

基于深度强化学习的机械臂避障路径规划研究

李广创, 程良伦

(广东工业大学自动化学院, 广东 广州 525000)

摘 要: 为了解决现有的机械臂焊接系统调整动作的难度大, 缺乏灵活性的问题, 本文采用了深度强化学习算法来解决机械臂的路径规划问题; 该方法使用一个三层的DNN网络, 输入为机械臂的状态信息, 输出为机械臂的运动关节角度, 通过离线训练, 机械臂能够自行训练出一条接近于最优的运动轨迹, 能够成功地避开障碍物到达目标点; 仿真在一个三自由度点焊机器人的模拟平台上进行, 仿真实验表明, 采用深度强化学习技术的机械臂能为焊接机械臂规划出一条无碰撞的路径, 具有较强的避障能力。

关键词: 机械臂; 强化学习; 碰撞检测; 路径规划

中图分类号: TP391.9 **文献标识码:** A

Research on Obstacle Avoidance Path Planning of the Mechanical Arm Based on Deep Reinforcement Learning

LI Guangchuang, CHEN Lianglun

(College of Automation, Guangdong University of Technology, Guangzhou 525000, China)

Abstract: In order to solve the difficulty of adjusting the existing mechanical arm welding system and the lack of flexibility, this paper adopts a deep reinforcement learning algorithm to solve the path planning problem of the mechanical arm; the method uses a three-layer DNN network with the input of mechanical state information and the output of the joint angle of the arm. Through offline training, the mechanical arm can be trained to move a nearly optimal motion trajectory and successfully avoid obstacles to reach the target point; the simulation is carried out in a three-degree-of-freedom welding robot simulation platform. The simulation experiment shows that the mechanical arm with deep reinforcement learning technology can plan a collision-free path for the welding mechanical arm and has strong obstacle avoidance ability.

Keywords: mechanical arm; reinforcement learning; collision detection; path planning

1 引言(Introduction)

在集装箱的焊接作业中, 为了实现高效的作业需求, 焊接机械臂大规模地应用到集装箱生产线中, 而作业的空间存在较多的障碍物, 需要在自动化焊接的过程中考虑碰撞的避免, 为了顺利地执行焊接作业, 对机械臂的避障路径规划的研究十分有必要。

针对机械臂的避障路径规划的问题, 国内外很多学者都进行了深入的研究。贾庆轩等人^[1]将笛卡尔空间障碍向构型空间障碍进行转换, 然后使用A*算法进行无碰撞路径规划。张云峰等人^[2]采用将机械臂从J空间(Joint Space)中随机采样后

通过正向运动学映射到C空间(Configuration Space), 然后采用改进的RRT算法进行路径规划。姬伟等人^[3]采用人工势场法进行机械臂的路径规划, 在障碍物处人工模拟一个斥力场, 在目标点处模拟一个引力场, 机械臂在这两个场的作用下运动。韩涛等人^[4]采用遗传算法来进行机械臂的路径规划。尹建军等人^[5]将机械臂工作空间映射到平面的二维空间, 再在平面空间上使用A*算法进行路径规划, 但是该方法只能使用与指定的场景, 缺乏灵活性。以上的这些方法需要精确的环境构建模型, 在复杂的工业环境下不能很好地适用。

在本文中, 采用强化学习的方法来解决避障路径规划

的问题,这个方法可以使用一个通用的框架来解决指定的任务。对于高维空间的复杂任务,例如机械臂系统,基于强化学习的无模型方法可以解决这样的问题。如利用DQN技术对三个关节的机械手臂进行端对端的控制^[6],用将深度强化学习技术应用于七自由度机械臂的插孔任务中^[7]。q-learning是一个经典的强化学习方法,由于q-learning算法需要一个巨大的Q表,导致在高维空间占用的内存巨大,并且不易收敛,因此,本文使用一种DQN的算法。与以往大多数碰撞检测算法不同,这个方法是一个无模型的方法,不需要针对每一种场景来建模,因此具有良好的通用性,该方法分为离线轨迹生成和在线应用。最后通过仿真,对本文提出的方法的有效性进行了验证。

2 强化学习基础(Foundation of reinforcement learning)

2.1 马尔科夫决策过程

大多数强化学习算法可以在同一个框架下描述,这个框架称为马尔科夫决策过程,简称MDP。一个完整的马尔科夫决策过程可以利用元组 (S, A, P, R, γ) 来表示。其中 S 是状态集, A 是动作集, P 是状态转移概率,也就是对应着环境和智能体模型, R 为回报函数, γ 为折扣因子,用于计算累计回报 $R^{[8,9]}$, 累计回报的计算公式为 $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$, 其中 $0 \leq \gamma \leq 1$ 。

定义策略 π 为在给定状态 S 时, 动作集上的分布 $\pi(a|s) = P[A_t = a|S_t = s]$ 。

定义状态-动作值函数为: 累计回报在状态 s 和动作 a 处的期望值:

$$q_{\pi}(s, a) = E_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a \right]$$

强化学习的目标是给定一个马尔科夫决策过程, 寻找最优策略 π , 使该决策下的累计回报期望最大, 即:

$$\pi^*(s, a) = \arg \max_a q_{\pi}(a, s)$$

2.2 q-learning算法

机械臂的避障属于强化学习的控制问题, 控制问题可以表示为: 给定强化学习的五个要素: 状态集 S 、动作集 A 、即时奖励 R 、衰减因子 γ 、探索率 ϵ , 求解最优的动作价值函数和最优策略。

q-learning算法是一种使用时序差分求解强化学习控制问题的方法, 这一类强化学习的问题求解不需要环境的状态转化模型, 是不基于模型的强化学习问题求解方法。它属于时序差分中的离线控制方法, 也即会使用两个控制策略, 一个策略用于选择新的动作, 另一个策略用于更新价值函数^[9,10]。

第一个策略, 它先用 ϵ -greedy策略在当前状态 S 选择出动作 A , 执行动作 A 得到 S' ; 第二个策略, 基于状态 S' , 使用贪婪法选择 A' , 也就是说, 选择使 $Q(S', a)$ 最大的动作 A' 来更新价值函数。用数学公式表示为:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r_t + \gamma \max_a Q(s', a) - Q(s, a)]$$

此时选择的动作只会参与价值函数的更新, 不会真正的

执行。价值函数更新后, 新的执行动作需要基于状态 S' , 基于策略一, 用 ϵ -greedy策略重新选择得到用表格表示的动作价值函数 Q^* , 使用贪婪策略, 即每次寻找最优的动作便可以得到机械臂的控制。

2.3 深度Q网络(DQN)

在q-learning算法中, 动作价值函数 $Q(S, a)$ 对应着一张表, 索引对应着状态-行为对。值函数的迭代更新对应实际上就是这张表的迭代更新, 当状态空间的维数很大, 或者状态空间是连续空间, 则此时值函数无法用一张表格来表示, 此时可以用函数逼近的方法来表示价值函数, 这个函数可以用参数 w 来表示, 最常见的表示方法是使用神经网络^[11], 具体的形式为: $\hat{Q}(s, a, w) \approx Q(s, a)$ 。

一般情况, 使用的网络结构为输入是状态 S 的特征向量 $\phi(S)$, 动作集合有多少个动作, 神经网络就有多少个输出 $\hat{Q}(s, a_i, w)$, 中间的结构可以DNN、CNN、RNN等, 没有特别的限制。

神经网络的训练需要带标签的样本, DQN使用经验回放技巧来解决这个问题, 也即将每次和环境交互得到的奖励与状态更新情况都保存在经验池 D 中, 然后从经验池 D 中采样 m 个样本用于后面目标 Q 值的更新。

在神经网络中使 $\phi(S)$ 用作为输入, 得到所有动作对应的 Q 值输出。用 ϵ -greedy策略在当前 Q 值输出中选择对应的动作 A , 在状态 S 执行当前动作 A , 得到新状态 S' 下, 对应的特征向量 $\phi(S')$, 奖励 R 和终止状态标志 is_end ; 也即5元组: $\{\phi(S), A, R, \phi(S'), is_end\}$, 它对应的目标 Q 值为:

$$y_i = \begin{cases} R_j, & \text{if } is_end = True \\ R_j + \gamma \max_{j'} Q(\phi(S'_{j'}), A'_{j'}, w), & \text{if } is_end = False \end{cases}$$

由此, 根据反向传播规则, 可以使用均方差损失函数

$$\frac{1}{m} \sum_{j=1}^m (y_j - Q(\phi(S_j), A_j, w))^2$$

来更新神经网络的参数。同时, 为了减少目标 Q 值和当前 Q 值的相关性, DQN使用两个结构完全相同的网络, 一个“当前网络 Q ”用来选择动作和更新模型参数, 另一个“目标网络 Q' ”用来计算目标 Q 值。目标网络 Q' 的网络参数不需要迭代更新, 而是每隔一段时间从当前网络 Q 复制过来, 即延时更新。

3 机械臂避障系统(Obstacle avoidance system of the mechanical arm)

3.1 空间机械臂模型

本文将深度强化学习技术应用于机械臂的路径规划中, 首先要对机械臂进行建模, 机械臂采用D-H法建立数学模型, 如图1和图2所示。机械臂相邻连杆的关系用一个4*4的齐次变换矩阵来描述, 旋转关节的齐次变换矩阵的一般形式为:

$$A_i = \begin{bmatrix} c\theta_i & -s\theta_i c\alpha_{i-1} & s\theta_i s\alpha_{i-1} & \alpha_{i-1} c\theta_i \\ s\theta_i & c\theta_i c\alpha_{i-1} & -c\theta_i s\alpha_{i-1} & \alpha_{i-1} s\theta_i \\ 0 & s\alpha_{i-1} & c\alpha_{i-1} & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

机械臂末端位置和姿态的求解是正运动学问题, 表示为

各个相邻齐次变换矩阵的连乘,机械臂的正运动学方程为:

$$T_3^0 = T_1^0 T_2^1 T_3^2.$$

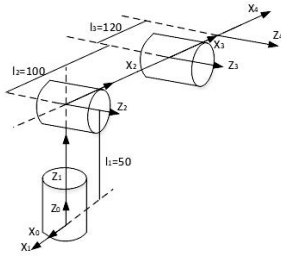


图1 3自由度串联机械臂模型

Fig.1 3-DOF tandem mechanical arm model

杆件参数	θ	d	a	α
杆件 1	θ_1	l_1	0	90°
杆件 2	θ_2	0	l_2	0
杆件 3	θ_3	0	l_3	0

图2 机械臂的DH参数

Fig.2 DH parameters of the mechanical arm

3.2 碰撞检测

三维空间障碍物一般具有不规则的形状,为了分析专注于核心问题的分析,将三维障碍物简化为空间球体。强化学习每一步需要检测当前机械臂与环境的状态,也即需要判断机器人的当前位置是否与障碍物发生碰撞,设球形障碍物的中心为 $P_0(x_0, y_0, z_0)$,半径为 r_0 ;圆柱体机械臂的半径为 r_1 ;机械臂轴线的两个端点分别为 $A(x_1, y_1, z_1)$ 、 $B(x_2, y_2, z_2)$,如图3所示。



图3 碰撞检测示意图

Fig.3 Schematic diagram of collision detection

设障碍物到直线的投影点为: $P_c(x_c, y_c, z_c)$

直线的方向向量为:

$$(x_1 - x_2, y_1 - y_2, z_1 - z_2) = (m, n, p)$$

根据AB的空间直线方程,得:

$$x_c = mt + x_1, y_c = nt + y_1, z_c = pt + z_1$$

根据障碍物与投影点的连线与直线相交得:

$$m(x_0 - x_c) + n(y_0 - y_c) + p(z_0 - z_c) = 0$$

根据上面两式可以求出投影点坐标 $P_c(x_c, y_c, z_c)$;进而确定障碍物到机械臂的距离:

$$d_1 = \sqrt{(x_1 - x_c)^2 + (y_1 - y_c)^2 + (z_1 - z_c)^2}$$

障碍物到端点的距离:

$$d_2 = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}$$

碰撞检测算法的描述为:

(1)对于每个轴

a.若障碍物的投影在线段区域内,即 $(x_c \in [x_1, x_2] \& y_c \in [y_1, y_2] \& z_c \in [z_1, z_2])$,则计算球心 r_0 到线段 AB 的距离 d_1 ,如果 $d_1 > r_0 + r_1$,则判断为安全,否则判断为碰撞。

b.若障碍物的投影不在线段的区域内,即 $(x_c \notin [x_1, x_2] \& y_c \notin [y_1, y_2] \& z_c \notin [z_1, z_2])$,则计算球心 r_0 到端点的距离 d_2 ,如果 $d_2 > r_0 + r_1$,则判断为安全,否则判断为碰撞。

(2)若三个轴都不与障碍物发生碰撞,则判断为安全,否则判断为碰撞。

3.3 机械臂避障系统相关参数的设定

奖励 r :机械臂每执行一个动作,环境会产生一个奖励,奖励函数的设置至关重要,直接影响着算法是否能收敛,针对机械臂路径规划问题,希望机械臂能顺利避开障碍物并到达目标点,同时产生的步数尽量少,因此奖励函数的设置为:

$$\begin{cases} \text{到达目标点: } R = R + 100 \\ \text{碰撞: } R = R - 100 \\ \text{其他: } R = 0.6 \times \|p_1 - p_2\| \end{cases}$$

状态 s :机械臂与环境的交互产生的状态,分别用下面的变量表示:末端点的坐标,第三个端点的坐标,是否碰撞,是否到达目标点,末端点与目标点的距离,第三个端点与目标点的距离。将状态量用一个向量的形式表示为:

$$s = (x_4, y_4, z_4, x_3, y_3, z_3, ds, ms, d_4, d_3).$$

行为 a :机械臂的行为是关节的转动的角度,用一个向量的形式表示为: $a = (\Delta\theta_1, \Delta\theta_2, \Delta\theta_3)$ 。

3.4 基于DQN的机械臂路径规划算法的实施

本文采用三层DNN的网络作为DQN的网络主体,其具体的机械臂避障路径规划算法描述如下:

随机初始化所有的动作和状态对应的价值 Q ,随机初始化当前 Q 网络的所有参数 w ,初始化目标网络 Q' 的参数 $w' = w$,清空经验回放集合 D

2.for episode=1,M do:

1)从环境中获得机械臂的初始状态 S 。

2)for $t=1, T$ do:

a.根据 ϵ -greedy策略选择动作 a ;

b.机械臂执行动作,然后从环境中获取奖励 R_t ,获取下一个时刻的状态 S' ;

c.存储机械臂与环境的交互过程:将 (S, a, R, S') 存在 D 中;

d.从 D 中随机采样 n 个最小批次的样本 (S_j, a_j, R_j, S'_j) , $j=1, \dots, n$;

e.如果机械臂到达该episode的终止状态,令 $y_j = R_j$,否则 $y_j = R_j + \gamma \max Q(\phi(S'_j), A'_j, w)$;

f.通过损失函数 $\frac{1}{m} \sum_{j=1}^m (y_j - Q(\phi(S_j), A_j, w))^2$,执行一次梯度下降,并更新参数 w ;

g.每隔 C 步,更新一次目标网络的权值参数 $w' = w$ 。

4 实验仿真(Experiment and simulation)

4.1 实验配置

本文的仿真环境在MATLAB和python下搭建,在

MATLAB下完成机械臂及碰撞检测模型的搭建,在python下完成深度强化学习模型的搭建。

神经网络的输入是机械臂的状态,用一个10维向量表示: $s = (x_4, y_4, z_4, x_3, y_3, z_3, ds, ms, d_4, d_3)$; 输出为机械臂的角度转动量 $a = (\Delta\theta_1, \Delta\theta_2, \Delta\theta_3)$, 本文允许机械臂的转角为 $\Delta\theta = \{-5^\circ, 0^\circ, 5^\circ\}$ 。

DQN迭代的最大轮数设置为10000轮,每一轮最大的步数设置为250步;本文设置的机械臂3个杆长为[50,100,100],起点坐标为(118.3,-68.3,13.3);终点坐标为(50,70,30)。障碍物1的坐标为(40,30,30),障碍物2的坐标为(70,-20,60),障碍物的半径为20,为了简化起见,将机械臂设置为点杆模型。

4.2 实验结果与分析

机械臂经过训练好的神经网络,每一步输出一个最大Q值得动作,然后使用正运动学来控制机械臂的行走;机械臂的避障结果如图4所示:机械臂经过强化学习模型的训练,成功地避开两个障碍物,到达目标点,机械臂的运动步数为32步,接近于最优;从图5可以看出,机械臂每一轮的累计回报逐渐增加,大约在6300轮迭代后收敛;从图6可以看出,机械臂每一轮的迭代步数逐渐减少,并最终稳定收敛。

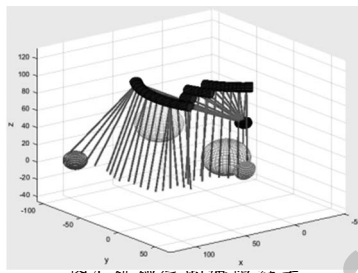


图4 机械臂的避障结果

Fig.4 Obstacle avoidance results for the mechanical arm

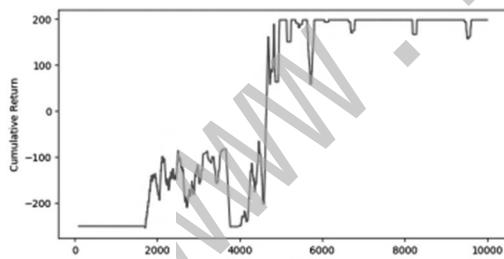


图5 每回合的累计回报

Fig.5 Cumulative return per round

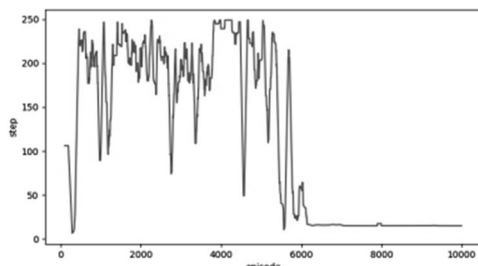


图6 每回合实际走的步数

Fig.6 Actual steps taken per round

5 结论(Conclusion)

本文以点焊机械臂为研究对象,将机械臂在三维空间的自动点焊转化为机械臂的避障路径规划问题。本文首先介绍了强化学习的基本原理,然后通过建立机械臂模型,采用深度强化学习的方法进行机械臂避障路径规划的求解,分别在MATLAB和TensorFlow平台上建立机械臂模型和强化学习模型;仿真结果表明,通过一定的训练步骤,机械臂能顺利规划出一条无碰撞的焊接路径。相比于传统的规划方法,本文的基于强化学习的规划方法简单,而且具有较高的通用性。

参考文献(References)

- [1] 贾庆轩,陈钢,孙汉旭,等.基于A*算法的空间机械臂避障路径规划[J].机械工程学报,2010,46(13):109-115.
- [2] 张云峰,马振书,孙华刚,等.基于改进快速扩展随机树的机械臂路径规划[J].火力与指挥控制,2016,41(05):25-30.
- [3] 姬伟,程凤仪,赵德安,等.基于改进人工势场的苹果采摘机器人机械手避障方法[J].农业机械学报,2013,44(11):253-259.
- [4] 韩涛,吴怀宇,杜钊君,等.基于遗传算法的机械臂实时避障路径规划研究[J].计算机应用研究,2013,30(05):1373-1376.
- [5] 尹建军,武传宇,Yang Simon X,等.番茄采摘机器人机械臂避障路径规划[J].农业机械学报,2012,43(12):171-175;157.
- [6] Wonchul Kim,Taewan Kim,et al.Three-link Planar Arm Control Using Reinforcement Learning[C].International Conference on Ubiquitous Robots and Ambient Intelligence,2017:424-428.
- [7] Tadanobu Inoue,Giovanni De Magistris,et al.Deep Reinforcement Learning for High Precision Assembly Tasks[C].IEEE International Conference on Intelligent Robots and Systems,2017:819-825.
- [8] 刘全,翟建伟,章宗长,等.深度强化学习综述[J].计算机学报,2018,41(01):1-27.
- [9] Kober,Jens,J.Andrew Bagnell,et al.Reinforcement learning in robotics:A survey[C].The International Journal of Robotics Research,2013:1238-1274.
- [10] 方小菊.基于强化学习的采摘机器人采摘臂避障设计[J].农机化研究,2017,39(11):198-202.
- [11] 王翌,胡立生.基于深度Q学习的工业机械臂路径规划方法[J].化工自动化及仪表,2018,45(02):141-145;171.

作者简介:

李广创(1994-),男,硕士生.研究领域:机器学习,机器人.
程良伦(1965-),男,博士,教授.研究领域:人工智能,工业大数据.