基于未知环境状态新定义及知识启发的机器人导航 Q 学习算法①

童小龙, 姚明海, 张灿淋

(浙江工业大学 信息工程学院, 杭州 310023)

摘 要: 由于强大的自主学习能力, 强化学习方法逐渐成为机器人导航问题的研究热点, 但是复杂的未知环境对算法的运行效率和收敛速度提出了考验. 提出一种新的机器人导航Q学习算法, 首先用三个离散的变量来定义环境状态空间, 然后分别设计了两部分奖赏函数, 结合对导航达到目标有利的知识来启发引导机器人的学习过程. 实验在 Simbad 仿真平台上进行, 结果表明本文提出的算法很好地完成了机器人在未知环境中的导航任务, 收敛性能也有其优越性.

关键词: 强化学习; 状态定义; 知识启发; Simbad 平台

A Q-Learning Algorithm for Robot Navigation Based on a New Definition of an Unknown Environment States and Knowledge Heuristic

TONG Xiao-Long, YAO Ming-Hai, ZHANG Can-Lin

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Due to powerful self-learning ability, reinforcement learning has become a research hot spot about robot navigation problems, but the operating efficiency and convergence speed of the algorithm are tried by the the complex unknown environment. A new Q-learning algorithm for robot navigation was proposed in this paper. First, three discrete variables were used to define the space states of the environment, and then two parts of the reward functions were designed, combining the beneficial knowledge for reaching the target to inspire and guide the robot's learning process. The experiment was executed on Simbad simulation platform. The results show that the proposed algorithm is well done in an unknown environment robot navigation task, and has a better convergence speed.

Key words: reinforcement learning; states definition; knowledge heuristic; Simbad platform

未知环境下移动机器人导航主要面临解决两大问题, 即机器人定位和路径规划[1,2]. 定位是指机器人根据所处的环境判断自己的位置和方向, 它需要识别出每一个物体是目标还是障碍物; 路径规划是指机器人需要找到一条从开始位置到目标位置的无碰撞路径, 为此机器人需要运行合适的路径规划算法, 计算出任意两点之间的路径[3]. 移动机器人的路径规划可分为基于地图的全局路径规划和基于传感器的局部路径规划. 针对环境已知的离线全局路径规划方法, 已经取得了大量成果. 基于传感器的局部路径规划是实现机器人在未知环境中探索的重要技术, 很多传统的人工智能算法在这方面做了大量的工作. 随着机器人应用领域的不断拓展, 机器人所面临的任务也越来越复杂,尽管很多情况下研究人员可以对机器人可能执行的重复行为进行预编程,但为实现整体的期望行为而进行行为设计变得越来越困难,设计人员往往不可能事先对机器人的所有行为做出合理的预测,而且当工作环境发生变化时,预先设计好的行为将不再具有适应性[4]. 不管是传统的智能控制方法还是针对特定任务的预编程, 都需要对环境知识和执行任务有比较清楚的了解. 基于强

① 基金项目:国家自然科学基金(61070113)

收稿时间:2013-06-08;收到修改稿时间:2013-07-09

化学习的机器人导航开始成为国内外学者研究的热点，优点主要体现在：无须建立精确的环境模型，简化了智能体(agent)的编程；无须构建环境地图，强化学习可以把避障、路径规划、协作等问题统一解决. 但是传统的强化学习方法应用在机器人导航任务时，由于环境的复杂性和学习任务的多样性，传统的 Q 学习算法往往存在学习时间长、收敛速度慢等问题. 改进的 Q 学习算法主要可以分为以下四类：(1)Q 值更新策略；(2)动作选择策略；(3)Q 值初始化策略；(4)减少状态空间大小策略. 文献[5]采用模拟退火(SA)的 Q 学习算法，利用 Metropolis 准则来平衡探索和利用. 文献[6]引入先验知识指导“探索”，提高强化学习速率. 文献[7]提出了一种新的环境状态空间的定义来减少状态的数量，提高了算法的收敛速度，同时提高了对未知动态环境的适应性. 文献[8]提出一种基于路径引导知识启发的强化学习方法，算法利用在线学习获得的路径知识来指导和加速机器人以后的强化学习过程，以减少机器人学习过程的盲目性.

本文在文献[7]的基础上，模拟人类的推理过程，用机器人与目标或者距离最近的障碍物之间的大概距离和方向来表示环境的状态信息，提出一种机器人探索环境状态空间的定义方法，同时结合对导航达到目标有利的知识来启发引导机器人的学习过程，大大提高了算法的性能和学习速率. 同时相对于很多强化学习的算法都只是在 Matlab 仿真环境中实现，与真实环境差别太大，本文的算法在 Simbad 仿真平台上运行，算法采用 Java 编写，实现机器人在未知环境中的探索，跟真实环境更加贴近，提高了算法的可移植性.

1　Simbad仿真平台介绍

Simbad 能够按照 GPL 开源许可证的方式使用. 它是一个基于 Java 3D 技术、使用 Java 编程语言编写的三维机器人仿真器，因此它可以在 Linux 或其他支持 Java 虚拟机(JVM)的平台上运行. 使用这个仿真器，我们可以创建环境或对环境进行裁减，然后使用各种传感器来开发自己的机器人控制器. 可用的传感器包括视觉传感器(彩色单镜相机)、范围传感器(声波和红外探测器)以及碰撞检测的缓冲等[9].

Simbad 仿真器是个很适合测试智能机器人算法的环境. Simbad 设计用来研究自治机器人环境中的人工智能(AI)算法，它包括了一个 rich GUI(图形用户界面)进行可视化操作，它不但可以对机器人的动作进行可视化，而且还可以从机器人的角度来进行可视化. Simbad 仿真器界面如图 1 所示.

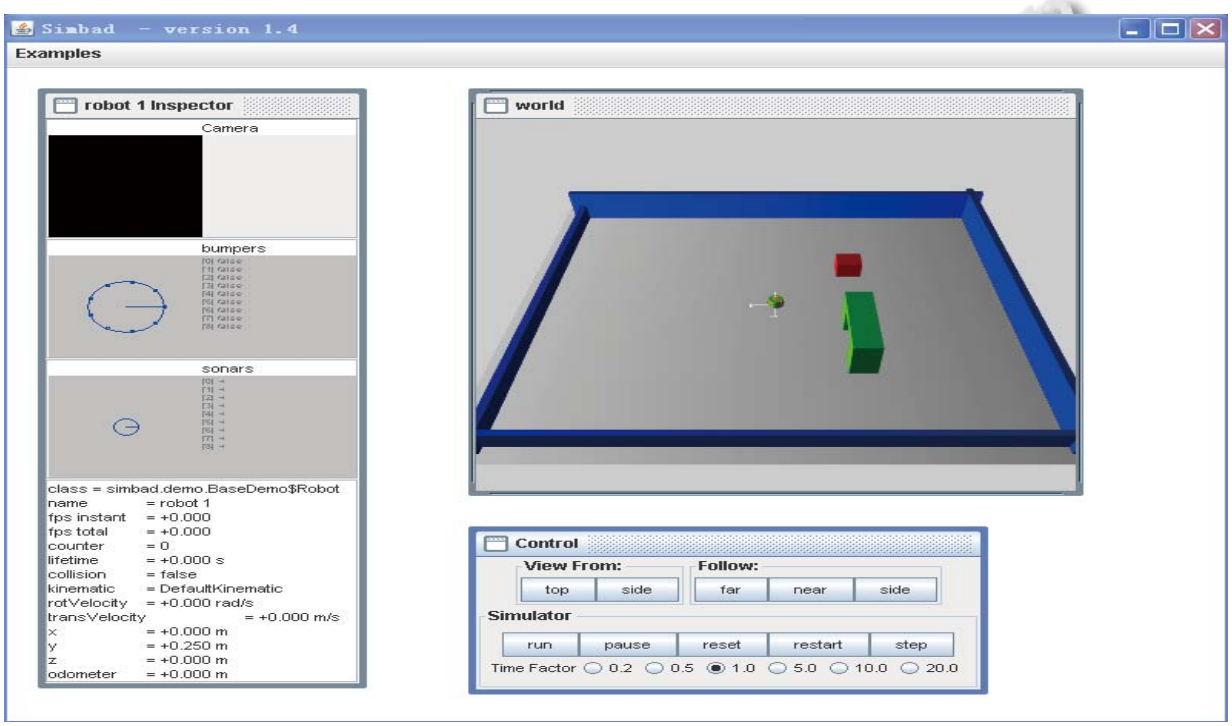

图 1　Simbad 仿真器界面

2 导航算法设计

2.1 经典的 Q 学习算法

Q 学习是一种模型无关的基于瞬时策略的强化学习方法, 是强化学习中应用最广泛的方法之一. Q 学习迭代时采用状态-动作对的奖赏和 Q(*s*,a) 作为估计函数, 因此 agent 的每一次学习迭代时都需要考察每一个行为. 在 Q 学习算法中, Q(*s*,a) 是指在状态 s 执行完动作 a 后获得的累积回报, 它取决于当前的立即回报和期望的延迟回报[10]. 所有状态-动作对的 Q 值存放在一张二维的 Q 表中, 其值在每个时间步被修改一次.

Q 学习的一般步骤如下.

Step 1: 初始化所有的 $Q(s,a)$;

Step2: 循环以下步骤, 直到满足结束条件.

1) 观察当前环境状态, 设为 s;

2) 利用 Q 表按照一定的动作选择策略(例如?-贪心)选择动作 a;

3) 执行该动作 a;

4) 设 r 为在状态 s 执行动作 a 后获得的立即回报;

5) 更新 $Q(s,a)$:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_{t+1}+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\right]$$

同时进入下一状态.

2.2 改进的机器人导航 Q 学习算法

机器人探索的环境中包括各种障碍物, 以及目标物体. 障碍物和目标有可能是静态的, 也有可能是动态的. 在一个未知环境中, 尤其是一个复杂的动态环境中, 机器人的各种状态信息可以通过各种传感器获得, 比如声纳、红外或者视觉传感器等, 合理的状态空间定义和奖赏函数设计都有利于加快 Q 学习算法的收敛速度.

2.2.1 环境状态空间的定义

为了克服复杂未知环境状态空间太大造成的"维数灾难", 不同于直接利用机器人的坐标位置来表示机器人的环境状态信息, 本文提出一种新的环境状态定义方式. 模拟人的推理过程, 当一个人走进一个未知的房间时, 并不关注与障碍物或者目标的具体位置或者确定距离, 只关注与障碍物或者目标的相对方向或者大概的距离.

基于以上的思想, 机器人环境状态用三个状态量来表征, 分别是相对于机器人来说目标所在的区间 Ri、相对于机器人来说距离机器人最近的障碍物所在的区间 Rj、相对于机器人最近障碍物与目标之前的角度 θ 所代表的区间 Gn. 其中 θ 与 Gn 的对应关系定义为当 θ 区间为[0°,45°)时, 区间为 G1, θ 区间为[45°,90°)时, 区间为 G2, 依次类推, 当 θ 区间为[315°,360°)时, 区间为 G8.

如图 2 所示, 当前机器人的环境状态表示为 $s=(R1,R4,G2)$

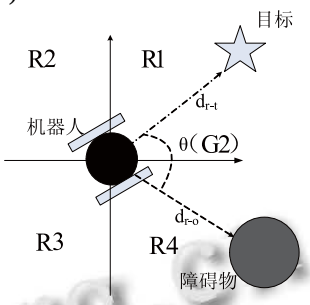

图 2 机器人环境状态描述

这样原来连续的状态空间就用三个离散的状态量表示出来, 大大减少了状态空间的维数.

2.2.2 奖赏函数的设计

奖赏信号可以对智能体执行动作的好坏进行评价, 所以, 奖赏函数设计的好坏直接影响强化学习算法本身的性能和收敛速度.

为了保证算法的收敛性, 同时加快强化学习的速度, 本文设计的奖赏函数由两部分组成, 第一部分是由机器人所处的状态安全等级决定的, 帮助机器人完成对环境空间的探索, 寻找一条合理的无碰路径, 第二部分是为了加速学习过程引入的, 利用机器人前进方向与目标所在位置的关系做为启发学习的知识, 引导机器人向目标靠近.

为了设计第一部分的奖赏函数, 首先定义机器人反映环境状态安全等级的状态集合:

1) SS(Safe States)集合: 机器人在当前的环境状态下几乎没有可能或者很低的概率才能撞到障碍物;

2) NS(Non-Safe States)集合: 机器人有比较高的概率碰到障碍物;

3) WS(Winning State)集合: 机器人达到目标所在位置的状态;

4) FS(Failure State)集合: 机器人与任何障碍物相撞.

根据以上的状态集合设计第一部分的奖赏函数 r_1, 定义如下

$$r_1=\begin{cases}3, S\subset SS\rightarrow WS\text{或者}S\subset NS\rightarrow WS\\ 1, S\subset NS\rightarrow SS\\ -3, S\subset NS\rightarrow FS\\ -1, S\subset SS\rightarrow NS\\ -1, S\subset NS\rightarrow NS\text{并且}d_{r\text{-}o}(n+1)<d_{r\text{-}o}(n)\\ 0, S\subset NS\rightarrow NS\text{并且}d_{r\text{-}o}(n+1)>d_{r\text{-}o}(n)\end{cases}$$

其中 $d_{r-o}(n)$ 和 $d_{r-o}(n+1)$ 分为为机器人当前时刻和下一时刻与距离它位置最近的障碍物的距离.

为了减少机器人学习过程的盲目性，针对机器人在未知环境中导航这一特定任务，第二部分即知识启发引导学习部分的奖赏函数 r_2 设计如下:

$$r_2=\begin{cases}1,\theta\in[0,\frac{\pi}{8})\text{或者}\theta\in[\frac{15\pi}{8},2\pi)\\0,\theta\in[\frac{\pi}{8},\frac{\pi}{4})\text{或者}\theta\in[\frac{7\pi}{4},\frac{15\pi}{8})\\-1,\theta\in[\frac{\pi}{4},\frac{\pi}{2})\text{或者}\theta\in[\frac{3\pi}{2},\frac{7\pi}{4})\\-2,\theta\in[\frac{\pi}{2},\frac{3\pi}{2})\end{cases}$$

其中 θ 是相对于机器人最近障碍物与目标之前的角度.

总的奖赏函数 r 设计为第一部分和第二部分之和，即 $r=r_1+r_2$.

2.2.3 改进的强化学习导航算法

基于以上的环境状态空间定义以及奖赏函数的设计，本文提出的基于未知环境状态空间新定义及知识启发的机器人导航 Q 学习算法流程图如图 3 所示.

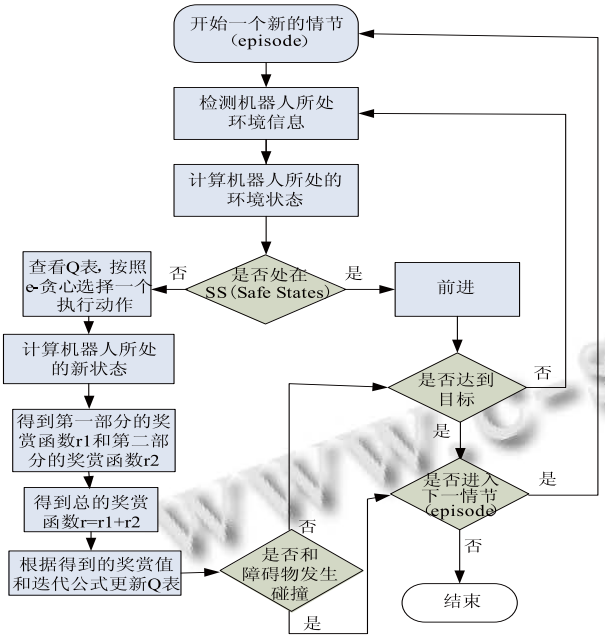

图 3　导航算法流程图

要完成机器人在未知环境中的导航任务，寻找一条最优的无碰路径，首先要通过机器人与环境的不断交互获得状态-动作对的 Q 值. 机器人在环境中的一次探索过程称为情节(episode)，在每一个情节中，机器人利用自身的传感器检测未知环境的信息，得到距离机器人最近的障碍物的距离和方位，首先判断是否处在 SS(Safe States)状态，是的话直接执行一次直行的动作，否则查看 Q 表得到一个合适的动作，动作执行完毕重新计算自身所处的状态，同时利用第一部分和第二部分的奖赏函数得到一个总的奖赏值，更新 Q 表. 直到发生碰撞或者达到目标位置，再开始下一情节.

当算法收敛时，机器人可以直接利用 Q 表的信息得到机器人在特定状态下执行的动作，指导机器人完成导航任务.

3　实验仿真与结果

3.1 实验环境的设置

本文在 Simbad 机器人仿真平台上搭建了机器人导航的未知环境，环境中有箱子和墙等障碍物，同时有一个红色小球作为达到的目标，整个环境是一个三维空间，其中平面大小是 20m×20m. 实验中相关参数设置如下: 学习率 α= 0.1，折扣因子γ= 0.9，最大情节数 maxepisodes = 1000，最大时间步 maxstep = 5000. 编程语言采用 Java，集成开发环境采用 eclipse.

3.2 结果与分析

将机器人置于一个特定的起始位置，每次情节(episode)开始都从这一位置开始，当训练次数达到最大情节数(maxepisodes)的时候结束训练，然后利用训练的结果，即收敛的 Q 表指导机器人再次完成导航任务，得到一条合理的光滑的无碰路径，如图 4 所示.

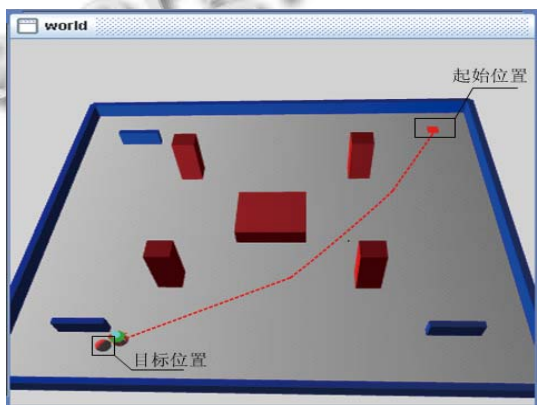

图 4　机器人导航路径

由获得的导航路径可以看出，本文提出的强化学习算法可以非常成功地完成机器人在一个复杂未知环境中的导航任务，获得的路径也接近最优的无碰路径.

为了了解本文算法设计中采用两部分奖赏函数对

导航任务完成情况的影响, 针对同一个探测环境, 同时采用相同的方式定义环境状态空间, 只是在奖赏函数的设计时只有第一部分的奖赏函数, 我们用这样设计的算法与本文提出的算法进行比较. 本文将结合两部分奖赏函数的 Q 学习算法称为算法一, 将只有第一部分奖赏函数的算法称为算法二. 采用算法二时, 同样当训练次数达到最大情节数(maxepisodes)时停止训练, 训练完成后同样可以得到一条机器人的导航路径, 如图 5 所示. 采用算法一完成机器人导航时, 机器人从起始位置运动到目标位置时行走的距离是22.07m, 当采用算法二完成时, 机器人的行走距离经过测量是25.03m, 距离比算法一有所增加, 尽管两者之间的差别不是很明显, 但是也说明算法一得到的无碰路径更加接近最优路径.

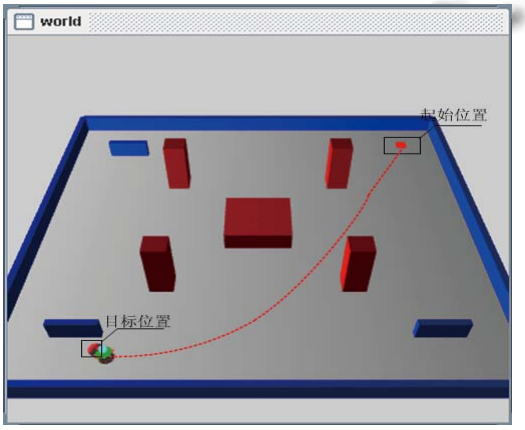

图 5 算法二的机器人导航路径

相对于行走距离的缩短, 算法一与算法二相比, 在收敛速度上的优势更加明显, 表 1 列出了两种算法收敛速度的比较情况.

表 1 算法收敛速度的比较

	第一次达到目标耗时	前 500 次训练中到达运动目标次数	前 100 次达到目标的情节中平均耗时
算法一	12096ms	237 次	1453ms
算法二	37632ms	78 次	1765ms

通过表 1 可以看出, 相对于只采用第一部分奖赏函数的 Q 学习算法, 本文提出的算法由于利用机器人前进方向与目标所在位置的关系做为启发学习的知识, 引导机器人向目标靠近, 减少了机器人对环境盲目探索的概率, 大大加快了学习速度.

4 结语

本文针对机器人在未知环境中的导航问题, 提出一种基于状态空间新定义方式和知识启发学习的机器人导航 Q 学习算法, 首先用三个离散的变量来定义大规模和连续的环境状态, 克服了大空间容易造成的“维数灾难”问题, 然后分别设计了第一部分和第二部分奖赏函数, 在完成对环境空间探索的基础上, 利用导航任务本身提供的知识来引导学习, 减少盲目搜索. 实验结果表明提出的算法很好地完成了机器人未知环境的导航任务, 算法收敛速度快, 解决了探索和利用的平衡问题, 找到的路径也基本上接近最优的无碰路径, 算法有比较好的性能.

参考文献

1 Filliat G, Mayer J. Map based navigation in mobile robots: I. A review of localization strategies. Cognitive System Research, 2003, 4: 243–82.

2 Mayer J, Filliat G. Map based navigation in mobile robots: II. A review of map learning and path planning strategies. Cognitive System Research, 2003, 4: 283–317.

3 Malki A, Lee J, Lee S. Vision based path planning for mobile robot using extrapolated artificial potential field and probabilistic obstacle avoidance. ASME International Mechanical Engineering Congress and Exposition, 2002: 133–9.

4 宋勇.机器人群体行为数学建模与定量分析方法研究[博士学位论文].济南:山东大学,2012.

5 Guo M, Liu Y, Malec J. A new Q-learning algorithm based on the metropolis criterion. IEEE Trans. Syst. ManCybern. B, 2004, 34(5): 2140–2143.

6 Framling K. Guiding exploration by pre-existing knowledge without modifying reward. Neur. Networks, 2007, 20(6): 736–747.

7 Jaradat K, Al-Rousan MAM, Quadan L. Reinforcement based mobile robot navigation in dynamic environment. Robotics and Computer-Integrated Manufacturing, 2011, 27(1): 135–149.

8 刘智斌,曾晓勤.基于路径引导知识启发的强化学习方法. 四川大学学报(工程科学版),2012,5:136–142.

9 Hugues L, Bredeche N. Simbad: an autonomous robot simulation package for education and research. From Animals to Animats 9. Springer Berlin Heidelberg. 2006. 831–842.

10 蒋艳凰,赵强利.机器学习方法.北京:电子工业出版社, 2009.