# 一种基于 Dueling DQN 改进的低轨卫星路由算法

许向阳，李京阳，彭文鑫
（河北科技大学信息科学与工程学院，河北 石家庄 050000）

**摘要：**卫星网络具有高动态性、节点处理能力不足，流量负载不均等问题。现有的地面路由算法并不能很好的解决卫星网络存在的问题。针对此问题，提出一种改进 Dueling DQN 的低轨卫星路由算法。首先，在路由算法中引入决斗网络的思想；然后在经验回放进行改进，将随机经验采样和优先经验采样进行融合，设置分层采样方法来进行采样；最后对网络进行参数的设置并且进行训练。从仿真和分析表明，从网络传输时延、系统吞吐量、丢包率方面有明显的提升。

**关键词：**卫星路由；分层经验回放；决斗网络
**中图分类号：**TN927.2　　**文献标识码：**A　　**文章编号：**2096-9759(2023)07-0056-04

## A Low-Orbit Satellite Routing Algorithm Based on Improved Dueling DQN

**XU Xiangyang, LI Jingyang, PENG Wenxin**
（School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050000, China）

**Abstract:** Satellite networks face challenges such as high dynamics, limited node processing capacity, and uneven traffic load. Existing ground routing algorithms cannot effectively solve these problems. To address this issue, an improved low-orbit satellite routing algorithm based on Dueling DQN is proposed. First, the concept of Dueling Networks is introduced into the routing algorithm. Then, the experience replay is improved by fusing random experience sampling and prioritized experience sampling, and a layered sampling method is set up for sampling. Finally, the network parameters are set and the model is trained. Simulation and analysis show significant improvements in network transmission delay, system throughput, and packet loss rate.

**Key words:** satellite routing, layered experience replay, dueling network

## 0 引言

近年来，随着无线通信技术的发展与技术的迭代，卫星通信以其覆盖范围广、抗毁性强、不受地理环境约束等优势在移动通信领域得到广泛的应用[1]。早期的卫星通信受限于技术，其主要功能为空中中继，且卫星数量较少，无法有效覆盖地球表面，随着运载火箭技术的不断发展，卫星成本不断降低，以多颗卫星组成的卫星星座网络应运而生，与早期相比，星座网络的优势在于对地球的覆盖性更高，系统抗毁性更强，当单颗卫星出现故障时，对卫星星座网络的数据转发与传输影响不大。相比于中、高轨卫星，低轨卫星具有传输时延更小、链路损耗较低、整体制造成本低等优势。使得构建低轨卫星星座网络成为研究重点。

与地面网络相比，低轨卫星网络存在诸多优势，但仍存在网络节点变化过快、节点处理能力不足[2]，数据包传输效率较低等问题。网络拓扑结构存在高动态和时变性[3]，使得地面网络的路由方法无法有效应用于低轨卫星网络。为了应对低轨卫星星座网络的特点，需要设计一种低轨卫星路由算法，它应该具备抗移动性、抗毁性和负载均衡能力[4]。低轨卫星路由算法的性能对低轨卫星网络的表现有着直接的影响。因此，设计一种高效、稳定的低轨卫星路由算法变得尤为重要。

当前低轨卫星路由算法的主要研究有：

文献[5]提出了 DT-DVTR(Discrete Time Dynamic Virtual Topology Routing)路由算法，使用虚拟拓扑的思想（Dynamic Virtual Topology）来处理卫星网络中复杂的动态变化。该算法将时间分割成多个片段，用最短路径算法计算每个静态拓扑网络任意两个卫星之间的路径集合，在提升网络传输数据性能的同时，会对卫星造成较大的星上存储开销；

文献[6]提出了 DRA(Datagram Routing Algorithm)算法。这种方法采用空间虚拟化的方式，设置虚拟节点以屏蔽卫星的运动。这种方法的优点是可以有效地屏蔽卫星的运动对路由转发的影响，并且提升了抗毁性；

文献[7]提出了介绍了一种名为 SLSR 算法，该算法提供了一种全局优化方案。该算法优势在于网络流量不会过度拥挤到最短路径，而是在多条轻负载链路中得到均衡转发。此外，SLSR 算法还采用了曼哈顿街区网络的思想，降低了算法复杂度和网络开销；

文献[8]考虑到卫星提供的网络资源与用户业务需求，提出了一种基于业务流分类的算法，该算法将根据预设的业务类型，对当前网络中的业务进行分类，根据业务的不同，选择相应的路由表进行路由转发。

## 1 决斗深度 Q 网络

DQN 算法使用单一的神经网络来估计每个状态下所有动作的 Q 值。这种网络结构称为全连接层。在训练过程中，DQN 使用经验回放和固定目标网络来提高学习效率和稳定性[9]。但单一神经网络对状态-动作值进行估计，会高估动作的 Q 值，从而造成某些状态下次优动作奖励值会优于最优动作，从而找不到最优策略，导致算法训练不稳定。针对此情况，本文引入 Dueling DQN 算法作为路由算法的基础。

Dueling DQN 算法通过将网络拆分为两个部分来进一步提高学习效率和稳定性[10-11]。这两个部分分别是状态价值函数

**收稿日期：**2023-03-09
**作者简介：**许向阳（1967-），男，河北石家庄人，硕士，河北科技大学，副教授，硕士导师，主要研究方向：无线自组网、卫星通信；李京阳（1997-），男，河北沧州人，研究生，硕士，主要研究方向：卫星路由；彭文鑫（1999-），男，河北唐山人，研究生，硕士，主要研究方向：卫星路由。

和优势函数。价值函数$V(s;\alpha,\theta)$用来估计每个状态的价值，仅与状态有关，与将要执行的动作无关，优势函数$A(s,a;\beta,\theta)$来估计每个动作的优劣程度，与状态动作都有关，最后的输出则将两者相加，如下所示：

$$Q(s,a;\alpha,\beta,\theta)=V(s;\alpha,\theta)+A(s,a;\beta,\theta) \quad (1)$$

其中$\boldsymbol{\alpha}$为价值函数支路的网络参数，$\boldsymbol{\beta}$为优势函数支路的网络参数，$\boldsymbol{\theta}$为卷积层公共部分的网络参数。但直接训练会存在将 V 值训练为固定值时，则算法变为 DQN 算法，则需要为神经网络增加一个约束条件，则 $Q$ 值函数计算如下所示：

$$Q(s,a;\alpha,\beta,\theta)=V(s;\alpha,\theta)+(A(s,a;\beta,\theta)-\frac{1}{|A|}\sum_{a'}A(\mathrm{s},a';\beta,\theta)) \quad (2)$$

## 2 改进的 Dueling DQN 算法

### 2.1 经验回放机制的改进

在原始的 DQN 算法中，每次使用完一个样本$(s_t,a_t,r_t,s_{t+1})$就丢弃，造成经验的的浪费，并且相关性强，不利于模型的学习。对此情况，Nature DQN 算法设置了经验池[12]，为了更新深度神经网络的参数，从经验池中随机选择一小批样本作为训练更新的样本。使用经验回放机制，每次采样时都从经验池中随机选择样本。这种随机选择样本的方式可以有效打破样本之间的相关性，从而更好地训练深度神经网络。

在 DRL 中，一般采用随机采样来随机均匀地抽取经验池中的经验。但随机采样是在经验池中等概率的随机选择样本，没有考虑到样本的重要性，可能会存在信息价值较高的样本在神经网络的训练中使用率比较低甚至没有被使用过的情况，会使得训练次数增加导致训练效率变低。为了解决这个问题，人们提出了一种名为优先经验回放的方法。在这种方法中，智能体会优先选择最有价值的一批样本来训练。为了防止过拟合，低价值的样本也会有一定的概率被选择。但采用优先经验回放机制后，采样时间会比较长，容易产生局部最优解。针对此情况，本文采取了两者的优势，提出一种随机分层抽取优势样本的采样方法，利用随机经验抽取的方法在经验池抽取随机均匀的抽取样本后，将样本分成多层，分层数取值为最大不超过样本数量的一半正整数，这种取值方式可以保证采样方式不会变为原先的随机经验采样。在每层中对比样本的采样概率，选择其中最大价值的样本来作为最后抽取的样本。这种方法既可以保留样本的多样性，又可以选取到相对的样本，加快收敛速度。采样图如下：

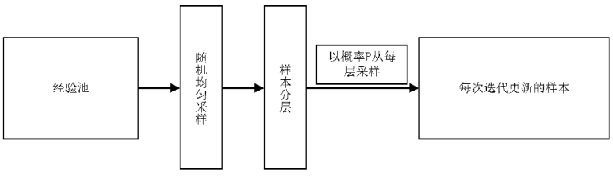


**图　改进采样流程图**

优先经验回放样本采样公式 $P(i)$ 为：

$$P(t)=\frac{e^{p_t}}{\sum_t e^{p_t}} \quad (3)$$

$$p_t=|\delta_t+\varepsilon| \quad (4)$$

其中，$\boldsymbol{\delta}_t$为时间差分误差（Temporal Difference Error，TD error），又称为时序误差，通常用于衡量代理在执行动作后，对于预期奖励的估计值与实际奖励之间的差异。时序误差一般是用来更新行为值函数，时序误差绝对值越大，说明其学习价值越高。优先经验回放采用时序误差，可以使得增加价值高的样本的重要性和减少错误行为发生的概率。而$\boldsymbol{\varepsilon}$ 的设置则是为了保证时序误差为零时采样概率不为零，时序误差计算公式如下：

$$\delta_t=R_t+\gamma*\max_{A_t}Q(S_t,A_t,\theta)-Q(S_{t-1},A_{t-1},\theta) \quad (5)$$

此采样机制的存在可以保证每个选择后的样本可以在迭代训练至少使用一次，从而避免网络的过拟合问题。

### 2.2 算法描述

Dueling DQN 模型的训练旨在学习最佳行为策略，以便在当前状态下选择具有最大价值的动作，即选择最优下一跳节点到目标节点。在模型训练期间，主要是调整 Dueling DQN 模型中的值函数 Q 网络参数和目标 Q 网络参数，以使网络参数朝着更优方向拟合。

值函数 Q 网络是 Dueling DQN 模型中的重要组件，代表着当前的学习能力，负责探索最新的路由环境并将每次探索环境过程中的经验$(S_t,A_t,R_t,S_{t+1})$存放在经验池 D 中。周期性地从经验池 $D$ 中随机选取一批样本，将状态 $S_t$ 和动作 $A_t$ 作为输入，输入到 $Q$ 网络中，分别计算出优势函数 $A(\mathbf{S_t},\mathbf{A_t},\boldsymbol{\theta})$ 和价值函数 $V(S^t,\boldsymbol{\theta})$，并通过组合优势函数 $A(\mathbf{S_t},\mathbf{A_t},\boldsymbol{\theta})$ 和状态价值函数 $V(S^t,\boldsymbol{\theta})$ 计算当前 $QA(\mathbf{S_t},\mathbf{A_t},\boldsymbol{\theta})$ 值。然后将下一个状态 $S_{t+1}$ 输入到值函数 $Q$ 网络中计算下一步的最优动作 $A_{t+1}$。这个过程将持续进行，直到网络收敛并产生最优的路由策略。

目标 Q 网络是已经学习到的经验的代表，负责给出基于历史经验的 Q 值评估。用来计算下一状态 $S_{t+1}$ 中选择所有可能动作 $A_{t+1}$ 的最大 Q 值，利用下一状态 $S_{t+1}$ 下一次动作 $A_{t+1}$，当前奖励 $R_t$ 以及折扣因子$\gamma$ 可计算 $Q\_target$。计算方法如下：

$$Q\_target=R_t+\gamma*\max_{A_t}Q'(S_{t+1},A_{t+1},\theta') \quad (6)$$

其中，$max_{A_t}$ 表示下一个状态 $S_{t+1}$ 中选择所有可能动作 $A_{t+1}$ 的最大 Q 值，$Q'(S_{t+1},A_{t+1},\theta')$表示目标 Q 网络对下一个状态 $S_{t+1}$ 中选择动作 $A_{t+1}$ 的预测值。

在训练过程中，通过运用均方误差（Mean Squared Error，MSE）的方法，使用目标 Q 值与值函数 Q 值的方差的期望和优势函数与价值函数之间的方差的期望计算 $Loss$ 函数，并利用 $Loss$ 函数的梯度下降法和 DNN 的反向传播过程完成值函数 Q 网络参数的更新，使得值函数 Q 网络能够更好地拟合目标 Q 值。

在算法中，需要对模型的状态$S$、动作$A$、奖励函数$R$、状态转移概率 $P$、折扣率$\gamma$ 进行设置。其中，状态 $S$ 设置为链路状态，动作 $A$ 设置为下一跳卫星节点，状态转移概率 $P$ 为选择下一跳节点的概率，奖励函数 $R$ 设置为当前路由到下一跳卫星节点的奖励，折扣率$\gamma$为未来期望奖励权重。

由于预先设置的星间链路的带宽相同，则 $t$ 时刻路由节点 $i$ 与路由节点 $j$ 之间的数据传输最大速率为：

$$C_{i,j}(t)=B\log_2(1+\frac{s}{n}) \quad (7)$$

其中，$B$ 为信道带宽，$s$ 为所传信号的平均功率，$n$ 为高斯噪声功率，由此可以得出链路传输时延：

$$T_{i,j}(t)=\frac{D_{i,j}(t)}{C_{i,j}(t)}+m \quad (8)$$

其中，$D_{ij}(t)$为 时刻路由节点 $i$ 与路由节点 $j$ 之间距离，$m$ 为固定发送时延。

奖励函数 R 由链路传输时延 $T_{ij}(t)$、源节点与目的节点距离 $D_{ij}(t)$构成，具体设置为：

$$R_t=\begin{cases}-\left\{\mu T_{i,j}(t)+\eta D_{i,j}(t)\right\}, 下一跳不为目的节点\\ 0, \quad 下一跳为目的节点\end{cases} \tag{9}$$

其中$\mu$，$\eta$为权重参数，令其$\mu$+$\eta$=1 ，将最大奖励设置为 0，当下一跳为目的节点时，奖励最大，其他情况奖励均为负值。在奖励函数设置时，必须考虑链路传输时延，时延越小，其奖励值越大，但只考虑时延特性，则可能使得数据传输偏离目的节点，则应该考虑节点的距离，节点距离 $D_{ij}(t)$ 可由升交点赤经的差值和真近地点角差值来表示。

在算法中，Q 值函数被拆分成价值函数和优势函数两部分，其中价值函数 $V(\mathbf{S_t},\boldsymbol{\theta})$是通过 Q 值函数网络得到每个动作 $A$ 的 Q 值，对所有动作的 Q 值求平均则可得到价值函数 $V(\mathbf{S_t},\boldsymbol{\theta})$，其公式如下：

$$V(S_t,\theta)=\frac{1}{|\mathrm{A}|}\sum_{A}Q(S_t,A_t,\theta) \tag{10}$$

其中，$|A|$表示为当前可选择动作个数。由此可以得出当前状态的优势函数：

$$A(S_t,A_t,\theta)=Q(S_t,A_t,\theta)-V(S_t,\theta) \tag{11}$$

在算法中，$Loss$ 函数使用均方误差指标进行运算，包括两部分：预测值与目标值之间的平方误差和优势函数与价值函数之间的平方误差，计算公式如下：

$$Loss=E[(Q\_target-Q\_value)^2+\lambda*(A(S_t,A_t,\theta)-V(S_t,\theta))^2] \tag{12}$$

基于 Dueling DQN 改进的算法完整流程如下：

| 算法 | 基于 Dueling DQN 算法 |
|---|---|
| 输入： | 网络拓扑 $G(V,E)$、状态空间 $S$、学习率 $\lambda$、动作空间 $A$、折扣率 $\gamma$、目标网络更新参数频率 $F$、回合 $M$、迭代次数 T |
| 初始化： | 经验池 $D$ |
| | Q 网络参数 $\theta$ |
| | 目标 Q 网络 $\theta'$ |
| for episode =1 to M do | |
| | 初始化环境,得到初始状态 $S_t$ |
| | for iteration t =1 to T do |
| | 采用 $\varepsilon$ 贪婪策略选择动作，随机选择一个动作 |
| | 以(1-$\varepsilon$)概率选择动作 $A_t=\arg\max_a Q(s_t,a;\theta)$ |
| | 执行动作 $A_t$，观察奖励 $R_t$ 和下一个状态 $S_{t+1}$ |
| | 将 $(S_t,A_t,R_t,S_{t+1})$ 存储到经验池 $D$ 中 |
| | 用改进的经验机制在经验池 $D$ 中进行采样均分 |
| | 对于每个 $(S_t,A_t,R_t,S_{t+1})$ 执行以下步骤 |
| | if $S_{t+1}$ 是终止状态，则 $Q\_target=R_t$ |
| | else 用目标 Q 网络计算 |
| | $Q\_target=R_t+\gamma(V(S_{t+1},\theta')+A(S_{t+1},A_{t+1},\theta')-\frac{1}{\lvert\mathrm{A}\rvert}\sum_{A_{t1}}A(S_{t+1},A_{t+1},\theta'))$ |
| | 计算当前状态 Q 值与优势函数，$Q\_value=Q(S_t,A_t,\theta)$， |
| | $A(S_t,A_t,\theta)=Q(S_t,A_t,\theta)-V(S_t,\theta)$ |
| | 计算 $Loss$ 函数 |
| | 使用梯度下降算法更新 Q 值网络：$\theta=\theta-\lambda*\nabla\theta Loss$ |
| | 每隔 F 步更新目标 Q 网络：$\theta'\leftarrow\theta$ |
| | 更新状态 $S=S_{t+1}$ |
| | end for |
| end for | |
| 输出：训练完成的 Dueling DQN 模型 | |

## 3 仿真

### 3.1 仿真参数设置

为了评估算法性能，利用NS3 仿真软件，在一个类似铱星卫星网络中构建仿真。其中，66 颗卫星分布在六个平面上。每颗卫星有两个层内链路和两个层间链路，层内链路一直连接，层间链路在反向缝区域断开。设置同层卫星链路带宽为 25Mbps，层间链路带宽为 1.5Mbps，队列缓冲大小为 50kb。数据包大小设置为 1kb，Hello 包发送周期仿真时间设置为 90s。仿真参数和算法训练参数如下：

**表 1　系统仿真参数设置**

| 参数 | 值 |
|---|---|
| LEO 卫星数量 | 66 |
| 轨道数 | 6 |
| 轨道高度 | 780km |
| 轨道倾角 | 86.4° |
| 星间链路带宽 | 25Mbps |
| 排队队列缓冲大小 | 50kb |

**表 2　算法仿真参数设置**

| 参数 | 值 |
|---|---|
| 回合 $M$ | 2000 |
| 迭代次数 $T$ | 50 |
| 折扣率γ | 0.9 |
| 探索率ε | 0.95 |
| 学习率 λ | 0.005 |
| 目标网络θ更新参数频率 | 200 步/次 |
| 经验池 容量 $D$ | 5000 |
| 路由最大跳数 $N$ | 15 |
| 奖励函数中时延权重$\mu$ | 0.6 |
| 奖励函数中跳数权重$\eta$ | 0.4 |

### 3.2 仿真结果分析

与本文算法进行对比的路由算法为SPF路由算法和ELB路由算法，将通过网络平均传输时延和丢包率以及系统吞吐量三个方面进行对比。

（1）网络平均传输时延

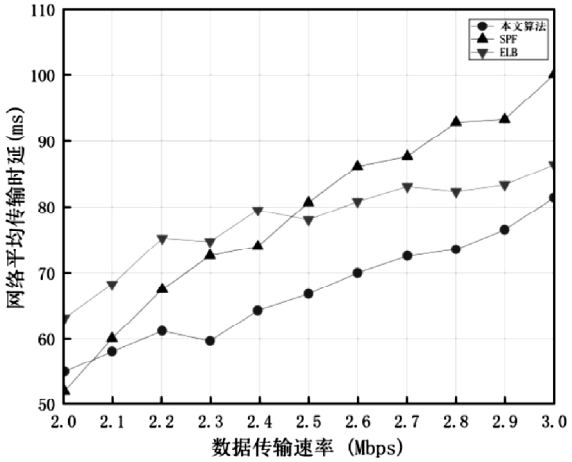


**图 2　网络平均传输时延对比**

随着数据传输速率不断增大,网内传输数据包数量增加,逐渐达到链路带宽上限,所以网络中数据包的传输平均时延上升。通过仿真结果可以看出,本文路由算法在网络传输时延上略优于 SPF 算法和 ELB 算法。由于采用了 Dueling DQN 模型进行路由计算,训练后的模型遵循最短路径的原则去选择路径,并在链路状态变化时,对链路进行切换从而减少节点拥塞问题。

(2)丢包率

仿真结果表明,本文算法在丢包率上当链路状态到达拥塞时,将会将数据转发到次优链路,有效地降低了丢包率。

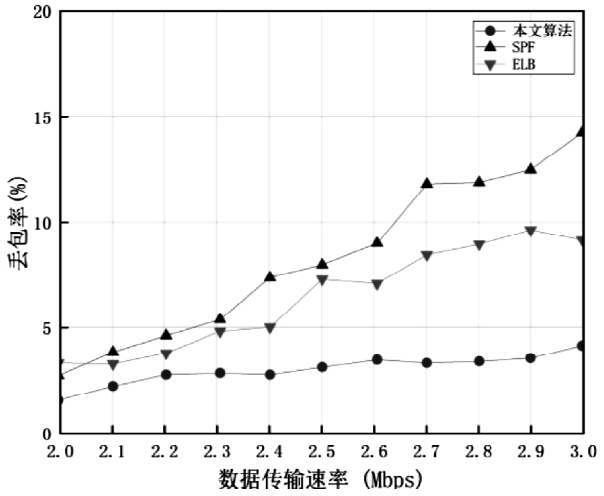


图 3 不同算法丢包率对比

(3)吞吐量

吞吐量是对卫星网络中路由算法在一定时间内的数据传输总容量的衡量标准。系统仿真结果如图所:

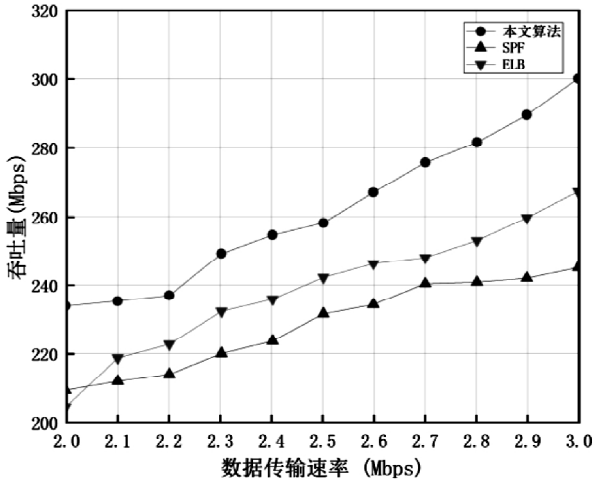


图 4 不同算法吞吐量对比

由仿真结果可知,在网络吞吐量性能方面比 SPF 算法和 ELB 算法要高,并且随着数据传输的速率的增大,系统吞吐量也在提升。由于奖励函数的设置,使得数据包会往距离最近,时延最短的方向进行转发,算法可以快速为卫星提供下一跳节点的选择,从而使得同样的时间内,卫星节点可以处理更多的任务,提升系统的吞吐量。

## 4 结语

本文针对低轨卫星网络存在的高动态性的问题,提出了一种基于深度强化学习的卫星路由算法,并通过设置相应的参数以及改进的采样机制来实现更好的效果。仿真结果表明,基于分层采样机制的深度强化学习算法有效的降低了数据发送时延,降低了丢包率,提高了系统吞吐量。

**参考文献:**

[1] Gounder V V, Prakash R, Abu-Amara H. Routing in LEO-based satellite networks[C]//1999 IEEE Emerging Technologies Symposium. Wireless Communications and Systems (IEEE Cat. No. 99EX297). IEEE, 1999: 22.1-22.6.

[2] 倪少杰,岳洋,左勇.卫星网络路由技术现状及展望[J].电子与信息学报,2022,44:1-13.

[3] Yan Y, Han G, Xu H.A survey on secure routing protocols for satellite network[J].Journal of Network and Computer Applications, 2019, 145:102415.

[4] 郑爽,张兴,王文博.低轨卫星通信网络路由技术综述[J].天地一体化信息网络,2022,3(03):97-105..

[5] Werner M. A dynamic routing concept for ATM-based satellite personal communication networks[J]. IEEE journal on selected areas in communications, 1997,15(8):1636-1648.

[6] Ekici E,Akyildiz I F,Bender M D. Datagram routing algorithm for LEO satellite networks[C]//Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies(Cat. No.00CH37064). IEEE, 2000,2:500-508.

[7] Yan H, Zhang Q, Sun Y. A novel routing scheme for LEO satellite networks based on link state routing[C]//2014 IEEE 17th International Conference on Computational Science and Engineering. IEEE, 2014:876-880.

[8] Mohorcic M, Svigelj A, Kandus G. Traffic class dependent routing in ISL networks[J].IEEE transactions on aerospace and electronic systems, 2004,40(4): 1160-1172.

[9] Li X, Liu H, Wang X. Solve the inverted pendulum problem base on DQN algorithm[C]//2019 Chinese Control And Decision Conference (CCDC). IEEE, 2019:5115-5120.

[10] Wang Z,Freitas N D,Lanctot M.Dueling Network Architectures for Deep Reinforcement Learning[J].JMLR.org, 2015.

[11] Ban T W . An Autonomous Transmission Scheme Using Dueling DQN for D2D Communication Networks [J]. IEEE Transactions on Vehicular Technology, 2020, PP(99):1-1.

[12][1] Guan Y, Liu B, Zhou J,et al. A New Subsampling Deep Q Network Method [C]//2020 International Conference on Computer Network, Electronic and Automation (ICCNEA). 2020.