引用格式:段建民,陈强龙. 利用先验知识的 Q-Learning 路径规划算法研究[J]. 电光与控制,2019,26(9):29-33. DUAN J M, CHEN Q L. Prior knowledge based Q-Learning path planning algorithm[J]. Electronics Optics & Control, 2019, 26(9):29-33.

利用先验知识的 *Q*-Learning 路径规划算法研究

段建民, 陈强龙
(北京工业大学,北京 100124)

摘 要: 强化学习中基于马尔可夫决策过程的标准 Q-Learning 算法可以取得较优路径,但是方法存在收敛速度慢及规划效率低等问题,无法直接应用于真实环境。针对此问题,提出一种基于势能场知识的 Q-Learning 移动机器人路径规划算法。通过引入环境的势能值作为搜索启发信息对 Q 值进行初始化,从而在学习初期便能引导移动机器人快速收敛,改变了传统强化学习过程的盲目性,适用于真实环境中直接学习。仿真实验表明,与现有的算法相比,所提算法不仅提高了收敛速度,而且还缩短了学习时间,使得移动机器人能够迅速找到一条较优的无碰撞路径。

关键词: 强化学习; 路径规划; 先验知识; 移动机器人; Q-Learning
中图分类号: TP242 **文献标志码**: A **doi**:10.3969/j.issn.1671-637X.2019.09.007

Prior Knowledge Based *Q*-Learning Path Planning Algorithm

DUAN Jian-min, CHEN Qiang-long
(Beijing University of Technology, Beijing 100124, China)

Abstract: The standard Q-Learning algorithm based on Markov decision process in reinforcement learning can obtain an optimum path, but the method has the shortcomings of slow convergence rate and low planning efficiency, and thus can not be directly applied to the real environment. This paper proposes a Q-Learning path planning algorithm for mobile robots based on the potential energy field knowledge. By introducing the potential energy value into the environment as the search heuristic information to initialize the Q value, the rapid convergence of the mobile robot can be guided in the early stage of learning, and the blindness of the traditional reinforcement learning process is avoided, which makes it suitable for direct learning in a real environment. The simulation result shows that: Compared with existing algorithms, the proposed algorithm not only improves the convergence speed, but also shortens the learning time, which can make the mobile robot find a better collision-free path quickly.

Key words: reinforcement learning; path planning; prior knowledge; mobile robot; Q-Learning

0 引言

如今,随着人工智能研究领域的不断拓深,移动机器人已经被广泛应用于各服务领域,对于移动机器人智能化等性能的要求也逐渐提高。导航技术是实现移动机器人智能化等性能的关键技术,而路径规划是导航技术中的最基本问题[1-3]。路径规划是指移动机器人在有障碍物的环境中,根据最短路径和最短规划时间等评估标准,从初始状态到目标状态找到一条较优的无碰撞路径[4-5]。

当移动机器人不能根据先验知识或通过模仿人来完成任务时,则需要尝试和学习的方法。尝试的成功或失败应该有助于改变"正确"方向的行为,而强化学习(Reinforcement Learning)正是研究此类行为的科学方法之一。针对复杂环境下的移动机器人路径规划问题,许多学者做了大量研究并取得了一定的成果[6-11],其中最具有代表性的是基于 Q-Learning 的路径规划算法。Q-Learning 作为一种监督式学习方法,使移动机器人能够利用学习机制,根据环境的变化规划出一条较优的无碰撞路径。

目前传统的强化学习方法因在学习初始阶段对环境没有先验知识,在机器人导航规划应用中往往存在收敛速度慢、学习时间长等问题[12-14]。针对此问题,很多学者强调了在学习期间利用先验知识进行指导的

收稿日期:2018-09-05 **修回日期**:2019-05-27
基金项目:北京市属高等学校人才强教计划资助项目(038000543117004)
作者简介:段建民(1959—),男,北京人,博士,教授,研究方向为汽车电子控制及智能化技术。

效果,FRAMLING 等[15]通过短时记忆和长时记忆的概念结合,指导先验知识在强化学习任务中的探索,从而提高强化学习速率;OH 等[16]使用模糊规则适当地指定标准的 Q-Learning 初始查找表来加速 Q-Learning;LILLICRAP 等[17]以神经网络来拟合 Q-Learning 中的 Q 函数 $Q(s,a)$,然后采用经验回放和目标网络的方法来改善 Q-Learning 收敛稳定性。

本文在标准 Q-Learning 算法基础上,借鉴先验知识启发搜索的思想,提出一种基于势能场知识的环境状态空间的定义方法。通过引入环境状态空间的先验知识作为搜索启发信息对 Q 值进行初始化,以解决在学习初始阶段的盲目搜索问题,使移动机器人从一开始就带有目的性地选择下一路径点,从而在学习初期便能引导移动机器人快速收敛,提高初始阶段学习效率,且大幅度地提升算法的收敛速度。

1 Q-Learing 算法

1.1 马尔可夫决策过程

马尔可夫决策过程(Markov Decision Process,MDP)可看成是一个元组 $M=<S,A,T,R>$,其中:S 是一组有限的状态集合;$A=\{a_1,\cdots,a_k\}$ 是一组 $k\geqslant 2$ 的动作集合;$T=[P_{sa}(\cdot)\mid s\in S,a\in A]$ 是下一状态的转移概率,$P_{sa}(s')$ 是状态 s 选择动作 a 后转换到状态 s' 的概率;R 表示奖赏函数,例如 $r(s,a,s')$ 是在状态 s 下执行动作 a 导致向状态 s' 过渡时所获得的立即奖赏值,其规定在不同状态 $s\in S$ 中给出的奖励值。

由上述,若有立即奖赏值 r 是根据当前状态 s 和动作 a 获得的,则 MDP 状态转移过程如图 1 所示。

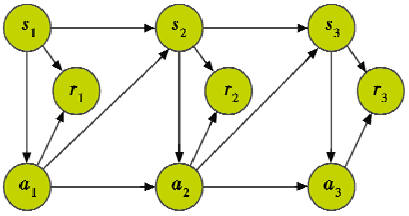


图 1 MDP 状态转移过程

Fig. 1 MDP state transfer process

1.2 标准 Q-Learning 原理

Q-Learning 是由 WATKINS 等提出的一种基于瞬时策略的无模型强化学习方法[18],它采用状态-动作对的值函数 $Q(s,a)$ 作为其迭代过程时的估计函数,所以在每次迭代学习时都要检查每一个动作。Q-Learning 作为一种 TD 控制算法,又与 TD 算法有区别:其采用状态-动作对的奖赏和 $Q(s,a)$ 作为其迭代时的估计函数,而不是 TD 算法中的独立状态值函数 $V(s)$。

在 Q-Learning 中,Q-Learning 最优值函数被定义为在当前状态 s 下执行动作 a,此后按照策略 π 执行最优序列时所获得的折算累积回报,即

$$Q^{\pi}(s,a)=r+\gamma\sum_{s'}P^{a}_{ss'}\max_{a'\in A}Q^{\pi}(s',a') \tag{1}$$

式中:γ 表示折扣因子;$P^{a}_{ss'}$ 表示状态 s 选择动作 a 后转换到状态 s' 的概率。

Q-Learning 的过程如下:移动机器人先在状态 s 下在所有可能的动作中选择动作 a 并执行,再根据获得动作 a 的立即奖赏值以及接收当前的状态动作值的估计来评估动作的结果。通过重复所有状态下的所有动作,移动机器人就可以通过判断长期折扣回报来学习总体上的最佳行为。Q 函数的值最终将收敛到最优值,其更新式为

$$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'\in A}Q^{\pi}(s',a')-Q(s,a)] \tag{2}$$

式中:α 表示学习率;$Q(s,a)$ 表示在状态 s 处执行动作 a 迭代时的估计值函数。

由文献[19]可知,满足一定条件下,对于任意给定的状态 s 及动作 a,在第 k 次($k\to\infty$)进行迭代更新的值函数最终将以概率 1 收敛于函数 $Q(s,a)$。

2 基于势能场知识的 Q-Learning 移动机器人路径规划算法

在复杂的未知环境中,不能直接利用移动机器人的坐标位置来表示当前的环境状态信息,避免因状态空间增大而出现维数灾难。本文提出一种基于势能场新的环境状态定义方式,在 Q-Learning 初始阶段引入对环境的先验知识,通过引入环境状态空间的先验知识(即势能值)作为搜索启发信息对各状态的 Q 值进行初始化,使得 Q-Learning 在学习初期即有一定的目标性,缩短学习时间,从而大幅度提高算法速率。

2.1 基于势能场知识的初始化状态值函数

人工势场法的基本原理是将移动机器人在环境中的运动视为在虚拟的人工受力场中的运动。在目标点附近产生一个引力势场,移动机器人在任意位置上都受到来自目标点的引力吸引,引导移动机器人朝向其运动;而在障碍物附近产生一个斥力势场,障碍物则对移动机器人产生斥力,阻止移动机器人向其靠近,避免发生碰撞。其势场函数如下所述。

引力势场函数为

$$U_{att}(s)=\frac{1}{2}k_{att}\rho^{2}_{goal}(s) \tag{3}$$

斥力势场函数为

$$U_{rep}(s)=\begin{cases}\dfrac{1}{2}k_{rep}\left(\dfrac{1}{\rho_{obs}(s)}-\dfrac{1}{\rho_0}\right)^2 & \rho_{obs}\leqslant\rho_0\\ 0 & \rho_{obs}>\rho_0\end{cases} \tag{4}$$

式中:k_{att},k_{rep} 分别表示引力和斥力比例系数;$\rho_{goal}(s)=$

$\| s - s_{goal} \|$ 和 $\rho_{obs}(s) = \| s - s_{obs} \|$ 分别表示移动机器人与目标点、障碍物之间的欧氏距离；ρ_0 表示障碍物影响系数。

Q-Learning 算法是一种基于瞬时策略的无模型强化学习算法，在学习过程中，其行为策略是向 Q 值递增的方向移动，因此，应该在移动机器人状态空间生成一个新势能场，该新势能场中，每一个状态势能值倒数代表每一个状态所获得的最大折算累积回报值，并满足距离目标点越近其引力势能场 $U_{att}(s)$ 越大，且使得障碍物区域势能值为零，目标点的势能值为全局最大的条件。为了保证算法的实时性，不考虑障碍物的斥力势场，有 $U_{rep}(s) = 0$，同时也能减小公式计算复杂度。根据上述关系式有

$$U(s) = U_{att}(s) + U_{rep}(s) = U_{att}(s) \tag{5}$$

$$W(s') = \begin{cases} 0 & s'\text{为障碍物} \\ 2 & s'\text{为目标点} \\ |1/U(s')| & s'\text{为其他} \end{cases} \tag{6}$$

其中：s' 为移动机器人在当前状态 s 下、在所有可能的动作中选取一个动作 a 更新后的状态；$|U(s')|$ 为由已知的环境信息构成的人工势能场中状态 s' 的势能值；$W(s')$ 为根据最优策略得到状态 s' 所能获得的最大累积收益。对于较复杂环境的 Q-Learning 算法路径规划，动作状态空间庞大，迭代速度慢，本文只考虑引入目标点改进引力势场初始化状态值函数来提供初始目标位置，移动机器人具有目标趋向性，能迅速朝目标方向移动，而最大累积收益 $W(s')$ 不考虑障碍物的斥力势场，可减小公式计算复杂度，同时，算法在实时优化过程中由于采用随机选择策略保证其不陷入局部最优解，不会出现移动机器人与障碍物碰撞或穿越的现象。

由上述可知，状态 - 动作对的初始 Q 值可定义为在当前状态 s 下、在所有可能的动作中选取动作 a 所获得的立即奖赏值 r 以及状态 s' 的最大累积回报两者之和，即有

$$Q(s,a) = r + \gamma W(s') \tag{7}$$

式中，$Q(s,a)$ 表示状态 - 动作对 (s,a) 的初始 Q 值。

2.2　基于势能场知识的 Q-Learning 算法步骤

1）确定起始点坐标（start）和目标点坐标（goal），在初始化状态值函数中以目标点为势场中心建立引力势能场，将势能值作为环境先验信息初始化 $W(s)$ 表；

2）将在势能场中后继状态 s' 的势能值定义为后继状态 s' 的最大折算累积回报，即

$$W(s') = \begin{cases} 0 & s'\text{为障碍物区域点} \\ 2 & s'\text{为目标点} \\ |1/U(s')| & s'\text{为其他} \end{cases};$$

3）利用改进的环境状态值函数来更新状态 - 动作值函数表，并对 Q 值初始化，即 $Q(s,a) = r + \gamma W(s')$，$r = \begin{cases} -0.2 & s'\text{为障碍物区域点} \\ 1 & s'\text{为目标点} \\ -0.1 & s'\text{为其他} \end{cases}$；

4）对于每一次尝试执行循环，并初始化当前状态 s；

5）对于每一次尝试中的每一次迭代执行循环；

6）根据随机选择策略从当前状态 s 下所有可能的动作中选取动作 a；

7）移动机器人通过执行选取的动作 a，将当前环境状态 s 更新转移至状态 s'，然后反馈所获得的立即奖赏值 r 给移动机器人；

8）观察更新后的新状态 s'；

9）根据所获得的立即奖赏值 r 更新状态 - 动作对的 $Q(s,a)$ 值，也就是 $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in A} Q^{\pi}(s',a') - Q(s,a)]$；

10）判断 s' 是否为移动机器人的目标状态，如果不是，则结束本次迭代学习过程，$iteration + 1$，继续返回到 6），否则返回到 11）；

11）结束该次尝试学习，$trials + 1$，并返回到 4），继续进行下一次尝试学习；

12）最后判断移动机器人是否到达目标，或学习系统是否已经达到设定的最大尝试次数，两个条件中只要有一个满足，则结束整个学习过程。

上述算法的学习过程由尝试和迭代组成，其中：尝试即学习周期，指一个从初始状态到目标状态的学习过程；迭代指每次尝试中选择一个动作及其具体执行的实现过程。

3　仿真实验与结果分析

3.1　实验环境

通过仿真实验来验证引入先验知识后算法的有效性和先进性。编译工具采用 Matlab（R2011a），实验采用图 2 所示的由 20 × 20 个方格组成的栅格环境。以其左下角为坐标原点，建立一个以水平方向为 X 轴，竖直方向为 Y 轴的坐标系。

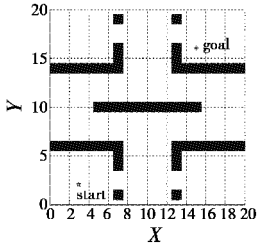


图 2　栅格环境

Fig. 2　Gridding environment

图 2 中，蓝色五角星 start 点表示移动机器人所处的起始位置，红色星形 goal 点表示目标位置，白色部分

为移动机器人的自由活动区域，黑色的实心方块为移动机器人无法穿越的障碍物区域，移动机器人动作空间集合 A 包含有前进、后退、左移、右移 4 个动作。

3.2 实验分析

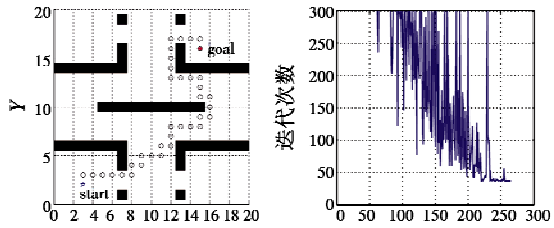
在仿真实验中，Q-Learning 算法所有学习过程的实验参数及关键参数值如表 1 所示。

表 1 仿真中的参数设置

Table 1 Parameters used in simulation

参数名称	参数值	参数名称	参数值
尝试次数	500	学习率 α	0.30
迭代次数	300	折扣因子 γ	0.95
迭代次数标准误差 δ	0.25	人工势场引力比例系数 k_{att}	2.00

标准 Q-Learning 算法的路径规划仿真结果如图 3 所示，首先初始化 Q 值为 0，然后利用 Matlab 在图 1 所示的 20×20 栅格环境中对标准 Q-Learning 算法进行路径规划仿真。



a 基于标准Q-Learning的移动机器人路径规划　b 标准Q-Learning算法收敛过程

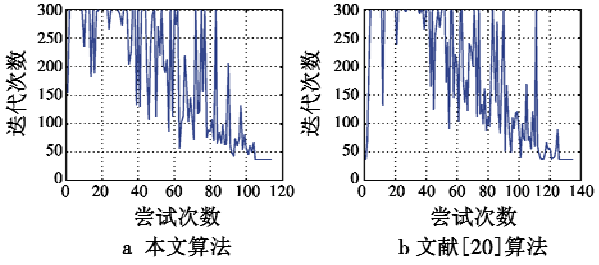
图 3 基于标准 Q-Learning 算法的移动机器人路径规划仿真结果

Fig. 3 Simulation results of mobile robot path planning based on standard Q-Learning algorithm

移动机器人从起始位置(3,2)开始，能够找到一条如图 3a 所示的最短路径并到达目标位置(15,16)，该算法的收敛过程如图 3b 所示，由图 3b 可知，传统算法在经过 265 次尝试后最终收敛于目标，但在学习初始阶段，移动机器人在设置的最大迭代次数范围内无法找到路径到达目标位置，这是因为在学习初始阶段 Q 值被初始化为 0，导致没有先验知识的移动机器人只能随机地选取动作，进而影响移动机器人在学习初始阶段规划效率，算法的收敛速度也明显变慢。

为了验证本文所提改进 Q-Learning 算法的优越性，在相同栅格环境地图下进行实验验证。本文改进后的 Q-Learning 算法收敛过程仿真结果如图 4a 所示，而图 4b 是在相同环境下文献[20]中提出的基于监督神经 Q-Learning(SNQL)算法的仿真结果。

图 4a 为基于势能场知识的 Q-Learning 移动机器人强化学习收敛过程，通过仿真实验发现，改进后算法在经过 105 次尝试后最终收敛于目标，而且迭代次数在收敛前稳步减少。在学习初始阶段，移动机器人经过十几次尝试后，基本上都能够在设置的最大迭代次数范围内找到目标点，表明改进算法明显改善学习过程中的收敛速度。与图 3b 标准 Q-Learning 算法的收敛过程比较可知，通过引入先验知识可以改善 Q-Learning 算法的数据传递滞后性，在学习初期引导移动机器人快速收敛，并缩短学习时间，从而提高收敛速度。



a 本文算法　b 文献[20]算法

图 4 两种改进的 Q-Learning 算法收敛过程

Fig. 4 Convergence process of two improved Q-Learning algorithms

通过图 4b 可以看出，算法在经过 126 次尝试后才能够最终收敛于目标，在学习初始阶段，移动机器人在设置的最大迭代次数范围内基本上无法到达目标位置(前 35 次尝试)，但移动机器人在初始阶段前 5 次尝试能够找到路径并到达目标点，以后的 30 次尝试中系统却无法找到路径抵达目标位置，说明该算法初始阶段的收敛稳定性较差。与图 4a 仿真结果比较可知，本文提出的改进算法能更加明显地提高算法学习初始阶段的学习效率，改善机器人路径规划强化学习算法的性能。

移动机器人每次尝试学习的迭代次数直接体现了当前学习的效果，为了能更加直观地突出改进算法的优越性，本文通过对学习过程中每次尝试的连续 10 次迭代次数的标准差进行计算，分析比较上述 3 种 Q-Learning 算法的学习收敛速度以及稳定性，其计算式为

$$\delta_k = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(n_{i+k-1} - \bar{n})^2} \quad k = 1,2,3,\cdots \quad (8)$$

式中：δ_k 表示第 k 个连续 10 次迭代次数的标准差；n_{i+k-1} 表示第 $i+k-1$ 次迭代次数；$\bar{n}$ 表示连续 10 次迭代次数的平均值。利用 Matlab 作 3 种 Q-Learning 算法标准差的演化过程对比结果如图 5 所示。

从图 5 中蓝色曲线可看出，在学习初始阶段基于标准的 Q-Learning 算法的标准差为零，在设置的最大迭代次数范围内无法到达目标位置。文献[20]算法和本文提出的基于势能场知识的 Q-Learning 算法在学习初始阶段标准差大于零，说明移动机器人在学习初始阶段已经能够到达目标点；而且在学习的后期阶段，基于势能场知识的 Q-Learning 算法的标准差曲线相对于标准 Q-Learning 算法的标准差曲线更平滑，收敛速度更快。由图 4 中曲线也能看出，基于势能场知识进

行路径搜索的算法迭代速度明显快于文献[20]算法，且文献[20]算法在初始阶段的收敛稳定性较差，从而表明，基于势能场的环境先验知识的 Q-Learning 算法使得移动机器人在整个训练过程中不会移动至陷阱区域，且更加明显地提高算法的规划效率。

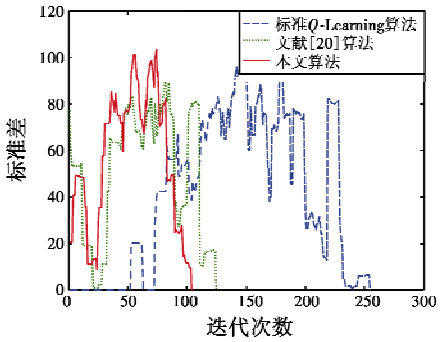


图 5　迭代次数标准差

Fig. 5　Standard deviation of iteration times

4　结束语

在强化学习中，基于马尔可夫决策过程的标准 Q-Learning 算法可以取得较优路径，但移动机器人在进行路径规划时，由于标准的 Q-Learning 算法缺乏对环境的先验知识，导致在学习过程中训练速度变慢，迭代效率变低，所以标准的 Q-Learning 算法很难直接应用。本文提出了基于势能场知识的 Q-Learning 算法，该算法将引力势场势能值作为环境先验信息初始化 Q 值，移动机器人利用该算法进行路径规划得到更快的收敛速度和更优的移动路径，同时，通过初始化可避免对障碍物的试错探索，减小了移动体有效状态空间，在明显少于传统强化学习的训练次数情况下达到更好的路径搜索效果，具有更快的收敛速度和更好的寻优能力。

参 考 文 献

[1]　谈自忠. 机器人学与自动化的未来发展趋势[J]. 中国科学院院刊，2015，30(6)：772-774.

[2]　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.

[3]　胡洋洋. 移动机器人楼层内定位与导航研究[D]. 南京：南京理工大学，2017.

[4]　李斯定. 基于增强学习的移动机器人动态路径规划算法研究[D]. 长沙：国防科学技术大学，2015.

[5]　王钦钊，程金勇，李小龙. 复杂环境下机器人路径规划方法研究[J]. 计算机仿真，2017，34(10)：296-300.

[6]　REN H G, YIN R J, LI F J, et al. Research on Q-ELM algorithm in robot path planning[C]//Control and Decision Conference, IEEE, 2016：5975-5979.

[7]　TAI L, LIU M. A robot exploration strategy based on Q-Learning network[C]//International Conference on Real-Time Computing and Robotics, IEEE, 2016：57-62.

[8]　于乃功，默凡凡. 基于深度自动编码器与 Q 学习的移动机器人路径规划方法[J]. 北京工业大学学报，2016，42(5)：668-673.

[9]　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)：1623-1628.

[10]　ZHANG Y F, LI W L, DE SILVA C W. RSMDP-based robust Q-Learning for optimal path planning in a dynamic environment[J]. IAES International Journal of Robotics & Automation, 2014, 31(4)：290-300.

[11]　KLIDBARY S H, SHOURAKI S B, KOURABBASLOU S S. Path planning of modular robots on various terrains using Q-Learning versus optimization algorithms[J]. Intelligent Service Robotics, 2017, 10(2)：121-136.

[12]　GASKETT C. Q-Learning for robot control [D]. Canberra：The Australian National University, 2002.

[13]　DUNG L T, KOMEDA T, TAKAGI M. Reinforcement learning for POMDP using state classification [J]. Applied Artificial Intelligence, 2008, 22(7/8)：761-779.

[14]　马朋委. Q-Learning 强化学习算法的改进及应用研究[D]. 合肥：安徽理工大学，2016.

[15]　FRAMLING K. Guiding exploration by pre-existing knowledge without modifying reward [J]. Neural Networks, 2007, 20(6)：736-747.

[16]　OH C H, NAKASHIMA T, ISHIBUCHI H. Initialization of Q-values by fuzzy rules for accelerating Q-Learning[C]//Proceedings of the IEEE World Congress on Computational Intelligence, 2002：2051-2056.

[17]　LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6)：1-14.

[18]　WATKINS C, CHRISTDPHER J, DAYAN P. Q-Learning [J]. Machine Learning, 1992, 8(1)：279-292.

[19]　SUTTON R S, BARTO A G. Reinforcement learning：an introduction[M]. Cambridge：The MIT Press, 1998.

[20]　LIN L X, XIE H B, ZHANG D B, et al. Supervised neural Q-Learning based motion control for bionic underwater robots[J]. Journal of Bionic Engineering, 2010, 7(s)：177-184.