DOI:10.19557/j.cnki.1001-9944.2019.09.001

结合神经网络和 $Q(\lambda)$ -learning 的路径规划方法

王　健[1],张平陆[2],赵忠英[1],程晓鹏[1]

(1.沈阳新松机器人自动化股份有限公司 特种机器人 BG,沈阳 110169;2.沈阳科技学院 机械与交通工程系,沈阳 110167)

摘要:Q-learning是一种经典的增强学习算法,简单易用且不需要环境模型;广泛应用于移动机器人路径规划。但在状态空间和动作空间较大时,经典的Q-learning算法存在学习效率低、收敛速度慢,容易陷入局部最优解等问题。通过引入神经网络模型,利用地图信息计算状态势值,从而优化了设计奖励函数。合理奖励函数为$Q(\lambda)$-learning算法提供了先验知识,避免训练中的盲目搜索,同时奖励函数激励避免了陷入局部最优解。仿真试验表明,改进的路径规划方法在收敛速度方面有很大的提升,训练得到的路径为全局最优。

关键词:路径规划;神经网络;强化学习;移动机器人;奖励函数

中图分类号:TP24;TP18　　**文献标志码**:A　　**文章编号**:1001-9944(2019)09-0001-04

Path Planning Method Based on Neural Network and $Q(\lambda)$-learning

WANG Jian[1],ZHANG Ping-lu[2],ZHAO Zhong-ying[1],CHENG Xiao-peng[1]

(1.Special Robot BG,Shenyang SIASUN Robot & Automation Co.,Ltd.,Shenyang 110169,China;2.Department of Mechanical and Traffic Engineering,Shenyang Institute of Science and Technology,Shenyang 110167,China)

Abstract:*Q*-learning is a classical reinforcement learning algorithm,which is simple to use and does not need environment model. It is widely used in mobile robot path planning. However,when the state space and action space are large,the classical *Q*-learning algorithm has the problems of low learning efficiency,slow convergence speed and easy to fall into local optimal solution. By introducing the neural network model and using map information to calculate the state potential value,the design reward function is optimized. Reasonable reward function provides prior knowledge for $Q(\lambda)$-learning algorithm,avoiding blind search in training,and reward function incentive avoids falling into local optimal solution. The simulation results show that the improved path planning method improves the convergence speed greatly,and the trained path is globally optimal.

Key words:path planning;neural network;reinforcement learning;mobile robot;reward function

路径规划是移动机器人的一项重要功能,用于引导移动机器人在地图中自主运动。路径规划的优劣直接影响移动机器人的运动效率、机器损耗和工作效率。与其它机器学习方法不同,增强学习方法无需监督信号,而是通过智能体与环境之间的信息交互进行“试错”,以极大化评价反馈信号为目标,通过学习得到最优或次优的搜索策略。总体来说,增强学习的主要目标就是将状态映射到动作的同时,最大化期望回报。

随着增强学习理论和算法的不断发展与成熟,应用增强学习方法解决移动机器人路径规划问题正成为路径规划的研究热点[1]。文献[2]提出了改进的 *Q*-learning 方法用于解决单个机器人的路径规划问题,通过引入标志位减少了学习过程的收敛时

收稿日期:2019-05-31;修订日期:2019-07-23

作者简介:王健(1986—),男,硕士,工程师,研究方向为机器人控制。

间，提高了算法的效率；文献[3]提出用神经网络模型来解决最短路径规划问题；文献[4]提出基于分层强化学习的路径规划方法；文献[5]对基于神经网络和 POS 的机器人路径规划方法做了较为深入的研究。强化学习是目前机器学习中富有挑战性和广泛应用前景的研究领域之一[6]。

目前，传统的强化学习方法在移动机器人路径规划应用中，由于其学习初始阶段对环境没有先验知识，往往存在收敛速度慢、学习时间长等问题。故在此通过引入神经网络方法，对传统 Q-learning 算法进行改进，优化设计奖励函数，提出了基于神经网络的改进 Q-learning 学习算法。

1 $Q(\lambda)$-learning 算法

Q-learning 是一种与模型无关的强化学习方法，在环境未知条件下，通过不断试错和探索，对所有可能的状态和动作进行多次尝试，采用数值迭代方法逼近最优解。

它以状态-动作对应的 $Q(s,a)$ 为估计函数，逐渐减小相邻状态间 Q 值估计的差异达到收敛条件，即

$$Q^*(s,a)=R(s,a)+\gamma\sum T(s,a,s')\max Q^*(s',a'),\quad s,s'\in S;a,a'\in A \tag{1}$$

式中：S 为状态集；A 为动作集；$T(s,a,a')$ 为状态 s 下执行动作 a 后转换到状态 s' 的概率；$R(s,a)$ 为状态 s 下执行动作 a 的奖励；γ 为折扣因子。寻找最优 Q 值 $Q^*(s,a)$ 的搜索策略为

$$\pi^*(s)=\arg\max Q^*(s,a) \tag{2}$$

更新公式为

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[R_{t+1}+\gamma\max_a Q(s_{t+1},a')-Q(s_t,a_t)] \tag{3}$$

在 Q-learning 算法中引入迹的思想，能够记录状态被访问的次数，在更新前一时刻的状态值函数时，也能对之前的状态值函数进行更新，即 $Q(\lambda)$-learning 算法。其更新公式为

$$e_t(s,a)=I_{ss_t}I_{aa_t}+\begin{cases}\gamma\lambda e_{t-1}(s,a) & Q_{t-1}(s_t,a_t)=\max_a Q_{t-1}(s_t,a)\\ 0 & \text{其他}\end{cases} \tag{4}$$

式中：I_{ss_t} 和 I_{aa_t} 为指数函数，如果 $s=s_t$ 则值为 1，反之为 0。误差项为

$$\delta_t=R_{t+1}+\gamma\max_a Q_t(s_{t+1},a)-Q_t(s_t,a_t) \tag{5}$$

更新动作公式为

$$Q_{t+1}(s,a)=Q_{t+1}(s,a)+\alpha\delta_t e_t(s,a) \tag{6}$$

搜索函数为

$$\pi(s)=\begin{cases}\arg\max_a Q(s,a) & \sigma>\varepsilon\\ a\sim A(s) & \sigma\leqslant\varepsilon\end{cases} \tag{7}$$

式中：σ 为 0~1 之间的随机数；ε 为探索因子，为 0~1 之间的数。

2 受生物启发的神经网络方法

受生物神经系统中 Hodgkin 和 Huxley 细胞膜模型[7]与 Grossberg 分流细胞模型[8]的启发，文献[9]提出了受生物启发的神经网络方法，用于解决移动机器人路径规划问题。该神经网络方法状态方程为

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=-Ax_i+(B-x_i)\left([I_i]^+\sum_{j=1}^{k}+\omega_{ij}[x_j]^+\right)-(D+x_i)[I_i]^- \tag{8}$$

其中

$$I_i=\begin{cases}E, & \text{目标位置}\\ -E, & \text{障碍物位置}\\ 0, & \text{其他}\end{cases}$$

式中：x_i 为第 i 个神经元的神经活动（神经元细胞膜的电势）；A，B，D 分别为被动衰减率、神经活动的上限和下限，均为非负常数；$[I_i]^+\sum_{j=1}^{k}+\omega_{ij}[x_j]^+$ 和 $[I_i]^-$ 为第 i 个神经元的刺激性、抑制性输入；I_i 为第 i 个神经元内部输入；E 为正整数，且 $E>>B$；$[\,]^-$，$[\,]^+$ 分别为取正、取负函数。该取正、取负函数的功能为

$$[a]^+=\max(a,0),[a]^-=\max(-a,0) \tag{9}$$

状态方程(8)中，第 i 个神经元与第 j 个神经元之间的权值连接 ω_{ij} 为一个距离函数。其表达式为

$$d_{ij}=|q_i-q_j|,\omega_{ij}=f(d_{ij}) \tag{10}$$

其中

$$f(a)=\begin{cases}\mu/a, & 0<a<r_0\\ 0, & a\geqslant r_0\end{cases}$$

式中：d_{ij} 为状态 q_i 和 q_j 之间的欧氏距离；μ 和 r_0 为正整数。

该神经网络方法不需要学习过程，根据神经元细胞之间的信息传递，可以求出神经元细胞所在状态的势值函数。通过实时更新势值函数，移动机器

人从初始位置沿着势值增大的方向到达目标位置，从而得到规划路径。

3 结合神经网络和 $Q(\lambda)$-learning 算法

3.1 优化设计奖励函数

在 Q-learning 算法中，奖励 $R(s,a)$ 表示状态 s 下执行动作 a 得到的奖励。奖励值的大小直接影响动作选择的正确性和误差传递的效率。因此，奖励函数设计的好坏，直接影响算法的收敛速度和最优解的质量。

传统的 Q-learning 算法，一般将目标状态的奖励设为很大的正整数，障碍物状态设为很小的负整数，其余状态处的回报值均为 0。这种方式的奖励函数没有启发性，机器人在初期学习阶段很难到达目标，导致收敛速度很慢。

在此，奖励函数的设计采用文献[9]所提神经网络方法。令状态方程(8)中，$A=10$，$B=D=1$，$E=100$，$\mu=1$，$r_0=2$，则状态方程转化为

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=-10x_i+(1-x_i)\left([I_i]^+ +\sum_{j=1}^{k}+\omega_{ij}[x_j]^+\right)-(1+x_i)[I_i]^- \tag{11}$$

该神经网络模型可以确保从目标状态发出的刺激性信息，通过神经元之间的横向连接，传递给该工作空间的所有状态，而从障碍物传出的抑制性信息只在有限的范围内传播。

通过状态方程(11)可以得到每个状态势值，势值矩阵为 X。将奖励函数定义为

$$R(s,a)=100(X(S')-X(S)) \tag{12}$$

式中：$X(S')$为执行下一个动作的状态势值；$X(S)$为当前状态势值。

3.2 算法实现流程

结合神经网络和 $Q(\lambda)$-learning 算法，具体的实现步骤如下：

步骤 1　利用状态方程(11)，经过 k 次迭代，求出状态势值矩阵；

步骤 2　根据式(12)，计算出奖励函数 $R(s,a)$；

步骤 3　进行第 i 次迭代计算，最大迭代次数为 n；

步骤 4　生成随机数 σ，执行式(7)搜索策略 $\pi(s)$；

步骤 5　执行动作 a，得到奖励 $R(s,a)$，转移到新状态 s'；

步骤 6　判断 s'是否为终止状态，如果是则跳到步骤 3 进行下一次迭代($i=i+1$)，不是则跳到步骤7；

步骤 7　根据式(5)，计算误差；

步骤 8　根据式(4)，更新动作状态迹，并更新其他动作状态迹；

步骤 9　根据式(6)，更新动作值函数；

步骤 10　$s\leftarrow s'$，跳到步骤 3，搜索下一个状态。

算法完成一次训练的流程如图 1 所示。该算法在迭代学习之前对根据环境地图信息进行状态势值计算，获得先验知识，以指导奖励函数设计，从而提高算法的收敛速度和最优解的质量。

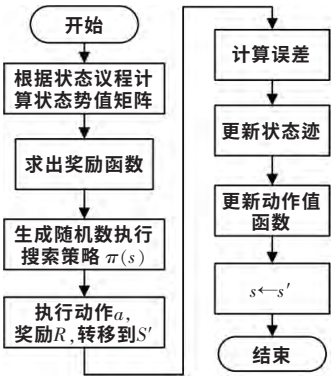

图 1　结合神经网络和 $Q(\lambda)$-learning 算法流程

Fig.1　Algorithm based on neural network and $Q(\lambda)$-learning flow chart

4 仿真试验与结果分析

通过仿真试验来验证改进方法的有效性，试验环境采用 20×20 栅格地图(如图 2 所示)，以图中左上角 S 为移动机器人的起点，以右下角 E 为目标。图中，白色部分为移动机器人的自由运动区间；黑色部分为障碍物，移动机器人无法穿越该区域。

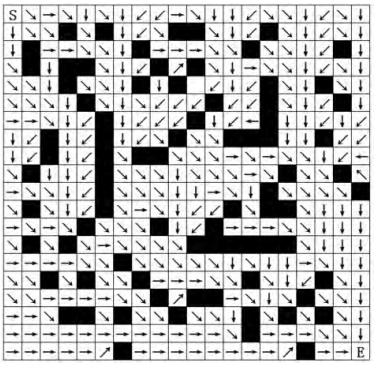

图 2　最优策略示意图

Fig.2　Optimal strategy schematic diagram

移动机器人动作集 A 包括上移、右上、右移、右下、下移、左下、左移、左上等 8 个动作；状态集包括 400 个位置，障碍物和目标状态为终止状态。当移动机器人移动到终止状态，则本次训练循环结束，重新进行下一次训练。

采用神经网络方法经过 1000 次迭代计算得到的势值分布如图 3 所示。根据式(12)处理状态势值可以得到奖励函数 $R(s,a)$。

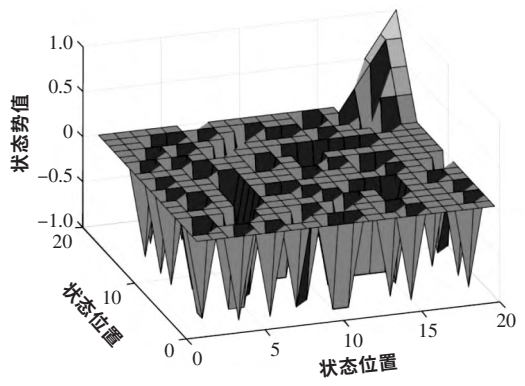

图 3　状态势值分布

Fig.3　State potential value distribution

算法中几个重要的参数会直接影响收敛速度。仿真试验中，折扣因子 γ 初始化为 0.8，学习速率 α 初始化为 0.05，探索因子 ε 初始化为 0.5，最大探索步数初始化为 400。当搜索步数超过最大步数时，仍未到达终止状态，则认为此次训练失败，重新进入下一次训练。

采用经典 Q-learning 方法，在训练 32000 次时收敛，而本文方法仅需 15000 次训练达到收敛状态，可见收敛速度有很大的提升。训练完成的最优策略如图 2 所示，从起点到终点的最优路径如图 4 所示。

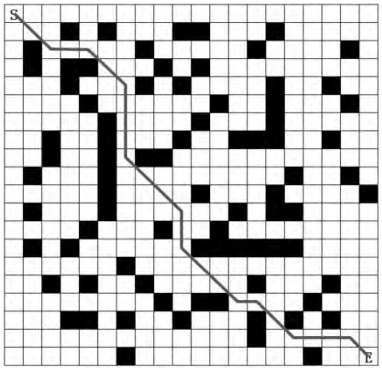

图 4　从起点到终点最优路径

Fig.4　Optimal path from start to end

采用同样的方法，生成另外 3 个障碍物分布不同的栅格地图，使用同样参数完成训练。统计移动机器人在地图所有状态下到达目标状态的平均步数，见表 1。

表 1　两种方法平均步数的对比

Tab.1　Comparison of average steps between the two methods

序号	经典方法	本文方法	序号	经典方法	本文方法
1	16.69	12.88	4	16.01	13.13
2	17.52	13.97	5	17.64	12.82
3	15.85	11.64	—	—	—

移动机器人在所有状态移动到终点的平均步数越少，说明策略越优。由表可知，本文方法在 5 次试验中，平均次数均明显低于经典 Q-learning 方法。

5　结语

所提出的结合神经网络和 $Q(\lambda)$-learning 算法的移动机器人路径规划算法，通过优化设计奖励函数，为增强学习提供了先验知识，解决了强化学习中存在的收敛速度慢和解的局部最优问题。通过仿真试验，本文方法与经典学习方法相比较，验证了该方法的有效性。

参考文献：

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

[2] Konar A，Chakraborty I G，Singh S J，et al. A deterministic improved Q-learning for path planning of a mobile robot[J]. IEEE Transactions on Systems，Man，and Cybernetics，2013，43(5)：1141-1153.

[3] Nazemi A，Omidi F.An efficient dynamic model for solving the shortest path problem[J].Transportation Research Part C Emerging Technologies，2013，26：1-19.

[4] 付成伟.基于分层强化学习的移动机器人路径规划[D].哈尔滨：哈尔滨工程大学，2008.

[5] 成伟明，唐振民，赵春霞，等.基于神经网络和 PSO 的机器人路径规划研究[J].系统仿真学报，2008，20(3)：608-611.

[6] 乔俊飞，侯占军，阮晓钢.基于神经网络的强化学习在避障中的应用[J].清华大学学报，2008，48(S2)：1747-1750.

[7] Hodgkin A L，Huxley A F.A quantitative description of membrane current and its application to conduction and excitation in nerve[J].Phys Lond，1952，117：500-544.

[8] Grossberg S.Contour enhancement，short term memory，and constancies in reverberating neural networks[J].Stud Appl Math，1973，52：217-257.

[9] Yang S X，Meng M. Neural network approaches to dynamic collision-free trajectory generation[J].IEEE Transactions on Systems，2001，31(3)：302-318.