引用本文：侯临风，何荣希，吴梓敬. 弹性光网络中基于 DRL 的 RMSA 算法[J]. 光通信技术，2024，48(3)：57-63.

弹性光网络中基于 DRL 的 RMSA 算法

侯临风，何荣希 *，吴梓敬
（大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要：为了更好地解决弹性光网络(EON)的路由、调制格式与频谱分配(RMSA)问题，进一步降低网络阻塞率，提出一种基于深度强化学习(DRL)的 RMSA 算法。该算法在奖励设计中将考虑影响 RMSA 决策的资源占用度和频谱邻接度这 2 个指标，以鼓励智能体优先选择资源占用度低、频谱邻接度高的路径来建立光路，并对比该算法与其它算法在不同网络中的性能。仿真结果表明：与几种典型的 DRL 算法相比，所提算法的网络阻塞率更低。

关键词：弹性光网络；路由、调制格式与频谱分配；网络阻塞率；深度强化学习；奖励设计
中图分类号：TN929.11　**文献标志码**：A　**文章编号**：1002-5561(2024)03-0057-07
DOI：10.13921/j.cnki.issn1002-5561.2024.03.010　**开放科学(资源服务)标识码(OSID)**：

RMSA algorithm based on DRL in elastic optical networks

HOU Linfeng, HE Rongxi*, WU Zijing
(College of Information Science and Technology, Dalian Maritime University, Dalian Liaoning 116026, China)

Abstract: In order to better solve the routing, modulation format and spectrum allocation (RMSA) problems of elastic optical networks (EON), and further reduce the network blocking rate, an RMSA algorithm based on deep reinforcement learning (DRL) is proposed. This algorithm will consider two indicators, resource occupancy and spectral adjacency, which affect RMSA decision making in reward design, to encourage agents to prioritize selecting paths with low resource occupancy and high spectral adjacency to establish optical paths, and compare the performance of this algorithm with other algorithms in different networks. The simulation results show that compared with several typical DRL algorithms, the proposed algorithm has a lower network blocking rate.

Key words: elastic optical network, routing, modulation format and spectrum allocation, network blocking rate, deep reinforcement learning, reward design

0 引言

传统波分复用(WDM)光网络采用固定栅格，这限制了其在未来高速率、大容量、可扩展的光传送网中的应用。为此，业界提出了基于光正交频分复用技术的弹性光网络(EON)[1-2]作为解决方案。然而，EON 在实际应用中面临一个关键问题，即路由、调制格式与频谱分配(RMSA)，这在一定程度上限制了 EON 的进一步优化和应用[3]。文献中常用整数线性规划(ILP)求其最优解，但由于 RMSA 是一个 NP-hard 问题，当网络复杂度较高时，基于 ILP 的算法可能无法在合理的时间内找到最优解。因此，一些启发式算法[2-4]被提出，用于解决 RMSA 问题中路由选择和频谱分配这 2 个子问题。另外，也有文献考虑了 EON 的频谱碎片问题。文献[5]首先提出二维资源模型，用以描述链路和路径上时频资源的可用性，并提出最小资源消耗选路策略和二维碎片感知频谱分配策略，这 2 种策略有利于减少频谱碎片并降低带宽阻塞率。文献[6]提出一种最大化业务承载力的碎片感知算法，在频谱分配时根据空闲频谱块大小及相邻业务剩余持续时间等因素评估其业务承载力，选取业务承载力最大的方案建立连

收稿日期：2024-01-06。
基金项目：国家自然科学基金项目(61371091、61801074、62371085)资助；大连市科技创新基金项目(2019J11CY015)资助。

作者简介：侯临风(2000—)，男，硕士研究生，现就读于大连海事大学信息科学技术学院电子信息专业，主要从事光通信网络研究，参与了国家自然科学基金项目，曾连续 2 个学年荣获大连海事大学硕士学业奖学金。
*** 通信作者**：何荣希(1971—)，男，博士，教授，主要研究方向为光网络和无线网络技术。

接，该算法能够降低网络阻塞率并提高频谱利用率。文献[7]针对抗毁 EON 中，在业务保护路径建立过程中可能出现的保护碎片问题，提出了一种基于自适应调制的碎片感知共享通路保护算法，该算法能够有效减少空闲碎片和保护碎片的数量。然而，上述策略都是基于人工提取特征的简单经验方法，缺乏对整体 EON 状态的全面感知，无法在具有时变特性的 EON 中实现真正的自适应服务。为此，文献[8]对异步优势表演者-评论家(A3C)算法进行改进，提出了基于经验集的 DeepRMSA 架构，并进一步提出基于窗口的灵活训练机制，取得了更显著的效果。文献[9]通过引入基于多智能体的 NC&M 体系结构和应用多智能体深度强化学习(MADRL)，将自治网络框架扩展到多域 EON。文献[10]提出一种基于深度强化学习(DRL)的非线性 EON RMSA 方案，使用高斯噪声模型计算非线性效应对传输质量的影响，显著降低了网络阻塞率。文献[11]使用图卷积神经网络和循环神经网络进行状态特征的提取，让智能体感知与 RMSA 相关的关键信息，以做出更好的决策。但是，上述方案在奖励设计上，仅考虑了业务请求是否服务成功，这种单一的奖励机制可能导致智能体在探索时表现出过大的随机性。为此，文献[12]提出启发式的奖励设计方案，在奖励设计中分别结合了频谱切割度和频谱碎片大小这 2 个因素，让智能体在训练过程中能够明确地识别并减少频谱碎片化影响，从而进一步降低网络阻塞率。但由于该方案仅考虑对候选频谱相邻资源的碎片化影响，其算法在奖励空间的设计上仍存在一定局限性。

奖励设计对于智能体的训练至关重要，而现有的大多数基于 DRL 的 RMSA 算法在奖励空间上的设计较为简单，限制了算法的性能。为了克服现有局限性并有效降低网络阻塞率，本文引入 2 个关键指标：资源占用情况和频谱邻接情况，并将这些指标融入奖励设计中，进而提出一种改进奖励设计 (IRD)DRL 的 RMSA 算法(下文简称 DRL-IRD 算法)。

1 网络模型

将 EON 建模为有向图 $G(V,E,F)$。其中，V 表示节点集，E 表示光纤链路集，F 表示每条光纤链路的频谱槽(频隙)集，则节点数、链路数和每条链路的频隙数分别为 $|V|$、$|E|$ 和 $|F|$。业务请求随机到达网络，表示为 $R_t(s,d,b,\tau)$。其中，s、d 分别表示该业务请求的源节点和目的节点；b 表示速率，单位为 Gb/s；τ 表示业务持续时间，t 表示第 t 个到来的请求。

当业务请求到达时，网络需要进行 RMSA 决策，具体过程为：

首先，在源、目的节点间用 K 最短路由(KSP)算法计算出 K 条最短路径。然后，根据任意一条路径 k 的距离确定调制等级 $M_k \in \{1,2,3,4\}$，该 4 个等级分别对应二进制相移键控(BPSK)、正交相移键控(QPSK)、8 阶正交振幅调制(8QAM)和 16QAM 调制格式，对应的传输距离分别为 5 000、2 500、1 250、650 km[8]。其次，根据路径 k 的调制等级 M_k 和业务请求 R_t 需求的速率 b，可以确定在该路径的每条链路上需要分配的频谱槽数 N_{need}^{k}[13]，即

$$N_{\text{need}}^{k}=\left\lceil \frac{b}{M_k C_{\text{BPSK}}} \right\rceil + N_{\text{P}} \tag{1}$$

其中，N_{P} 表示保护带宽 (设为 1 个频谱槽)，$\lceil\cdot\rceil$表示向上取整，C_{BPSK} 表示在 BPSK 调制格式下链路上每个频谱槽所能支持的速率。

最后，在所选路径的每条链路上选择满足频谱邻接性、频谱连续性和频谱不重叠性约束[5]的频谱资源进行分配。如果分配成功，则利用所选路径和频谱槽建立光路(业务连接)，连接持续时间为 τ。如果不能找到满足以上约束的分配方案，则阻塞该请求。制定 RMSA 策略需要尽可能降低网络阻塞率，以成功服务更多的业务请求。

2 DRL-IRD 算法

DRL-IRD 算法的框架与基于 DRL 的 RMSA 算法[8]类似，如图 1 所示。当业务请求到达时，DRL 智能体的深度神经网络(DNN)以状态数据为输入，根据策略(可用服务方案的概率分布)输出动作，然后确定 RMSA 方案对请求进行服务。接着，根据服务提供的结果，反馈相应的奖励，并将当前请求服务过程中产生的状态、动作和奖励存储在经验集中。每当存储了足够数量的样本时，就发出训练信号更新 DNN 的参数。

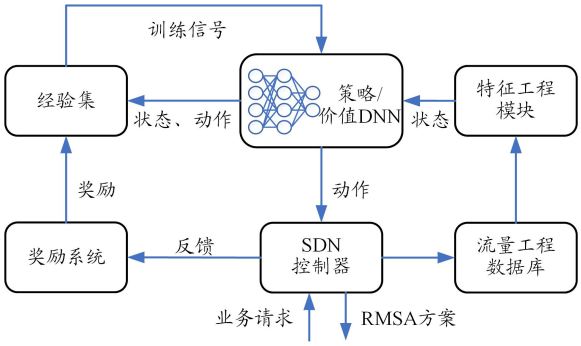


图 1 基于 DRL 的 RMSA 算法框架

于是，通过调整 DNN，DRL 智能体能够逐渐学习到有效的策略，最终得到一个具有自我学习和自适应能力的自主 EON 系统。通过这样重复的服务训练，DRL 智能体能够获得更高的长期累积奖励。

2.1 状态空间

状态空间是一个 $1\times[2|V|+1+5K]$ 的矩阵，包含了业务请求 R_t 及其 K 条备选路径上的信息，定义为

$$s_t=\left\{s,d,\tau,\left[I_{\text{first}}^k,N_{\text{first}}^k,N_{\text{need}}^k,B_k,N_{\text{average}}^k\right]\Big|_{k\in[1,K]}\right\} \quad (2)$$

其中，I_{first}^k、N_{first}^k 分别表示第 k 条备选路径上首个可用频谱块的起始索引及其大小，N_{need}^k 表示第 k 条备选路径上该请求所需频谱槽数，B_k 和 N_{average}^k 分别表示备选路径 k 上可用频谱块的个数及其平均大小。

状态空间矩阵中，通过 $|V|$ 个元素表示该请求的源节点 s，s 用 1 表示，其余节点用 0 表示。同理，通过 $|V|$ 个元素表示请求的目的节点 d，d 用 1 表示，其余节点用 0 表示。本文用 1 个元素表示该请求的持续时间 τ。

2.2 动作空间

DRL-IRD 算法通过感知网络拓扑状态，从而与环境进行交互，以自动配置 RMSA 方案。在动作空间中，通过 KSP 算法选取 K 条最短路径作为备选路径，智能体将从 K 条备选路径中为每一个业务请求 R_t 选择一条路径 k，分别确定其调制格式后，采取首次命中(FF)策略选择首个可用频谱块进行分配。因此，动作空间是一个 $1\times K$ 的矩阵。

2.3 奖励空间

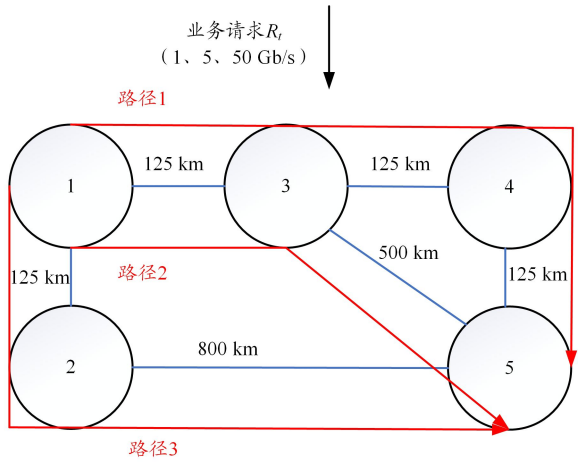
对于 DRL 来说，奖励的设置至关重要，它能够指导智能体选择更好的动作以最大化预期累积奖励。本文引入资源占用和频谱邻接指标来评估 RMSA 决策的好坏，并将这些指标结合到奖励设计中，以更好地指导智能体训练。

1)资源占用度 O_k：用以反映使用路径 k 建立连接时的资源占用情况，定义为

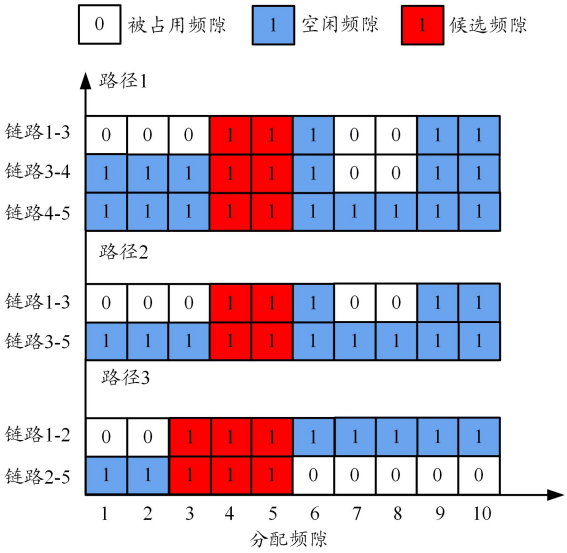
$$O_k=\frac{\ln\left(1+\frac{M_k}{H_k}\right)}{\ln\left(1+\frac{M_{\max}}{H_{\min}}\right)} \quad (3)$$

式(3)中，M_k 的取值越大，表示所选调制等级越高，在该路径建立连接时，每条链路上所需频谱槽就越少，越有利于减少频谱资源使用量。H_k 为所选路径 k 的跳数，H_k 越小，说明该请求将在更少的链路上占用频谱槽，这有利于节约资源。因此，将 M_k 与 H_k 相除后的值越大，表明选择路径 k 将占用更少的资源。然后，通过取对数和除以最大值减小指标波动性并归一化；并且为了归一化，再除以可能取到的最大值。$M_{\max}$ 为最高调制等级，其值为 4；$H_{\min}$ 为路径的最少跳数，其值为 1。资源占用度 O_k 越大，表明在所选路径建立光路所需的资源越少，故应优先考虑选择 O_k 值更大的路径建立光路。

网络拓扑及频谱使用情况如图 2 所示。图 2(a)为一个源、目的节点为 1 和 5，速率为 50 Gb/s 的业务请求 R_t 到达网络，为其找到 3 条距离最短的路径；图 2(b)为各路径上链路的频谱使用情况。其中，路径1(1–3–4–5)的跳数 H_1=3，路径距离为 375 km，按照调制等级、



(a) 网络拓扑



(b) 路径及其频谱使用情况

图 2 网络拓扑及频谱使用情况

调制格式与传输距离的对应关系[8]为其选择 16QAM 格式，即 M_1=4，根据式(3)可以计算出其资源占用度 O_1=0.526 5。同理，可计算出路径 2(1–3–5)和路径 3(1–2–5) 的资源占用度 O_2=0.682 6、O_3=0.569 3。由于 O_2 最大，使用路径 2 建立光路占用的频谱槽数少于其它 2条路径的资源占用度，因此应鼓励选择路径 2 建立光路。

2)频谱邻接度 C_k：用以反映在路径 k 上分配频谱资源后剩余的空闲频谱槽的邻接程度，定义为

$$C_k=\frac{\ln\left(\frac{1}{H_k}\sum_{l\in P_k}\frac{N_l}{B_l}\right)}{\ln(|F|)} \tag{4}$$

其中，l 为构成路径 k(用 P_k 表示)的任意链路，N_l 和 B_l 分别表示在路径 k 建立业务连接后链路 l 上的空闲频谱槽数和可用频谱块的个数，则 N_l/B_l 表示链路 l 上可用频谱块的平均大小。计算所选路径 k 的每条链路上可用频谱块平均大小之和，并除以路径的跳数 H_k，可以得到路径 k 的所有链路上可用频谱块大小的平均值，再取对数以减小该值的波动。频谱邻接度 C_k 越大，表明在路径 k 建立连接后各链路上空闲频谱资源的邻接程度越大，更容易为后续业务请求找到满足频谱邻接性约束条件的可用频谱资源，因而应鼓励选择 C_k 值更大的路径建立连接。

图 2(b)中的 3 条路径按照 FF 策略，都能成功为该请求找到可用频谱资源：为路径 1 分配频隙 4–5，并通过式(4)可以算出 C_1=0.397 9；同理，为路径 2 分配频隙 4–5 后，可得 C_2=0.439 3；为路径 3 分配频隙 3–5 后，可得 C_3=0.544 1。由于 C_3 最大，故从提高频谱邻接度的角度考虑，应鼓励选择路径 3 建立连接。

综合上述 2 个指标，奖励 r_t 可定义为

$$r_t=\begin{cases}\frac{1}{2}(O_k+C_k)，为\ R_t\ 成功建立光路\\ -1，R_t\ 被阻塞\end{cases} \tag{5}$$

当业务请求 R_t 找到频谱块被阻塞时，反馈奖励 r_t=–1；当业务请求 R_t 服务成功时，依据式(3)和式(4)计算资源占用度和频谱邻接度，并计算两者的算术平均值作为 r_t 反馈给智能体，从而鼓励智能体选择资源占用度和频谱邻接度大的路径建立连接，有利于降低后续请求被阻塞的概率。

2.4 训练模型

本文基于 A3C 算法[14]设计训练过程，该算法在环境的副本中训练多个局部 A3C 对，然后将训练参数更新为全局的 A3C 对，可以有效利用资源，提高训练效率。当智能体服务于业务请求时，基于策略的表演者在动作空间中选择合适的动作，通过在网络环境中执行该动作获得奖励。基于价值的评论家为这个动作打分，表演者根据评论家的评价修改选择动作的概率。智能体的目标是学习最优 RMSA 策略 $\pi(a_t|s_t)$，使累积的长期折扣奖励最大化。为此，在采样期间，将智能体与环境交互产生的样本转换存储在经验集 Λ 中。当经验集包含 2N–1 组样本时，则选取前 N 组样本计算累积奖励 G_t，G_t 的表达式为

$$G_t=\sum_{i=0}^{N-1}\gamma^i r_{t+i} \tag{6}$$

其中，γ 为折扣因子，$\gamma\in[0,1]$。

在 A3C 中，有一个策略网络 $\pi(a_t|s_t;\theta_\pi)$和一个价值网络 $v(s_t;\theta_v)$。θ_π 和 θ_v 分别表示每个局部 A3C 对的策略网络参数和价值网络参数。按照式(6)计算长期累积奖励 G_t 后，通过以下梯度更新全局参数 θ_π^* 和 θ_v^*。

$$\partial\theta_\pi^*\leftarrow\partial\theta_\pi^*+\nabla_{\theta_\pi}\ln\pi(a_t|s_t;\theta_\pi)\times[G_t-v(s_t;\theta_v)]+\beta\nabla_{\theta_\pi}H[\pi(a_t|s_t;\theta_\pi)] \tag{7}$$

$$\partial\theta_v^*\leftarrow\partial\theta_v^*+\partial[G_t-v(s_t;\theta_v)]^2/\partial\theta_v \tag{8}$$

式(7)中，$[G_t-v(s_t;\theta_v)]$反映实际动作优于平均水平的程度。同时，为了增加动作选择的多样性，引入策略熵项 $\beta\nabla_{\theta_\pi}H[\pi(a_t|s_t;\theta_\pi)]$提高智能体探索环境的能力，$\beta$ 用于控制熵正则化项的强度[12]。

2.5 算法流程

DRL–IRD 算法流程如算法 1 所示。

算法 1：DRL–IRD 算法流程

输入：请求及其备选路径的信息

输出：RMSA 方案

1)初始化 $\Lambda=\phi$ 和 ε=1；

2)for 每个请求 R_t do

3) If $\Lambda==\phi$ then

4) 更新参数 $\theta_\pi=\theta_\pi^*$ 和 $\theta_v=\theta_v^*$；

5) end

6) 释放过期请求占用的频谱；

7) 计算状态 s_t；

8) 计算策略网络 $\pi(a_t|s_t;\theta_\pi)$和价值网络 $v(s_t;\theta_v)$；

9) 以 ε 的概率根据 $\pi(a_t|s_t;\theta_\pi)$选择动作 a_t，否则令 $a_t=\text{argmax}_{a_t}\pi(a_t|s_t;\theta_\pi)$；

10) 计算奖励 r_t；

11）　存储$[s_t, a_t, v(s_t;\theta_v), r_t]$至 Λ；
12）　if $|\Lambda| == 2N-1$ then
13）　　for $t \in \{1, N\}$ do
14）　　　计算累积奖励 G_t；
15）　　end
16）　　更新 θ_π^* 和 θ_v^*；
17）　　设置 $\Lambda=\phi$, $\varepsilon=\max\{\varepsilon-\varepsilon_0, \varepsilon_{\min}\}$；
18）　end
19）end

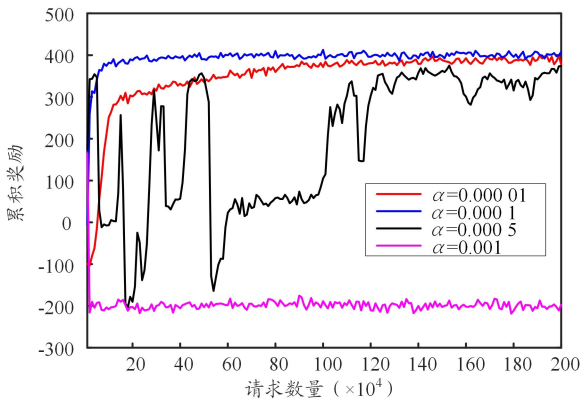
3 仿真及数据分析

3.1 仿真环境

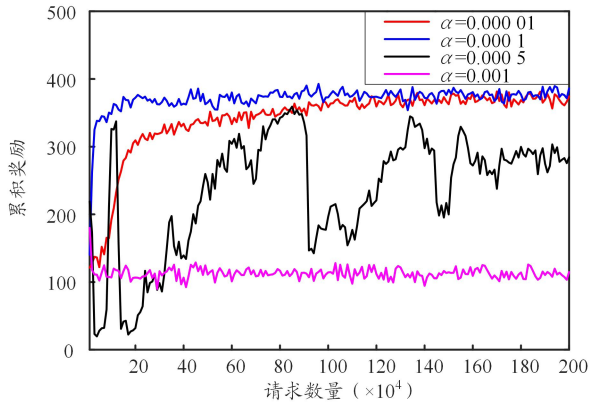
本文在 14 节点的 NSFNET 网络和 24 节点的 USNET 网络[5]中进行仿真实验。假设每条光纤链路包含 100 个频谱槽，每个频谱槽大小为 12.5 GHz。业务请求的源、目的节点在网络中随机选择，所需速率 b 均匀地分布在 25~100 Gb/s。业务请求的到达服从参数为 λ 的泊松分布，业务持续时间服从均值为 $1/\mu$（仿真时设置 μ=0.4）的负指数分布，因此网络负载为 λ/μ Erlang。仿真中，设置 K=5，N=200，β=0.01，探索率 $\varepsilon_0=10^{-5}$，最低探索率 $\varepsilon_{\min}$=0.05。策略网络和价值网络均有 5 个隐藏层，每层有 128 个神经元，并使用指数线性单位（ELU）作为激活函数。

3.2 仿真结果分析

首先，在 NSFNET 和 USNET 网络中分别设置网络负载为 200、300 Erlang，令折扣因子 γ=0.95，然后比较本文算法在不同学习率 α 下的学习曲线，如图 3 所示。可以看出，当 α 从 0.000 01 提高到0.000 1 时，累积奖励曲线上升更快，并且经过 80×10^4 个请求的训练后，在 2 个网络中均能收敛。当 α=0.000 5 时，学习曲线出现很大震荡，难以收敛；当 α=0.001 时，学习曲线迅速收敛到一个较低的值，甚至可能为负，这表明智能体难以执行那些能带来高奖励回报的动作。因此，在后续仿真实验中将学习率 α 设置为 0.000 1。



（a）NSFNET

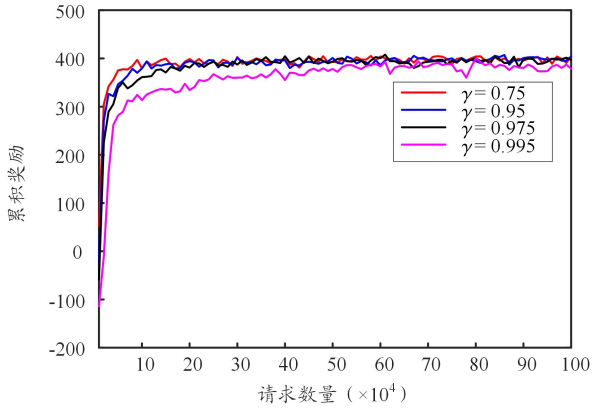


（b）USNET

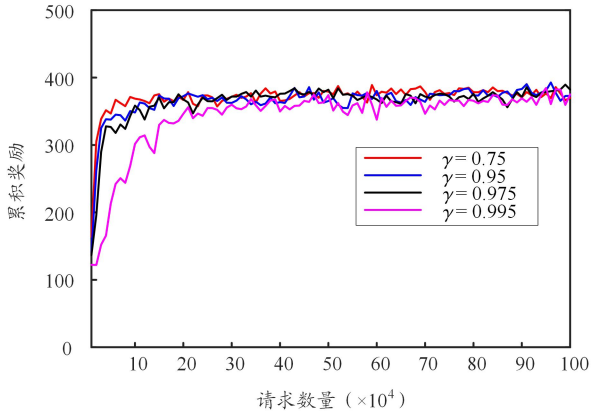
图 3　DRL-IRD 算法在2 种网络环境下，不同学习率 α 的累积奖励

同样，在 NSFNET 和 USNET 网络中分别设置网络负载为 200、300 Erlang，比较本文算法在不同折扣因子 γ 下的学习曲线，如图 4 所示。可以看出，本文算法的学习曲线在 γ 取 0.995 时收敛速度相对较慢，累积奖励也相对较低，而当 γ 取值在 0.975 以下时，曲线差异较小，不过仍呈现 γ 越小，上升和收敛速度越快的趋势。考虑到 γ 取 0.75 时，算法表现出令人满意的上升速度和收敛性能，因此，在后续的仿真实验中将折扣因子 γ 设置为 0.75。

为了检验算法的有效性，对本文算法与 KSP-FF 算法、DeepRMSA 算法[8]、DRL-Cut 算法和 DRL-Fragsize 算法[12]在不同网络中的性能进行比较，仿真指标为网络阻塞率，定义为被阻塞的业务请求数与总的业务请求数的比值。网络阻塞率越小，说明网络能成功服务



（a）NSFNET

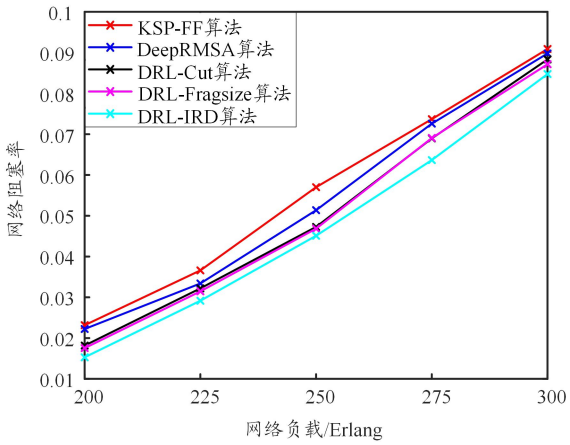


（b）USNET

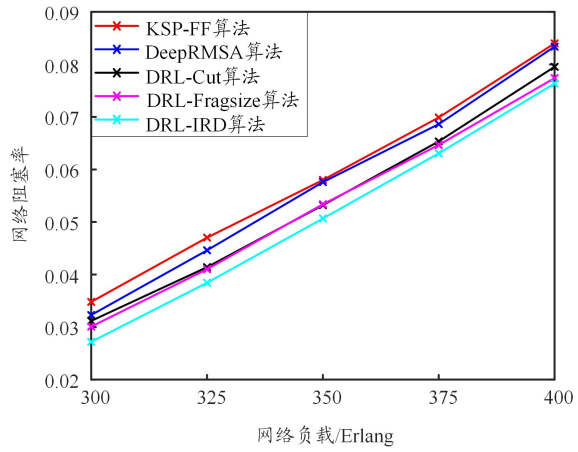
图 4　DRL-IRD 算法在2 种网络环境下，不同折扣因子 γ 的累积奖励

更多业务请求，算法的性能越好。

每次仿真产生 200×10⁴ 个请求。其中，前 100×10⁴ 个请求确保 DRL 算法模型达到稳定状态，而后 100×10⁴ 个请求则用于统计网络阻塞率，仿真结果如图 5 所示。可以看出，随着网络负载增加，5 种算法的网络阻塞率均呈现出上升趋势。这是因为随着网络负载的增加，网络中更多的频谱资源被占用，这使得后续的业务请求更难找到空闲的频谱资源来建立连接。此外，KSP-FF 算法在不同的网络负载下的网络阻塞率最高，DRL-IRD 算法最低，而 DeepRMSA 算法、DRL-Cut 算法、DRL-Fragsize 算法的网络阻塞率居于前两者之间。其中，DRL-Cut 算法和 DRL-Fragsize 算法的网络阻塞率接近，并且都优于 DeepRMSA 算法。在 NSFNET 网络中，当网络负载为 200 Erlang 时，DRL-IRD 算法比基于简单奖励设计的 DeepRMSA 算法的网络阻塞率降低了 31.23%，与性能最好的 DRL-Fragsize 算法相比，也降低了 12.88%；在 USNET 网络中，当网络负载为 300 Erlang 时，DRL-IRD 算法比 DeepRMSA 算法和 DRL-Fragsize 算法的网络阻塞率分别降低了15.81%和9.62%。主要原因是 DeepRMSA 算法在 RMSA 中应用了 DRL，能够对 EON 的状态进行较为全面的感知，因此经过大量业务请求的训练后，其网络阻塞率低于基于规则的传统 KSP-FF 算法。由于 DRL-Cut 算法和 DRL-Fragsize 算法在奖励设计中考虑了频谱碎片化影响，使智能体在训练过程中能捕获这些关键信息，因此相比基于简单奖励设计的 DeepRMSA 算法，这 2 种算法也能够降低网络阻塞率。而本文提出的 DRL-IRD 算法能够进一步捕获 RMSA 决策在资源占用方面的影响，并能感知与分配频谱不相邻的空闲资源的碎片情况，因此在训练达到稳定后的网络阻塞率更低。相比之下，DRL-IRD 算法均表现出了最佳的性能。



（a）NSFNET



（b）USNET

图 5　5 种算法在 2 种网络环境和不同网络负载下的网络阻塞率

4 结束语

本文提出了 DRL-IRD 算法，该算法将资源占用度和频谱邻接度 2 个指标结合到 DRL 的奖励设计中。鼓励智能体选择资源占用程度更小，且在频谱分配后空闲频谱资源邻接程度更大的路径建立连接，可以提高成功服务后续业务请求的概率，从而降低网络阻塞率。仿真结果表明：对学习率和折扣因子的调整可以影响 DRL-IRD 算法的学习效率，与 KSP-FF 算法和几种典型 DRL 算法相比，本文所提算法具有更低的网络阻塞率。

参考文献：

[1] MANDLO A. Routing and dynamic core allocation with fragmentation optimization in EON-SDM[J]. Optical Fiber Technology, 2024, 83: 103658-1–103658-13.

[2] CHATTERJEE B C, BA S, OKI E. Fragmentation problems and management approaches in elastic optical networks: a survey[J]. IEEE Communications Surveys & Tutorials, 2017, 20(1): 183–210.

[3] ABKENAR F S, RAHBAR A G. Study and analysis of routing and spectrum allocation (RSA) and routing, modulation and spectrum allocation (RMSA) algorithms in elastic optical networks (EONs)[J]. Optical Switching and Networking, 2017, 23: 5–39.

[4] RUIZ L, BARROS R J D, DE MIGUEL I, et al. Routing, modulation and spectrum assignment algorithm using multi-path routing and best-fit[J]. IEEE Access, 2021, 9: 111633–111650.

[5] LIU Y, HE R, WANG S, et al. Temporal and spectral 2D fragmentation-aware RMSA algorithm for advance reservation requests in EONs[J]. IEEE Access, 2021, 9: 32845–32856.

[6] 王世成，陈晓静，何荣希. 弹性光网络中基于业务承载力的碎片感知路由与频谱分配算法[J]. 激光与光电子学进展，2022，59(7)：0706007-1–0706007-9.

[7] 刘鑫，何荣希，陈晓静. 弹性光网络中基于自适应调制的碎片感知共享通路保护算法[J]. 光通信技术，2020，44(2)：52–58.

[8] CHEN X, LI B, PROIETTI R, et al. DeepRMSA: a deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks[J]. Journal of Lightwave Technology, 2019, 37(16): 4155–4163.

[9] CHEN X, PROIETTI R, YOO S J B. Building autonomic elastic optical networks with deep reinforcement learning[J]. IEEE Communications Magazine, 2019, 57(10): 20–26.

[10] SHI C, ZHU M, GU J, et al. Deep-reinforced impairment-aware dynamic resource allocation in nonlinear elastic optical networks [C]//2021 Opto-Electronics and Communications Conference, July 3–7, 2021, Hong Kong, China. New York: IEEE, 2021: 1–3.

[11] XU L, HUANG Y C, XUE Y, et al. Deep reinforcement learning-based routing and spectrum assignment of EONs by exploiting GCN and RNN for feature extraction [J]. Journal of Lightwave Technology, 2022, 40(15): 4945–4955.

[12] TANG B, HUANG Y C, XUE Y, et al. Heuristic reward design for deep reinforcement learning-based routing, modulation and spectrum assignment of elastic optical networks[J]. IEEE Communications Letters, 2022, 26(11): 2675–2679.

[13] ZHU Z, LU W, ZHANG L, et al. Dynamic service provisioning in elastic optical networks with hybrid single-/multi-path routing [J]. Journal of Lightwave Technology, 2013, 31(1): 15–22.

[14] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C]//PMLR. Proceedings of International conference on machine learning. Now York: PMLR, 2016: 1928–1937.