

基于模型的强化学习在无人机路径规划中的应用

杨思明^{1,2}, 单征¹, 曹江², 郭佳郁¹, 高原², 郭洋², 王平²,
王景², 王晓楠²

(1. 数学工程与先进计算国家重点实验室, 郑州 450001; 2. 军事科学院, 北京 100091)

摘要: 针对当前强化学习算法在无人机升空平台路径规划任务中样本效率低、算法鲁棒性较差的问题, 提出一种基于模型的内在奖励强化学习算法。采用并行架构将数据收集操作和策略更新操作完全解耦, 提升算法学习效率, 并运用内在奖励的方法提高智能体对环境的探索效率, 避免收敛到次优策略。在策略学习过程中, 智能体针对模拟环境的动态模型进行学习, 从而在有限步内更好地预测状态、奖励等信息。在此基础上, 通过结合有限的规划计算以及神经网络的预测, 提升价值函数的预测精准度, 以利用较少的经验数据完成智能体的训练。实验结果表明, 相比同样架构的无模型强化学习算法, 该算法达到相同训练水平所需的经验数据量减少近600幕数据, 样本效率和算法鲁棒性都有大幅提升, 相比传统的非强化学习启发类算法, 分数提升接近8 000分, 与MVE等主流的基于模型的强化学习算法相比, 平均分数可以提升接近2 000分, 且在样本效率和稳定性上都有明显提高。

关键词: 无人机; 升空平台; 路径规划; 强化学习; 深度学习

开放科学(资源服务)标志码(OSID):



中文引用格式: 杨思明, 单征, 曹江, 等. 基于模型的强化学习在无人机路径规划中的应用[J]. 计算机工程, 2022, 48(12): 255-260, 269.

英文引用格式: YANG S M, SHAN Z, CAO J, et al. Application of model-based reinforcement learning in path planning of unmanned aerial vehicle[J]. Computer Engineering, 2022, 48(12): 255-260, 269.

Application of Model-Based Reinforcement Learning in Path Planning of Unmanned Aerial Vehicle

YANG Siming^{1,2}, SHAN Zheng¹, CAO Jiang², GUO Jiayu¹, GAO Yuan², GUO Yang², WANG Ping²,
WANG Jing², WANG Xiaonan²

(1. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China;

2. Academy of Military Sciences, Beijing 100091, China)

[Abstract] This paper focuses on the problem of low sample efficiency and poor algorithm robustness of the current reinforcement learning algorithms used for path planning of Unmanned Aerial Vehicle (UAV) platforms. Furthermore, this paper proposes a model-based reinforcement learning algorithm with intrinsic rewards. The algorithm adopts a parallel architecture, completely decouples data collection operations and policy update operations, and improves the learning efficiency of the algorithm. Moreover, intrinsic reward improves the agent's exploration efficiency and prevents convergence to sub-optimal strategies. In the strategy learning process, the agent learns based on the dynamic model of the simulated environment, so that information such as the state and reward can be better predicted within a limited step. Finally, by combining finite planning calculation steps and neural network prediction, the prediction accuracy of the value function is improved. This reduces the amount of empirical data required to complete the training of the agent. The experiment results show that our algorithm, compared with the model-free reinforcement learning algorithm of the same architecture, requires approximately 600 fewer empirical data to achieve the same training level. The sample efficiency and algorithm robustness are also greatly improved. Compared with traditional heuristic algorithms, the score improves by nearly 8 000 points. Compared with mainstream model-based reinforcement learning algorithms such as MVE, the average score of the algorithm can improve by approximately 2 000 points and the agent had obvious advantages in sample efficiency and stability.

基金项目: 国家自然科学基金(61971092, 61701503)。

作者简介: 杨思明(1994—), 男, 硕士研究生, 主研方向为深度学习、强化学习; 单征, 教授; 曹江, 研究员; 郭佳郁, 硕士研究生; 高原, 副教授; 郭洋、王平, 助理研究员; 王景, 研究员; 王晓楠, 助理研究员。

收稿日期: 2021-11-08 **修回日期:** 2021-12-23 **E-mail:** 972856350@qq.com

【Key words】Unmanned Aerial Vehicle(UAV); aerial platform; path planning; reinforcement learning; deep learning
DOI: 10.19678/j.issn.1000-3428.0063156

0 概述

随着当前城市内移动通信终端数量的快速增长以及物联网、云计算、高清视频等新应用新技术的迅速发展,大型城市中数据月均流量消耗增长迅猛^[1]。无人机升空平台作为辅助地面基站,可为城市提供无线覆盖保障。当前无人机升空平台多采用低空无人机,如何根据环境信息和用户位置信息实时规划路径,以规避建筑物对于信号的遮挡以及调整合适的飞行方向、速度以避免发生多普勒频移造成的快衰落,是当前无人机升空平台在提供无线通信保障任务中亟待解决的问题。

解决上述问题的传统方法是通过对目标区域进行建模,然后使用最优控制算法进行路径规划。ROMERO等^[2]利用地面用户和无人机基站之间发送的控制信息,提出一种基于随机梯度下降法的分布式自适应无人机轨迹优化算法。ZENG等^[3]研究在已知地面用户位置的情况下使用无人机升空平台为地面用户提供数据传输服务的内容,进行圆形飞行轨迹设计,以在固定时间内最大化地面用户的上行速率。LYU等^[4]提出一种高效的螺旋式无人机布局算法,意在使用最少的无人机升空平台,保证每一个地面用户都能被有效覆盖,但是该算法需要无人机平台在固定高度悬停。ALZENAD等^[5]设计一个无人机升空平台在三维空间中的评估模型,以利用最小的发射功率实现对于目标区域的覆盖。KALANTARI等^[6]提出一种粒子群优化框架,使得可以利用最少数量的无人机完成对目标区域的无线覆盖。AL-HOURANI等^[7]根据地面静态用户的位置信息,将无人机升空平台的部署问题表示为一个二次约束混合整数非线性问题,用以得到最优的三维部署方案,最大化地面静态用户的下行速率。但上述算法主要存在以下问题:一是需要对环境进行复杂且精确的建模,而精确建模需要耗费大量时间以及计算资源,并且当前很多实际问题并不能准确地建模;二是当前算法更多考虑的是为地面静态用户提供通信覆盖的场景。目前对于地面多移动用户的无人机升空平台实时路径规划方法的研究还处于初期阶段。

基于深度强化学习(Deep Reinforcement Learning, DRL)的方法通过将路径规划任务建模为时序决策优化问题,利用神经网络的泛化性能以及强化学习的优化思想最大化累积收益,使智能体学习到最优策略。文献[8-9]使用DQN算法^[10]对无人机升空平台进行路径规划,以最大化数据传输速率。但该算法只能应用于离散动作空间任务,并且存在价值函数估值过高的问题,对智能体学习路径规划策略造

成了偏差。对此,WANG等^[11]使用Double DQN算法^[12]优化无人机平台飞行轨迹,用以在对地面所有用户进行覆盖的前提下最大化下行速率。Double DQN算法弥补了DQN价值函数估值过高的问题,但仍然不能应用在连续动作空间任务中。同时,由于智能体探索能力随着策略更新次数的增加而下降,智能体会出现收敛到局部最优策略的情况。文献[13-14]使用DDPG算法^[15]成功地将深度强化学习应用在连续动作空间的路径规划任务中,但是该算法超参数过多,在复杂问题中训练速度慢且不稳定。可见,当前DRL算法在处理路径规划这一类高维状态动作空间任务时,存在探索性能差、训练过程不稳定、样本效率低等问题。针对上述问题,文献[16]提出了基于内在奖励的强化学习算法,使得智能体可以高效地对环境进行探索,并且单调提升策略性能。

目前提升样本效率的方法主要有 off-policy 类算法^[15,17]以及基于模型的算法。前者由于行动策略与目标策略不同,需要设计合理的重要性采样方法,并对超参数进行反复调整,否则会使学习过程出现较大偏差,导致智能体学习不稳定,收敛到局部最优策略;后者通过使智能体学习环境的动态模型,从而提升样本效率,但当前仍存在探索能力低下^[18-19]、数据收集效率较低^[20-21]、价值函数预测偏差较大^[22-23]的问题。本文研究利用基于模型的方法结合内在奖励强化学习算法,提出基于模型的强化学习算法在无人机升空平台路径规划中的应用,在保证最终性能的前提下提升样本效率,以使用较少数据完成对于智能体的训练。

1 模拟环境构建

本节主要阐述无人机升空平台通信保障任务的模拟环境构建工作,该模拟环境不仅为智能体提供用于训练的经验数据,同时可以作为一个算法验证平台,用于比较各类算法在任务中的性能。为了使得模拟环境贴近实际环境,首先建立城市环境中的空对地信道模型,用于估算不同情况下的路径损耗值。在此基础上,将任务归纳为一个时序决策问题,并使用OpenAI-GYM架构搭建环境。

1.1 空对地信道建模

本文基于城市环境建立一个空对地信道路径损耗模型,主要考虑城市建筑物对信号遮挡造成的路径损耗。国际电信联盟(ITU)在其官方标准文件中提出一种基于建筑物遮挡对无线电信号传输造成损耗的通用模型^[24]。该模型可适用于多种城市环境,将发射机和接收机之间的视距通信及非视距通信传输概率定义为仰角和环境参数的函数,并且通过数学推导,可以得到通过Sigmoid渐进化简后的公式:

$$P(L_{\text{LoS}}, \theta) = \frac{1}{1 + a \exp(-b[\theta - a])} \quad (1)$$

其中: a, b 为 S-curve 参数。

无人机升空平台与用户之间发生非视距传输的概率为:

$$P(N_{\text{NLoS}}, \theta) = 1 - P(L_{\text{LoS}}, \theta) \quad (2)$$

因此,传播模型的路径损耗为:

$$P_{\xi}^{\text{PL}} = F_{\text{FSPL}} + \eta_{\xi} \quad (3)$$

其中: F_{FSPL} 为自由空间损耗,是针对理想全向天线传输计算得到的损耗公式; η_{ξ} 是由环境决定的过度路径损耗, ξ 代表传播组。本文将传播模型分为视距通信和非视距通信模型,即 $\xi \in \{L_{\text{LoS}}, N_{\text{NLoS}}\}$ 。

总的路径损耗模型可以写为:

$$P^{\text{PL}} = P(L_{\text{LoS}}, \theta) \times P_{L_{\text{LoS}}}^{\text{PL}} + P(N_{\text{NLoS}}, \theta) \times P_{N_{\text{NLoS}}}^{\text{PL}} \quad (4)$$

其中: P^{PL} 是信道模型的总路径损耗,可以计算无人机升空平台与每个地面移动用户之间信号的路径损耗。

1.2 任务优化方程

无人机升空平台通信保障任务的目标是使无人机升空平台在应急通信保障任务期间最大化所有用户的下行速率之和,同时需要保证任何用户的下行速率高于预设的门限速率,并保证每个用户不会出现由多普勒频移造成的快衰落。

无人机升空平台与一个地面移动用户的三维关系如图1所示。在图1中,参数 h 和 L 分别表示无人机升空平台的飞行高度以及用户之间的水平面距离,参数 \mathbf{V}_f 和 \mathbf{V}_m 为无人机升空平台及用户的速度向量, \mathbf{d} 是三维坐标系中无人机平台位置指向用户位置的向量。

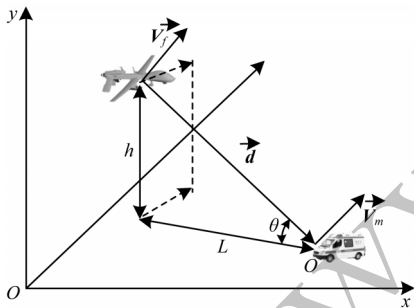


图1 无人机升空平台与用户的关系

Fig.1 Relationship between UAV aerial platform and user

此外,定义光速为 c ,信号频率为 f ,基站发射功率为 P_s ,带宽为 W ,高斯白噪声的功率为 N 。由此,根据多普勒频移定理,可以得到用户 m 在时隙 t 收到的信号频率为:

$$f_{mt} = f \left(\frac{c - \mathbf{v}_m \cdot \mathbf{d} / |\mathbf{d}|}{c - \mathbf{v}_f \cdot \mathbf{d} / |\mathbf{d}|} \right) \quad (5)$$

通过式(5)可以计算得到路径损耗 P^{PL} (单位为 dB)。所以,用户 m 在时隙 t 收到的信号功率为:

$$Pr_{mt} = 10 \lg(1000 \times P_s) - P^{\text{PL}} \quad (6)$$

通过香农公式可以得到理论上用户的最大下行速率:

$$C_{mt} = W \lg \left(1 + \frac{Pr_{mt}}{N} \right) \quad (7)$$

其中: C_{mt} 是用户 m 在时隙 t 的下行速率。

定义模拟环境在时隙 t 的奖励值为:

$$R_t = \begin{cases} \sum_{m=1}^M C_{mt}, & C_{mt} > f_{mt}, C_{mt} \geq C_{th} \\ 0, & C_{mt} < f_{mt}, C_{mt} \leq C_{th} \end{cases} \quad (8)$$

其中: M 和 C_{th} 分别为用户的数量和任务预设的用户最小门限下行速率。为了防止用户接收信号发生快衰落,需要确保符号时间大于相关时间,即 $C_{mt} > f_{mt}$ 。同时,要保证每个用户的下行速率高于设定的门限速度,所以要设置 $C_{mt} \geq C_{th}$,如果这两个条件都满足,则时隙 t 的奖励值是所有用户下行速率之和,否则为 0。设任务总的收益为:

$$G_t = \sum_{t=0}^T R_t, R_t = 0, T = t \quad (9)$$

即设置总的收益为所有时隙奖励值的和,但如果某个时隙的奖励值为 0,即触发了约束条件,则任务直接结束。基于上述分析,将无人机升空平台的应急通信保障问题概括为一个马尔科夫时序决策问题,可以采用强化学习的手段进行求解,目标就是最大化累积收益 G_t 。

在得到时序决策优化方程后,使用 OpenAI-Gym 架构^[25]进行环境构建。任务设置如下:在尺寸为 50 km×50 km×5 km 的城区范围内,随机分布着一些高度在 50~150 m 的建筑物。无人机升空平台为地面随机分布的 10 个移动目标提供通信保障,无人机升空平台可以在 0°~360°范围内调整飞行方向,在 0°~180°方位内调整飞行仰角,在每小时 180~300 km 范围内调整飞行速度。无人机升空平台需要保证每个用户的下行速率大于门限速率,同时防止由于多普勒频移造成的快衰落。在此前提下,任务的目标是最大化用户的总下行速率。任务中如果出现飞机碰撞到建筑物,则判定实验结束,并返回-100的奖励值,如果出现任何一个用户的下行速率低于阈值速率或由于多普勒频移出现了快衰落现象,则判定实验结束,并返回-50的奖励值;如果在通信保障任务期间未发生上述问题,则返回奖励值 100。

2 算法设计

在利用无模型算法进行学习时,为了准确估计价值函数,根据任务的复杂性不同,需要采样上万幕的数据才能得到较为准确的价值估计网络。因此,本文借鉴 MVE 算法^[23]的思想,采用基于模型的算法对动态模型进行学习,其中包含 3 个重要的待学习函数:状态转移函数 $T_c(s, a)$ 用来预测后继状态;状态终止预测函数 $d_c(s)$ 用来预测状态 s 为终止状态的概

率;奖励预测函数 $r_\phi(s, a, s')$ 用来预测返回的奖励值。状态价值函数被设定为结合了短期和长期价值函数的形式,短期价值函数是通过学习到的环境动态模型经过数步规划得到的奖励值之和,而长期价值函数则是通过神经网络直接预测得到的价值函数,形式如下:

$$T_H^{\text{MVE}}(r, s') = r + \left(\sum_{i=1}^H D^i \gamma^i r_\phi(s'_{i-1}, a'_{i-1}, s'_i) \right) + D^{H+1} \gamma^{H+1} Q_\theta^*(s'_H, a'_H) \quad (10)$$

其中: $s'_0 = s'$, $a'_i = \pi_\phi(s'_i)$ 为状态 s'_i 在策略 π 下得到动作; $s'_i = T_\zeta(s'_{i-1}, a'_{i-1})$ 为根据状态转移函数得到的后继状态; $D^i = d(s') \prod_{j=1}^i (1 - d_\zeta(s'_j))$ 用来判断任务是否终止。从式(10)可以看出,算法根据收集到的经验数据使用两段式的方法对状态 s' 进行评估, r 为状态转移到 s' 所获得的即时奖励值,而中间一项则表示算法根据学习到的动态模型进行 H 步规划得到的奖励值之和, H 的选择不宜过大,否则规划会产生较大方差,影响价值函数的计算,最后一项则是利用神经网络预测得到的 H 步之后的动作价值函数。短期规划和长期预测的结合使得价值函数的计算更加准确。

但是MVE算法只有在当模型复杂度不高,并且在所有学习到的动作价值函数具有相似的误差时具有较好性能。当模型较为复杂时,MVE算法难以调整固定的超参数 H ,而模型误差的累积会导致价值函数评估出现严重偏差。为了解决上述问题,需要综合考量 $H+1$ 个不同预测步长的MVE形式的状态价值来计算得到一个合适的价值函数。候选的TD目标为: $T_0^{\text{MVE}}, T_1^{\text{MVE}}, \dots, T_H^{\text{MVE}}$,即考量从0步规划到 H 步的 $H+1$ 种不同状态价值。传统的方法是使用对于候选目标的平均或者以指数衰减的方法对候选目标值进行加权的方法,本文选择通过平衡 Q 函数学习中的误差以及规划模型的误差,得到对于候选目标更好的加权方式。针对每个候选 T_i^{MVE} ,其在规划中有3个重要参数,分别为 Q 函数预测参数 θ 、奖励函数预测参数 ϕ 、状态转换函数预测参数 ζ ,如式(10)所示,它们共同作用组成一个 $H=i$ 步的TD目标 T_i^{MVE} 。为了增强算法的鲁棒性,设置一个候选的TD目标中有 L 个预测参数 $\theta = \{\theta_1, \theta_2, \dots, \theta_L\}$, N 个奖励函数预测参数 $\phi = \{\phi_1, \phi_2, \dots, \phi_N\}$, M 个状态转移预测参数 $\zeta = \{\zeta_1, \zeta_2, \dots, \zeta_M\}$ 。

算法的概述图如图2所示。图2展示了 $M=N=L=2$ 情况下 (s_0, a_0) 的TD目标值 T_2^{MVE} 的估计值,可以通过这些数据求得 T_2^{MVE} 的均值 T_2^μ 和方差 $T_2^{\sigma^2}$ 。为了找到合适的权值 w ,使得加权后的TD目标值之和与

真实的动作价值的均方误差最小,将两者的泛化误差进行分解得到:

$$\begin{aligned} E \left[\left(\sum_{i=0}^H w_i T_i^{\text{MVE}} - Q^\pi(s, a) \right)^2 \right] = \\ \text{Bias} \left(\sum_i w_i T_i^{\text{MVE}} \right)^2 + \text{Var} \left(\sum_i w_i T_i^{\text{MVE}} \right) \approx \\ \text{Bias} \left(\sum_i w_i T_i^{\text{MVE}} \right)^2 + \sum_i w_i^2 \text{Var} (T_i^{\text{MVE}}) \quad (11) \end{aligned}$$

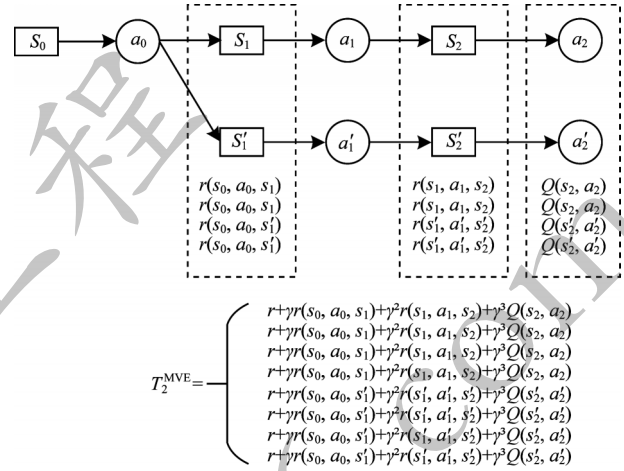


图2 基于模型算法的概述图

Fig.2 Overview figure of model-based algorithm

为使得均方误差最小,使用经验数据中估计得到的方差来估计方差项,并最小化方差项。采用逆方差权重法,将 w_i 设置为 $\text{Var}(T_i^{\text{MVE}})$ 的倒数,并对最终结果进行规范化,最终得到加权后的状态价值函数为:

$$T(r, s) = \frac{\sum_{i=0}^H w_i}{\sum_j w_j} T_i^\mu \quad (12)$$

其中: $w_i^{-1} = T_i^{\sigma^2}$ 。

将算法与内在奖励RL算法以及impala并行架构结合,最终得到基于模型的内在奖励强化学习算法,算法流程架构如图3所示。可以看到,算法采用并行架构完全解耦了数据采集和策略更新过程。Worker独立地进行经验数据收集,在结束一幕数据交互后,同步Learner最新的策略,并将收集到的数据存入Buffer。Learner周期地从Buffer中提取数据进行更新,通过V-trace方法对行动策略采集到的数据进行重要性采样,得到适合目标策略学习的价值函数预测值,分内部、外部奖励两个部分使用上述基于模型的方法对价值函数进行评估,最终合并内部奖励和外部奖励预测得到的价值函数,并利用PPO的方法对策略进行更新。实验结果表明,该方法在智能体取得相同性能的情况下提高了样本效率。

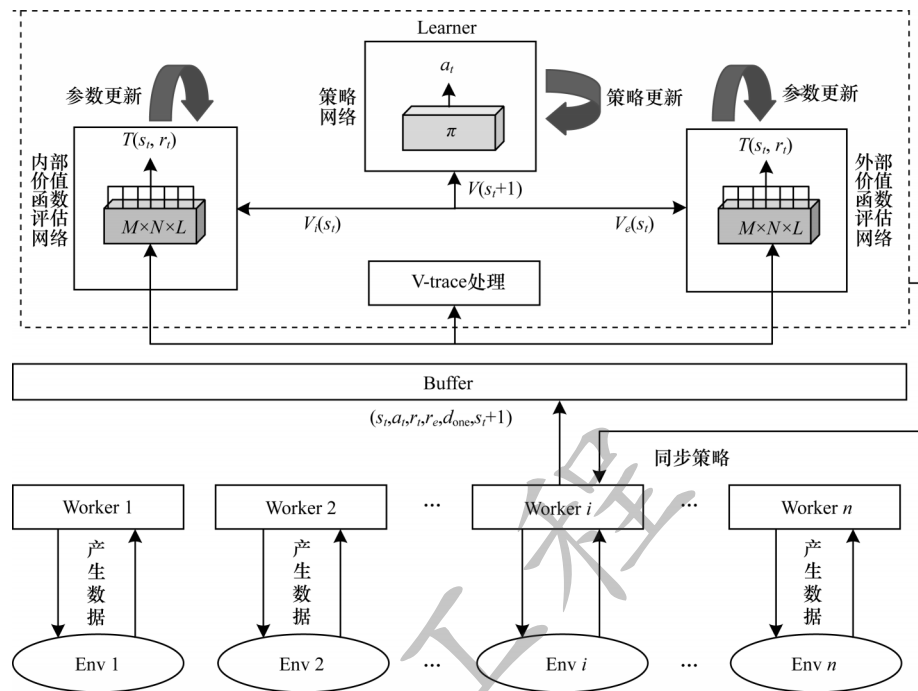


图3 基于模型的内在奖励算法结构

Fig.3 Structure of model-based intrinsic reward algorithm

3 实验结果与分析

本程序使用 python3.8 编写,运行环境为 Win 10 操作系统,装有 2 块 NVIDIA 3090 显卡以及 64 GB 内存。实验中神经网络均由全连接网络和 ReLu 网络组成,使用 32 个并行的实验环境进行数据采集。本文提出的基于模型的内在奖励算法与基于 Impala 架构的无模型内在奖励算法的性能对比如图 4 所示。

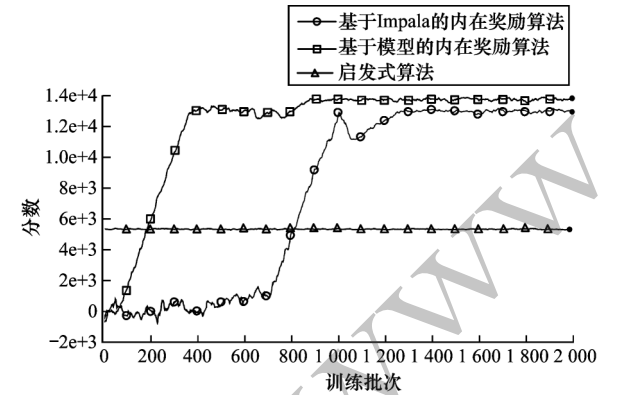


图4 不同算法的性能对比

Fig.4 Performance comparison of different algorithms

从图4可以看出,本文算法相较于拥有相同架构但不使用对环境动态模型进行学习的算法具有更好的性能,可以利用很少的经验数据快速完成对于策略的学习,并且学习过程更加稳定。为了比较本文算法与非强化学习启发式算法的性能,基于文献[3-5]的思想,构建一套简化的启发式算法。该算法将当前分布在地面的多个用户包含在一个最小的

圆内,要求无人机始终保持在圆心位置,速度方向则为所有用户当前速度向量之和的方向。可以看到,启发式算法在环境中可以达到近 6 000 分的水平,微小的波动是由于地面用户在遇到障碍物时进行随机避障,速度方向并不保持一致,从而导致无人机飞行方向发生偏移,进而影响最终得分情况。相较于启发式算法,本文算法在前期学习过程得分较差,但当智能体能够对状态价值函数进行准确评估后,最终算法的得分远高于启发式算法。

此外为了说明的本文算法相较于其他基于模型算法的优势,在模拟环境中采用了多种算法进行测试比较,结果如图5所示。

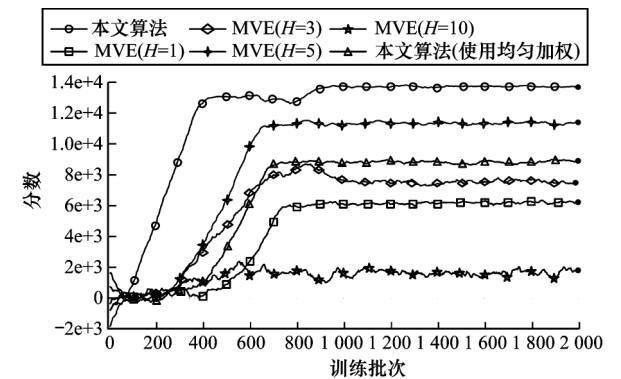


图5 本文算法与不同视界 MVE 算法的性能对比

Fig.5 Performance comparison between the proposed algorithm and MVE algorithm in different horizons

从图5可以看出,相比于 MVE 采用固定规划值 (H) 的情况,基于组合规划值的方法训练速度和效果更好,同时训练过程更为平稳,并且对于 MVE 类

规划值固定的算法,如何调节超参数 H 也是一个难题,从图5可以看出,当 H 从1提高到5的过程中,规划值的增大减小了价值函数预测的方差,而准确的价值函数提高了算法的学习速率,也决定了最终收敛到的策略性能。而当 H 取10时,智能体在整个训练过程中波动很大,并且最终无法学习到一个较好的策略。原因在于:在训练初期,当预测模型没有得到准确学习时,过长的规划值会导致价值函数方差、偏差都较大,在这种情况下由于方差、偏差的累积,智能体始终无法学到准确的预测模型参数以及价值函数,这就使得智能体在训练过程中全程无法进行有效的策略迭代。所以,对于固定规划值类的算法,超参数的调整是一个难题。而使用均匀加权训练算法与本文算法有着相同的架构,但在组合规划值时,权值使用的是均匀加权算法。可以看出,该算法的速度和最终性能都与本文算法有差距。

实验中还对算法对于不同超参数集的鲁棒性进行了研究,利用20组有较大差异的超参数集对算法进行了测试,并且对最终得分求均值,结果如图6所示。

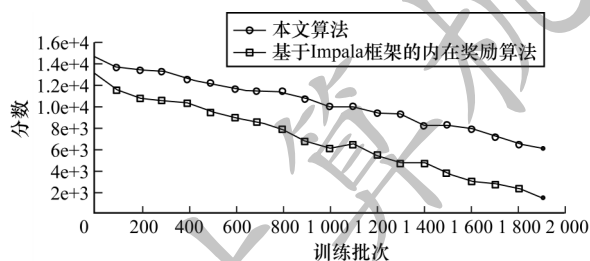


图6 不同算法的鲁棒性对比

Fig.6 Robustness comparison of different algorithms

图6比较了基于模型的权值组合规划值扩展算法与基于Impala框架的内在奖励算法在20组不同超参数集下作用于模拟环境中的平均得分。从图6可以看出,基于Impala框架的内在奖励算法在使用接近20组超参数集时,其得分均值已低于2000分,而基于模型的权值组合规划值扩展算法稳定在6000分左右。实验结果表明,基于模型的算法针对不同超参数具有更强的鲁棒性。原因在于:基于模型的权值组合规划值扩展算法在训练过程中对于环境动态模型的学习,在一定程度上弥补了超参数设置带来的价值函数预测偏差。

4 结束语

本文针对强化学习算法在无人机升空平台路径规划任务中存在的样本效率低的问题,提出基于模型的内在奖励强化学习算法。通过将任务概述为一个时序决策优化问题,基于OpenAI-GYM构建模拟环境,并结合规划与预测的方法提高价值函数的评估准确性。实验结果表明,该算法在保证智能体性能的前提下,在样本效率、学习速度、算法鲁棒性上都有较大提升。下一步将研究提升算法的迁移能

力,并结合迁移学习和元学习的思想对算法进行改进,以将训练完毕的智能体投入到相似的场景中执行任务。

参考文献

- [1] PEPPER R. Cisco visual networking index global mobile data traffic forecast update[EB/OL]. [2021-09-30]. https://www.gsma.com/spectrum/wpcontent/uploads/2013/03/Cisco_VNI-global-mobile-data-traffic-forecastupdate.pdf.
- [2] ROMERO D, LEUS G. Non-cooperative aerial base station placement via stochastic optimization[C]//Proceedings of the 15th International Conference on Mobile Ad-Hoc and Sensor Networks. Washington D. C., USA: IEEE Press, 2019: 131-136.
- [3] ZENG Y, ZHANG R. Energy-efficient UAV communication with trajectory optimization[J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3747-3760.
- [4] LYU J B, ZENG Y, ZHANG R, et al. Placement optimization of UAV-mounted mobile base stations[J]. IEEE Communications Letters, 2017, 21(3): 604-607.
- [5] ALZENAD M, EL-KEYI A, LAGUM F, et al. 3-D placement of an unmanned aerial vehicle base station for energy-efficient maximal coverage[J]. IEEE Wireless Communications Letters, 2017, 6(4): 434-437.
- [6] KALANTARI E, YANIKOMEROGLU H, YONGACOGU A. On the number and 3D placement of drone base stations in wireless cellular networks[C]//Proceedings of the 84th IEEE Vehicular Technology Conference. Washington D. C., USA: IEEE Press, 2016: 1-6.
- [7] AL-HOURANI A, KANDEEPAN S, LARDNER S. Optimal LAP altitude for maximum coverage[J]. IEEE Wireless Communications Letters, 2014, 3(6): 569-572.
- [8] GUO J L, HUO Y H, SHI X J, et al. 3D aerial vehicle base station (UAV-BS) position planning based on deep Q-learning for capacity enhancement of users with different QoS requirements[C]//Proceedings of the 15th International Wireless Communications & Mobile Computing Conference. Washington D. C., USA: IEEE Press, 2019: 1508-1512.
- [9] BAYERLEIN H, DE KERRET P, GESBERT D. Trajectory optimization for autonomous flying base station via reinforcement learning[C]//Proceedings of the 19th IEEE International Workshop on Signal Processing Advances in Wireless Communications. Washington D. C., USA: IEEE Press, 2018: 1-5.
- [10] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. Computer Science, 2013, 25: 253-262.
- [11] WANG Q, ZHANG W Q, LIU Y W, et al. Multi-UAV dynamic wireless networking with deep reinforcement learning[J]. IEEE Communications Letters, 2019, 23(12): 2243-2246.
- [12] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. Artificial Intelligence, 2016, 30(1): 14-20.
- [13] LIU C H, MA X X, GAO X D, et al. Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning[J]. IEEE Transactions on Mobile Computing, 2020, 19(6): 1274-1285.

(下转第269页)

(上接第260页)

- [14] QI H, HU Z Q, HUANG H, et al. Energy efficient 3-D UAV control for persistent communication service and fairness; a deep reinforcement learning approach[J]. IEEE Access, 2020, 8: 53172-53184.
- [15] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1509.02971>.
- [16] YANG S M, SHAN Z, CAO J, et al. Path planning of UAV base station based on deep reinforcement learning[J]. Procedia Computer Science, 2022, 202: 89-104.
- [17] FUJIMOTO S, VAN HOOFF H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1802.09477>.
- [18] ANTHONY T, TIAN Z, BARBER D. Imagination-augmented agents for deep reinforcement learning[C]// Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017: 5360-5370.
- [19] NAGABANDI A, KAHN G, FEARING R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning[C]// Proceedings of IEEE International Conference on Robotics and Automation. Washington D. C., USA: IEEE Press, 2018: 7559-7566.
- [20] BUCKMAN J, HAFNER D, TUCKER G, et al. Sample-efficient reinforcement learning with stochastic ensemble value expansion[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1807.01675>.
- [21] KURUTACH T, CLAVERA I, DUAN Y, et al. Model-ensemble trust-region policy optimization[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1802.10592>.
- [22] FEINBERG V, WAN A, STOICA I, et al. Model-based value estimation for efficient model-free reinforcement learning[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1803.00101>.
- [23] CLAVERA I, ROTHFUSS J, SCHULMAN J, et al. Model-based reinforcement learning via meta-policy optimization[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1809.05214>.
- [24] Recommendation ITU-R. Propagation data and prediction methods required for the design of terrestrial broadband millimetric radio access systems operating in a frequency range of about 20~50 GHz[R]. Geneva, Switzerland, 2001.
- [25] BROCKMAN G, CHEUNG V, PETERSSON L, et al. OpenAI Gym[EB/OL]. [2021-09-30]. <https://arxiv.org/abs/1606.01540>.