·热点与综述· 文章编号：1000-3428(2021)10-0016-10 文献标志码：A 中图分类号：TP242

基于强化学习的路径规划技术综述

闫皎洁[1,2]，张锲石[1,2]，胡希平[1,2]

（1.中国科学院深圳先进技术研究院，广东 深圳 518055；2.中国科学院大学 深圳先进技术学院，广东 深圳 518055）

摘 要：路径规划作为移动机器人自主导航的关键技术，主要是使目标对象在规定范围内找到一条从起点到终点的无碰撞安全路径。阐述基于常规方法和强化学习方法的路径规划技术，将强化学习方法主要分为基于值和基于策略两类，对比时序差分、Q-Learning等基于值的代表方法与策略梯度、模仿学习等基于策略的代表方法，并分析其融合策略和深度强化学习方法方法的发展现状。在此基础上，总结各种强化学习方法的优缺点及适用场合，同时对基于强化学习的路径规划技术的未来发展方向进行展望。

关键词：路径规划；强化学习；深度强化学习；移动机器人；自主导航

开放科学（资源服务）标志码（OSID）：

中文引用格式：闫皎洁，张锲石，胡希平.基于强化学习的路径规划技术综述[J].计算机工程，2021，47(10)：16-25.

英文引用格式：YAN J J，ZHANG Q S，HU X P.Review of path planning techniques based on reinforcement learning[J].Computer Engineering，2021，47(10)：16-25.

Review of Path Planning Techniques Based on Reinforcement Learning

YAN Jiaojie[1,2]，ZHANG Qieshi[1,2]，HU Xiping[1,2]

（1.Shenzhen Institute of Advanced Technology，Chinese Academy of Sciences，Shenzhen，Guangdong 518055，China；
2.Shenzhen College of Advanced Technology，University of Chinese Academy of Sciences，Shenzhen，Guangdong 518055，China）

【Abstract】Path planning is one of the key technologies for autonomous navigation of mobile robots.It aims at planning a collision free optimal path from the current position to the destination in real time.This paper introduces the path planning techniques that are based on Reinforcement Learning(RL) and common methods，and categorizes the methods based on RL into two types：the value-based methods and the strategy-based methods.Then the paper compares value-based representation methods(including Timing Difference(TD)，Q-Learning，etc.) and the strategy-based representation methods(including Strategy Gradient(SG) and Imitation Learning(IL)，etc.)，and analyzes the development status of its fusion strategy and Deep Reinforcement Learning(DRL).On this basis，the paper summarizes the advantages，disadvantages and application scenarios of the RL-based methods.Finally，the future development trends of the path planning techniques based on RL are discussed.

【Key words】path planning；Reinforcement Learning(RL)；Deep Reinforcement Learning(DRL)；mobile robot；autonomous navigation

DOI：10.19678/j.issn.1000-3428.0060683

0 概述

随着计算机技术、人工智能技术及自动化控制技术的发展，移动机器人的智能化程度不断提高，路径规划作为实现机器人自主导航的核心技术受到广泛关注[1-2]。路径规划就是使目标对象在最小的时间或距离代价下，在规定区域范围内找到一条从起点到终点的安全无碰撞路径。目前，路径规划的核心方法[3-4]按其特点及关键技术可分为常规方法与强化学习(Reinforcement Learning，RL)方法两类。常规方法可分为传统方法、图形学方法及智能仿生学方法。近年来，又由强化学习方法衍生出深度强化学习方法，深入研究强化学习方法及其衍生方法对路径规划技术的发展有重要的意义。本文简述基于常规方法的路径规划技术，重点分析基于强化学习的路径规划技术并将其分为基于值和基于策略两类，再由强化学习引出基于深度强化学习的路径规划技术，同时对强化学习代表方法的原理、特点、优

基金项目：国家自然科学基金(U1913202，U1813205)；深圳科技计划基础研究项目(JSGG20191129094012321，JCYJ20180507182610734)。
作者简介：闫皎洁(1998—)，女，硕士研究生，主研方向为移动机器人路径规划；张锲石(通信作者)，高级工程师、博士；胡希平，教授、博士。
收稿日期：2021-01-23 **修回日期：**2021-04-26 **E-mail：**jiaojie_yan@163.com

缺点、适用场合及改进策略进行深入探讨。

1 基于常规方法的路径规划技术

常规方法分为传统方法、图形学方法、智能仿生学方法等3类:

1)传统方法主要包括模拟退火法[5]、人工势场法[6]和模糊逻辑法[7]。这类方法最早应用于路径规划技术,具有描述简单、易于实现的特点,但不能充分利用先验知识和全局信息,求解时容易陷入局部最优解或遇到目标不可达的问题。

2)图形学方法主要包括A*算法[8]、栅格法[9]等。这类方法可提供建模方法,解决了传统方法建模难的问题,但由于搜索效率低下,导致其难以应用在实际系统中。

3)智能仿生学方法主要包括遗传算法[10]、人工神经网络算法[11]、蚁群算法[12]、粒子群优化(Particle Swarm Optimization,PSO)算法[13]等。这类方法的原理与自然生物的性质或生态机制非常接近,如模仿生物遗传进化、人体神经网络系统、蚂蚁觅食等行为,故统称为智能仿生学方法。由于仿生特点,这类方法更加智能、效率更高,但在路径规划应用中存在容易陷入局部最优解、收敛速度慢等问题。

为更加清晰直观地对比各类各种常规方法,表1给出了各种常规方法的优劣势对比结果。

表1 应用于路径规划技术的常规方法优劣势对比

Table 1 Comparison of advantages and disadvantages of conventional approaches applied to path planning techniques

常规方法	代表方法	优势	劣势
传统方法	模拟退火法 人工势场法 模糊逻辑法	描述简单且易于实现	易陷入局部最优解或目标不可达
图形学方法	A*算法 栅格法	可提供建模方法	搜索效率低下
智能仿生学方法	遗传算法 人工神经网络算法 蚁群算法 粒子群优化算法	具有仿生学特点且更加智能高效	易陷入局部最优解且收敛速度慢

2 基于强化学习的路径规划技术

2.1 强化学习基本原理与研究历程

强化学习[14-15]的基本原理为智能体在环境反馈奖励或惩罚的刺激下持续学习,根据反馈不断调整策略,最终达成奖励最大化或实现特定目标。强化学习方法主要包括状态、策略、行动、奖励等4个要素。智能体在状态s_t下,根据策略π选择动作a_t,并从状态s_t转移到新的状态s_{t+1},同时获得环境反馈的奖励r,根据获得的奖励r获得最优策略π^*。

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, E_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,a_t)|S_0=S\right] \tag{1}$$

其中:$\gamma\in(0,1)$为折扣率。

强化学习思想最早可追溯到行为心理学研究。1911年THORNDIKE提出效果律(Law of Effect):一定情景下让动物感到舒服的行为,就会与此情景增强联系(强化),当此情景再现时,动物的这种行为也更易再现;反之则相反。20世纪50年代中期,最优控制理论被提出,基本原理为从控制方案中寻找最佳方案。1956年BELLMAN[16]提出动态规划方法,1977年WERBOS[17]提出自适应动态规划方法。直到20世纪80年代末90年代初,人工智能、机器学习等技术开始得到广泛应用,强化学习开始受到关注。1988年SUTTON等[14]提出时序差分(Temporal Difference,TD)算法,1992年WATKINS等[18]提出Q-Learning算法,1994年RUMMERY等[19]提出SARAS算法,1995年BERSEKAS等[20]提出解决随机过程中优化控制的神经动态规划方法,2006年KOCSIS等[21]提出置信上限树算法,2009年LEWIS等[22]提出反馈控制自适应动态规划算法,2014年SILVER等[23]提出确定性策略梯度(Deterministic Policy Gradient,DPG)算法,2016年Google DeepMind[24]提出A3C方法。

2.2 强化学习方法分类

求解强化学习问题的方法分为基于值、基于策略以及基于值与基于策略相结合的方法。基于值的方法定义了值函数,根据值函数的大小选择动作;基于策略的方法将策略进行参数化,通过优化参数使策略的累计回报最大。

当移动机器人在复杂未知环境下进行路径规划时,由于初期探索策略存在盲目性,导致强化学习存在收敛速度慢的问题,在机器人训练过程中需要花费大量时间。此外,随着环境复杂度和系统状态维度的增加,需要训练的参数呈指数级增长,因此会消耗大量训练时间和存储空间,最终导致维数灾难[25]。此外,强化学习的可移植性和通用性差,训练过的机器人无法直接在新的环境中按照期望规划移动。

2.2.1 基于值的强化学习方法

基于值的方法主要适用于离散动作空间,目标是通过最大化每个状态的值函数来得到最优策略。

值函数用来衡量当前状态下机器人选择策略的优劣程度。根据自变量的不同,值函数可以分为状态值函数$V(s)$和状态-动作对值函数$Q(s,a)$,如式(2)和式(3)所示:

$$V^{\pi}(s)=E_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r(S_t,a_t)|S_0=S\right] \tag{2}$$

$$Q^{\pi}(s_t,a_t)=r(s_t,a_t)+\gamma V^{\pi}(s_{t+1}) \tag{3}$$

从式(2)和式(3)可知:状态值函数是某状态下的奖励反馈值,状态-动作对值函数是状态-动作对下的奖励反馈值,因此只需最大化值函数就可达成最终奖励最大化。基于值的方法主要包括TD[14]、Q-Learning[18]、SARSA[19]、Dyna[26]等方法。

1)TD算法

TD算法是一类无模型的强化学习算法,从环境中取样并学习当前值函数的估计过程,原理为通过借助时间的差分误差来更新值函数,误差计算公式和值函数更新公式分别如式(4)和式(5)所示:

$$\delta_t=r_{t+1}+\gamma V(s_{t+1})-V(s_t) \tag{4}$$

$$V(s_t)\leftarrow V(s_t)+\alpha\delta_t \tag{5}$$

其中:α为学习率。

TD算法基于蒙特卡罗思想和动态规划思想,可直接学习初始体验,无需环境动态模型,同时基于学习更新,需等待最终学习结果。NAIR等[27]提出一种针对静态障碍物路径规划和避障的修正时序差分算法,降低了TD算法的计算复杂度。MARTIN等[28]将时序差分算法的更新过程简化为高斯回归过程,提高了机器人在海洋环境的路径规划中的数据处理效率。

2)Q-Learning算法

在TD算法的基础上,WATKINS等[18]提出Q-Learning算法。Q-Learning算法是强化学习发展的里程碑,是基于值的强化学习算法中应用最广泛的算法,也是目前应用于移动机器人路径规划最有效的算法之一。Q-Learning算法属于在线强化学习算法,基本思想为定义一个状态-动作对值函数$Q(s,a)$,将某时刻的数据代入式(6)和式(7)中更新值函数$Q(s,a)$。

$$Q(s_t,a_t)=Q_t(s_t,a_t)+\alpha_t\delta_t \tag{6}$$

$$\delta_t=r_{t+1}+\gamma\max_{a'}Q_t(s_{t+1},a')-Q(s_t,a_t) \tag{7}$$

其中:a_t为学习率;δ_t为误差;a'为状态s_{t+1}执行的动作。

Q-Learning算法采用离线策略(off-policy)来产生动作,根据该动作与环境的交互得到的下一个状态及奖励来学习得到另一个最优的$Q(s,a)$。目前,关于Q-Learning算法的改进是学者们研究的重点方向,主要包括以下4个方面:

(1)引入启发式思想。启发式思想通常是对某一训练环节进行趋势性启发以提高学习效率。启发式搜索策略具有较强的搜索能力,LI等[29]在改进的Q-Learning算法中引入了启发式搜索策略来加快学习过程,通过限制方向角的变化范围,缩小搜索空间。刘智斌等[30]利用shaping函数参与Q表的更新,对Q-Learning的趋势性进行启发,大幅提高收敛效率。JIANG等[31]提出一种基于经验重放和启发式知识的深度Q-Learning算法来实现智能机器人的路径规划。一方面,启发式知识可以指导机器人的动作选择,减少智能机器人系统中的随机性;另一方面,启发式知识提高了神经网络的训练效率,使其可以更快地收敛到最优的行动策略。

(2)引入分层思想。分层强化学习[32]致力于将一个大规模问题分解为若干个子问题,以分而治之的思想逐个解决,分层思想可以很好地解决传统强化学习中的维数灾难问题。刘智斌等[30]提出的基于BP神经网络的双层启发式强化学习算法,引入了双层强化学习模型:第一层为定量层,通过Q-Learning算法训练得到精确结果;第二层为定性层,具有较好的泛化能力,提供大方向上的预见。BUITRAGO-MARTINEZ等[33]提出一种基于选择的双层Q-Learning学习方法:第一层用来训练机器人的基本行为且每一种行为在训练阶段相互独立;第二层通过训练机器人并协调这些基本行为来解决路径规划问题。刘志荣等[34]建立双层网络结构,使用Q-target神经网络来计算目标Q值,减少了目标状态对当前状态的依赖,大幅提高了收敛效率。

(3)引入模糊逻辑思想[35]。在生活中的许多概念都具有模糊性,如远和近、快和慢等,模糊逻辑用隶属度取代布尔数值来标识程度,在人工智能领域起到了重要的作用。LUVIANO等[36]将模糊逻辑应用于连续时间的多智能机器人路径规划中进行以下改进:①模糊量化状态空间;②将模糊逻辑与WoLF-PHC[37]算法结合,使Q函数通过模糊状态空间进行分离;③将模糊Q-Iteration模型用于智能体的次优策略,解决了传统Q-Learning算法的维数灾难问题。WEN等[38]在优化的Q-Learning算法基础上提出模糊Q-Learning(Fuzzy Q-Learning,FQL)算法并将其应用于路径规划的避障问题,进一步提高了训练的收敛速度。葛媛等[39]提出一种基于模糊RBF网络的Q-Learning算法,使模糊神经系统具有自适应性,对未知动态环境中移动机器人的自主路径规划具有一定的应用价值。

(4)多算法结合思想。一个单独的路径规划算法在实际应用中或多或少都存在一定的缺陷,设计新的算法难度大,因此可通过多种算法的结合来解决问题。除上述三大类的改进思路外,学者们还将Q-Learning与其他类型的算法相结合进行优化。朴松昊等[40]用遗传算法初步规划出全局最优路径,并结合Q-Learning算法实现机器人的避障行为,两种算法取长补短,满足

了路径规划的高实时性要求。MEERZA等[41]提出一种基于Q-Learning和粒子群优化的路径规划算法，利用PSO改进Q表的迭代，在速度和精度上相比单独使用这两种算法性能更优。SHI等[42]将Q-Learning算法与蚁群算法中的信息素机制进行融合，机器人之间通过信息素进行信息交换，解决了多智能体路径规划中的信息共享问题，并且在Q值的作用下，机器人做出状态更新和决策选择。YAO等[43]在Q-Learning算法的基础上，结合人工势场法，以黑洞势场为环境，使机器人在没有先验知识的情况下可跳出局部最优解。LIU等[44]将RRT与Q-Learning算法进行结合，提出一种基于Q-Learning的分区启发式RRT算法，利用Q-Learning改进奖励函数，获得全局最优路径，此算法可以获得更平滑的结果，并且提高了搜索和避障的能力。为解决Q-Learning奖励函数定义宽泛导致学习效率低下的问题，王子强等[45]提出一种基于详细回报分类的Q-Learning算法，根据移动机器人与障碍物间的距离，对每个时刻机器人获得的奖励值分配安全等级，使机器人学习过程的安全等级更高，选择的路径更合理。

3）SARSA算法

SARSA算法与Q-Learning算法相似，也是一种在线强化学习算法。区别在于SARSA算法采用在线策略（on-policy），迭代的是$Q(s,a)$的实际值，误差计算公式如下：

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{8}$$

由式（8）可知：SARSA值函数$Q(s,a)$的更新涉及到(s,a,r,s_{t+1},a_{t+1})这5个部分，它们构成了该算法的名字SARSA。

在机器学习中，若智能体在线学习且注重学习期间所获奖励，则SARSA算法的适用性会更强。SARSA算法是单步更新算法，即SARSA（0）。在获得奖励后，仅更新上一步状态和动作对应的Q值，但每一步得到的奖励都会影响最终得到的奖励，因此将此算法优化为多步更新的SARSA算法，即SARSA（λ）。ZOU等[46]针对复杂动态环境提出一种基于SARSA（λ）优化的RRT路径规划算法，通过该优化可增加扩展点时的选择并减少无效节点数，从而提高算法性能。XU等[47]通过SOM神经网络获得位置信息并产生R值，再使用SARSA（λ）并基于产生的R值来更新Q值，从而使路径规划更加精准高效。FATHINEZHAD等[48]提出监督模糊SARSA学习算法，将监督学习和强化学习的优点相结合，训练找出每个模糊规则的最佳动作，通过模糊SARSA学习在线微调模糊控制结论部分的参数，减少了学习时间和训练失败次数。

SARSA算法与Q-Learning算法有相同点但又有所区别，这两种算法的相同点包括：（1）在TD算法的基础上改进；（2）使用ε-greedy选择新的动作；（3）均为在线强化学习算法。这两种算法的区别包括：（1）Q-Learning算法使用off-policy，迭代内容为$Q(s,a)$的最大值；（2）SARSA算法使用on-policy，迭代内容为$Q(s,a)$的实际值。

4）Dyna算法

Dyna算法并不是一个具体的强化学习算法，而是一类算法框架的总称。将基于模型的强化学习和与模型无关的强化学习相结合，既从模型中学习，也从与环境交互的经验中学习，从而进行函数更新。DABOONI等[49]利用直接启发式动态规划（Heuristic Dynamic Programming，HDP）改进Dyna算法，采用HDP策略学习构造Dyna-agent，并提出一种新的在线规划学习算法Dyna-HDP，可以更快得到近似最优路径，并具有一定的稳定性。VIET等[50]将移动机器人的学习过程分为两阶段：第一阶段通过Dyna-Q算法加速获取最优策略，并训练机器人躲避障碍物；第二阶段训练机器人获得平滑的路径。该方法能够有效解决障碍物密集的未知环境下移动机器人的路径规划问题。HWANG等[51]将一种基于树的自适应模型学习方法与Dyna-Q算法相结合，利用模型训练产生的经验加速迭代，训练效率得到明显提升。

2.2.2　基于策略的强化学习方法

基于策略的方法通过直接优化策略得到最优策略。基于策略的方法主要包括策略梯度（Policy Gradient，PG）[52]、模仿学习（Imitation Learning，IL）[53]等方法。

1）策略梯度法

策略梯度法是基于策略的算法中最基础的一种算法[52]，基本思路为通过逼近策略来得到最优策略。策略梯度法分为确定性策略梯度法和随机性策略梯度法（Stochastic Policy Gradient，SPG）。在确定性策略梯度法中，动作被执行的概率为1，而在随机性策略梯度法中，动作以某概率被执行。与随机性策略梯度法相比，确定性策略梯度法在连续动作空间求解问题中性能更好。假设需要逼近的策略是$\pi(s,a;\theta)$，策略π对参数θ可导，定义目标函数和值函数如式（9）和式（10）所示。从初始状态s_0开始，依据策略π_θ选取动作的分布状态如式（11）所示。根据式（9）~式（11）得到的策略梯度公式如式（12）所示。

$$J(\pi_\theta)=E\left[\sum_{t=1}^{\infty}\gamma^{t-1}r_t|s_0,\pi_\theta\right] \tag{9}$$

$$Q^{\pi_\theta}(s,a)=E\left[\sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}|s_t=s,a_t=a,\pi_\theta\right] \tag{10}$$

$$d^{\pi_\theta}(s)=\sum_{t=1}^{\infty}\gamma^t P(s_t=s|s_0,\pi_\theta) \tag{11}$$

$$\nabla_\theta J(\pi_\theta)=\sum_s d^{\pi_\theta}(s)\sum_a\nabla_\theta\pi_\theta(s,a)Q^{\pi_\theta}(s,a) \tag{12}$$

LIU等[54]以学习曲线理论为基础，构造经验池容量变化函数，在传统深度确定性策略梯度法（Deep Deterministic Policy Gradient，DDPG）中加入学习曲线，从而实时调整回放缓冲容量，改进后的算法奖励值更高，学习能力更强。PAUL等[55]将DDPG算法应用于机械臂的路径规划中，即应用于连续动作空间，使用该方法进行训练，简化了学习步骤并提高了成功率。ZHENG等[56]提出一种改进的多智能体深度确定性策略梯度法（Improved Multi-Agent Deep Deterministic Policy Gradient，IMADDPG），通过增加平均场网络（Mean Field Network，MFN）最大化智能体的返回值，使所有的智能体在训练时能最大限度地提高协作性，最终可求解全局最优路径。

2）模仿学习法

与策略梯度法相同，模仿学习也是一种直接策略搜索方法。模仿学习的基本原理是从示范者提供的范例中进行学习，示范者一般提供人类专家的决策数据，通过模仿专家行为得到与专家近似的策略。

在线性假设下，反馈信号可由一组确定基函数$\varphi_1,\varphi_2,\cdots,\varphi_k$线性组合而成，因此策略的价值可表示如下：

$$E_{s_0\sim D}[V^\pi(s_0)]=E\left[\sum_{t=0}^{\infty}\gamma^t\varphi(s_t)|\pi\right]=$$
$$E\left[\sum_{t=0}^{\infty}\gamma^t\omega\varphi(s_t)|\pi\right]=\omega E\left[\sum_{t=0}^{\infty}\gamma^t\varphi(s_t)|\pi\right] \tag{13}$$

若有策略π的特征期望满足$[\mu(\pi)-\omega^t\mu_E]_2\leqslant\varepsilon$时式（14）成立，则该策略$\pi$是模仿学习法的一个解。

$$\left|E\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)|\pi_E\right]-E\left[\sum_{t=0}^{\infty}\gamma^t R(s_t)|\pi\right]\right|=$$
$$|\omega^t\mu(\pi)-\omega^t\mu_E|\leqslant|\omega|_2|\mu(\pi)-\mu_E|\leqslant\varepsilon \tag{14}$$

以上求解过程与通过计算积累奖励值获得最优策略的直接学习方法有本质区别。在多步决策中，基于积累奖励值的学习方法存在搜索空间过大、计算成本过高的问题，模仿学习法能够很好地解决多步决策中的这些问题。PFEIFFER等[57]将模仿学习与强化学习结合起来进行模型训练，通过模仿学习的目标驱动演示，可以显著提高强化学习过程中的探索能力，同时在避障方面也显现出较好的性能。HUSSEIN等[58]提出一种深度模仿学习方法来学习三维环境中的路径规划任务，利用主动学习对监督策略进行改进，以便将其推广到未知的动态环境中。传统的模仿学习不支持一个模型学习多个任务，XU等[59]提出共享多头模仿学习（Shared Multi-headed Imitation Learning，SMIL），使移动机器人在不同模型切换的情况中，使用一个模型学习多个任务。该方法将每个任务建模为子策略，并设计一个多头策略激活所有的子策略，以学习相关任务的共享信息。

2.2.3 基于值和策略相结合的强化学习方法

Actor-Critic算法[60]将基于值的算法和基于策略的算法的优点相融合，相比传统策略梯度法效率更高，是一种性能较好的强化学习算法。Actor-Critic算法分为Actor和Critic两个部分，其中，Actor由策略梯度法衍生而来，Critic由基于值的算法衍生而来。该算法的原理为：Actor根据概率选择行动，Critic为选择的行动反馈奖励，Actor再根据Critic的反馈修改选择行动的概率。Actor策略函数的参数更新公式如下：

$$\theta=\theta+\alpha\nabla_\theta\log_a\pi_\theta(s_t,a_t)Q(s,a,\omega) \tag{15}$$

使用均方差损失函数来更新Critic的网络参数ω：

$$\omega=\sum(r+\gamma V(s')-V(s,\omega))^2 \tag{16}$$

MUSE等[61]将Actor-Critic框架用于与平台无关的移动机器人路径规划任务中，使不同机器人平台的功能具有一定的可移植性，并用于新型机器人平台的快速原型设计。LACHEKHAB等[62]提出一种模糊Actor-Critic学习算法（Fuzzy Actor-Critic Learning Algorithm，FACL），使用基于模糊逻辑的控制器控制机器人从起点到终点的路径规划，并在模糊规则中基于概率选择机器人的下一行动，因此可在已知环境下较好地完成路径规划任务。SHAO等[63]提出一种基于广义优势估计（Generalized Advantage Estimator，GAE）的Actor-Critic算法，使智能体可以从多个过程中进行学习以节约训练时间，并利用GAE估计优势函数减少方差，从而提高了策略梯度的估计精度。

2.3 强化学习方法的局限性

虽然基于值的强化学习方法收敛速度快、学习效率高，但存在以下局限性：

1）当动作空间是连续动作空间时，如果采用基于值的方法，需要对动作空间离散化，进而会导致连续空

间到离散空间指数级的映射，从而产生维数灾难问题。

2）由基于值的方法最终得到的是一个确定性的策略，而最优策略可能是随机的，此时值函数法不适用。

3）值函数的一个微小的变动通常会导致一个原本被选择的动作反而不能被选择，这种变化会影响算法的收敛性。

与基于值的方法不同，基于策略的方法适用于高维或连续动作空间，并具有更好的收敛性，但也存在以下局限性：

1）需要完全序列样本才可以做算法迭代，训练慢、方差高，在多数情况下没有基于值的方法有效。

2）优化的梯度方向可能不是策略梯度的最优方向，因此易收敛到局部最优解，而非最优策略。

3）移动机器人的路径规划通常应用在离散动作空间中，基于策略的算法优势并不能显现出来，因此目前基于策略的算法在移动机器人路径规划上的应用较少。

表2给出了强化学习方法的代表方法、特点和优劣势对比结果。

表2　强化学习方法的代表方法、特点和优劣势对比

Table 2　Comparison of representative methods，characteristics and advantages and disadvantages of reinforcement learning methods

强化学习方法	代表方法	特点	优势	劣势
基于值的方法	TD	属于无模型的强化学习方法	1.结合动态规划和蒙特卡罗思想； 2.为后续算法的基础算法	应用效果一般
	Q-Learning	1.采用贪心策略； 2.采用off-policy，使用目标策略选取动作； 3.属于在线强化学习方法	1.所需参数少； 2.探索性强，注重获得最优结果时更适用； 3.采用贪心策略，可保证收敛性	1.收敛速度慢； 2.可能会导致维数灾难
	SARSA	1.采用贪心策略； 2.采用on-policy，使用行为策略选取动作； 3.属于在线强化学习方法	1.利用性强，注重获得的奖励时更适用； 2.采用贪心策略，可保证收敛性	收敛速度慢
	Dyna	1.为一类算法框架的总称； 2.结合基于模型和无模型的强化学习	建立环境模型代替真实环境，迭代虚拟样本函数值，实现实时学习规划	收敛速度慢
基于策略的方法	策略梯度法	对收益期望求梯度，沿梯度方向优化策略	适用于三维动作空间	1.需要全部的序列样本才能迭代参数； 2.易陷入局部最优解
	模仿学习法	从专家决策中学习得到最终最优决策	在多步决策中应用效果更好	1.误差积累； 2.分布不匹配
基于值与策略相结合的方法	Actor-Critic算法	1.Actor衍生于策略梯度法，Critic衍生于基于值的强化学习算法； 2.可以进行单步更新	相比传统的策略梯度法效率更高	容易产生局部最优解

3　基于深度强化学习的路径规划技术

3.1　深度强化学习与路径规划

强化学习的最终目的是通过最大化奖励值来获得最优策略，具有较强的决策能力。在越来越复杂的现实场景应用中，需要利用深度学习从原始大规模数据中提取高级特征，深度学习具有较强的感知能力，但缺乏一定的决策能力。深度强化学习[64]（Deep Reinforcement Learning，DRL）将强化学习的决策能力与深度学习的感知能力相结合，可以直接根据输入的信息进行控制，是一种更加接近人类思维的人工智能方法。

2013年，谷歌的人工智能研究团队DeepMind[65]将Q-Learning算法与卷积神经网络相结合，创新性地提出深度Q网络（Deep Q-Network，DQN）。DQN基础模型为一个卷积神经网络，并使用Q-Learning的变体进行训练。DQN对Q-Learning主要做了以下改进：

1）用卷积神经网络替代状态-动作对值函数$Q(s,a)$。具体地，使用参数为θ_i的值函数$Q(s,a;\theta_i)$，迭代i次后的损失函数表示如下：

$$L_i(\theta_i)=E_{s,a,r,s'}[(Y_i-Q(s,a;\theta_i))^2] \tag{17}$$

其中：Y_i近似表示值函数的优化目标。Y_i的计算公式如下：

$$Y_i=r+\gamma\max_{a'}Q(s',a;\theta^-) \tag{18}$$

在学习过程中通过θ_i更新θ^-，具体学习过程为对θ_i求偏导得到梯度：

$$\nabla_{\theta_i}L_i(\theta_i)=E_{s,a,r,s'}[(r+\gamma\max_{a'}Q(s',a;\theta^-)-Q(s,a;\theta_i))\nabla_{\theta_i}Q(s,a;\theta_i)] \tag{19}$$

2）使用经验回放技术。在每个时间步t时，存储智能体的经验样本$e_t=(s_t,a_t,r_t,s_{t+1})$到回放记忆单元$D=\{e_1,e_2,\cdots,e_t\}$中，通过重复采样历史数据增加样本使用率，可有效避免学习时参数震荡。

3)随机小批量从记忆单元D中取样。由于样本之间相关性高,因此直接从连续样本中学习是低效的,随机小批量取样降低了样本间的关联性,从而提升了算法的稳定性。

为更有效地将深度强化学习方法应用于路径规划,学者们尝试了许多改进思路。TAI等[66]针对没有障碍物地图和距离信息稀疏的情况,提出一个基于学习的无地图运动规划器,以稀疏的10维测距结果和目标相对于移动机器人坐标系的位置作为输入,连续转向命令作为输出,通过异步深度强化学习方法训练规划器,使训练和样本收集可以并行执行。该方法在极端复杂的环境中具有更好的稳定性。王珂等[67]基于深度强化学习提出一种基于最小深度信息选择的训练模式,通过运动学方程约束,优化状态空间的搜索与采集,提高了训练速率。李辉等[68]将深度卷积神经网络的特征提取能力与强化学习的决策能力相结合,提出一种基于深度强化学习的改进算法,该算法用近似值函数代替Q-Learning中的动作值函数,设计包含4层结构的深度卷积神经网络,以网络的输出代替传统的Q值表,解决了Q-Learning在状态空间较大时产生的维数灾难问题。

3.2 深度强化学习方法的局限性

目前,深度强化学习方法存在以下3个方面的局限性:

1)深度强化学习理论支撑不够。谷歌的DeepMind团队于2015年在《自然》杂志上发表的文章[69]虽然取得了较好的应用效果,但没有证明DQN的收敛性,并且到目前为止在DQN或其他深度强化学习方法的基础上的改进工作也没有很好地解决该问题。

2)样本采样率低。样本采样率低使得深度强化学习方法有时在实际应用中效果不佳。导致该问题的主要原因有两个:一是完成任务需要收集大量数据;二是训练过程中利用当前数据的有用信息效率低。

3)在连续动作空间中应用有限。目前主流的深度强化学习方法大多适用于离散动作空间,对于机器人的机械臂路径规划等连续动作空间[70]的任务还处于初步研究阶段,理论支撑不够,因此应用十分有限。

3.3 强化学习与深度强化学习的异同

强化学习通常使用马尔可夫决策过程进行描述,具体而言:机器处在一个环境中,每个状态为机器对当前环境的感知;机器只能通过动作来影响环境,在机器执行一个动作后,会使当前环境按某种概率转移到另一个状态;当前环境会根据潜在的奖赏函数反馈给机器一个奖赏。

深度强化学习是深度学习与强化学习的结合,具体而言是结合了深度学习的结构和强化学习的思想,但它的侧重点更多在强化学习上,解决的仍是决策问题,只是借助神经网络强大的表征能力拟合Q表或直接拟合策略以解决状态-动作空间过大或连续状态-动作空间问题。

4 基于强化学习的路径规划技术展望

虽然移动机器人路径规划技术已取得了大量的科研成果,并广泛地应用于实际场景中,但随着移动机器人应用领域的扩大和应用场景的复杂化,从目前的发展现状和未来的发展需求来看,基于强化学习方法的路径规划技术的下一步研究方向主要包括以下4个方面:

1)设计有效的奖励函数。强化学习通过最大化奖励值来获得最优策略,那么策略是否最优取决于奖励函数。现阶段奖励函数由专家学者凭借专业知识设计,面对路径规划领域日益复杂多变的应用环境,不合理的奖励函数也会使得到的最优策略不合理。有学者提出元学习(Meta Learning,ML)等方式,让智能体尝试在面对环境或任务变换的情况下,从合理的策略中不断完善奖励函数。因此,设计有效的奖励函数是未来发展的热点之一。

2)解决强化学习的探索-利用困境。一方面,探索的目的是通过不断探索新的环境信息来获得更高的奖励值,以此避免陷入局部最优解。另一方面,利用探索是指用已学习到的信息来选择奖励值最高的动作。探索-利用的困境即使得探索信息与利用信息两者之间得到平衡,目前常用的解决方法为ε-greedy算法[69]。该算法的基本原理是使智能体以ε为概率随机探索信息,并以$1-\varepsilon$为概率利用信息,通过不断的学习,ε会不断衰减以保证后期的学习效率。ε-greedy算法简单易实现,但随机探索效率低,因此如何解决强化学习的探索-利用困境有待进一步研究。

3)研究强化学习方法与常规方法的结合方法。每种强化学习方法在路径规划应用中都存在自身局限性,为了弥补单一方法的不足,通过不同方法之间相互结合的优势互补可以得到一些性能更好的方法,如传统路径规划算法、图形学算法、智能仿生学算法以及强化学习算法之间的有效结合,相互取长补短后均具有一定的发展前景。

4)将强化学习算法应用于多智能体协作的路径

规划研究。多智能体协作路径规划技术具有高灵活性、易部署、高协调性等优点,被广泛应用于机器人双臂协作路径规划、机器人足球赛、多无人机竞速赛等实际场景中。目前,对于单机器人路径规划研究的成果较多,而多机器人协作路径规划的成果相对较少,对应用中出现的碰撞、路径死锁、协调配合、花费代价大等一系列问题有待进一步解决。

5 结束语

本文阐述基于常规方法、强化学习方法及深度强化学习方法的路径规划技术,分类并对比强化学习方法的特点、优劣性及适用场合。针对强化学习方法应用于路径规划技术时存在的局限性,重点研究了将启发式思想、分层思想、模糊逻辑思想及多算法结合思想融入强化学习算法的改进思路。面对未来更加复杂的应用环境,下一步将从设计有效的奖励函数、解决强化学习的探索-利用困境等方面入手,对强化学习在路径规划技术中的应用进行更深入的研究。

参考文献

[1] 戴博,肖晓明,蔡自兴. 移动机器人路径规划技术的研究现状与展望[J]. 控制工程,2005,12(3):198-202.
DAI B,XIAO X M,CAI Z X. Current status and future development of mobile robot path planning technology[J]. Control Engineering of China,2005,12(3):198-202. (in Chinese)

[2] RAJA P. Optimal path planning of mobile robots:a review[J]. International Journal of Physical Sciences,2012,7(9):1314-1320.

[3] 王春颖,刘平,秦洪政. 移动机器人的智能路径规划算法综述[J]. 传感器与微系统,2018,37(8):5-8.
WANG C Y,LIU P,QIN H Z. Review on intelligent path planning algorithm of mobile robots[J]. Transducer and Microsystem Technologies,2018,37(8):5-8. (in Chinese)

[4] 张广林,胡小梅,柴剑飞,等. 路径规划算法及其应用综述[J]. 现代机械,2011(5):85-90.
ZHANG G L,HU X M,CHAI J F,et al. Summary of path planning algorithm and its application[J]. Modern Machinery,2011(5):85-90. (in Chinese)

[5] TAVARES R S,MARTINS T C,TSUZUKI M S G. Simulated annealing with adaptive neighborhood:a case study in off-line robot path planning[J]. Expert Systems with Applications,2011,38(4):2951-2965.

[6] LIU Y C,ZHAO Y J. A virtual-waypoint based artificial potential field method for UAV path planning [C]// Proceedings of 2016 IEEE Chinese Guidance,Navigation and Control Conference. Washington D. C. ,USA:IEEE Press,2016:949-953.

[7] GARCIA M A P,MONTIEL O,CASTILLO O,et al. Optimal path planning for autonomous mobile robot navigation using ant colony optimization and a fuzzy cost function evaluation[J]. Applied Soft Computing,2009,9(3):1102-1110.

[8] 周滔,赵津,胡秋霞,等. 复杂环境下移动机器人全局路径规划与跟踪[J]. 计算机工程,2018,44(12):208-214.
ZHOU T,ZHAO J,HU Q X,et al. Global path planning and tracking for mobile robot in cluttered environment[J]. Computer Engineering,2018,44(12):208-214. (in Chinese)

[9] LEE T K,BAEK S H,CHOI Y H,et al. Smooth coverage path planning and control of mobile robots based on high-resolution grid map representation [J]. Robotics and Autonomous Systems,2011,59(10):801-812.

[10] 刘传领. 基于势场法和遗传算法的机器人路径规划技术研究[D]. 南京:南京理工大学,2012.
LIU C L. Researches on technologies for robot path planning based on artificial potential field and genetic algorithm[D]. Nanjing:Nanjing University of Science and Technology,2012. (in Chinese)

[11] ZHU A,YANG S X. A neural network approach to dynamic task assignment of multirobots [J]. IEEE Transactions on Neural Networks,2006,17(5):1278-1287.

[12] RASHID R,PERUMAL N,ELAMVAZUTHI I,et al. Mobile robot path planning using ant colony optimization[C]// Proceedings of the 2nd IEEE International Symposium on Robotics and Manufacturing Automation. Washington D. C. ,USA:IEEE Press,2016:1-6.

[13] 胡章芳,孙林,张毅,等. 一种基于改进QPSO的机器人路径规划算法[J]. 计算机工程,2019,45(4):281-287.
HU Z F,SUN L,ZHANG Y,et al. A robot path planning algorithm based on improved QPSO[J]. Computer Engineering,2019,45(4):281-287. (in Chinese)

[14] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine Learning,1988,3(1):9-44.

[15] 赵冬斌,邵坤,朱圆恒,等. 深度强化学习综述:兼论计算机围棋的发展[J]. 控制理论与应用,2016,33(6):701-717.
ZHAO D B,SHAO K,ZHU Y H,et al. Review of deep reinforcement learning and discussions on the development of computer go[J]. Control Theory & Applications,2016,33(6):701-717. (in Chinese)

[16] BELLMAN R. Dynamic programming and lagrange multipliers[J]. Proceedings of the National Academy of Sciences,1956,42(10):767-769.

[17] WERBOS P J. Advanced forecasting methods for global crisis warning and models of intelligence [J]. General Systems Yearbook,1977,22(12):25-38.

[18] WATKINS C J C H,DAYAN P. Q-learning[J]. Machine Learning,1992,8(3/4):279-292.

[19] RUMMERY G A,NIRANJAN M. On-line q-learning using connectionist systems[M]. Cambridge,UK:University of Cambridge,1994.

[20] BERTSEKAS D P,TSITSIKLIS J N. Neuro-dynamic programming:an overview[C]//Proceedings of the 34th IEEE Conference on Decision and Control. Washington D. C. ,USA:IEEE Press,1995:560-564.

[21] KOCSIS L,SZEPESVARI C. Bandit based Monte-Carlo planning[C]//Proceedings of 2016 European Conference on Machine Learning. Berlin,Germany:Springer,2006:282-293.

[22] LEWIS F L, VRABIE D. Reinforcement learning and adaptive dynamic programming for feedback control[J]. IEEE Circuits and Systems Magazine, 2009, 9(3): 32-50.
[23] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of 2014 International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2014: 387-395.
[24] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2016: 1928-1937.
[25] ROUGIER J. Comment on "ensemble averaging and the curse of dimensionality"[J]. Journal of Climate, 2018, 31(21): 9015-9016.
[26] SUTTON R S. Generalization in reinforcement learning: successful examples using sparse coarse coding [C]// Proceedings of 1996 International Conference Neural Information Processing Systems. Cambridge, USA: MIT Press, 1996: 1038-1044.
[27] NAIR D S, SUPRIYA P. Comparison of temporal difference learning algorithm and Dijkstra's algorithm for robotic path planning [C]//Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems. Washington D. C., USA: IEEE Press, 2018: 1619-1624.
[28] MARTIN J, WANG J K, ENGLOT B. Sparse Gaussian process temporal difference learning for marine robot navigation[EB/OL]. [2020-12-11]. https://arxiv. org/abs/1810. 01217.
[29] LI S D, XU X, ZUO L. Dynamic path planning of a mobile robot with improved Q-learning algorithm[C]//Proceedings of 2015 IEEE International Conference on Information and Automation. Washington D. C., USA: IEEE Press, 2015: 409-414.
[30] 刘智斌,曾晓勤,刘惠义,等. 基于BP神经网络的双层启发式强化学习方法[J]. 计算机研究与发展,2015,52(3):579-587.
LIU Z B, ZENG X Q, LIU H Y, et al. A heuristic two-layer reinforcement learning algorithm based on BP neural networks[J]. Journal of Computer Research and Development, 2015, 52(3): 579-587. (in Chinese)
[31] JIANG L, HUANG H Y, DING Z H. Path planning for intelligent robots based on deep Q-learning with experience replay and heuristic knowledge[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(4): 1179-1189.
[32] 周文吉,俞扬. 分层强化学习综述[J]. 智能系统学报,2017,12(5):590-594.
ZHOU W J, YU Y. Summarize of hierarchical reinforcement learning[J]. CAAI Transactions on Intelligent Systems, 2017, 12(5): 590-594. (in Chinese)
[33] BUITRAGO-MARTINEZ A, DE LA ROSA R F, LOZANO-MARTINEZ F. Hierarchical reinforcement learning approach for motion planning in mobile robotics [C]//Proceedings of 2013 Latin American Robotics Symposium and Competition. Washington D. C., USA: IEEE Press, 2013: 83-88.
[34] 刘志荣,姜树海,袁雯雯,等. 基于深度Q学习的移动机器人路径规划[J]. 测控技术,2019,38(7):24-28.
LIU Z R, JIANG S H, YUAN W W, et al. Robot path planning based on deep Q-learning[J]. Measurement & Control Technology, 2019, 38(7): 24-28. (in Chinese)
[35] 裴道武. 关于模糊逻辑与模糊推理逻辑基础问题的十年研究综述[J]. 工程数学学报,2004,21(2):249-258.
PEI D W. A survey of ten years' studies on fuzzy logic and fuzzy reasoning[J]. Chinese Journal of Engineering Mathematics, 2004, 21(2): 249-258. (in Chinese)
[36] LUVIANO D, YU W. Continuous-time path planning for multi-agents with fuzzy reinforcement learning[J]. Journal of Intelligent & Fuzzy Systems, 2017, 33(1): 491-501.
[37] BOWLING M, VELOSO M. Multiagent learning using a variable learning rate[J]. Artificial Intelligence, 2002, 136(2): 215-250.
[38] WEN S H, CHEN J H, LI Z, et al. Fuzzy Q-learning obstacle avoidance algorithm of humanoid robot in unknown environment[C]//Proceedings of 2018 Chinese Control Conference. Washington D. C., USA: IEEE Press, 2018: 5186-5190.
[39] 葛媛,布朋生,刘强. 模糊强化学习在机器人导航中的应用[J]. 信息技术,2009,33(10):127-130.
GE Y, BU P S, LIU Q. Application of fuzzy Q-learning in robot navigation[J]. Information Technology, 2009, 33(10): 127-130. (in Chinese)
[40] 朴松昊,洪炳熔. 一种动态环境下移动机器人的路径规划方法[J]. 机器人,2003,25(1):18-21,43.
PIAO S H, HONG B R. A path planning approach to mobile robot under dynamic environment[J]. Robot, 2003, 25(1): 18-21, 43. (in Chinese)
[41] MEERZA S I A, ISLAM M, UZZAL M M. Q-learning based particle swarm optimization algorithm for optimal path planning of swarm of mobile robots[C]//Proceedings of 2019 International Conference on Advances in Science, Engineering and Robotics Technology. Washington D. C., USA: IEEE Press, 2019: 1-5.
[42] SHI Z G, TU J, ZHANG Q, et al. The improved Q-Learning algorithm based on pheromone mechanism for swarm robot system [C]//Proceedings of the 32nd Chinese Control Conference. Washington D. C., USA: IEEE Press, 2013: 6033-6038.
[43] YAO Q F, ZHENG Z Y, QI L, et al. Path planning method with improved artificial potential field—a reinforcement learning perspective[J]. IEEE Access, 2020, 8: 135513-135523.
[44] LIU Z Y, LAN F, YANG H B. Partition heuristic RRT algorithm of path planning based on Q-learning[C]//Proceedings of 2019 Advanced Information Technology, Electronic and Automation Control Conference. Washington D. C., USA: IEEE Press, 2019: 386-392.
[45] 王子强,武继刚. 基于RDC-Q学习算法的移动机器人路径规划[J]. 计算机工程,2014,40(6):211-214.
WANG Z Q, WU J G. Mobile robot path planning based on RDC-Q learning algorithm[J]. Computer Engineering, 2014, 40(6): 211-214. (in Chinese)
[46] ZOU Q J, ZHANG Y, LIU S H. A path planning algorithm based on RRT and SARSA(λ) in unknown and complex conditions[C]//Proceedings of 2020 Chinese Control and

Decision Conference. Washington D. C. ,USA:IEEE Press, 2020:2035-2040.

[47] XU D, FANG Y C, ZHANG Z Y, et al. Path planning method combining depth learning and sarsa algorithm[C]// Proceedings of the 10th International Symposium on Computational Intelligence and Design. Washington D. C. , USA:IEEE Press, 2017:77-82.

[48] FATHINEZHAD F, DERHAMI V, REZAEIAN M. Supervised fuzzy reinforcement learning for robot navigation[J]. Applied Soft Computing, 2016, 40:33-41.

[49] DABOONI S, WUNSCH D. Heuristic dynamic programming for mobile robot path planning based on Dyna approach[C]// Proceedings of 2016 International Joint Conference on Neural Networks. Washington D. C. , USA:IEEE Press, 2016:3723-3730.

[50] VIET H H, AN S H, CHUNG T C. Dyna-Q-based vector direction for path planning problem of autonomous mobile robots in unknown environments[J]. Advanced Robotics, 2013, 27(3):159-173.

[51] HWANG K S, JIANG W C, CHEN Y J. Adaptive model learning method for reinforcement learning[C]//Proceedings of SICE'12. Washington D. C. , USA:IEEE Press, 2012: 1277-1280.

[52] 刘建伟,高峰,罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报,2019,42(6):1406-1438.
LIU J W, GAO F, LUO X L. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6):1406-1438. (in Chinese)

[53] WANG Q Z, XU D, SHI L Y. A review on robot learning and controlling: imitation learning and human-computer interaction[C]//Proceedings of the 2013 Chinese Control and Decision Conference. Washington D. C. , USA:IEEE Press, 2013:2834-2838.

[54] LIU Y D, ZHANG W Z, CHEN F M, et al. Path planning based on improved deep deterministic policy gradient algorithm[C]//Proceedings of the 3rd Information Technology, Networking, Electronic and Automation Control Conference. Washington D. C. , USA:IEEE Press, 2019:295-299.

[55] PAUL S, VIG L. Deterministic policy gradient based robotic path planning with continuous action spaces[C]// Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Washington D. C. , USA: IEEE Press, 2017:725-733.

[56] ZHENG S F, LIU H. Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation[J]. IEEE Access, 2019, 7:147755-147770.

[57] PFEIFFER M, SHUKLA S, TURCHETTA M, et al. Reinforced imitation: sample efficient deep reinforcement learning for Mapless navigation by leveraging prior demonstrations[J]. IEEE Robotics and Automation Letters, 2018, 3(4):4423-4430.

[58] HUSSEIN A, ELYAN E, GABER M M, et al. Deep imitation learning for 3D navigation tasks[J]. Neural Computing and Applications, 2018, 29(7):389-404.

[59] XU J H, LIU Q W, GUO H, et al. Shared multi-task imitation learning for indoor self-navigation[C]//Proceedings of 2018 IEEE Global Communications Conference. Washington D. C. , USA:IEEE Press, 2018:1-7.

[60] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of Actor-Critic reinforcement learning: standard and natural policy gradients[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2012, 42(6):1291-1307.

[61] MUSE D, WERMTER S. Actor-Critic learning for platform-independent robot navigation[J]. Cognitive Computation, 2009, 1(3):203-220.

[62] LACHEKHAB F, TADJINE M. Goal seeking of mobile robot using fuzzy actor critic learning algorithm[C]// Proceedings of the 7th International Conference on Modelling, Identification and Control. Washington D. C. , USA:IEEE Press, 2015:1-6.

[63] SHAO K, ZHAO D B, ZHU Y H, et al. Visual navigation with Actor-Critic deep reinforcement learning[C]// Proceedings of 2018 International Joint Conference on Neural Networks. Washington D. C. , USA:IEEE Press, 2018:1-6.

[64] 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(1):1-27.
LIU Q, ZHAI J W, ZHANG Z Z, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1):1-27. (in Chinese)

[65] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[EB/OL]. [2020-12-11]. https://arxiv. org/abs/1312. 5602v1.

[66] TAI L, PAOLO G, LIU M. Virtual-to-real deep reinforcement learning: continuous control of mobile robots for mapless navigation[C]//Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. Washington D. C. , USA:IEEE Press, 2017:31-36.

[67] 王珂,卜祥津,李瑞峰,等. 景深约束下的深度强化学习机器人路径规划[J]. 华中科技大学学报(自然科学版),2018,46(12):77-82.
WANG K, BU X J, LI R F, et al. Path planning for robots based on deep reinforcement learning by depth constraint[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2018, 46(12):77-82. (in Chinese)

[68] 李辉,祁宇明. 一种复杂环境下基于深度强化学习的机器人路径规划方法[J]. 计算机应用研究,2020,37(S1): 129-131.
LI H, QI Y M. Robot path planning method based on deep reinforcement learning in complex environment[J]. Application Research of Computers, 2020, 37(S1):129-131. (in Chinese)

[69] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.

[70] GU S, LILLICRAP T, SUTSKEVER I, et al. Continuous deep Q learning with model-based acceleration[EB/OL]. [2020-12-11]. https://www. cnblogs. com/wangxiaocvpr/p/5664795. html.

编辑 陆燕菲