基于深度增强学习和多目标优化改进的卫星资源分配算法

张沛[1,2]，刘帅军[3]，马治国[2]，王晓晖[1]，宋俊德[1]
（1. 北京邮电大学计算机学院，北京 100876；2. 中国信息通信研究院，北京 100191；
3. 中国科学院软件研究所，北京 100190）

摘 要：针对多波束卫星系统中资源分配序列决策的多目标优化（MOP）问题，为了在提升卫星系统性能的同时，提高用户业务需求的满意度，提出了一种基于深度增强学习（DRL）的 DRL-MOP 算法。所提算法基于 DRL 和 MOP 技术，对动态变化的系统环境和用户到达模型建模，以归一化处理后的频谱效率、能量效率和业务满意度指数的加权和作为优化目标，实现了系统和用户累计性能的优化。仿真对比表明，所提算法可以更好地解决面向多波束卫星系统的多目标优化问题，系统性能和用户满意度优化结果较好，且收敛快、复杂度低。

关键词：多波束卫星系统；资源分配；序列决策；深度增强学习；多目标优化
中图分类号：TN927+.2
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2020117

Improved satellite resource allocation algorithm based on DRL and MOP

ZHANG Pei[1,2], LIU Shuaijun[3], MA Zhiguo[2], WANG Xiaohui[1], SONG Junde[1]
1. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. China Academy of Information and Communications Technology, Beijing 100191, China
3. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China

Abstract: In view of the multi-objective optimization (MOP) problem of sequential decision-making for resource allocations in multi-beam satellite systems, a deep reinforcement learning(DRL) based DRL-MOP algorithm was proposed to improve the system performance and user satisfaction degree. With considering the normalized weighted sum of spectrum efficiency, energy efficiency, and satisfaction index as the optimization goal, the dynamically changing system environments and user arrival model were built by the proposed algorithm, and the optimization of the accumulative performance in satellite systems based on DRL and MOP was realized. Simulation results show that the proposed algorithm can solve the MOP problem with rapid convergence ability and low complexity, and it is obviously superior to other algorithms in terms of system performance and user satisfaction optimization.

Key words: multi-beam satellite system, resource allocation, sequential decision-making, deep reinforcement learning, multi-objective optimization

1 引言

卫星通信网络由于其覆盖广、部署快、不受地面情况影响的优点，已经被用于多个商用系统，在空天互联网和 5G 网络中也是研究热点之一。随着卫星通信的快速发展，多点波束以高增益的点波束覆盖和频分复用的优点，在卫星通信系统中获得了广泛应用。而在多波束卫星通信系统中，由于频谱、

收稿日期：2019–12–20；**修回日期**：2020–05–20
通信作者：马治国，mazhiguo@caict.ac.cn
基金项目：国家重点研发计划基金资助项目（No.2018YFB0105105）；国家科技重大专项基金资助项目（No.2018ZX03001016）
Foundation Items: The National Key Research and Development Program of China (No.2018YFB0105105), The National Science and Technology Major Project of China (No.2018ZX03001016)

功率等资源受限特性，资源分配问题一直是备受关注的研究热点[1-2]。

基于传统的优化算法，针对卫星系统中的资源分配问题，国内外研究机构及学者已经做了大量研究工作[3-9]。其中，文献[3-5]建立了卫星系统功率分配优化问题，采用拉格朗日对偶理论进行优化，以达到系统总容量的最优化[3-4]以及能量效率（EE, energy efficiency）和频谱效率（SE, spectral efficiency）的多目标联合优化[5]。然而一个主要的影响因素，即同频信道干扰（CCI, co-channel interference）被忽略。文献[6]指出资源分配优化问题在考虑 CCI 的场景中被证明是 NP 难题，在此背景下，包括遗传算法（GA, genetic algorithm）、模拟退火（SA, simulated annealing）算法、快速非支配排序遗传算法（NSGA, non-dominated sorting genetic algorithm）等启发式算法被广泛应用。文献[7]提出一种基于 SA 算法改进的算法，实现卫星系统吞吐量最大化，同时兼顾波束间公平。为了实现多目标优化，文献[8]提出了一种两阶段的功率优化方法，利用 GA-SA 算法和 NSGAII 算法实现容量最大化和功率最小化的折中优化。文献[6]利用 NSGAII 算法求解多波束卫星系统中 EE 和 SE 多目标优化问题的帕累托解。为了进一步实现用户和系统性能的综合优化，文献[9]提出一种基于 SA 算法和 NSGAII 改进的算法，实现了在星上缓存限制下用户业务满意度指数（SI, satisfaction index）和系统 SE 的联合优化。然而，传统的资源优化主要解决当前时刻的系统性能，很难适应动态变化的复杂环境。

随着新一代人工智能技术的日趋成熟和广泛应用，深度学习算法被初步应用于无线资源分配的研究，并证明了其有效性，如解决蜂窝通信系统的资源分配问题[10]、密集 WLAN 的切换管理[11]、无线通信网络中的功率优化[12]等。深度增强学习（DRL, deep reinforcement learning）等深度学习方法在卫星系统的资源优化领域中也有许多研究工作。Ferreira 等[13]提出了一种基于深度神经网络集成的多目标增强学习，求解认知卫星通信资源分配多目标优化问题。Hu 等[14-15]利用 DRL 算法进行多波束卫星系统和下一代宽带卫星系统中的跳波束（BH, beam hopping）动态决策，相较传统算法具有较低的复杂度。文献[16]中提出了一种多波束卫星系统中的资源分配框架。Liu 等[17-18]指出 DRL 算法能够很好地解决序列决策分配问题，并基于 DRL 提出了一种解决动态信道分配（DCA, dynamic channel allocation）问题的 DRL-DCA 算法，降低了卫星系统的系统业务阻塞率。在这些文献中，DRL 等深度学习算法被应用于解决资源优化等动态决策问题，相较传统算法具有更好的动态环境感知能力和序列决策能力。但是这些研究没有考虑用户性能的提升，如在传统优化中提到的业务满意度指数[9]。

本文为了解决多波束卫星系统中资源分配序列决策的累计性能优化和多目标优化（MOP, multi-objective optimization）问题，在提升卫星系统性能的同时，提高用户业务需求的满意度指数，尽可能满足用户的业务需求，提出了一种基于深度增强学习和多目标优化改进的 DRL-MOP 算法。首先基于 DRL 算法，将卫星和用户建模为智能体，将信道环境建模为交互环境，进行了状态、动作和收益的设计；将归一化处理后的能量效率、频谱效率和满意度指数的加权和作为优化目标；在智能体和环境的交互过程中，依据目标函数的增量给出智能体的即时奖励；进一步地，基于 Q-learning 的思想，利用误差函数和随机梯度下降法训练和更新网络，实现动作值函数的优化。

2 系统模型和资源分配问题

2.1 系统模型

考虑多波束卫星系统的下行链路，用户初始时刻随机地分布在不同波束中。系统的总波束数目为 M_b，初始时刻的下行用户总数为 N。系统采用时分复用和全频复用，不同波束之间产生不同程度的同频干扰，系统链路采用高斯白噪声信道。考虑连续时刻下的资源分配，假设在时刻 t_j，已服务用户集合为 U_{t_j}，且每个用户只占用系统中的一个载波资源。则 U_{t_j} 占用的卫星载波分配矩阵为 $\boldsymbol{W}_{t_j}$。在时刻 t_j，下行用户 u_{t_j} 按照泊松分布随机到达或离开系统，间隔时间为 Δt。多波束卫星系统模型如图 1 所示。

2.2 资源分配问题

考虑多波束卫星通信系统中，载波和功率联合分配的优化问题。给定多波束卫星系统总的带宽为 B，每载波的带宽为 B_m，则系统中可用的最大载波数量为 $M_c = \dfrac{B}{B_m}$。考虑动态载波分配，即每个波

束可以在可用载波集合中分配载波，则每波束的带宽取决于分配给该波束的载波个数。

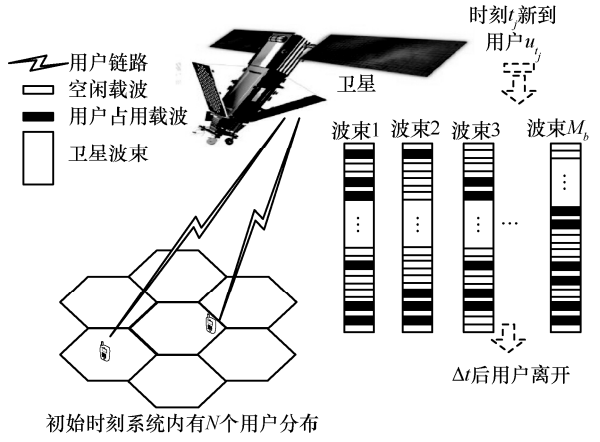

图 1　多波束卫星系统模型

系统的波束和载波分配矩阵 $\boldsymbol{W}_{t_j}$ 可以表示为

$$\boldsymbol{W}_{t_j}=\begin{bmatrix} {w_{11}}^{t_j} & {w_{12}}^{t_j} & \cdots & {w_{1b}}^{t_j} & \cdots & {w_{1M_b}}^{t_j} \\ {w_{21}}^{t_j} & {w_{22}}^{t_j} & \cdots & {w_{2b}}^{t_j} & \cdots & {w_{2M_b}}^{t_j} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ {w_{M_c1}}^{t_j} & {w_{M_c2}}^{t_j} & \cdots & {w_{M_cb}}^{t_j} & \cdots & {w_{M_cM_b}}^{t_j} \end{bmatrix} \tag{1}$$

其中，$\boldsymbol{W}_{t_j}$ 为元素取值为 0 或 1 的矩阵，大小为 $M_c\times M_b$。矩阵 $\boldsymbol{W}_{t_j}$ 的列代表某波束，总列数即总波束数目 M_b；行代表某载波，总行数即每波束可以分配到的最大载波个数 M_c。若 ${w_{mb}}^{t_j}$=1，代表在时刻 t_j，第 m 个（行）载波被分配给对应列的第 b 个（列）波束；若 ${w_{mb}}^{t_j}$=0，代表该载波没有分配给对应列的波束。不同时刻下，$\boldsymbol{W}_{t_j}$ 会随着下行用户 u_{t_j} 的到达或离开而变化，分配或释放相应的载波资源。

时刻 t_j，系统的总功率为 ${P_{\text{tot}}}^{t_j}$，每波束的功率为 $P_b^{t_j}$，其中，$b=1,2,\cdots,M_b$ 代表波束数目编号。假设波束内每载波的功率取固定值 P_c，则每波束分配的功率随载波数目的不同而不同。则依据式(1)，其功率 $P_b^{t_j}$ 可表示为

$$P_b^{t_j}=P_c\sum_{m=1}^{M_c}w_{mb}^{t_j} \tag{2}$$

则系统的总功率为

$${P_{\text{tot}}}^{t_j}=\sum_{b=1}^{M_b}P_b^{t_j}+P_o \tag{3}$$

其中，$\sum_{b=1}^{M_b}P_b^{t_j}$ 为系统总的有效载荷功率；P_o 为卫星平台本身的功率，主要包括用户链路和控制链路的功率放大器等引起的功率损耗。根据香农公式，系统中用户 i 的最大速率 $R_i^{t_j}$ 可以表示为

$$R_i^{t_j}=B_m\det\left(1\text{b}\left(\boldsymbol{I}_{M_c}+\mathbf{SINR}_i^{t_j}\right)\right) \tag{4}$$

其中，用户编号 $i=1,2,3,\cdots,N_{t_j}$，N_{t_j} 是时刻 t_j 系统中的下行用户总数；$\boldsymbol{I}_{M_c}$ 是 M_c 阶单位矩阵；$\mathbf{SINR}_i^{t_j}$ 是用户 i 的信干噪比矩阵，考虑了波束间的同频干扰和系统高斯白噪声。

假设系统在理想状态下，即用户可以达到最大速率 $R_i^{t_j}$，则时刻 t_j 系统的能量效率 EE_{t_j} 为

$$\text{EE}_{t_j}=\frac{\sum_{i=1}^{N_{t_j}}R_i^{t_j}}{\sum_{b=1}^{M_b}P_b^{t_j}+P_o} \tag{5}$$

其中，$\sum_{i=1}^{N_{t_j}}R_i^{t_j}$ 为系统总的下行吞吐量。

系统的频谱效率 SE_{t_j} 可表示为

$$\text{SE}_{t_j}=\frac{\sum_{i=1}^{N_{t_j}}R_i^{t_j}}{B} \tag{6}$$

通常为了最优化多波束卫星系统的吞吐量和功率性能，需要对 EE 和 SE 进行优化。然而，由于它们都是从卫星系统的角度出发，并没有从用户的角度考虑，比如是否能最大限度地满足用户业务需求。因此，本文引入满意度指数的优化目标，满意度指数指的是系统内所有下行用户业务需求的满足程度。假设每个用户的业务需求速率为 R_i^{req}，则总的满意度指数为

$$\text{SI}_{t_j}=\sum_{i=1}^{N_{t_j}}\min\left(\frac{R_i^{t_j}}{R_i^{\text{req}}},1\right) \tag{7}$$

其中，$\min\left(\frac{R_i^{t_j}}{R_i^{\text{req}}},1\right)$ 为用户 i 的满意度指数，当 $R_i^{t_j}>R_i^{\text{req}}$ 时，$\min\left(\frac{R_i^{t_j}}{R_i^{\text{req}}},1\right)$=1，表示该用户的业务需求满足；否则，$\min\left(\frac{R_i^{t_j}}{R_i^{\text{req}}},1\right)=\frac{R_i^{t_j}}{R_i^{\text{req}}}$，表示业务需

求不满足，其取值范围为 0~1，表征了相应的满足程度。

进一步地，可以将多波束卫星系统的资源分配问题的优化目标表示为式(8)~式(11)。

$$\text{opt.}\quad \mathcal{P}_1=\max\sum_{t_j\in\mathcal{T}}\mathrm{SE}_{t_j} \tag{8}$$

$$\mathcal{P}_2=\max\sum_{t_j\in\mathcal{T}}\mathrm{EE}_{t_j} \tag{9}$$

$$\mathcal{P}_3=\max\sum_{t_j\in\mathcal{T}}\mathrm{SI}_{t_j} \tag{10}$$

$$\begin{aligned}\text{s.t.}\ &\sum_{i=1}^{N_{t_j}}R_i^{t_j}\geqslant R_{\mathrm{tot}}^{\mathrm{req}},\forall t_j\\ &\sum_{b=1}^{M_b}P_b^{t_j}\leqslant P_{\mathrm{tot}}^{\max},\forall t_j\\ &P_b^{t_j}\leqslant P_b^{\max},\forall b,t_j\end{aligned} \tag{11}$$

本文优化的是 T 时间段的系统累计性能，其中 $\mathcal{T}=\{t_j\mid t_j\in[0,T]\}$。$\mathcal{P}_1$ 为最大化总累计频谱效率；$\mathcal{P}_2$ 为最大化总累计能量效率；$\mathcal{P}_3$ 为总的累计满意度指数。任意时刻 t_j 下，系统总的分配功率、每波束的功率应不高于系统总功率门限 $P_{\mathrm{tot}}^{\max}$ 和单波束功率门限 $P_b^{\max}$。此外，任意时刻，系统总的吞吐量应该满足最小需求速率要求 $R_{\mathrm{tot}}^{\mathrm{req}}$，以保障系统的最低性能要求。

3 算法

3.1 DRL-MOP 算法架构

本文所提 DRL-MOP 算法中，为了达到卫星和用户性能同时优化的多目标，将智能体建模为卫星和用户的整体，将信道环境建模为交互环境，通过将多波束卫星系统中的资源分配问题建模为智能体与环境的交互过程，并通过 Q 网络的反复训练和学习，达到最大收益。所提算法架构如图 2 所示。状态输入网络后，得到相应的动作和收益，同时状态数据更新，得到经验数据 $\Phi(j)=\{s_j,a_j,r_j,s_{j+1}\}$。采用经验回放的机制，当经验数据达到回放门限时，利用误差函数 $L(\theta)$ 进行网络训练和目标网络参数更新。基于 Q-learning 思想，实现累计收益的优化。

3.2 DRL-MOP 设计

首先，构建马尔可夫决策过程（MDP, Markov decision process）。

1) 状态 s_j。DRL-MOP 算法的智能体由卫星和用户组成，相应的状态信息包含卫星的载波分配矩阵 $\boldsymbol{W}_{t_j}$、已服务用户集合 U_{t_j} 和新到用户信息 u_{t_j}。可以表示为

$$s_j=\{\boldsymbol{W}_{t_j},U_{t_j},u_{t_j}\} \tag{12}$$

2) 动作 a_j。在时刻 t_j，新到用户 u_{t_j} 随机接入系统的波束中。当接入波束有空闲载波且满足相应的功率和吞吐量限制时，由智能体的状态数据 s_j 经过预处理后输入 Q 网络，输出为 M_c 个 Q 值，对应不同载波的 Q 值。通过 $\varepsilon-$贪心算法，智能体可以随机选择空闲载波，或者将最大的 Q 值作为动作值。选择第 m 个 Q 值，对应为选择第 m 个载波的动作。动作可以表示为

$$a_j=m \tag{13}$$

其中，m 代表被选择的载波编号。

3) 收益 r_j。将优化目标的增量 ΔF 作为收益判

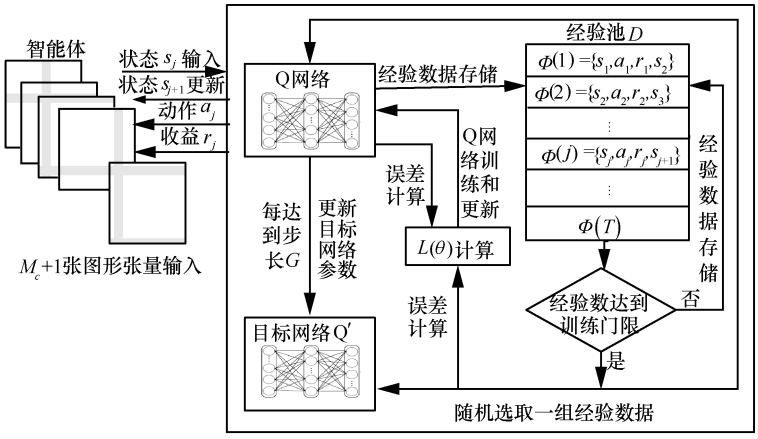

图 2 所提算法架构

定的依据。任务目标 F_j 由式(14)中的 3 个指标（频谱效率 SE_{t_j}、能量效率 EE_{t_j} 和业务满意度指数 SI_{t_j}）进行归一化处理后的加权和组成。首先，为了保障各指标对总的目标函数的影响公平，将 3 个指标值都归一化到[0,1]。接着，根据优化目标的侧重对各指标变量赋予相应的权重，再通过加权和求得优化目标 F_j。

$$F_j=\omega_1 a_1 \mathrm{SE}_{t_j}+\omega_2 a_2 \mathrm{EE}_{t_j}+\omega_3 a_3 \mathrm{SI}_{t_j} \tag{14}$$

其中，a_1、a_2、a_3 为 3 个指标的归一化参数，ω_1、ω_2、ω_3 为权重参数。则优化目标的增量 ΔF 可以表示为

$$\Delta F=F_{j+1}-F_j \tag{15}$$

根据激活函数 Sigmoid 函数，当优化目标的增量 $\Delta F>0$ 时，表示新的动作分配收益增大，则 $r_j=r_h$；否则，表示收益降低，则 $r_j=r_l$。其中 $r_h>r_l$，且 $r_h,r_l\in[0,1]$。收益计算式为

$$r_j=\begin{cases} r_h, \Delta F>0 \\ r_l, \Delta F\leqslant 0 \end{cases} \tag{16}$$

接着，对输入数据进行一定的处理，并且构建相应的 Q 网络训练和更新过程。

1) 输入重构。通常 Q 网络的输入需要满足一定的图形张量形式，因此需要进行状态重构。设定一个用户占用一个载波信道，因此载波分配矩阵 $\boldsymbol{W}_{t_j}$ 可以涵盖已服务用户集合信息 U_{t_j}。首先将载波分配矩阵其拆分为 M_c 个一维行向量。然后通过补 0 和矩阵变换操作，将其转换为 M_c 张 $L\times L$ 维图形张量。其中，每张图映射时刻 t_j 的一个载波占用情况。最后构造第(M_c+1)个全 0 向量，将向量中用户 u_{t_j} 到达的波束序号对应的位置置 1，并将其映射为一张 $L\times L$ 维图形张量。则状态可以映射为共(M_c+1)张图形张量，其中前 M_c 张图分别表征了时刻 t_j 系统中各载波被波束占用的信息 $\boldsymbol{W}_{t_j}$ 和服务的用户集合信息 U_{t_j}，第(M_c+1)张图表征了新到用户 u_{t_j} 的波束位置信息。

2) Q 网络训练和更新。Q 网络作为动作价值函数，通过卷积神经网络（CNN, convolutional neural network）提取状态特征，可以将不同状态的图形张量输入，映射为不同动作对应的一组 Q 值。例如，Q 值 $Q(s_j,a_j)$ 表示在当前图形张量 s_j 输入下，执行动作 a_j 后的累计效益。

DRL-MOP 算法基于 Q-learning 的思想，通过利用经验池中存储的标签时延的数据，计算目标网络 Q' 函数的估计值和网络预测的 Q 值之间的误差值，进一步通过随机梯度下降法更新目标网络的权重和偏置等参数，达到动作值函数和累计收益的最优化。

其中，目标网络 Q' 函数值的计算通过如式(17)所示的贝尔曼（Bellman）方程计算，表征当前状态的动作值函数 $Q'(s_j,a_j)$ 由此时刻的即时奖励 r_j 加上下一个状态的最大化将来奖励值 $\max(Q'(s_{j+1},a_{j+1}))$。折扣系数 $\gamma\in[0,1]$ 是对将来奖励的一个折扣[19]。

$$Q'(s_j,a_j)=r_j+\gamma\max(Q'(s_{j+1},a_{j+1})) \tag{17}$$

误差函数为

$$L(\theta)=\mathrm{E}[(r_j+\gamma\max(Q'(s_{j+1},a_{j+1},\theta'))-Q(s_j,a_j,\theta))^2] \tag{18}$$

由式(18)可以看到，为了实现动作值函数的优化和逼近，必须尽可能地让误差函数趋近于 0，即 $\min(L(\theta))$。

3.3　步骤和流程

为了解决多波束卫星系统资源分配的序列决策问题，达到系统和用户的最大累计收益优化，本文提出了 DRL-MOP 算法。所提算法基本步骤如下。

1) 参数初始化。初始化仿真场景参数，初始分布 N 个用户；初始化所提算法 Q 网络参数。

2) 在时刻 $t_j\in\mathcal{T}$，判断如果用户 u_{t_j} 到达，且当接入波束有空闲载波时，则跳转到步骤 3)进行动作选择；当没有空闲载波时，拒绝该用户接入，跳转到时刻 t_{j+1}。如果用户 u_{t_j} 离开，则跳转到步骤 7)。

3) 构建马尔可夫过程。输入重构状态，求得 Q 值向量。根据 $\varepsilon-$贪婪算法选择动作 m 作为接入载波，更新状态数据 s_{j+1}，根据式(16)计算收益 r_j。

4) 经验回放。将经验信息 $\Phi(j)=\{s_j,a_j,r_j,s_{j+1}\}$ 存入经验池 D。当经验池的数据达到可以训练的数量时，随机抽取一组经验数据进行 Q 网络训练。

5) Q 网络训练。通过式(18)计算得到的误差函数 $L(\theta)$ 和随机梯度下降法反向修正 Q 网络参数。

6) Q 网络更新。每达到步长 G，更新目标网络参数 θ'。跳转到下一时刻，返回步骤 2)。

7) 删除用户占用的载波，更新状态。跳转到下

一时刻，返回步骤 2)。

8) 到达结束时刻 T，算法结束。返回优化结果。

3.4 算法复杂性

通常 DRL 算法能够通过卷积、降采样、权值统一等操作降低数据维度和计算复杂度，从而解决高纬度和复杂输入的问题，缩减计算开销和计算时间，以应用到解决在线资源分配问题中。所提算法采用 CNN 描述，包含卷积层和全连接层。使用卷积层和全连接层的时间复杂度来进行算法总体复杂度的描述。其中，卷积层的时间复杂度为

$$O\left(\sum_{l=1}^{L} K_l^2 H_l^2 C_{l-1} C_l\right) \tag{19}$$

其中，L 是卷积层数；K_l 是第 l 层的卷积核大小；H_l 是第 l 层的输出维度；C_l 是第 l 层的输出通道大小，即卷积核个数；C_{l-1} 是输入通道大小。

全连接层的时间复杂度为

$$O\left(\sum_{l=1}^{L'} 2X_l Y_l\right) \tag{20}$$

其中，L' 是全连接的层数，X_l 是第 l 层全连接层的输入，Y_l 是全连接层的输出。

算法总体复杂度为

$$O\left(\sum_{l=1}^{L} K_l^2 H_l^2 C_{l-1} C_l + \sum_{l=1}^{L'} 2X_l Y_l\right) \tag{21}$$

4 仿真和分析

4.1 场景和参数设置

仿真场景为 L 频段多波束卫星系统，采用单颗 GEO 卫星，主要参数依据 GEO 移动无线接口规范[20]给定。系统内初始随机分布 200 个用户，仿真开始后用户业务到达模型服从参数为 70 次/分钟的泊松分布。采用 MATLAB 平台进行仿真，使用其 DeapLearnToolbox 工具箱进行算法的实现和仿真。DRL-MOP 算法采用 CNN 描述，包含 2 个卷积层和 2 个全连接层。并采用经验回放的机制，经验池容量为 20 000，经验回放门限为 2 000。激活函数为 Sigmoid 函数。系统和算法仿真参数如表 1 所示。

表 1 系统和算法仿真参数

仿真参数	取值
下行工作频率/ MHz	1 542
波束个数/个	37
系统带宽/MHz	5
载波带宽/kHz	312.5
载波个数/个	16
最大天线增益/dBi	41.6
用户终端 EIRP/dBW	7、11
接收天线增益 G 与接收系统噪声温度 T 之比/(dB/K)	–24、–22
用户业务到达率/(次·分钟$^{-1}$)	70
卷积核 K_1	7
卷积核 K_2	2
卷积层 1 H_1	3
卷积层 2 H_2	2
卷积层 1 $C_0 \times C_1$	1×16
卷积层 2 $C_1 \times C_2$	16×32
全连接层 1 $X_1 \times Y_1$	128×128
全连接层 2 $X_2 \times Y_2$	128×16
采样批量	4
折扣因子	0.9
更新步长	50
学习率	0.001
初始探索概率	1
最终探索概率	0.01

4.2 Q 网络计算过程

本文所提算法的 Q 网络计算过程如图 3 所示。相关参数如表 2 所示。

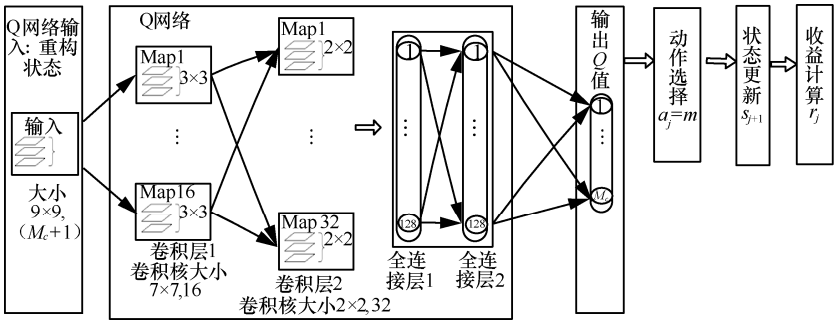

图 3 Q 网络计算过程

表 2　　Q 网络参数

网络层	输入	卷积核	偏置	权值	输出
卷积层 1	9×9，M_c+1	7×7，16	3×3，16	—	3×3，16
卷积层 2	3×3，16	2×2，32	2×2，32	—	2×2，32
全连接层 1	128	—	128×1	128×128	128
全连接层 2	128	—	16×1	16×128	M_c

Q 网络的第一层是输入层，输入是依据 3.2 节由状态重构得到的(M_c+1)张 9×9 大小的图形张量。输入数据通过 Q 网络的第二层和第三层卷积层，得到 32 张 2×2 大小的图层。第四层和第五层为全连接层，分别由 128 个神经元组成，最后与 M_c 个输出连接。第三层的输出经过向量化处理后，输入第四层和第五层，得到 M_c 个 Q 值的输出。接着，算法依据 ε – 贪心算法从输出中选择第 m 个 Q 值，执行为用户分配所接入波束的第 m 个载波的动作。同时更新状态数据，并依据式(14)~式(16)计算该动作的即时收益。

4.3　仿真结果分析

本节对比了本文所提 DRL-MOP 算法和传统的 SA 算法、GA，从不同角度验证了算法的综合性能。其中对比算法具体介绍如下。

SA 算法。在保障目标函数 SI 门限要求的情况下最优化 EE，算法流程参考文献[7,9]。算法初始温度 T_o =500，每个温度迭代次数 N_S =20，降温系数 α =0.97，结束温度 $T_{\text{end}}=1.75\times10^{-24}$ 。

GA。优化目标为 SE、EE、SI 的归一化权重和。算法父代个数 N_G =100，变异概率 p_m =0.001，交叉概率 p_c =0.06，迭代次数 M_G =500。

1) 系统和用户性能提升效果

不同功率分配下各目标的优化结果如图 4 所示。其中，所提算法 SE、EE、SI 的权重因子取值均为 $\frac{1}{3}$。可以看到，随着系统总功率限制从 300 W 增加到 1 000 W，不同算法下系统和用户性能的仿真结果随着系统功率分配的增加均呈现升高的趋势，可以得到更好的 SE、EE 和 SI。

图 4 中，本文所提 DRL-MOP 算法的结果最优，在功率限制为 1 000 W 时，该算法系统 SE 为 13.11 bit/(s·Hz)，高于 SA 算法的 8.71 bit/(s·Hz)和 GA 的 6.94 bit/(s·Hz)；EE 为 62.62 kbit/(s·W)，高于 SA 算法的 51.40 kbit/(s·W)和 GA 的 38.51 kbit/(s·W)；SI 达到 0.97，高于 SA 算法的 0.86 和 GA 的 0.78。

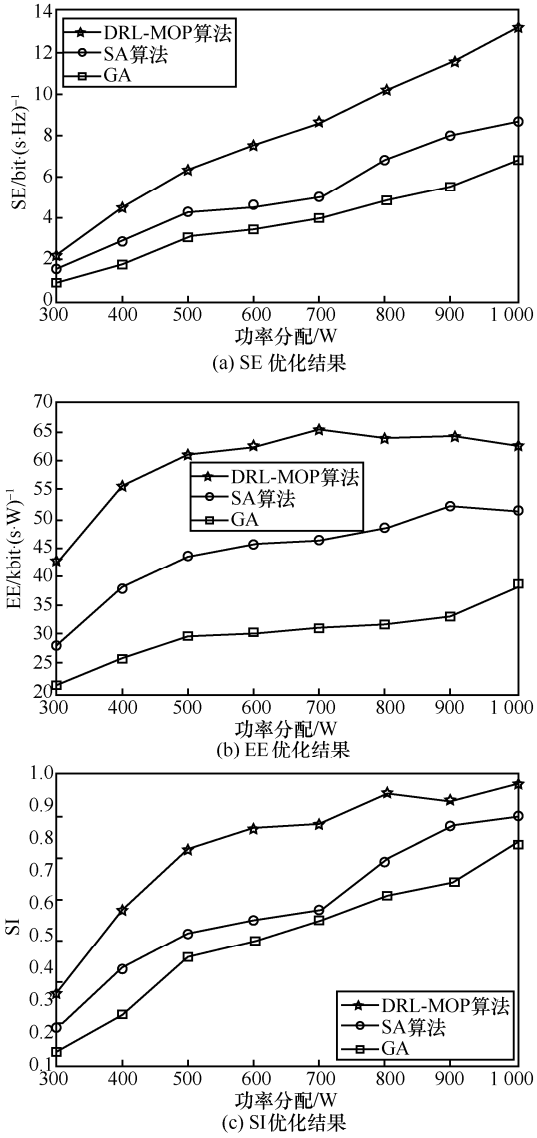

图 4　不同功率分配下各目标的优化结果

2) 算法收敛性

图 5 给出了 DRL-MOP 算法在系统总功率为 900 W 时目标函数值的收敛效果。

图 5 中横轴为业务到达次数，纵轴为依据式(14)求得的算法的总目标函数值。从图 5 可以看到，

算法在约 2 000 次时，性能未提高。这是由于经验值回放门限为 2 000，当经验条目数超过此阈值之后，算法才开始训练过程。在约 19 000 次以后，由于探索概率降低到最终探索概率，收敛曲线抖动变小。随着业务到达次数增加，当算法运行到约 55 000 次之后，目标函数值逐渐收敛，趋于稳定。

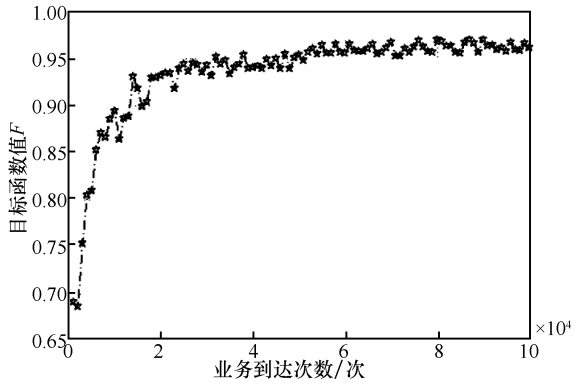

图 5　DRL-MOP 算法收敛趋势

3) 不同权重取值对优化结果的影响

图 6 给出了所提算法在 SE、EE、SI 的权重因子取值分别为 $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$（第一组）和 $\left(\frac{1}{4},\frac{1}{2},\frac{1}{4}\right)$（第二组）时，系统的吞吐量和功率性能。同时也给出了 SA 算法和 GA 的对比结果。从图 6 可以看到，随着系统分配功率的增加，各算法的总吞吐量均呈现增加的趋势，同时消耗的功率也都有所增加。

由图 6(a)可知，所提算法在不同权重设置下的总吞吐量均较高，明显优于 SA 算法和 GA，且在权重取值均为 $\frac{1}{3}$ 下的吞吐量最高。当功率分配为 1 000 W 时，第一组权重下的吞吐量为 65.56 Mbit/s，优于第二组权重下的 60.72 Mbit/s。这是由于第一组权重下，SE 和 SI 的权重更大，因此获得的 SE 和 SI 更高。由于仿真采用的用户业务到达模型和业务需求相同，在第一组权重设置下，用户的业务需求能够得到更好的满足。

由图 6(b)可知，所提算法的系统功耗在不同权重因子的设置下表现不同。当权重因子取值均为 $\frac{1}{3}$ 时，所提算法功耗高于 SA 算法和 GA。这是由于此时各权重相同，侧重于综合性能的优化。虽然功耗增加，但是吞吐量还是优于其他算法。从图 6 的仿真中可以看到，第一组权重设置下的各性能指标相比 SA 算法和 GA 具有明显优势。当功率分配为 1 000 W 时，DRL-MOP 算法功耗为 896.76 W，GA 为 751.90 W；SA 算法为 697.01 W。当权重取值改变为 $\frac{1}{4}$、$\frac{1}{2}$、$\frac{1}{4}$ 后，DRL-MOP 算法功耗有所降低，且当功率分配为 1 000 W 时，DRL-MOP 功耗为 731.08 W。仿真表明，当 EE 权重增大后，第二组权重下的系统 EE 更高，系统的功耗下降，并低于 GA 的功耗。可以看到，通过设置不同的权重，可以达到系统目标函数和性能的优化侧重。

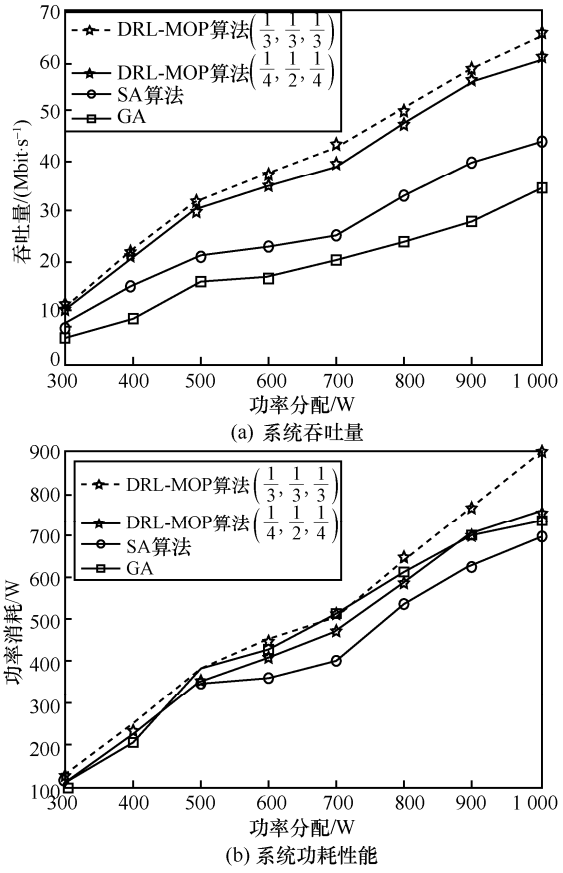

图 6　不同算法和参数设置下的系统吞吐量和功耗性能

4) 不同收益值对优化结果的影响

图 7 为系统总功率为 900 W、折扣因子为 0.9 时，在不同奖惩值取值下，目标函数值 F 的优化曲线。横坐标为奖励 r_h，取值范围为 0.1~0.9；纵坐标为 F 值。当 r_l=0.01 时，随着奖励 r_h 增大，所提算法更快地收敛到更优的结果。

4.4　复杂性分析

根据仿真场景和参数的设置，可以由 3.4 节中的式(19)~式(21)，计算得到所提算法的总体复杂度为 5.2×10^4，计算过程为

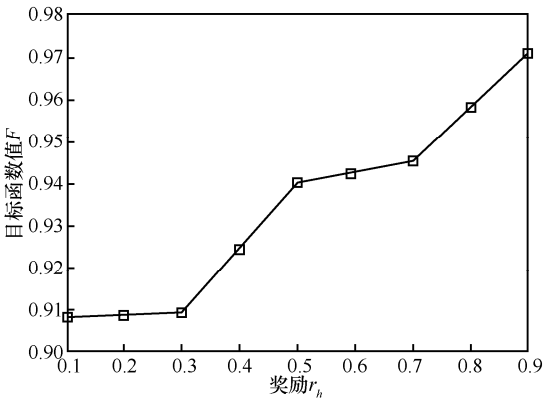

图 7　不同的奖惩值取值下的优化结果（r_f=0.01）

$$O\left(\sum_{l=1}^{L}K_l^2H_l^2C_{l-1}C_l+\sum_{l=1}^{L'}2X_lY_l\right)=$$
$$O(K_1^2H_1^2C_0C_1+K_2^2H_2^2C_1C_2+2X_1Y+2X_2Y)=$$
$$7^2\times3^2\times1\times16+7^2\times3^2\times1\times16+2\times128\times128+$$
$$2\times128\times16\approx5.2\times10^4 \tag{22}$$

依据 4.2 节中的参数，GA 的复杂度可以表示为

$$O(M_GN_G^2)=500\times100^2=5\times10^6 \tag{23}$$

SA 算法的复杂度可以表示为

$$O\left(N_S(\ln(\alpha))^{-1}\ln\frac{T_{\text{end}}}{T_o}\right)=$$
$$20\times(\ln(0.97))^{-1}\times\ln\frac{1.75\times10^{-24}}{500}\approx4\times10^4 \tag{24}$$

可以看到，相较于传统算法 SA 算法、GA，所提算法由于涉及卷积、神经元等操作，复杂度高于 SA 算法，但单次业务到达时复杂度低于 GA，并且表现出来的总体性能是最优的。此外，随着训练过程的收敛，所提算法的复杂度会有所降低，更能适应动态变化环境下系统累计性能的优化。

5　结束语

针对多波束卫星系统和用户性能多目标优化的问题，考虑深度增强学习在解决资源分配序列决策和求解累计性能优化时的优势，本文提出了一种基于深度增强学习和多目标优化改进的 DRL-MOP 算法。所提算法可以有效解决多波束卫星系统资源分配的序列决策问题，能够通过权重设置实现系统多目标的优化侧重，且收敛性好、复杂度低。

参考文献：

[1] 易克初, 李怡, 孙晨华, 等. 卫星通信的近期发展与前景展望[J]. 通信学报, 2015, 36(6): 161-176.
YI K C, LI Y, SUN C H, et al. Recent development and its prospect of satellite communications[J]. Journal on Communications, 2015, 36(6): 161-176.

[2] WANG C, CUI G, WANG W, et al. Joint estimation of carrier frequency and phase offset based on pilot symbols in quasi-constant envelope OFDM satellite systems[J]. China Communications, 2017, 14(7): 1-11.

[3] 史煜, 张邦宁, 郭道省, 等. 一种改进的多波束卫星通信系统功率分配算法[J]. 通信技术, 2016, 49(10): 1355-1359.
SHI Y, ZHANG B N, GUO D X, et al. A modified water-filling algorithm of power allocation for multi-beam satellite systems[J]. Communications Technology, 2016, 49(10): 1355-1359.

[4] ARTIGA X, NUNEZ-MARTINEZ J, PEREZ-NEIRA A, et al. Terrestrial-satellite integration in dynamic 5G backhaul networks[C]//The 8th Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop. Piscataway: IEEE Press, 2016: 1-6.

[5] 阚茜, 许小东. 一种能量和频谱效率兼顾的多波束卫星系统功率分配策略[J].中国科学技术大学学报, 2016, 46(2): 138-147.
KAN X, XU X D. Power allocation based on energy and spectral efficiency in multi-beam satellite systems[J]. Journal of University of Science and Technology of China, 2016, 46(2): 138-147.

[6] 阚茜. 衰落信道下波束卫星系统功率分配策略研究[D]. 合肥: 中国科学技术大学, 2016.
KAN X. Power allocation of multi—beam satellite system in fading channel[D]. Hefei: University of Science and Technology of China, 2016.

[7] COCCO G, DE COLA T, ANGELONE M, et al. Radio resource management optimization of flexible satellite payloads for DVB-S2 systems[J]. IEEE Transactions on Broadcasting, 2018, 64(2): 266-280.

[8] ARAVANIS A I, SHANKAR M R B, ARAPOGLOU P, et al. Power allocation in multibeam satellite systems: a two-stage multi-objective optimization[J]. IEEE Transactions on Wireless Communications, 2015, 14(6): 3171-3182.

[9] ZHANG P, WANG X, MA Z, et al. Joint optimization of satisfaction index and spectrum efficiency with cache restricted for resource allocation in multi-beam satellite systems[J]. China Communications, 2019, 16(2): 189-201.

[10] 廖晓闽, 严少虎, 石嘉, 等. 基于深度强化学习的蜂窝网资源分配算法[J]. 通信学报, 2019, 40(2): 11-18.
LIAO X M, YAN S H, SHI J, et al. Deep reinforcement learning based resource allocation algorithm in cellular networks [J]. Journal on Communications, 2019, 40(2): 11-18.

[11] HAN Z, LEI T, LU Z, et al. Artificial intelligence based handoff management for dense WLANs: a deep reinforcement learning approach[J]. IEEE Access, 2019, 7: 31688-31701.

[12] FAN H, ZHU L, YAO C, et al. Deep reinforcement learning for energy efficiency optimization in wireless networks[C]//The 4th International Conference on Cloud Computing and Big Data Analysis. Piscataway: IEEE Press, 2019: 465-471.

[13] FERREIRA P V R, PAFFENROTH R, WYGLINSKI A M, et al. Multi-objective reinforcement learning for cognitive satellite communications using deep neural network ensembles[J]. IEEE Journal on Selected Areas in Communications, 2018, 36: 1030-1041.

[14] HU X, LIU S, WANG Y, et al. Deep reinforcement learning based beam hopping algorithm in multibeam satellite systems[J]. IET Communications, 2019, 13(16): 2485-2491.

[15] HU X, ZHANG Y, LIAO X, et al. Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems[J]. IEEE Transactions on Broadcasting, 2019, doi: 10.1109/TBC.2019.2960940.

[16] HU X, LIU S, CHEN R, et al. A deep reinforcement learning-based framework for dynamic resource allocation in multibeam satellite systems[J]. IEEE Communications Letters, 2018, 22(8): 1612-1615.

[17] LIU S, HU X, WANG W. Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems[J]. IEEE Access, 2018, 6: 15733-15742.

[18] 刘帅军. 卫星通信系统中动态资源管理技术研究[D]. 北京: 北京邮电大学, 2018.
LIU S J. The research on dynamic resource management techniques for satellite communication systems[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[19] 彭伟. 揭秘深度强化学习[M]. 北京: 中国水利水电出版社, 2018: 266-291.
PENG W. Exploring deep reinforcement learning[M]. Beijing: China Water & Power Press, 2018: 266-291.

[20] ETSI. GEO-mobile radio interface specifications (Release 1): V1.3.1[S]. TS.101 376-5-5, (2005-02-11)[2019-12-20].

[作者简介]

张沛（1986– ），女，河南三门峡人，北京邮电大学博士生，主要研究方向为卫星通信、深度增强学习、神经网络等。

刘帅军（1988– ），男，河北邢台人，博士，中国科学院软件研究所助理研究员，主要研究方向为低轨星座网络、卫星 5G 融合、动态资源管理。

马治国（1978– ），男，北京人，中国信息通信研究院高级工程师，主要研究方向为 5G 通信、卫星通信等。

王晓晖（1972– ），男，浙江建德人，北京邮电大学讲师，主要研究方向为 5G 通信、卫星通信等。

宋俊德（1938– ），男，河北沧州人，博士，北京邮电大学教授，主要研究方向为智慧城市、5G 通信、卫星通信等。