Computer Science and Application 计算机科学与应用, 2017, 7(7), 671-677
Published Online July 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.77078

Hans汉斯

Dynamic Planning Method Based on Time Delayed Q-Learning

Xia Zhuang

Civil Aviation Flight University of China, Guanghan Sichuan
Email: 16405540@qq.com

Received: Jul. 8th, 2017; accepted: Jul. 22nd, 2017; published: Jul. 25th, 2017

Abstract

Aiming at the unknown environment of the existing robot dynamic planning methods with the slow convergence, a robot planning method based on time delayed Q-Learning is proposed. Firstly, the robot planning is modeled as MDP model, and it is then transferred as the problem which can be solved by reinforcement learning method. Then, the goal function of dynamic planning is defined, and the planning algorithm based on time delayed Q-Learning is proposed. The Q value of every state action pair is initialized to Rmax, so that all the state action pairs are explored, via decreasing the number of updating for Q value, to improve the updating efficiency. The simulation experiment shows: this time delayed Q-Learning algorithm can achieve the path planning of the mobile robot; compared with the other methods, this method has the advantages of good convergence performance and quick convergence rate with big priority, thus it is an effective robot planning method.

Keywords

Robot, Dynamic Planning, Time Delayed Q Learning, Optimal Policy

基于时延Q学习的机器人动态规划方法

庄　夏

中国民用航空飞行学院，四川 广汉
Email: 16405540@qq.com

收稿日期：2017年7月8日；录用日期：2017年7月22日；发布日期：2017年7月25日

摘　要

主要针对现有机器人动态规划方法环境未知，且收敛性能欠佳的缺点，提出了一种基于时延Q学习的机

文章引用: 庄夏. 基于时延 Q 学习的机器人动态规划方法[J]. 计算机科学与应用, 2017, 7(7): 671-677.
https://doi.org/10.12677/csa.2017.77078

器人动态规划方法。首先，对机器人规划进行了MDP建模，将其转换为一个可以通过强化学习解决的问题。然后，定义了规划的目标函数，并描述了基于时延Q学习的机器人规划算法。在该算法中采用Rmax方法来初始化所有状态动作对的Q值，使得所有状态动作对都能被探索到，同时通过时延的Q值来减少Q值更新的次数，从而提高Q值更新的效率。仿真实验表明：文中设计的时延Q学习算法能有效地实现移动机器人的路径规划，较其它算法相比，具有收敛效果好和收敛速度快的优点，具有较大的优越性，是一种有效的机器人动态规划方法。

关键词

机器人，动态规划，时延Q学习，最优策略

1. 引言

机器人动态规划[1]是指在某一个给定的运行空间中，移动机器人通过路径的动态规划来获得一条从初始位置到目标位置的最优路径。环境未知的情况下的机器人路径规划是该领域的研究难点。解决路径规划的主要研究方法包含全局规划法[2] [3] [4]和局部规划方法[5] [6]，全局规划方法主要包括：神经网络和人工势场的方法、遗传算法和粒子群算法等。而局部规划算法主要包括含滚动路径规划和在线视点寻求方法等。全局规划方法主要解决环境已知的问题，在获取了先验知识后才进行规划，其优点是易于收敛；而局部规划法可以用于解决环境未知或部分可观察的问题，但是其难以收敛，且易于陷入局部最优。

为了更好地利用全局规划方法进行机器人最优规划。文献[7]提出了一种多步长的蚁群算法，采用栅格法对机器人的工作环境进行建模，采用启发信息来更新信息素，同时采用最大和最小蚂蚁来限制信息素从而防止算法陷入早熟收敛。文献[8]在经典蚁群算法的基础上调整转移概率，设定信息素的上下界，并在规划过程中，根据动态障碍物运行方向来避免碰撞，并对环境突发情况加入Follow_wall行为进行改进。文献[9]方法提出了一种改进型目标粒子群算法，实现算法在粒子群之间的信息传递，采用SPEA2的环境选择和配对选择策略来使粒子群收敛到Pareto边界，通过自适应原理来对速度权重进行计算来实现算法的全局和局部搜索能力。文献[10]方法设计了一种基于改进混合蛙跳算法实现移动机器人的路径规划，在经典的算法中引入欧氏距离和最优蛙群，通过可调控参数产生新个体来代替随机更新操作，将路径规范问题转换为最小化问题，并根据环境目标和障碍物定义青蛙的适应度，从而实现最优规划。文献[11]提出了一种改进的人工势场法求解机器人规划问题，通过对障碍物影响范围进行分层，使得机器人顺利达到目标，并将改进的势场法用于复杂环境的路径规划和路径选择，从而解决由局部极小导致的陷阱问题。

上述工作均研究了机器人的动态规划算法，但仍然存在一些不足，如在面对环境未知情况下容易陷入局部最优，因此，本文提出了一种基于时延Q学习算法的机器人动态规划算法，首先将机器人规划问题建模一个MDP模型，然后定义了基于时延Q学习算法的机器人规划算法，并通过实验证明了文中方法能有效地解决机器人路径规划问题，尤其是在环境未知的情况下，能实现比同类算法更好的性能，具有收敛速度快和求解效率高的优点，适合于环境未知情况下的机器人的最优路径规划。

2. 背景知识

2.1. MDP

机器人与未知环境的交互可以通过一个有限的马尔科夫决策过程(Markov Decision Process, MDP)来进行建模，一个 MDP 可以建模为一个四元组 $M=\langle X,U,\rho,f\rangle$，其中，$X=\{x_1,x_2,\cdots,x_n\}$ 为所有状态组成的状态空间，任意状态 $x_t\in X$ 表示 Agent 在 t 时刻所处的状态；$U=\{u_1,u_2,\cdots,u_n\}$ 是所有动作构成的动作空间，任意动作 $u_t\in U$ 表示 Agent 在 t 时刻采取的动作；$\rho: X\times U\to\mathbb{R}^n$ 是立即奖赏映射函数，$\rho(x_t,u_t,x_{t+1})$ 表示 Agent 在时刻 t 处在状态 x_t 处，采取动作 u_t 后，迁移到下一个状态 x_{t+1} 处获得的立即奖赏；$f: X\times U\times X\to[0,1]$ 表示状态转移函数，即移动机器人在当前状态 x 处采取动作 u 时转移到 x' 的概率。

2.2. Q 学习算法

Q 学习算法[12] [13]是由 Watkins 等人在 1989 年首次提出的一种离策略学习算法，属于时间差分(Temporal difference, TD)学习算法的一种，在 Q 学习算法中，产生样本的行为策略和用于评估的策略不是同一个策略。行为策略往往采用 ε-greedy 策略，ε-greedy 策略即 Agent 以 1-ε 的概率选择最优动作，而以 ε/|m|的概率选择其他动作，m 为动作的个数。评估策略采用的是 greedy 策略，即贪心策略，在 Q 学习中，动作值函数的更新可以表示为：

$$Q(z_t)=Q(z_t)+\alpha\left(r_{t+1}+\gamma\max_{u_{t+1}}Q(z_{t+1})-Q(z_t)\right) \tag{1}$$

3. 基于时延 Q 学习的机器人动态规划

3.1. 机器人规划的 MDP 模型

MDP 模型需要根据特定的场景进行建模，如对于图 1 所示机器人规划场景。

为了将图 1 所示的机器人规划场景进行建模：

状态空间：状态空间需要被离散化，当 x 轴和 y 轴都被分为 10 等份时，Agent 可以在这 100 个位置中的任意一个。因此，状态空间为：$X=\{1,2,\cdots,100\}$，Agent 在任意状态处的位置可以表示为 (x,y)。

动作空间：Agent 可以采取的动作上下左右四个动作，将动作空间表示为 $U=\{0,1,2,3\}$，0 表示向上的动作，1 表示向下的动作，2 表示向左的动作，3 表示向右的动作。

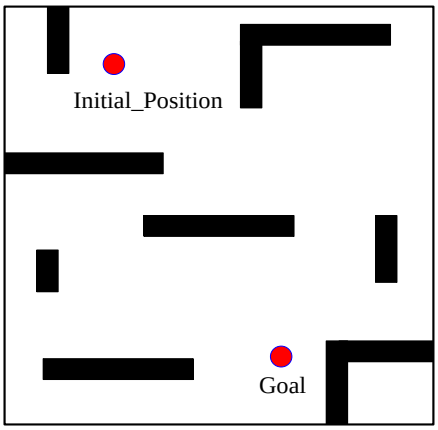

Figure 1. Robot programming experiment scene graph
图 1. 机器人规划实验场景图

奖赏函数：Agent 每移动一步，将会获得立即奖赏为−1，当 agent 到达目标时，能获得一个立即奖赏为 1，因此，奖赏函数可以被定义为：

$$\rho(x_t, y_t, u_t, x_{t+1}, y_{t+1}) = \begin{cases} 0 & (x_{t+1}, y_{t+1}) = \text{“Goal”} \\ -1 & \text{else} \end{cases} \tag{2}$$

迁移函数：Agent 每移动一步，当采用向上的动作 0 时，Agent 将会由当前状态 (x, y) 转移到状态 $(x, y-1)$；当采用向左的动作 2 时，(x, y) 会转移到状态 $(x-1, y)$；当采用向右的动作 3 时，(x, y) 会转移到状态 $(x+1, y)$；当采用向下的动作 1 时，(x, y) 会转移到状态 $(x, y\text{-}1)$。当遇到障碍物或者墙壁时，agent 将会保持在原地不会变化，因此转移函数可以表示为：

$$f(x_t, y_t, 0, x_t, y_t+1) = \begin{cases} 1 & y_t + 1 \neq \{\text{wall}, \text{barrier}\} \\ 0 & \text{else} \end{cases} \quad f(x_t, y_t, 1, x_t, y_t-1) = \begin{cases} 1 & y_t - 1 \neq \{\text{wall}, \text{barrier}\} \\ 0 & \text{else} \end{cases}$$

$$f(x_t, y_t, 2, x_t-1, y_t) = \begin{cases} 1 & x_t - 1 \neq \{\text{wall}, \text{barrier}\} \\ 0 & \text{else} \end{cases} \quad f(x_t, y_t, 2, x_t+1, y_t) = \begin{cases} 1 & x_t + 1 \neq \{\text{wall}, \text{barrier}\} \\ 0 & \text{else} \end{cases} \tag{3}$$

3.2. 机器人规划的目标

强化学习通过最大化长期的累积奖赏来实现状态到动作的最优映射。强化学习算法中通常采用状态值函数或动作值函数来近似表示累积奖赏，动作值函数即 Q 值函数可以表示为：

$$\forall x \in X : Q^h(x,u) = \rho(x,u) + \gamma \sum_{x' \in X} f(x,u,x') V^h(x') \tag{4}$$

最优行为策略 h^* 对应的最优 Q 值函数表示为：

$$\begin{aligned} &\forall x \in X, u \in h(x): \\ &Q^*(x,u) = \rho(x,u) + \gamma \sum_{x' \in X} f(x,u,x') \max_{u' \in U} Q^*(x',u') \end{aligned} \tag{5}$$

移动机器人规划的目标就是使得式(5)最大化。

3.3. 基于时延 Q 学习的机器人最优规划算法

经典的 Q 学习算法仅仅采用 ε-greedy 贪心策略来实现探索和利用的平衡，通常在算法的运行前期采用比较大的 ε 值，实现较大的探索，以尽可能学习到较为精确的 Q 值，而在算法运行的后期，设置比较小的 ε 值，尽可能地利用已经学习到的策略和 Q 值。

时延 Q 学习算法在 Q 学习算法的基础上主要进行下面两方面的改进：

1) 采用 Rmax 来对所有状态动作对的 Q 值设置为 $V_{\max}$，使得 Agent 在初始时刻就开始尽量地探索，因为没有被探索过的状态动作对 (x,u) 对应的 Q 值都初始化为最大值，因为会有更高的概率被探索到。

2) 状态动作对 (x,u) 被访问过 h 次后，才开始更新状态动作对 (x,u) 的 Q 值。

基于时间延 Q 学习的机器人路径规划算法可以描述为：

算法 基于时间延 Q 学习的移动机器人路径规划算法

初始化：动作总数 m，状态动作 n，构建机器人路径规划对应的 MDP 模型，初始化所有状态动作对 (x,u) 的 Q 值 $Q_0(x,u)$，折扣因子 γ，资格迹衰减因子 λ，学习率 α；

步骤 1：状态动作对 (x,u) 的 Q 值 $Q_0(x,u) = V_{\max}$，(x,u) 被访问的次数 $U(x,u) = 0$，$B(x,u) = 0$ 为 (x,u) 更新的初始时刻，$C(x,u) = 0$ 为 (x,u) 的访问次数，$L(x,u) = \text{false}$ 表示是否进行学习，当前时刻 $i = 0$，误差因子 ξ，计数器阈值 h；

步骤 2：观察当前的状态 x_t，然后选择在状态 x_t 执行的动作：

$$u = \arg\max_{u \in U} Q(x_t, u) \tag{6}$$

步骤 3：执行动作 u，得到下一个状态 x_{t+1}，得到立即奖赏 r_{t+1}：

If $B(x,u) \le i$ **then**

$L(x,u) = \text{true}$;

End If

If $L(x,u) = \text{true}$ **then**

If $C(x,u) = 0$ **then**

$B(x,u) = i$

End If

$C(x,u) += 1$

$U(x,u) += r_{t+1} + \gamma \max_{u \in U} Q(x_{t+1}, u)$

End If

If $C(x,u) = h$ **then**

$q = U(x,u)/h$

End If

If $Q(x,u) - q \ge \xi$ **then**

$Q(x,u) = q$

$i^* = i$

Else if $B(x,u) \ge i^*$ **then**

$L(x,u) = \text{true}$

$U(x,u) = 0$

$C(x,u) = 0$

End If

步骤 4： $i = i + 1$，

If $i \le I$

转移到步骤 2 继续执行

Else

转移到步骤 5 继续执行

End If

步骤 5：根据生成的最优状态动作值函数来映射到最优策略：

$$u = \arg\max_{u \in U} Q^*(x_t, u) \tag{7}$$

4. 仿真实验

为了对文中方法进行验证，机器人动态规划系统中机器人的初始状态如图 1 中的 Initial_Position，目标状态为 Goal。

算法中参数初始化为：状态动作对 (x,u) 的 Q 值 $Q_0(x,u) = V_{\max}$，状态动作对 (x,u) 的访问次数 $U(x,u) = 0$，状态动作对 (x,u) 的最后访问时间 $B(x,u) = 0$，状态动作对 (x,u) 的学习标志 $L(x,u) = \text{false}$，迭代次数 $i = 0$，误差因子 $\xi = 0.5$，计数器阈值 $h = 5$。

采用文中算法对机器人规划系统进行建模和仿真，最后得到最优策略，根据最优策略获得的任意一

条从初始状态 Initial_Position 到目标状态 Goal 的路径如图 2 所示。

为了验证文中算法的优越性，将文中方法与经典的 Q-Learning 算法以及文献[11]中基于蛙跳算法的方法进行比较，得到的三种方法的收敛情况如图 3 所示。

从图 3 中可以看出，文中方法不仅具有较快的收敛速度，同时具有最好的收敛解，这表现在文中方法在第 50 个情节处就开始收敛，收敛的解为 13。文献[11]所示的蛙跳算法从算法运行开始，收敛速度一直较另外两种方法更慢，在情节数为 90 时趋于收敛，但收敛的解为局部最优解 22 步。文献[11]采用基于 Q 学习的方法进行最优策略获取，其算法在运行过程中收敛速度较文中方法更慢，但其在情节为 70 时收敛，收敛的时间步为 23，仍然为局部最优解。显然，文中方法具有最快的收敛速度和最好的收敛解。文中方法没有陷入局部最优解是因为在算法中采用了 RMAX 方法来增加探索能力，使得初始状态处所有状态动作处都具有较大的 Q 值，从而使得那些没有访问过的状态动作对能有较大的概率被访问到。文中收敛速度快是因为状态动作对的更新需要根据状态动作对访问的次数来确定，从而使得状态动作更新的效率更高，最终加快了算法的收敛速度。

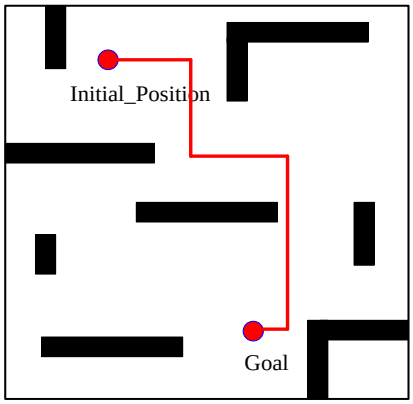

Figure 2. The robot plans the optimal path
图 2. 机器人规划最优路径

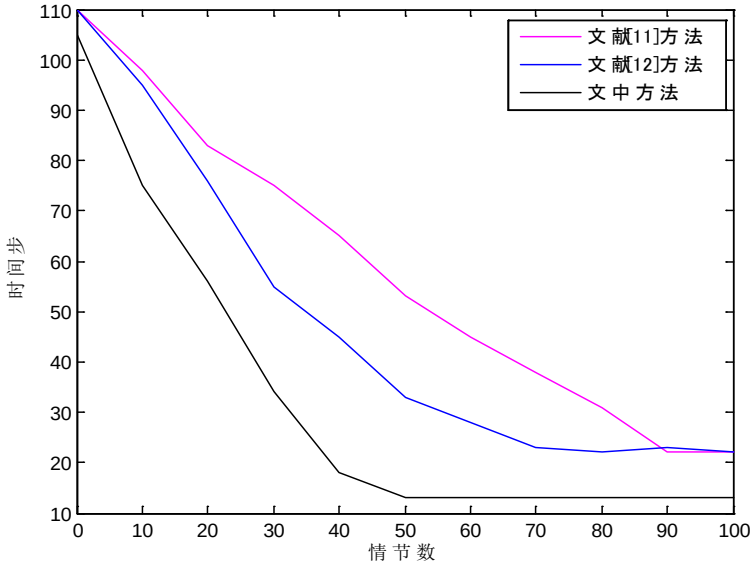

Figure 3. Three methods convergent performance comparison
图 3. 三种方法收敛性能比较

5. 总结

为了实现对机器人动态规划问题进行求解，设计了一种基于时延 Q 学习的机器人动态规划方法。首先，介绍了 MDP 和 Q 学习的相关背景知识，然后将机器人规划问题规划为 MDP 模型，并定义了其相关的目标函数。最后，定义了基于时延的 Q 学习算法来求解机器人行动规划。算法中通过采用 Rmax 初始化所有状态动作对的 Q 值来增加探索，同时采用计数器来记录任意状态动作对被访问的次数，从而加快状态动作对的更新。实验结果表明文中方法能有效获取最优路径，且较其它方法具有收敛速度快和收敛效果好的优点，具有很大的优越性。

基金项目

国家自然科学基金项目资助 U1433126。

参考文献 (References)

[1] Schaal, S. and Atkeson, C. (2010) Learning Control in Robotics. *IEEE Robotics & Automation Magazine*, **17**, 20-29. https://doi.org/10.1109/MRA.2010.936957

[2] 宋勇, 李贻斌, 李彩虹. 移动机器人路径规划强化学习的初始化[J]. 控制理论与应用, 2012, 12(29): 1623-1628.

[3] Bu, Q., Wang, Z. and Tong, X. (2013) An Improved Genetic Algorithm for Searching for Pollution Sources. *Water Science and Engineering*, **6**, 392-401.

[4] Deng, Z.Y. and Chen, C.K. (2006) Mobile Robot Path Planning Based on Improved Genetic Algorithm. *Journal of Chinese Computer Systems*, **27**, 1695-1699.

[5] Liu, C.M., Li, Z.B., Zhen, H., *et al.* (2013) A Reactive Navigation Method of Mobile Robots Based on LSPI and Rolling Windows. *Journal of Central South University* (*Science and Technology*), **44**, 970-977.

[6] Er, M.J. and Zhou, Y. (2008) A Novel Framework for Automatic Generation of Fuzzy Neural Networks. *Neurocomputing*, **71**, 584-591. https://doi.org/10.1016/j.neucom.2007.03.015

[7] 曾明如, 徐小勇, 罗浩, 徐志敏. 多步长蚁群算法的机器人路径规划研究[J]. 小型微型计算机系统, 2016, 2(37): 366-369.

[8] 屈鸿, 黄利伟, 柯星. 动态环境下基于改进蚁群算法的机器人路径规划研究[J]. 电子科技大学学报, 2015, 2(44): 260-265.

[9] 翁理国, 纪壮壮, 夏旻, 王安. 基于改进多目标粒子群算法的机器人路径规划[J]. 系统仿真学报, 2014, 12(26): 2892-2898.

[10] 潘桂彬, 潘丰, 刘国栋. 基于改进混合蛙跳算法的移动机器人路径规划[J]. 计算机应用, 2014, 34(10): 2850-2853.

[11] 温素芳, 郭光耀. 基于改进人工势场法的移动机器人路径规划[J]. 计算机工程与设计, 2015, 10(36): 2818-2822.

[12] Watkins, C.J.C.H. and Dayan, P. (1992) Q-Learning. *Machine Learning*, **8**, 279-292.

[13] Palanisamy, M., Modares, H., Lewis, F.L., *et al.* (2015) Continuous-Time Q-Learning for Infinite-Horizon Discounted Cost Linear Quadratic Regulator Problems. *IEEE Transactions on Cybernetics*, **45**, 165-176. https://doi.org/10.1109/TCYB.2014.2322116