深空探测器多智能体强化学习自主任务规划

孙泽翼[1]，王 彬[1,2]，胡馨月[1]，熊 新[1,2]，金怀平[1,2]

（1. 昆明理工大学 信息工程与自动化学院，昆明 650500；2. 云南省人工智能重点实验室，昆明 650500）

摘 要：针对深空探测器执行附着任务时各子系统协同规划自主性、快速性和自适应性的要求，提出一种基于近端策略优化方法的多智能体强化学习协同规划，将单智能体近端策略优化算法与多智能体混合式协作机制相融合，设计了一种多智能体自主任务规划模型，并引入噪声正则化优势值解决多智能体集中训练中协同策略过拟合的问题。仿真结果表明，多智能体强化学习自主任务规划方法能根据实时环境变化，对智能自主优化小天体附着任务的协作策略适时调整，与改进前的算法相比提高了任务规划成功率和规划解的质量，缩短了任务规划的时间。

关键词：多智能体强化学习；深空探测自主任务规划；近端策略优化；小天体附着

中图分类号：TP18 **文献标识码：**A **文章编号：**2096-9287(2024)03-0244-12

DOI:10.15982/j.issn.2096-9287.2024.20230159

引用格式：孙泽翼，王彬，胡馨月，等. 深空探测器多智能体强化学习自主任务规划[J]. 深空探测学报（中英文），2024，11（3）：244-255.

Reference format: SUN Z Y，WANG B，HU X Y，et al. Multi-agent reinforcement learning autonomous task planning for deep space probes[J]. Journal of Deep Space Exploration，2024，11（3）：244-255.

引 言

小天体探测在揭示生命起源、保护地球安全及开发天然资源等方面具有深远的意义，但小天体探测任务周期长、目标远且存在极大的不确定性。与月球、火星等地外天体着陆探测的任务不同，小天体本身引力较弱，且相对运动速度较慢，因此在执行小天体表面附着任务时深空探测器的导航制导控制系统面临着更大的挑战。其中任务规划系统由于小天体状况差异较大，可参考的先验经验较少，且深空环境多变，传统任务规划方法对小天体附着过程的适应性较差[1]。

在小天体附着过程中，需要探测器的多个子系统协同完成各项任务，不同子系统在各自不同的子任务目标下还需遵守协作规则，满足时序约束，同时兼顾资源的共享机制。针对这一问题，近年来很多机构和学者利用人工智能算法解决深空探测器任务规划问题展开了研究。姜啸等[2]将智能规划理论与约束可满足技术条件相结合，对多约束要求下深空探测器自主任务规划技术进行了深入研究。赵宇庭等[3]提出基于分布式求精搜索的多智能体规划空间规划方法，设计了动态智能体交互图引导多智能体协同规划，将时间资源约束处理抽象为约束满足问题并采用基于图论的方法进行处理。史谦郡等[4]针对空间站运营短期任务规划中的复杂约束满足问题，结合深度强化学习技术，提出基于深度确定性策略梯度算法的空间站复杂约束处理方法。柳景兴等[5]针对深空探测器任务规划多子系统协同机制多约束的问题，提出深空探测任务规划认知图谱构架及多属性约束冲突检测方法。毛维杨等[6]针对深空探测器自主任务规划多约束的需求，提出了基于动态奖励的强化学习深空探测器任务自主规划模型构建方法。

上述研究提升了深空探测器自主任务规划的能力，但由于探测器各子系统的并行分布特性和多约束属性，现有方法的协作机制和策略优化方法仍然难以满足探测器任务规划系统自主性、快速性和自适应性的需求。

针对上述研究现状，本文提出一种基于近端策略优化算法的深空探测器多智能体强化学习自主任务规划方法，首先给出了深空探测器多智能体自主任务规划问题的模型，并根据深空探测任务规划状态转移约束、资源约束及时间约束等构建深空探测器多智能体交互知识环境，将单智能体近端策略优化算法与多智能体混合式协作机制相结合，设计了一种多智能体强化学习自主任务规划模型，并在此基础上引入噪声正

收稿日期：2023-11-03 修回日期：2024-01-25

基金项目：空间碎片专项（KJSP2020020302）

则化优势值解决多智能体训练中的策略过拟合问题。

1　深空探测器自主任务规划多智能体系统模型

1.1　深空探测器任务规划问题描述

深空探测器是由多个子系统组成的整体，因此深空探测器自主任务规划问题是一个多约束、高冲突的复杂组合优化问题，其目标是在满足任务、资源和时间约束的前提下，根据任务目标自主规划动作序列及执行时间，以求达到资源消耗最少，时间消耗最少的目标或使多个目标函数综合达到最优。

根据该问题的特点，结合多智能体理论[7]，将深空探测器的子系统视作单独的智能体，深空探测器任务规划问题可描述为在给定初始状态、目标状态和知识域的条件下，自主生成满足时序、时间和资源约束的各个子系统动作序列。多智能体深空探测器任务规划问题描述为

$$P = (I, G, A, A_{\text{seq}}) \tag{1}$$

$$C = \{c_1, c_2, \cdots c_i, \cdots c_m\} \tag{2}$$

其中：I为任务的初始状态；G为任务的目标状态；A为探测器中可执行的操作集合；A_{seq}为任务规划的动作序列；C为含有m个约束的约束集，约束集包括知识域中动作的时序约束、时间要求和资源消耗等，其中c_i代表C的一个约束条件。任务规划的目标是由根据给定的初始状态I以及约束集C，生成一组有序任务规划的动作序列A_{seq}，以达到期望的目标状态G。

1.2　深空探测器自主任务规划多智能体系统建模

在深空探测器多智能体系统中，每个子系统智能体的决策过程可建模为一个马尔可夫序贯决策过程。单智能体的状态空间不仅包括自身的状态，还包括其它所有深空探测器子系统的状态。单智能体的动作空间是该智能体对应的深空探测器子系统可执行的全部动作集合，动作空间和状态空间均为离散空间。

深空探测器可表示为n个能力异构的子系统集合，其中每个子系统都具有不同的状态、动作空间。每个子系统的马尔科夫决策过程（Markov decision process，MDP）模型可用5元组表示，深空探测器的决策模型可表示为

$$M = (S, A, T, R, \gamma) \tag{3}$$

$$S = \{s^1, s^2, \cdots, s^i, s^n\} \tag{4}$$

$$A = \left\{a^1, a^2, \cdots, a^i, a^n\right\} \tag{5}$$

$$T = P\left(s_{t+1}^i | s_t^i, a_t^i\right) \tag{6}$$

其中：i为深空探测器子系统的编号；S描述探测器的整体状态，其中s^i为第i个子系统的状态信息；A为探测器的整体动作，a_i为第i个子系统可以采取的动作，$a^i \in A$；T为第i个子系统在t时刻状态s_t转移到s_{t+1}的状态转移函数，其中s_t^i为第i个子系统在t时刻的状态；P为第i个子系统在t时刻状态下执行a_t^i动作转移到s_{t+1}^i的概率，根据状态转移函数P输出下一个状态；R为强化学习环境给予的奖惩值；γ为折扣因子，其中$\gamma \in [0,1]$。

深空探测器多智能体任务规划采用完全合作型的策略，每个智能体子系统发出动作获得的奖励会受到其它智能体子系统动作的影响。每个智能体的策略可表示为

$$\pi^i = (a^i | s; \theta^i) \tag{7}$$

其中：θ^i为第i个智能体中策略网络的参数，深空探测器多智能体的联合策略可表示为

$$\pi = \{\pi^1, \pi^2, \cdots, \pi^i, \cdots, \pi^n\} \tag{8}$$

深空探测器多智能体系统的目标是找到一种联合策略使得整个系统的期望收益最大化，从而达到深空探测器任务规划整体目标的均衡稳态。

1.3　约束条件

与一般的规划问题不同，深空探测任务规划问题中存在着多种约束[1]，本文方法在规划过程中主要考虑以下约束。

1）时序约束：各子系统动作需要在一定的时序条件下执行。对于深空探测器任务规划存在时序约束问题定义第i个子系统的第u个动作$a^{i,u}$的前序动作集合为$\boldsymbol{PA}(a^{i,u})$，前序动作集合包含了该动作被执行以前必须先执行的动作，表示为

$$\boldsymbol{PA}(a^{i,u}) = \left\{a^{i,u1}, a^{i,u2}, \cdots, a^{i,u\left|\boldsymbol{PA}(a^{i,u})\right|}\right\} \tag{9}$$

2）时间约束：每个动作的执行不是瞬时完成，都需要一定的执行时间。深空探测器实际工作中动作切换同样需要一定的时间，这类事件与实际系统相关且多为固定时间，因此本文方法只考虑前者，暂时不考虑切换时间。定义$\text{time}(s^i, a^i)$为执行动作所消耗的时间，所有动作的时间$\sum \text{time}(s^i, a^i)$消耗要尽可能地少，同时不能超过整个系统的最大规划时间上限$t_{\max}$。

3）资源约束：执行每个动作需要消耗一定的资源。本文主要考虑存储空间、电量、燃料的消耗，每种资源的最大值是一定的，任何时刻消耗的资源总和不能超过资源最大值，存储空间、电量、燃料的最大

值分别为$C_{\max}$、$e_{\max}$、$f_{\max}$。

综合以上时序、时间和资源约束，对于这一多目标优化问题，可以用数学模型描述为

$$\begin{cases} f(x)=\begin{cases} \min \sum \text{time}(s^i,a^i) \\ \min \sum res_{\text{electricity}}(s^i,a^i) \\ \min \sum res_{\text{capacity}}(s^i,a^i) \\ \min \sum res_{\text{fuel}}(s^i,a^i) \end{cases} \\ \text{s.t.}\begin{cases} \sum \text{time}(s^i,a^i)<t_{\max} \\ \sum res_{\text{electricity}}(s^i,a^i)<e_{\max} \\ \sum res_{\text{capacity}}(s^i,a^i)<c_{\max} \\ \sum res_{\text{fuel}}(s^i,a^i)<f_{\max} \end{cases} \end{cases} \tag{10}$$

$f(x)$的目的是最小化规划序列消耗的资源和时间。其中：$res_{\text{capacity}}(s^i,a^i)$智能体在$s_i$状态下执行$a_i$动作所消耗的容量，$res_{\text{electricity}}(s^i,a^i)$为智能体在$s^i$状态下执行$a^i$动作所消耗的电量，$res_{\text{fuel}}(s^i,a^i)$为智能体在$s^i$状态下执行$a^i$动作所消耗的燃料。

2 基于多智能体强化学习的深空探测器任务规划方法

基于第一节所述的数学建模问题，本文提出一种多智能体强化学习深空探测器自主任务规划方法，该方法将深空探测器每一个子系统看作一个智能体，为了在多智能体系统中实现高效协作，完成复杂任务。拟将强化学习的自适应能力与任务规划的高层次决策相结合。基于多智能体强化学习的深空探测器自主任务规划框架如图1所示。

基于多智能体强化学习的深空探测器自主任务规划框架主要包括深空探测器多智能体交互环境、训练部分和应用部分3大部分。深空探测器多智能体交互环境包括深空探测器任务规划知识。训练模型部分的目的是寻找一个最优策略，用于生成深空探测器各个子系统的最优动作，该部分采用混合式训练方法，即集中式训练，分布式执行[7]。每个智能体基于局部观测和策略去生成一个动作来最大化折扣累计奖励，所有智能体基于全局状态学习中心价值函数。应用部分针对规划任务（通常包括事先存储的任务，地面临时上传的任务，运行过程中产生的临时任务，出现故障时要求重规划的任务）选用训练好的最优策略进行规划，生成规划指令并进行规划结果验证，如果结果合理，则输出规划序列传输给深空探测器系统，如果不符合时序约束或者存在资源冲突，则进行重新规划。

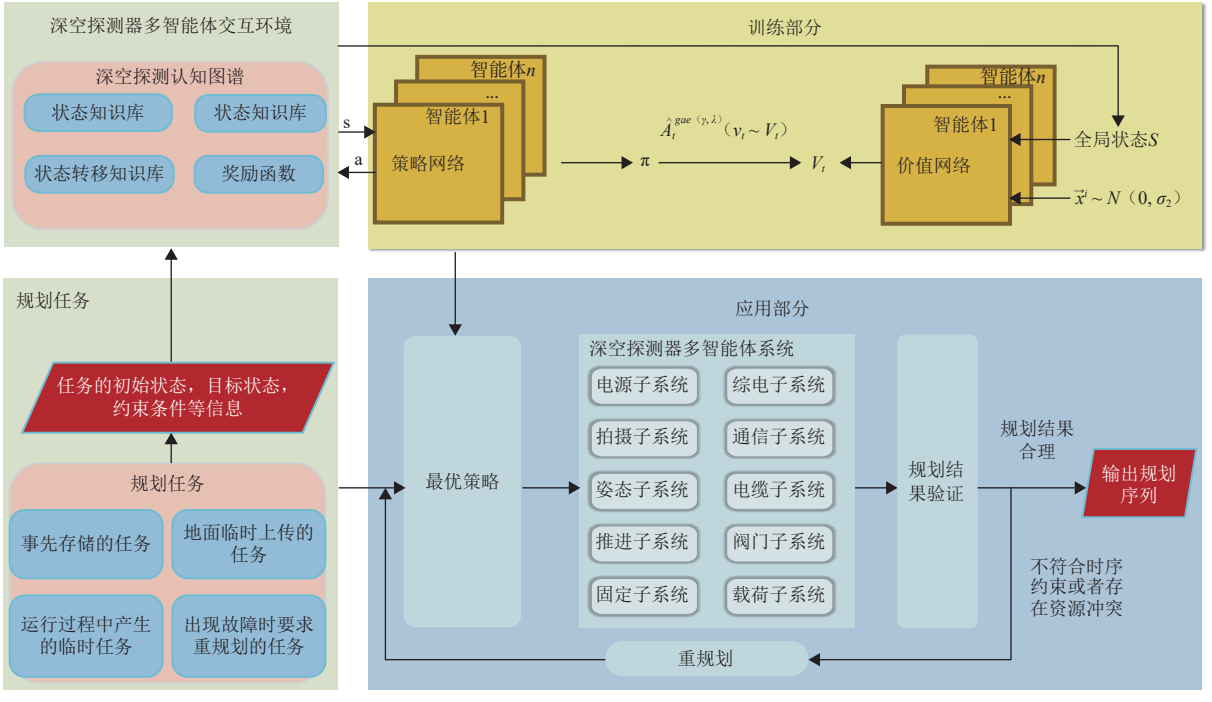

图 1 基于多智能体强化学习的深空探测器自主任务规划方法框架

Fig. 1 Framework for autonomous task planning of deep space probes based on multi-agent reinforcement learning

2.1　深空探测器多智能体强化学习环境

本文方法将深空探测器的每一个子系统看作一个智能体，多智能体强化学习环境是指所有智能体在训练和执行任务规划过程中交互所需用到的知识。包括状态空间知识库、动作空间知识库、实现时序约束状态转移空间知识库及融合资源和时间约束的奖励函数。

2.1.1　构建状态空间知识库

为实现成功的协调，智能体需要自身和其它智能体及其相互关系的信息，在深空探测器多智能体协同任务中，每个智能体不但要掌握自身的状态还要掌握其它智能体的状态，才能选取对整个系统更有利的动作。

深空探测器状态空间包含深空探测器中各子系统可达的状态。根据式（3）～（6）所构建的模型，智能体可观测的环境为探测器的整体状态，第i个子系统的状态集用$\boldsymbol{S}^i$表示，其状态有$|\boldsymbol{S}^i|$种，可表示为

$$\boldsymbol{S}^i=\left\{s^{i,1},s^{i,2},\cdots,s^{i,j},\cdots,s^{i,|S^i|}\right\} \tag{11}$$

其中：$s^{i,j}$为第i个子系统的第j个状态。

2.1.2　构建动作空间知识库

深空探测器动作空间包含深空探测器各子系统可执行的动作。多智能体交互环境中智能体的动作是将深空探测器各动作抽象化后的离散动作。每个动作的执行都会造成对应状态的变化，第i个子系统的可执行动作集用$\boldsymbol{A}^i$表示，其种类有$|\boldsymbol{A}^i|$种，其中$1\leqslant i\leqslant N$。$\boldsymbol{A}^i$可表示为

$$\boldsymbol{A}^i=\left\{a^{i,1},a^{i,2},\cdots,a^{i,u},\cdots,a^{i,|A^i|}\right\} \tag{12}$$

其中：$a^{i,u}$为第i个子系统的第u个可执行动作，每个动作执行都会消耗一定的资源以及时间res_{time}，资源包括容量res_{capacity}、电量$res_{\text{electricity}}$和燃料res_{fuel}。

2.1.3　构建状态转移空间知识库

深空探测器状态转移空间库包含深空探测器各子系统从当前状态执行某可执行动作后能到达的下一个状态的所有状态转移关系。

本文构建的三元组为$G=(s^{i,j},a^{i,u},s^{i',j'})$，其中$G$为深空探测器状态与动作的状态转换空间，$s^{i,j}$为深空探测器第$i$个子系统的第$j$个可达状态，$a^{i,u}$为深空探测器定义第$i$个子系统的第$u$个动作，$s^{i',j'}$为深空探测器第$i'$个子系统的第$j'$个可达状态。

2.1.4　设计奖励函数

深空探测器多智能体强化学习环境中的奖励函数定义为r，设计时融合了1.3节的约束条件，以便更好地向各智能体传达深空探测器任务规划的目标，同时也为各个智能体学习如何进行任务规划提供引导。深空探测器多智能体强化学习的总奖励函数R定义为每个智能体单独奖励值r_i之和的平均值，奖励值为所有智能体奖励值的集合，定义为

$$R=\frac{1}{n}\sum_{i=1}^{n}r^i \tag{13}$$

$$r^i=r_p+r_n \tag{14}$$

奖励包括正向奖励r_p和负向奖励r_n。为了引导多智能体强化学习在进行深空探测器自主任务规划的过程中保证任务规划的完整性，构造正向奖励r_p。对任务中必须要执行的关键动作集合定义为A_{nes}，且需要满足$A_{\text{nes}}\subseteq A_{\text{seq}}$。即所有需要执行的关键动作$A_{\text{nes}}$都应被包含有序任务规划的动作序列$A_{\text{seq}}$中，如果深空探测器达到目标状态且$A_{\text{seq}}$包含关键动作则给予正向奖励值$r_p$，涵盖的越多则给予的奖励值$r_p$越大。

但目标导向的离散性奖励设置难以区分单个智能体对系统性能的贡献，这可能会导致规划得到次优解。为了使深空探测器规划的任务序列为最优的动作序列，在深空探测器子系统的每步动作都给予一个负向奖励r_n。其定义为

$$\begin{aligned}r_n=&\mu_c res_{\text{capacity}}(s_t,a_t)+\mu_e res_{\text{electricity}}(s_t,a_t)+\\&\mu_f res_{\text{fuel}}(s_t,a_t)+\mu_t res_{\text{time}}(s_t,a_t)\end{aligned} \tag{15}$$

其中：μ_c、μ_e、μ_f、μ_t为负向奖励权重，一般根据具体任务要求计算赋值。

2.2　基于近端策略优化算法的多智能体学习框架

本文在多智能体近端策略优化算法（Multi Agent Proximal Policy Optimization，MAPPO）[7]的基础上采用Actor-Critic架构构建深空探测器小天体附着任务规划强化学习框架，如图2所示。

MAPPO 算法是将PPO 应用于多智能体任务的变种，其不同之处在于，在策略网络部分，为进一步降低优势函数的方差，由广义优势估计函数（Generalized Advantage Estimation，GAE）[8]代替，对优势函数进行估计，对其方差和偏差达到平衡，能够在一定偏差的情况下显著地降低评估的方差。深空探测器各个子系统作为单智能体，每个智能体有自己的策略网络（Actor）和价值网络（Critic），每个智能体和环境交互，得到经验数据存储到经验池中，从经验池中抽取小批次样本。根据全局状态用GAE方法计算全局的优势值。之后根据引导更新策略网络和价值网络，它们通过价值网络中心价值函数V共同学习一个联合策略π以实现协同规划，最终生成一个完整的动作序列集，并使得整个系统的期望收益最大化。

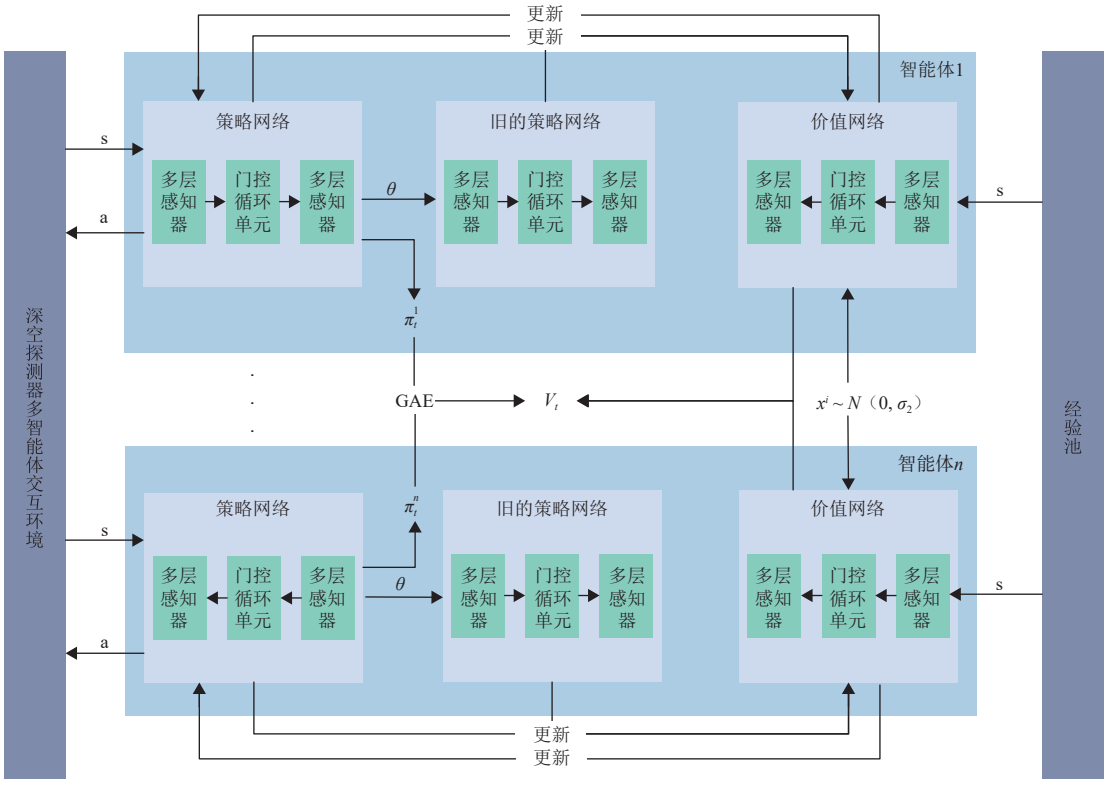

图 2 引入噪声正则化优势值的多智能体近端策略优化算法

Fig. 2 Multi-agent proximal policy optimization algorithm with noisy regularized advantage values

整个系统的期望收益定义为

$$J^{\pi}(s_0) = E_{a^i \sim \pi^i}\left[\sum_{t=0}^{\infty} \gamma_t r_t^i(S_t, A_t)\right] \tag{16}$$

其中：$J^{\pi}(s_0)$为从初始状态s_0开始在联合策略Π下，整个系统的期望收益；$a^i \sim \pi^i$为第i个智能体策略Π所映射的动作a^i；γ_t为t时刻的折扣系数，$\gamma \in [0,1]$；$r_t(S_t, A_t)$为t时刻各智能体子系统当前状态下采取动作$[a_t^1, \cdots, a_t^n]$所获得的奖励值r_t。

深空探测器小天体附着任务规划强化学习框架通过优化策略网络来引导子系统智能体做出更优的动作获取更高的奖励。算法采用重要性采样，利用新旧策略网络与环境互动，并用收集到的深空探测器数据训练策略网络的参数，同时为防止新旧网络参数差距过大，采用clip函数限制，每个智能体的策略网络目标函数$L^{\mathrm{CLIP}}(\theta_i)$为

$$L^{\mathrm{CLIP}}(\theta_i) = \hat{E}_t[\min(r_t(\theta_i)\hat{A}_t^{gae(\gamma,\lambda)}, clip(r_t(\theta_i), 1-\varepsilon, 1+\varepsilon)\hat{A}_t^{gae(\gamma,\lambda)}] \tag{17}$$

$$\hat{A}_t^{gae(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V \tag{18}$$

其中：θ_i为第i个智能体中策略网络的参数；$clip(r_t(\theta_i), 1-\varepsilon, 1+\varepsilon)$为当$r_t(\theta_i)$的比值大于$1+\varepsilon$则取为$1+\varepsilon$，小于$1-\varepsilon$则取为$1-\varepsilon$，$r_t(\theta_i)$为新旧策略网络的比值。策略网络中的优势函数$\hat{A}_t^{gae(\gamma,\lambda)}$采用GAE方法计算，其中$\lambda$是平衡均差和方差的参数，$\delta_{t+l}^V$是优势函数的无偏估计。

每个智能体的价值网络用于估计深空探测器子系统当前全局状态下执行某个动作的价值$V(s_t)$。价值网络的目标函数定义为

$$L(\phi) = \frac{1}{Bn}\sum_{i=1}^{B}\sum_{k=1}^{n} \max\left[(V_f(s_i^{(k)}) - \hat{R}_i)^2, clip(V_f(s_i^{(k)}), V_{f_{\mathrm{old}}}(s_i^{(k)}) - \varepsilon, V_{f_{\mathrm{old}}}(s_i^{(k)}) + \varepsilon) - \hat{R}_i^2\right] \tag{19}$$

其中：B为数据的批量大小；n为智能体的数量；$V_{\phi}(s_i^{(k)})$ $s_i^{(k)}$的状态价值函数；$V_{\phi_{\mathrm{old}}}(s_i^{(k)})$为旧策略网络的$s_i^{(k)}$的状态价值函数；$\hat{R}_i$为乘以折扣系数后的奖励。

价值网络采用集中式价值函数，兼顾了全部智能体信息，因此使得MAPPO可以更好地解决多智能体协作问题。但集中式函数导致价值函数无法对每个智能体对整体的贡献做出评价，难以判断是哪个智能体的哪个动作导致优势值$\hat{A}_t^{gae(\gamma,\lambda)}$提高。针对该问题，本文引入噪声正则化优势值进行缓解。

2.3 针对策略过拟合问题的噪声正则化优势值

集中式函数的优点在于可以更好地掌握全局状态

促进智能体之间的合作，但是也给中心化的价值网络带来了信息冗余。这种现象被称为策略过拟合[8]。针对策略过拟合问题，将全局的优势值分解每个智能体求和的形式为

$$A(s,\boldsymbol{a})=\sum_{i}^{N}\boldsymbol{A}^{i}(s,a^{i}) \tag{20}$$

这样共享策略的优势值就不会影响不相关的智能体。为了分解出正确的$A^{i}(s,\boldsymbol{a}^{i})$，本文构建了两种方法，直接引入噪声正则化优势值或通过引入噪声正则化状态价值间接影响优势值，以弱化多智能体合作中的策略过拟合问题的影响。

2.3.1　引入噪声正则化优势值

为每个智能体采样一个高斯噪声x^{i}为

$$x^{i}\sim N(0,1),\ \forall i\in N \tag{21}$$

其中：N为智能体的数量，本文在原优势值$A_{b}(s,\boldsymbol{a})$中加入一个α权重的噪声x^{i}，同时在原优势值$A_{b}(s,\boldsymbol{a})$中乘以$1-\alpha$的权值，构成新的优势值$A_{b}^{i}(s,a^{i})$来训练多智能体的策略，新的优势值$A_{b}^{i}(s,a^{i})$为

$$A_{b}^{i}(s,a^{i})=(1-\alpha)A_{b}(s,\vec{a})+\alpha x^{i},\ \forall i\in N,\ b\in B \tag{22}$$

其中：B为采样批量；b为采样批量的一个样本。

2.3.2　引入噪声正则化状态价值

随机采样一个高斯噪声向量$\boldsymbol{x}^{i}\sim N(0,\sigma^{2})$给第$i$个智能体，然后连接噪声和全局状态$S$输入到中心化价值网络中，可转化公式为

$$v^{i}=V\left(concat s,x^{i}\right),\ \forall i\in N \tag{23}$$

随机噪声传递给中心化价值网络并传播到优势值。高斯噪声x^{i}干扰了中心化价值网络，并传播到${}_{i}A(s,a^{i})$，用于扰动优势值。新的优势值${}_{i}A(s,a^{i})$为

$${}_{i}A(s,a^{i})=r+\gamma v^{i}s_{t+1}-v^{i}s_{t} \tag{24}$$

本文采用MAPPO算法对深空探测器小天体附着任务进行建模并优化。每个智能体具有自己的策略网络和价值网络，通过与深空探测器多智能体交互环境不断交互获取奖励，学习并更新状态的动作值[9]。此外算法针对策略过拟合问题引入了噪声正则化优势值，直到每个智能体在当前规划序列下不能通过改变策略获取更大的奖励值。在本文所提的框架中，各个智能体之间是合作的关系，因此规划的目的是使得多智能体系统总的奖励值最大。本文方法通过优势函数求累计奖励最大值并选取累计奖励最大的动作，在不断迭代的过程中学习最优联合策略。最后将训练好的规划策略所对应的规划序列作为最优规划序列[10]。

2.4　多智能体强化学习深空探测任务规划算法实现

深空探测器多智能体利用共享的交互环境计算并获取回报值，经过训练后得到了最优策略，将其用于深空探测器的任务规划。具体实现伪代码如表1所示。

表 1　算法实现

Table 1　Algorithm implementation

Algorithm 1 Noise-MAPPO

Input: Policy Network Parameters θ，Value Network Parameters φ,experience pool $D\leftarrow\{\}$,batch size B,Agent nums N,noise variance σ^{2};
Output： solution;

Sample random noise $\boldsymbol{x}^{i}\sim N(0,\ \sigma^{2}),\ \ \forall i\in N$ for agent i
while step $\leqslant step_{\max}$ do:
　For i=1 to *batch size* do:
　Interact all agents with environment following current policy to obtain $[s,a,r,s']$，add to trajectory τ
　Sample minibatch b from experience pool D
　End For
If use Noise Advantage-MAPPO:
Add noise value ε to each agent according to equation (22)
If use Noise Value-MAPPO:
Add noise value ε to each agent according to equation (24)
Calculate advantage value A for trajectory τ using GAE
Calculate reward value $\hat{R}$ based on trajectory τ
Split trajectory τ into batches and store in experience pool D
For each training epochs do:
　Update Policy network θ according to equation (17)
　Update Value network φ according to equation (19)
End for

3　实验设计与结果分析

3.1　实验案例说明

本文以小天体附着自主任务规划为背景，采用pytorch和gym搭建深度强化学习环境。对第2节所设计的深空探测器自主任务规划模型进行训练，规划目标是在尽可能短的时间用较少的资源完成附着任务自主规划[11-12]，本文使用文献[5]的基础之上构建基于认知图谱的多智能体强化学习环境，在本研究中，多智能体共享同一个环境，结果如图3所示。

多智能体环境中包含每个子系统的名称、涉及的状态、所属的子系统动作、每一个动作消耗的存储容量、电量、燃料、时间，以及根据专家知识设定不同智能体的动作所带来的状态转移关系[13-14]。其中不同颜色代表不同的子系统：橙色表示通信子系统，蓝色表示综电子系统，浅蓝色表示电源子系统，深绿表示载荷子系统，绿色表示推进子系统，棕色表示姿态子系统，黄色表示电缆子系统，粉色表示固定子系统，深粉色表示阀门子系统，红色表示成像子系统。所有子

系统的任务初始状态设置为随机状态，深空探测器要规划的总任务序列包括载荷相机拍照、姿态子系统调整姿态完成附着等动作[15-16]。各子系统需满足1.3节的时序约束、时间约束、资源约束等条件。

图 3 深空探测器多智能体强化学习环境图谱

Fig. 3 Multi-agent reinforce learning environment graph for deep space probes

3.2 对比实验及结果分析

为验证所提出方法的可行性，针对深空探测器小天体附着任务构建测试用例，针对上述10个子系统一共构建了6个不同规模的深空探测器规划任务算例，规划目标为满足资源和时序约束且奖励值最高。本文的仿真实验环境如表2所示，本实验的超参数设置如表3所示，实验算例结果如表表4所示。

为判断使用不同的神经网络与添加噪声的方式对本文算法的影响，将MAPPO中的策略网络中的网络模型分别设为卷积神经网络（Convolutional Neural Network，CNN）和回归神经网络（Recurrent Neural Network，RNN），与不添加噪声、添加噪声给优势值和添加噪声给价值函数的方法进行组合，一共设置6种算法，分别为采用卷积神经网络的MAPPO（下文用MAPPO指代）、采用循环神经网络的MAPPO（下文用RMAPPO指代）、引入噪声正则化优势值采用卷积

表 2　仿真环境

Table 2　Simulation environment

项目	配置
操作系统	Windows10家庭中文版 64位
编程语言	Python3.8.13
处理器	Intel®Core™i5-10400F CPU @3.60 Hz
显卡	GeForce RTX 2060 SUPER
内存	16 GB

表 3　超参数设置

Table 3　Settings of hyper-parameters

超参数	数值	超参数	数值
训练的总步数	200 000	Batch Size	128
最大规划步数	200	裁剪系数	0.2
折扣因子	0.99	GAE的参数λ	0.95
学习率	0.000 5	熵系系数	0.01

神经网络的MAPPO（下文用NA-MAPPO指代）、引入噪声正则化状态价值的采用循环神经网络的MAPPO（下文用NV-MAPPO指代）、引入噪声正则化优势值的采用循环神经网络的多智能体近端策略优化算法（下文用NA-RMAPPO指代）、引入噪声正则化状态价值的采用循环神经网络的多智能体近端策略优化算法（下文用NV-RMAPPO指代）,在6种不同的案例下进行对比实验。

表 4　实验算例说明

Table 4　Experimental case descriptions

案例	最大存储/GB	最大燃料/kg	最大电量/（kW·h）	最大时间/s
案例1	400	400	400	4 000
案例2	500	500	500	5 000
案例3	600	600	600	6 000
案例4	800	800	800	8 000
案例5	1 000	1 000	1 000	10 000
案例6	2 000	2 000	2 000	20 000

3.2.1　多智能体强化学习任务规划奖励值对比实验

根据3.1节和3.2节的设置，对模型进行训练评估，为了更好地展现训练过程中奖励值随着训练步数变化的情况，将每个案例的智能体奖励值之和随训练步数的变化趋势记录下来，如图4所示。

训练曲线表明在不同的案例中，6种算法可以满足深空探测器规划任务。在不同案例中， NV-RMAPPO和RMAPPO比其它算法的收敛速度更快。在案例3～5中NV-RMAPPO的收敛速度最快。说明引入噪声正则化价值函数和采用循环神经网络能更快地得到规划结果。

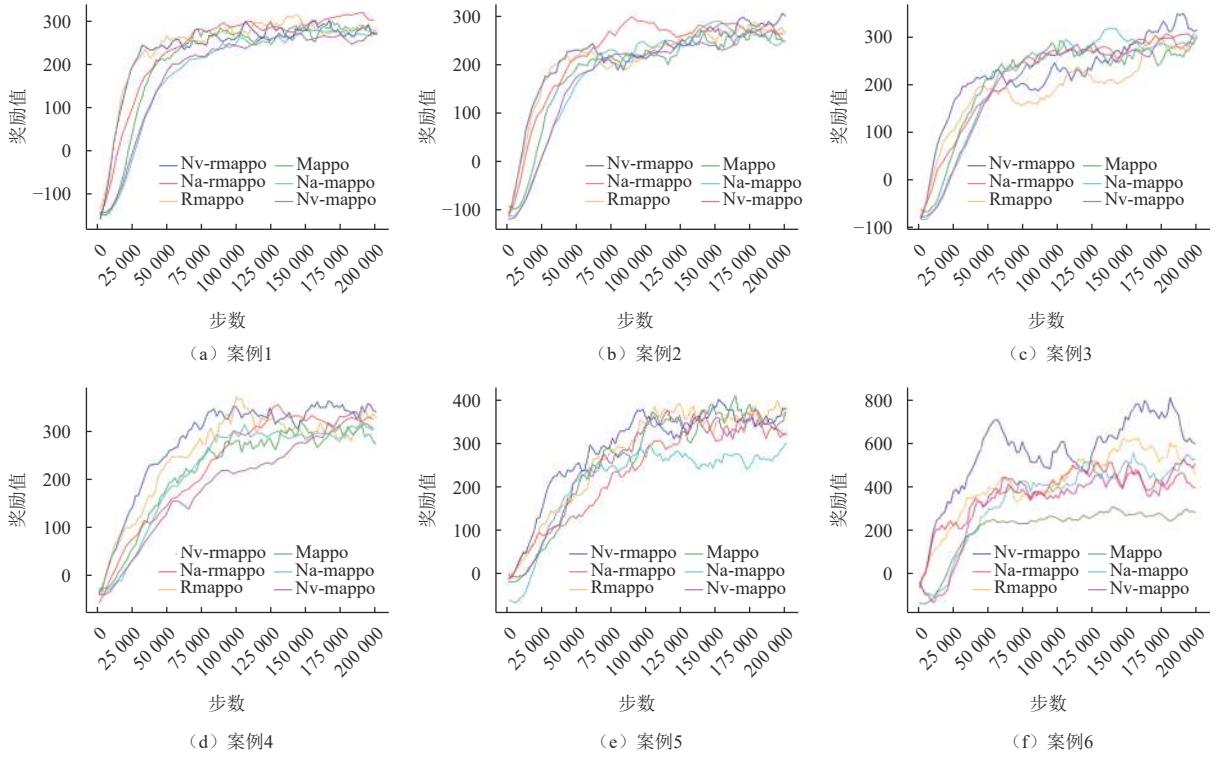

图 4　不同案例下不同算法的奖励值

Fig. 4　Reward values of different algorithms under different cases

3.2.2　多智能体强化学习任务规划适用性实验

为了验证深空探测器多智能体强化学习任务规划在不同规划条件下的适用性。采用6种算法对表4中所示的6个案例进行实验，分别进行10次实验取算术平均值，统计不同案例下不同算法在随机初始状态下的奖励值，规划步数，规划时长以及规划成功率。统计不

同案例下不同算法在随机初始状态下的奖励值、规划步数、规划时长以及规划成功率，结果如图5～8所示。

图5中所示规划奖励值是多次实验的算术平均值，NV-RMAPPO在6个案例中均有较好的表现，在奖励值设置方法相同的条件下比MAPPO有明显提升，平均提升14.16，最少提升6，最多提升33。可以看到改进后的算法相较于原算法的奖励值都有一定程度的提升。根据2.1.4节所述，奖励值的设置与规划序列消耗的资源以及规划的成功与否有关，说明规划解的质量和成功率都有提高。但不能判断规划解质量的提升和成功率的提升对奖励值提高的影响。为更好地判断该影响，因此后文列出了不同案例下不同算法的规划步数，规划时长、规划成功率以及消耗的资源。

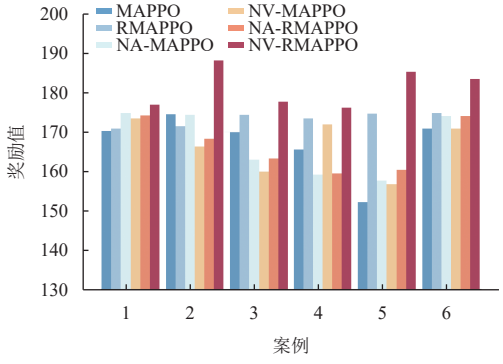

图 5 不同案例的规划奖励值

Fig. 5 Reward values of planning under different cases

图6中所示的规划步数是多次实验的算术平均值。在6个案例中，NV-RMAPPO均有较好的表现，规划步数相较于MAPPO算法，平均减少了1.52步，最少减少了0.7步，最多减少了4.8步。说明NV-RMAPPO算法能用更短的规划步数完成任务规划。

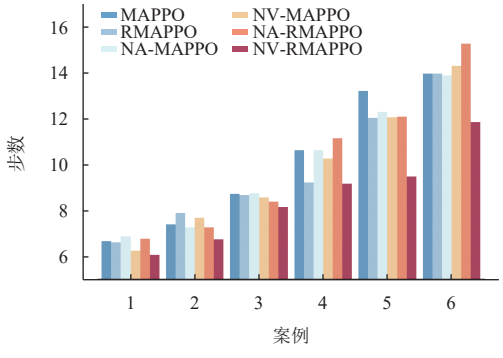

图 6 不同案例的规划步数

Fig. 6 Number of planning steps under different cases

图7为6个案例的规划时长平均值，NV-RMAPPO均有较好的表现， 规划时长较MAPPO平均减少了0.35 s，最少减少了0.07 s，最多减少了0.58 s。说明采用本文方法，引入噪声正则化价值函数和采用循环神经网络有助于缩短任务规划时间。

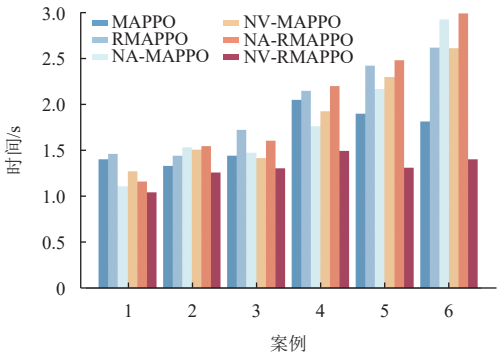

图 7 不同案例下的规划时长

Fig. 7 Planning time under different cases

图8为6个案例下的规划成功率平均值，NV-RMAPPO规划成功率最高， MAPPO规划成功率最低。NV-RMAPPO成功率较MAPPO平均提升11.3%，最低提升6%，最高提升16%，由此可见，本文算法通过引入噪声正则化价值函数和采用循环神经网络有效提高了自主任务规划成功率。

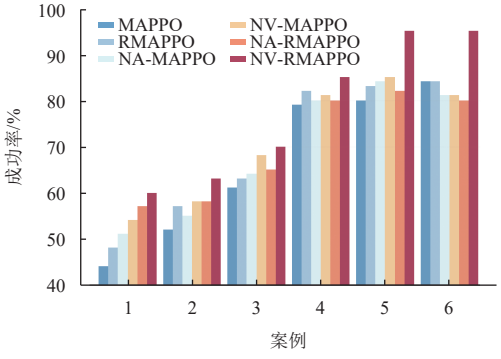

图 8 不同案例下的规划成功率

Fig. 8 Planning success rates under different cases

综合上述实验结果，通过对比不同案例下不同算法的奖励值、规划步数、规划时长以及规划成功率。可以看到本文所提出的NV-RMAPPO相较于其它算法在进行自主任务规划中规划所需时间更短，步数更少，规划成功率更高。

3.2.3 多智能体强化学习任务规划有效性实验

本文针对不同案例下的不同算法的规划序列的资源消耗以及时间消耗的平均值如表5所示，可以看到，引入噪声正则化优势值的算法相较于原算法各种资源的消耗有不同程度的减少，可以说明本文算法有效地提高了规划解的质量。

从有效性实验来看，在不同的案例中NV-RMAPPO算法规划解的质量更高， 综合有效性和适用性 NV-RMAPPO在不同的案例中均有较好的表现。因此本文

采用NV-RMAPPO进行深空探测器自主任务规划仿真实验，可得到包含各子系统状态与动作可行解集输出结果。为更清楚地展现使用本文方法规划的深空探测器多子系统协同工作过程，将规划结果以甘特图展示，规划结果的甘特图如图9所示。由仿真结果可以看出，采用本文算法能够实现深空探测器小天体附着自主规划任务，规划结果较为合理，证明了本文方法的有效性和适用性。

表 5　不同案例下不同算法的资源消耗情况

Table 5　Resource consumption of different algorithms under different cases

案例	算法	平均消耗容量/G	平均消耗电量/（kW·h）是	平均消耗燃料/kg	规划序列的平均总时间/s
案例1	MAPPO	262.75	231.31	36.67	815.94
	RMAPPO	236.37	236.37	32.39	727.39
	NA-MAPPO	258.13	215.62	32.87	771.66
	NV-MAPPO	260.26	209.47	32.50	679.61
	NA-RMAPPO	258.58	240.26	40.93	755.65
	NV-RMAPPO	**265.61**	**211.00**	**21.88**	**593.68**
案例2	MAPPO	281.84	236.95	34.36	909.42
	RMAPPO	261.07	262.83	39.88	840.00
	NA-MAPPO	270.89	272.25	38.75	787.34
	NV-MAPPO	280.41	245.38	35.08	834.63
	NA-RMAPPO	279.86	234.04	46.79	891.86
	NV-RMAPPO	**265.17**	**270.28**	**37.50**	**718.06**
案例3	MAPPO	283.12	296.66	60.08	1026.08
	RMAPPO	262.71	271.37	31.96	817.43
	NA-MAPPO	277.03	285.13	30.74	923.72
	NV-MAPPO	314.1	348.55	67.5	1320.35
	NA-RMAPPO	286.60	291.93	38.57	946.62
	NV-RMAPPO	**268.94**	**271.32**	**44.56**	**781.47**
案例4	MAPPO	334.08	431.08	66.73	1 375.75
	RMAPPO	291.12	323.44	46.91	1 047.15
	NA-MAPPO	318.32	359.28	55.74	1 230.43
	NV-MAPPO	299.61	364.82	50.32	1 211.06
	NA-RMAPPO	325.63	393.93	57	1 340.05
	NV-RMAPPO	**287.42**	**280.85**	**40.94**	**953.92**
案例5	MAPPO	381.06	421.67	85.37	1 624.07
	RMAPPO	331.32	430.28	55.83	1 194.75
	NA-MAPPO	355.2	518.57	89.78	1 781.93
	NV-MAPPO	361.46	478.71	76.99	1 532.79
	NA-RMAPPO	328.61	416.2	49.51	1 305.51
	NV-RMAPPO	**317.5**	**395.65**	**46.22**	**1 007.59**
案例6	MAPPO	495.58	534.84	40.04	1 433.35
	RMAPPO	475.75	535.68	40.45	1 305.51
	NA-MAPPO	430.35	594.57	40.08	1 514.45
	NV-MAPPO	452.39	570.13	76.99	1 532.79
	NA-RMAPPO	487.24	596.03	45.05	1 519.75
	NV-RMAPPO	**476.92**	**390.03**	**34.01**	**1 264.69**

注：加粗部分为每个案例中表现最优的算法及其资源消耗情况。

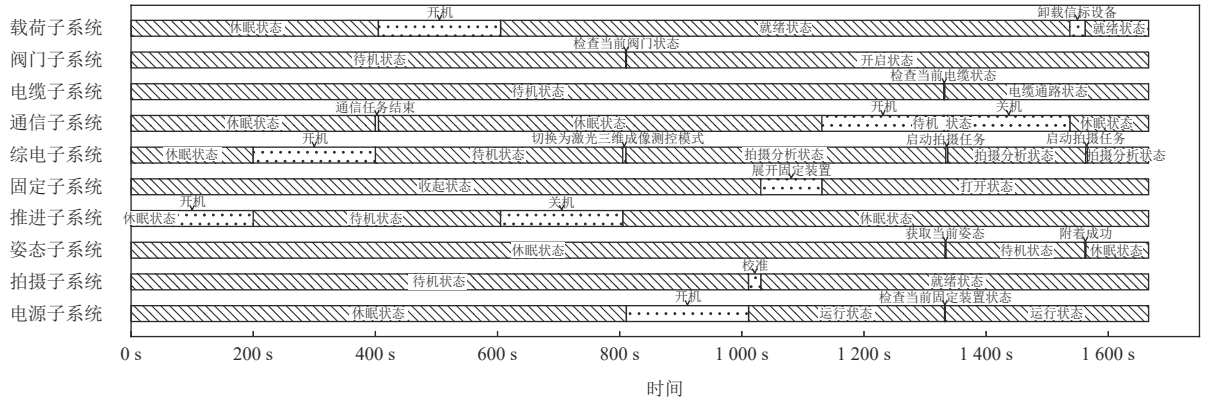

图 9　深空探测器小天体附着任务规划甘特图

Fig. 9　Gantt chart for asteroid rendezvous and sampling task planning of deep space probes

4　结　论

本文以深空探测器小天体附着任务自主规划为目标，针对深空探测器任务规划过程中子系统之间并行协作的多约束要求展开研究，提出了一种基于多智能体强化学习的深空探测器小天体附着任务规划方法，首先根据深空探测任务规划中的状态转移约束、资源约束及时间约束等构建深空探测器多智能体交互知识环境，然后将单智能体近端策略优化算法与多智能体混合式协作机制相结合，并在此基础上引入噪声正则化优势值解决多智能体集中训练中的策略过拟合问题，最终给出了多智能体强化学习自主任务规划模

型。仿真实验结果显示与现有算法相比，本文方法的任务规划成功率平均提高了11.3%，任务规划时间平均缩短了0.35 s，任务规划步数平均减少了1.52步，同时规划结果序列的资源消耗以及总执行时间显著减少。本文方法可以实现深空探测器的自主任务规划，并有效提高了任务规划质量、提升了规划成功率以及规划速度。

后续研究拟将深空探测器任务规划中的资源、时序等多约束条件与奖励函数的设置相融合，以优化本文方法中多智能体强化学习的协作策略，增强该自主任务规划模型对于不同任务的自适应能力，进一步提升规划效率和规划成功率。

参 考 文 献

[1] 徐瑞，李朝玉，朱圣英，等. 深空探测器自主规划技术研究进展[J]. 深空探测学报（中英文），2021，8(2)：111-123.
XU R，LI Z Y，ZHU S Y，et al. Progress in deep space explorer autonomous planning[J]. Journal of Deep Space Exploration，2021，8(2)：111-123.

[2] 姜啸. 基于约束可满足的深空探测器任务规划方法[D]. 北京：北京理工大学，2018.
JIANG X. Mission planning method for deep space probes based on constraint satisfaction [D]. Beijing：Beijing Institute of Technology，2018.

[3] 赵宇庭，徐瑞，李朝玉，等. 基于动态智能体交互图的深空探测器任务规划方法[J]. 深空探测学报（中英文），2021，8(5)：519-527.
ZHAO Y T，XU R，LI Z Y，et al. Mission planning method for deep space probe based on dynamic agent interaction diagram[J]. Journal of Deep Space Exploration，2021，8(5)：519-527.

[4] 史兼郡. 基于深度强化学习的空间站短期任务规划方法研究[D]. 长沙：国防科技大学，2020.
SHI J J. Research on short-term task planning method for space station based on deep reinforcement learning[D]. Changsha：National University of Defense Technology，2020.

[5] 柳景兴，王彬，毛维杨，等. 深空探测器任务规划认知图谱及多属性约束冲突检测[J]. 深空探测学报（中英文），2023，10(1)：88-96.
LIU J X，WANG B，MAO W Y，et al. Cognitive graph for autonomous deep space mission planning and multi-constraints collision detection[J]. Journal of Deep Space Exploration. 2023，10(1)：88-96.

[6] 毛维杨，王彬，柳景兴，等. 基于强化学习的深空探测器自主任务规划方法[J]. 深空探测学报（中英文），2023，10(2)：220-230.
MAO W Y，WANG B，LIU J X，et al. An autonomous planning method for deep space exploration tasks in reinforcement learning based on dynamic rewards[J]. Journal of Deep Space Exploration，2023，10(2)：220-230.

[7] YU C，VELU A，VINITSKY E，et al. The surprising effectiveness of MAPPO in cooperative，multi-agent games[EB/OL]. (2022-11-4) [2023-11-03]. https://arxiv.org/abs/2103.01955v1.

[8] WANG S Y，CHEN W Y，HU J，et al. Noise-regularized advantage value for multi-agent reinforcement learning[J]. Mathematics，2022，10(15)：2728.

[9] SCHULMAN J，WOLSKI F，DHARIWAL P，et al. Proximal policy optimization algorithms[EB/OL]. (2017)[2023-11-3]. http://arxiv preprint arxiv：1707.06347，2017.

[10] 司雪圆. 基于约束可满足的航天器自主任务规划方法研究[D]. 北京：北京理工大学，2015.
SI X Y. Autonomous mission planning method of spacecraft based on the constraint satisfaction[D]. Beijing：Beijing Institute of Technology，2015

[11] 徐雅男. 小行星附着机构的整机构型设计与动力学分析[D]. 南京：南京航空航天大学，2023.
XU Y N. Whole mechanism design and dynamic analysis of asteroid attachment mechanism[D]. Nanjing：Nanjing University of Aeronautics and Astronautics，2023.

[12] 崔平远，徐瑞，朱圣英，等. 深空探测器自主技术发展现状与趋势[J]. 航空学报，2014，35(1)：13-28.
CUI P Y，XU R，ZHU S Y，et al. Development status and trend of deep space probe autonomous technology[J]. Acta Aeromautica et Astaronautica Sinica，2014，35(1)：13-28.

[13] KOOTBALLY Z，SCHLENOFF C，LAWLER C，et al. Towards robust assembly with knowledge representation for the planning domain definition language (PDDL)[J]. Robotics and Computer-Integrated Manufacturing，2014，33(C)：42-45.

[14] GHALLAB M，NAU D S，TRAVERSO P . Automated planning：theory & practice[M]. Burlington：Morgan，2004.

[15] 徐文明. 深空探测器自主任务规划方法研究与系统设计[D]. 哈尔滨：哈尔滨工业大学，2006.
XU W M. Research and system design of autonomous mission planning methods for deep space probes [D]. Harbin：Harbin Institute of Technology，2006.

[16] 冯小恩，李玉庆，杨晨，等. 面向自主运行的深空探测航天器体系结构设计及自主任务规划方法[J]. 控制理论与应用，2019，36(12)：2035-2041.
FENG X E，LI Y Q，YANG C，et al. Architecture design and autonomous mission planning for autonomous deep space exploration spacecraft[J]. Control Theory and Application，2019，36(12)：2035-2041.

作者简介：

孙泽翼(1998−)，男，硕士研究生，主要研究方向：人工智能、深空探测器自主任务规划。

通信地址：云南省昆明市呈贡新区景明南路727号(昆明理工大学信自楼)(650504)

电话：18468005198

E-mail：2676537466@qq.com

王彬(1977−)，博士，副教授，主要研究方向：实时控制、智能控制、智能信息处理。**本文通信作者**。

通信地址：昆明理工大学信息工程与自动化学院(650500)

电话：18725152375

E-mail：wangbin@kust.edu.cn

Multi-Agent Reinforcement Learning Autonomous Task Planning for Deep Space Probes

SUN Zeyi[1]，WANG Bin[1,2]，HU Xinyue[1]，XIONG Xin[1,2]，JIN Huaiping[1,2]

（1. Faculty of Information Engineering & Automation，Kunming University of Science & Technology，Kunming 650500，China；
2. Yunnan key Laboratory of Artificial Intelligence，Kunming University of Science & Technology，Kunming 650500，China）

Abstract：To meet the requirements for autonomy，rapidity，and adaptability in the collaborative planning of each subsystem during the attachment mission of a deep space probe，a collaborative planning strategy based on proximal policy optimization method and multi-agent reinforcement learning was proposed. By combining the single-agent proximal policy optimization algorithm with the hybrid collaborative mechanism of multi-agent，a multi-agent autonomous task planning model was designed. The noise-regularized advantage value ws introduced to solve the problem of overfitting in the collaborative strategy of multi-agent centralized training. Simulation results show that the multi-agent reinforcement learning collaborative autonomous task planning method can intelligently optimize the collaboration strategy of small celestial body attachment missions according to real-time environmental changes，and compared with the previous algorithm，it improves the success rate of task planning and quality of planning solutions，and shortens the time of task planning.

Keywords：multi-agent reinforcement learning；autonomous task planning of deep space exploration；proximal policy optimization；small celestial body attachment

Highlights：

- Research has been carried out on multi-constraint requirements in parallel cooperation between multiple subsystems in the mission planning process for deep space probes.
- A multi-agent reinforcement learning based deep space probe asteroid attachment mission planning method has been proposed.
- Based on the distributed parallel features and multi-constraint properties of various subsystems of the deep space probe，we propose a multi-agent reinforcement learning coordinated planning strategy based on proximal policy optimization for deep space probes.
- To address the policy overfitting problems faced in centralized training of multi-agent systems，we introduced a regularization on policy’s advantage value using noise to optimize the multi-agent coordinated optimization strategy for deep space probes.

[责任编辑：宋宏，英文审校：宋利辉]