

面向高频动态到达的天文卫星机遇 目标任务规划方法*

王旭航^{1,2} 吴海燕¹

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

摘要 以巡天设备每天将探测到数以万计的变源天体以及对变源天体的观测需求增长为背景, 形成了由高频动态到达的机遇目标 (ToO) 及其后随观测任务组成的长序列任务规划问题. 该问题具有观测事件随机性、数据获取高时效性、可选择性多和约束复杂的特点, 常被视为 NP (非确定性多项式) 难题, 因此获取监督学习的标签数据不易. 而针对采用无监督学习的深度强化学习 (DRL) 方法求解长序列任务规划问题时, 卫星作为智能体难以快速收敛至全局最优策略. 为此本文借鉴局部注意力 (LA) 机制的思想对指针网络 (PN) 进行改进, 提出局部注意力指针网络 (LA-PN) 算法. 该算法通过引入时间窗口的方式, 使模型专注于对当前决策有重要影响的序列部分, 减少了无效探索. 通过仿真结果对比分析, 验证算法的收益性、实时性和泛化性.

关键词 机遇目标, 后随观测, 任务规划, 深度强化学习, 局部注意力指针网络

中图分类号 V474

Scheduling Methods for Astronomical Satellite Target of Opportunity Tasks with High-frequency Dynamic Arrivals

WANG Xuhang^{1,2} WU Haiyan¹

1(National Space Science Center, Chinese Academy of Sciences, Beijing 100190)

2(University of Chinese Academy of Sciences, Beijing 100049)

Abstract Based on the background where sky survey equipment detects tens of thousands of variable sources daily and the demand for observing these variable sources increases, a long-sequence task planning problem is formed. This problem comprises dynamically arriving high-frequency Targets of Opportunity (ToO) and their subsequent follow-up observation tasks. This type of problem is classified as an NP hard problem due to the randomness of observation events, strong timeliness of data acquisition,

* 中国科学院空间科学战略性先导科技专项资助 (XDA15040100)

2024-10-08 收到原稿, 2025-03-08 收到修定稿

©The Author(s) 2025. This is an open access article under the CC-BY 4.0 License
(<https://creativecommons.org/licenses/by/4.0/>)

high selectivity of decision paths, and the coupling effect of complex constraints such as Earth occlusion and observation time windows. Traditional supervised learning methods are difficult to apply due to the difficulty in obtaining high-quality labeled data, while Deep Reinforcement Learning (DRL) combines the perceptual ability of deep learning with the sequential decision optimization mechanism of reinforcement learning to construct end-to-end autonomous decision-making agents, demonstrating the universal advantage of handling complex planning problems. This article is based on the DRL framework and constructs a Markov Decision Process (MDP) model for long sequence task planning problems. State space encoding, action space encoding, and reward functions are defined in the task planning problem to guide reinforcement learning algorithms to converge towards optimizing the target set. However, classical DRL methods have significant limitations in unsupervised learning mode - when faced with long sequence dynamic task flows, agents have difficulty quickly converging to the global optimal strategy due to the large exploration space, and lack sensitivity to local key tasks. Therefore, this paper improves the Pointer Network (PN) by drawing on the concept of the Local Attention (LA) mechanism, proposing the Local Attention Pointer Network (LA-PN) algorithm. This algorithm innovatively introduces a sliding time window constraint on the attention range, allowing the model to dynamically focus on task subsequences related to the current decision, and filter high-value targets through local context awareness to reduce redundant computation. By comparing and analyzing simulation results, verify the profitability, real-time performance, and generalization of the algorithm.

Key words Target of Opportunity (ToO), Follow-up observation, Task scheduling, Deep Reinforcement Learning (DRL), Local Attention Pointer Network (LA-PN)

0 引言

暂现源和剧烈爆发天体的观测在时间和空间上都具有不可预知性^[1]. 对这些目标的观测称为机遇目标 (ToO) 观测. 空间的机遇目标, 例如引力波、伽马射线暴、X 射线暴等, 都是重要的天文观测现象, 蕴含着天文学和物理学发展的新规律^[2]. 由于 ToO 的重要性和不可预知性, 一旦探测到, 便需迅速开展后随观测. 后随观测不仅可以提供关于这些事件的更多细节, 而且还可以作为多信使天文学的关键部分, 与引力波探测等其他观测手段相结合, 为研究提供关于宇宙极端事件更全面的视角.

空间天文卫星的任务是通过在轨空间探测实现对 ToO 的观测和后随观测. 到 2030 年左右, 一系列巡天设备每天将探测到成千上万个 ToO, 例如中国空间科学先导专项二期已成功部署的怀柔一号卫星 (GECAM)、天关卫星 (EP)、中法合作的太空望远镜卫星 (SVOM) 等多个空间天文卫星任务, 以及已经开展概念研究的全变源追踪猎人星座计划 (CATCH), 美国计划中的罗曼太空望远镜, 以及欧洲空间局的欧几里得太空望远镜等. 观测任务可选择性多、观测事件随机性、数据获取高时效性是空间天文卫星任务的

重要特点, 这对空间天文卫星测运控提出了很高的快速响应要求, 因此需要及时制定出能满足需求的最优或近似最优的任务观测序列, 这有助于空间天文卫星及时捕捉到重大科学发现.

空间天文卫星的科学任务规划问题是一个复杂的多约束条件下的多目标优化问题, 目前有关该问题的研究不断深入, 并基于启发式算法、智能优化算法和机器学习算法进行求解.

美国国家航空航天局 (NASA) 在哈勃太空望远镜项目中研发了 SPIKE 任务规划系统, 建模为约束满足问题, 内置了许多启发式规划方法^[3]. Wu 等^[4] 针对天文观测任务建模为多目标优化数学模型, 设计实现了基于带精英策略的非支配排序遗传算法 (NSGA-II) 的任务规划方法. Huang 等^[5] 针对硬 X 射线调制望远镜 (HXMT) 长期任务规划问题采用贪婪算法和遗传算法进行求解. Yin 等^[6] 针对空间天文卫星的超额任务调度问题, 提出了一种改进的人工蜂群算法, 提高了算法寻优能力和收敛速度. 但这些启发式算法和智能优化算法存在启发式因子设计难度大和优化收敛速度较慢的问题, 不适合处理大量动态到达且需要快速响应的用户需求.

深度学习通过神经网络构建的端到端模型,可以避免传统算法的迭代优化过程,从而提高任务规划问题的求解效率.然而卫星任务规划是一个 NP-Hard 问题,深度学习通常需要大量标签数据,而这些数据往往难以获取.此外,深度学习的性能可能无法超越获取标签数据的算法性能.

与此不同,深度强化学习 (DRL) 适用于处理序贯决策问题,能够在不确定环境中快速学习决策策略. DRL 利用无标签数据进行训练,训练好的端到端模型实现对 ToO 和后随观测的快速响应.

Wang^[7] 针对卫星信息获取、灾害防治等应用中随机到达的观测需求,建模为一个马尔可夫决策过程问题,并利用异步优势演员-评论家 (A3 C) 算法学习在线调度问题中的最优决策策略,具有较高的收益性和时效性. Ma^[8] 针对密集观测场景下的敏捷成像卫星任务规划问题,提出多头注意力指针网络 (MHA-PN) 算法,采用演员-评论家 (AC) 算法训练模型,具有较高的求解效率. Zhao^[9] 针对成像卫星用户观测需求数量不固定且观测任务动态到达等特点,设计并实现了一种基于强化学习和循环神经网络的有效方法. 决斗双深度 Q 网络-长短期记忆模型 (D3 QN-LSTM) 来解决单星自主任务规划问题,提高决策的稳定性.

以上基于 DRL 的卫星任务规划研究存在优化目标单一、状态空间较大导致算法模型的训练难以收敛到最优策略的问题. 本文针对高频动态到达的 ToO 和后随观测任务组成的长序列观测场景,建立指针网络 (PN) 序贯决策模型,从任务优先级、应急任务响应度方面考虑多目标优化. 由于长序列任务规划的状态选择可能性较多,卫星作为智能体难以快速收敛至全局最优策略的问题,借鉴局部注意力 (LA) 机制的思想对指针网络 (PN) 进行改进,提出局部注意力指针网络 (LA-PN) 算法,通过引入时间窗口聚焦对当前决策有重要影响的序列部分,从而减少了无效探索.

1 数学模型

1.1 问题描述

卫星任务规划是一个 NP 问题,针对高频动态到达的 ToO 和后随观测任务组成的长序列任务观测场

景,天文卫星 ToO 和后随观测任务规划问题可以简要描述为,在接收到来自终端用户上传的 ToO 和后随观测需求后,计算每轨对需求目标的可见时间窗口,并对时间窗口进行排序,以此作为输入任务序列,考虑任务时间窗口及任务间转移机动时间、每轨调姿次数等约束条件,设计相应任务规划方法以实现卫星资源的有效利用和配置,在其能力范围内最大化科学观测收益,并尽快响应 ToO 观测任务^[10].

1.2 问题假设

ToO 和后随观测任务规划问题受多个因素的影响,为了聚焦问题进行以下场景假设.

假设 1 实际规划中需考虑卫星平台的电量等约束,这里假设卫星电量能够满足观测需求.

假设 2 随着天基通信网络的快速发展,预计未来 5~10 年后卫星实时通信将成为可能,这里不再考虑数传资源约束.

假设 3 将卫星的多个有效载荷抽象为一个有效载荷.

假设 4 卫星在执行一个观测任务时,不能被抢占和打断.

假设 5 观测对象的属性是 10 min 曝光时间就可以满足科学需求.

1.3 符号定义

为形式化描述天文卫星 ToO 和后随观测任务规划问题,定义和公式符号描述如下.

定义 1 $\mathbf{X} = \{\mathbf{x}_i^t = \{\mathbf{s}_i, \mathbf{d}_i^t\}, i = 1, \dots, M\}$ 表示一个规划周期内随时间已到达的任务的可见时间窗口排序,作为模型输入任务序列, M 为已到达待规划任务个数. $\mathbf{x}_i^t = \{\mathbf{s}_i, \mathbf{d}_i^t\}$ 表示在时间步 t 时编号为 i 的任务,将每个候选任务用静态属性和动态属性两部分进行描述,静态属性定义为 $\mathbf{s}_i = \{T_i^{\text{ws}}, D_i, T_i^{\text{we}}, A_i, P_i\}$, $\mathbf{s}_i$ 中的元素分别为任务的可见时间窗口开始时间、观测所需时间、可见时间窗口结束时间、观测所需卫星侧摆角度、优先级,动态属性定义为 $\mathbf{d}_i^t = \{W_i^t, A_i^t, O_i^t, N_i^t, J_i^t\}$, $\mathbf{d}_i^t$ 中的元素分别为在时间步 t 时,任务是否满足时间窗口约束、任务是否已经访问过、卫星侧摆角度、每轨剩余机动次数、执行任务是否需要调姿. $\mathbf{X}^t$ 表示在时间步 t 时剩余的候选任务序列^[8]. 在每个解码时间步 t ,静态元素不变,动态元素会根据上一个输出节点占用资源情况发生动态变化.

定义 2 规划得到的任务集合为 $\mathbf{Y} = \{\mathbf{y}_i,$

$y_2, \dots, y_N\}$, 其中 N 表示执行任务的数量. $\mathbf{Y}^t$ 表示在时间步 t 时已规划的执行观测任务序列.

定义 3 执行下一个观测活动 $\mathbf{x}_{i+1}^t$ 调姿距离.

如图 1 所示, 假设载荷的位置可以粗略地看为是等值坐标系的原点, 当前卫星侧摆角度 φ^t , 下一个任务需要卫星侧摆角度 A_{i+1} , 卫星的视场角为 θ_{FOV} . 通常观测时, 为保证效果, 要将目标源置于视场中心或近中心位置. 一旦目标源处于视场边缘, 卫星就得调姿, 重新指向目标源进行观测. 因此, 可依据卫星自身特点设定一个小于视场角 θ_{FOV} 的参数 $\theta_{\text{FOV}'}$, 以此辅助观测操作.

$$\Delta\theta_i = \left[\varphi^t - \frac{\theta_{\text{FOV}'}}{2}, \varphi^t + \frac{\theta_{\text{FOV}'}}{2} \right], \quad (1)$$

$$\Delta\theta_{i+1} = \left[A_{i+1} - \frac{\theta_{\text{FOV}'}}{2}, A_{i+1} + \frac{\theta_{\text{FOV}'}}{2} \right]. \quad (2)$$

若 $\Delta\theta_i \cap \Delta\theta_{i+1} = \emptyset$, $\emptyset$ 为非空集, 则执行任务 $\mathbf{x}_{i+1}^t$ 前需要调姿, 调姿距离为 $L_{i,i+1} = |\varphi_{i+1} - \varphi_i^t|$; 若 $\Delta\theta_i \cap \Delta\theta_{i+1} \neq \emptyset$, 则执行任务 $\mathbf{x}_{i+1}^t$ 前不需要调姿, 调姿距离为 $L_{i,i+1} = 0$.

1.4 预处理阶段

1.4.1 约束条件

空间天文卫星在规划 ToO 和后随观测任务时, 需要根据约束条件进行决策, 是一个复杂的多约束条件下的多目标优化问题. 随着天基通信网络的快速发展, 预计未来 5~10 年后卫星实时通信将成为可能, 这将消除数传和测控资源约束. 对卫星规划问题的其

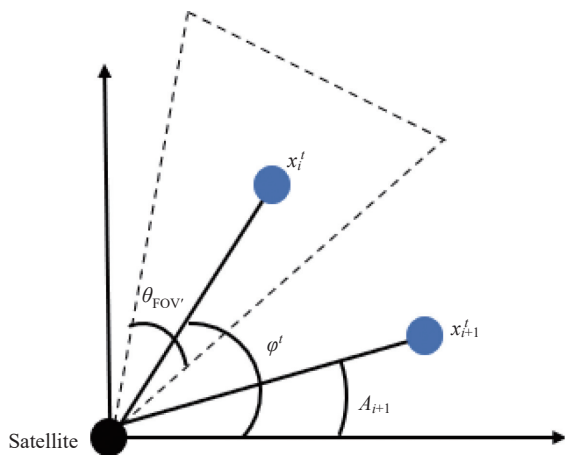


图 1 二维坐标系中卫星望远镜的视场

Fig. 1 FOV of satellite telescope in two-dimensional coordinate system

余约束条件进行分析.

约束 1 观测方向与太阳矢量夹角约束. 为了避免观测仪器被太阳直射, 在观测过程中通常要求观测方向与太阳矢量之间的夹角不小于设定的最小夹角约束.

约束 2 异常环境约束. 过境南大西洋异常区期间卫星不进行有效观测.

约束 3 地球遮挡约束. 当地球处于卫星观测方向与目标天体之间并形成遮挡时, 卫星需暂停对该目标天体的观测, 防止采集无效数据. 即卫星仅能在与观测目标彼此可见的时段内执行观测任务.

约束 4 任务间转移机动时间约束. 卫星调整每单位角度的姿态的时间消耗为 T_s . 下一个任务的开始时间必须大于当前任务结束时间与卫星侧摆机动时间之和, 即

$$T_{i+1}^{\text{ws}} > T_i^{\text{ws}} + d_i + t_{\text{con}}, \quad (3)$$

$$t_{\text{con}} = |L_{i,i+1}| T_s. \quad (4)$$

其中, T_i^{ws} 和 T_{i+1}^{ws} 分别表示任务 i 和任务 $i+1$ 的可见时间窗口开始时间, d_i 表示观测任务 i 所需时间, $|L_{i,i+1}|$ 表示卫星调姿距离.

约束 5 观测任务最短持续时间约束. 根据 1.2 节中的假设 5, 对于每个候选任务 $\mathbf{x}_i^t$, 单次观测时长不少于 10 min 就可以满足科学需求, 即

$$\forall i \in M, d_i \geq 600. \quad (5)$$

其中, M 表示已到达待规划任务个数, d_i 表示观测任务 i 所需时间.

约束 6 每轨最多调姿次数约束. 每轨最多调姿次数为 N 次.

1.4.2 卫星轨道参数

天文卫星轨道参数参照 SVOM 卫星设置 (见表 1). 在预处理阶段, 根据卫星轨道和约束条件计算观测目标可见时间窗口, 得到按照可见时间窗口排序的输入任务序列^[11].

1.5 优化准则

针对 ToO 和后随观测需求高频动态到达观测区域, 需要快速响应的任务规划需求, 智能体在每个时间步进行任务选择后的奖励函数考虑多目标优化, 优化准则如下^[11].

最大化科学观测收益. 由每个被规划执行的任务的优先级进行描述, 该值越大, 表明卫星的观测收益

越高, 有

$$\max(f_{\text{pri}}) = \frac{P_i}{P_{\max}}. \quad (6)$$

其中, P_i 表示任务 i 的优先级, $P_{\max}$ 表示已到达待规划的 M 个任务中最大的优先级。

最小化应急任务的响应时间. 通过调整不同优先级任务的观测顺序, 确保高优先级任务能够尽快得到访问, 即

$$\min(f_{\text{res}}) = \frac{o_i P_i}{M P_{\max}}. \quad (7)$$

其中, o_i 表示第 i 个规划执行任务在最终规划结果中的排序, M 表示已到达待规划任务个数。

表 1 卫星轨道参数
Table 1 Satellite orbit parameters

参数	取值
轨道高度/km	7012.2172
偏心率	0.0002961
轨道倾角/(°)	28.863529
升交点赤经/(°)	287.41907
近地点幅角/(°)	299.01349
平近点角/(°)	292.49033

最终模型在每个时间步选择一个任务后的奖励函数 R^t 由以上优化准则加权求得, 即

$$R^t = \max(\omega_1 f_{\text{pri}} - \omega_2 f_{\text{res}}). \quad (8)$$

其中, 两个权重值 ω_1 和 ω_2 可以根据任务需求分别进行设置。

2 算法模型

2.1 模型框架

利用 LA-PN 算法对密集观测场景下高频动态到达的 ToO 和后随观测任务自主规划问题进行求解. 算法的模型结构如图 2 所示, 分为 Input (输入)、Encoder (编码器)、Decoder (解码器) 和 Output (输出) 四部分。

Input 部分. $\mathbf{X} = \{\mathbf{x}_i^t = \{\mathbf{s}_i, \mathbf{d}_i^t\}, i = 1, \dots, M\}$ 作为模型的输入任务序列。

Encoder 部分. 使用 Embedding Layer (嵌入层) 和 Encoder-GRU (编码器-门控循环单元) 按时间步逐序读入输入任务序列中的每个任务 $\mathbf{x}_i^t$, $i \in [1, M]$, 直至读入整个输入序列, 生成包含整个输入序列隐藏信息的关联信息矩阵 $\bar{\mathbf{h}}^M$ 和向量 $\mathbf{C}$. Encoder-GRU 在编码时间步 t , 对应的编码单元基于上

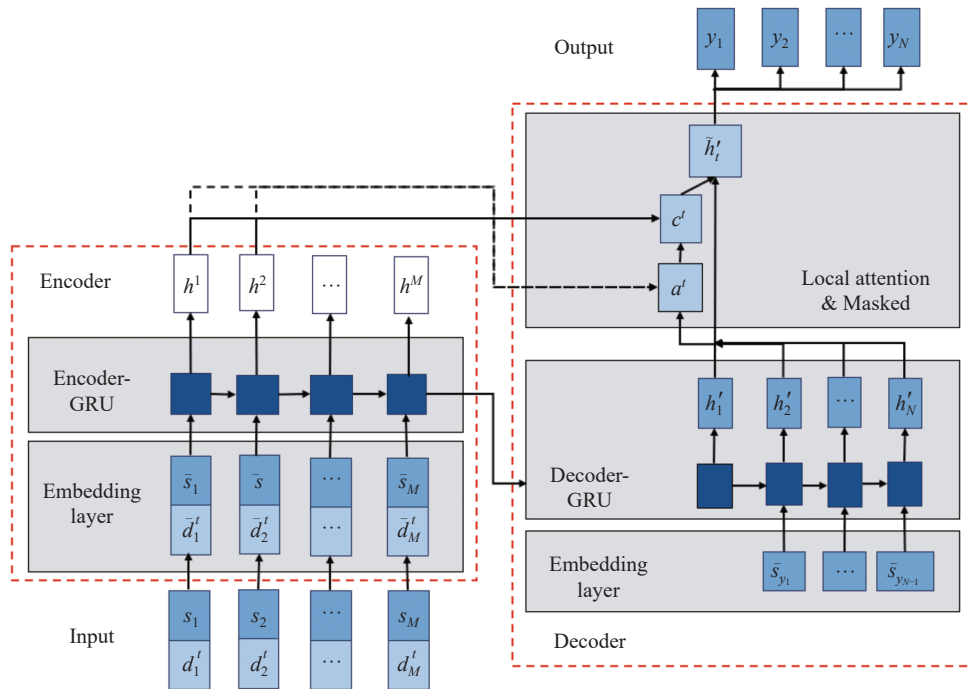


图 2 LA-PN 算法模型结构

Fig. 2 LA-PN algorithm model structure

一编码时间步的关联信息向量 $\mathbf{h}^{t-1}$ 和当前编码时间步 Embedding Layer 的输出元素 $\bar{\mathbf{x}}_i^t$, 按照

$$\mathbf{h}^t = f(\mathbf{h}^{t-1}, \bar{\mathbf{x}}_i^t) \quad (9)$$

计算元素关联信息向量, 并传递给下一步的编码单元. 将 Encoder-GRU 在每个时间步计算得到的关联信息向量拼接成的 $[\mathbf{h}^1; \mathbf{h}^2; \dots; \mathbf{h}^M]$ 记为矩阵 $\bar{\mathbf{h}}^M$. 其中, f 为神经网络 Encoder-GRU 参数化的编码函数, $t=1$ 时 $\mathbf{h}^0$ 为指定的超参数向量, 通常向量值取 0^[12].

Decoder 部分. Decoder 获得 Encoder 输出的向量 $\mathbf{C}$ 和矩阵 $\bar{\mathbf{h}}^M$, 使用 Embedding Layer 和 Decoder-GRU(解码器-门控循环单元)^[1,12] 按时间步逐序生成输出元素. Decoder-GRU 在解码时间步 t , 解码单元首先基于上一解码时间步的输出信息向量 $\mathbf{h}'_{t-1}$ 和当前解码时间步 Embedding Layer 的输出元素 $\bar{\mathbf{s}}_{y_{t-1}}$, 按照

$$\mathbf{h}'_t = g(\mathbf{h}'_{t-1}, \bar{\mathbf{s}}_{y_{t-1}}) \quad (10)$$

计算输出信息向量. 其中 g 为神经网络 Decoder-GRU 参数化的解码函数, $t=1$ 时 $\mathbf{h}'_0$ 为指定的超参数向量, 通常向量值取 0. 并将输出信息向量 $\mathbf{h}'_t$ 传递给下一个解码步^[12], 进而根据关联信息矩阵 $\bar{\mathbf{h}}^M$ 和输出信息向量 $\mathbf{h}'_t$, 基于 Local Attention 机制和 Mask 向量计算解码时间步 t 候选任务的概率分布.

Output 部分. 根据解码时间步 t 候选任务的概率分布, 选择概率值最大的任务作为输出任务 $\mathbf{y}_t$, 并将输出元素 $\mathbf{y}_t$ 传递给下一解码时间步.

2.2 指针注意力机制

指针注意力机制可以描述为在每个解码时间步 t , 使用注意力机制计算候选任务与当前输出的相关性, 并通过掩码 (Mask) 向量和 Softmax 函数确定候选任务的概率值, 概率值最大的候选任务即为下一个要观测的任务^[13]. 具体计算过程如下.

通过拼接映射点乘计算对齐向量. 根据 Encoder-GRU 在每个时间步计算得到的关联信息矩阵 $\bar{\mathbf{h}}^M$ 和解码时间步 t 的输出信息向量 $\mathbf{h}'_t$, 按照下式分别计算两者之间的相关性:

$$\mathbf{a}^t = \text{softmax}\{\mathbf{V}_a^T \tanh(\mathbf{W}_a [\bar{\mathbf{h}}^M; \mathbf{h}'_t])\}. \quad (11)$$

其中, $\mathbf{V}_a$ 和 $\mathbf{W}_a$ 为 Decoder-GRU 训练所得参数.

计算与当前输出紧密相关的上下文向量 $\mathbf{c}^t$. 按照

下式用对齐向量对关联信息矩阵 $\bar{\mathbf{h}}^M$ 进行加权累加得到上下文向量, 即

$$\mathbf{c}^t = \sum_{i=1}^M \mathbf{a}^t \bar{\mathbf{h}}^M. \quad (12)$$

将上下文向量 $\mathbf{c}^t$ 与当前步骤的输出信息向量进行拼接, 通过线性变换和 tanh 函数激活变换求得当前步骤最终表示向量, 即

$$\tilde{\mathbf{h}}'_t = \tanh(\mathbf{W}_c [\mathbf{c}^t; \mathbf{h}'_t]). \quad (13)$$

其中 $\mathbf{W}_c$ 为 Decoder-GRU 训练所得参数. 按照

$$P(\mathbf{y}^{t+1} | \mathbf{Y}^t, \mathbf{X}^t) = \text{softmax}\left[\mathbf{V}_c^T \tilde{\mathbf{h}}'_t + \ln(\mathbf{F}_{\text{Mask}})\right] \quad (14)$$

计算当前输出元素的条件概率. 其中, $\mathbf{V}_c$ 为训练所得参数, $\mathbf{F}_{\text{Mask}}$ 为 Mask 向量.

Mask 向量每位的取值为 1 或 0, 分别对应输入任务序列中每个候选任务在时间步 t 时是否满足约束条件. 若 Mask 向量的某一位为 0, 通过式 (14) 计算得到该位置的候选任务成为下一个输出的规划任务的概率为 0, 从而实现对该任务的排除. Mask 向量初始值设置为 $[1, 0, \dots, 0]$, 确保从输入序列的第一个任务开始执行.

2.3 LA-PN 算法

LA-PN 算法通过引入 LA 机制对 PN (指针网络) 方法进行改进. 该算法在按照式 (12) 计算上下文向量 $\mathbf{c}^t$ 时, 仅选取部分编码得到的关联信息向量序列进行计算, 选取方式为在关联信息矩阵内确定一个对应位置 $\mathbf{p}_t$, 本文 $\mathbf{p}_t$ 为上一解码步选择的观测任务 $\mathbf{y}_{t-1}$ 在输入序列中的序号, 然后选择该位置前后步长 D 区域, 即 $[\mathbf{p}_t - D, \mathbf{p}_t + D]$ 内的关联信息向量, 其中步长 D 为超参数, 可人工设置. 这种方式可以减少计算复杂度, 并提高算法收敛到最优策略的概率, 使其更适应不同长度的输入序列.

相比于 PN 方法和 MHA-PN (多头注意力-指针网络) 方法, 这种 LA 机制能够有效忽略长序列中距当前解码步较远的任务信息, 因为这些信息对当前选择任务的影响较小. 同时, 动态调整关注的序列范围也能够提高算法的适应性.

2.4 算法流程

在每个解码时间步 t 根据 LA-PN 算法模型计算得到输出节点 $\mathbf{y}_t$ 后, 根据 $\mathbf{y}_t$ 资源占用情况和星上资源限制, 对每个任务的动态元素和 Mask 向量进行

实时调整^[8]. 动态元素和 Mask 向量更新算法伪代码如下.

算法1 动态元素和Mask向量更新算法

输入: 静态元素 $\mathbf{s}$, 动态元素 $\mathbf{d}^{t-1}$, 输出节点 $\mathbf{y}_t$

输出: 更新后的动态元素 $\mathbf{d}^t$, $\mathbf{F}_{\text{Mask}}$

```

1: for  $i = 1, 2, \dots, M$  do
2:   update  $\text{win}_i^t$  according to  $w_{e_{y_t}}$ 
3:   update  $\text{acc}_i^t$  according to  $\mathbf{y}_t$ 
4:   calculate  $\text{pos}_i^t, \text{num}_i^t, \text{adj}_i^t$  by Eq.(1) and Eq.(2)
5: end for
6: initialize  $\mathbf{F}_{\text{Mask}}$  with  $[1, 1, 1, \dots, 1]$ 
7: for  $i = 1, 2, \dots, M$  do
8:   if  $\text{win}_i^t == 0 \parallel \text{acc}_i^t == 1$ 
9:     Mask  $[i] = 0$ 
10:  end if
11: update  $\mathbf{F}_{\text{Mask}}^i$  according to  $\text{pos}_i^t, \text{num}_i^t, \text{adj}_i^t$ 
12: end for
  
```

按照以上过程逐步解码, 直至 Mask 向量全部为 0, 即所有任务均已被规划, 或没有任务满足时间窗口要求. 整个算法的伪代码如下.

算法2 LA-PN算法

输入: 候选任务序列 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$

输出: 规划得到的任务序列 $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$

```

1: initialize  $\mathbf{F}_{\text{Mask}}$  with  $[1, 0, 0, \dots, 0]$ 
2: while not ( $\mathbf{F}_{\text{Mask}}$  all is 0) do
3:   calculate  $P(\mathbf{y}^{t+1} | \mathbf{Y}^t, \mathbf{X}^t)$  by Eq.(12)
4:   choose  $\mathbf{y}_{t+1}$  according to  $P(\mathbf{y}^{t+1} | \mathbf{Y}^t, \mathbf{X}^t)$ 
5:   update dynamic according to  $\mathbf{y}_{t+1}$ 
6:   update  $\mathbf{F}_{\text{Mask}}$  according to dynamic
7: end while
  
```

2.5 模型的训练

指针神经网络算法是依赖大量训练数据的算法, 而组合优化问题通常为 NP 难题, 难以获取监督学习的标签数据 (最优解). 因此, 本文采用深度确定性策略梯度算法 (DDPG) 实现指针神经网络的参数学习, 其主要原理是 Actor 网络根据 Critic 网络的输出, 通过迭代自主优化每个状态下选择的动作, 以最大化最

终收益^[14]. 结合前面提到的 LA-PN 算法模型, 可进一步进行模型训练的具体定义.

2.5.1 状态编码

状态编码为 2.1 节中的 Input 部分和 Mask 向量. 其中, Mask 向量整合约束条件, 表示每个状态下每个任务是否可以作为下一个观测任务.

2.5.2 动作编码

根据 1.3 节符号定义和 2.1 节算法模型结构, 从 Input 部分选取当前状态下的下一个观测任务 $\mathbf{y}_{t+1}$ 为动作编码, 其通过 Actor 网络实现.

Actor 网络即图 2 给出的 LA-PN 算法模型结构, 其基于状态进行动作选择形成方案.

Actor 目标网络参数为 θ' , 用于提供下一状态的策略 $\pi_{\theta'}(\mathbf{X}^{t+1})$, Actor 训练网络参数为 θ , 提供当前状态的策略 $\pi_{\theta}(\mathbf{X}^t)$, Actor 网络在参数更新时的策略梯度为

$$\nabla_{\pi_{\theta}} J(\theta) = E \{ \nabla_{\theta} [(T^t - Q_{\varphi}(\mathbf{X}^t, \mathbf{y}_{t+1})) \pi_{\theta}(\mathbf{X}^t)] \}. \quad (15)$$

其中, $\pi_{\theta}(\mathbf{X}^t)$ 为根据式 (14) 选择出 $\mathbf{y}_{t+1}$ 并计算 $\lg P(\mathbf{y}_{t+1} | \mathbf{Y}^t, \mathbf{X}^t)$, T^t 和 $Q_{\varphi}(\mathbf{X}^t, \mathbf{y}_{t+1})$ 分别为 Critic 目标网络和 Critic 训练网络输出值.

2.5.3 奖励函数

使用 DDPG 算法的最终目的是确定每个状态下的最优动作. 在本文中是在已选择的任务状态下, 找出下一个需要观测的任务. 在学习过程中, 算法在选出动作后, 需要一个奖惩值来引导其朝目标方向优化. 在本文中 Actor 网络输出下一个待观测任务后的奖励函数按照式 (8) 进行计算.

根据经验设计的 Critic 网络由三部分组成, 分别为与图 2 相同的 Encoder、一个 LSTM 处理过程和一个由两层 ReLU 函数组成的解码器, 负责对 Actor 网络生成的动作进行评价得到 Q 值.

Critic 目标网络参数为 φ' , 用于近似估计下一时刻状态-动作的 Q 值函数 $Q_{\varphi'}(\mathbf{X}^{t+1}, \pi_{\theta'}(\mathbf{X}^{t+1}))$. 于是可以得到当前时刻状态-动作的 Q 值函数的目标值为

$$T^t = R^t + \gamma Q_{\varphi'}(\mathbf{X}^{t+1}, \pi_{\theta'}(\mathbf{X}^{t+1})). \quad (16)$$

Critic 训练网络参数为 ϕ , 输出当前时刻状态-动作的 Q 值函数 $Q_{\varphi}(\mathbf{X}^t, \mathbf{y}_{t+1})$. Critic 网络损失函数为

$$L_{\text{loss}} = E [T^t - Q_{\varphi}(\mathbf{X}^t, \mathbf{y}_{t+1})]^2. \quad (17)$$

训练部分的伪代码如下。

算法3 DDPG算法

输入: 训练数据集
输出: LA-PN算法模型参数

```
1: initialize actor network with random weights  $\theta$ 
2: initialize critic network with random weights  $\varphi$ 
3:  $\theta' \leftarrow \theta, \varphi' \leftarrow \varphi$ 
4: for iteration = 1, 2, ..., max_iteration do
5:   reset gradients :  $d\theta \leftarrow 0, d\varphi \leftarrow 0$ 
6:   for instance = 1, 2, ...,  $I$  do
7:     while not ( $F_{\text{Mask}}$  all is 0) do
8:       choose  $y_{t+1}$  according to  $\pi_{\theta}(X^t)$ 
9:       compute  $R^t$  by Eq. (6)
10:      compute  $X^{t+1}$  by Mask
11:      push( $X^t, y_{t+1}, R^t, X^{t+1}$ ) into buffer
12:    end while
13:  end for
14: for episode = 1, 2, ...,  $E$  do
15:   sample  $N$  instances from replay buffer
16:    $d\theta \leftarrow$  compute  $\nabla_{\pi_{\theta}} J(\theta)$  by Eq. (13)
17:    $d\varphi \leftarrow$  compute loss by Eq. (15)
18:   update  $\theta$  using  $d\theta$ 
19:   update  $\varphi$  using  $d\varphi$ 
20:  $\theta' \leftarrow \tau\theta + (1 - \tau)\theta', \varphi' \leftarrow \tau\varphi + (1 - \tau)\varphi'$ 
21: end for
22: end for
```

3 实验验证

3.1 模型的训练

3.1.1 实验环境与数据

采用 LA-PN 算法模型对高频动态到达的天文卫星 ToO 和后随观测任务规划问题进行求解。使用 Pytorch 框架在 Matrox G200 eR2 GPU 上对 LA-PN 算法模型进行训练。在训练过程中, 根据经验将滑动窗口参数分别设定为 3, 5, 7, 10, 15。实验结果表明, 窗口参数为 5 时效果最佳, 其余超参数根据经验设定, 具体数值详见表 2。

生成 8000 个用于训练的调度周期, 每个调度周期包含固定数量随机产生的观测任务。观测任务的静态属性取值范围如表 3 所示, 其中 $[a, b]$ 表示对应元素数值随机产生, 满足 a 到 b 之间的均匀分布, 并根

表 2 LA-PN 算法训练参数设定

Table 2 LA-PN algorithm training parameter settings

参数名	参数值	含义
Batch_size	128	批样本数量
D	5	滑动窗口大小
lr_actor	1×10^{-5}	Actor网络的学习率
lr_critic	5×10^{-4}	Critic网络的学习率
hidden_size	512	隐含层维度

表 3 数据集中每个任务的元素设定

Table 3 Element settings for each task in dataset

元素	设定
T^{ws}	[0 s, 4000 s]
D	[600 s, 700 s]
T^{we}	$[T^{\text{ws}} + 700 \text{ s}, T^{\text{ws}} + 800 \text{ s}]$
A	$[-5^{\circ}, 5^{\circ}]$
P	[0.1, 0.9]

据需要进行归一化处理。

3.1.2 模型的训练

按照表 2 的模型超参数设定、表 3 的任务元素设定生成数据集, 使用算法 3 对 LA-PN 算法模型进行训练, 收敛曲线如图 3 所示。图 3(a) 中每个调度周期内的候选任务序列为 50 个, 图 3(b) 中每个调度周期内的候选任务序列为 100 个。可以看出, 在整个学习过程中, 观测总收益持续提升, 并最终趋于收敛, 这表明该算法可以学习到最大化优化目标式 (8) 的规划策略。

使用训练好的模型, 可以对长度为 100 的输入序列进行端到端的推理, 推理结果如图 4 所示。

3.2 算法对比

比较 LA-PN 算法模型与 ACO (蚁群优化)^[15]、PN 和 MHA-PN^[8] 算法在总收益、任务优先级、应急任务响应度和求解时间等指标上的表现, 通过仿真实验验证 LA-PN 算法的收益性、实时性和泛化性。

设计两种场景 A 和 B, 其中需求分布在卫星的星下点轨迹附近经纬度 $\pm 5^{\circ}$ 范围内。采用 STK11 软件对这两种场景进行了仿真分析, 根据 SVOM 卫星的载荷视角和约束条件, 计算了卫星对观测目标的过境时段。

3.2.1 收益性、实时性

在场景 A 中生成 200 个调度周期, 每个调度周期

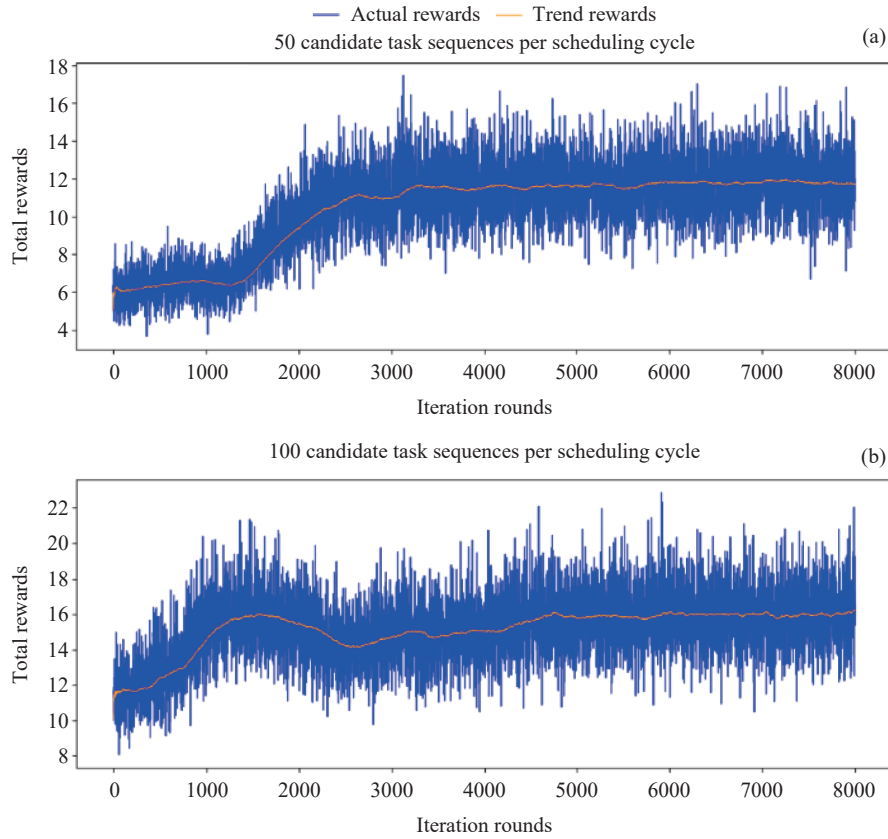


图3 LA-PN 算法模型训练总收益收敛曲线

Fig. 3 LA-PN algorithm model training total rewards

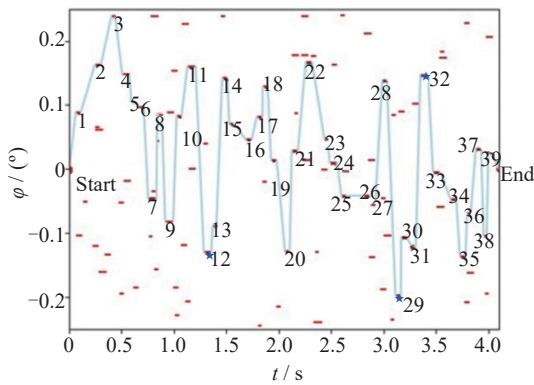


图4 模型的推理结果 (红色点表示观测任务时间窗口, 编号表示规划结果任务执行顺序, 蓝色五星表示执行任务前需要调姿)

Fig. 4 Schematic diagram of the inference results of the model (Red dots represent the time window of observation tasks, numbers represent the sequence of planned result tasks execution, and blue stars represent the need for attitude adjustment before task execution)

包含 100 个随机产生的观测需求, 主要验证算法的收益性和实时性. 图 5 给出了 LA-PN 算法与其余算法在单个调度周期内各项指标的对比, 图 5(a) 为总收益对比, 图 5(b) 为任务优先级比例对比, 其计算方式为

$$f_{\text{pri}} = \frac{\sum_{i=1}^N P_i}{\sum_{j=1}^M p_j} \quad (18)$$

其中, P_i 为任务 i 的优先级; M 为输入任务的个数; N 为规划执行任务的个数, 即规划任务总优先级占输入任务总优先级比例. 图 5(c) 给出了应急任务响应度对比, 表 4 为算法执行时间及 LA-PN 算法与其余算法总收益提高率的对比. 可以看出, LA-PN 算法相较于其他算法在多个周期内所获得的任务规划总收益和任务优先级比例更高, 且应急任务响应度是明显优于其他算法的.

3.2.2 泛化性

在场景 B 中生成 4 个调度周期, 每个调度周期待规划任务数量从 50 泛化到 200, 主要验证算法的泛化性. 如图 6 所示, 随着单个调度周期内待规划的任务

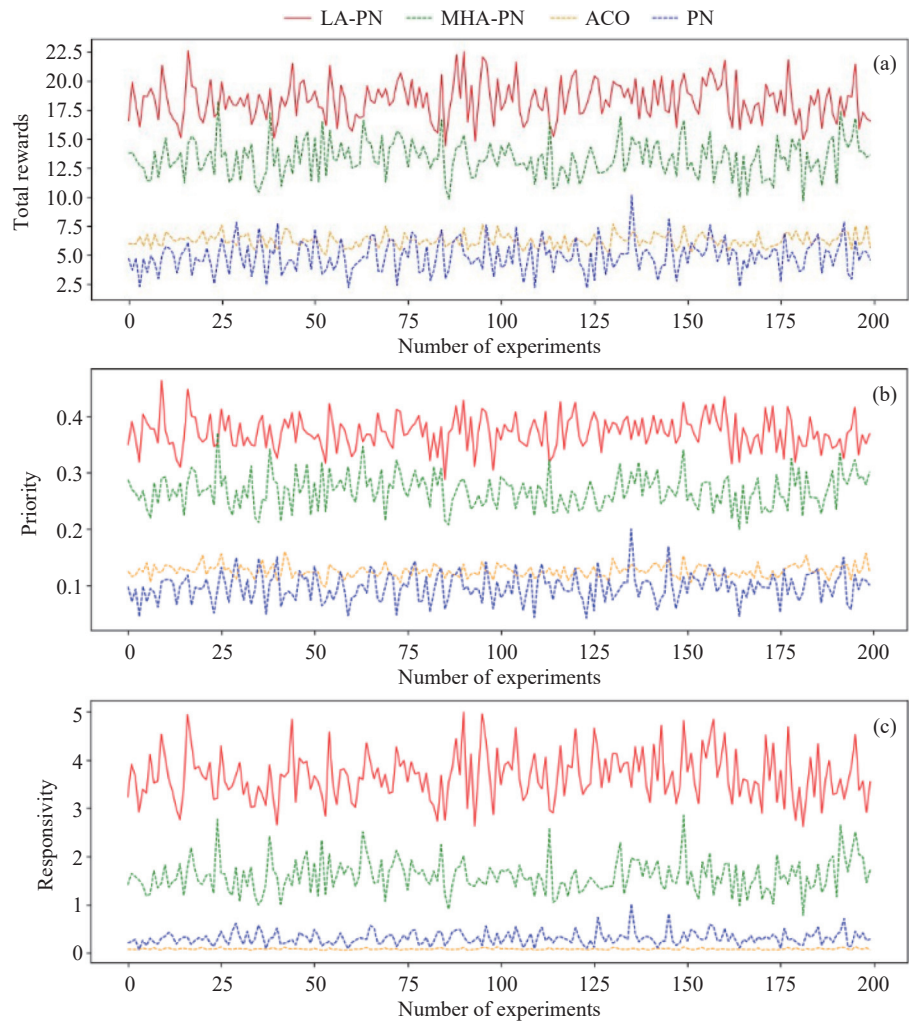


图 5 LA-PN 与其他算法各项指标对比

Fig. 5 Comparison of various indicators between LA-PN and other algorithms

表 4 周期内算法平均执行时间对比
Table 4 Comparison of average execution time of algorithms within a cycle

算法	$T/(\text{s} \cdot \text{epoch}^{-1})$	$R_{\text{rate}}/(\%)$
LA-PN	0.396318	—
PN	0.308217	273.83
MHA-PN	0.297375	37.21
ACO	13.027278	192.80

数量逐渐增多, LA-PN 算法能有效对其求解, 并且比其他算法获得的任务规划收益更高; 当任务规模超过 150 个时, 由于每轨调姿次数限制以及时间窗口约束, 总收益趋于饱和, 符合实际情况. LA-PN 算法随着任务规模的增加表现出较强的适应性和优越性.

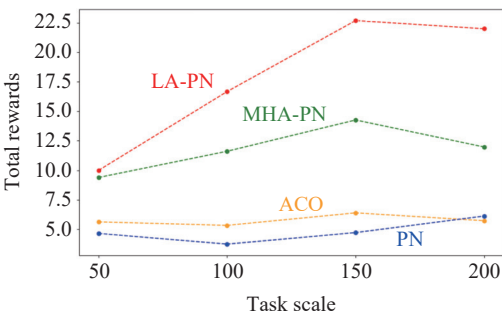


图 6 算法泛化能力对比

Fig. 6 Comparison of algorithm generalization ability

4 结论

针对高频动态到达的 ToO 和后随观测任务组成的长序列任务规划问题, 建立基于 PN 的序列决策模

型,考虑最大化科学观测收益、应急任务响应度等多目标优化,并对PN进行改进,提出LA-PN算法.该算法通过引入时间窗口的方式,使模型专注于对当前决策有重要影响的序列部分,提高了求解效率.与其他算法相比,LA-PN算法可以获得更高的总收益,在应急任务响应度指标上有显著提高.随着任务规模的增加,LA-PN算法显示出优越的泛化性.

参考文献

- [1] LI Dalin. Research on Observation Scheduling Model and Method of Astronomy Satellite[D]. Harbin: Harbin Institute of Technology, 2020 (李大林. 天文观测卫星任务规划模型与方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2020)
- [2] XU Ziling, LIU Yurong, FENG Zhun. Mission planning for astronomical satellite based on genetic algorithm under tiling coverage strategy[J]. *Chinese Journal of Space Science*, 2022, **42**(2): 321-328 (徐子羚, 刘玉荣, 冯准. 基于遗传算法的Tiling覆盖策略天文卫星任务规划[J]. *空间科学学报*, 2022, **42**(2): 321-328)
- [3] XIANG Shang, CHEN Yinguo, LI Guoliang, *et al.* Review on satellite autonomous and collaborative task scheduling planning[J]. *Acta Automatica Sinica*, 2019, **45**(2): 252-264 (向尚, 陈盈果, 李国梁, 等. 卫星自主与协同任务调度规划综述[J]. *自动化学报*, 2019, **45**(2): 252-264)
- [4] WU Haiyan, MENG Xin, ZHANG Yuzhu, *et al.* Research on the planning method for astronomy observation mission[J]. *Chinese Journal of Space Science*, 2013, **33**(5): 561-568 (吴海燕, 孟新, 张玉珠, 等. 面向天文观测的空间科学卫星任务规划方法研究[J]. *空间科学学报*, 2013, **33**(5): 561-568)
- [5] HUANG Yue, QU Jinlu, JIA Shumei, *et al.* Long-term planning algorithm for the HXMT mission[J]. *Chinese Journal of Space Science*, 2017, **37**(6): 766-772 (黄跃, 屈进禄, 贾淑梅, 等. HXMT卫星长期任务规划算法[J]. *空间科学学报*, 2017, **37**(6): 766-772)
- [6] YIN Xiaodan, BAI Meng, LI Zhuoheng. Clustering-scheduling methods for oversubscribed short-term tasks of astronomical satellites[J]. *Transactions of Nanjing University of Aeronautics and Astronautics*, 2023, **40**(3): 307-322 (尹晓丹, 白萌, 李卓恒. 超额订购下天文卫星短期任务的聚类规划方法[J]. *南京航空航天大学学报(英文版)*, 2023, **40**(3): 307-322)
- [7] WANG Haijiao. Massive Scheduling Method Under Online Situation for Satellites Based on Reinforcement Learning[D]. Beijing: National Space Science Center, Chinese Academy of Sciences, 2018 (王海蛟. 基于强化学习的卫星规模化在线调度方法研究[D]. 北京: 中国科学院大学(中国科学院国家空间科学中心), 2018)
- [8] MA Yifan. Agile Imaging Satellite Task Planning Method Based on Deep Reinforcement Learning[D]. Hangzhou: Zhejiang University, 2021 (马一凡. 基于深度强化学习的敏捷成像卫星任务规划方法[D]. 杭州: 浙江大学, 2021)
- [9] ZHAO Shijie. Research and Implementation of Satellite Autonomous Task Planning Method Based on Reinforcement Learning[D]. Shanghai: East China Normal University, 2024 (赵世杰. 基于强化学习的卫星自主任务规划方法研究与实现[D]. 上海: 华东师范大学, 2024)
- [10] LIU Xiaolu, XU Yingjie, HE Renjie, *et al.* Satellite-oriented on-orbit service mission planning method[J]. *Journal of National University of Defense Technology*, 2020, **42**(5): 143-150 (刘晓路, 许英杰, 贺仁杰, 等. 面向卫星的在轨服务任务规划方法[J]. *国防科技大学学报*, 2020, **42**(5): 143-150)
- [11] ZENG Erkang, WU Haiyan, FENG Zhun. Sky area target of opportunity mission planning method[J]. *Chinese Journal of Space Science*, 2023, **43**(2): 361-368 (曾而康, 吴海燕, 冯准. 天区机遇目标空间探测任务规划算法[J]. *空间科学学报*, 2023, **43**(2): 361-368)
- [12] CHUNG J, GULCEHRE C, CHO K, *et al.* Empirical evaluation of gated recurrent neural networks on sequence modeling[OL]. arXiv preprint arXiv: 1412.3555, 2014
- [13] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015
- [14] BELLO I, PHAM H, LE Q V, *et al.* Neural combinatorial optimization with reinforcement learning[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview. net, 2017
- [15] ZHAO Fanyu. Mission Scheduling and Planning of Multi-Target Observations for Spacecraft[D]. Beijing: Beijing Institute of Technology, 2015 (赵凡宇. 航天器多目标观测任务调度与规划方法研究[D]. 北京: 北京理工大学, 2015)

作者简介



王旭航 女, 1999年2月出生于河北省石家庄市, 现为中国科学院国家空间科学中心研究生, 主要从事卫星智能运控等方面的研究。
E-mail: wangxuhang22@mails.ucas.ac.cn



吴海燕(通信作者) 女, 1970年9月出生于江苏省南京市, 硕士研究生导师, 现为中国科学院国家空间科学中心正高级工程师, 主要从事卫星地面应用系统、卫星智能运控技术等方面的研究。
E-mail: why@nssc.ac.cn