基于强化学习的路径规划算法研究

22级电子信息 罗梦丽 20222007110438

**摘要：**在自动驾驶汽车技术的诸多研究中，路径规划是自动驾驶汽车的重要研究内容，也是当前自动驾驶汽车领域研究的热点和难点之一，其目的是在各种复杂环境下能够获得一条连接起始点与目标点的最优无碰撞路径。文章介绍了几种基于强化学习的路径规划算法及其基本原理与特性，回顾了不同算法在路径规划算法领域的研究和应用，并对自动驾驶汽车路径规划的未来研究方向做出了展望。

**关键词：**自动驾驶汽车；路径规划；强化学习

**Abstract:** Among many researches on self-driving car technology, path planning is an important research content of self-driving cars, and also one of the hot spots and difficulties in the current research field of self-driving cars. Its purpose is to obtain an optimal collision-free path connecting the starting point and the target point in various complex environments.This paper introduces several path planning algorithms based on reinforcement learning and their basic principles and characteristics, and reviews the research and application of different algorithms in the field of path planning algorithms.The future research direction of autonomous vehicle path planning is also presented.

**Key words:** Autonomous vehicle; Path planning; Reinforcement learning

1. **绪论**

**1.1、研究背景**

自动驾驶汽车使用传感器感知环境，并依照合理的算法在复杂环境中实现自主运动，使其能在道路上安全、高效地行驶。作为自动驾驶汽车研究的一个重要环节，路径规划就是根据给定的环境模型，在一定的约束条件下，利用路径规划算法规划出一条连接车辆当前位置和目标位置的无碰撞路径。在此基础上可按照不同需求选择一条最优路径。

自动驾驶汽⻋的路径规划算法已发展多年，为应对各种场景和问题，诞⽣了多种规划算法。路径规划的核心就是算法的设计，路径规划算法目前已经得到了广泛的关注。随着各个领域的不断发展，传统路径规划算法在复杂环境模型下存在规划速度慢、数据计算量大等问题，难以满足路径规划实时性、准确性的要求。在此前，针对高级驾驶辅助系统(ADAS)和自动驾驶应用，已经开展了大量基于深度学习的研究活动。随着人工智能的不断进步，强化学习领域已经成为一个强大的学习框架具有学习复杂的高维环境的能力，使用强化学习或深度强化学习应用于自主驾驶或自动驾驶领域是一个新兴热点领域。

**1.2、路径规划算法分类**

根据对环境信息的把握程度可把路径规划划分为基于先验完全信息的全局路径规划和基于传感器信息的局部路径规划。其中，从获取障碍物信息是静态或是动态的角度看，全局路径规划属于静态规划(又称离线规划) ，局部路径规划属于动态规划(又称在线规划)。全局路径规划需要掌握所有的环境信息，根据环境地图的所有信息进行路径规划; 局部路径规划只需要由传感器实时采集环境信息，了解环境地图信息，然后确定出所在地图的位置及其局部的障碍物分布情况，从而可以选出从当前结点到某一子目标结点的最优路径。

依照路径规划算法在原理上的不同，可将算法大致分为基于搜索的路径规划算法、基于采样的路径规划算法、基于智能算法的路径规划算法以及基于强化学习的路径规划算法等。其中，基于搜索的路径规划算法主要包括 Dijkstra 算法、A\* 算法等；基于采样的路径规划算法主要包括 RRT 算法、PRM算法等；基于智能算法的路径规划算法主要包括遗传算法、蚁群算法等；基于强化学习的路径规划算法主要包括Q-Learning、DPG、DDPG等。

**1.3、路径规划的一般步骤**

路径规划的一般步骤主要包括环境建模、路径搜索、路径平滑三个环节。

1. 环境建模。环境建模是路径规划的重要环节，目的是建立一个便于计算机进行路径规划所使用的环境模型，即将实际的物理空间抽象成算法能够处理的抽象空间，实现相互间的映射。
2. 路径搜索。路径搜索阶段是在环境模型的基础上应用相应算法寻找一条行走路径，使预定的性能函数获得最优值。
3. 路径平滑。通过相应算法搜索出的路径并不一定是一条运动体可以行走的可行路径，需要作进一步处理与平滑才能使其成为一条实际可行的路径。

**1.4、强化学习的概念和模型**

强化学习（Reinforcement Learning ,RL）是一种重要的机器学习方法，机器学习分为监督学习、无监督学习、强化学习三种。强化学习是系统从环境到行为映射的学习，目的是使奖励信号（强化信号）函数值最大。换句话说，强化学习是一种学习如何从状态映射到行为以使得获取的奖励最大的学习机制。一个动作需要不断在环境中进行实验，环境对动作做出奖励，系统通过环境的奖励不断优化行为，反复实验，延迟奖励。

主体(Agent)通过与环境(environment)交互进行学习，交互包括行动（action），奖励(reward)，状态(state)。交互过程表述如下：每一步主体都根据策略选择一个行动方式执行，比如说向前走、向后走，然后感知做出该行动后的立即奖励，还有下一步的状态，该状态也就是环境的状态，通过已有的经验再修改自己的策略做出下一个动作，经验就是根据这一步一步的动作学习来的，每一步都积累一些奖励值(立即奖励)，目标就是让累积的奖励值最大。

1. **研究现状**

强化学习根据策略优化的方式不同，可分为基于价值的方法，基于策略的方法和Actor-Critic方法。

**2.1、基于价值的方法**

**2.1.1、Q-Learning**

1. 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，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。求解和蒙特卡罗法类似，都是价值迭代，即通过价值函数的更新，来更新策略，通过策略来产生新的状态和即时奖励，进而更新价值函数。一直进行下去，直到价值函数和策略都收敛。由于其简单和易于实现的优点，Q-Learning已成为移动机器人的路径规划领域使用最广泛的算法。

Q-Learning算法对于小型强化学习问题非常灵活有效，但是对于大数据异常复杂的状态和可选动作，使其需要维护的Q-table异常大而陷入维度爆炸。

在路径规划中容易出现的非凸性问题中，Roozbeh Bazargani等人提出了一种Q-learning和MPC结合的在线规划算法[1]，在无障碍物路径生成过程中引入Q-learning，将reward函数由-1改为-||a||2使路径更加平滑，从而使MPC在有障碍物的情况下数值求解中非凸性优化问题得到了缓解，得以利用终端成本和约束条件保证了稳定性和可行性。

在复杂未知的环境下，传统Q-learning算法会因为状态和动作异常复杂使Q表异常庞大而陷入维度爆炸。针对这一问题，Yuxiang Wang等人设计了一种改进的Q-learning算法[2]，在网格地图环境下，建立了改进的动作步距与奖励值之间的定量函数关系来作为奖励函数，将重力势场应用到Q-learning算法的更新步骤中，并且采用深度神经网络代替Q-table，将其改进为DQN算法，利用其经验回放和固定目标网络两个核心元素，改进了路径规划的结果，解决了维度爆炸的问题。

针对RRT\*效率低，所需时间长的问题，Tomoya Kawabe等人在文中提出了一种RRT\*和Q-learning结合的方法[3]，用于多个机器臂无碰撞轨迹规划，使用RRT\*算法寻找轨迹目标姿势，通过Q-learning算法可以避免机械臂之间的碰撞，经过仿真实验，验证出该方法轨迹规划时间比单独使用RRT\*算法减少了20%。

**2.1.2、DQN**

Deep Q-Learning算法的基本思路来源于Q-Learning。但是和Q-Learning不同的地方在于，它的Q值的计算不是直接通过状态值s和动作来计算，而是通过的Q网络来计算的。DQN主要使用的技巧是经验回放（experience replay）,即将每次和环境交互得到的奖励与状态更新情况都保存起来，用于后面目标Q值的更新。通过经验回放得到的目标Q值和通过Q网络计算的Q值肯定是有误差的，可以通过梯度的反向传播来更新神经网络的参数w，当w收敛后，我们的就得到的近似的Q值计算方法，进而贪婪策略也就求出来了。

DQN对价值函数做了近似表示，因此有了解决大规模强化学习问题的能力。但是它并不一定能保证Q网络的收敛，这导致训练出的模型效果很差。

传统DQN算法在通过神经网络快速计算Q值来优化最优状态时不可避免存在过度估计的问题，针对这一情况，文献[4]中将行为选择和目标Q值计算解耦，选择当前网络当前动作的最大Q值用于目标Q值的计算，考虑到动作值函数对静态环境的影响，最终结果为神经网络价值函数的输出和环境状态值两部分的线性组合，改进后的方法具有更明显的优势。

**2.1.3、Double DQN**

Double DQN算法在DQN算法的基础上，使用两个Q网络，一个当前Q网络Q用来选择动作，更新模型参数，另一个目标Q网络Q′用于计算目标Q值。目标Q网络的网络参数不需要迭代更新，而是每隔一段时间从当前Q网络Q复制过来，即延时更新，这样可以减少目标Q值和当前的Q值相关性，并且通过解耦目标Q值动作的选择和目标Q值的计算这两步，来达到消除过度估计的问题。

针对传统DQN算法存在的动作值过度估计的问题，[5]中通过解耦目标Q值动作的选择和目标Q值的计算的方法，将Double DQN算法应用于USV的路径规划中，并指出Double DQN算法能有效地处理复杂的环境信息,使路径规划最优。[21]中将基于经验的启发式搜索算法（EBHS）和Double DQN结合在一起，既克服了错误率的问题，又保留了预学习最优策略的优点，缩短了规划时间。

针对传统DQN算法收敛速度慢，训练时间长这一问题，[6]中将Population Base Training(PBT)超参数搜索算法应用于Double DQN中，可以在线调整学习速率，使收敛更快。

为了能够适应未知复杂多变的环境，并使规划出的路径更加光滑，[7]中将Double DQN算法中卷积神经网络设置了两个输入层，提高了不同维度的环境信息的输入适应性，并且使用NURBS算法使DDQN输出的离散动作更加平滑。

**2.1.4、Dueling DQN**

Dueling DQN算法在DQN的基础上，通过优化神经网络的结构得到。其将Q网络分成两部分，一部分仅仅与状态有关，与具体要采取的动作无关，叫做价值函数，另一部分同时与状态和动作有关，叫做优势函数，最终的价值函数表示为两部分函数的线性组合，相比于单独动作价值的评估更准确。

针对DQN算法存在的抽样方法没有充分利用存储信息和在训练过程中行动选择的随机性太大,收敛太慢的问题，[8]中Dueling DQN网络中引入优先级采样机制，建立具有优先级采样的Dueling DQN模型，使收敛速度变快，效率得到提高。

为了解决收敛速度慢、过度估计的问题,[9]中提出了一种Variant of Double dueling DQN（V-D D3QN）算法，将Doubel DQN和Dueling DQN结合在一起，随机选择一个dueling DQN网络在每个时间步长更新其参数,利用剩余的dueling DQN网络确定更新目标。文献[10]中提出了一种基于分数位回归的改进DQN算法（QR-D3QN），在QR-DQN的基础上，对目标Q值的计算方法进行了改进，以减小值高估的影响，结合对抗网络和添加奖励经验取样,提高有效数据的利用效率，使机器人可以迅速避免障碍找到最优路径。

**2.2、基于策略的方法**

**2.2.1、 TRPO**

由于随机策略梯度方法存在学习速率难以确定的问题，置信域策略优化（Trust Region Policy Optimization,TRPO）被提出，用来确定一个合适的更新步长，使得策略更新后，回报函数的值单调不减。

存在缺陷是无法处理大参数矩阵，二阶优化很慢，难以实现和调试。因此在路径规划算法中难以得到应用。

**2.2.2、 PPO**

PPO是对TRPO算法的改进，通过检验KL散度值来决定λ是增大还是减小，形成PPO-Penalty算法,直接截断用于策略梯度的目标函数，从而得到更保守的更新，形成PPO-Clip算法。

针对传统PPO算法高方差的问题，[11]中用自适应参数来计算重采样概率，通过限制重采样范围，提高数据利用率和算法收敛速度，采用延迟策略更新的方法，产生较小的方差。

为了提高稳定性和收敛速度，[12]中对传统PPO算法的网络结构进行了改进，由原来的两个全连接层改为三个全连接层，神经元个数分别为128、128、64，激活函数选择tanh，并且在目标函数中引入当前策略的转换概率和原策略下的转换概率之比。[13]中提出了基于频率分解的PPO算法（FD-PPO），并设计了启发式奖励函数，该方法在各种复杂环境表现出优势。[14]中仅在测试阶段使用PPO算法，在训练阶段只是用来给出路径点，有效的减少了训练时间，使算法收敛更快。

**2.3、Actor-Critic方法**

Actor-Critic方法的主要思想是用近似的值函数来指导策略参数更新。首先，Actor针对当前地状态，通过运行策略函数选择一个行为；其次，Critic采用值函数对Actor的行为进行评价；然后，Actor根据Critic的评价，调整自己的策略；最后，Critic根据环境给出的回报计算出一个更新的目标值，来调整自己的评价策略。该方法的引入主要是为了解决Policy Gradient算法中回合更新效率低的问题，即它仅在一个回合完成之后，才能更新参数，Actor-Critic算法可以实现单步更新，收敛要快很多。

**2.3.1、DPG**

确定策略梯度（Deterministic Policy Gradiant,DPG）是最常用的连续控制方法。DPG是一种Actor-Critic方法，它有一个策略网络，一个价值网络。策略网络控制智能体做运动，它基于状态s做出动作a。价值网络不控制智能体，只是基于状态s给动作a打分，从而指导策略网络做出改进。[15]中将DPG算法用于机器人连续行动空间的路径规划，克服了传统算法在非结构化环境中难以获得目标位置的缺陷。

确定策略梯度可以处理连续动作空间的任务，但是无法直接从高维输入中学习策略。

**2.3.2、DDPG**

在DPG算法的基础上引入DQN算法，就形成了深度确定性策略梯度算法（Deep Deterministic Policy Gradiant,DDPG）。DDPG分别用神经网络逼近行为值函数，实现直接从原始数据中进行端对端的学习，还使用了批量标准化的深度学习技术来应对不同量纲问题，实现对不同任务、不同类型的数据进行有效学习。

为了提高稳定性和收敛速度，[16]中运用一种multi-critic DDPG算法，用k值的平均值近似代替行为价值函数，改进后的算法泛化能力优于传统DDPG，成功率也增加了。[17]中提出了TPR-DDPG算法，用不同的归一化预处理方法改进奖励函数，并在策略网络中加入BatchNorm层，使机器人能够在复杂环境中迅速到达目标位置，并且算法更加稳定。[18]中通过引入Prioritized Experience Replay (PER)策略，提高了算法的收敛速度，减少了路径长度。[19]中提出用权重区分不同样本，不再是均匀的随机选择，减少算法试错的次数，从而节约了计算时间。

传统的DDPG算法的回放缓冲容量固定，无法实现多特征样本数据学习能力的提高，抑制了算法的学习能力。针对这一问题，[20]中构造了基于学习曲线的经验池容量变化函数，可以动态的调整回放缓冲区的容量，从而提高了算法的学习能力。

1. **比较与分析**

基于值和策略的方法对比:

1. 策略搜索方法具有更好的收敛性。有些值函数在后期会一直围绕最优值持续小幅震荡而不收敛。而基于策略的方法遵循的是梯度法，它会一直朝着优化策略的方向进行更新，试图收敛。
2. 策略搜索方法更简单，在某些情况下，使用基于值函数方法求解最优策略非常复杂甚至无效。
3. 策略搜索方法可以学到随机策略，往往是最优策略。而使用基于值函数的方法学到的大多数策略都是确定性策略，无法获得最优解。
4. 基于值函数的方法只适用于离散动作空间，而基于策略的方法适用于连续动作空间。
5. **展望**

随着科学技术的不断发展，路径规划技术面对的环境将更为复杂多变。这就要求路径规划算法要具有迅速响应复杂环境变化的能力。同时随着人工智能的不断进步，强化学习领域已经成为一个强大的学习框架，具有学习复杂的高维环境的能力，使用强化学习或深度强化学习应用于自主驾驶或自动驾驶领域是一个新兴热点领域。现有的路径规划算法在静态全局路径规划及已知环境信息的动态路径规划问题中进行了大量的研究，不同的算法在其进行路径规划时都具有一定优势。然而，对于未知环境下的路径规划，关于不确定环境因素与多车协同的路径规划研究较少。这不是目前单个或单方面算法所能解决问题，因此在未来的路径规划技术中，除了研究发现新的路径规划算法外，还有两个方面值得关注：一方面是可以将路径规划算法进行有效结合( 即混合算法) 。任何的单一路径规划算法都不可能解决所有实际应用中的路径规划问题，特别是在面对交叉学科的新问题时，研究新算法的难度大，路径规划算法间的优势互补为解决这一问题提供了可能。另一方面是可以将环境建模技术和路径规划算法的结合。面对复杂的二维甚至三维连续动态环境信息时，算法所能做的是有限的，好的建模技术和优秀路径规划算法相结合将成为解决这一问题的一种方法。因此，通过多种算法融合，在高维、动态、复杂以及多变的环境中进行路径规划的研究，将成为自动驾驶汽车算法研究的主流方向。

参考文献

1. R. Bazargani, M. S. Nadrabadi, E. Firouzmand, I. Sharifi and H. A. Talebi, "Motion Planning of Autonomous Vehicles Using Obstacle-Free Q-Learning Path Generator and Model Predictive Control," 2021 9th RSI International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, Islamic Republic of, 2021, pp. 230-235, doi: 10.1109/ICRoM54204.2021.9663449.
2. Y. Wang, S. Wang, Y. Xie, Y. Hu and H. Li, "Q-learning-based Collision-free Path Planning for Mobile Robot in Unknown Environment," 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 2022, pp. 1104-1109, doi: 10.1109/ICIEA54703.2022.10006304.
3. Tomoya Kawabe;Tatsushi Nishi.A Flexible Collision-Free Trajectory Planning for Multiple Robot Arms by Combining Q-Learning and RRT[A].2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)[C],2022
4. Cheng Yi;Meng Qi.Research on Virtual Path Planning Based on Improved DQN[A].2020 IEEE International Conference on Real-time Computing and Robotics (RCAR)[C],2020
5. Zhijian Huang;Huiqun Lin;Guichen Zhang.The USV Path Planning Based on an Improved DQN Algorithm[A].2021 International Conference on Networking, Communications and Information Technology (NetCIT)[C],2021
6. Xu Hou;Meng Joo Er;Tianhe Liu.Path Planning of Unmanned Surface Vehicle Port Docking Based on Improved Double Deep Q-Network[A].2022 5th International Conference on Intelligent Autonomous Systems (ICoIAS)[C],2022
7. Zhenzhong Chu;Fulun Wang;Tingjun Lei;Chaomin Luo.Path Planning Based on Deep Reinforcement Learning for Autonomous Underwater Vehicles Under Ocean Current Disturbance[J].IEEE Transactions on Intelligent Vehicles,2023,Vol.8(1): 108-120
8. Zhijian Huang;Sen Liu;Guichen Zhang.The USV Path Planning of Dueling DQN Algorithm Based on Tree Sampling Mechanism[A].2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)[C],2022
9. Ying Huang，GuoLiang Wei，YongXiong Wang.V–D D3QN:The Variant of Double Deep Q-Learning Network with Dueling Architecture[A].第37届中国控制会议[C],2018
10. Lun Zhou;Ke Wang;Hang Yu;Zhen Wang.Path planning of improved DQN based on quantile regression[A].2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT)[C],2022
11. J. Xu, X. Yan, C. Peng, X. Wu, L. Gu and Y. Niu, "UAV Local Path Planning Based on Improved Proximal Policy Optimization Algorithm," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096457.
12. Z. Wang, Z. Hu, Y. Yang and Y. Yin, "Research on PPO algorithm in solving AUV path planning problems," 2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shanghai, China, 2021, pp. 73-79, doi: 10.1109/AINIT54228.2021.00024.
13. Chenyang Qi;Chengfu Wu;Lei Lei;Xiaolu Li;Peiyan Cong.UAV path planning based on the improved PPO algorithm[A].2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE)[C],2022
14. Xinming Tang;Yi Chai;Qie Liu.A 2D UAV Path Planning Method Based on Reinforcement Learning in the Presence of Dense Obstacles and Kinematic Constraints[A].2022 IEEE 11th Data Driven Control and Learning Systems Conference (DDCLS)[C],2022
15. Paul, Somdyuti;Vig, Lovekesh.Deterministic Policy Gradient Based Robotic Path Planning with Continuous Action Spaces[J].2017 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION WORKSHOPS (ICCVW 2017),2017,: 725-733
16. Runjia Wu;Fangqing Gu;Jie Huang.A multi-critic deep deterministic policy gradient UAV path planning[A].2020 16th International Conference on Computational Intelligence and Security (CIS)[C],2020
17. Yaping Zhao;Xiuqing Wang;Ruiyi Wang;Yunpeng Yang;Feng Lv.Path Planning for Mobile Robots Based on TPR-DDPG[A].2021 International Joint Conference on Neural Networks (IJCNN)[C],2021
18. Yijian Tan;Yunhan Lin;Tong Liu;Huasong Min.PL-TD3: A Dynamic Path Planning Algorithm of Mobile Robot[A].2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)[C],2022
19. Yuansheng Dong;Xingjie Zou.Mobile Robot Path Planning Based on Improved DDPG Reinforcement Learning Algorithm[A].2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)[C],2020
20. Yandong Liu;Wenzhi Zhang;Fumin Chen;Jianliang Li.Path planning based on improved Deep Deterministic Policy Gradient algorithm[A].2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)[C],2019
21. Julian Bernhard;Robert Gieselmann;Klemens Esterle;Alois Knol.Experience-Based Heuristic Search: Robust Motion Planning with Deep Q-Learning[A].2018 21st International Conference on Intelligent Transportation Systems (ITSC)[C],2018