Journal of System Simulation

Volume 35 | Issue 2 Article 18

2-16-2023

DQN-based Joint Scheduling Method of Heterogeneous TT&C Resources

Naiyang Xue
1.Graduate School, Space Engineering University, Beijing 101416, China;, 2163628670@qq.com

Dan Ding
2.Department of Electronic and Optical Engineering, Space Engineering University, Beijing 101416, China;, ddnjr@163.com

Yutong Jia
1.Graduate School, Space Engineering University, Beijing 101416, China;

Zhiqiang Wang
1.Graduate School, Space Engineering University, Beijing 101416, China;

See next page for additional authors

DQN-based Joint Scheduling Method of Heterogeneous TT&C Resources

Abstract

Abstract: Joint scheduling of heterogeneous TT&C resources as research object, a deep Q network (DQN) algorithm based on reinforcement learning is proposed. *The characteristics of the joint scheduling problem of heterogeneous TT&C resources being fully analyzied and mathematical language being used to describe the constraints affecting the solution, a resource joint scheduling model is established. From the perspective of applying reinforcement learning, two neural networks with the same structure and the action selection strategies based on* ε*greedy algorithm are respectively designed after Markov decision process description, and DQN solution framework is established.* The simulation results show that DQN-based heterogeneous TT&C resources scheduling method can identify a TT&C scheduling scheme with better scheduling revenue than the genetic algorithm.

Keywords

telemetry, track and command (TT&C), joint scheduling of heterogeneous TT&C resources, deep Q network, scheduling revenue, reinforcement learning

Authors

Naiyang Xue, Dan Ding, Yutong Jia, Zhiqiang Wang, and Yuan Liu

Recommended Citation

Naiyang Xue, Dan Ding, Yutong Jia, Zhiqiang Wang, Yuan Liu. DQN-based Joint Scheduling Method of Heterogeneous TT&C Resources[J]. Journal of System Simulation, 2023, 35(2): 423-434.

第 35 卷第 2 期 系统仿真学报© Vol. 35 No. 2
2023 年 2 月 Journal of System Simulation Feb. 2023

基于DQN的异构测控资源联合调度方法

薛乃阳[1]，丁丹[2*]，贾玉童[1]，王志强[1]，刘渊[3]
(1. 航天工程大学 研究生院，北京 101416；2. 航天工程大学 电子与光学工程系，北京 101416；
3. 中国人民解放军61646部队，北京 100192)

摘要： *以异构测控网资源联合调度为研究对象，提出一种基于强化学习的深度Q网络(deep Q network, DQN)算法。在充分分析异构测控资源联合调度问题特点后，用数学语言对影响问题求解的约束条件进行描述，建立了资源联合调度模型；从应用强化学习解决问题的角度，对求解的问题进行马尔科夫决策过程描述后，分别设计了2个结构相同的神经网络和基于ε贪婪算法的动作选择策略，并建立了DQN求解框架。*仿真结果表明：基于DQN的异构测控资源调度方法较遗传算法能够找到调度收益更优的测控调度方案。

关键词： *航天测控；异构测控资源联合调度；深度Q网络；调度收益；强化学习*

中图分类号：TP273+.1 文献标志码：A 文章编号：1004-731X(2023)02-0423-12

DOI: 10.16182/j.issn1004731x.joss.21-0879

引用格式: 薛乃阳, 丁丹, 贾玉童, 等. 基于DQN的异构测控资源联合调度方法[J]. 系统仿真学报, 2023, 35(2): 423-434.

Reference format: Xue Naiyang, Ding Dan, Jia Yutong, et al. DQN-based Joint Scheduling Method of Heterogeneous TT&C Resources[J]. Journal of System Simulation, 2023, 35(2): 423-434.

DQN-based Joint Scheduling Method of Heterogeneous TT&C Resources

Xue Naiyang[1], *Ding Dan*[2*], *Jia Yutong*[1], *Wang Zhiqiang*[1], *Liu Yuan*[3]

(1. Graduate School, Space Engineering University, Beijing 101416,China; 2. Department of Electronic and Optical Engineering, Space Engineering University, Beijing 101416, China; 3. PLA 61646 Troops, Beijing 100192, China)

Abstract: Joint scheduling of heterogeneous TT&C resources as research object, a deep Q network (DQN) algorithm based on reinforcement learning is proposed. *The characteristics of the joint scheduling problem of heterogeneous TT&C resources being fully analyzied and mathematical language being used to describe the constraints affecting the solution, a resource joint scheduling model is established. From the perspective of applying reinforcement learning, two neural networks with the same structure and the action selection strategies based on ε greedy algorithm are respectively designed after Markov decision process description, and DQN solution framework is established.* The simulation results show that DQN-based heterogeneous TT&C resources scheduling method can identify a TT&C scheduling scheme with better scheduling revenue than the genetic algorithm.

Keywords: telemetry, track and command (TT&C); joint scheduling of heterogeneous TT&C resources; deep Q network; scheduling revenue; reinforcement learning

0 引言

随着越来越多不同属性和种类的航天器进入太空[1]，尤其是低成本微小卫星星座型系统的大量出现[2]，在轨航天器数量急剧增长，使国家航天测控网面临越来越严重的资源冲突和争用的问题。同时，民用、商用航天测控公司建立并运营的测

收稿日期：2021-08-31 修回日期：2021-10-11
第一作者：薛乃阳(1997-)，男，硕士生，研究方向为测控与通信技术。E-mail：2163628670@qq.com
通讯作者：丁丹(1980-)，男，副研究员，博士，研究方向为测控与通信技术。E-mail：ddnjr@163.com

控网也正在不断发展壮大。国有测控网采用封闭的服务器/客户端模式的服务器架构，而商业测控网通常采用开放的卫星任务中心支持对外联网的新模式[3]。对资源归属属性不同且异构的测控网中测控资源进行合理调度和管理，可以提高对现有测控资源的利用效率。

异构测控资源联合调度是对国有或民用、商用测控资源进行分析的基础上，在测控资源有限并且分别属于不同构型测控网的条件下，根据多类测控网络的能力特点和服务偏好，合理分配测控资源和时间，以达到最大化满足所有测控需求，并发挥出不同测控网资源组合利用的最大效益。但是目前大多数研究都集中在国家航天测控网单个管理体制下的资源调度规划，针对包含国有和商用测控资源组成的异构测控资源联合调度研究较少。文献[4]通过建立的异构测控资源调度模型，用遗传算法求解并证明了对异构测控资源采用联合调度方法较不采用联合调度，可以显著提高测控需求满足率和生成的最优调度方案的测控收益。Stottler等[5]针对美国空军卫星控制网任务调度问题，通过引入人类调度员的启发式经验，研制了MIDAS (managed intelligent deconfliction and scheduling)系统，提高了任务规划和解冲突效率。张天娇等[6]对天地基测控资源联合调度问题采用遗传–蚁群混合优化算法进行求解，提高了求解性能。李长德等[7]针对多星测控调度建立并训练相应的深度神经网络模型，提高了调度方法的任务满足率并缩减了程序运行时间。

从已有的研究可以看出，虽然遗传算法及其改进算法已经可以应用于异构测控资源联合调度问题的求解，但是随着测控场景的不断扩大以及实际测控任务中可能出现的应急测控任务等特殊情况，同时遗传算法还存在算法早熟的问题[8]，使得问题的求解效果不能满足实际需求。强化学习(reinforcement learning, RL)是机器学习的重要分支之一[9]。本文提出了一种基于强化学习的DQN (deep Q network)算法来解决异构测控资源联合调度问题。

1 异构测控资源特点分析与建模

由于RL具有与环境灵活交互实现自我寻优的特性，在使用DQN算法求解异构测控资源联合调度问题时，异构测控资源调度环境建模的质量对最终算法的求解效果有重要的影响。为了能够建立准确的异构测控资源调度环境模型，本文在归纳出异构测控资源特点的基础上，对其涉及到的主要实体要素进行定义和建模，分析主要约束条件，并对约束目标进行数学化描述。

1.1 异构测控资源特点分析

异构测控资源由国家航天测控基础设施与民用、商用测控网这些不同归属属性和架构的测控资源构成。国家测控基础设施是我国专门由国家机构设立的航天测控网。民营测控资源主要由商业公司或者民间组织运营的地面测控网。由于不同测控网的主要服务对象和目的不同，测控网运行架构、数据传输协议以及测控任务优先级排序等方面都有较大的差异，造成其网络的异构性。与单个测控网内的测控资源调度相比，具体的不同点主要体现在4个方面[10]：

(1) 测控网运营管理模式不同。虽然国有测控网与民商测控网都是由测控中心与分布在各地的测控站构成，但是商业测控网的测控管理模式更多元。传统的国有测控网采用封闭的服务器/客户端模式的服务器架构，测控中心与测控站通过专线相连接，执行任务的安全性和保密性更高；商用测控网采用“软件即服务”的基于云计算的开放式架构，将测控中心建立在公共互联网上，网络节点增多，较专用网络传输时延变长，但是建设和运营成本更低。

(2) 不同测控网对同一个测控任务的优先级排序不同。不同测控网对不同属性的测控任务存在不同偏好，因此表现为对测控任务优先级排序存

DOI: 10.16182/j.issn1004731x.joss.21-0879

在差异。比如，国有测控网不考虑执行任务时的效费比，并且在同等条件下要优先保障国家级的航天任务，确保其可靠、安全地执行。而非国有测控资源在注重任务可靠性的同时，也同样重视执行任务的成本和收益，以期能够提高公司的市场竞争力。因此，同一个测控任务在不同属性的测控网中被选择的优先级不同。

(3) 建站范围和方式不同。我国的国有测控资源主要依靠自建，由陆基、天基(中继卫星)和海基(测量船)测站组成。其中，陆基站主要分布在我国境内，建立的海外站较少，因此，覆盖的圈次和范围比较有限。虽然测量船可以作为海外机动站，但是成本和费用较高，受海况和天气影响，不能长期固定在某区域执行任务。而民商用测控公司可以较方便地选择在境外的全球范围内建站，在建站的方式上不仅可以自建测控站，还可以租用或者开展国际合作请求测控支持。

(4) 不同测控网的测控能力存在差异。目前，我国航天测控网具备S频段、C频段、X频段、Ka频段和超短波等5种频段测控支持能力。非国有测控网使用的频率范围受法律等政策的限制，比如，在我国S频段上行信号属于国家资源，主要用于国防或政府主导的民用卫星测控，而民用和商业测控公司只能发送X频段的上行指令，非国有测控网所属测站不能在没有官方允许的情况下私自向卫星发射S频段上行信号。

1.2 变量及符号定义

参考对异构测控资源联合调度已有的研究成果[6-7, 10]，测控调度场景中包含的主要变量及符号进行如下定义：

(1) 卫星

卫星集合$S=\{S_1, S_2, \cdots, S_s\}$中包含$s$颗卫星。$\forall S_k \in S$，$S_k=(SID_k, ST_k, SA_k, SF_k, SD_k, SN_k, SP_k)$。$SID_k$为卫星标识符；$ST_k$为卫星归属属性，$ST_k \in \{1, 2\}=\{$国有卫星，商用卫星$\}$；$SA_k$为此卫星过境所有测站的测控弧段集合；$SF_k$为用户卫星测控频段，有$SF_k \in \{1, 2, 3\}=\{$S，X，VHF$\}$，分别对应S、X和VHF频段；$SD_k$为卫星测控任务的最短持续时间；$SN_k$为该卫星一个调度周期(圈次)内所要求的测控次数；$SP_k$为卫星优先级。

(2) 测控站

测控资源集合$R=\{R_1, R_2, \cdots, R_m, \cdots, R_r\}$，测控站总数为$r$。其中，前$m$个测控站为国有测控站，其余为商用测控站。$\forall R_k \in R$，$R_k=(RID_k, RT_k, RA_k, RF_k, RC_k)$。其中，$RID_k$为测控站代号；$RT_k$为测控资源类型，$RT_k \in \{1, 2\}=\{$国有测控站，商业测控站$\}$；$RA_k$为过境此测站的所有可用测控弧段；$RF_k$为测控频段，对应用户卫星测控频段；$RC_k$为测控站设备最短转换时间。

(3) 测控需求

测控需求集合$Req=\{Req_1, Req_2, \cdots, Req_u\}$，卫星提出的所有测控需求的总个数为$u$，$\forall Req_k \in Req$，$Req_k=(reNo_k, reSnum_k, reT_k, reS_k, rePri_k, reF_k)$。其中，$reNo_k$为测控需求代号；$reSnum_k$为提出需求的卫星编号；$reT_k$为任务执行期限，由起始时间与终止时间定义的时间段来表示；reS_k为是否属于S频段上行任务，$reS_k \in \{1, 0\}=\{$S频段上行任务，非S频段上行任务$\}$；$rePri_k$为测控任务优先级；$reF_k=\{1, 2\}=\{$国有测控网，商业测控网$\}$，为测控需求固定的资源属性偏好，表示由于测控能力或者测控成本等约束条件的限制，只能由国有测控网或商用测控网提供测控服务。

(4) 测控可用弧段

测控弧段集合$A=\{A_1, A_2, \cdots, A_s\}$。$\forall A_k \in A$，$A_k=(AID_k, AS_k, AR_k, AW_m^s, AW_m^e, A_kt_l^s, A_kt_l^e, APr_k)$。其中，$AID_k$为测控弧段编号值；$AS_k$和$AR_k$分别为此弧段中卫星和测控资源代号；$AW_m^s$和$AW_m^e$分别为计算出的弧段中可见时间窗口理论开始与结束时间；APr_k为对该测控弧段进行调度的收益，也称为测控弧段被选择优先级，其数值由弧段中的

卫星与测控站属性来确定。

(5) 调度问题决策集合

调度问题决策集合 $Ad=\{Ad_1, Ad_2, \cdots, Ad_u\}$。$\forall Ad_k \in Ad$，$Ad_k=(\xi_{ik}, \theta_{ikl}, A_i t_l^{\mathrm{s}}, A_i t_l^{\mathrm{e}}, A_i)$。其中，$\xi_{ik}=1$ 表示选定测控弧段 A_i 执行测控需求 Req_k 的测控任务，否则 $\xi_{ik}=0$；$\theta_{ikl}=1$ 表示测站 R_l 完成测控需求 Req_i 的测控任务后，经过设备转换的时间后即转入对 Req_k 的测控任务，否则 $\theta_{ikl}=0$；$A_i t_l^{\mathrm{s}}$ 与 $A_i t_l^{\mathrm{e}}$ 分别表示测控弧段 A_i 在测控站 R_l 中测控任务实际开始时间和结束时间。

由于测控任务弧段与卫星、测控资源与任务执行时间窗口都是唯一对应的，且可见弧段中基本包含了调度场景中的所有信息，因此，本文选择测控弧段作为直接调度对象。

1.3 约束模型

在异构测控资源联合调度过程中要考虑的约束具体如下：

(1) 时间窗口有效性约束。所有的测控任务都必须在卫星与测控站可见时间窗口内进行；同时，对于每一个测控站，当执行完一个测控任务后，需要设置设备复位与下次任务的准备时间 RT_m，才能执行下一个测控任务。

$$C_1=\{\forall A_i \in A, \exists J_k \in J: AW_i^{\mathrm{s}} \leqslant A_k t_l^{\mathrm{s}} \leqslant A_k t_l^{\mathrm{e}} \leqslant AW_k^{\mathrm{e}}\} \tag{1}$$

$$C_2=\{\forall R_k \in R, m\text{为整数}, \text{且 } \forall 0<m<\left|RJ_k\right|: A_k t_{m+1}^{\mathrm{s}}-A_k t_m^{\mathrm{e}}+RT_m \geqslant 0\} \tag{2}$$

(2) 测控频段一致性约束。在测站 R_k 的所有可测弧段 RA_k 中，测控资源 R_k 可发送与接收的测控频段 RF_k 要包括卫星 S_l 所需频段 SF_l，在此弧段内才能建立测控链路。

$$C_3=\{\forall A_i \in A, A_i \in SA_k, A_i \in RA_l, \text{则: } SF_k \in RF_l\} \tag{3}$$

(3) 测控设备独占性约束。每一个测控设备在同一时刻只能对一颗卫星提供测控服务。对于决策集合中的任意 2 个测控可用弧段，若这 2 个弧段中测控站相同，那么这 2 个弧段的实际调度时间窗口不能有任何交集。

$$C_4=\{\forall A_i, A_j \in Ad, \text{若}\xi_{ik}=1 \text{且}\xi_{jk}=1, \text{同时}AR_i=AR_j, \text{则: } [A_i t_l^{\mathrm{s}}, A_i t_l^{\mathrm{e}}] \bigcap [A_j t_l^{\mathrm{s}}, A_j t_l^{\mathrm{e}}]=\varnothing\} \tag{4}$$

(4) 测控任务执行次数约束。每一个测控任务最多只能被完成一次，不能再参与后续的调度任务；每个测控任务的执行时间必须不能短于最短测控时间。

$$C_5=\{\forall S_i \in S, A_p, A_q \in SA_i: (AS_p=AS_q) \bigcap (AR_p=AR_q) \bigcap (AW_p^{\mathrm{s}}=AW_q^{\mathrm{s}})=\varnothing\} \tag{5}$$

$$C_6=\{\forall J_k \in SJ_k, SD_k \leqslant A_k t_l^{\mathrm{e}}-A_k t_l^{\mathrm{s}}\} \tag{6}$$

(5) 异构测控网服务偏好约束。在异构测控资源联合调度中，卫星属性与测控资源属性相同的测控弧段优先级要比卫星与测站属性不同的弧段优先级高。

$$C_7=\{\forall A_k \in A: \text{if } RT_k=ST_k, APr_k \in A_k; \text{if } RT_k \neq ST_k, APr_l \in A_k; \text{则} APr_k>APr_l\} \tag{7}$$

(6) 异构测控网测控能力约束。异构测控网中不同设备具有不同的性能和权限，比如，对于需要执行 S 频段上行信号的测控任务，只能由国有测控资源来完成；同时，由于测控需求对网络安全性、响应时间、测控成本的特殊需要，某些测控任务只能由指定的符合条件的测控网下属测控站执行。

$$C_8=\{\text{if } reS_k=1, \exists A_l \in SA_k, A_l \in RA_m, SF_k=1: RF_k=1, reF_l=1\} \tag{8}$$

$$C_9=\{A_k \in RA_l, \text{且}A_k \in SA_m, \text{当测控需求对执行任务的测控网有特殊需求时: } reF_l=i\}, i=1, 2 \tag{9}$$

异构测控资源联合调度的目标是根据异构测控网络能力特点和测控需求，尽可能提高异构测控网之间任务安排的合理性，从而发挥出不同测控网资源组合利用的最大效益。本文建立了以调度问题决策集合中测控弧段的任务综合优先级 JPr_k 数值之和为基础的测控调度方案评价指标。

DOI: 10.16182/j.issn1004731x.joss.21-0879

JPr_k的定义如式(10)所示，表示卫星S_i在测控弧段A_l下的测控任务综合优先级。

$$JPr_k = SP_i + APr_l,\ A_l \in Ad \tag{10}$$

考虑到当测控资源有限而不足以满足所有测控需求，即调度过程中出现的资源冲突的情况，本文对调度问题决策集合中没有被满足的测控任务定义了负的调度收益。设调度方案中被满足的测控需求数量为*actSched*，而测控需求总数量为*u*，测控需求未被满足的卫星优先级为*SP*。综上，在基于上述约束的基础上，该异构测控资源联合调度问题的目标函数为

$$Obj = \sum_{k=1}^{actSched} JPr_k - 5 \times \sum_{s=1}^{u-actSched} SP_s \tag{11}$$

2 基于DQN的异构测控资源联合调度算法实现

异构测控资源联合调度可以视为给卫星提出的测控需求依次分配满足条件的最优测控弧段的问题，因此是一个多阶段的序贯决策问题。由于RL方法可以实时获取当前环境的反馈信息，并且还能在没有事先启发式经验的条件下自主学习，根据当前环境情况灵活调整最优决策方案，很适合于解决序贯决策问题。本文研究的问题具有状态空间巨大和资源属性繁杂的特点，故采用DQN算法来进行问题求解。

2.1 DQN方法概述

DQN是一种基于值的时序差分学习算法，通过引入神经网络代替反馈评价当前状态下所有可选动作好坏的数值表格(Q值表格)，克服了Q学习在状态空间较大的问题上运行效率较低的缺点，更适合于求解大规模状态空间的问题。强化学习的基本要素包括状态、动作、策略、奖励函数，通过智能体与环境的交流反馈，从随机的不断尝试中学习到最优策略，并以此来获取最大化长期累积回报的机器学习方法[11]，其智能体与环境的交流反馈过程如图1所示。

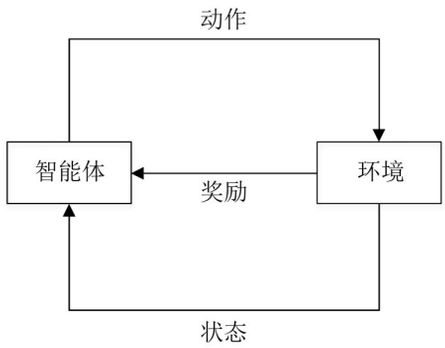

图1 智能体与环境的交流反馈过程
Fig. 1 Communication and feedback process between agent and environment

RL中智能体与环境的交互过程如下：首先，智能体通过感知环境得到环境的一个状态s；其次，智能体根据某个决策规则选择一个动作a；最后，动作执行完后改变环境状态，下一个时刻智能体通过从环境中获取一个奖励r来对其决策规则进行修正。由于每一个状态下的每一个行为都将获得累积奖励，即为对应的$Q(s, a)$，在复杂问题中为了避免过于庞大的Q值表格造成的“维度灾难”，在DQN算法中，将状态和动作作为神经网络输入，将训练好后神经网络的输出作为动作的Q值。需要注意的是，DQN的学习过程不再是训练Q表，而是训练神经网络中的权重参数，其训练方式为最小化目标Q值与预测Q值的损失函数：

$$L(\omega) = \mathrm{E}[(r(s, a, s') + \gamma \max_{a'} Q(s', a'; \omega') - Q(s, a; \omega))^2] \tag{12}$$

2.2 马尔科夫决策过程描述

利用DQN方法研究异构测控资源联合调度问题，首先需进行异构测控资源联合调度问题向马尔科夫决策过程的转化，从而方便应用DQN算法求解。异构测控资源联合调度问题具有求解空间巨大，资源争用冲突高，约束条件繁杂的特点。本文从RL的角度对调度问题进行了抽象，为了将其转化成一个序贯决策问题，引入“零动作”的概念，将复杂的异构测控资源调度问题抽象成卫星测控需求按照固定次序和优化规则依次选择与其要求相匹配的测控弧段，从而达到总收

益值最高的优化问题。此类序贯决策问题可以用马尔科夫决策过程进行描述，进行马尔科夫决策过程描述需要根据调度问题特点，定义决策时刻集，并分别设计测控状态、测控动作和即时奖励函数。

2.2.1 决策时刻集

在传统的测控资源调度问题的求解中，通常按照一定的启发式经验选择某些特殊的测控弧段优先或者推后为其安排测控任务[12]，但是本文利用DQN算法智能寻优的特点，设计了描述调度问题求解的决策时刻集，在测控动作中引入“零动作”，降低了问题求解的难度。

决策时刻集是指在一次完整的序贯决策问题迭代周期中，智能体所要选择动作的所有决策时刻。本文将所有卫星测控需求都赋予其唯一的ID值 $reNo_k$，按照编号顺序，逐次选择与其要求相匹配的测控弧段，将这种每次需要开始选择测控可用弧段的时刻称为一个决策时刻 t，这样从第一颗卫星的第一个测控需求开始到最后一颗卫星的最后的一个测控需求选择可用弧段的所有决策时刻构成了决策时刻 T。同时，为了克服按照固定顺序选择测控可用弧段方法的缺点，本文在每一个需要选择测控动作的决策时刻，都在其可选动作空间中添加了“零动作”，这样就可以避免选择先后顺序对算法生成优化调度方案的影响。若设所有的测控需求个数为 $|Req|$，则 T 可以表示为 $T=\{1, 2, \cdots, |Req|\}$。

2.2.2 测控状态设计

状态空间是调度过程中所有状态的集合，主要用来描述测控调度环境的主要特征和变化。测控状态的设计需要与调度目标相关，状态特征作为状态属性的数值表示，需要便于计算和表示。因此，本文从测控弧段可用时间、已规划测控任务分布情况和每个测控站的时间利用率三个方面入手，用一个维数为 r(场景中测控站数量)的列向量组 $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3$ 来表示当前情况下的测控状态 $\boldsymbol{s}$，即 $\boldsymbol{s}=[\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3]$。

(1) $\boldsymbol{f}_1$：当前状态下，每一个测控站在其所有可用弧段中的总时间列向量，即 $\boldsymbol{f}_1=\{\alpha_1, \alpha_2, \cdots, \alpha_r\}^{\mathrm{T}}$。列向量中元素 α_i 表示测控站 R_i 在当前状态下与已规划测控任务不冲突的所有可用测控弧段时间窗口长度的总和。设 RA_i' 为测控站 R_i 当前可用弧段集合，其弧段个数为 $|RA_i'|$。则 α_i 计算式为

$$\alpha_i=\sum_{i=1}^{|RA_i'|}(AW_i^{\mathrm{e}}-AW_i^{\mathrm{s}}) \tag{13}$$

(2) $\boldsymbol{f}_2$：当前状态下，目前调度决策中每个测控站已规划的测控任务数量列向量，即 $\boldsymbol{f}_2=\{\beta_1, \beta_2, \cdots, \beta_r\}^{\mathrm{T}}$。列向量中元素 β_i 表示测控站 R_i 在目前状态下已经执行的测控动作个数。设当前状态下的决策时刻时间步长为 t、目前 R_i 已经规划的测控动作个数为 $|R_iA_t'|$，则 β_i 计算式为

$$\beta_i=|R_iA_t'| \tag{14}$$

(3) $\boldsymbol{f}_3$：当前状态下，在已有的测控调度决策序列中，每个测控站执行任务时间段长度与当前总调度时间的比值列向量，即 $\boldsymbol{f}_3=\{\eta_1, \eta_2, \cdots, \eta_r\}^{\mathrm{T}}$。列向量中元素 η_i 表示测控站 R_i 在当前决策时刻 t 下，此测控站已规划的测控任务时间长度与所有测控站执行任务的总时长的比值。设当前状态下，测站 R_i 已规划任务集合为 R_iA_t'，集合中元素个数为 $|R_iA_t'|$。则 η_i 计算式为

$$\eta_i=\frac{\sum_{s=1}^{|R_iA_t'|}(A_st_i^{\mathrm{e}}-A_sW_i^{\mathrm{s}})}{\sum_{k=1}^{r}\sum_{s=1}^{|R_kA_t'|}(A_st_k^{\mathrm{e}}-A_sW_k^{\mathrm{s}})} \tag{15}$$

2.2.3 测控动作设计

整个异构测控系统运行过程中，测控环境所有可能经历的状态构成了状态集 S，对于状态集中任意一个状态 s，都应该对应着可以选择的动作集合 $A(s)$。本文将每个测控可用弧段按照时间先后关系都赋予了唯一代号 AID_k，作为测控动作集合。但是，在测控资源有限，不能全部满足提出的测控

DOI: 10.16182/j.issn1004731x.joss.21-0879

需求时，需要在测控动作中定义放弃满足某个测控需求的情况。同时，由于舍弃不同的测控需求对总调度收益造成的影响也不一样，为了保证每个测控需求都有可能被舍弃，在每一个决策时刻可选的动作中都要加上放弃当前测控需求的选项，将此时的舍弃动作代号定义为0，即为“零动作”，代表放弃满足当前测控需求。因此，定义动作$a_k=AID_k$。

2.2.4 即时奖励函数设计

即时奖励函数选取适当影响着DQN算法的求解效果，短期奖励可以视为采取的动作对调度方案的当前短期影响，累积奖励是每个动作对整个调度过程产生的奖励总和，DQN算法的最终目的就是最大化累积奖励。

本文的调度目标是在提高测控任务满足率的同时尽可能提高异构测控资源的利用效率，这一目标可以通过最大化测控计划总收益值函数进行转换。收益函数的定义如式(10)所示，由于当有测控需求没有被满足(即$a_m=0$)时，当前收益值为负；同时，每个测控动作的收益值设置也遵循了约束C_7。因此，测控总收益值越大，测控任务满足率越高，对异构测控资源的调度方案也就越合理。而对于每个测控决策时刻选择的动作而言，不仅仅需要考虑当前动作的测控收益，同样也要考虑到当前动作对后续测控动作可选测控弧段个数的影响，因此，奖励函数设计为

$$R=JPr_k+\mu\times\sum_{i=1}^{|R|}\sum_{j=1}^{|RA_i|}\left|AW_j^{\mathrm{e}}-AW_j^{\mathrm{s}}\right| \tag{16}$$

其中，JPr_k为当前动作的测控任务综合优先级，μ为一个常系数，用来调节权重大小，后面的多项式表示采取该动作，排除动作空间中与决策序列相冲突的所有可用测控弧段后，所有剩余的测控弧段时间段总长度。

2.3 算法流程

2.3.1 神经网络的输入与输出

根据2.2节中设计的测控状态和动作的编码方式，测控状态用向量组$\boldsymbol{s}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3]$表示，本文建立了BP神经网络结构，如图2所示。神经网络由输入层、隐藏层、输出层3部分组成。输入层神经元分别对应着状态$\boldsymbol{s}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3]$的所有特征向量以及测控动作$a_k=AID_k$；输出层神经元对应着动作空间，计算输出每一个状态–动作对的价值$Q(s,a)$。

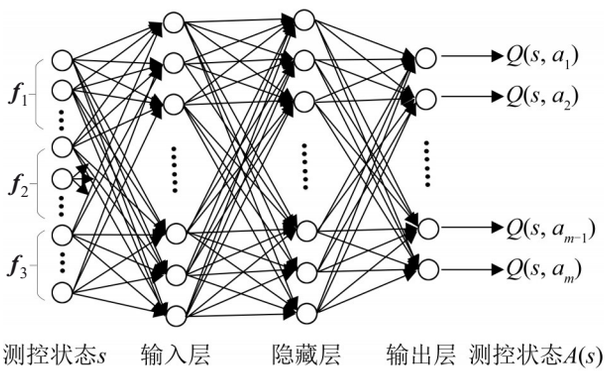

图2 神经网络结构
Fig. 2 Neural network structure

2.3.2 动作选择策略

在RL算法中，动作选择策略要处理好探索与利用的关系。探索是指智能体随机尝试不同的动作，寻找探索是否存在更有利的信息；利用是指智能体根据当前已知的信息进行最佳决策。为了使探索和利用次数达到平衡，获得最大化的累积奖励，本文使用ε–贪婪算法作为平衡探索与利用的平衡策略，其数学表达式为

$$A=\begin{cases}\mathrm{argmax}_a Q(s|a),\ p=1-\varepsilon\\ \text{随机选择动作}a,\ p=\varepsilon\end{cases} \tag{17}$$

本文采用不断衰减的ε-greedy策略，在每一次环境处于初始状态时，定义$\varepsilon=1$，即智能体用100%的概率随机选择搜索动作，以满足智能体对环境的感知和探索需要。当训练迭代次数增加时，ε值不断减小至0，即智能体按照100%的概率按照已有的学习经验对下一步采取的动作进行决策。ε值的衰减：

$$\varepsilon=\mathrm{e}^{-\rho E_i} \tag{18}$$

式中：E_i为第i个测控决策时刻；ρ为0~1之间的

常数，用来调节ε值衰减的快慢。

2.3.3 算法求解框架

本文提出的基于DQN算法的异构测控资源调度流程如图3所示，具体的算法步骤如下。

step 1：通过STK仿真软件计算所有测控弧段过境信息，并将其存入测控弧段集合A中。

step 2：定义算法中学习阈值N、目标神经网络更新频率T_renew；初始化动作值函数Q，设置ε贪婪的衰减权重ρ；设置学习率ga，经验库容量$rmemo$等参数。

step 3：构建并初始化2个相同结构的神经网络，预测Q网络Q_eval和目标Q网络Q_target，并对其设置相同的网络参数，将神经网络的训练方法设置为随机梯度下降法。

step 4：开始一次任务调度过程，初始化测控环境转态s，从第一颗卫星的第一个测控需求开始分配测控弧段。

step 5：按照2.3.2中的策略选择测控动作a，由于设置了代号为0的测控动作，消除了测控需求调度先后次序对最优方案的影响。

step 6：根据选择的测控动作和当前的测控状态，刷新测控弧段集合，删除与已选择测控弧段相冲突的待选测控弧段，使调度方案满足约束条件，并计算新状态s_1。

step 7：计算环境对动作a的反馈奖励r，同时把获得的经验(s, a, r, s_1)插入到经验库中，若经验库达到最大容量$rmemo$，则令最先插入的经验被最新的经验覆盖。

step 8：当经验库中的经验数量达到开始学习阈值N时，从经验库中随机抽取小批量样本，计算预测Q值Q_{eval}和目标Q值Q_{target}；若没有达到，转到step 5。

step 9：根据Q_{eval}、Q_{target}以及式(12)，计算损失函数；更新Q_eval参数，学习次数加1。

step 10：当学习次数达到更新频率T_renew时，同步预测Q网络与目标Q网络的参数；否则，若此时已经没有待分配的测控需求，迭代次数增加1，即完成了一套测控调度方案，当总迭代次数还没达到最大迭代次数时，转至step 4；若达到最大迭代次数，选择所有迭代过程中形成的所有调度方案中测控总收益最高的方案，算法结束。

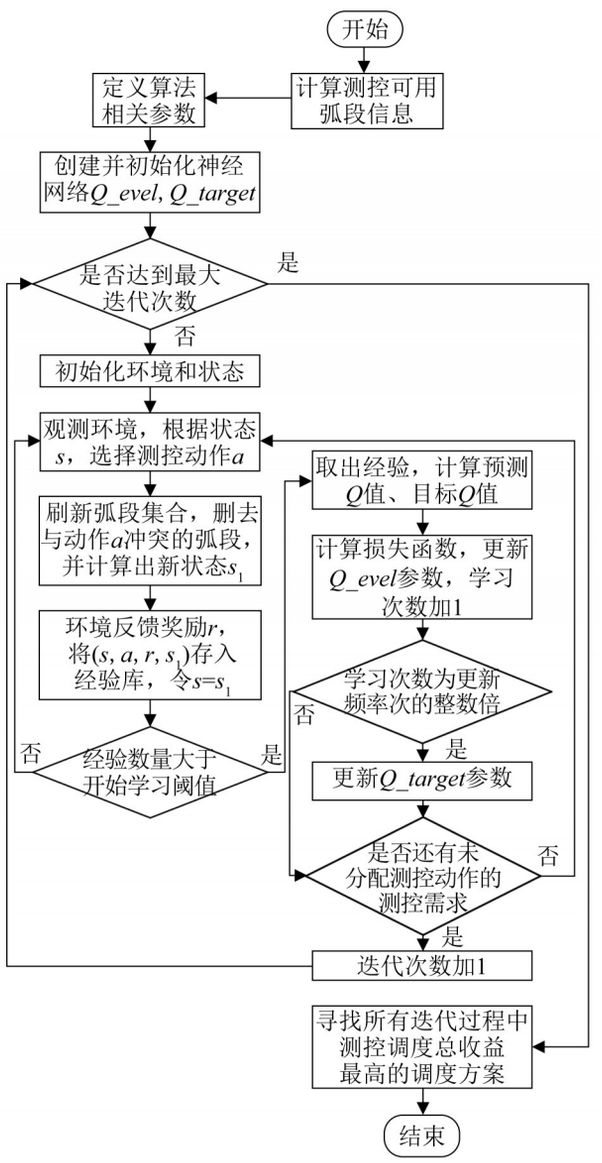

图3 算法流程
Fig. 3 Algorithm flowchart

3 实验仿真和分析

本文结合目前异构测控网中陆基测控站分布情况对算法进行实验仿真。同时，通过对比遗传

算法与DQN算法求解异构测控资源联合调度问题的求解性能，验证DQN的效能。其中，遗传算法的评价指标与本文相同，都为式(11)，遗传算法是测控资源调度问题中的一种常用算法[4]。通过比较基于DQN算法与遗传算法中随着算法迭代次数变化的测控总收益值变化情况以及最终求得的测控资源最佳调度方案，可以验证DQN算法的可行性和有效性。

3.1 实验场景

创建一个仿真场景，总调度时间区间为2021-08-20 T 00:00:00—21 T 00:00:00。利用STK软件模拟我国24颗真实运行中的低轨卫星，同时在场景中建立6个地面站，并通过此软件计算可见弧段信息。在场景中设置的24颗卫星中，有16颗是国有卫星，8颗属性定义为商用卫星，测控弧段最短持续时间SD_k设置为8 min，根据约束C_7，定义每个测控可用弧段的调度收益(弧段被选择优先级)如表1所示。每颗卫星在调度时间内设置4~6个测控需求，时间窗口设置与场景总调度时间相同，每颗卫星的优先级SP_k设置为范围在3~5的常数，所有卫星一共提出了128条测控需求；同时，随机选择5颗国有卫星需要执行S频段上行任务，测控任务只能由国有测控网测站执行，选择4颗商业卫星由于运营成本约束，各有2次任务只能由商业测控网测站执行。在建立的6个测站中，佳木斯站、喀什站和三亚站是国有测站，帕劳站、呼伦贝尔站和老挝站是商业测控站。由式(11)可知，测控任务综合优先级数值JPr_k由可用弧段调度收益值与卫星优先级数值相加而成。

表1　测控弧段调度收益
Table 1　TT&C act income statement

测控资源属性ST_k	卫星属性RT_k	弧段调度收益APr_k
1	1	10
1	2	5
2	1	6
2	2	10

3.2 模型参数设置

将提出的异构测控资源联合调度DQN算法在MATLAB2019a中编程实现，硬件配置为Inter Core i7-10510U CPU @2.3 GHz，RAM 24GB。根据设计的仿真场景，经过多次实验对比，确定神经网络结构及神经元配置，如表2所示。

表2　神经网络结构及神经元配置
Table 2　Neural network structure and neuron configuration

神经网络结构名称	神经元个数
输入层	19
隐藏层1	40
隐藏层2	30
输出层	1

神经网络输入为测控状态s以及动作a，通过深度神经网络计算，输出其相对应的状态–动作对的价值$Q(s, a)$。通过数据验算，采用表2进行参数配置的网络的期望Q值与预测神经网络计算产生的Q值之间的总体均方误差为14.265 5，其神经网络拟合效果示意图如图4所示，从图中可以看出，神经网络的拟合效果较好。

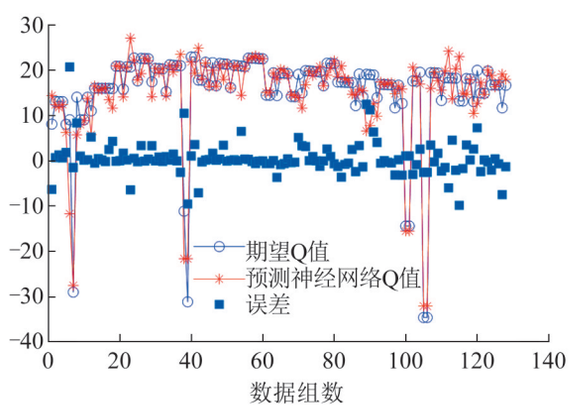

图4　神经网络拟合效果
Fig. 4　Schematic diagram of neural network fitting effect

在强化学习中，参数设置对算法的求解性能有较大的影响，经过多次实验探索，本文模型的部分参数值设置情况如表3所示。

表3 DQN模型部分参数设置
Table 3 Partial parameter settings of DQN model

参数名称	参数值
迭代次数	1 050
观察期迭代次数	50
经验池大小	8 000
随机梯度下降采样样本大小	400
目标Q网络更新频率	1 200
学习率	0.000 15
ε贪婪算法的初始探索值	1
ε贪婪算法的最终探索值	0.06
ε衰减因子	0.4

3.3 实验结果与分析

基于上述场景和实验环境，首先用STK计算出场景仿真时间内所有可用弧段信息，得到了679条测控可用弧段，并将其按照2.3.3中的算法流程自动输入到算法中，并接着按照求解程序进行实验仿真。本文选择最大化异构测控资源测控调度总收益为评价指标，算法求得的最优调度方案可以满足所有卫星提出的测控需求，即 $actSched=128$。DQN算法求解得到的测控调度总收益值与其总 Q 值之和随算法迭代次数增加的变化趋势分别如图5~6所示；为了更好地展示算法求解效果，将求解得到的所有方案的调度总收益按照算法迭代次数不断累加，并除以相应的迭代次数，得到的测控收益平均值的变化趋势如图7所示。

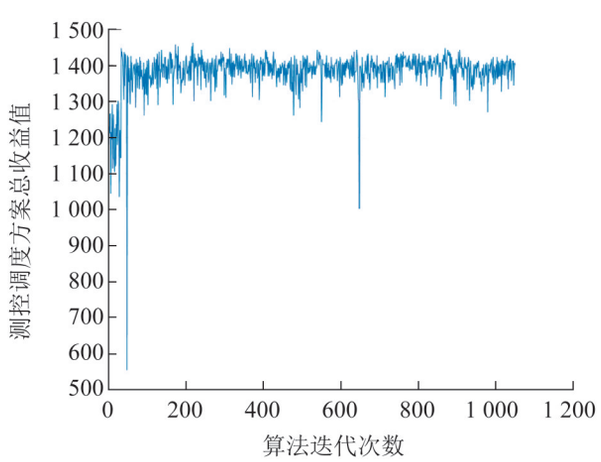

图5 测控调度总收益变化趋势
Fig. 5 Trend chart of total revenue of TT&C

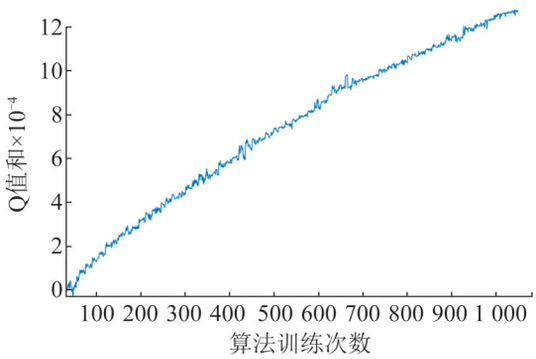

图6 Q值和变化趋势
Fig. 6 Change trend graph of sum of Q value

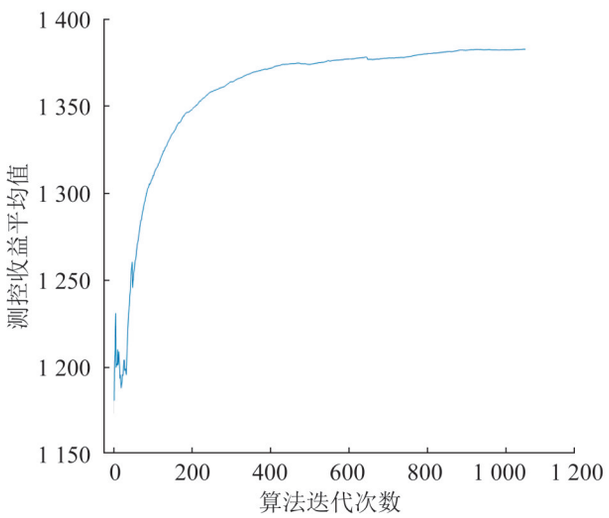

图7 测控调度收益平均值变化趋势
Fig. 7 Change trend graph of average value of TT&C income

DQN求解程序最终生成了满足全部测控调度需求的调度方案。从图5~6中可以看出，最开始随机生成的测控调度方案收益明显低于后续的总收益值与 Q 值之和，这是由于程序将最开始的50次迭代设置为观察期，按照 ε 贪婪算法动作选取策略，最开始每一次智能体的动作都是随机选择的，并将这些动作和环境反馈出的奖励和下一个状态储存入经验池。随着迭代次数的增加，记忆库中积累的经验信息越来越多，忽略掉个别代次的极值影响，生成调度方案调度收益值与 Q 值和都在稳步提高。通过图7可以看出，最开始随机生成的调度计划平均值较小且变化无规律，但是随着算法迭代次数的增加，ε 贪婪值调整变小，经验库

https://dc-china-simulation.researchcommons.org/journal/vol35/iss2/18
DOI: 10.16182/j.issn1004731x.joss.21-0879

越来越充实，测控收益平均值随着算法迭代次数的变化在不断升高，测控动作的选择越来越智能，也验证了本文提出的DQN算法的有效性。为了检验本文提出的DQN算法的求解效果，在测控场景相同的条件下，设置了基于遗传算法求解异构测控资源调度问题的对照组，将2种算法在迭代过程中求得的最高测控总收益值变化趋势进行了对比，求解结果如图8所示。

通过观察图8可知，DQN算法在218次迭代时得到的测控调度方案的收益值最高，为1 463，后面迭代的方案一直没有超过此收益值的情况出现；而遗传算法在178次迭代时，取得了最高的调度收益值1 408。从总体趋势可以看出，虽然最终2种算法的测控任务满足率都是100%，但是DQN算法在寻找最优调度方案方面要明显优于遗传算法。由于遗传算法寻优时存在“算法早熟”的现象，虽然遗传算法收敛速度较快，但是寻找到最优解的性能要弱于DQN算法。图9为DQN算法寻找到的最佳异构测控资源调度方案甘特图，每颗卫星的测控任务用不同颜色的色块表示，每段色块上的X/Y代号表示第X颗卫星在测控弧段集合中的代号$AID_x=Y$。

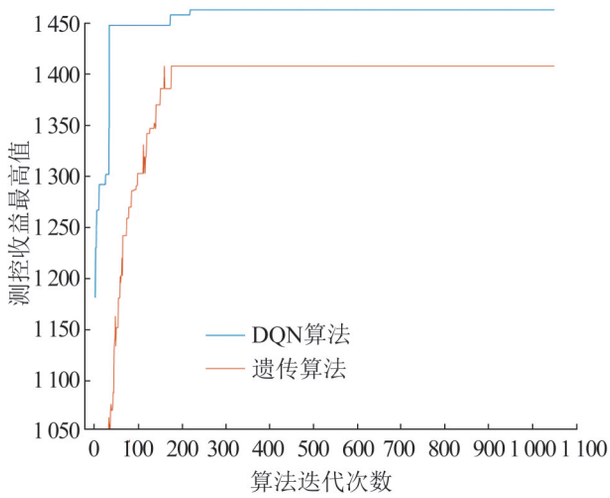

图8　DQN算法与遗传算法最优调度收益变化趋势
Fig. 8　Change trend of DQN algorithm and genetic algorithm optimal scheduling income

从图9中可以直观地看出，最终求得的最优测控调度方案可以满足模型中设置的异构测控资源调度所有约束条件。综上，通过仿真实验可以验证本文提出的基于DQN算法可以在一定程度上提高对异构测控资源联合调度问题的求解效果。

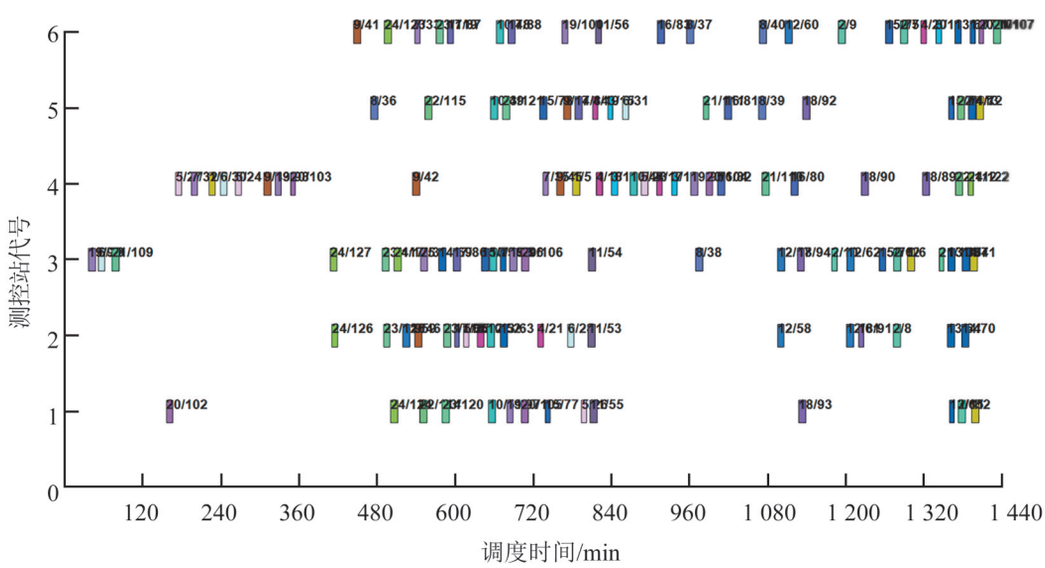

图9　最优调度方案的任务甘特图
Fig. 9　Task Gantt chart of optimal scheduling plan

4 结论

本文利用DQN算法，从深度强化学习的角度对异构测控资源联合调度问题进行了研究求解，验证了通过DQN算法求解异构测控资源联合调度问题的可行性，并通过仿真算例证明了算法对最优值的求解效果要优于以遗传算法为代表的传统求解算法。进一步的研究工作可以关注以下几个方面，在异构测控资源联合调度中，不同的测控资源管理部门对测控方案的要求会有不同的重点，需要进一步设计更能体现实用价值的测控资源调度收益评价函数；本文是利用DQN算法求解异构测控资源联合调度的初步尝试，下一步可以从测控状态优化设计和算法流程优化入手，提高算法的求解性能。

参考文献：

[1] 于志坚. 我国航天测控系统的现状与发展[J]. 中国工程科学, 2006, 8(10): 42-46.
Yu Zhijian. Status Quo and Development of Spaceflight TT&C Systems[J]. China Engineering Science, 2006, 8 (10): 42-46.

[2] 宋永生, 李铎, 陈劭睿. 商业航天测控管理[J]. 数字通信世界, 2019(7): 29.

[3] 郭夏锐. 商业卫星测控发展现状及趋势[J]. 国际太空, 2019(10): 44-48.

[4] 薛乃阳, 丁丹, 王红敏, 等. 基于改进遗传算法的多类测控资源调度方法[J]. 系统工程与电子技术, 2021, 43(9): 2535-2543.
Xue Naiyang, Ding Dan, Wang Hongmin, et al. Multi-type Measurement and Control Resource Scheduling Method Based on Improved Genetic Algorithm[J]. System Engineering and Electronic Technology, 2021, 43 (9): 2535-2543.

[5] Stottler R, Richards R. Managed Intelligent Deconfliction and Scheduling for Satellite Communication[C]//2018 IEEE Aerospace Conference. IEEE, 2018: 1-7.

[6] 张天骄, 李济生, 李晶, 等. 基于混合蚁群优化的天地一体化调度方法[J]. 系统工程与电子技术, 2016, 38(7): 1555-1562.
Zhang Tianjiao, Li Jisheng, Li Jing, et al. Space-ground Integrated Scheduling Based on the Hybrid ant Colony Optimization[J]. Systems Engineering and Electronics, 2016, 38(7): 1555-1562.

[7] 李长德, 徐伟, 徐梁, 等. 基于深度神经网络的多星测控调度方法[J]. 中国空间科学技术, 2022, 42(1): 65-72.
Li Changde, Xu Wei, Xu Liang, et al. Multi-satellite TT&C Scheduling Method Based on DNN[J]. Chinese Space Science and Technology, 2022, 42(1): 65-72.

[8] 朱凤龙. 遗传算法"早熟"现象的探究及改进策略[D]. 重庆: 西南大学, 2010.
Zhu Fenglong. Research on the Premature Convergence of Genetic Algorithm and Its Improved Politics[D]. Chongqing: Southwest University, 2010.

[9] 陈仲铭, 何明. 深度强化学习原理与实践[M]. 北京: 人民邮电出版社, 2019: 6-7.
Chen Zhongming, He Ming. Deep Reinforcement Learning: Principles and Practices[M]. Beijing: Posts & Telecom Press, 2019: 6-7.

[10] 薛乃阳, 丁丹, 王红敏, 等. 引入微元法思想的混合测控资源联合调度方法[J]. 系统仿真学报, 2022, 34(4): 826-835.
Xue Naiyang, Ding Dan, Wang Hongmin, et al. Idea of Infinitesimal Method-introduced Hybrid TT&C Resources Joint Scheduling[J]. Journal of System Simulation, 2022, 34(4): 826-835.

[11] 杨子璇. 基于深度Q网络的带返工汽车涂装作业重排方法[D]. 大连: 大连理工大学, 2020.
Yang Zixuan. A Resequencing Method for Automotive Painting Operations Considering Rework Based on Deep Q Network[D]. Dalian: Dalian University of Technology, 2020.

[12] 安元元, 李伟超, 王伟, 等. 一种低轨卫星星座测控地面站调度策略研究[J]. 时间频率学报, 2021, 44(2): 120-131.
An Yuanyuan, Li Weichao, Wang wei, et al. Research on Schedule Strategy of Ground Stations for LEO Satellites [J]. Journal of Time and Frequency, 2021, 44(2): 120-131.

DOI: 10.16182/j.issn1004731x.joss.21-0879