DOI:10.19651/j.cnki.emt.2315067

基于深度 Q 网络的海上环境智能路径规划*

李鹏程[1] 周远国[1] 杨国卿[2]
(1.西安科技大学通信与信息工程学院 西安 710054; 2.杭州电子科技大学电子信息学院 杭州 310018)

摘 要: 深入研究了融合航海优先级(NP)和优先级经验回放(PER)策略的深度 Q 网络(DQN)算法在海上环境智能路径规划问题上的应用。不同于传统路径规划算法,本优化算法能够自主探索并学习海上环境的规律,无需依赖人工构建的海洋环境全局信息。本研究开发了基于 Gym 框架的海上仿真环境,用以模拟和验证改进的 DQN 模型。该模型融合了航海优先级和优先级经验回放机制,通过调整学习过程中经验样本的利用频率,提升了算法对重要决策的学习效率。此外,引入新的奖赏函数,进一步增强了模型对路径规划问题的适应能力和稳定性。仿真实验结果证明,该模型在避免障碍物及寻找最佳路径方面相较于基准方法有显著提升,展现了一定的泛化性和优秀的稳定性。

关键词: 改进深度 Q 网络;海上模拟仿真环境;航海优先级;奖赏函数
中图分类号: TP242.6 **文献标识码**: A **国家标准学科分类代码**: 510.8050

Intelligent maritime path planning based on deep Q-Networks

Li Pengcheng[1] Zhou Yuanguo[1] Yang Guoqing[2]
(1. College of Communication and Information Engineering, Xi′an University of Science and Technology, Xi′an 710054, China;
2. College of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: This study delves into the application of a deep Q-Network (DQN) algorithm, which integrates strategies of Navigational Priority (NP) and Prioritized Experience Replay (PER), for intelligent path planning in maritime environments. Unlike conventional path planning algorithms, our optimized model autonomously explores and learns the patterns of the maritime environment without relying on manually constructed global maritime information. We have developed a maritime simulation environment based on the Gym framework to simulate and validate our improved DQN model. This model incorporates the mechanisms of Navigational Priority and Prioritized Experience Replay, enhancing the algorithm′s learning efficiency for critical decisions by adjusting the frequency of experience sample utilization during the learning process. Additionally, the introduction of a novel reward function has further strengthened the model′s adaptability and stability in addressing path planning issues. Simulation experiments demonstrate that our model significantly outperforms baseline methods in avoiding obstacles and finding optimal routes, showcasing notable generalizability and exceptional stability.

Keywords: improved deep Q-Network; maritime simulation environment; navigational priority; reward function

0 引 言

在全球经济高度一体化的大背景下,海上运输已然成为全球物流供应链中不可或缺的重要组成部分。然而,海洋环境的内在不稳定性——包括不稳定的气候状况、复杂的海洋流动、不明确的海况和繁忙的航海通道——给航行路径规划提出了前所未有的挑战[1]。因此,精确、高效且安全的海上路径优化不仅是确保航行安全的必要条件,而且也是提高整体运输效率的关键变量。传统的路径规划方法多依赖于静态全局地图,这种依赖导致感知和决策之间存在明显的隔离,限制了其在海上环境中的应用性。

针对海上环境信息部分或完全已知的场景下,国内外研究界已进行了大量关于航船路径规划方面的研究。例如,文献[2]详细描述了一种基于改进的 A* 算法的路径规划方法,并将其作为战争游戏模拟的参考框架。该研究建立了一种映射机制,通过这一机制实现了 A* 算法在战争游戏模拟环境中的初步应用,并构建了一种能够同时满足多目标需求并保证生成最优路径的评估函数。文献[3]提出了一种海上应

收稿日期:2023-11-27
* 基金项目:国家自然科学基金(61801009)、陕西省自然科学基金面上项目(2024JC-YBMS-556)资助

急物流路径规划算法，该算法采用矩阵式数据库整合航行条件属性。基于数据库，算法定量评估各航线的可通行性，并实施最短配送路径的静态规划。同时，考虑到航行环境的动态变化，研究还设计了一套动态规划流程，以实现路径的实时优化。文献[4]提出了一种融合 A* 算法和蚁群算法的最短巡逻路径优化方法。该方法在传统 A* 的八方向搜索基础上增加了多角度搜索，用以构建巡逻点间的最短路径网络。结合这一网络，研究者设定了多点巡逻路径规划的目标函数，并运用蚁群算法进行全局最优求解。这些算法在解决环境先验信息不足问题方面取得了一定的成效，能显著提升路径规划问题求解的收敛速率，并能生成相对优良的导航路径。然而，由于算法自身与环境的交互性不足，它们在复杂环境中的应用受到限制。具体来说，这些算法无法充分获取和利用环境信息，导致在面对更为复杂的环境时，必须投入大量时间进行非目标导向的搜索，从而降低了求解速度，并增加了陷入局部最优解的风险。

在应对海洋环境的复杂性方面，机器学习（machine learning，ML）技术为研究者们开辟了新的研究途径[5]。在这一领域内，监督学习方法能够通过利用历史数据训练神经网络，为航船构建高效的预测模型，从而确保其能够沿着最优路径航行。强化学习，作为一种尖端的机器学习技术，允许智能体在无需预设特定模型或控制规则的情况下，自主地学习控制策略[6]。2013 年，DeepMind 推出的深度 Q 网络（deep Q-Network，DQN）算法，将神经网络与强化学习相结合，标志着深度强化学习（deep reinforcement learning，DRL）领域的一个重大突破[7]。该方法在处理复杂任务时的卓越表现，源于其对深度学习和强化学习优势的有效融合。当前，对 DQN 的改进主要集中在解决原始 DQN 算法在实际应用中遇到的诸多挑战。其中，双重 DQN（double DQN）通过将动作选择网络和动作评估网络分离，旨在减少估计误差，从而有效解决了 DQN 中的过估计问题。优先经验回放（prioritized experience replay，PER）策略通过为经验回放池中的每个样本分配优先级，确保重要的学习经验能够被更频繁地回放，从而提高了样本利用的效率。此外，杜宾（Dueling）网络架构在 DQN 中引入了两个独立的网络[8]，一个负责估计状态值函数，另一个负责估计每个动作的优势函数，这种结构的改进促进了状态-动作值函数的更为精确的估计[9]。

传统深度 Q 网络（DQN）算法主要适用于相对静态的环境，其中环境状态转移主要依赖于智能体的行为。在海洋环境中，外部因素如海流和风速等呈现出动态变化的特性，而这些变化并非完全由智能体的行为所驱动。这一环境的不确定性和动态性对传统 DQN 算法的预测和适应能力提出了挑战。在海洋环境中，算法需要有效平衡探索（exploration）与利用（exploitation）之间的关系。传统 DQN 算法可能受限于其过度依赖已知策略，而不足以探索新的或更优的路径选择。此外，传统 DQN 算法在应对海洋环境中不断变化的状态时，可能未能充分利用历史经验样本，导致学习效率低下，需要大量迭代才能收敛至有效策略。针对以上存在的问题，本文首先构建了一个基于 Gym 框架的海上环境仿真平台，用于模拟船舶在海洋中可能遇到的各种动作和状态，并在此仿真环境中开展了一系列实验以验证所提出方法的有效性。研究中引入的 NP 和 PER 机制允许算法在学习过程中对经验样本的重要性进行动态调整，进而提升了路径规划的效率和准确性。新设计的奖赏函数进一步促进了模型适应各种海上条件的能力，强化了规划路径的稳定性和鲁棒性。实验结果证实，该改进模型在避免障碍物和优化路径选择方面明显超越了传统 DQN 算法，显示出更高的学习效率和更强的适应性。

1　相关工作

1.1　Q-learning 算法

强化学习（reinforcement learning，RL）是一类机器学习技术，它使得智能体能够在环境中通过探索与尝试的方法来学习最优策略，旨在最大化累积获得的奖励[10]。区别于传统的监督学习，强化学习不依靠预先标注的数据集，而是基于从环境交互中获得的奖励或惩罚进行学习[11]。如图 1 所示，在强化学习框架中，核心概念包括智能体（Agent）、环境（Environment）、行为（Action）、状态（State）以及奖励（Reward）。智能体观察所处的环境，并依据其策略做出行动，进而获得相应的奖励或惩罚。策略指导智能体在特定状态下的行动选择，而价值函数用于预测从当前状态开始的预期累积奖励[12]。强化学习面临的主要挑战之一是在探索新策略与利用已知最优策略间寻求平衡。此外，折扣因子则反映了智能体对即时奖励与未来奖励的偏好程度。目前流行的强化学习算法包括 Q 学习、深度 Q 网络（DQN）和策略梯度方法，这些算法已在多个领域如游戏、机器人技术和自动驾驶等方面取得成功，显示出强化学习处理复杂问题的巨大潜力[13]。

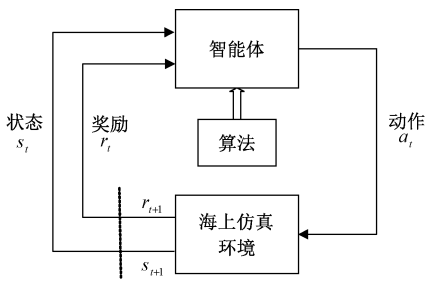

图 1　智能体与海上仿真环境交互示意图

Q-learning 算法是一种无模型的强化学习的算法，使用 Q 表格来存储物体的不同状态 s 和动作集合 A 之间的关系[14]。Q 表格是一个二维表格，其中行表示状态，列表示动作，每个单元格存储了对应状态和动作组合的 Q 值，表示智能体采取该动作后预期获得的累积奖赏。Q 表格通过不断迭代更新来逼近最优的 Q 值。更新 Q 表格的方式

使用了贝尔曼方程,其更新规则如下:

$$Q(s,a)=(1-\alpha)\times Q(s,a)+\alpha\times(r+\gamma\times\max(Q(s',a'))) \tag{1}$$

式中:$Q(s,a)$ 是状态 s 下采取动作 a 的 Q 值;α 是学习率(learning rate),控制着新旧 Q 值 γ 之间的权重;γ 是折扣因子(discount factor),表示未来奖赏的重要性;$\max(Q(s',a'))$ 是在下一个状态 s' 下所有可能动作的 Q 值中的最大值。

通过上述更新规则,Q 表格在每次与环境进行交互时根据当前的状态、动作和奖赏信息进行更新。通过不断迭代,Q 表格中的 Q 值会逐渐逼近最优策略的预期累积奖赏。

1.2 奖赏函数

奖赏函数是在强化学习中用于评估智能体行为优劣的函数[15]。奖赏函数根据智能体的动作和环境的状态,给出一个标量值作为奖励信号,用于指导智能体在环境中采取哪些动作以最大化未来累积的奖赏值。奖赏函数通常用一个数学公式来表示。一般而言,奖赏函数的公式如下:

$$R(s,a)=f(s,a) \tag{2}$$

式中:$R(s,a)$ 是在状态 s 下采取动作 a 所获得的即时奖赏;$f(s,a)$ 是根据具体任务和环境特点设计的奖赏函数。

奖赏函数的设计可以根据任务需求进行灵活调整,常见的奖赏函数设计准则如下:

1)目标导向性:奖赏函数应该与任务目标相关,鼓励智能体朝着实现任务目标的方向前进。

2)正负反馈:正向奖励可以加强智能体正确的行为,并促使其重复类似的行动;负向奖励可以惩罚智能体错误或不合理的行为,引导其避免这样的行动。

3)折扣因子:奖赏函数应该考虑智能体的长期收益,通过设置适当的折扣因子,奖赏函数可以引导智能体在决策过程中权衡当前和未来的奖励。

4)稳定性和连贯性:奖赏函数应该在不同的环境状态下保持稳定和连贯,相似的环境状态应有相似的奖励值,不同的环境状态则应有不同的奖励值。智能体在决策过程中权衡当前和未来的奖励。

1.3 经验回放机制

经验回放是深度 Q 网络算法中的重要的环节,用于解决强化学习中数据相关性和非稳定性问题。在传统的在线学习中,智能体会按顺序经历状态转换并从这些连续的经验中学习。这种连续的学习方式可能导致数据之间的强相关性,从而影响学习的稳定性。

为解决这个问题,在深度 Q 网络中引入经验回放机制。具体来说,与环境互动时,智能体不是立即从其经验中学习,而是将每一个状态转换 (s,a,r,s') 存储在一个数据缓冲区中。然后,在学习时,智能体会从这个缓冲区中随机抽取一批经验,从而打破数据之间的相关性。经验回放的学习更新公式如下:

$$L(\theta)=\mathrm{E}_{(s,a,r,s')\sim U(D)}[(r+\gamma\max_{a'}Q(s',a';\theta^-)-Q(s,a;\theta))^2] \tag{3}$$

式中:D 是经验回放缓冲区,存储过去的转换 (s,a,r,s');U(D)表示从 D 中随机抽取的一个转换。

2 海上仿真环境的设计与构建

2.1 建立环境

在本研究中,选用栅格地图来对海上环境进行建模和描述。该地图构成一个二维离散网格,每个格子代表环境中的一个特定区域,其状态可能为:空闲、障碍物或目标区域。此栅格地图可以采用直角坐标系或序号法进行表示。值得注意的是,地图的尺寸对于决策速度和环境信息存储量均有显著影响。如图 2 所示,在此采用了直角坐标系进行表示。在本文的研究框架下,利用二维矩阵来呈现栅格地图,其中每个矩阵元素的值标明了其对应位置的状态:0 表示空闲,−1 表示障碍,1 表示起始点,而 3 则表示目标位置。总的来说,通过应用栅格地图,得以将环境进行离散化,并精确标注出空闲区域、障碍及目标等关键信息,进而为智能体在此离散环境中进行高效的路径搜索和规划提供了基础。

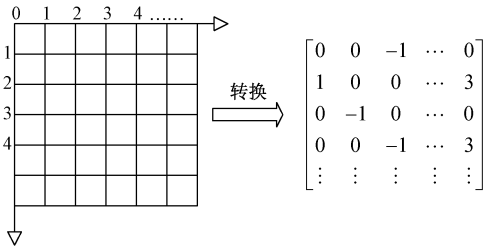

图 2 采用二维矩阵表示的栅格地图示意图

2.2 状态定义和动作定义

为了更好的表征船舶在海上的信息,状态定义为当前船舶的位置,当前位置的海流和风力等信息。具体来说,将海域通过网格(grid)来划分,每个网格的属性包括位置坐标 (x,y)、海流的速度 v_{sea} 和风力 v_{wind}。状态用如下向量进行表示:

$$\boldsymbol{S}=[x,y,v_{sea},v_{wind}] \tag{4}$$

在构建海上环境模型时,需要定义环境的动作空间。动作空间定义为船舶从当前位置移动到下一个位置的方向。在智能路径规划中,动作空间定义了智能体可以采取的所有可能动作。

2.3 奖赏函数设计

奖赏函数是改进深度 Q 网络算法的核心部分,它提供了航船在海上环境下动作好坏的反馈,能够在模型训练过程中激励算法更好的适应环境变化,从而有效避开突发的强海流。在海洋环境的路径规划中,需要设计一个奖赏函数来引导航船进行有效的路径选择。按照如下奖励函数进行设计:1)基本移动奖励,当航船进行正常移动时,给予一个小的正奖励,鼓励其继续探索未知环境 R_{move}。2)到达目标奖励,当航船成功到达目标位置时,给予一个大的正奖励 R_{goal}。3)碰撞惩罚,如果航船与障碍物发生碰撞,给予一个

负奖励 $R_{\text{collision}}$。 4)环境因素加成:为了引导航船有效躲避强海流,定义了环境难度系数 $E_{difficulty}$。 5)时间惩罚:为了鼓励航船尽快达到目标区域,每走一步给予一个小的负奖励 R_{time}。 对应的奖励函数公式如下:

$$R(s,a,s')=\begin{cases}R_{goal}, & \text{是终点}\\ R_{collision}, & \text{碰到障碍物}\\ 1+\alpha\times I_{current}, & \text{强海流区域}\\ R_{move}+R_{time}, & \text{正常移动}\end{cases} \tag{5}$$

其中,s 是当前状态,a 是采取的动作,s' 是动作后的新状态,α 是一个比例系数,$I_{current}$ 代表当前海上的海流强度。

3　改进深度 Q 网络智能路径规划算法

在本研究中,构建了一个专门的海上仿真环境,以便更好地模拟和研究航船的智能路径规划问题。该环境是基于二维栅格地图构建的,这种地图形式提供了一个清晰且易于处理的框架,以表示复杂的海洋地理特征。每个栅格单元代表海洋环境中的一个区域,包括可能影响航行的各种要素,如海流、风速、障碍物等。同时,状态的表征是通过一维元组实现的,这样做的目的是为了简化状态空间,使得状态信息更加直观且易于神经网络处理。

在深度 Q 网络(DQN)算法的设计中,如果使用卷积神经网络(CNN)来提取环境特征,虽然能够有效处理图像型数据,但也可能引起模型的过拟合问题。过拟合会导致模型在训练集上表现良好,但在新的、未见过的数据上表现不佳。这是因为 CNN 可能会过度学习训练数据中的特定细节,而忽略了更广泛的、一般性特征。

3.1　神经网络策略

在深度 Q 网络(DQN)算法中,神经网络策略发挥着至关重要的作用。它不仅指导智能体的行为决策,更是为强化学习提供了强大的函数逼近能力。这一点在复杂多变的海上环境中尤为重要,它使得航船能够有效地学习和操作,应对各种海洋状况。如图 3 所示,神经网络的设计通常采用三层结构:输入层、隐藏层和输出层。在这里,输入层的神经元数量需要与环境状态的维度相匹配。考虑到输入类型为一维数组,本研究选择了全连接层作为隐藏层的基础结构。全连接层由三层隐含层组成,每层包含若干神经元。隐含层的神经元数量是决定模型拟合效果的关键因素,数量越多,模型对环境状态的表示能力就越强,能够捕捉更加微妙的环境变化。然而,神经元数量的增加也会带来过拟合的风险,即模型可能过度适应训练数据,而在未见过的数据上表现不佳。为了缓解这一问题,本研究在全连接层中加入了批归一化(batch normalization, BN)层。BN 层通过规范化每层输入的分布,有助于加速训练过程,同时降低过拟合的风险。在网络的输出层,神经元的数量则与环境动作空间的动作数量相匹配。这些神经元的输出表示了在当前状态下,采取不同动作的预期 Q 值。Q 值是强化学习中的核心概念,代表了采取某一行动后获得的预期回报。通过对输出层的 Q 值进行比较,DQN 能够选择出当前状态下最优的行动策略。

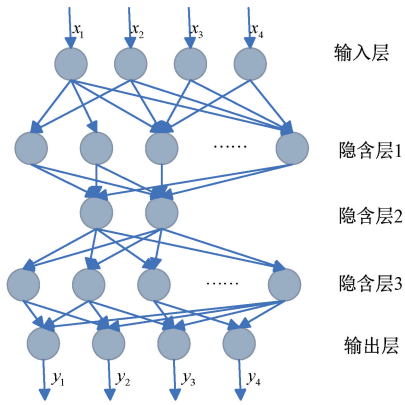

图 3　神经网络结构

3.2　优先级经验回放

为了能更高效的学习和获得更强的鲁棒性,在算法中引入优先级经验回放(prioritized experience replay,PER)。在传统的经验回放中,每个经验被均匀地从回放缓冲区中采样。而在 PRE 中,经验被按照优先级进行采样,并且这些优先级是基于经验的(temporal difference error, TD)误差。具体来说,每个经验的优先级 p 被定义为:

$$p=|\delta|+\epsilon \tag{6}$$

式中:δ 是 TD 误差,表示为:

$$\delta=r+\gamma\max_{a'}Q(s',a';\theta^{-})-Q(s,a;\theta) \tag{7}$$

式中:r 是奖励,γ 是折扣因子,$Q(s,a)$ 是当前的 Q 值估计,而 $\max_{a'}Q(s',a')$ 是在下一个状态 s' 下所有可能动作的最大 Q 值。ε 是一个很小的正数,确保每个经验都有采样的机会。然后,采样的概率 $P(i)$ 与优先级 p 成正比:

$$P(i)=\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}} \tag{8}$$

为了减少优先级引入的偏差,使用重要性采样权重 w 来调整学习更新:

$$w_i=\left(\frac{1}{N}\cdot\frac{1}{P(i)}\right)^{\beta} \tag{9}$$

式中:N 是回放缓冲区的大小,而 β 是一个介于 0～1 的参数,随着时间的推移逐渐增加。最后,Q 值的更新为上面权重的乘积:

$$\Delta Q=w_i\cdot\delta \tag{10}$$

这样,PER 会优先采样那些具有高 TD 误差的经验,从而加速模型的学习并提高性。

3.3　航海优先级

在海上路径规划研究中,航海优先级是一个至关重要的量化指标。为了评估航船在多障碍物和多约束条件下路径规划的性能,在算法中引入航海优先级来综合考虑各种影响因素,如碰撞风险、航道拥挤程度和海上情况等因素,航海优先级 NP 公式表示如下:

$$\boldsymbol{NP} = \varpi_1 \times \boldsymbol{D} + w_2 \times \boldsymbol{V} - \varpi_3 \times \boldsymbol{C} \tag{11}$$

式中:D 是航船发生碰撞风险的概率,V 是航道的拥挤程度,C 是海上环境约束的累计成本,如风向和洋流因素,而 ϖ_1,ϖ_2,ϖ_3 是权重系数,用于平衡各个因素的相对重要性。

为了将优先级经验回放和航海优先级结合,在本研究中提出一个综合优先级 P_{total},它不仅考虑了海面环境和约束条件,还考虑了学习算法的内在需求。这样的设计旨在提高智能路径规划算法在海上环境中的性能,通过更有效地从高优先级地经验中学习,加快求解最优路径的速度。

$$\boldsymbol{P}_{total} = \alpha \times \boldsymbol{NP} + (1-\alpha) \times \boldsymbol{P}_{PER} \tag{12}$$

其中,α 是一个权重系数,用于平衡航海优先级和优先级经验回放在算法的重要性。

3.4 动态适应ϵ-退火策略

在强化学习领域,尤其是针对复杂和动态变化的海上环境智能路径规划,本文提出了一种新的动态适应ϵ-退火策略。该策略综合了ϵ-贪心策略和退火策略的优势,旨在通过动态调整探索率来有效平衡探索和利用的需求。动态适应ϵ-退火策略考虑了海上环境智能路径规划的复杂性和不确定性,通过参数化的方法(如初始ϵ值、最终ϵ值和退火速率)来调整策略,确保模型在不同训练阶段都能获得适当的探索与利用平衡。此外,策略的设计还考虑到了环境反馈,允许根据模型的实际表现来适时调整退火速率,从而更好地适应动态变化的海上环境。

$$\epsilon(t) = \epsilon_{start} - \left(\frac{\epsilon_{start} - \epsilon_{end}}{N}\right) \times t \tag{13}$$

其中,t 是当前的训练步骤,N 是总训练步骤数。当 $t = N$ 时,$\varepsilon(t)$ 将等于 ε_{ends}。

3.5 模型训练

在模型训练的每一次迭代中,改进后的 DQN 算法的网络更新是通过根据经验优先级从回放缓冲区中抽取一批数据来进行的。将这些数据传入到策略网络和目标网络中,然后求两个网络参数 θ 的梯度,以便使用优化算法(adaptive moment estimation, Adam)进行参数更新,损失函数如式(14)所示。

$$L(\theta) = \mathrm{E}_{(s,a,r,s')\sim U(D)}[(r + \gamma \max_{a'} Q(s',a';\theta^-) - Q(s,a;\theta))^2] \tag{14}$$

为了求梯度,首先考虑损失函数中的平方项的梯度。通过式(5)将损失函数简化为:

$$L(\theta) = \mathbb{E}_{(s,a,r,s')\sim U(D)}[w_i \delta^2] \tag{15}$$

然后对得到的损失函数求梯度,对于单个样本,损失函数的梯度为:

$$\nabla_{\theta} \boldsymbol{L}(\boldsymbol{\theta}) = 2\boldsymbol{w}_i \delta \nabla_{\theta} \boldsymbol{Q}(\boldsymbol{s},\boldsymbol{a};\boldsymbol{\theta}) \tag{16}$$

最后,使用适应性矩估计算法(Adam)和学习率 η 更新网络参数:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\theta} \boldsymbol{L}(\boldsymbol{\theta}) \tag{17}$$

图 4 展示了改进 DQN 算法的更新流程,其中将当前环境状态输入到策略网络中,然后通过动作选择机制与环境进行交互,得到该动作的奖励和下一步的状态。这四元组被存储在回放缓冲区中,当经验数量达到 batch_size 时,从其中根据优先级抽取一小批量样本进行训练。每个 N 时间步后,策略网络会将网络参数拷贝给目标网络,并根据损失函数对策略网络的参数进行更新。

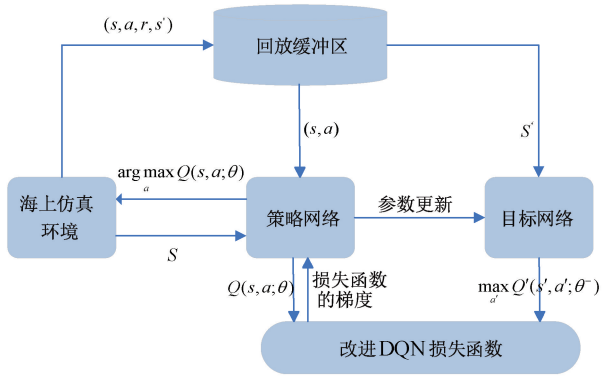

图 4 改进 DQN 算法更新流程

伪代码 1 基于改进 DQN 的海上环境智能路径规划算法

1)初始化:

初始化策略 Q 网络 $Q(s,a;\theta)$;

初始化目标 Q 网络 $Q'(s,a;\theta^-)$ 与 $Q(s,a;\theta)$ 相同;

初始化经验回放缓冲区 D 并为每个经验分配优先级;

初始化航海优先级 NP 权重 ϖ_1,ϖ_2,ϖ_3

2)For episode=1 to k do

初始化状态 s

For step=1 to n do

根据当前 Q 网络选择动作 a(使用 ε-贪心策略)

执行动作 a,观察奖励 r 和新状态 s';

If rect=target or n >max_step:

break;

计算 TD 误差:

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

根据公式计算航海优先级 NP;

根据公式计算综合优先级

用 δ 更新优先级 P

将经验 (s,a,r,s') 和对应的优先级存储到 D 中;

从 D 中根据优先级抽取一个小批量的经验;

计算权重 ω,用于调整 loss 的影响;使用权重 ω 和 TD 误差 δ 更新 Q

其中,rect 表示航船当前的位置;target 表示目标区域的位置;max_step 表示回合最大执行的步数。根据上面伪代码的改进 DQN 算法模型流程图如图 5 所示。

4 实验及验证分析

4.1 实验设置和参数

在本项研究中,为了评估基于深度 Q 网络(DQN)的海上环境智能路径规划算法的有效性,构建了一个基于 Gym

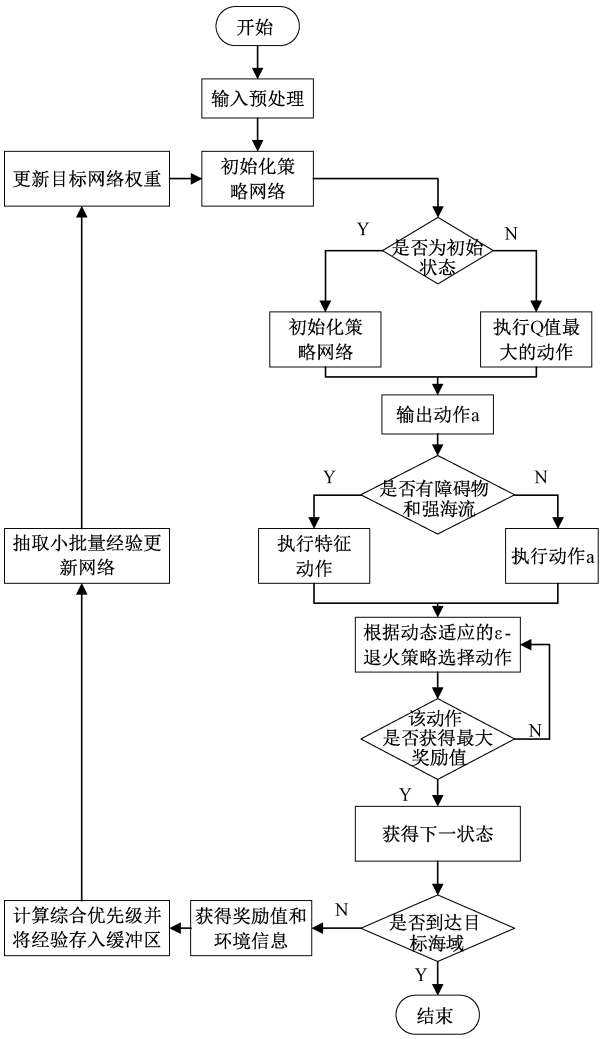

图 5　改进 DQN 算法模型流程图

的仿真环境，并将所提出的改进 DQN 算法与传统的 DQN 及随机策略进行了细致的对比分析。以下为具体的实验设计及参数设置：

1)仿真环境配置：研究所用的仿真环境由一个二维栅格地图构成，如图 6 所展示。在这个场景中，黑色方块标定了航船的起始点，红色方块标记了目标区域，而蓝色方块则模拟了海面上的障碍物。此外，每个栅格单元还模拟了海流速度和风力等属性，进一步细化了海上环境的模拟，如图 7 所示的示意图。

2)智能体描述：仿真环境中的智能体被定义为一艘在海面上航行的船舶，其初始位置被设在起点。智能体的主要任务是规划出一条既安全又高效的航线，从而成功绕过障碍物并抵达目标区域。

3)算法配置：在算法的配置上，该研究主要基于改进后的深度 Q 网络，同时选取了 PRE-DQN 算法和 Dueling-DQN 作为基准，以进行性能对比。

4)训练策略：整个训练过程包括了 1 000 轮迭代。在

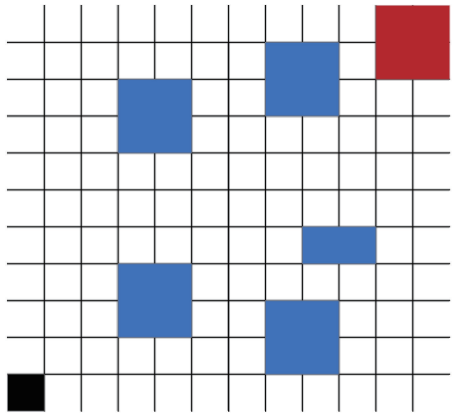

图 6　海上仿真环境

图 7　海上环境模型示意图

每一轮迭代中，智能体均从设定的起点出发，并努力避开障碍物以抵达目标区域。为了增强训练效率和算法的鲁棒性，引入了优先级经验回放和航海优先级机制。

5)评估准则：评估指标包括平均奖励值、平均到达目标区域所需的步数，以及均方误差(mean squared error，MSE)，以量化算法的性能表现。

本实验的具体参数设置如表 1 所示。

表 1　参数设置

参数	值
回合	1 000
一个回合中时间步数	600
经验池的尺寸	500
经验回放批次大小	16
折扣因子 γ	0.98
学习率 α	0.01
最大探索概率 ε_{max}	0.3
最小探索概率 ε_{min}	0.01
衰减因子 ς	3×10^{-6}
目标网络更新频率 N	50
神经元数量	64

4.2　实验结果

根据所述实验设置与参数框架,仿真实验采取了 1 000 个训练周期,每个周期称为一个 Episode。实验过程中,以每 50 个训练步骤为周期,将策略网络中的权重传递至目标网络,以此来实现后者的更新。在训练启动前,智能路径规划算法对策略网络和目标网络进行了初始化,并且创建了回放经验缓冲区以存储经验元组。初始化完成后,智能体首先着重于对环境的探索活动,并不立即开始网络训练。这一策略旨在积累初始的环境数据集,为后续的训练学习奠定基础。在初期阶段,由于智能体尚未完全探索并理解环境,其所采取的航线往往并非最直接或最有效,有时可能会绕行较远路线或不慎撞上障碍物。此外,智能体在此阶段找到的路径可能并非最优,如图 8 所揭示的路径选择情形。随着经验的积累和学习的深入,预期智能体会逐步优化其路径选择策略,从而提高规划的效率和安全性。

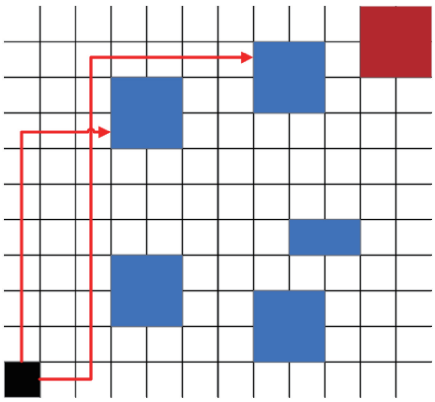

图 8　智能体在训练初期进行路径规划结果

随着充分的经验积累和持续的学习过程,智能体逐渐能够有效的进行避障,并学会寻找一条安全且高效的路径。在本研究中,通过比较不同 episode 所获得的总回报,可以发现平均回报值最高的回合通常对应于最短的航线。一旦智能体发现了这一最短航线,它便开始沿此路线进行智能规划,并持续这一行为直到学习曲线稳定。因此,如图 9 所示,智能体在避免障碍物的同时,能够快速地辨认并规划出最短的航行路径。此现象不仅展现了智能体的学习能力,也证明了其在海上环境下进行高效路径规划的能力。

图 10 中的数据展示了传统改进的深度 Q 网络(DQN)和基于综合优先级改进的 DQN 算法在训练过程中的均方误差(MSE)变化趋势。从图中可以清晰地看出,改进 DQN(用绿色曲线表示)在整个训练周期中的 MSE 均低于传统改进 DQN(用蓝色曲线和灰色曲线表示)。这一观察表明,改进的 DQN 在预测奖励与实际奖励之间的误差更小,从而使学习过程更加准确和有效。特别是在训练的初期阶段,改进 DQN 的 MSE 降低速度明显快于传统改进的 DQN,这表明改进 DQN 能够更迅速地逼近最优策略。随着训练的继续进行,改进 DQN 的 MSE 趋于稳定并保持在

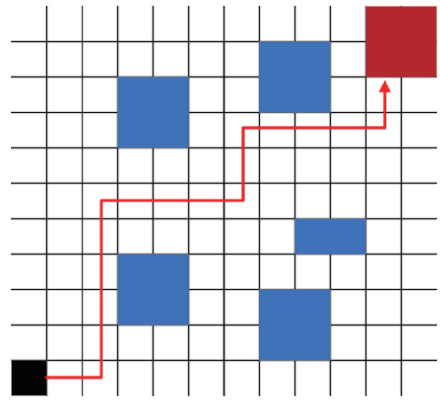

图 9　智能体完成路径规划的结果

较低水平,而传统改进的 DQN 的 MSE 虽然也呈现下降趋势,但降低速度相对较慢,且波动性更加显著。这种差异可能是由于改进 DQN 中引入的优先级经验回放策略,该策略使得模型能够更加关注那些预测偏差较大的经验,从而实现更快速的调整和优化。综合来看,图 10 中的实验数据强有力地支持了结合优先级经验回放的改进 DQN 在减少预测误差和提高学习稳定性方面,相较于传统改进的 DQN 具有明显优势的结论。

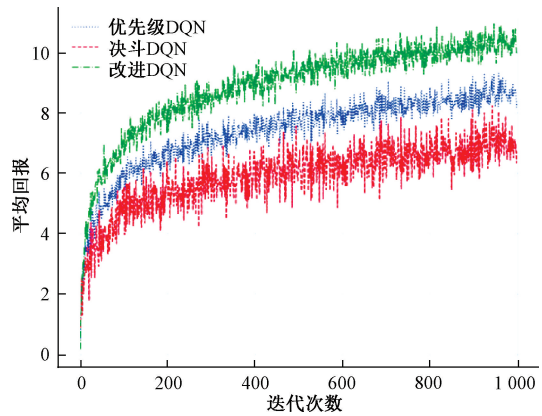

图 10　智能体在 3 种算法下平均回报的表现

图 11 中展示了传统改进的 DQN 与引入融合优先级经验回放的改进 DQN 在训练过程中的 MSE 变化情况。明显地,改进的 DQN(绿色曲线)在整个训练过程中的 MSE 都低于传统改进的 DQN(蓝色曲线和灰色曲线)。这表示改进的 DQN 在预测和实际奖励之间的误差更小,从而使得学习过程更为准确。尤其是在早期训练阶段,改进的 DQN 的 MSE 下降速度更快,这意味着该算法能够更迅速地逼近最优策略。另外,随着训练的进行,改进的 DQN 的 MSE 趋于稳定并保持在一个较低的水平,而传统改进的 DQN 的 MSE 虽然也在下降,但下降的速度相对较慢,且波动性更大。这可能是由于优先级经验回放策略,使得模型更加关注那些预测偏差较大的经验,从而更快地调整和优化。总之,图中数据强烈表明了引入优先级经验回放的改

进 DQN 在减少预测误差和提高学习稳定性上都优于传统改进的 DQN。

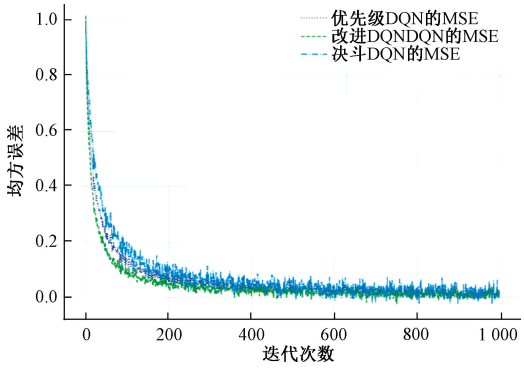

图 11　智能体在 3 种算法下均方误差表现

图 12 对比了传统改进的 DQN 与改进 DQN 在不同回合数下到达目标所需的平均步数。首先，从整体趋势上看，两种算法在随着回合数增加时，到达目标所需的平均步数都呈现下降趋势。这表明无论是传统改进的 DQN 还是改进的 DQN，它们都在学习和优化其策略，使得智能体更高效地完成任务。然而，更为明显的是，改进的 DQN 在几乎所有回合中的平均步数都显著低于传统改进的 DQN。这意味着，与传统改进的 DQN 相比，改进的 DQN 能够更快地找到一条到达目标的路径。这种效率的提高可能是由于优先级经验回放的引入，使得算法更加关注那些预测偏差较大的经验，从而更快地调整其策略。此外，随着回合数的增加，两种算法之间的差距逐渐扩大。在早期回合，两者的差异虽然存在，但并不显著。然而，随着更多的训练，改进的 DQN 的效益更加明显，这进一步突出了其相对于传统改进 DQN 的优势。

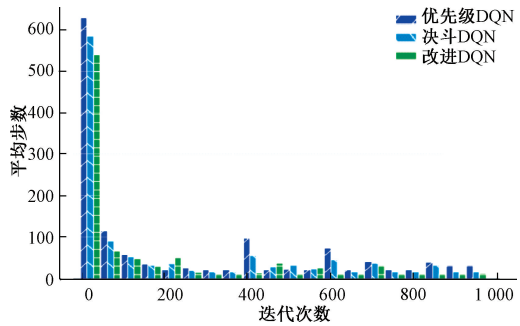

图 12　智能体在 3 种算法下到达目标所需的平均步数

通过表 2 的数据比较分析可知，在探索具有较高复杂性和众多障碍物的海域环境时，PRE-DQN 与 Dueling DQN 算法表现出一定的局限性。在这些情况下，智能体有时会陷入局部最优解，未能学习到全局最优的决策策略。这种现象可能导致航船无法有效规划出一条从起始点到目标点的安全路径。相比之下，本文提出的改进 DQN 算法显著提高了路径规划的效果，能够有效地找到一条安全且相对优化的路径。该算法的改进显著增强了智能体的决策能力，从而为海域航行提供了更为可靠的导航支持。

表 2　强海流环境下 3 种环境对比表

算法	躲避成功率/%
PRE-DQN	35
Dueling DQN	30
Improved DQN	75

5　结　　论

本研究提出了基于改进深度 Q 网络(DQN)的海上环境的智能路径规划算法，并在 Gym 中构建了一个海上环境仿真模型，模拟了真实海洋环境中的海流和风速等因素。使用改进的 DQN 作为智能路径规划的算法，并进行了大量的实验验证。结果表明，相比于传统的 DQN 算法，改进的 DQN 能够在海上环境模型中有效地学习到路径规划策略。在训练过程中，航船的行驶路径逐渐变得稳定，碰到障碍物的次数逐渐减少，到达目标的时间也逐渐缩短。这表明改进深度 Q 网络的智能路径规划算法在海上环境下，具有更高学习效率、更高的智能路径规划能力和更好的鲁棒性。

参考文献

[1]　高天航，吕靖，赖成寿. 考虑船舶偏好的海上风险规避路径规划研究[J]. 运筹与管理，2018，27(11)：43-49.
[2]　张韬，项祺，郑婉文，等. 基于改进 A* 算法的路径规划在海战兵棋推演中的应用[J]. 兵工学报，2022，43(4)：960-968.
[3]　李杰，和娅. 移动网络海上应急物流路径规划算法[J]. 舰船科学技术，2020，42(14)：181-183.
[4]　张丹红，陈文文，张华军，等. A* 算法与蚁群算法相结合的无人艇巡逻路径规划[J]. 华中科技大学学报(自然科学版)，2020，48(6)：13-18.
[5]　王奇，黎海涛. 基于机器学习与惯性导航的室内定位技术研究[J]. 电子测量技术，2016，39(8)：138-143.
[6]　邓修朋，崔建明，李敏，等. 深度强化学习在机器人路径规划中的应用[J]. 电子测量技术，2023，46(6)：1-8.
[7]　TANG X，YANG Y，LIU T，et al. Path planning and tracking control for parking via soft actor-critic under non-ideal scenarios [J]. IEEE/CAA Journal of Automatica Sinica，2024，11(1)：181-195.
[8]　王军，杨云霄，李莉. 基于改进深度强化学习的移动机器人路径规划[J]. 电子测量技术，2021，44(22)：19-24.
[9]　武曲，张义，郭坤，等. 基于 DPES Dueling DQN 的路径规划方法研究[J]. 计算机应用与软件，2023，40(6)：147-153.
[10]　王志伟，邹艳丽，刘唐慧美，等. 基于改进 Q-learning 算法和 DWA 的路径规划[J]. 传感器与微系统，2023，42(9)：148-152.
[11]　邓修朋，崔建明，李敏，等. 深度强化学习在机器人路径规划中的应用[J]. 电子测量技术，2023，46(6)：1-8.
[12]　李腾，曹世杰，尹思薇，等. 应用 Q 学习决策的最优攻击路径生成方法[J]. 西安电子科技大学学报，2021，48(1)：160-167.
[13]　卫玉梁，靳伍银. 基于神经网络 Q-learning 算法的智能车路径规划[J]. 火力与指挥控制，2019，44(2)：46-49.
[14]　刘俊利. 基于 TensorFlow 的 Q-Learning 算法研究与实现[J]. 现代计算机，2019(29)：26-28，34.
[15]　周瑶瑶，李烨. 基于排序优先经验回放的竞争深度 Q 网络学习[J]. 计算机应用研究，2020，37(2)：486-488.

作者简介

李鹏程，硕士研究生，主要研究方向为强化学习、深度学习、机器人路径规划。
E-mail：lipc@stu. xust. edu. cn
周远国(通信作者)，副教授，博士，主要研究方向为计算电磁学、机器学习。
E-mail：zyg@xust. edu. cn