province Higher Education Science and Technology Research Youth Fund Project (QN2016102, QN2016105)

大地减小了搜索范围；文献[4]在圆锥引力场函数基础上，优化了人工势场法完成了移动机器人的避障和路径规划；文献[5]设计了一种有知识引导的遗传算法且能够有效地提高求解实际路径规划问题的能力；文献[6]将强化学习和模糊逻辑两者结合起来完成了机器人路径规划；文献[7]结合静电势场理论和标量势能理论提高了机器人路径规划的鲁棒性。

机器人的内在动机是机器人在交互过程中自发的对新的环境、新的现象、新的刺激的一种应对机制，其最初概念来源于发展型心理学。传统移动机器人的学习依靠的是开发者对外界环境的设定和建模，一旦环境出现变化，就需要开发者重新进行建模与编程，这就导致机器人自主学习能力低，在线实时性差，制约了其智能程度的提高。针对以上问题，文献[8]经过研究观察生物的自我意识，提出了内在动机（IM）理论思想，并成功运用在机器人对未知环境的主动探索学习中。文献[9]利用内在动机使机器人能自主学会趋光移动技能。文献[10]基于内在情感动机的强化学习框架，成功对智能体进行了情景模拟仿真实验。

目前将内在动机与强化学习算法相结合用以实现机器人路径规划的研究极少，本文受心理学内在动机思想启发，提出一种基于内在动机的Q学习算法，另外，为改进Q学习算法的随机初始化而导致的网络训练迭代次数过多，效率较低的问题，加入引力势场对未知环境进行初始化，该算法以Q学习为框架，利用内在动机信号作为内部奖励，与外部信号生成评价信号引导机器人进行自主学习。通过“感知-内在动机与Q学习-动作-感知”的认知过程，渐进的自主学习，实现移动机器人避障和最优路径规划。最后与传统强化学习算法进行对比，体现了该算法的稳定性和高效性。

1 基于内在动机的引力势场Q学习算法

1.1 控制器结构

为了更好地实现移动机器人自主避障和路径规划，设计一种基于IM-GPF-Q算法的智能控制模型，如图1所示。在工作过程中，机器人的基本行为动作有很多种，基于控制理论的思想，本文旨在研究机器人路径规划。本文提出的算法可以提高机器人自主学习能力，加快学习速度，机器人可以根据执行的动作从环境中获得评价，之后根据评价来调整行为策略，并实现针对路径规划的最佳动作。

总体来说，机器人从环境中每感知一个状态信息，就会产生一个相应动作。随着运行次数的增加，移动机器人逐渐学习到最佳的动作组合以应对各种环境。可以根据所处的环境选择合适的行为动作，实现最优的路径规划。

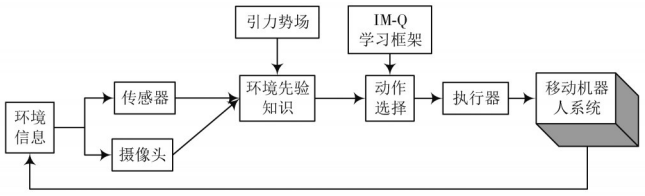

图1 移动机器人智能控制结构模型

Fig. 1 Intelligent control structure model of mobile robot

1.2 内在动机

内在动机理论最初是由心理学领域专家提出，它是运动感觉和认知发育过程中最重要的机制，属于人类行为动机学的一个重要研究课题，同时它在发育心理学中具有深远影响。关于内在动机的思想和理论很多，但是大部分的相关学者都把动机分成两类：内在动机和外在动机。不同心理学家对内在动机具体定义和主要产生原由也有着明显的差别。Alderfer认为人类为适应周围环境不断进行摸索和开发的需要是内在动机的主要组成部分；McGregor提出的Y理论和McClelland提出的成就动机理论主要强调人们为追求成功而在活动工作中产生的动力；而Deci和Ryan等人比较重视个人能力、意愿以及与人之间的关系等三种内在需要；Amabile根据前者对内在动机的研究，给出了组成内在动机的五大要素：事务参与程度、胜任感、自我定位、好奇心和兴趣。综合以上学者和其他一些研究者的观点，内在动机主要是由人们的某些内在精神需要而产生，推动人们不断去发现或探索周围环境，引导其产生好奇心，促使人们自愿地参与到感兴趣的活动中。因此，内在动机主要受个体的兴趣、需要和情感的影响[11-12]。

如图2所示是内在动机机制模型。首先感受器接收外界环境发出的感知信号并将其传送到神经中枢，然后神经中枢中的信号经过在小脑中整合转为由运动神经系统给出运动的控制命令，最后控制命令指导效应器执行动作，进而引起环境的变化。内在动机机制是在不需要对外界环境建立模型的情况下，将外界状态经过整合转变为相应操作，同时把由环境发出的强化信号转化为评价指标，生成取向信息，从而调整和改善对环境执行的动作，最终完成“感知-评价整合-动作-感知”的认知过程，促使生物体逐步拥有高超的运动支配能力。

1.3 Q学习算法

强化学习（Reinforcement Learning），是机器学习方法中的一个重要分支，是多学科多领域交叉的一个产物，它的本质是解决decision making问题，即自动进行决策，并且可以做连续决策，在机器人智能发育、控制及

分析预测等领域有重要应用。图3所示为强化学习的基本原理框架，强化学习是模拟人类学习的过程，将其看作是一个试探评价过程，智能体在环境中选择执行一个动作，环境受其影响状态产生改变，环境同时会产生一个奖励或惩罚的信号反馈给智能体，该信号即为强化信号，智能体通过整合该信号和环境当前的状态再给出下一个动作，依据的原则是使该动作作用于环境后接收到奖励信号的概率增大。智能体执行的动作不仅会改变下一刻环境的状态，还会影响下一刻接收的强化值和最终的强化值。

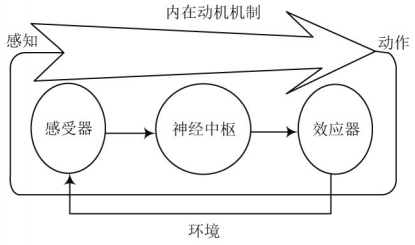

图2　内在动机机制结构模型

Fig. 2　Structural model of intrinsic motivation mechanism

图3　强化学习原理框架

Fig. 3　Principle framework of reinforcement learning

在标准的强化学习框架中，包含智能体(agent)、环境及四要素：

1) 策略(Policy)：策略π是智能体与环境交互时的行为选择依据，$S \to A$是状态空间到动作空间的映射，智能体在状态s_t下根据π选择动作 $a_t + 1 \in A$。

2) 奖惩函数(Reward)：奖惩函数r反映当前环境状态 $s_t \in S$(S为状态空间集)下的动作 $a_t \in A$ (A为动作空间集)对达成目标的贡献度，r值越大代表当前环境下的动作越有利于达成目标。

3) 值函数(Value Function)：值函数V是agent在环境下行动的期望回报，强化学习通过不断试错来迭代值函数，以获得较大期望回报，用于提高高回报值的状态动作选择概率，而降低低回报值的环境状态选择概率。

通过对值函数评估来不断改变对应状态的行动策略，以达到最大化回报。

4) 环境模型(Model of Environment)：强化学习的基本运行环境包括环境状态及其对应的动作映射空间。

本文所用的Q学习算法是从瞬时差分(TD)算法演变而来，是强化学习算法中的经典算法。该算法结合了动态规划与动物心理学知识，从而可以实现具有回报的机器在线学习。该算法利用Markov过程进行建模，迭代得出最优解。

选择策略为：

$$\pi^*(s) = \arg\max Q^*(s,a) \tag{1}$$

迭代计算公式如下：

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \kappa\left[R(s_t,a_t) + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right] \tag{2}$$

式中：s_t表示在t时刻外部环境状态集；a_t表示第t时刻的动作集；$Q(s_t,a_t)$表示在s_t状态下执行动作a_t后所得到的值函数大小；$Q^*(s,a)$是最优值函数；γ是折扣因子；κ是学习因子，且$0 < \kappa < 1$；$R(s_t,a_t)$表示系统在t时刻，外部状态为s_t时所表现出来的外部动作a_t后使状态转移到s_{t+1}后的奖赏信号。

Q学习算法步骤如下：

Step1：初始化$\boldsymbol{Q}(s_t,a_t)$为零矩阵；

Step2：从当前状态s_{t+1}出发并选择执行一个动作a_t；

Step3：获得下一状态s_{t+1}并同时得到奖励信号R；

Step4：根据式(2)更新$\boldsymbol{Q}$值；

Step5：从Step2重复以上动作，直到$\boldsymbol{Q}$矩阵更新完内部所有数值。

2　基于内在动机的引力势场Q学习算法

在经典的Q学习算法中，由于存在白板学习，导致学习速度缓慢，故在学习初期加入引力势场，其策略是往Q值增大的方向移动。因此对于初始状态的值函数表，应当满足距离目标点越近其引力势场越大，故其与距离成反比，这与传统的引力势场有所不同，传统的是与距离成正比，为了保证算法的实时性，暂且不考虑障碍物的斥力势场，同时为减少公式计算的复杂程度，引入参量L用于调整初始状态值的变化范围：

$$V_{\text{init}}(s) = L - D(p_{\text{init}},p_{\text{goal}}) \tag{3}$$

式中：$L \geqslant \sqrt{x^2 + y^2}$，$x$为环境的水平长度，$y$为环境的竖直长度；$D(p_{\text{init}},p_{\text{goal}}) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$为起始点和目标点间的距离。

确定起始点坐标a和目标点坐标b，在初始化状态值函数中以目标点为引力势场中心建立势场，根据目标点的位置作为环境先验信息初始化$V(s)$表，设定初始化后的$V(s)$值均大于等于0。此时算法的动作选择函数为：

$$\pi^*(s) = \arg\max\left[Q(s,a) + \varepsilon V(s)\right] \tag{4}$$

接着将内在动机加入Q学习算法中，假设当前感知的未知环境输入为S：

$$S=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\cdots)$$

式中：$\boldsymbol{x}_t$表示第t时刻的感知输入向量，它的个数是由环境所有的状态决定，此时假设N为内在激励函数，其等于当前t时刻的外部环境感知输入$\boldsymbol{x}_t$与机器人期望输出值$\hat{x}_E$的差，即$N=\|\boldsymbol{x}_t-\hat{x}_E\|$，并将其奖励信号更新为内在动机的奖励机制：

$$R=\xi r_{\text{in}}+\eta r_{\text{ex}}=\xi N+\eta r_{\text{ex}}=\xi\|\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t\|+\eta r_{\text{ex}} \quad (5)$$

式中：r_{in}为内在动机函数；r_{ex}为外在动机函数；ξ和η分别表示r_{in}和r_{ex}的训练加速度因子。此时改进后的IM-GPF-Q学习算法迭代公式为：

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\kappa\left[(\xi\|\boldsymbol{x}_t-\hat{\boldsymbol{x}}_t\|+\eta r_{\text{ex}})+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\right] \quad (6)$$

3 仿真实验和实验结果

为了验证此算法的可行性，运用Matlab R2014a平台进行实验。实验环境为Windows 7，64位操作系统，Intel® Core™ i5-3470 CPU @3.20 GHz处理器，4 GB内存。

实验总共分三组进行，每组设置都包含一个目标点，一个起始点和不同形式的障碍物若干，分别验证本文所提出的IM-GPF-Q算法的有效性，并与典型的Q学习算法进行对比，证明IM-GPF-Q算法的高效性与智能性。

如图4所示，环境中黑色方块为障碍物用来模拟客厅中沙发、茶几和电视机等家具摆放，蓝色五角星是起始点，红色圆点是目标点，小圆圈为机器人所走的路径。从图4可知，改进后的IM-GPF-Q算法能使机器人很好地避开黑色障碍物到达目的地且路径平滑。

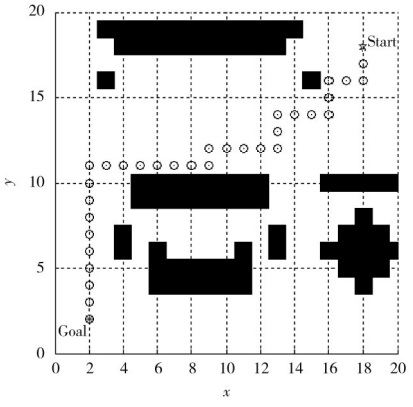

图4 环境A机器人运动轨迹

Fig. 4 Robot motion trajectory in environment A

如图5所示，障碍物设置成具有一定挑战的类似圆形区域，起始点在圆内，目标点在圆外右下方，根据图内路径可知，机器人能成功地避开障碍物抵达目标点，并且所走路径为最优路径。

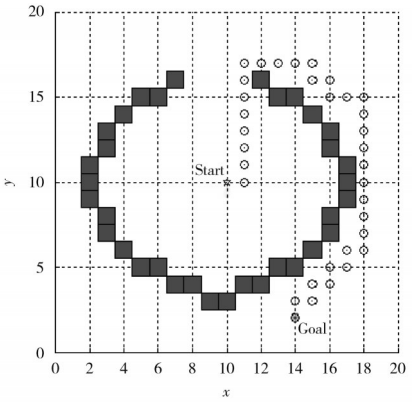

图5 环境B机器人运动轨迹

Fig. 5 Robot motion trajectory in environment B

如图6所示，图中障碍物设置成具有一定陷阱的模式，显然，从起始点到目标点的两条可行路径上障碍物的数量是一样的，但两条路径所需要的步数是不一样的，而机器人可以做出如图6所示的最优路径，证明了所提出的算法可以有效地使机器人到达目标位置。

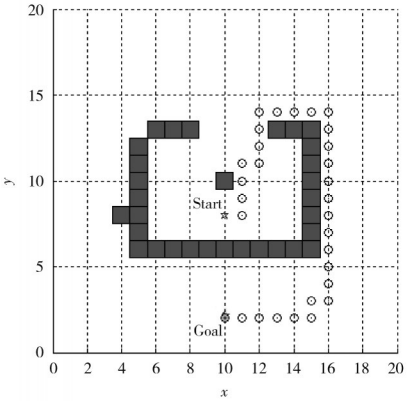

图6 环境C机器人运动轨迹

Fig. 6 Robot motion trajectory in environment C

图7为机器人在环境A中运动时改进算法和未改进算法的标准偏差对比图。

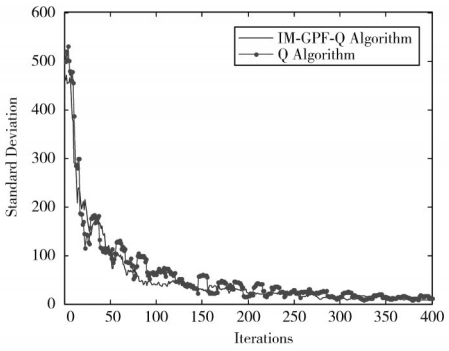

图7 环境A机器人运动标准差

Fig. 7 Standard deviation of robot motion in environment A

由图7可知，随着步数的增加，利用IM-GPF-Q算法进行仿真的轨迹图更快地趋于稳定，说明优化后的算法

在加快学习进程和收敛速度的高效性。实验结果表明，IM-GPF-Q算法在机器人避障和路径规划中得到很好的效果，完成了全局最优路径规划，使机器人自主学习能力和适应环境能力得以提高。

4 结　论

本文提出一种基于内在动机的Q学习算法，即IM-GPF-Q算法，将其应用到移动机器人避障和路径规划中，使其渐进地学会了行为与奖惩信号之间的对应关系，自主学习和适应周围未知环境，并做出相应的动作避开障碍物，按照最优路径到达目标位置，实验证明此算法的可靠性、稳定性和高效性。内在动机的加入加快了Q学习在机器人执行避障和路径规划的效率，节省了大量学习时间。且该算法不仅仅只能运用到机器人避障，还可以在机器人跟踪及导航等方面灵活运用，能实现机器人智能化与自主化。

参 考 文 献

[1] 徐兆辉.移动机器人路径规划技术的现状与发展[J].科技创新与应用，2016(3)：43.
XU Zhaohui. Current status and development of mobile robot path planning technology [J]. Technological innovation and application, 2016(3): 43.

[2] 江杰，任恒靓.基于改进人工势场法的移动机器人路径规划的研究[J].自动化应用，2017(8)：80-81.
JIANG Jie, REN Hengliang. Research on Path Planning of Mobile Robot Based on Improved Artificial Potential Field Method [J]. Automated application, 2017(8): 80-81.

[3] CHOSET H. Simultaneous mapping, path planning, and localization using topological and range sensor information [C]// Proceedings of the 31st International Symposium on Robotics. Ottawa: Canadian federation for robotics, 2000: 299-305.

[4] 樊晓平，李双艳，陈特放.基于新人工势场函数的机器人动态避障规划[J].控制理论和应用，2005，22(5)：703-707.
FAN Xiaoping, LI Shuangyan, CHEN Tefang. Dynamic obstacle avoidance planning of robot based on new artificial potential field function [J]. Control theory and application, 2005, 22(5): 703-707.

[5] 王雪松，高阳，程玉虎.知识引导遗传算法实现机器人路径规划[J].控制与决策，2009，24(7)：1043-1049.
WANG Xuesong, GAO Yang, CHENG Yuhu. Knowledge-guided genetic algorithm for robot path planning [J]. Control and decision, 2009, 24(7): 1043-1049.

[6] 梁泉.未知环境中基于强化学习的移动机器人路径规划[J].机电工程，2012，29(4)：477-481.
LIANG Quan. Mobile robot path planning based on reinforcement learning in unknown environment [J]. Mechanical and electrical engineering, 2012, 29(4): 477-481.

[7] BAYAT F, NAJAFINIA S, ALIYARI M. Mobile robots path planning: Electrostatic potential field approach [J]. Expert systems with applications, 2018, 100: 68-78.

[8] CEDERBORG T, OUDEYER P Y. From language to motor Gavagai: unified imitation learning of multiple linguistic and non-linguistic sensorimotor skills [J]. IEEE transactions on autonomous mental development, 2013, 5(3): 222-239.

[9] 庞涛，阮晓钢，陈静.基于内发动机机制的机器人趋光控制[J].北京工业大学学报，2014，40(1)：32-37.
PANG Tao, RUAN Xiaogang, CHEN Jing. Robotic phototaxis control based on internal engine mechanism [J]. Journal of Beijing University of Technology, 2014, 40(1): 32-37.

[10] 鲁成祥.基于动机的强化学习及其应用研究[D].曲阜：曲阜师范大学，2016.
LU Chengxiang. Motivation-based reinforcement learning and its application research [D]. Qufu: Qufu Normal University, 2016.

[11] KANFER R. Motivation theory and industrial and organizational psychology [M]. Palo Alto: Counsulting Psychologist Press, 1990.

[12] CORDOVA D I, LEPPER M R. Intrinsic motivation and the process of learning [J]. Journal of educational psychology, 2016, 88(4): 715-730.

作者简介：李福进(1957—)，男，河北唐山人，博士，教授，硕士研究生导师，主要研究方向为智能控制与智能仪表。
张俊琴(1992—)，女，河北张家口人，硕士研究生在读，主要研究方向为模式识别与智能装置。
任红格(1979—)，女，河北石家庄人，博士，副教授，主要研究方向为模式识别、智能控制和认知机器人。