基于 DQN 的 VDES 异构星座兼容策略研究*

王雪帆 [1,2,3]， 李宗旺 [1,3]， 梁旭文 [1,3†]

(1 中国科学院微小卫星创新研究院，上海 201203；2 上海科技大学信息科学与技术学院，上海 201210；
3 中国科学院大学，北京 100049)

（2022年4月20日收稿；2022年6月21日收修改稿）

王雪帆，李宗旺，梁旭文. 基于DQN的VDES异构星座兼容策略研究[J]. 中国科学院大学学报，DOI:10. 7523/j.ucas. 2022. 071.

摘要 异构 VDES（VHF Data Exchange System）星座采用相同的通信频率和时分多址（Time Division Multiple Access，TDMA）通信机制，使得异构星座重复覆盖区域内存在大量由时隙冲突造成的同频干扰，严重影响通信质量。针对这个问题，提出了一种基于深度 Q 网络（Deep Q-net，DQN）的星座间兼容策略。基于 VDES 通信流程，设置船站作为资源信息中转节点，赋予卫星对通信环境的感知能力。在此基础上，将异构星座场景下的资源分配问题建模为强化学习 (Reinforcement Learning，RL) 问题，提出了一种基于 DQN 的时隙资源分配算法。通过重构历史资源信息和当前资源信息，规划最优时隙资源分配方案，并根据结果对算法迭代优化。仿真结果表明，所提出的策略可以有效提高通信性能。

关键词 VDES; DQN; 卫星通信; 干扰规避; 资源分配

中图分类号：TN92 **文献标志码**：A **DOI**:10. 7523/j.ucas. 2022. 071

Research on compatible strategy of VDES heterogeneous constellation based on DQN

WANG Xuefan[1,2,3]， LI Zongwang[1,3]， LIANG Xuwen[1,3]

(1 *Innovation Academy for Microsatellites of CAS, Shanghai* 201203, *China*;
2 *School of Information Science and Technology, ShanghaiTech University, Shanghai* 201210, *China*;
3 *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract As heterogeneous VDES (VHF Data Exchange System) constellations use the same communication frequency and time division multiple access (TDMA) communication mechanism, a large number of co-channel interference is caused by slot conflicts in the overlapping area of constellations，reducing the communication quality. To tackle this problem, an inter-constellation compatibility strategy for deep Q-net (DQN) is proposed. Based on the VDES communication process, the ship station is set as the resource information transfer node, which enables the satellite to perceive the communication environment. On this basis, the resource allocation problem in the heterogeneous constellation scenario is modeled as a Reinforcement Learning (RL) problem, and a DQN-based slot resource allocation algorithm is proposed. By reconstructing the historical resource information and current resource information, the optimal slot resource allocation scheme is planned and the algorithm is iteratively optimized according to the results. Simulation results show that the proposed strategy can effectively enhance the communication performance.

Keywords VDES; DQN; satellite communications; interference avoidance; resource allocation

*中国科学院青年创新促进会（2019293）资助

†通信作者，E-mail: liangxw@microsate.com

由于正在使用的海上信息服务——船舶自动识别系统（Automatic Identification System, AIS）和特殊应用报文(Application-Specific Messages, ASM)的信道容量已经趋于饱和，国际海事组织（International Maritime Organization，IMO）、国际航道标志协会（International Association of Lighthouse Authorities，IALA）和国际电信联盟（International Telecommunication Union，ITU），共同提出建设未来海上通信平台 VDES。为了争夺未来全球海上通信系统标准的制高点，依照 VDES 技术建议书[1]和 VDES 技术规范草案[2]，各个海洋强国积极投入到 VDES 系统的研究和建设中[3-5]。超过 10 个国家正在计划或建造多颗 VDES 卫星，而出于国家安全需要，难以实现国家间的星间链路。因此，VDES 异构星座兼容共存是未来发展的必然趋势。现有协议草案中所有 VDES 星座采用相同的频率、时间对齐方式和时隙表[2]，导致重复覆盖区的链路造成大量冲突，严重影响系统整体性能，因此 VDES 多星座兼容是 VDES 建设关键问题之一。

现有机制通过扩频的调制编码方式和两个卫星专用信道使得重复覆盖区最多能同时收到 4 个不同星座的扩频信号[2]。然而扩频信号的数据量是非扩频信号的 1/131 到 1/2.5[2]，无法充分满足应用数据的交互需求。

星座间兼容有硬兼容和软兼容两种方式，硬兼容指通过只改变硬件参数以达到兼容。星座间兼容已在全球导航卫星系统（Global Navigation Satellite System，GNSS）中提出，解决方法主要有功率控制、卫星姿态控制和天线覆盖范围控制等。VDES 卫星具有单波束覆盖范围广、重复覆盖区面积大、重复覆盖率高的特点，导致功率控制、姿态控制的运算和控制开销大，无法使用。为了兼容地面通信，VDES 卫星天线采用 60 度斜装，也无法控制覆盖范围。

软兼容指通过只改变软件协议，以达到兼容。随着卫星产业的商业化和规模化，地球静止轨道（Geostationary Earth Orbits，GEO）卫星和低轨（Low Earth Orbit，LEO）卫星逐渐成为星座间兼容的焦点。LEO 卫星之间信息交互协作，在保证 GEO 卫星通信质量的同时，提高时间利用率和吞吐量[6-8]。为了进一步提升性能，在 GEO 卫星性能不明显下降的情况下，通过 LEO 和 GEO 星座联合资源控制，提高频谱效率[9-10]。然而 GEO 和 LEO 共存时，GEO 性能更加优先，受干扰更少，而 VDES 星座的优先程度相同。

近年来，由于深度强化学习（deep reinforcement learning，DRL）在序列决策问题上的优异性能[11]，被广泛应用到通信资源分配上。针对 LEO 卫星的波束间资源分配问题，通过将通信请求的特征作为状态的一部分[12]，降低了链路切换时的延迟，在此基础上增加对波束内资源的统筹[13] ，进一步提高整体性能。但这些策略仅针对单一系统的场景，目前缺乏针对多星座共存时通信链路特点的时隙分配研究。

基于以上分析，结合VDES通信的高时变性的特点，提出一种基于改进DQN算法的VDES多星座兼容策略。首先，在现有VDES通信机制的基础上，设置船站作为时隙资源感知节点，使卫星能够获得所有其他卫星的历史资源分配信息。然后，将VDES通信资源分配过程建模成强化学习问题，以最大化长期动作收益总和为目标。最后通过状态重构及DQN算法对强化学习问题进行求解，得到近似最优的资源分配策略，提高VDES通信的性能。

1 时隙分配 DQN 模型

1.1 异构 VDES 星座兼容策略

VDE-SAT 是 TDMA 系统，一帧时长一分钟，以一分钟的开始为一帧的开始。一帧有 3 个子帧，每个子帧包含发送、处理、传输的完整通信流程，有 6 个 DC（Data Channel）时隙用于传输应用数据。卫星获取请求后，开始规划请求的分配，由于下行的 DC 时隙一定排在上行的 DC 时隙前面，导致所有卫星给同方向的请求分配的 DC 时隙具有趋同性[2]。然后卫星广播分配方案，DC 时隙被相应的终端占用传输应用数据。同序号的 DC 时隙被同一个终端占用直至应用数据全部传输完毕。

在自组织网络中，通过侦听信道能够感知其他终

端的传输情况，降低碰撞[14]，而 VDES 异构星座之间无法感知互相的时隙占用情况。在此启发下，针对星船通信特性，提出一种自组织星座间信息交互策略：卫星指定在重复覆盖范围内的某个船站作为时隙资源感知节点，然后向卫星转发其他星座卫星上个子帧的时隙分配信息。在不改变现有通信体制的情况下，达到星座间交互资源分配信息的目的。虽然卫星只能获得其他卫星上个子帧的分配情况，但历史 DC 时隙分配情况能为当前子帧的分配提供指引。上个子帧传输成功的 DC 时隙意味着该种分配方式具有较高的成功率，应为待发送请求选择最相似的分配方式。

根据其他消息的设计，设置资源分配信息消息为上行信令，指定船站作为节点的消息为下行信令。资源分配信息消息需要增加一次突发作为开销。上行信令采用扩频因子为 8 的扩频信号[2]，因此增加资源分配信息不会对正常的上行信令传输造成过大影响。另一方面，指定船站作为节点的消息开销仅为 6 个字节，而卫星一个子帧最多传输 234 个字节的下行行令[2]，即使有多个星座重复覆盖，该消息对下行信令的影响也有限。

1.2 系统模型

VDES 时隙分配是一种序列决策问题，而当前子帧的分配会对未来产生影响，应着眼于长远利益，最大化未来的总传输成功时隙数量。

性能评估函数设为平均传输成功时隙数量 P_S 和时隙利用率 P_U，P_S 用于衡量策略的整体性能，P_U 用于衡量策略对空闲时隙的利用程度。由于一个子帧是一个完整通信流程，P_S 和 P_U 都基于子帧衡量，P_S 定义为：

$$P_S = \sum_{t=1}^{T} n_{suc}^{t} / (6 \times T), \tag{1}$$

其中，T 为总子帧数，n_{suc}^{t} 为第 t 个子帧中传输成功的 DC 时隙数量。P_U 定义为：

$$P_U = \sum_{t=1}^{T} n_{usd}^{t} / \sum_{t=1}^{T} n_{vac}^{t}, \tag{2}$$

其中 n_{vac}^{t} 为第 t 个子帧中空闲 DC 时隙数量，n_{usd}^{t} 为 n_{vac}^{t} 中被分配了请求且传输成功的 DC 时隙数量。

由于 VDES 系统时变性高，尚处于预研阶段，具有较少的先验信息，无法针对通信过程建立有效的系统模型。这使得传统的动态规划算法[15-16]无法解决此类资源分配问题。因此，考虑通过基于无模型的强化学习算法[17]获得最大化 P_S 的策略。由于 P_S 为离散值，无法直接作为目标函数进行优化，需要通过优化一种与 P_S 具有较强相关性的目标函数间接优化 P_S。假如传输成功的 DC 时隙价值高，传输碰撞的 DC 时隙价值低，则 P_S 与所有 DC 时隙的总价值呈正相关。由于不分配时隙的操作也能通过降低碰撞率来提升 DC 时隙的价值，所以 DC 时隙的价值转换成操作的价值，将未来动作的总价值作为目标函数。基于以上分析，将时隙分配问题建模为 RL 问题，模型如图 1 所示。

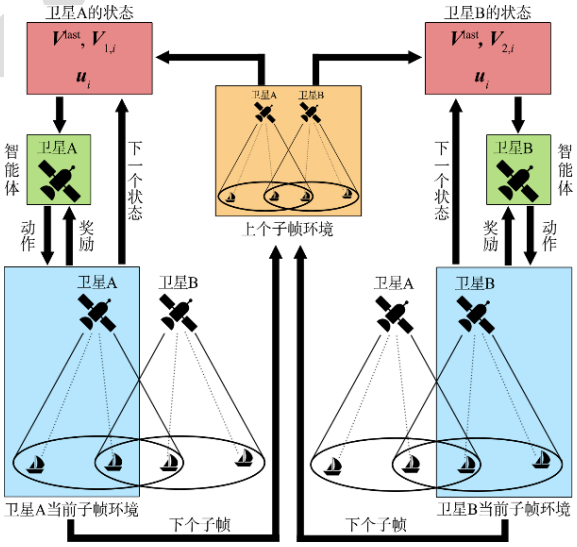

图 1　时隙分配问题的 RL 模型

Fig.1 RL model of slot allocation problem

1）状态：状态是环境高度精简概括。假设在某个子帧，第 k 个卫星收到的请求有 N 个。请求队列 $\boldsymbol{u} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_i, ..., \boldsymbol{u}_N\}$，其中第 i 个请求 $\boldsymbol{u}_i = \{u_i^1, u_i^2\}$，$u_i^1$ 表示该请求的通信方向，u_i^2 船站终端位置，因为碰撞由同一 DC 时隙两个链路的通信方向和船站终端位

置决定。

时隙状态矩阵(slot state matrix，SSM)定义为 $\boldsymbol{V}=\{\boldsymbol{V}_1,\boldsymbol{V}_2\}$，其中 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 分别表示卫星 A 和卫星 B 的时隙状态，$\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2\in\mathbb{R}^{6\times4}$，其定义为：

$$\boldsymbol{V}_k=\begin{bmatrix} v_1^1 & v_1^2 & \cdots & v_1^4 \\ v_2^1 & v_2^2 & \cdots & v_2^4 \\ \cdots & \cdots & \cdots & \cdots \\ v_m^1 & v_m^2 & \cdots & v_m^4 \\ \cdots & \cdots & \cdots & \cdots \\ v_6^1 & v_6^2 & \cdots & v_6^4 \end{bmatrix}, \tag{3}$$

其中 $k\in\{1,2\}$，$m\in\{1,2,...,6\}$，v_m^1 代表是否被占用，v_m^2 表示传输方向，v_m^3 表示船站位置，v_m^4 表示传输是否发生冲突。

针对 VDE-SAT 的高时变性和星座间的信息交互，状态应该包含请求信息 $\boldsymbol{u}_i$、处理 $\boldsymbol{u}_i$ 时第 k 个卫星的时隙状态 $\boldsymbol{V}_{k,i}$ 和处理完上个子帧的请求后所有卫星的时隙状态 $\boldsymbol{V}^{\text{last}}$，状态 $\boldsymbol{s}_{k,i}$ 定义为：

$$\boldsymbol{s}_{k,i}=\{\boldsymbol{V}^{\text{last}},\boldsymbol{V}_{k,i},\boldsymbol{u}_i\}. \tag{4}$$

基于 VDE-SAT 的通信流程，对于同一子帧的请求，上个子帧传输结果一致，所以这些请求对应 $\boldsymbol{s}_{k,i}$ 中的 $\boldsymbol{V}^{\text{last}}$ 也一致。

2）动作：动作是对正在处理的请求执行的操作。对于每个请求有 7 种操作选择，即不分配时隙和分配 6 个 DC 时隙中的某一个。强化学习经过运算后，获得每个动作的预期价值，将其中价值最高的动作的索引输出。用分配时隙的索引表示动作 a_i，动作 a_i 定义如下式：

$$a_i=\begin{cases} 0 & ，\quad \text{不分配DC时隙} \\ i & ，\quad \text{分配第}i\text{个DC时隙}, \end{cases} \tag{5}$$

3）奖励：奖励表示当前请求的各个动作对当前状态有利程度。不同动作对时隙分配的影响不同，奖励应分为不分配时隙时和分配 DC 时隙时两种情况进行判断。

当前子帧空闲 DC 时隙较少时，即 $\sum_m v_m^1\geq4$，传输成功的时隙数量可能较大。不分配时隙能够降低碰撞、减少对其他卫星的影响，价值较高，应给予正数奖励。而且空闲时隙越少，再分配请求造成碰撞的可能性越大，所以不分配的奖励越大。当前子帧无空闲 DC 时隙时，$\sum_m v_m^1=6$，只能执行不分配的动作，不分配的价值极高，应获得极大的奖励。当前子帧空闲 DC 时隙较多时，即 $\sum_m v_m^1\leq3$，为了提高传输成功时隙数量，应鼓励给请求分配时隙，不分配的动作价值低，应不予以奖励。当 $a_i=0$ 时，奖励 r_i 定义如下式：

$$r_i=\begin{cases} 0 & ，\ a_i=0\,\&\,\sum_m v_m^1\leq3 \\ \alpha_1 & ，\ a_i=0\,\&\,\sum_m v_m^1=4 \\ \alpha_2 & ，\ a_i=0\,\&\,\sum_m v_m^1=5 \\ 1 & ，\ a_i=0\,\&\,\sum_m v_m^1=6 \end{cases},0<\alpha_1<\alpha_2<1. \tag{6}$$

当 $a_i\neq0$ 时，首先需要考虑时隙分配的结果。为了提高传输成功时隙数量，传输是否成功决定奖励的正负。初始奖励 r^o 判断奖励的正负，由是否传输发生碰撞决定，定义如下式：

$$r^o=\begin{cases} \beta & ，\ a_i\neq0\&\text{传输无碰撞} \\ -\beta & ，\ a_i\neq0\&\text{传输有碰撞} \end{cases},\beta>0. \tag{7}$$

为了进一步提升传输成功时隙数量，基于星座间信息交互策略，当前子帧的链路与上个子帧的链路越相似，请求传输成功的可能性越高，这样分配的动作价值也越高。在传输成功的基础上，$\boldsymbol{V}^{\text{last}}$ 中的 $\boldsymbol{v}_{k,a_i}$ 的方向和船站位置与 $\boldsymbol{u}_i$ 的一致性越强，奖励应越大，设置每增加一个一致的特征，奖励翻倍，奖励 r_i 定义如下式：

$$r_i=\begin{cases} r^o, a_i\neq0\,\&\,v_{k,a_i}^2\odot u_i^1+v_{k,a_i}^3\odot u_i^2=0 \\ \delta r^o, a_i\neq0\,\&\,v_{k,a_i}^2\odot u_i^1+v_{k,a_i}^3\odot u_i^2=1 \\ \delta^2 r^o, a_i\neq0\,\&\,v_{k,a_i}^2\odot u_i^1+v_{k,a_i}^3\odot u_i^2=2 \end{cases},\delta>0. \tag{8}$$

2 基于 DQN 的时隙分配算法

2.1 DQN 构架

为了有效计算当前子帧的时隙分配结果对未来环境的影响，需要最大化未来动作的总价值。根据贝尔曼方程，在策略 π 下动作 a_i 的价值定义为：

$$q_\pi(\boldsymbol{s},a)=E_\pi(r_i+\gamma r_{i+1}+\gamma^2 r_{i+2}+\ldots\,|\,\boldsymbol{s}=\boldsymbol{s}_{k,i},a=a_i)$$
$$=r_i+\gamma\sum_{s'\in S}P_{ss'}\sum_{a'\in A}\pi(a'\,|\,s')q_\pi(s',a'), \tag{9}$$

其中，S 是所有状态的集合，A 是所有动作的集合；$\sum_{s'\in S}P_{ss'}\sum_{a'\in A}\pi(a'\,|\,s')q_\pi(s',a')$ 是在策略 π 下未来获得的奖励，用来衡量一次请求分配对未来的影响；γ 是折扣因子，用于调整未来奖励的重要性，$\gamma\in(0,1)$；$P_{ss'}$ 是状态 $\boldsymbol{s}$ 变为状态 $\boldsymbol{s}'$的概率；$\pi(a'\,|\,s')$ 是当状态为 $\boldsymbol{s}'$时选择动作 a'的概率。为了使长期利益最大化，应选择未来奖励最大的动作 a_{opt}，动作 a_{opt} 定义如下式：

$$a_{opt}=\arg\max_{a_i\in A}q_*(\boldsymbol{s}_i,a_i), \tag{10}$$

其中，行为价值函数 $q_*(\boldsymbol{s}_i,a_i)$ 是当前状态对应的所有动作中的价值最大的，行为价值函数的计算公式如下：

$$q_*(\boldsymbol{s}_i,a_i)=r_i+\gamma\sum_{\boldsymbol{s}_{i+1}\in S}P_{\boldsymbol{s}_i\boldsymbol{s}_{i+1}}\max_{a_{i+1}}q_*(\boldsymbol{s}_{i+1},a_{i+1}). \tag{11}$$

通过计算式(11)可以针对当前处理的请求得到最优决策，但由于状态空间过大，而且状态中的历史资源信息导致连续的状态之间具有强相关性，使得传统的 RL 算法计算量巨大又低效。而 DQN 算法[18]的记忆池加抽样不仅能应对状态空间过大，而且能破坏强相关性，使强化学习快速有效，可以有效解决此类问题。

整个问题已经简化成寻找行为价值函数 $q_*(\boldsymbol{s}_i,a_i)$ 最优值。Q 网络用于输出特定状态下动作的价值，即：$(\boldsymbol{s}_i,a_i)\to Q(\boldsymbol{s}_i,a_i\,|\,\theta)$，其中 θ 是 Q 网络的参数。为了使状态更加精准地反映通信所需信息，根据特征对当前子帧时隙状态 $\boldsymbol{V}_{k,i}$ 和上个子帧时隙状态 $\boldsymbol{V}^{\text{last}}$ 以及当前处理的请求 $\boldsymbol{u}_i$ 进行重构，如图 2 所示。首先将两种传输方向和两种船站位置编码成 4 位独热码。然后 $\boldsymbol{u}_i$ 在独热码后方填 0 重构成 1×8 的向量。$\boldsymbol{V}_{k,i}$ 结合每个 DC 时隙是否被分配，选择 4 位 0 或者 4 位独热码，并在 4 位数据后面填 0，6 个当前子帧 DC 时隙对应的重构向量组成 6×8 的矩阵。$\boldsymbol{V}^{\text{last}}$ 在重构前需要先根据传输分析获得上个子帧的传输结果，确定 v_m^4 的值，而发生碰撞的 DC 时隙当成空闲时隙处理。每一个 DC 时隙再根据卫星自身是否传输成功，判断独热码是由上个子帧 $\boldsymbol{V}^{\text{last}}$ 中卫星自身的 SSM $\boldsymbol{V}_k$ 产生还是由其他卫星的 SSM $\boldsymbol{V}_{k'}$ 产生，确定独热码的位置，并在空余的地方填 0。6 个上个子帧 DC 时隙对应的重构向量组成 6×8 的矩阵。为了使每个请求处理时的 $\boldsymbol{V}^{\text{last}}$ 差异化，将 $\varphi(\boldsymbol{u}_i)$ 作为一个元素加入 $\varphi(\boldsymbol{V}^{\text{last}})$ 中。重构后的状态定义如下式：

$$\varphi(\boldsymbol{s}_{k,i})=\varphi(\boldsymbol{V}^{\text{last}},\boldsymbol{V}_{k,i}). \tag{12}$$

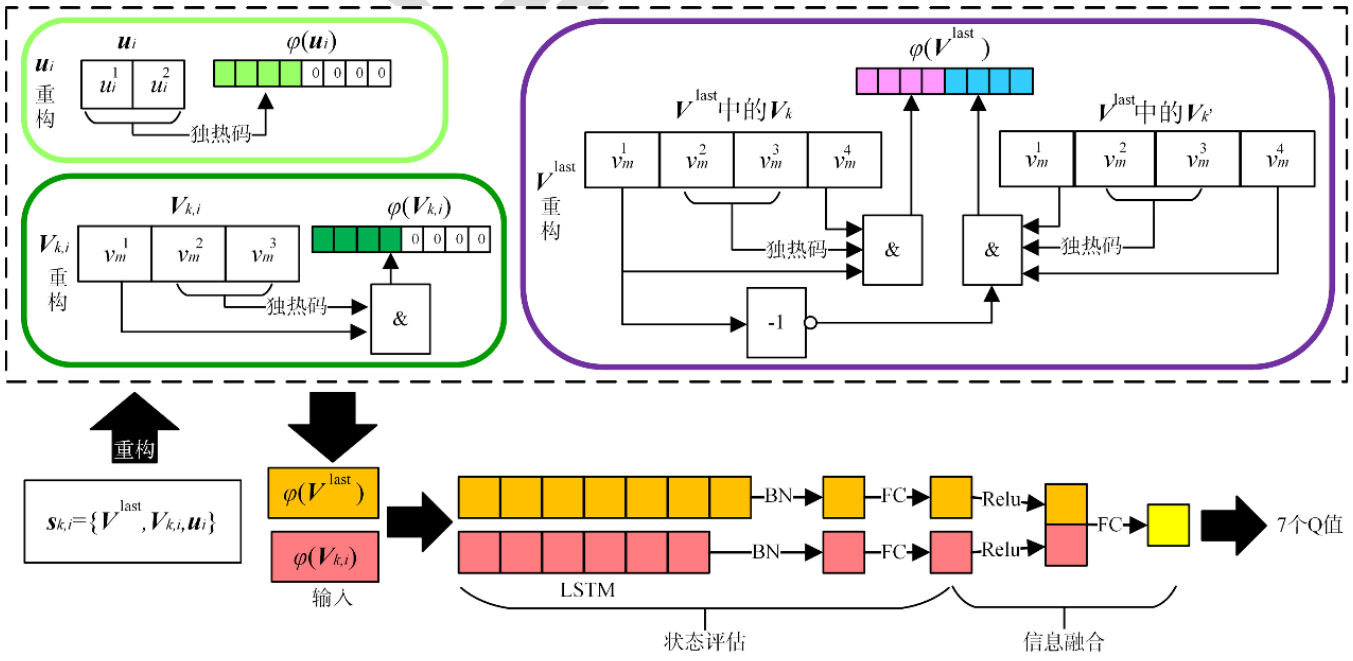

图 2　状态的重构与 Q 网络结构

Fig.2 Reformulation of state variables and structure of Q network

如图 2 所示，Q 网络的输入是具有 2 个元素的 $\varphi(\boldsymbol{s}_i)$，输出是 7 个动作的 Q 值。Q 网络的结构分为状态评估和信息融合两部分。由于时隙分配具有有序、时间相关的特点，而且同时需要长期的通信环境特征和当前子帧的通信环境特征，所以使用 2 个长短期记忆 (long-short term memory，LSTM) 层作为状态评估的网络，用以分析 $\varphi(\boldsymbol{s}_i)$ 的每个元素。为了融合分析结果，信息融合部分使用了 2 个全连接（fully connected，FC）层。

2.2 DQN 参数更新

为了优化策略，需要通过迭代优化 Q 网络的参数 θ。值得注意的是，与一般的强化学习问题不同，时隙分配问题没有终止状态。若当前子帧没有空闲时隙时，应拒绝以后的请求。该状态被认为是当前子帧的分配终止，设置为无效状态，其目标 Q 值的计算方式与其他状态不同。

损失函数 $L(\theta)$ 表示当前状态下能够获得的最大未来奖励与 Q 网络输出动作对应未来奖励之间的差距，即目标 Q 值 y_i 与实际 Q 值之差，定义如下所示：

$$L(\theta)=E[(y_i-Q(\varphi(\boldsymbol{s}_i),a_i;\theta))^2]. \tag{13}$$

目标 Q 值实际上只有在每次迭代结束后才能更新。在实际应用中，为了降低运算量，引入时序差分（1-step TD）学习的概念，用当前奖励加上折扣后的未来奖励作为目标 Q 值，因为受过部分训练的网络已经可以提供一个近似最优的策略。目标 Q 值 y_i 的计算公式为：

$$y_i=\begin{cases} r_i+\eta & ,\boldsymbol{s}_{i+1}\text{为无效状态} \\ r_i+\gamma\max_{a_{i+1}}\overline{Q}(\varphi(\boldsymbol{s}_{i+1}),a_{i+1};\overline{\theta}) & ,\boldsymbol{s}_{i+1}\text{不为无效状态,} \end{cases} \tag{14}$$

其中，$\overline{\mathrm{Q}}$ 为具有参数 $\overline{\theta}$ 的目标 Q 网络，η 用于给无效状态赋值。

改进的 DQN 算法的主要实现过程如算法 1 所示，分为初始化阶段和训练阶段。初始化阶段包括场景参数和模型参数的初始化。在更新 Q 网络参数的过程中，使用 ε-贪心算法用以提高动作空间探索程度。通过网络参数不断的更新迭代，对决策评估越来越准确，能得到近似最优的决策。

算法 1 基于改进 DQN 的时隙分配算法

初始化

初始化当前子帧的 SSM、上个子帧的 SSM、Q 网络和经验池

训练

For 请求 i=1:N **do**

只保留 3 个子帧之内达到的请求

由 $\boldsymbol{V}_{k,i}$ 获得 $\boldsymbol{s}_i$

if $\boldsymbol{s}_i$ 是无效状态 **then**

令 $a_i=0$

else

重构 $\boldsymbol{s}_i$ 为 $\varphi(\boldsymbol{s}_i)$

使用 ε-贪心算法获得动作 a_i

更新 $\boldsymbol{V}_{k,i}$

通过下一个请求获得 $\boldsymbol{s}_{i+1}$，并重构为 $\varphi(\boldsymbol{s}_{i+1})$

end if

标记 $\boldsymbol{s}_{i+1}$ 是否是无效状态

if 当前子帧的请求处理完毕 **then**

通过各个卫星的时隙分配，获得传输结果

根据传输结果获得动作 a_i 对应的奖励 r_i

更新 $\boldsymbol{V}^{\text{last}}$ 和 $\boldsymbol{V}_{k,i}$

将 $\{\varphi(\boldsymbol{s}_i),a_i,r_i,\varphi(\boldsymbol{s}_{i+1})\}$ 存入经验池

end if

从经验池中取一批元组计算 $L(\theta)$

根据 $L(\theta)$ 的值执行梯度下降

每 G 个子帧用 $\overline{\theta}=\theta$ 更新 $\overline{\theta}$

End for

3 仿真结果

依照 AIS 和 ASM 的实际运行过程中的真实数据，VDES 仿真参数设置如表 1 所示。重复覆盖率为在重复覆盖区与卫星建立通信链路的船站在所有与卫星建立通信链路的船站中的占比，假设所有卫星的重复覆盖率相同。针对 VDES 海洋通信容量更大的特点，增加

了业务传输所需时隙数量。业务的传输方向和船站位置服从平均分布。新请求到达率服从独立的泊松分布。策略参数如表 2 所示。

表 1　仿真参数

Table 1 Simulation parameters

仿真参数	参数值
业务请求到达率/（个/船/子帧）	0.02
业务占用时隙数量/个	1~3
单个卫星覆盖船站数量/个	25~1800
重复覆盖率	0~1

表 2　策略参数

Table 2 Strategy parameters

策略参数	参数值
仿真时间 T/子帧	2000
折扣因子 γ	0.8
α_1	0.125
α_2	0.1875
β	0.2
δ	2
学习率	0.001
经验池存储状态数量/个	10000
探索概率 ε	0.1
批大小	128

由于分配实际由卫星执行，对分配产生影响的是请求的总数以及各类请求的比例，因此将平均一个子帧里卫星收到的新请求数量作为卫星请求到达数量，而重复覆盖率反映了船站位置的分布。为了保证请求的实时性，每个请求最多被尝试分配 3 次，之后将作为旧请求被剔除。此外，将随机分配方法作为对照。随机分配方法为每个请求随机选择不分配时隙或者分配某个 DC 时隙。现有分配方法[2]和随机分配方法均仅依靠卫星自身的分配信息，不具有通信环境感知能力。

图 3 显示了 DQN 算法的迭代收敛过程，在仿真过程中，卫星请求到达个数为 12，重复覆盖率为 0.5。从图 3 中可以看出，DQN 算法在经过 200 多次迭代后逐渐收敛，表明该算法能够通过历史资源分配信息和当前资源分配信息合理分配资源。

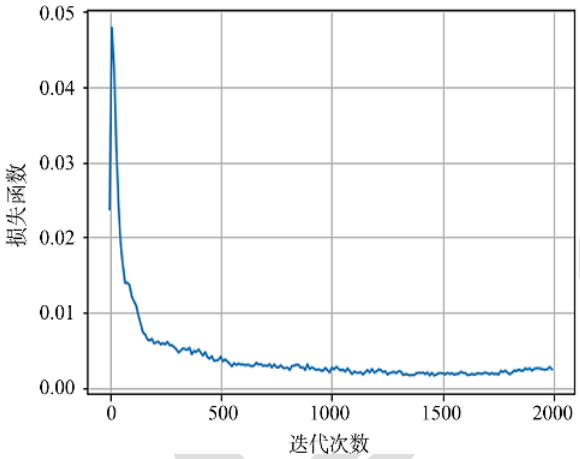

图 3　DQN 算法损失函数收敛速度

Fig.3 The convergence speed of the DQN algorithm

传输成功时隙数量 P_S 与卫星请求到达数量的仿真结果如图 4(a)所示。时隙利用率 P_U 与卫星请求到达数量的仿真结果如图 4(b)所示。由图 4 可知，相比其他两种方法，基于 DQN 的分配算法在不同卫星请求到达数量、不同重复覆盖率的情况下都表现出更好的性能，证明了所提出策略的有效性。现有机制的分配方法性能最差，是因为现有机制倾向于给同一 DC 时隙分配同向请求，提升了碰撞发生概率。当卫星请求到达率超过 8 时，所有算法达到性能上限，对时隙的利用能力不再随请求的增加而提高。对于基于 DQN 的分配算法，再增加请求个数只会为了降低碰撞而不为新请求分配时隙；对于现有分配方法和随机选择，因为此时几乎所有时隙已经被分配完，再增加请求个数不会对结果造成影响。性能上限表明，当卫星请求到达个数超过 8 时，此时卫星应广播限制低重要性的请求发送频率，保证重要信息的传输顺畅。此外，每秒浮点数运算次数（floating point operations per second，FLOPS）反应神经网络算法对硬件性能的要求。目前市面上的 DSP 和 FPGA 的 FLOPS 大多为 1G-80G。经统计，基于 DQN 的分配算法的单次运行所需浮点数运算次数为 40573，则运算用时小于 1ms，卫星硬件水平能够承受算法的时间开销。

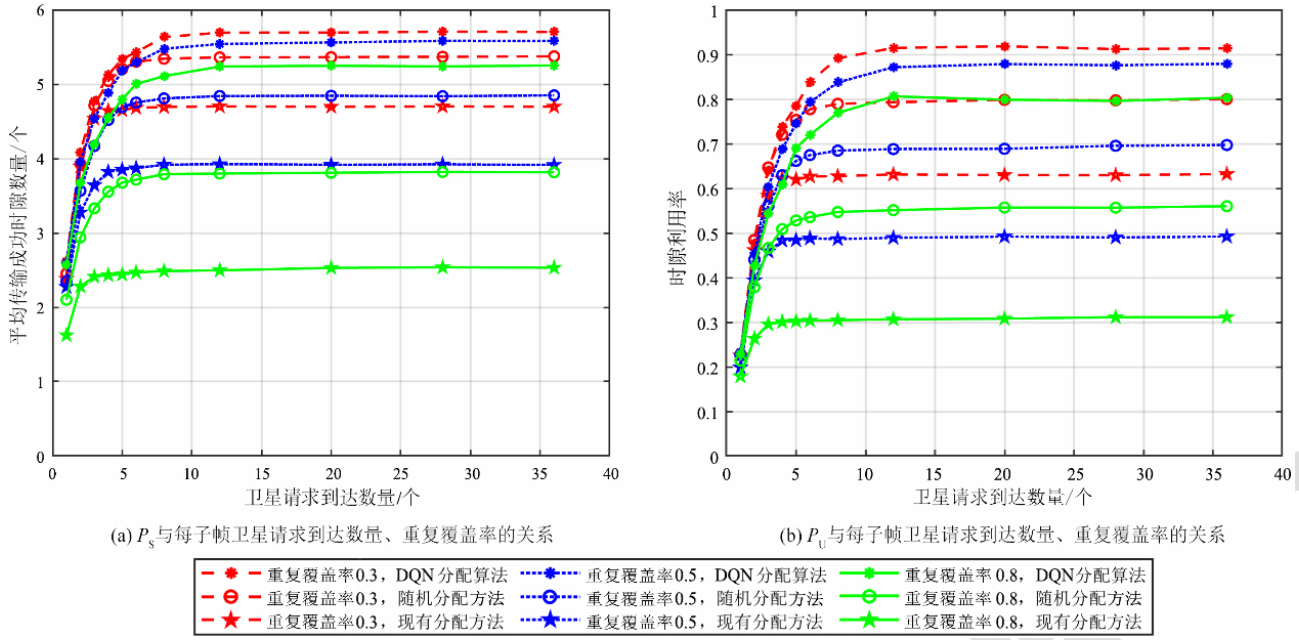

图 4　不同时隙分配方法的通信性能对比

Fig.4 Communication performance comparison of different slot allocation methods

4 结论

针对 VDES 多星座重复覆盖场景中的同频碰撞问题，本文提出了一种基于改进 DQN 的星座间兼容策略。基于 VDES 通信流程，通过设置船站作为卫星时隙资源感知节点，提升星座间的信息交互性。在此基础上提出一种基于改进 DQN 的时隙分配算法，将重构后的请求信息、历史时隙分配结果和当前子帧的时隙分配方案作为关键因素输入网络，获得时隙分配方案，并通过分配结果更新优化算法。仿真结果表明，提出的策略能够显著提高传输成功时隙数量。下一阶段的研究可以从重复覆盖率的动态变化入手，模拟真实场景中的卫星运动，为 VDES 协议的进一步完善提供建议。为了实现工程化应用，在维持性能不变的同时，降低 DQN 参数精度和模型压缩以满足卫星的计算能力也是研究方向之一。

参考文献

[1] ITU-R. Recommendation ITU-R M.2092-1-Technical characteristics for a VHF data exchange system in the VHF maritime mobile band [S/OL]. Geneva: International Telecommunication Union:(2022-02-23)[2022-04-20].https://www.itu.int/rec/R-REC-M.2092-1-202202-I/en/.

[2] IALA. Guideline G1139 The Technical Specification of VDES [S/OL]. Saint Germain en Laye: International Association of Marine Aids to Navigation and Lighthouse Authorities:(2019-06-21)[2022-04-20].https://www.iala-aism.org/product/g1139-technical-specification-vdes/.

[3] 姚治萱. VDES 通信技术应用及其发展趋势[J]. 世界海运, 2019, 42(2): 34-38. DOI:10.16176/j.cnki.21-1284.2019.02.007.

[4] 胡旭, 林彬, 王珍. 基于 VDES 的空天地海通信网络架构与关键技术[J]. 移动通信, 2019, 43(5): 1-8. DOI:10.3969/j.issn.1006-1010.2019.05.001.

[5] 王福斋, 胡青, 姚高乐, 等. 甚高频数字交换系统发展现状及推进工作建议[J]. 中国海事, 2021(2): 18-21. DOI:10.16831/j.cnki.issn1673-2278.2021.02.005.

[6] Wang Y F, Ding X J, Zhang G X. A novel dynamic spectrum-sharing method for GEO and LEO satellite networks[J]. IEEE Access, 2020, 8: 147895-147906. DOI:10.1109/ACCESS.2020.3015487.

[7] Gu P, Li R, Hua C Q, et al. Dynamic cooperative spectrum sharing in a multi-beam LEO-GEO co-existing satellite system[J]. IEEE Transactions on Wireless Communications, 2022, 21(2): 1170-1182. DOI:10.1109/TWC.2021.3102704.

[8] Gu P, Li R, Hua C Q, et al. Cooperative spectrum sharing in a co-existing LEO-GEO satellite system[C]//GLOBECOM 2020 - 2020 IEEE Global

Communications Conference. December 7-11, 2020, Taipei, China. IEEE, 2020:1-6. DOI:10.1109/GLOBECOM42002.2020.9347950.

[9] Jia M, Li Z, Gu X M, et al. Joint multi-beam power control for LEO and GEO spectrum-sharing networks[C]//2021 IEEE/CIC International Conference on Communications in China (ICCC). July 28-30, 2021, Xiamen, China. IEEE, 2021: 841-846. DOI:10.1109/ICCC52777.2021.9580210.

[10] 李壮. 基于干扰控制的 LEO 和 GEO 频谱共享方法[D]. 哈尔滨: 哈尔滨工业大学, 2021.

[11] Bu G J, Jiang J. Reinforcement learning-based user scheduling and resource allocation for massive MU-MIMO system[C]//2019 IEEE/CIC International Conference on Communications in China (ICCC). August 11-13, 2019, Changchun, China. IEEE, 2019: 641-646. DOI:10.1109/ICCChina.2019.8855949.

[12] Li Z W, Xie Z C, Liang X W. Dynamic channel reservation strategy based on DQN algorithm for multi-service LEO satellite communication system[J]. IEEE Wireless Communications Letters, 2021, 10(4): 770-774. DOI:10.1109/LWC.2020.3043073.

[13] 李子煜. 多波束低轨卫星通信系统切换与资源管理算法研究[D].重庆:重庆邮电大学,2021.

[14] Seo J, Cho K, Cho W, et al. A discovery scheme based on carrier sensing in self-organizing Bluetooth Low Energy networks[J]. Journal of Network and Computer Applications,2016,65:72-83. DOI:10.1016/j.jnca.2015.09.015.

[15] 韩存武, 周慧, 刘蕾, 等. 基于数据的无线通信网络功率和速率控制[J]. 计算机仿真, 2022, 39(2): 375-379, 511. DOI:10.3969/j.issn.1006-9348.2022.02.072.

[16] 曾欢, 张灿, 陈德元. 空间通信网中音视频传输的应用层 QoS 控制与测试方法[J]. 中国科学院研究生院学报, 2011, 28(1): 108-115. DOI:10.7523/j.issn.2095-6134.2011.1.016.

[17] Strehl A L, Li L H, Wiewiora E, et al. PAC model-free reinforcement learning[C]//ICML '06: Proceedings of the 23rd international conference on Machine learning. June 25-29, 2006, Pittsburgh, USA. New York: ACM, 2006: 881–888. DOI:10.1145/1143844.1143955.

[18] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533. DOI:10.1038/nature14236.