文章编号：2095-2163(2022)09-0094-07 中图分类号：TN929.5 文献标志码：A

基于强化学习和神经网络的导频功率动态优化

李 烨，肖梦巧
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：为了更好地实现用户对移动通信网络的无线接入，合理分配基站导频功率十分重要，本文研究了无线接入网中基站导频功率的动态优化问题，设计了一种结合强化学习和神经网络的导频功率优化模型，以感知无线接入网的变化。其次，利用 Q 学习算法来维持基站与外部环境的连接和信息交互；在 Q 学习算法中，利用神经网络学习 Q 值，避免了状态爆炸问题；最后设计了关键性能指标保护机制和回退机制，以满足工程要求。仿真结果表明，提出方案能很好地适应无线网络频繁变化，并获得显著的性能增益。

关键词：无线接入网；导频功率；Q 学习；神经网络；关键性能指标保护机制；回退机制

Dynamic optimization of pilot power based on reinforcement learning and neural network

LI Ye, XIAO Mengqiao
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 In order to better realize the wireless access of users to the mobile communication network, it is very important to reasonably allocate the pilot power of the base station. This paper studies the dynamic optimization of the pilot power of the base station in the radio access network. Firstly, a pilot power optimization model combining reinforcement learning and neural network is designed to sense the changes of the radio access network. Secondly, the Q - learning algorithm is used to maintain the connection and information interaction between the base station and the external environment. In the Q - learning algorithm, the neural network to learn the Q value is used for avoiding the state explosion problem. Finally, a key performance indicator protection mechanism and a rollback mechanism are designed to meet engineering requirements. Simulation results show that the proposed scheme can well adapt to the frequent changes of wireless networks and achieve significant performance gains.

【Key words】 radio access network; pilot power; Q - learning; neural network; key performance indicator protection mechanism; rollback mechanism

0 引 言

由于 5G 研究的全面发展、无线接入技术的多样化和频谱资源的整合，无线接入网[1]（Radio Access Network，RAN）在处理性能、协调能力和业务部署等方面面临着新的挑战。因此，RAN 架构需要取得显著进展，以满足未来的各种需求。同时，云计算、大数据和虚拟化技术在核心网络中的应用，为 RAN 架构的演进提供了坚实的基础。

为了更好地实现用户对移动通信网络的无线接入，合理分配基站导频功率十分重要，因其影响着网络的覆盖。作为下行链路功率的一部分，导频功率与其它下行信道共享额定的基站功率。一方面，过多的导频功率分配会增加小区重叠区域，从而导致下行链路干扰和小区重叠区域的增加，这也可能将导致导频污染问题[2]；另一方面，导频能力不足将导致覆盖漏洞，从而减少所支持的业务。

为了实现网络性能的最大化，一些专家学者对导频功率分配优化问题进行了研究。Ma 等人[3]以渐近信干噪比为目标，将导频分配问题表述为最小权重多指标分配问题。该方案提高了系统性能，但算法复杂度较高。为了降低导频分配算法的复杂度，Omid 等人[4]提出一种低复杂度的导频分配策略，采用 SCP 的迭代，构造求解局部优化的非凸问题，有效降低了迭代算法的复杂度。Jang 等人[5]提出的中下行多用户、多输入、多输出系统的节能设

基金项目：华为技术有限公司合作项目（YBN2019115054）。

作者简介：李 烨（1974-），男，博士，高级工程师，主要研究方向：机器学习、图像处理、移动通信；肖梦巧（1994-），女，硕士研究生，主要研究方向：无线接入网、大规模 MIMO、强化学习。

通讯作者：李 烨 Email：liye@usst.edu.cn

收稿日期：2022-02-28

计，考虑了导频功率、数据功率和速率自适应。Liu 等人[6]通过优化导频功率配置，使基站总功率更加合理。在 RAN 中，导频功率一般是依靠人工经验进行配置，后期再根据需求逐步进行人工优化。由于小区导频功率变化后，会同步影响周边邻区。如果导频功率配置过大，会对邻区造成干扰；导频功率配置过小，又会造成覆盖空洞。因此，导频功率优化不能仅针对单小区进行处理，还要对整网或整片区域进行联合动态优化。

基于此，本文提出了一种基于强化学习和神经网络的导频功率动态优化方案，设计了一种结合强化学习和神经网络的新型模型。该模型研究了导频功率与网络性能增益之间的关系，通过最大化网络性能来适应连续变化的 RAN 环境；以网络环境状态和导频功率调整值作为 Q 学习的输入，网络流量和容量作为输出。由于 Q 表不适用连续状态空间，因此结合神经网络，将状态和动作映射到 Q 值，使得整个系统更加灵活。此外，为了确保网络的稳定性和连续调整导频功率的可行性，通过有效分析历史数据，并充分利用所获得的实时数据，提出了关键性能指标（Key Performance Indicator，*KPI*）保护机制和回退机制，以满足工程要求。

1　系统模型

假设一个覆盖区配置一个中心小区和被动联动调整小区。中心小区根据本小区配置和负载状态以及被动联动调整小区的负载状态，进行导频功率联动调整，从而优化覆盖区的网络性能，实现覆盖区内基站间的负载均衡。

覆盖区包含数据模块和导频模块。其中，数据模块负责采集各类数据（如基站配置数据等），在 Q 学习算法和 *KPI* 保护机制中使用，历史数据也用于 *KPI* 基线计算；导频模块与数据模块交互，获取运行环境中所有网络状态信息，实时识别神经网络模块的状态。Q 学习算法在每次迭代中向导频模块提供最优的导频功率调整动作，从而根据神经网络的输出获得良好的 RAN 性能增益。

2　算法描述

2.1　参数

Q 学习是最流行的强化学习算法之一，旨在处理马尔科夫决策过程问题[7]。本文将每个小区的基站建模为智能体，每个基站维护自己的 Q 值表，以降低优化复杂度。结合 Q 学习模型中智能体、环境、动作、状态及奖励五大元素，对该问题进行建模[8]。与 Q 学习相关的所有参数定义如下：

（1）智能体：智能体通过与环境进行交互获取奖励值（*reward*），来学习改善自己的策略，从而获得该环境下最优策略。在导频功率优化问题中，将每个小区基站作为一个智能体。

（2）环境：本文将 RAN 作为与智能体进行交互的环境。

（3）状态：每个智能体都有各自的状态向量。本文基站的状态向量可定义为如下五元组：

$$s_i = [z_1, z_2, z_3, z_4, z_5] \tag{1}$$

其中，z_1 为小区网络 TCP 负载；z_2 为用户设备数量；z_3 为当前导频功率；z_4、z_5 分别表示参考信号接收功率（Reference Signal Receiving Power，RSRP）分布的均值和方差；s_i 表示基站采用某一个动作后，同覆盖区内所有小区中用户导频功率分配状态。

（4）动作：每个智能体都有一个动作集合 A，即每个小区基站对本小区用户进行导频功率分配的调整值集合，定义为：

$$a_i \in A \tag{2}$$

导频功率的最大值和最小值限制可表示为：

$$\begin{cases} Maximum = MaxBSpower \cdot r_{\max} \\ Minimum = MaxBSpower \cdot r_{\min} \end{cases} \tag{3}$$

其中，$r_{\max}$ 和 $r_{\min}$ 分别表示导频功率与基站功率的最大和最小比值。调整后的导频功率应限制在一定范围内，即：$[Minimum, Maximum]$。对于超出最大值或最小值的值，将其调整为最大值或最小值。

（5）奖励：表示智能体在当前状态下选择动作获得的收益、即网络增益，由流量和容量两部分组成。由于接入 RAN 的用户设备数量在不断变化，系统需要消除网络波动和附加增益（正/负增益）的影响，因此在奖励计算中引入相对增益的概念，以保证算法带来增益。

定义 1　相对 RAN 流量增益　数学定义式可写为：

$$r_T = (T_{i+1}/L_{i+1} - T_i/L_i)/(T_i/L_i) \tag{4}$$

其中，r_T 是各状态 i 到状态 $i+1$ 之间的相对流量增益；T_i 表示网络业务（如呼叫建立）数量的网络流量；L_i 为 BS 的 TCP 负载，表示网络资源的利用率；T_i/L_i 反映单位资源占用下，BS 支持的业务数量。

定义 2　相对 RAN 容量增益　数学定义式可写为：

$$r_C = (C_{i+1} - C_i)/C_i \tag{5}$$

其中，r_C 表示状态 i 到状态 $i+1$ 的相对容量增

益，C_i 为网络容量，描述了基站支持的最大网络吞吐量。

因此，奖励由 RAN 相对流量增益 r_T 和 RAN 相对容量增益 r_C 共同计算，即：

$$r_i(s_i,a_i,s_{i+1}) = \omega \cdot r_T + (1-\omega) \cdot r_C \quad (6)$$

其中，$\omega \in [0,1]$ 量化了 2 部分重要性之间的权衡。

2.2 Q 学习与神经网络联合优化算法

在 Q 学习中，估计动作值函数 $Q(s,a)$ 用来学习最优导频功率分配方案，从而在执行动作 a 的状态中获得最大期望奖励。换言之，在每个步骤 a 处选择使函数 $Q(s,a)$ 最大化的动作。$Q(s,a)$ 的更新为：

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (7)$$

其中，a 表示当前动作；s 表示当前状态；a' 表示状态中任何可能的动作；s' 表示采取行动后的新状态；r 是在状态 s 下根据特定动作 a 获得的立即奖励；$\alpha \in [0,1]$ 表示学习率；$\gamma \in [0,1]$ 表示延迟与立即奖励的相对值的折扣因子。

动作选择机制，负责选择代理执行的操作。在本文中，采用 ε - 贪婪策略，对应的数学公式如下：

$$a_i = \begin{cases} random\ a\ from\ A & \text{if } \xi < \varepsilon \\ arg\max_{a \in A(s)} Q(s,a) & otherwise \end{cases} \quad (8)$$

其中，$\varepsilon \in [0,1]$ 为固定概率；$\xi \in [0,1]$ 表示时间步长 i 上的一致随机数；A 为可选择的动作集。该规则利用概率 $(1-\varepsilon)$ 选取最佳动作，利用概率 ε 进行探索。

在迭代过程中，Q 学习算法通常使用 Q 表来储存不同时刻的状态动作值。这一算法在面对大规模数据空间或连续数据的任务时非常低效。因此，在导频功率优化问题中，采用 Q 表单独存储每个 Q 因子是并不现实的。本文利用非线性函数来近似 $Q(s,a;\theta)$，这里的 θ 描述了近似的可调参数。在此情况下，通常利用神经网络处理状态空间爆炸问题[9]，神经网络结构如图 1 所示。

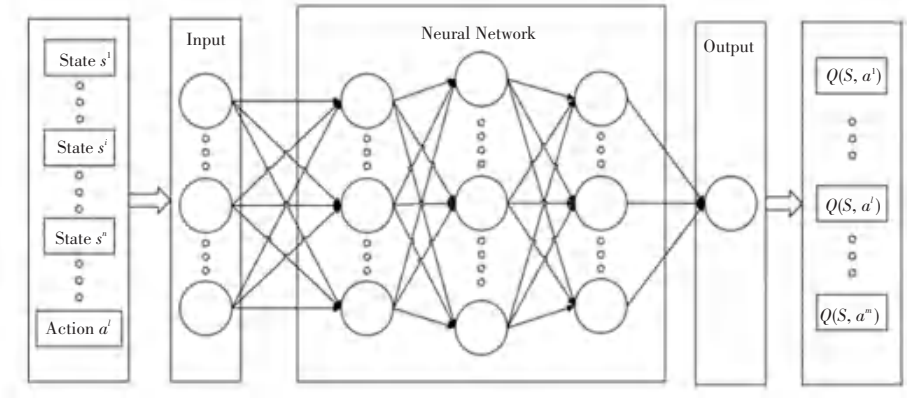

图 1 神经网络结构

Fig. 1 The structure of Neural Network

由图 1 可以看出，神经网络的输入是模型和导频功率调整前的状态，输出是导频可以采取的每个动作的 Q 值。根据 Q 学习算法选择导频功率最优动作，智能体从环境中获得真正的收益。

综合前述可知，通过实际回报与预测回报之间的误差训练算法权重，并在迭代过程中利用梯度下降法进行更新。

研究推得，神经网络模型的输入为：

$$X = [s^1,\cdots,s^i,\cdots,s^n,a^l] \quad (9)$$

其中，$s = (s^1,\cdots,s^i,\cdots,s^n)$ 表示实际状态映射到状态空间 s，$a^l \in A$ 是智能体在该状态下可以采取的动作。

神经网络模型的输出为基于状态 s 的 Q 学习算法的 Q 值。此外，神经网络采用直接梯度下降法更新参数。在 Q 学习中，通过最小化样本上的损失函数来训练神经网络，其损失函数公式见如下：

$$L(\theta) = \{((s,a) + \alpha \cdot [r + \gamma \max_{a'} Q(s',a') - Q(s,a)]) - Q'(s,a)\}^2 \quad (10)$$

其中，$Q(s,a)$ 是预测值，$Q'(s,a)$ 是真实值。

利用神经网络估计每个动作的价值函数 $Q(s,a^l)$，采用动作的函数值 a^l 进行估计。进而传统的 Q 学习算法中的 Q 表被替换为：

$$Q(s,a) = [Q(s,a^1),\cdots,Q(s,a^i),\cdots,Q(s,a^m)] \quad (11)$$

2.3 KPI 保护机制

实际应用中，在不出现急剧恶化的情况下，网络性能通常表现为某些 *KPI* 指标。该模型提供各种 *KPI* 保证，并为每个 *KPI* 定义一个基线。如果导频功率调整后，*KPI* 的计算值低于基线，则奖励为 0。因此，奖励功能改进为：

$$r_i(s_i,a_i,s_{i+1})=\{\omega\cdot r_T+(1-\omega)\cdot r_C\}\cdot\prod f(KPI_{i,j}) \tag{12}$$

其中，$f(x)$ 为单位步长函数，使 *KPI* 服从正态分布。

因此，ξ_{KPI_j} 为 KPI_j 的基线，基线 ξ_{KPI_j} 为：

$$\xi_{KPI_j}=\mu_j-3\sigma_j \tag{13}$$

其中，μ_i 和 σ_i 分别是从数据模块中得到的历史数据，经计算得出的 *KPI* 平均值和标准差，结果值见表 1。

表 1　6 个关键性能指标

Tab. 1　Six key performance indicators

类别	中文名称	英文缩略语
SER	RRC 建立成功率	*RRC_SER*
	PS 无线接入承载建立成功率	*PS_RAB_SER*
	CS 无线接入承载建立成功率	*CS_RAB_SER*
CDR	PS 业务掉话率	*PS_CDR*
	CS 业务掉话率	*CS_CDR*
——	AMR 语音话务量	*AMR. Erlang*

2.4 协同优化

在 RAN 中，由于用户设备的位置在不断移动，同时导频功率的调整将影响基站服务范围，因此需要进行软切换操作。软切换比例[9-10]在一定程度上能较好地反映基站的活跃度，基站的软切换比例越高，用户在基站的覆盖范围内进行的通信越多，基站对覆盖区域网络性能的考量就越重要。在进行整体性能优化时，需要考虑软切换比例所连接的所有基站之间的协同优化。则奖励函数计算为：

$$r_i(s_i,a_i,s_{i+1})=\sum_{t=1}^{k}\eta^t\cdot r^t(s_i^t,a_i^t,s_{i+1}{}^t) \tag{14}$$

其中，$r^t(s_i^t,a_i^t,s_{i+1}{}^t)$ 是 BS 的奖励（可按式(12)计算）；k 为覆盖区内 BS 的个数；η^t 为 BS t 的软切换比例，可按式(15)计算得到：

$$\eta^t=\frac{N_t}{N_{all}} \tag{15}$$

其中，N_t 是用户设备从其它 BS 到 BS t 与 BS 到其它 BS 的软切换次数之和；N_{all} 是所有 BS 之间的软切换次数之和。

结合式(12)、(14)和(15)，整个覆盖区（所有相邻 BS）的奖励函数为：

$$r_i(s_i,a_i,s_{i+1})=\sum_{t=1}^{k}\frac{N_t}{N_{all}}\{\omega[(T_{i+1,t}/L_{i+1,t}-T_{i,t}/L_{i,t})/(T_{i,t}/L_{i,t})]+(1-\omega)[(C_{i+1,t}-C_{i,t})/C_{i,t}]\}\cdot\prod f(KPI_{i,j,t}) \tag{16}$$

2.5 KPI 回退机制

回退是指模型将任务回退到原来的状态，是动态调整系统中的重要机制之一[11]。回退机制增强了系统的灵活性，提高了模型在复杂情况下的动态处理能力。本文结合 Q 学习算法中的回退机制，通过式(16)评估 RAN 性能，若调整后 RAN 流量和容量下降 20%或 *KPI* 下降到基线以下，则执行回退机制。如果采用回退机制，模型状态不做调整，基站导频功率会恢复到上次的时间 i 值。换言之，实际动作 a_i'' 与 Q 学习算法直接给出的动作 a_i 不同，使用实际发送到导频功率模型的动作更新，调整神经网络损失函数：

$$L(\theta)=\{(Q(s,a'')+\alpha[r+\gamma\max Q(s',a'')-Q'(s,a'')])-Q'(s,a'')\}^2 \tag{17}$$

Q 值定义为：

$$Q(s,A)=[Q(s,a^1),\cdots,Q(s,a'),\cdots,Q(s,a^m)] \tag{18}$$

其中，$Q(s',a'')$ 为最佳动作选择 a^l 和状态 s 下新的 Q 值。在有回退机制时，用 $Q(s',a'')$ 替换 $Q(s',a^l)$。

3 仿真结果与分析

3.1 仿真参数设置

本文通过冷启动仿真和模型试验，给出了导频功率动态优化仿真中的设置参数：导频功率与基站功率的最小和最大比值 $r_{\min}$ 和 $r_{\max}$ 分别为 5%和 20%，系统模型的生命周期 T 为 24 h，折扣因子 ω 为 0.7，覆盖区基站数量 k 为 10。神经网络的输入范围与 *KPI* 的均值和标准方差分别见表 2、表 3。

表 2　神经网络的输入范围

Tab. 2　The input ranges of the Neural Networks

输入	范围
TCP 负载	(0,1]
用户设备数量	[0,150]
导频功率/dBm	[30,36]
RSRP 均值	[50,150]
RSRP 标准方差	—
导频功率动作 A/dB	[-2, -1,0,1,2]

表 3 *KPI* 的均值和标准方差

Tab. 3 The means and standard variances of the *KPIs*

KPI	中文名称	英文缩略语
RRC_SER	0.998 407	0.008 846
PS_RAB_SER	0.999 376	0.000 629
CS_ RAB_SER	0.999 477	0.003 353
PS_CDR	0.049 596	0.003 578
CS_CDR	0.001 163	0.000 395
AMR. Erlang	1 397.846 500	897.265 600

3.2 性能分析

图 2 给出了冷启动期间的每代通信量。图 2 中,实线对应使用预训练得到的权重初始化神经网络的情况,虚线对应随机初始化神经网络的情况。

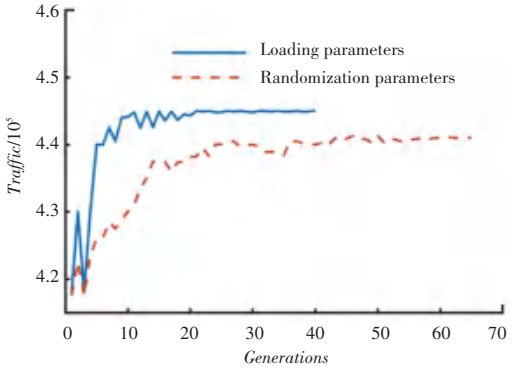

图 2 导频功率优化

Fig. 2 Pilot power optimization

由图 2 可见,冷启动过程中使用的离线数据集合是在密集区域中心,负荷高。用户设备位置在基站中随机均匀分布,因此容量是恒定的,冷启动效果可以通过对比流量来体现。具有随机参数的神经模型经过约 45 次迭代后几乎收敛,神经网络的收敛速度明显提高。从图 2 中实线可以看出,经过 10 次左右的迭代后,网络流量增益可以提升约 6.2%。

导频功率的结果对比如图 3 所示。从图 3 中可以看出:导频功率是下调的,调整后的大多数基站导频功率为 30 dBm,该结果与预期一致,说明导频功率配置更高效、更稳定。冷启动方法可以作为神经网络的初始权重,从而提高早期模型的效率。

具有神经网络的 Q 学习算法进行导频功率优化的结果如图 4、图 5 所示。图 4、图 5 中包括不同网络波动和用户数。

从图 4、图 5 可以看出:相对流量增长 7%,相对容量增长 16%。结果表明,该模型能够有效地解决基站导频功率动态实时调整问题,并在获得更多话务量和充足容量的同时,获得了更好的网络性能。

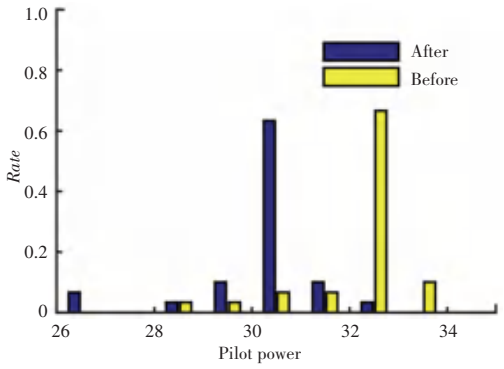

图 3 导频功率对比

Fig. 3 The comparison of pilot power

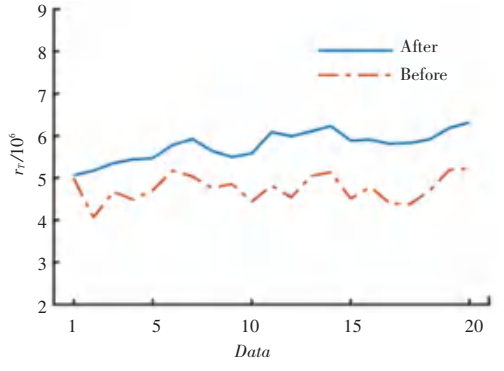

图 4 相对流量增益 r_T 前后对比

Fig. 4 The comparison of relative traffic gain r_T

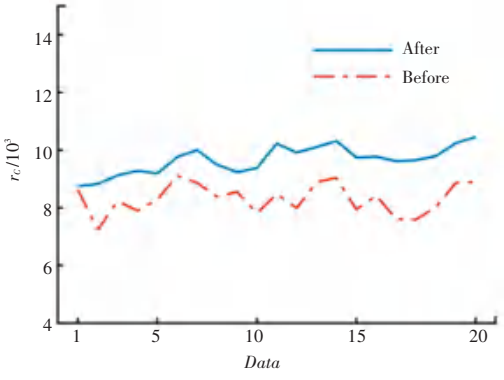

图 5 相对容量增益 r_C 前后对比

Fig. 5 The comparison of relative capacity gain r_C

研究得到的各指标的 *KPI* 对比如图 6 所示。由图 6 中每个 *KPI* 的比较显示可以看出,本文选取的 *KPI* 指标,可以有效反映网络性能的稳定性和用户接入的可靠性。显然,部署后 *KPI* 值更稳定。

在测试期间,各指标均值和标准差见表 4、表 5,可见各指标值均得到改善,表明本文提出的系统模型在保证关键性能指标稳定性的同时,提高了当前网络的性能,进而为智能基站的发展打下基础。

表 4　*KPI* 的均值对比

Tab. 4　The average comparison of *KPI*s

	RRC_SER	*PS_RAB_SER*	*CS_RAB_SER*	*PS_CDR*	*CS_CDR*	*AMR. Erlang*
Before	0.998 417	0.999 39	0.999 366	0.050 622	0.001 073	13 468.803 6
After	0.998 777	0.999 40	0.999 657	0.049 200	0.000 881	13 674.345 9

表 5　*KPI* 标准方差对比

Tab. 5　The standard variances comparison of *KPI*s

	RRC_SER	*PS_RAB_SER*	*CS_RAB_SER*	*PS_CDR*	*CS_CDR*	*AMR. Erlang*
Before	0.000 874	0.000 729	0.000 305	0.003 569	0.000 385	927.687 6
After	0.000 120	0.000 547	0.000 045	0.001 401	0.000 055	827.265 6

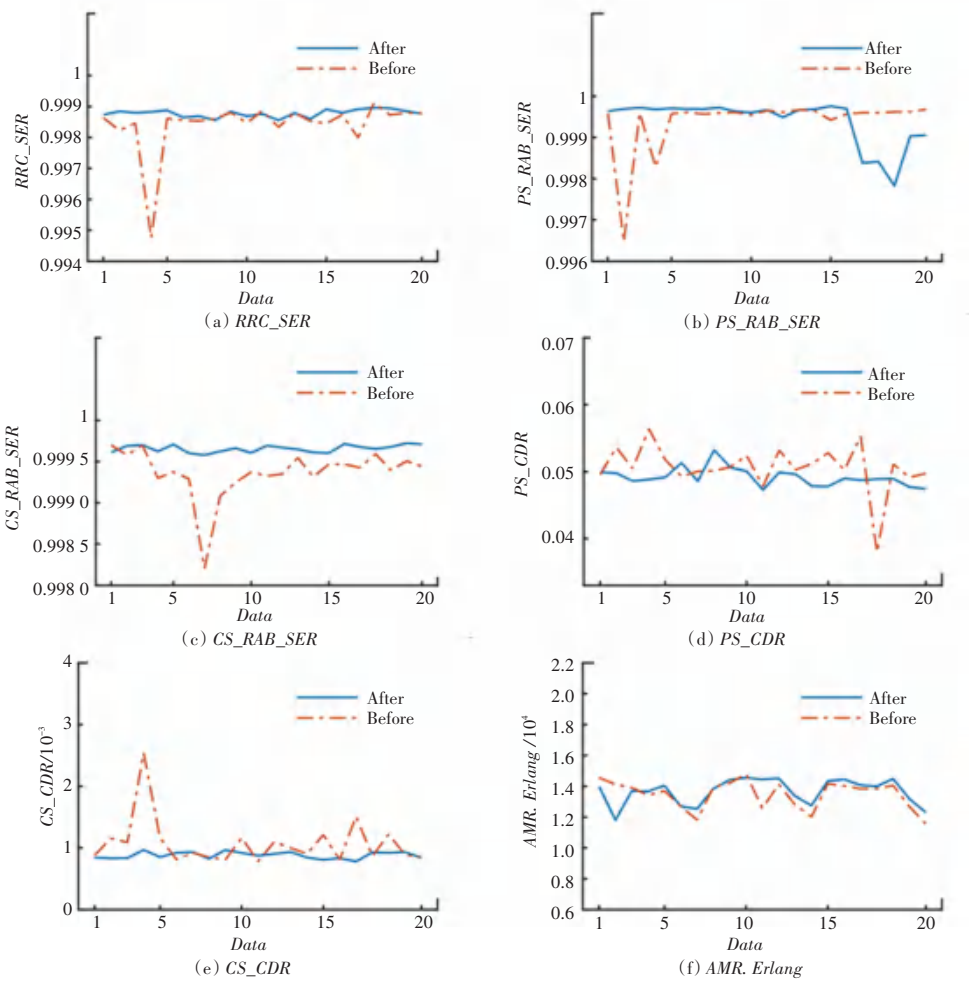

图 6　*KPI* 对比

Fig. 6　The comparison of *KPI*s

4　结束语

本文研究了强化学习与神经网络联合优化导频功率的方法。在 RAN 中，设计了一个基站覆盖区系统模型，建立了导频功率与网络性能的关系，使得网络流量和容量最大化；在 Q 学习奖励计算中提出了相对增益的概念，并利用软切换比例将同覆盖区基站进行协同优化；利用神经网络解决了 Q 表状态空间爆炸问题；增加冷启动程序，以减少算法参数随机化的影响。此外，提出了 *KPI* 保护和回退机制，保证导频功率部署的稳定性和可靠性。仿真结果表明，所提算法能够很好地解决基站导频功率的动态调整问题，在 RAN 环境变化中取得了很大的优势。后续将考虑导频功率与小区实际覆盖情况和小区边缘用户分布的影响，进一步优化基站导频功率。

（下转第 104 页）