星地融合网络中基于异质图表征的多智能体协作切换方法

付一阳[1)] 胡 博[1)] 刘人鹏[1)] 陈山枝[2)]

[1)](北京邮电大学网络与交换技术全国重点实验室 北京 100876)

[2)](中国信息通信科技集团有限公司无线移动通信全国重点实验室 北京 100191)

摘 要 随着卫星通信技术的快速演进,星地融合网络已成为构建下一代全球网络的重要发展方向。低轨卫星作为其重要组成部分,以低延迟和广覆盖等优势,正成为全球产业界与学术界的焦点。低轨卫星在近地轨道上高速运动,导致星地融合网络拓扑结构的高动态变化,这使得星地和星间的频繁移动切换成为影响通信连续性的重要技术挑战。当前,研究者们基于多智能体深度强化学习,利用分布式架构与多目标优化等技术实现卫星切换控制算法,但在状态建模和特征提取等方面仍有不足,限制了切换决策效果。本文提出一种星地融合网络的基于异质图表征的多智能体协作切换方法MA-HGDQN,旨在提升强化学习方法的状态表达能力与决策性能。该方法将星地网络抽象为多种类型的节点与边构成的异质图,通过异质图注意力网络提取网络拓扑特征并生成低维嵌入向量,再将该向量作为各智能体的状态输入。在此基础上,构建多智能体协同机制,结合贡献量化与同质智能体参数共享策略,实现通信性能优化与系统负载均衡的协同演化。在拟合性能方面,所提方法聚焦"预切换"阶段,在提升切换经验占比的同时显著降低计算量,仿真耗时较A2C减少88%。在切换性能方面,MA-HGDQN较未引入异质图的MA-DQN算法,累计切换次数减少10%、通信中断次数减少45%,系统负载公平性提升11%以上,性能优势明显。此外,本文验证了所提算法在实际应用中的适用性,其决策用时受星座构型和用户规模的影响小,且切换频率和接入成功率表现稳定,工程应用前景广阔。

关键词 星地融合网络;异质图表征;切换控制;负载均衡;多智能体深度强化学习

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00347

A Multi-Agent Heterogeneous Graph Representation-Based Handover Method in Satellite-Terrestrial Integrated Network

FU Yi-Yang[1)] HU Bo[1)] LIU Ren-Peng[1)] CHEN Shan-Zhi[2)]

[1)](State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876)

[2)](State Key Laboratory of Wireless Mobile Communications, China Information Communication Technologies Group Corporation, Beijing 100191)

Abstract With the rapid development of the satellite communication technology, satellite-terrestrial integrated networks (STINs) have become a key development direction for developing the next-generation global network. As a key component of this technology, low earth orbit (LEO) satellites are becoming a focus of global industry and academia, owing to their distinct advantages such as low latency and broad coverage. It is expected to be combined with terrestrial

收稿日期:2025-04-22;在线发布日期:2025-10-24。本课题得到国家自然科学基金重点项目(No. 61931005)资助。**付一阳**,博士研究生,主要研究领域为星地融合通信、移动性管理、深度强化学习和生成式人工智能。E-mail:fuyiyang@bupt.edu.cn。**胡 博**(通信作者),博士,教授,主要研究领域为星地融合通信、低空智联网、网络人工智能、泛在移动计算。E-mail:hubo@bupt.edu.cn。**刘人鹏**,博士研究生,主要研究领域为星地融合通信、移动性管理、深度强化学习和生成式人工智能。**陈山枝**,博士,教授级高工,IEEE Fellow,国家杰出青年科学基金获得者,主要研究领域为B5G/6G移动通信网络架构、星地融合通信、车联网。

communication networks to achieve a global coverage. However, satellite communication still faces several challenges, especially the frequent changes in the STIN topology caused by the high-speed movement of LEO satellites in near-Earth orbits. Frequent mobility handovers between space-to-ground and inter-satellite links have become a key technology challenge impacting the continuity of communications. To address the problem, previous studies have used multi-agent deep reinforcement learning approaches, which leverage multi-objective optimization and distributed architectures to adapt to large-scale satellite network scenarios. Although these methods have made progress, existing studies still have limitations in state modeling and feature extraction, particularly in not fully utilizing network structural information, which restricts the effectiveness of handover decision-making. This limitation makes it difficult for existing methods to fully exploit their potential in the face of complex network structures. This paper proposes a multi-agent collaborative handover method based on heterogeneous graph representation for STINs, termed MA-HGDQN, aiming to enhance state representation and decision-making performance. The core idea of this method is to express the STIN as a heterogeneous graph composed of multiple types of nodes and relationships, effectively capturing the network′s complex topology. By using a heterogeneous graph attention network to extract low-dimensional embedding vectors of topological features as the state input for each agent, the system′s understanding of network structures and decision-making accuracy is significantly improved. On this basis, this paper also proposes a multi-agent collaboration mechanism, combining agent contribution quantification and homogeneous agent parameter sharing strategies to achieve communication performance optimization and system load balancing. Through this collaboration mechanism, different agents can better cooperate to address the handover issue in large-scale networks, ultimately optimizing communication performance and enhancing system stability. In terms of fitting performance, the proposed method focuses on the "pre-handover" stage, increasing the proportion of handover experience while significantly reducing computational load, which results in a simulation time reduction of 88% compared to the A2C algorithm. In terms of handover performance, compared to the MA-DQN algorithm without the introduction of heterogeneous graphs, MA-HGDQN shows significant improvements in the number of handovers, communication interruption times, and system load fairness. Specifically, the cumulative number of handovers decreases by 10%, communication interruptions decrease by 45%, and system load fairness improves by more than 11%, demonstrating clear advantages. Additionally, this paper also validates the applicability of the proposed algorithm under actual application scenarios. Experimental results demonstrate that the proposed MA-HGDQN algorithm maintains stable performance under various network conditions. Experimental results demonstrate that the algorithm maintains a stable decision time under different constellation configurations and user scales, with consistent handover frequency and access performance, demonstrating its broad adaptability and effectiveness in practical applications.

Keywords satellite-terrestrial integrated network; heterogeneous graph representation; handover control; load balancing; multi-agent deep reinforcement learning

1 引 言

为了满足日益增长的全球通信业务需求，在第五代移动通信标准（the Fifth Generation Mobile Communication，5G）的基础上，第六代移动通信标准（the Sixth Generation Mobile Communication，6G）提出了沉浸式通信、超大规模连接、超高可靠低时延通

信、泛在连接、通信与人工智能(Artificial Intelligence，AI)、通信感知一体化等六大场景，增进网络与多领域技术的融合，实现全球、全域、全时网络覆盖[1]。目前，5G通信的连接对象主要集中在陆地区域海拔10 km以内，考虑到建设成本和维护成本，难以全面覆盖偏远地区、海洋以及空中的通信设备[2]。为实现全球、全域、全时网络覆盖，具有受地理条件的影响小、覆盖范围广的非地面网络(Non-terrestrial Network，NTN)，可以作为地面网络(Terrestrial Network，TN)的有力补充，弥补其通信盲区。产业界与学术界正共同推进地面移动通信与卫星通信的融合，致力于实现随时随地的无缝覆盖与高质量服务[3]。其中，部署高度在2000 km以下的低地球轨道(Low Earth Orbit，LEO)卫星凭借其低时延、广覆盖、高容量等核心优势，在通信、导航、遥感等领域展现出显著优势。例如，在远距离通信时，低轨卫星网络能为地面用户提供比现有光纤传输更短的端到端时延[4]，这导致在世界上的高科技企业争相设计和建造自己的巨型星座系统，如Starlink、OneWeb等[5]。

卫星轨道高度的降低不止减少了信号的传播时延，也导致单星覆盖面积的缩减以及卫星与用户终端之间的高速相对运动，单颗LEO卫星通常只能给地面固定用户提供几分钟到十几分钟间的持续服务，导致用户需要频繁进行切换以保证通信的连续性[6]。这不仅会带来切换时延，还可能造成传输中断、信令开销增加等问题。特别地，当卫星高速移动引发多个用户和业务切换时，触发大量切换请求导致网络资源竞争加剧，进而影响切换成功率[1]。例如，大量终端同时选择信号最强的同一卫星作为接入点，这会导致有限的星上资源无法满足所有终端的切换请求，部分终端连接中断[7]。因此，如何设计切换判决方案实现平稳可靠的切换决策，始终是LEO卫星网络研究中的核心议题。

根据用户是否变更卫星节点，卫星网络的切换决策可以划分为波束切换和卫星切换[8]。早期LEO卫星网络中卫星数量有限，对LEO卫星网络切换方案的研究主要集中在波束切换上[9-11]；随着大量LEO星座依次建成，卫星间切换的频率大幅增长，切换方案的研究重点转向卫星切换方案的设计。截至目前，研究人员分别从属性切换方法[12-15]、图模型方法[16-20]以及强化学习方法[21-24]等角度出发展开研究，并取得诸多优秀成果。其中，属性切换方法是基于预设属性阈值进行切换决策，例如根据信号强度或仰角等单一指标设定阈值触发切换。其优点是实现简单、开销低，但缺点是在复杂动态环境中表现欠佳，难以应对多种目标间的权衡。图模型方法需要把切换问题抽象为最短路径问题或是将网络转化为二分图求解。此类方法可以通过调整边的权重，灵活调整优化目标，但随着网络规模的扩大，算法复杂度飞速提升，难以在快速变化的卫星网络环境中实时应用。针对前两种方法面临的问题，强化学习方法可以通过离线训练智能体，让其在模拟环境中反复试错学习复杂策略，训练完成后在线决策速度快、可实时响应变化；通过多智能体协作来实现分布式策略，避免大规模网络中算法复杂度过高的问题。然而，强化学习方法因其隐式学习的方式，导致拟合速度慢、决策性能不稳定等问题。本文聚焦于利用异质图模型捕获网络拓扑和节点异构性，给智能体提供有效特征，从而提升DRL卫星切换决策的稳定性，此研究具有重要的理论意义和现实价值。

本文主要的研究工作如下：

(1) 本文提出了一种面向多智能体决策的局部网络的界定方法，根据节点间可见关系、连接关系和空间距离关系计算关联分数，评估节点在智能体切换决策中的重要程度，筛减掉弱关联节点，从而避免其影响到智能体的推理过程。

(2) 本文提出了一种MA-HGDQN(Multi-agent Heterogeneous-graph-representation-based Deep Q Network)算法。针对每个智能体，引入异质图来显式定义局部网络中的卫星节点和地面用户间的关系，提出一种多重注意力机制从可见关系、连接关系、距离关系三方面分别迭代聚合节点特征，减少局部观测噪声的干扰，提升智能体决策推理的稳定性。此外，提出了一种基于系统负载平衡和切换性能的长短期全局奖励方法，促进多智能体间的协作优化，在减少累积切换次数和连接中断次数的同时，维持卫星网络整体负载的均衡。

(3) 仿真结果表明，本文所提算法MA-HGDQN在拟合速度、累积切换次数、连接失败概率和系统公平性方面全面优于MA-DQN算法，证明了异质图网络能提取星地网络特征，进而提升强化学习算法拟合稳定性。

本文后续组织如下：第2节介绍了现有卫星切换方法；第3节描述了系统模型的构建方式；第4节具体展开介绍所提算法的异质图表征、多智能体协同机制等关键技术；第5节通过仿真与性能分析验证了所提算法的性能优势以及实际应用中的适用性；第6节对全文进行了总结。

2 相关工作

早期的卫星切换方案主要依据一条或多条切换准则判断切换动作，分别称为单属性决策方法和多属性决策方法[12-15]。其中，单属性决策方法的核心思想是以单一属性因子作为判决依据，根据预先设定的切换属性阈值判断切换决策，常见的切换依据包括仰角、剩余服务时间、信道状态等。面对复杂多变的卫星网络环境，单属性决策方法容易导致个别属性优异但其余属性较差，难以同时应对多种性能需求[12]。针对这一问题，研究者们围绕多属性决策开展研究工作。文献[13]提出根据卫星和用户的实际情况赋予属性因子不同的权重，从而选出最合适的切换目标。文献[14]利用熵值法对各个切换条件进行加权，将多目标优化问题转化为单目标优化问题。可见，熵值法只利用了卫星系统的瞬时信息，难以保证用户长期的切换性能优化。但由于算法复杂度较高，难以直接部署在LEO卫星星座上。文献[15]提出了一种基于分组和集群的移动性管理架构，在超密集低轨道卫星网络中实现灵活的功能配置和轻量级的移交流程。相比单属性决策，多属性决策可以在优化单项属性的同时，避免其他属性性能过度劣化，实现多属性间的平衡。然而，上述方法往往需要针对特定场景人为设计属性的权重分配，难以灵活适配不同场景。

为了更好适配具有动态拓扑的卫星网络，研究者们提出了基于图的卫星切换方案[16-20]。文献[16]提出了一种基于图的基本卫星切换框架，其中将卫星的每个覆盖时段为节点，存在重叠的覆盖时段间构建有向边，从而把切换决策问题转化为最短路径选择问题。文献[17]进一步将有向图扩展为时间扩展图，并更新每个时隙的最短路径以提高切换预测成功率。另一种常见思路则是构建卫星和地面用户的二分图，根据节点和边的特征优化边的权重，从而选择最优链路[18-20]。文献[18]使用二部图的形式表示低轨卫星与地面网关之间的通信，并基于Kuhn-Munkres算法研究了最大权重匹配问题，在发射功率约束下最大限度提高整体通信质量和平衡卫星系统负载，并应用MIMO技术提高传输速率。文献[19]提出了一个基于图的可定制切换框架，把边的权重设定为可定制的切换标准，在选择保持服务质量（Quality of service，QoS）的切换序列时同时考虑切换时间和目标。文献[20]研究了卫星服务航空交通场景下的切换问题，利用基于有向图的切换框架和多属性决策算法构建多属性动态图，对系统总体吞吐量进行优化。然而，图的规模会随着网络中卫星节点的增加而迅速扩大，导致训练时间过长，因此如何将其应用于巨型星座是一项挑战。

为了让切换方案适配高动态的星地网络，研究者将强化学习（Reinforcement Learning，RL）方法应用到卫星切换设计中，利用智能体与环境交互的特性，有效感知系统状态，根据动作和奖励实时更新策略，不断更新迭代[20-23]。在地面移动通信系统中，智能切换决策倾向于以网络侧为中心的集中式决策。这是由于网络侧对邻近小区的资源状态（如负载、干扰、信干噪比等）和网络拓扑具有全局可见性。与之相对，在卫星通信系统中，卫星难以实时获取所有终端的精确位置信息、遮挡情况、干扰状态等参数，而终端可以基于星历预测卫星的运行轨迹和可见时间。以终端为中心的切换决策方式更有利于信息的获取。此外，低轨卫星网络规模不断扩大，由最初数十颗逐渐增长至数千甚至数万颗，面向整个网络的集中式决策需要获取和处理大量数据。这不仅导致训练模型的复杂度极高，在实际应用中单智能体也难以实时观测到整个网络环境的状态信息。因而，面向星地网络的切换决策逐渐向着以用户为中心、自主智能的方向发展。文献[21]对比基于集中式切换方案和分布式切换方案的Deep Q-learning算法，证明了分布式智能切换方案可通过减小状态空间和动作空间，解决大规模星地网络切换决策的高复杂度问题。文献[22]综合考虑信道质量、空闲信道数和剩余服务时间，把切换决策中的多目标优化问题转化为RL问题，采用深度Q网络（Deep Q Network，DQN）求解，在避免切换失败的同时减少切换次数。文献[23]提出了一种基于TAD-RL的切换控制方法。该方法通过引入期望值替代与自适应梯度调整机制，有效缓解了时序差分目标噪声问题，提升了高动态变负载环境下的策略稳定性。文献[24]提出了一种基于并行模糊神经网络的切换判决方法，引入并行输入模块来提高决策性能，实现了平滑且可靠的切换，并有效降低了切换时延、呼叫阻塞率和切换次数。

然而，上述RL算法均使用隐式学习的方式提取状态空间里的特征，迭代优化模型的动作策略，提取特征的过程处于“黑箱”状态，这导致了以下问题。一方面，智能体需要大量样本数据试错来推断环境、动作、奖励之间的关系，未能利用先验结构知

识，导致收敛速度缓慢。另一方面，隐式学习采用神经网络隐式编码策略，若Q值估计因隐式学习的不确定性产生高方差，策略更新方向将不稳定。为了提高模型的稳定性，一种思路是引入图来建模环境中实体之间的显式关系。

本文将异质图网络用于提取星地网络特征，优化DRL的卫星切换决策，提出了一种面向星地融合网络的基于异质图表征的多智能体协作切换方法，此研究具有重要的理论意义和现实价值。

3　问题描述

3.1　系统模型

本文符号的含义见表1。在LEO卫星通信网络，地面用户终端往往位于多个卫星节点的可见范围内，如图1所示。假设地面用户终端配备单天线，网络中有N个卫星节点和K个地面用户终端，分别用集合$\forall n\in\mathcal{N}=\{1,2,\cdots,N\}$和$\forall k\in\mathcal{K}=\{1,2,\cdots,K\}$表示。卫星按照预先配置的轨迹连续运动，并在有限的时间内覆盖自身可见区域内的用户。用户可以使用全球定位系统（Global Position System，GPS）确定自身确切的位置，并根据星历预测自身和卫星之间的覆盖关系[25]。

表1　系统模型变量

变量名称	含　义
B_{load}	卫星网络的负载平衡度
$c_{n,k}^{t}$	在时间步t下卫星n与用户k的连接关系
$(CNR)_{\min}$	满足通信需求的CNR最小值
$EIRP$	有效全向辐射功率
G_r	用户终端接收天线增益
T	时隙数量
l_n^t	在时间步t下卫星n的负载
N	卫星节点数量
K	用户终端数量
n_0	系统噪声功率
$q_{n,k}^{t}$	在时间步t下卫星n与用户k间的信号强度
$v_{n,k}^{t}$	在时间步t下卫星n与用户k间的剩余服务时间
$v_{\min}$	剩余服务时间的限定值
$U_{\max}$	每颗卫星最大可服务用户数量
$v_{n,k}^{t}$	在时间步t下卫星n与用户k的可见关系

使用$v_{n,k}$表示卫星n与用户k之间的覆盖情况，定义为

$$v_{n,k}=\begin{cases}1,\text{用户}k\text{在卫星}n\text{的覆盖范围内}\\0,\text{其他}\end{cases}\tag{1}$$

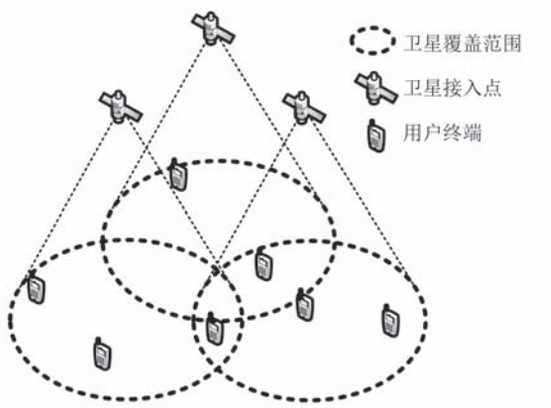

图1　LEO卫星通信网络

类似地，卫星n与用户k之间的连接关系可定义为

$$c_{n,k}=\begin{cases}1,\text{用户}k\text{连接到卫星}n\\0,\text{其他}\end{cases}\tag{2}$$

3.2　切换因素

3.2.1　卫星节点负载

为了降低切换失败率，需要尽量保证卫星网络的负载均衡。用户节点应倾向于选择具有较多空闲资源的卫星作为新的接入点。我们使用卫星节点当先服务的用户数量和最大的可服务用户数量之比表示卫星节点的负载，表示为

$$l_n=\frac{\sum_{k\in\mathcal{K}}c_{n,k}}{n_{\max}}\tag{3}$$

其中，$n_{\max}$表示单颗卫星连接用户的最大数量，所有卫星节点的负载集合表示为$\mathbb{L}=\{l_1,l_2,\cdots,l_N\}$。当用户选择连接到$l_n=1$的卫星节点时，我们假定卫星和用户之间的连接中断。

我们定义同用户节点间存在连接关系的所有卫星节点负载的方差作为网络负载的失衡程度，并使用其倒数表示网络负载的公平性，网络负载的失衡程度表示为

$$B_{load}=\frac{\sum_{n\in\mathcal{N}}\left(l_n-\bar{l}\right)^2\left(1-\prod_{k\in\mathcal{K}}(1-c_{n,k})\right)}{\sum_{n\in\mathcal{N}}\left(1-\prod_{k\in\mathcal{K}}(1-c_{n,k})\right)}\tag{4}$$

其中，$\bar{l}$表示给用户提供服务的所有卫星节点的平均负载。

3.2.2　剩余服务时间

根据星历信息，我们可以推测出用户和卫星在所有时刻的仰角数值，并进一步分析出卫星进入或离开用户可见范围的时刻，从而知晓卫星和用户间

的剩余服务时间。为了便于表示时刻，本文把系统时间离散化为T个时间步，假定通信链路状态每个时隙内不发生变化。在时间步t下卫星n与用户k间的剩余服务时间定义为

$$\upsilon_{n,k}^{t}=\begin{cases}t_{n,k}^{out}-t, t\in\left(t_{n,k}^{in}, t_{n,k}^{out}\right)\\0, \text{其他情况下}\end{cases} \tag{5}$$

其中，$t_{n,k}^{in}$和$t_{n,k}^{out}$分别表示卫星n进入和离开用户k可见范围的时刻。当用户选择连接到$\upsilon_{n,k}^{t}=0$的卫星节点时，我们假定卫星和用户之间的连接中断。

3.2.3 信号强度

由于LEO卫星网络的高动态性，卫星节点和用户节点之间的信道质量会快速变化，本文中使用载波噪声比（Carrier to Noise Ratio，CNR）评估信道质量的度量。通信链路的CNR计算为

$$CNR=EIRP+G_r-PL-n_0 \tag{6}$$

其中，有效全向辐射功率（Effective Isotropic Radiated Power，EIRP）反映发射机发送功率和发射天线增益的综合效果，G_r是接收天线的增益，PL是通信链路的总路径损耗，n_0是系统噪声功率。

3.2.4 连接中断

本文使用卫星节点负载、剩余服务时间、信号强度来模拟当前时刻卫星n和用户k间通信链路的中断概率$p_{n,k}^{t}$。

$$p_{n,k}^{t}=\left(1-\mathcal{F}_1\left(l_n^t\right)\right)\left(1-\mathcal{F}_2\left(\upsilon_{n,k}^{t}\right)\right)\left(1-\mathcal{F}_3\left(CNR_{n,k}^{t}\right)\right) \tag{7}$$

其中，$\mathcal{F}_1(\cdot)$、$\mathcal{F}_2(\cdot)$、$\mathcal{F}_3(\cdot)$分别表示因单个切换因素导致连接中断的概率函数，定义方式如下：

$$\mathcal{F}_1(l)=\begin{cases}1, l=1\\0, l<1\end{cases} \tag{8}$$

$$\mathcal{F}_2(\upsilon)=\begin{cases}1, \upsilon=0\\0, \upsilon>0\end{cases} \tag{9}$$

$$\mathcal{F}_3(x)=\begin{cases}\dfrac{(CNR)_{\min}-x}{1+(CNR)_{\min}-x}, x<(CNR)_{\min}\\0, \quad x\geqslant(CNR)_{\min}\end{cases} \tag{10}$$

3.3 问题建模

用户周期性地检测LEO卫星网络信息，并根据信息决定是否切换卫星接入点。为了保证用户通信服务的连续性，本文考虑以下三种切换准则。

第一，切换次数增加会导致连接中断概率提升和信号开销增长。因此，需要设置一个正值p_1作为切换成本，当通信链路的剩余服务时间大于限定值$r_{\min}$且发生切换过程时，产生切换成本p_1，即

$$\zeta_{n,k}^{t}=\begin{cases}p_1, c_{n,k}^{t-1}\neq c_{n,k}^{t}\cap\upsilon_{n,k}^{t}>\upsilon_{\min}\\0, \text{其他情况下}\end{cases} \tag{11}$$

第二，使用信号强度、负载状态、剩余服务时间判断当前连接是否中断。令$(CNR)_{\min}$表示满足通信需求的CNR最小值，当信号强度低于限定值、剩余服务时间不足或卫星节点负载过高，判断为连接中断，此时产生连接中断成本$f_{n,k}^{t}=p_2$（p_2为正值），其他情况下设置$f_{n,k}^{t}=0$，即

$$f_{n,k}^{t}=\begin{cases}p_2, q_{n,k}^{t}<(CNR)_{\min}\text{或}\upsilon_{n,k}^{t}=0\text{或}l_n=1\\0, \text{其他情况下}\end{cases} \tag{12}$$

第三，为避免用户大量同时接入同一颗卫星，保持系统负载平衡，使用系统负载平衡度B_{load}设置惩罚函数$\mathcal{F}(\cdot)$，即

$$\psi^t=\mathcal{F}\left(B_{load}\right) \tag{13}$$

本文切换方案的目标是最小化系统的长期切换次数，同时保证用户的通信质量以及系统负载平衡，多目标优化问题可以表示为

$$\begin{aligned}&\min_{\left\{c_{n,k}^{t}\right\}}\sum_{t=0}^{T}\left(\psi^t+\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}c_{n,k}^{t}\left(\zeta_{n,k}^{t}+f_{n,k}^{t}\right)\right)\\&\text{s.t. } \mathrm{C}_1: c_{n,k}^{t}\in\{0,1\}, \forall n\in\mathcal{N}, \forall k\in\mathcal{K}\\&\quad \mathrm{C}_2: \sum_{n\in\mathcal{N}}c_{n,k}^{t}=1, \forall k\in\mathcal{K}\\&\quad \mathrm{C}_3: \sum_{k\in\mathcal{K}}c_{n,k}^{t}\leqslant n_{\max}, \forall n\in\mathcal{N}\\&\quad \mathrm{C}_4: \zeta_{n,k}^{t}\in\left\{0,p_1\right\}, \forall n\in\mathcal{N}, \forall k\in\mathcal{K}\\&\quad \mathrm{C}_5: f_{n,k}^{t}\in\left\{0,p_2\right\}, \forall n\in\mathcal{N}, \forall k\in\mathcal{K}\end{aligned} \tag{14}$$

约束条件C_1、C_2和C_3用于控制卫星节点和用户节点的连接状态，保证用户节点在每个时隙仅和单个卫星节点保持连接，且同时接入同一卫星节点的用户数量不超过卫星节点的总信道数。C_4和C_5分别是切换成本和连接中断成本的设置。

4 方法描述

本文的关键问题是确定在切换事件期间每个用户应该连接到哪颗卫星以保持通信连续，从而在平衡卫星之间的负载的同时最小化随后不必要的切换事件的数量。本文将多目标优化问题转换为MDP网络，并提出了一种MA-HGDQN算法进行切换决策。

4.1 DRL类型分析

RL是一种用于分析和自动化目标导向学习的计算方法，通过与环境的多次直接交互进行的智能

体学习，从而不需要完美的监督或完整的环境模型[26]。任何RL问题都可以由五元组$\{k,S,A,\mathcal{R},\pi\}$定义。$k$表示智能体，是在每个时间步与环境交互的组件。智能体观察环境的当前状态$s^t\in S$(S是k的所有可能状态的集合)，然后在采取动作$a^t\in A$(A是k的所有可能动作的集合)后，智能体k通过奖励函数$\mathcal{R}$获得瞬时奖励r^{t+1}，当前状态s^t在的动作a^t的影响下转换到新的状态s^{t+1}。π是指导代理根据其当前状态选择合适动作的策略。

我们围绕无模型(Model-free)的DRL展开研究工作，其不需要环境的动态信息，而是通过与黑箱环境的试错交互从头开始学习策略[27]。无模型DRL可以根据是否直接学习策略进一步分为两类，即基于值的方法和基于策略的方法。

对于基于策略的DRL方法，参数θ更新依赖于策略梯度估计

$$\nabla_\theta \mathcal{J}(\theta)=E_{\tau\sim\rho_\theta}\left[\sum_{t=0}^{T}\nabla_\theta \ln\pi_\theta\left(s^t,a^t\right)\mathcal{R}(t)\right] \tag{15}$$

其中，轨迹$\tau=(s_0,a_0,r_1,s_1,a_1,\cdots,s_T)$，未来奖励的折扣总和$\mathcal{R}(t)=\sum_{i=t}^{T-1}\gamma^{i-t}r^i$。在LEO卫星切换问题中，每颗卫星给用户提供的最长服务时间通常在3分钟～5分钟。对应到按秒生成的星地网络切片中，合理的“切换动作”通常每隔上百个时间步才会出现一次。多数步的$r^{t+1}\approx 0$，致使有效梯度来源极少。

对于基于值的DRL方法，可以结合经验回放池，对稀疏切换事件进行优先采样，利用同一事件多次学习，极大提升稀疏信号的样本利用率，缓解事件稀缺带来的学习停滞，因而更加适合用于本文问题的解决。

DQN是基于值的DRL模型的基础。DQN算法允许智能体根据评估状态-动作对的动作值来学习最优策略π^*。不同于注重当前可用数据的权重分配的熵值法，DQN使用基于预期长期回报的决策策略，更新当前的行动策略时会考虑到未来潜在的奖励，适用于解决长期规划问题，其原理如下：

$$\begin{aligned}&Q\left(s,a^t\right)\leftarrow Q(s,a)+\\&\alpha\left(r+\gamma\max_a Q\left(s',a'\right)-Q(s,a)\right)\end{aligned} \tag{16}$$

其中，$Q(\cdot)$表示由状态s和动作a到Q值的映射，选取当前最大Q值的动作作为最优动作，α是学习率，γ是折扣因子。使用$Q(\cdot)$来预测，在下一状态s'选择最优动作的条件下，智能体可获得的最大回报。

基于上述方式，智能体可以学习在给定状态下选择哪个动作能最大化累积的未来奖励，最终学习到一个近似最优策略。计算目标值的$Q(\cdot)$参数。目标值表示为

$$\hat{Q}(s,a)=r+\gamma\max_{a'}Q\left(s',a';\theta^-\right) \tag{17}$$

Q网络每次更新都会改变目标值，这可能影响参数优化的方向，进而导致训练振荡或直接发散。为此，DQN引入了Target网络$Q_T(\cdot)$，每隔一定的步同步主网络$Q(\cdot)$权重，从而获得一个相对静止的目标值，即

$$\hat{Q}(s,a)=r+\gamma\max_{a'}Q_T\left(s',a';\theta^-\right) \tag{18}$$

DQN使用的损失函数是预测Q值与目标Q值之间的均方误(Mean Squared Error，MSE)，表示为

$$\mathcal{L}(\theta)=\mathbb{E}\left[\left(\hat{Q}(s,a)-Q(s,a;\theta)\right)^2\right] \tag{19}$$

4.2　MA-HGDQN算法

针对LEO卫星切换问题，卫星节点数量通常在数百至数千颗，且随着卫星通信技术的发展，星座规模会持续增大。这导致集中式决策的状态s和动作a的维度不断增大。状态需要包含所有链路的状态信息，即

$$s^t=\left[s_{n,k}^t\right]_{N\times K} \tag{20}$$

其中，s^t表示时间步t下整个卫星网络的状态，$s_{n,k}^t$表示时间步t下卫星n和用户k间链路的状态，包括节点间的相对位置关系$L_{n,k}^t$、信号强度$q_{n,k}^t$、剩余服务时间$\upsilon_{n,k}^t$、卫星节点负载l_n^t和连接状态$c_{n,k}^t$，表示为

$$s_{n,k}^t=\left[L_{n,k}^t\quad q_{n,k}^t\quad \upsilon_{n,k}^t\quad l_n^t\quad c_{n,k}^t\right] \tag{21}$$

其中，$L_{n,k}^t$表示卫星和用户节点之间的动态位置关系，使用用户仰角值α、一阶导数和二阶导数来准确描述，即

$$L_{n,k}^t=\left[\varphi_{n,k}^t\quad \Delta\varphi_{n,k}^t\quad \Delta^2\varphi_{n,k}^t\right] \tag{22}$$

联合动作表示为

$$a^t=\left[a_1^t,a_2^t,\cdots,a_K^t\right] \tag{23}$$

显然，状态和动作空间的尺寸与用户数量和卫星数量均线性相关，难以实际应用于大型LEO卫星网络。

为了降低状态和动作空间的尺寸，一种有效的方法是把决策方式由集中式转化为分布式，将模型调整为多智能体架构。具体而言，考虑K个智能体$k\in\mathcal{K}$(每个用户对应一个智能体)，对应到元组$\{k,S_k,A_k,\mathcal{R}_k,\pi_k\}$。其中，$S_k$和$A_k$分别是$S$和$A$的子集，

表示仅和智能体k相关的状态和动作。所有组件可以定义如下。

(1) 智能体k:每个用户被认为是独立采取动作的智能体,智能体的集合等同于用户集合$\mathcal{K}$。

(2) 状态s_k^t:在时间步t下智能体k由环境得到的当前观测。具体定义见第4.2.1节。

(3) 动作a_k^t:在时间步t下智能体k选择连接到一个卫星节点$n\in\mathcal{N}_k$,其中$\mathcal{N}_k$表示用户的候选卫星集合。具体定义见第4.2.2节。

(4) 奖励函数$\mathcal{R}_k$:使用局部奖励和全局奖励结合的方式,分别评价切换节点的好坏和对网络平衡的影响。具体定义见第4.2.3节。

4.2.1 局部状态

针对单个用户节点的切换决策,星地网络中节点的重要程度不同。例如,处于用户可见范围内的卫星节点与决策具备强关联,而其他卫星节点的关联强度较弱。如果决策过程中考虑大量弱相关节点的特征,一方面会引入噪声,降低模型的拟合速度,另一方面状态空间的过大导致模型无法准确提取稀疏信息,降低决策的准确性。针对这一问题,我们定义关联分数对节点进行重要度排序,筛选节点并构建以决策用户节点为中心的局部网络。由于星地网络随时间显著变化,为了简化星地决策过程,通常按照一定时间间隔构建网络快照。本文把局部网络构建为异质图,并通过异质图注意力网络(Heterogeneous Graph Attention Network, HGAT)网络聚合网络的拓扑和节点信息,作为智能体决策的状态输入,增加算法的性能和可解释性。

(1) 网络节点重要程度排序

对于卫星节点,与决策用户间具有“可见关系”和“连接关系”的节点给予高关联分数,与决策用户相邻节点具有“连接关系”的节点给予低关联分数,其他卫星节点的分数设置为0。

$$\mathrm{Score}_{\mathrm{sat},n}=\begin{cases}v_{high}, \text{节点与决策用户具有可见/连接关系}\\ v_{low}, \text{节点与决策用户的邻居具有连接关系}\\ 0, \text{其他情况下}\end{cases} \tag{24}$$

对于用户节点,与决策用户连接到同一卫星节点的用户节点给予高关联分数,处于决策用户一定距离内的用户节点给予低关联分数,其余用户节点的分数设置为0。

$$\mathrm{Score}_{\mathrm{ue},k}=\begin{cases}v_{high}, \text{节点与决策用户连接到同一卫星节点}\\ v_{low}, \text{节点处于决策用户一定距离内}\\ 0, \text{其他情况下}\end{cases} \tag{25}$$

卫星节点和用户节点分别根据关联分数进行排序,筛选强关联节点构建局部星地网络。

(2) 构建局部网络异质图

根据节点关联分数排序筛选出局部卫星集合$\mathcal{N}_k$(包含V_s个卫星节点);以及用户集合$\mathcal{K}_k$(包含用户k节点和V_u-1个邻居节点)。

用户k的局部网络的状态空间表示为

$$s_k^t=\left[s_{n,i}^t\right]_{\mathrm{V}_s\times\mathrm{V}_u},\ n\in\mathcal{N}_k,\ i\in\mathcal{K}_k \tag{26}$$

上述方式只是把局部网络的信息堆叠构成状态,不包含星地网络节点间复杂的拓扑信息,导致解释性较差。为避免上述问题,本文使用异质图来表示局部网络,将其定义为有向图$\mathcal{G}=(V, E, T_v, T_e)$。其中,$V=V_u\cup V_s$表示节点集合,满足$\forall v\in V$,$E$表示边集合,满足$\forall e\in E$。$T_v$表示节点类型集合,包括“卫星节点”和“用户节点”。卫星节点的特征包括空间位置和卫星节点负载,用户节点的特征包括空间位置和当前信号强度。T_e表示边关系类型集合,具体包括:

① 可见关系:存在于仰角大于设定的最小值的卫星节点和用户节点之间,使用剩余可见时间、信号强度、相对位置关系作为边的特征。

② 连接关系:存在于当前时刻通信的卫星节点和用户节点之间,使用信道系数作为边的特征。

③ 距离关系:存在于距离小于设定的最大值的用户节点之间,使用距离作为边的特征。

(3) 使用异质图表征作为局部状态

节点和节点类型之间存在映射关系:$\Phi(v)$:$V\to T_v$,边和边的类型之间存在映射关系:$\Psi(e)$:$E\to T_e$。在异质图中,使用源节点和目标节点的二元组来定义边$e=(u,v)$,其中$u\in V$表示源节点,$v\in V$表示目标节点,r表示关联类型,定义为三元组$\left(\Phi(u),\Psi(e),\Phi(v)\right)$,例如“可见关系”的关联$r_{visible}$表示为

$$r_{visible}=\left(\text{"卫星节点"},\text{"可见关系边"},\text{"用户节点"}\right) \tag{27}$$

之后,根据不同的关联r用来换划分子图,如图2所示。

HGAT来聚合局部网络异质图中的节点和边信息,从而实现输入状态的高效表征。具体而言,单

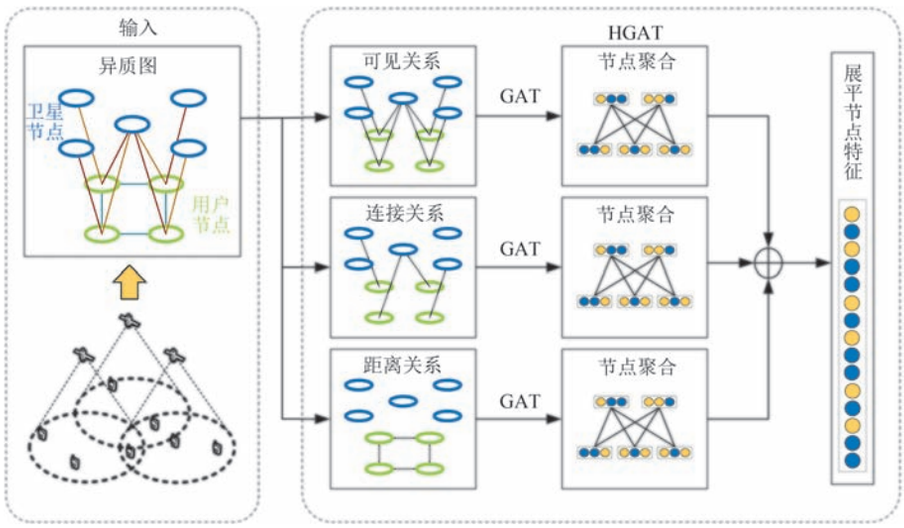

图2　基于多维关联的异质图表征

层HGAT网络会根据关联r把异质图划分为多个子图，分别使用注意力机制聚合邻域节点的特征，再将所有子图中的节点特征求和，单层的节点特征聚合特征为

$$x_i^{(l+1)}=\sum_{r\in\mathcal{R}}x_{r,i}^{(l+1)} \tag{28}$$

其中，$x_{r_n,i}^{(l+1)}$是基于关联r_n的子图使用注意力机制聚合得到的特征聚合，表示为

$$x_{r,i}^{(l+1)}=\sum_{j\in\mathcal{N}(i)\cup\{i\}}\alpha_{r,i,j}\boldsymbol{\Theta}_{r,l}x_j^{(l)} \tag{29}$$

其中，$x_j^{(l)}$表示第l层节点j的特征，注意力系数$\alpha_{i,j}$可计算[28]为

$$\alpha_{r,i,j}=\frac{\exp\left(a^{\mathrm{T}}\mathrm{LeakyReLU}\left(\boldsymbol{\Theta}_s x_i+\boldsymbol{\Theta}_t x_j+\boldsymbol{\Theta}_e e_{r,i,j}\right)\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}}\exp\left(a^{\mathrm{T}}\mathrm{LeakyReLU}\left(\boldsymbol{\Theta}_s x_i+\boldsymbol{\Theta}_t x_k+\boldsymbol{\Theta}_e e_{r,i,k}\right)\right)} \tag{30}$$

其中，$e_{i,j}$是以节点i为起点和以节点j为终点的边的特征，$\boldsymbol{\Theta}_s$、$\boldsymbol{\Theta}_t$和$\boldsymbol{\Theta}_e$均是可学习的参数矩阵。通过注意力权重的计算过程中引入边特征，实现节点特征和边特征的聚合。

用户节点的二阶邻域包括与用户节点相关的卫星节点以及与卫星节点相关的其他用户节点。通过使用不少于2层的HGAT进行消息传递和聚合，节点v包含二阶邻域内所有节点的信息，可以描述局部网络的状态。使用更新后的所有卫星节点和用户k的特征展平处理得到智能体k的输入状态s_k^t。由于更新机制的复杂性，HGAT网络训练产生的算力和时间成本远高于DQN网络，可先通过无监督或自监督方式进行预训练，然后将嵌入固定用于DQN训练，从而应用于计算资源受限的星地网络。

4.2.2　动作空间

使用独热向量(One-hot Vector)表示智能体k的动作a_k^t，“1”表示用户k连接到的卫星节点，向量维度是$1\times\mathcal{N}_k$，动作空间A_k中包含$\mathcal{N}_k$个动作。联合动作表示为$a^t=(a_1^t,a_2^t,\cdots,a_K^t)\in A$，联合动作空间表示为$A=A_1\times\cdots\times A_k\times\cdots\times A_K$。

智能体k根据当前Q值选择最佳动作，即用户k选择最优卫星接入点。然而，始终选择最优动作可能导致探索不足，进而陷入局部最优。本文采用ε-greedy 策略，在前期尝试更多先前未探索的动作，探索更广阔的动作空间。

$$\pi\left(a_k^t\middle|s_k^t\right)\leftarrow\begin{cases}1-\varepsilon+\varepsilon/\left|A_k\left(s_k^t\right)\right|, & a\in\{a^*\}\\ \varepsilon/\left|A_k\left(s_k^t\right)\right|, & a\in A_k\backslash\{a^*\}\end{cases} \tag{31}$$

其中，ε是探索率，a^*是当前状态的最优动作。智能体以概率ε随机选择动作(探索)，以概率$1-\varepsilon$选择当前已知价值最大的动作(利用)。

训练初期需要更多探索，随着对环境了解增加应逐渐减少探索概率ε。本文将初始探索概率设置为0.9，并随着训练次数的增加指数衰减到0.01。

4.2.3　奖励函数

为实现多目标优化中的三重目标(最小化切换次数、避免通信中断、保持负载平衡)，本文使用长短期奖励结合的方式设置奖励函数：

$$r_k^t=R_{global}\left(s^t,a^t\right)+R_{local}\left(s_k^t,a_k^t\right) \tag{32}$$

其中，R_{global}是评估系统整体性能的全局奖励函数，根据系统负载平衡情况计算奖励，R_{local}是评估本次动作的局部奖励函数，对过早切换和连接中断的情

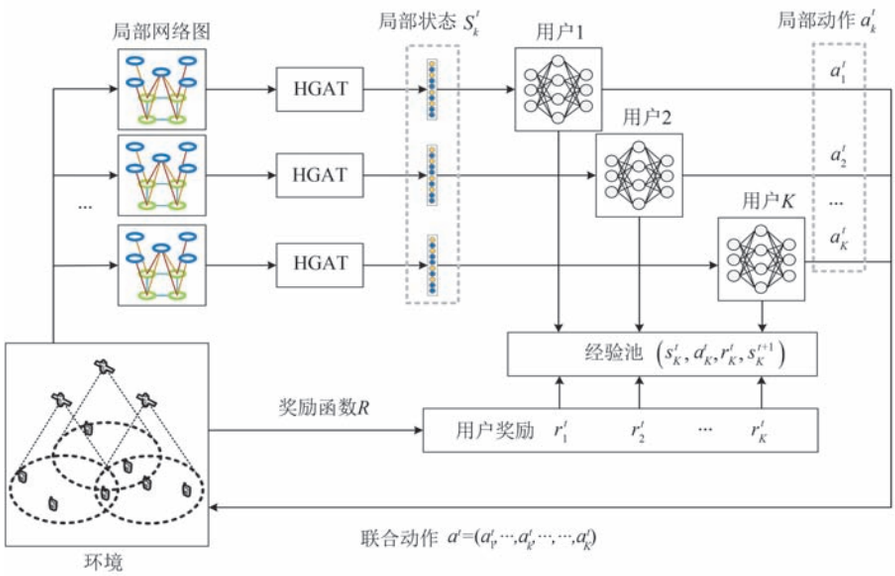

图3 基于异质图表征的多智能体协作切换模型

况进行惩罚。

R_{local}用于评价本次切换动作的好坏，仅与智能体k的状态s_k^t和动作a_k^t有关，设置方式如下：

$$R_{local}(s_k^t, a_k^t)=\begin{cases} value_p, \text{用户不切换卫星节点，且连接未中断} \\ f\left(r_{m,n}^{(k-1)}, r_{m,n}^{(k)}, l_{m,n}^{(k-1)}, r_{m,n}^{(k)}, c_{m,n}^{(k-1)}, c_{m,n}^{(k)}\right), \text{用户切换卫星节点，且连接未中断} \\ value_n, \text{连接中断} \end{cases} \tag{33}$$

其中，$value_p$表示正值，用于奖励用户和卫星间的保持连接；$value_n$表示负值，用于惩罚用户无法正常连接到选中的卫星节点；当用户切换卫星且连接未中断的情况下，需要根据环境综合判断切换节点带来的性能增益和切换成本，生成奖励值。

R_{global}用于促进智能体间的合作[29]，本文使用网络节点的负载平衡程度定义联合奖励，常规方式是所有智能体共享一个全局奖励，设置方式如下：

$$R_{global}(s^t, a^t)=\frac{1-4B_{load}}{1+4B_{load}} \tag{34}$$

其中，$B_{load}\in[0, 0.25]$是所有可见卫星节点负载的方差，通过公式(34)将B_{load}的值映射到$[0,1]$之间，定义为系统负载平衡系数，值随系统负载方差的减小而增加。

然而，直接将全局奖励分配给各智能体进行学习，无法明确判断各个智能体的行为对该全局奖励贡献的大小。随着智能体增多，全局奖励与单个智能体的行为之间的关联愈发稀薄。针对这一问题，需要引入信用分配。

为了精确量化每个智能体对团队回报的贡献，研究者们提出了来源于合作博弈论的Shapley值。对给定联合动作a^t下的全局回报$R_{global}(s^t, a^t)$，Shapley值的定义为

$$\phi_k=\frac{1}{k!}\sum_{\mathbb{S}\subseteq K\setminus\{k\}}|\mathbb{S}|!(K-|\mathbb{S}|-1)!\left[R(\mathbb{S}\cup\{k\})-R(\mathbb{S})\right] \tag{35}$$

其中，$\mathbb{S}$是除了智能体k以外的部分或全部智能体构成的集合，$R(\mathbb{S})$表示集合$\mathbb{S}$中所有智能体采取联合动作，其余智能体取某基线动作时的全局奖励。

由于需要遍历所有子集或排列，Shapley值的计算复杂度为指数级，在智能体较多的情况下难以实际应用。作为替代方案，使用蒙特卡洛法采样子集，近似估计Shapley值，改进后的公式如下

$$\phi_k\approx\frac{1}{U}\sum_{u=1}^{U}\left[R(\mathbb{S}_m\cup\{k\})-R(\mathbb{S}_m)\right] \tag{36}$$

随机抽样U次，每次的智能体子集为$\mathbb{S}_m$，智能体k的边际贡献是子集加入k前后的回报的差值的平均值。使用ϕ_k代替公式(32)中的R_{global}，智能体k奖励函数改进为

$$r_k^t=\phi_k(s^t, a^t)+R_{local}(s_k^t, a_k^t) \tag{37}$$

4.3 训练和执行方式

在具有部分可观察性和高协调要求的环境中，

集中培训与分散执行（Centralized Training with Decentralized Execution，CTDE）是最好的选择[27]。在CTDE框架下，策略网络或价值网络参数共享在理论和实践上都具有优势。

同构智能体是应用参数共享的重要前提，即各智能体具有相同的动作空间、观测维度和目标函数。对于LEO星座的切换决策问题，每个用户（智能体）面临相同类型的决策，即从候选卫星节点集合中选择最佳卫星节点；具有类似的目标，即减少切换次数、提升负载平衡度。尽管不同用户所处位置和信道状态不同，但其决策结构和奖励机制可以被设计为相同，这意味着其属于“同质”或“部分同质”的智能体。在此假设下，共享参数的策略网络或Q网络是合理的。

本文提出的MA-HGDQN架构见图3，其中所有智能体共享一套DQN参数。所有智能体的交互经验被视为等价且可共用的训练样本，丰富了经验回放池中的样本多样性，避免单个智能体“切换”经验匮乏的问题，从而提高样本效率和收敛速度。

为了进一步增加“切换事件”在经验池中的比例，我们定义了“预切换”状态，包括以下两种情况：(1)用户的剩余服务时间不足10秒；(2)用户上一时刻连接中断。在满足“预切换”条件时，智能体根据自身观测的局部网络状态s_k^t执行独立动作a_k^t；其他时刻保持连接状态不变，节约系统算力。

智能体将“切换事件”的经验经由卫星节点统一反馈至地面上具备充足算力的服务器，服务器按照一定周期将训练好的神经网络参数反馈给用户终端，优化智能体的切换决策。

4.4　算法流程

算法1.　MA-HGDQN切换决策算法

输入：仿真总时隙数T，用户数量K，用户位置，卫星星历数据。每个episode包含的时隙数N_e，训练轮数J，学习率α

输出：累积切换次数N_h、连接中断次数N_f、系统负载失衡度B_{load}

BEGIN
初始化经验池
FOR $e=1,2,\cdots,N_e$ DO
　$N_h=0, N_f=0$
　用户随机初始接入
　FOR $t=1,2,\cdots,T$ DO
　　FOR $k=1,2,\cdots,K$ DO
　　IF 用户k进入“预切换”状态
　　根据4.2.1节构建局部网络异质图
　　根据4.2.1节使用HGAT聚合特征
　　$a_k^t=\arg\max_a Q(s_k^t,a)$
　　$r_k^t=\phi(s^t,a^t)+R_{local}(s_k^t,a_k^t)$
　　存储经验
　　END IF
　END FOR
　计算并记录系统负载失衡度B_{load}
　FOR $k=1,2,\cdots,K$ DO
　　IF 用户k连接的卫星节点发生变化
　　$N_h=N_h+1$
　　END IF
　　IF 用户k连接中断
　　$N_f=N_f+1$
　　END IF
　END FOR
　记录当前episode的N_h和N_f
　FOR $j=1,2,\cdots,J$ DO
　　随机抽取经验
　　训练智能体
　END FOR
END FOR

算法1的第4～30步是一个连续情节（episode）的完整流程。其中，第5步表示随机生成episode的初始状态。第8～14步是智能体切换决策的具体流程，通过第8步的“预切换”判断机制，在调整“切换”经验占比的同时，大幅缩减MA-HGDQN算法的仿真用时。第16～23步是记录切换性能的评价指标，包括系统负载方差、累积切换次数和累积中断次数。第27～30步是智能体的训练过程，使用随机抽取经验训练一套DQN参数，之后同步到所有智能体的DQN部分。

5　仿真与性能分析

本节先介绍实验环境设置，包括实验数据集、对比方法、超参数设置等。然后从算法的拟合效果、模型性能、切换频率、中断频率以及负载平衡度等角度评估本文提出的基于异质图表征的多智能体协作切换算法MA-HGDQN，并通过消融实验验证通过异质图表征提取网络拓扑信息的有益性。最后，从实际应用的角度，评估了算法对不同卫星系统之间的兼容性。

5.1　实验设置

本实验的代码运行环境基于STK、Python 3.9、PyTorch 2.10.0及CUDA 11.8构建，硬件配置采

用Intel Xeon Silver 4210R CPU(主频2.39 GHz,2个处理器)、NVIDIA RTX 3080显卡、64 GB内存。仿真系统将星地网络建模为等间隔的网络切片,并在此基础上开展离散事件仿真。其中,智能体训练的连续情节数量(Number of Episode,NoE)设置为200,每个episode对应300 s的仿真过程,星地网络切片之间的间隔设置为1 s,星地之间的信道根据3GPP的信道模型[30-31]生成,地面用户设置为在指定区域内随机分布,并在每个episode中重新生成。为了证明所提出的策略的效率,对比以下算法的仿真性能:

(1) 最低负载策略(Load):单准则切换策略,当用户进入"预切换"状态后,用户切换到可见范围内负载最低的卫星。

(2) 最短距离策略(Distance):单准则切换策略,当用户进入"预切换"状态后,用户切换到可见范围内空间距离最近的卫星。

(3) A2C:基于策略梯度的方法,采用Actor-Critic结构进行联合训练。A2C将环境状态作为输入,按时间切片分别生成切换动作的概率分布(策略)和状态值,利用优势函数(Advantage)进行策略更新。

(4) MA-DQN:根据局部网络生成状态空间,再输入DQN隐式学习状态间的关联,进行切换决策。DQN仅在用户进入"预切换"状态时进行切换决策,其余时刻保持连接状态不变。

(5) MA-HGDQN:根据卫星节点和用户节点间可见关系、连接关系、距离关系构建局部网络的异质图数据,使用HGAT模块提取网络信息,再输入DQN进行切换决策。MA-HGDQN仅在用户进入"预切换"状态时进行切换决策,其余时刻保持连接状态不变。

仿真系统的详细参数设置参照表2。

表2 仿真系统参数

参数名称	数 值
最小仰角值	30°
有效全向辐射功率	43 dBW
用户接收天线增益	29 dBi
用户数量	30
用户区域	以(0°N, 0°W)为中心 半径为2°的球面区域内
系统带宽	20 MHz
NoE	200
episode长度	300 s
时隙间隔	1 s

卫星星座分布参考Starlink星座,设置多种星座架构用于评估算法的适应性,用于评估的星座配置参数见表3。

表3 星座配置参数

	基准星座	星座2	星座3
轨道数	72	36	72
每轨卫星数	22	22	44
轨道倾角	53°	53°	53°
轨道高度	550 km	550 km	550 km

智能体训练的超参数配置见表4。

表4 超参数配置

	HGAT	DQN	A2C
学习率	-	0.0001	0.01
batch size	-	64	30
γ	-	0.9	0.9
输入维度	节点特征:7 边特征:6/3/1	MA-HGDQN:96 MA-DQN:35	35
隐藏层神经元数量	[64, 64, 16]	[256, 256, 64]	[256, 64]
输出维度	96	5	actor:5 critic:1

5.2 算法性能评估

5.2.1 拟合性能

本节分别评估了基于值和基于策略的DRL算法的拟合性能。其中,基于值的DRL算法包括本文提出的MA-HGDQN算法和MA-DQN算法,两者的区别在于是否使用HGAT模块聚合网络信息作为状态输入,且仅在用户进入"预切换"状态时进行动作决策。基于策略的DRL算法采用A2C算法,智能体在每个时间片上进行动作决策。

图4展示了不同算法在训练过程中的累积奖励值,其中累积奖励定义为每个episode内所有智能体的奖励值总和。如图4(a)所示,MA-HGDQN与MA-DQN两种基于值函数的方法在约10个episode后逐渐收敛,表现出较快的拟合速度。由于此类方法仅针对"预切换"阶段实施动作决策,故其累积奖励值能直接地反映算法在切换性能上的优劣。从模型收敛后的表现来看,MA-HGDQN的累积奖励值相比MA-DQN提高约2000,表明其在切换决策质量上具有更强优势。相比之下,A2C作为基于策略的方法,在约20个episode后逐步趋于稳定,拟合所需的训练轮数为基于值函数方法的两倍。此外,得益于基于值函数的方法仅需在少量的时间步上完成

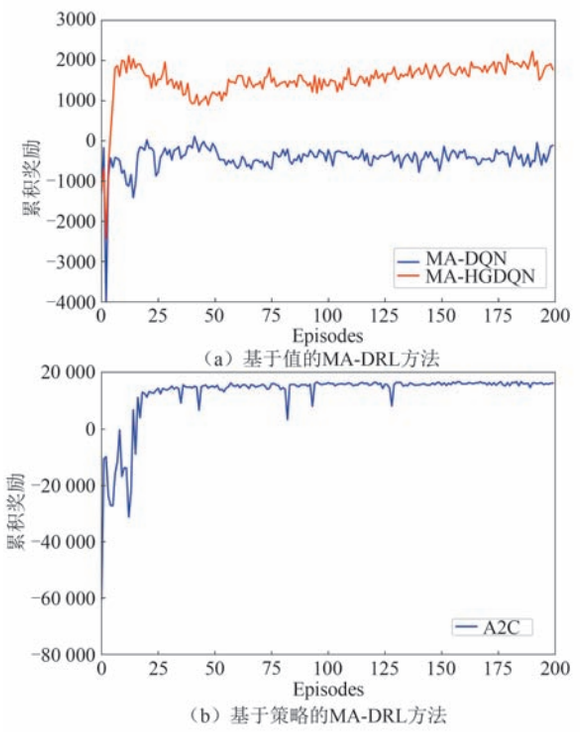

图4　强化学习算法的累积奖励

动作决策分析。如表5所示，MA-HGDQN的仿真用时相比A2C减少了约88%，展现出更高的训练效率。

表5　单Episode仿真用时

	仿真用时
MA-HGDQN	1分10秒
MA-DQN	1分6秒
A2C	10分10秒

综上所述，本节对所评估智能算法的拟合效果与训练效率进行了系统分析。结果表明，两种基于值函数的方法（MA-HGDQN和MA-DQN）在提升收敛速度与降低计算开销方面具有明显优势。其中，MA-HGDQN通过引入异质图表征，有效增强了状态建模能力，实现了累积奖励的显著提升，展现出更优的训练性能与决策效果。

5.2.2　切换决策性能

本节研究了不同算法的切换决策性能，主要评估以下三项指标：累积切换次数、累积中断次数、系统负载平衡程度。其中，累积切换次数是所有用户在单个episode中的切换总次数，用于评估切换动作的频率；累积中断次数是所有用户在单个episode中的连接中断总次数，用于评估用户切换卫星节点的优劣；系统负载平衡程度使用公式（34）定义的系统负载平衡系数作为量化指标，用于评估算法能否均衡分配卫星节点的接入资源。

鉴于频繁切换对通信连续性的负面影响，本节首先在相同环境配置下对比分析了不同算法的累积切换次数表现。图5展示了各算法在训练过程中累积切换次数随episode变化的趋势。约经过50个episode后，各算法的累积切换次数趋于稳定，其中MA-HGDQN的性能波动显著低于其余四种对比算法，表现出更好的稳定性。为进一步量化算法的长期性能，计算了后150个episode内的平均累积切换次数，结果如图6所示。所提MA-HGDQN算法取得了最低的平均累积切换次数，较MA-DQN算法减少了10.3%。

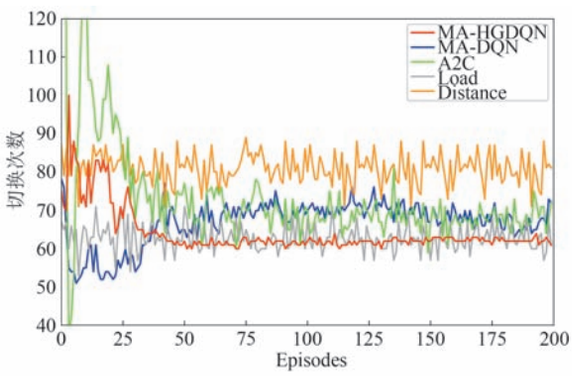

图5　累积切换次数

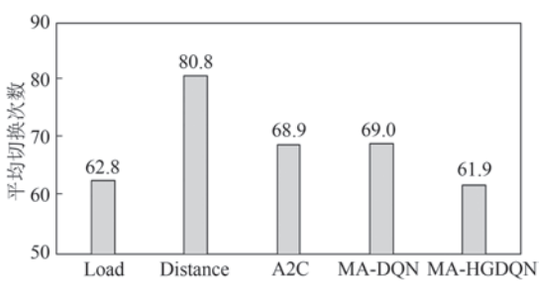

图6　平均累积切换次数

为了评估切换决策中卫星节点选择的合理性，本文进一步比较了不同算法在相同环境配置下的累积通信中断次数表现。图7中“Load”和“Distance”曲线显示，基于属性切换的算法在应对多因素导致的连接中断时效果较差，其中最短距离策略的平均中断次数甚至超过MA-DRL方法的150倍以上。对于三种MA-DRL算法，基于值函数的MA-HGDQN与MA-DQN在约50个episode后性能趋于稳定，而基于策略的A2C算法则需约100个episode才能达到相对稳定的表现。如图8所示，从平均性能角度来看，基于值函数的MA-DRL方法

(MA-HGDQN和MA-DQN)通过针对“切换事件”经验的重点训练,显著优于基于策略的方法,平均中断次数较A2C减少超过80%。此外,借助异质图表征技术,MA-HGDQN较MA-DQN进一步降低了45.8%的平均累积中断次数,彰显出其在提升切换决策质量方面的显著优势。

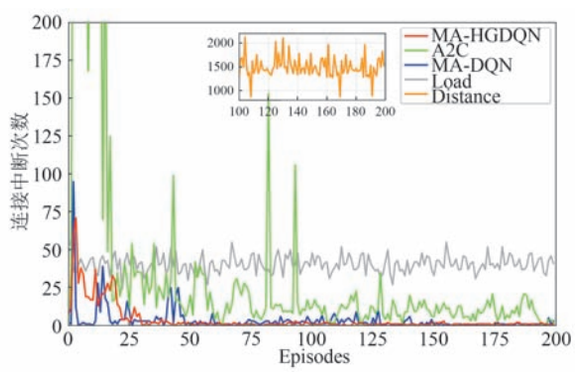

图7 累积中断次数

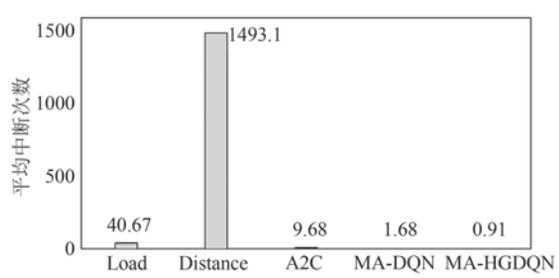

图8 平均累积中断次数

最后,针对系统负载平衡性能进行分析,以验证多智能体间的协作效果。由于各算法初始状态存在差异,无法直接比较实时负载性能,故采用多个时间点系统负载方差的累积分布函数(Cumulative Distribution Function,CDF)曲线进行性能展示,如图9所示。结果表明,所提MA-HGDQN算法的系统负载方差整体低于MA-DQN,且与A2C表现相近。为进一步量化性能,采用公式(34)将系统负载方差的均值转换为系统平衡系数,数值越大代表系统负载分布越均衡。图10显示,除专门针对负载优化的最低负载策略外,MA-HGDQN与A2C算法表现最佳,两者在其他性能指标上均显著优于最低负载策略。引入异质图表征后,MA-HGDQN实现了约11.4%的系统平衡性能提升。

综上所述,所提MA-HGDQN算法在累积切换次数、累积通信中断次数及系统负载均衡三个关键指标上均展现出显著优势。异质图表征的消融实验

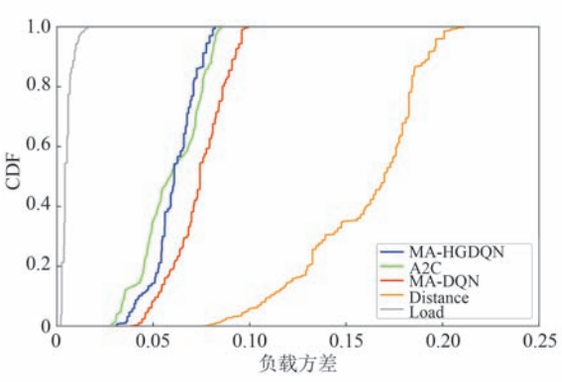

图9 系统负载方差(CDF)

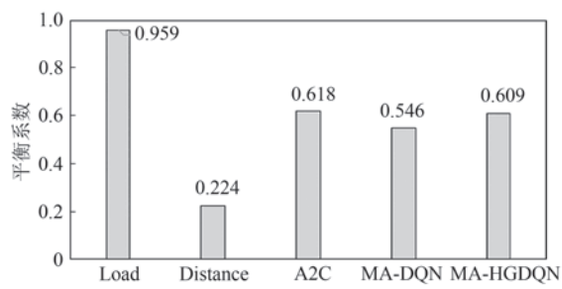

图10 系统负载平衡系数

进一步验证了其在提升切换性能方面的关键作用。具体而言,MA-HGDQN较基线方法分别减少了10.3%的累积切换次数、45.8%的累积中断次数,并提升了11.4%的系统负载公平性,充分体现了该方法在多智能体协同切换优化中的有效性与优越性。

5.3 实际应用的适用性分析

5.3.1 MA-HGDQN算法复杂度分析

(1) DQN复杂度

智能算法的复杂度主要取决于神经网络的复杂程度,即隐藏层的数量以及隐藏层中神经元的数量。对于DQN模型,其复杂度可以表示为

$$\Psi_{\mathrm{DQN}}=O\left(\tilde{S}\cdot W_1+\sum_{l=1}^{L-1}W_lW_{l+1}+W_L\cdot\tilde{A}\right) \tag{38}$$

其中,$\tilde{S}$表示状态空间S包含的特征数量,$\tilde{A}$表示输出动作数量,W_l表示第l个隐藏层包含的神经元数量。

如果使用单智能体进行联合切换决策,输入维度表示为$M\cdot N\cdot\tilde{S}_{m,n}$,输出维度表示为$M\cdot N$,系统整体的复杂度表示为

$$\Psi_{\mathrm{SA-DQN}}=O\left(M\cdot N\cdot\tilde{S}_{m,n}\cdot W_1+\sum_{l=1}^{L-1}W_lW_{l+1}+W_L\cdot M\cdot N\right) \tag{39}$$

为了保证神经网络性能，神经元的数量需要根据输入规模进行调整。随着网络整体规模的扩大，系统的计算复杂度会快速提升，难以实际应用在大规模星地网络中。

如果使用多智能体进行切换决策，使用局部网络筛选节点信息后，系统整体的复杂度表示为

$$\Psi_{\mathrm{MA-DQN}}=O\left(N\cdot\left(\tilde{S}_{m,n}\cdot W_1+\sum_{l=1}^{L-1}W_lW_{l+1}+W_{\mathrm{L}}\cdot M_n\right)\right) \tag{40}$$

在不改变局部网络规模的前提下，MA-DQN结构可保持不变，单个智能体的计算复杂度不变，系统整体的复杂度与用户节点的数量的关系近似为线性。

(2) HGAT复杂度

HGAT聚合局部网络节点特征的计算复杂度可表示为

$$\Psi_{\mathrm{HGAT}}=O\left(\sum_{l=1}^{L}\sum_{v\in\mathcal{M}_n\cup\mathcal{N}_n}\sum_{r\in\mathcal{R}}F_{v,r,l}\cdot E_{v,r}\right) \tag{41}$$

其中，$F_{v,r,l}$表示节点v在第l层和关联r下的源节点特征数量，$E_{v,r}$表示节点v在关联r下的边特征数量。

系统中所有HGAT的复杂度可表示为

$$\Psi_{\mathrm{HGAT},total}=O(N\cdot\Psi_{\mathrm{HGAT}}) \tag{42}$$

可见，在不改变局部网络规模的前提下，单个智能体内HGAT更新局部网络信息的计算复杂度不会发生明显变化，HGAT在整个系统中的计算复杂度与用户节点的数量间的关系近似为线性。

(3) MA-HGDQN复杂度

在所提MA-HGDQN算法中，每个智能体的动作决策计算复杂度可视为单个HGAT模块与DQN模块复杂度的叠加。在局部网络规模固定的前提下，单个智能体的计算复杂度相对稳定，整个系统的计算复杂度与智能体的数量之间呈线性正比关系。

5.3.2　MA-HGDQN决策用时评估

本节针对不同星座规模及用户配置，评估了MA-HGDQN算法中单个智能体动作决策的计算时延，以验证第5.3.1节中对算法计算复杂度的理论分析。

表6展示了MA-HGDQN算法中单个智能体单次切换决策的计算时延稳定维持在40 ms左右。该结果验证了所提算法的实际计算开销在用户终端数量及卫星星座规模扩大时保持相对恒定，支持其在大规模卫星网络中的可行性和应用潜力。

表6　不同环境配置下单次动作决策用时（单位ms）

	10用户	20用户	30用户
基准星座	40.9	40.7	41.5
星座2	39.7	41.1	40.6
星座3	37.4	39.8	41.3

5.3.3　算法适应性评估

为验证所提算法在实际应用场景中的适应性，本节针对多种星座拓扑结构及不同规模的用户终端，测试了MA-HGDQN算法的性能表现。

图11和图12分别展示了不同环境配置下的累积切换次数与累积中断次数。经过约50个训练episode后，累积切换次数趋于稳定，累积中断次数接近于零，表明算法能够实现稳定的切换频率，并保证切换到满足通信需求的卫星节点。所提MA-HGDQN算法展现出良好的适应性，能够有效应对不同星座架构和用户规模的变化。

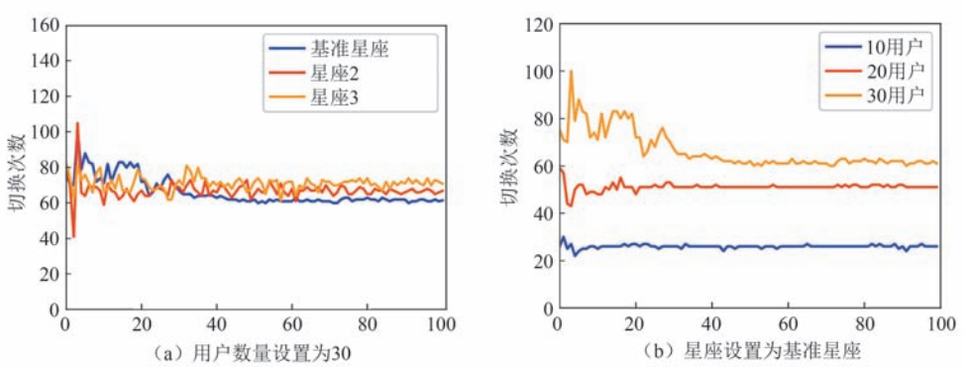

图11　不同环境配置下的累积切换次数

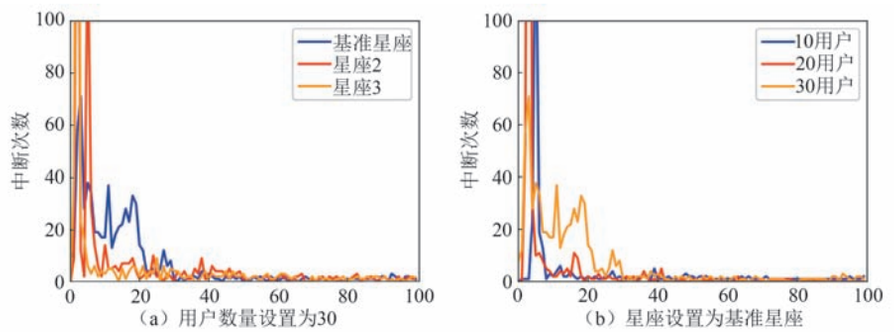

图12 不同环境配置下的累积中断次数

6 结 论

本文提出了一种基于异质图表征的多智能体协作切换方法MA-HGDQN，用于解决星地融合网络中频繁切换所带来的通信中断与负载失衡问题。该算法融合图神经网络与强化学习的优势，提升切换决策的智能性与系统整体性能。具体而言，针对强化学习算法在提取网络特征时难以充分利用网络拓扑与先验结构信息的问题，本文引入异质图表征思想，将星地融合网络建模为包含多类型节点与关系的异质图数据结构，并基于HGAT提取融合拓扑语义的低维向量表示，作为各智能体DQN的状态输入，从而提升模型的表达能力与收敛性能。在此基础上，本文构建了一种面向星地网络切换场景的多智能体协同机制，一方面引入全局奖励分配机制，量化个体智能体对系统性能的边际贡献；另一方面针对同质智能体实行经验和参数共享策略，促进策略的一致性与泛化能力。通过上述设计，不仅能够实现最优的个体通信效能，还能在系统层面协同优化网络负载分布，实现通信连续性与系统公平性的统一。最后，通过仿真实验评估结合上述技术的MA-HGDQN算法性能。在拟合性能方面，所提算法着重分析“预切换”阶段，在提高“切换”经验占比的同时大幅节约算力，相比基于策略的A2C算法仿真用时减少88%。在切换性能方面，所提MA-HGDQN显著优于其他对比算法，仅在系统负载均衡性能上略逊色于最低负载策略。对比未引入异质图表征的MA-DQN算法，MA-HGDQN算法能够减少10%的累积切换次数，减少45%的累积中断次数，负载公平性提升11%以上。

针对在实际应用场景中的适应性，本文首先分析了在多智能体架构下，单个智能体中HGAT与DQN模块的计算复杂度随用户终端数量与星座规模增长的变化趋势。理论分析与仿真结果均表明，其计算开销在大规模网络场景下依然保持稳定，智能体动作决策的平均计算时延维持在约40 ms，具备良好的实时响应能力。进一步地，本文在多种卫星星座构型与用户规模配置下，对算法的切换决策性能进行了系统测试，结果显示所提方法能够有效确保用户接入满足通信服务质量要求的卫星节点，并维持稳定的切换频率。这证明了算法在复杂动态环境中的鲁棒性与适应性。需要指出的是，本文的实验设置聚焦于小范围地理区域内的地面用户分布，尚未充分考虑大范围区域内用户分布的空间异质性与聚集性特征。因此，未来研究将致力于进一步拓展理论模型与算法适应性，以应对更复杂场景下的星地融合协同通信需求。

参 考 文 献

[1] Chen Shan-Zhi, Sun Shao-Hui, Kang Shao-Li, et al. Key technologies for 6G integrated satellite-terrestrial mobile communication. Science China: Information Sciences, 2024, 54(05): 1177-1214 (in Chinese)
(陈山枝, 孙韶辉, 康绍莉, 等. 6G星地融合移动通信关键技术. 中国科学: 信息科学, 2024, 54(05): 1177-1214)

[2] Zhang Hai-Jun, Chen An-Qi, Li Ya-Bo, et al. Key technologies of 6G mobile network. Journal of Communications, 2022, 43(07): 189-202 (in Chinese)
(张海君, 陈安琪, 李亚博, 等. 6G移动网络关键技术. 通信学报, 2022, 43(07): 189-202)

[3] Chen Shan-Zhi, Sun Shao-Hui, Kang Shao-Li. System integration of terrestrial mobile communication and satellite communication—the trends, challenges and key technologies in B5G and 6G. China Communications, 2021, 17(12): 156-171

[4] Toyoshima M. Recent trends in space laser communications for small satellites and constellations. Journal of Lightwave Technology, 2021, 39(3): 693-699

[5] Ma Ting, Qian Bo, Qin Xiao-Han, et al. Satellite-terrestrial integrated 6G: An ultra-dense LEO networking management

architecture. IEEE Wireless Communications, 2022, 31(1): 62-69

[6] Juan E, Lauridsen M, Wigard J, et al. 5G New Radio mobility performance in LEO-based non-terrestrial net-works// Proceedings of the 2020 IEEE Globecom Workshops (GC Wkshps). Taiwan, China, 2020: 1-6

[7] Wang Zhi, Li Li-Hua, Xu Yue, et al. Handover control in wireless systems via asynchronous multiuser deep reinforcement learning. IEEE Internet of Things Journal, 2018, 5(6): 4296-4307

[8] Chowdhury P K, Atiquzzaman M, Ivancic W. Handover schemes in satellite networks: State-of-the-art and future research directions. IEEE Communications Surveys & Tutorials, 2006, 8(4): 2-14

[9] Del Re E, Fantacci R, Giambene G. Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks. IEEE Journal on Selected Areas in Communications, 2002, 13(2): 397-405

[10] Maral G, Restrepo J, Del Re E, et al. Performance analysis for a guaranteed handover service in an LEO constellation with a" satellite-fixed cell" system. IEEE Transactions on Vehicular Technology, 1998, 47(4): 1200-1214

[11] Del Re E, Fantacci R, Giambene G. Handover queuing strategies with dynamic and fixed channel allocation techniques in low earth orbit mobile satellite systems. IEEE Transactions on Communications, 1999, 47(1): 89-102

[12] Zhou Di, Sheng Min, Li Jian-Dong, et al. Aerospace integrated networks innovation for empowering 6G: A survey and future challenges. IEEE Communications Surveys & Tutorials, 2023, 25(2): 975-1019

[13] Li Feng, Wan Qiu-Hua, He Qi-En, et al. An improved many-objective evolutionary algorithm for multi-satellite joint large regional coverage. IEEE Access, 2023, 11: 45838-45849

[14] Zhang Sen-Bai, Liu Ai-Jun, Liang Xiao-Hu. A multi-objective satellite handover strategy based on entropy in LEO satellite communications//Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC). Chengdu, China, 2020: 723-728

[15] Qin Xiao-Han, Ma Ting, Zhang Xin, et al. A lightweight hierarchical mobility management architecture for ultra-dense LEO satellite network//Proceedings of the ICC 2023-IEEE International Conference on Communications. Rome, Italy, 2023: 679-684

[16] Wu Zhao-Feng, Jin Feng-Lin, Luo Jian-Xin, et al. A graph-based satellite handover framework for LEO satellite communication networks. IEEE Communications Letters, 2016, 20(8): 1547-1550

[17] Hu Xin, Song Hang-Yu, Liu Shuai-Jun, et al. Velocity-aware handover prediction in LEO satellite communication networks. International Journal of Satellite Communications and Networking, 2018, 36(6): 451-459

[18] Feng Lang, Liu Yi-Fei, Wu Liang, et al. A satellite handover strategy based on MIMO technology in LEO satellite networks. IEEE Communications Letters, 2020, 24(7): 1505-1509

[19] Hozayen M, Darwish T, Kurt G K, et al. A graph-based customizable handover framework for LEO satellite networks// Proceedings of the 2022 IEEE Globecom Workshops (GC Wkshps). Rio de Janeiro, Brazil, 2022: 868-873

[20] Lv Xi-Yu, Wu Shao-Hua, Li Ai-Min, et al. A weighted graph-based handover strategy for aeronautical traffic in LEO SatCom networks. IEEE Networking Letters, 2022, 4(3): 132-136

[21] Liu Hao-Tian, Wang Yi-Chen, Li Pei-Xuan, et al. A multi-agent deep reinforcement learning-based handover scheme for mega-constellation under dynamic propagation conditions. IEEE Transactions on Wireless Communications, 2024, 23(10): 13579-13596

[22] Wang Jie, Mu Wei-Qing, Liu Ya-Nan, et al. Deep reinforcement learning-based satellite handover scheme for satellite communications//Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP). Changsha, China, 2021: 1-6

[23] Liu Ren-Peng, Wang Xiao-Zhe, Hu Bo, et al. Twin Adaptive Distributional Reinforcement Learning for Robust Handover Control in LEO Satellite Networks. IEEE Wireless Communications Letters, 2025, Early Access

[24] Li Ning, Gong Bin, Deng Zhong-Liang. A handoff algorithm based on parallel fuzzy neural network in mobile satellite networks. Journal of Communications, 2017, 12(7): 395-404

[25] Ali I, Al-Dhahir N, Hershey J E. Predicting the visibility of LEO satellites. IEEE Transactions on Aerospace and Electronic Systems, 1999, 35(4): 1183-1190

[26] Sutton R. S., Barto A. G.. Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 2018

[27] Wang Yi-Qin, Wang Yu-Feng, Tian Feng, et al. Intelligent games meeting with multi-agent deep reinforcement learning: a comprehensive review. Artificial Intelligence Review, 2025, 58(6): 165

[28] Brody S, Alon U, Yahav E. How attentive are graph attention networks?. arXiv preprint arXiv:2105.14491, 2021

[29] Agogino A K, Tumer K. A multiagent approach to managing air traffic flow. Autonomous Agents and Multi-Agent Systems, 2012, 24(1): 1-25

[30] 3rd Generation Partnership Project (3GPP). Study on new radio (NR) to support non-terrestrial networks (NTN). 3GPP TS 38.811, version 15.4.0, 2020

[31] 3rd Generation Partnership Project (3GPP). Study on channel model for frequencies from 0.5 to 100 GHz. 3GPP TS 38.901, version 17.1.0, 2023

FU Yi-Yang, Ph. D. candidate. His research interests include satellite-terrestrial integrated networks, mobility management, deep reinforcement learning and generative AI.

HU Bo, Ph. D., professor. His research interests include satellite-terrestrial integrated communications, low-altitude intelligent network, network artificial intelligence and ubiquitous mobile computing.

LIU Ren-Peng, Ph. D. candidate. His research interests include satellite-terrestrial integrated networks, mobility management, deep reinforcement learning and generative AI.

CHEN Shan-Zhi, Ph. D., professor. His research interests include B5G/6G mobile communication network architectures, satellite-terrestrial integrated communications and vehicular communication networks.

Background

BackgroundIn recent years, the integrated satellite and terrestrial systems have been widely discussed by industries and academics. Driven by the demand for ubiquitous coverage, massive connectivity, and high data rates in the forthcoming 6G era, communication networks aim to merge terrestrial infrastructure, airborne platforms, and satellite constellations, creating a multi-layer communication network architecture. A central and persistent challenge in managing the highly dynamic and heterogeneous networks is mobility management, specifically the handover control.

In this paper, we summarize the existing handover control methods in satellite-terrestrial integrated networks (STINs) and conduct an in-depth comparison of the strengths and weaknesses of various methods. Existing multi-agent deep RL algorithms, such as MA-DQN, often suffer from noisy local observations and unstable policy convergence due to implicit learning of inter-node relationships. To address this issue, we proposed an intelligent handover control method based on heterogeneous graph representation and multi-agent collaboration. Heterogeneous graphs can simultaneously characterize multiple types of nodes (such as satellites, ground stations, and users) in STINs and their heterogeneous relationships (such as inter-satellite links and satellite-to-ground links), thereby more comprehensively reflecting the complex topological structure and multi-dimensional interaction characteristics of the network. The representation vector generated by the heterogeneous graph attention network can effectively optimize the state input of reinforcement learning. Besides, we design a multi-agent architecture to achieve collaborative optimization, making the algorithm adaptable to applications in large-scale satellite-terrestrial integrated networks.

We gratefully acknowledge support from the National Natural Science Foundation of China, No. 61931005.