基于深度强化学习的蜂窝网资源分配算法

廖晓闽[1,2]，严少虎[3]，石嘉[1]，谭震宇[1]，赵钟灵[1]，李赞[1]

（1. 西安电子科技大学综合业务网理论及关键技术国家重点实验室，陕西 西安 710071；
2. 国防科技大学信息通信学院，陕西 西安 710106；3. 中国电子科技集团公司第二十九研究所，四川 成都 610036）

摘 要：针对蜂窝网资源分配多目标优化问题，提出了一种基于深度强化学习的蜂窝网资源分配算法。首先构建深度神经网络（DNN），优化蜂窝系统的传输速率，完成算法的前向传输过程；然后将能量效率作为奖惩值，采用 Q-learning 机制来构建误差函数，利用梯度下降法来训练 DNN 的权值，完成算法的反向训练过程。仿真结果表明，所提出的算法可以自主设置资源分配方案的偏重程度，收敛速度快，在传输速率和系统能耗的优化方面明显优于其他算法。

关键词：蜂窝网；资源分配；深度强化学习；神经网络

中图分类号：TN929.5

文献标识码：A

doi: 10.11959/j.issn.1000−436x.2019002

Deep reinforcement learning based resource allocation algorithm in cellular networks

LIAO Xiaomin[1, 2], YAN Shaohu[3], SHI Jia[1], TAN Zhenyu[1], ZHAO Zhongling[1], LI Zan[1]

1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an 710071, China
2. School of Information and Communications, National University of Defense Technology, Xi’an 710106, China
3. The 29th Research Institute of China Electronics Technology Group Corporation, Chengdu 610036, China

Abstract: In order to solve multi-objective optimization problem, a resource allocation algorithm based on deep reinforcement learning in cellular networks was proposed. Firstly, deep neural network (DNN) was built to optimize the transmission rate of cellular system and to complete the forward transmission process of the algorithm. Then, the Q-learning mechanism was utilized to construct the error function, which used energy efficiency as the rewards. The gradient descent method was used to train the weights of DNN, and the reverse training process of the algorithm was completed. The simulation results show that the proposed algorithm can determine optimization extent of optimal resource allocation scheme with rapid convergence ability, it is obviously superior to the other algorithms in terms of transmission rate and system energy consumption optimization.

Key words: cellular networks, resource allocation, deep reinforcement learning, neural network

1 引言

随着无线网络中通信设备数量的急剧增加和业务需求的多样化，有限的频谱资源与人们日益增长的无线频谱需求之间的矛盾日渐突出和加剧。当前无线通信领域面临着智能化、宽带化、多元化、综合化等诸多技术挑战，无线网络环境变得日益复杂多样和动态多变，此外，绿色网络和智慧网络等新概念的提出，使频谱资源管理的优化目标日趋多样化，因此，如何优化频谱利用，最大限度地实现

收稿日期：2019−01−19；**修回日期**：2019−02−15

通信作者：严少虎，youngtiger@263.net

基金项目：国家自然科学基金重点资助项目（No.61631015）

Foundation Item: The Key Project of National Natural Science Foundation of China (No.61631015)

频谱资源的高效管理是当前急需解决的重点问题。

传统蜂窝网资源分配方法主要有博弈理论、拍卖机制、图论着色理论、遗传算法等。Huang 等[1]将博弈理论应用于小区间蜂窝网的频谱分配，假设基站预先获得且共享信道状态信息（CSI，channel state information），将 2 个通信设备放置于相邻小区的重叠区域，采用静态重复的古诺博弈模型来求解纳什均衡解，获得最优的频谱效率，仿真模拟 3 种典型场景，通过求解一系列优化方程式来获得最优分配策略。Wang 等[2]提出了一种安全的频谱拍卖机制，该机制综合考虑频谱属性和拍卖特性，采用自适应竞价、信息加密和拍卖协议等方式，在提高频谱利用率的同时，极大地提升频谱拍卖机制的安全性。Yang 等[3]采用图论着色理论对全双工设备到设备（D2D，device-to-device）蜂窝网进行频谱和功率分配，构造干扰感知图，提出了一种基于图论着色理论的资源共享方案，该方案以网络吞吐量为优化目标，算法收敛速度快，时间复杂度低。Takshi 等[4]基于遗传算法实现 D2D 蜂窝网中的频谱和功率分配，通过同时搜索不同区间，获得全局最优的频谱效率和干扰性能，而且蜂窝网用户的信干噪比保持最低，对 D2D 用户数量没有限制，并且采用信道预测方法来减少 CSI 信息的过载，算法具有较强的搜索性能。然而，随着未来无线网络向高密集、大数据、动态化、多目标优化等方向发展，传统的蜂窝网资源分配方法不再适用，例如，传统方法主要进行静态优化，很难适应动态变化的环境；当多目标优化问题为 NP-hard 问题时，求解困难；没有发挥出大数据优势，无法充分挖掘隐藏在数据中的信息等。

当前，以机器学习、深度学习为代表的新一代人工智能技术已广泛应用于医疗、教育、交通、安防、智能家居等领域，从最初的算法驱动逐渐向数据、算法和算力的复合驱动转变，有效地解决了各类问题，取得了显著成效。目前，机器学习在无线资源分配的研究还处于早期探索阶段。例如，文献[5]提出采用深度学习方法对 LTE 中未授权频谱进行预分配，利用长短期记忆（LSTM，long short-term memory）神经网络来学习历史经验信息，并利用学习训练好的 LSTM 网络对未来某一窗口的频谱状态进行预测；文献[6]采用深度神经网络（DNN，deep neural network）对认知无线电中次用户使用的频谱资源和传输功率进行分配，最大化次用户频谱效率的同时，尽可能地减少对主用户造成的干扰；文献[7]将卫星系统中的动态信道分配问题建模成马尔可夫决策过程，采用深度卷积神经网络提取有用特征，对信道进行动态分配，有效地减少阻塞率，提高了频谱效率。目前，机器学习方法可以充分利用大数据的优势，模拟人类的学习行为，挖掘数据隐藏信息，以获取新的知识，然后对已有的知识结构进行重组，不断地改善自身的性能。此外，机器学习还可以实现动态实时交互，具有很强的泛化能力，在无线资源分配应用中凸显优势。

本文考虑优化蜂窝网的传输速率和系统能耗，基于深度强化学习提出了一种全新的蜂窝网资源分配算法，该算法分为两部分，即前向传输过程和反向训练过程。在前向传输过程中，考虑优化蜂窝网传输速率，采用增广拉格朗日乘子法，构建频率分配、功率分配和拉格朗日乘子的迭代更新数据流，在此基础上，构造 DNN。在反向训练过程中，将能量效率作为奖惩值，构建误差函数来反向训练 DNN 的权值。前向传输过程和反向训练过程反复迭代，直到满足收敛条件时，输出最优资源分配方案。本文所提算法可以通过调整误差函数中的折扣因子来自主设置频谱分配策略的偏重程度，收敛速度快，在传输速率和系统能耗的优化方面明显优于其他算法，能够有效地解决多目标优化问题。

2 系统模型

考虑蜂窝网的下行链路，假设蜂窝移动通信系统中有 M 个微基站和 N 个授权移动用户，用户随机分布在小区内，所有基站和用户都为单天线系统。在每个小区内采用正交频分复用（OFDM，orthogonal frequency division multiplexing），每个频率只分配给一个用户使用，其他小区可以重复使用频率，即采用完全频率重用方案，因此从实际出发，综合考虑蜂窝网中所有基站对移动用户造成的干扰情况。系统采用集中式控制，信道增益、噪声功率等精确信道状态信息未知，每个授权移动用户仅将位置信息、干扰和传输速率通过导频信号传输给中心控制节点，由中心控制节点制定频谱分配方案。为了建设绿色网络，系统在最大化传输速率的过程中，还需要考虑能耗问题，具体的系统模型如图 1 所示。

假设 $m=\{1,2,\cdots,M\}$ 表示微基站的集合，$n=\{1,2,\cdots,N\}$ 表示移动用户的集合，$k=\{1,2,\cdots,K\}$ 表示可用频率的集合。基站 m 中的移动用户 n 使用频率 k 通信时，干扰信号为

$$I_{m,n}^{k}=\sum_{i=1,i\neq m}^{M}\sum_{j=1}^{N}L_{i,j}D_{i,j}^{k}p_{i,j}^{k}\left|h_{i,n}^{k}\right|^{2} \quad (1)$$

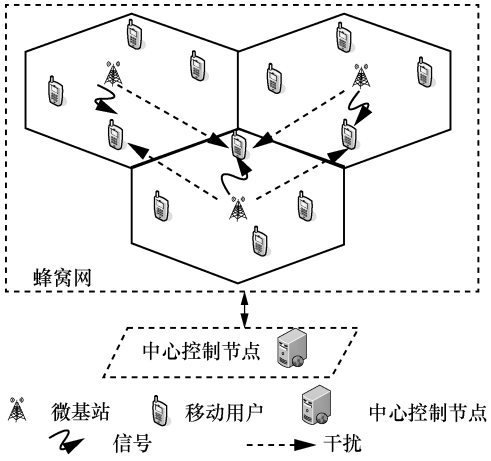


图 1　系统模型

其中，$L_{i,j}$ 表示移动用户 j 与基站 i 的接入关系，若移动用户 j 接入基站 i，则 $L_{i,j}=1$，反之 $L_{i,j}=0$；$D_{i,j}^{k}$ 表示基站 i 内频率 k 的分配情况，若基站 i 把频率 k 分给移动用户 j，则 $D_{i,j}^{k}=1$，反之 $D_{i,j}^{k}=0$；$p_{i,j}^{k}$ 表示基站 i 使用频率 k 与用户 j 通信时的功率；$h_{i,n}^{k}$ 表示基站 i 使用频率 k 与用户 n 通信时的信道增益。

系统总体的传输速率可以表示为

$$R=\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{lb}\left(1+\frac{L_{m,n}D_{m,n}^{k}p_{m,n}^{k}\left|h_{m,n}^{k}\right|^{2}}{\sigma_{m,n}^{2}+I_{m,n}^{k}}\right) \quad (2)$$

其中，$\sigma_{m,n}^{2}$ 表示基站 m 与用户 n 通信时的噪声。

采用文献[8]提出的能量效率来衡量系统能耗，即将每焦耳的能量最多能携带多少比特（单位为 bit/J）作为衡量标准，则系统总体的能量效率可以表示为

$$H=\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{B_{m,n}^{k}\mathrm{lb}\left(1+\dfrac{L_{m,n}D_{m,n}^{k}p_{m,n}^{k}\left|h_{m,n}^{k}\right|^{2}}{\sigma_{m,n}^{2}+I_{m,n}^{k}}\right)}{p_{m,n}^{k}} \quad (3)$$

其中，$B_{m,n}^{k}$ 表示基站 m 使用频率 k 与用户 n 通信时的带宽。

根据系统优化目标，在基站子载波发射功率之和满足最大发射功率约束的条件下，要解决的多目标优化问题描述如式(4)~式(6)所示。

$$\max_{\{D_{m,n}^{k},p_{m,n}^{k}\}} \quad R \quad (4)$$

$$\max_{\{D_{m,n}^{k},p_{m,n}^{k}\}} \quad H \quad (5)$$

$$\text{约束条件为}\sum_{n=1}^{N}\sum_{k=1}^{K}L_{m,n}D_{m,n}^{k}p_{m,n}^{k}\leqslant P_{m}^{\max},\ \forall m\in\{1,2,\cdots,M\} \quad (6)$$

其中，$P_{m}^{\max}$ 表示基站 m 的最大发射功率。

3　基于深度强化学习的资源分配算法

本文除了考虑传输速率外，还综合考虑能耗，于是资源分配问题变成了 NP-hard 问题，难以求得最优解。目前研究热点是将该问题转化为求解其次优解，但是求解复杂度高，影响系统运行效率[9]，本文采用深度强化学习方法来求解该问题。

3.1　算法框架

深度强化学习将深度学习的感知能力和强化学习的决策能力相结合，不断以试错的方式与环境进行交互，通过最大化累积奖赏的方式来获得最优策略[10]。本文采用深度 Q 网络（DQN，deep Q-network）来具体求解资源分配问题，核心思想是将值网络作为评判模块，基于值网络来遍历当前观测状态下的各种动作，与环境进行实时交互，将状态、动作和奖惩值存储在记忆单元中，采用 Q-learning 算法来反复训练值网络，最后选择能获得最大价值的动作作为输出。基于深度强化学习的资源分配算法的基本框架如图 2 所示。

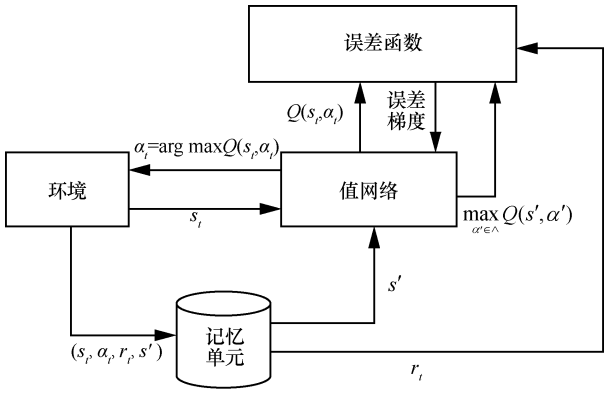


图 2　基于深度强化学习的资源分配算法的基本框架

在图 2 中，s_t 为算法进行到第 $t(t=1,2,\cdots,T)$步时所对应的观测，a_t 为观测 s_t 下所执行的动作，r_t 为观测 s_t 下执行动作 a_t 后，所获取的奖赏/惩罚，值网络采用 DNN 来描述，即将 DNN 作为动作状态值函数 $Q(s_t,a_t)$。

算法采用Q-learning学习机制[11]，主要根据如式(7)所示的迭代式来实现动作状态值函数的优化学习。

$$\begin{cases} Q_{k+1}(s_t,a_t)=Q_k(s_t,a_t)+\alpha_k E_k \\ E_k=r_t+\gamma\max\limits_{a'\in\wedge}Q_k(s',a')-Q_k(s_t,a_t) \end{cases} \tag{7}$$

其中，α_k 是学习速率，$\gamma\in(0,1)$ 为折扣因子，s' 为执行动作 a_t 后获得的观测值，a' 为动作集合 $\wedge$ 中使得第 k 次迭代下的动作状态值函数在观测值 s' 下可执行的动作。从式(7)可以看出，要实现动作状态值函数的逼近，即

$$Q_{k+1}(s_t,a_t)\approx Q_k(s_t,a_t) \tag{8}$$

则 $r_t+\gamma\max\limits_{a'\in\wedge}Q_k(s',a')-Q_k(s_t,a_t)\to 0$。对于每一次迭代 k，可以通过最小化如式(9)所示的目标函数来实现参数更新。

$$\min E=\min(r_t+\gamma\max_{a'\in\wedge}Q_k(s',a')-Q_k(s_t,a_t)) \tag{9}$$

因此，本文将式（9）作为误差函数，通过求解误差梯度，即采用梯度下降法来更新 DNN 中的参数，求得动作状态值函数的最优解。

3.2 算法流程

对于系统模型中给出的多目标优化问题，基于深度强化学习的资源分配算法主要分成 2 个过程来求解，分别是前向传输过程和反向训练过程。在前向传输过程中，本文以传输速率最大化为优化目标，利用式(4)和式(6)构造 DNN；在反向训练过程中，将能量效率作为奖惩值，利用式(9)来反向训练 DNN。

3.2.1 前向传输过程

构造 DNN 是前向传输过程的核心，主要分成以下 7 个步骤。

1) 考虑到每个微基站在所有信道上的发射功率之和不能超过其最大发射功率，依据式(4)和式(6)，系统传输速率最优化问题表示为

$$\min_{\{D_{m,n}^k,p_{m,n}^k\}} -R \tag{10}$$

约束条件为

$$P_m^{\max}-\sum_{n=1}^{N}\sum_{k=1}^{K}L_{m,n}D_{m,n}^k p_{m,n}^k\geqslant 0,\quad \forall m\in\{1,2,\cdots,M\} \tag{11}$$

2) 采用增广拉格朗日乘子法将约束优化问题转化为无约束优化问题，构造的增广拉格朗日函数表示为

$$\varphi(D_{m,n}^k,p_{m,n}^k,\mu_m,\eta)=-R+\frac{1}{2\eta}\sum_{m=1}^{M}\{[\max(0,\mu_m-\eta(P_m^{\max}-\sum_{n=1}^{N}\sum_{k=1}^{K}L_{m,n}D_{m,n}^k p_{m,n}^k))]^2-\mu_m^2\} \tag{12}$$

其中，$\mu=\{\mu_{\mathrm{m}},\forall m\in\{1,2,\cdots,M\}\}$ 为拉格朗日乘子，η 为惩罚因子，从而把求解约束优化问题转化为求解无约束优化问题，即

$$\min\varphi(D_{m,n}^k,p_{m,n}^k,\mu_m,\eta) \tag{13}$$

3) 构造的增广拉格朗日函数分别对 $D_{m,n}^k$ 和 $p_{m,n}^k$ 求偏导。

$$\begin{cases} \nabla_{D_{m,n}^k}\varphi(D_{m,n}^k,p_{m,n}^k,\mu_m,\eta)=0 \\ \nabla_{p_{m,n}^k}\varphi(D_{m,n}^k,p_{m,n}^k,\mu_m,\eta)=0 \end{cases} \tag{14}$$

求得的 $D_{m,n}^k$ 和 $p_{m,n}^k$ 迭代方程式如式（15）所示。

$$\begin{cases} D_{m,n}^{k(l+1)}=\dfrac{\sqrt{[\eta(\sigma_{m,n}^2+I_{m,n}^k)+\left|h_{m,n}^k\right|^2\xi_{m,n}^{k(l)}]^2-4\eta\left|h_{m,n}^k\right|^2[(\sigma_{m,n}^2+I_{m,n}^k)\xi_{m,n}^{k(l)}-\dfrac{\left|h_{m,n}^k\right|^2}{\ln 2}]}-\eta(\sigma_{m,n}^2+I_{m,n}^k)-\left|h_{m,n}^k\right|^2\xi_{m,n}^{k(l)}}{2\eta L_{m,n}p_{m,n}^{k(l)}\left|h_{m,n}^k\right|^2} \\ p_{m,n}^{k(l+1)}=\dfrac{\sqrt{[\eta(\sigma_{m,n}^2+I_{m,n}^k)+\left|h_{m,n}^k\right|^2\xi_{m,n}^{k(l)}]^2-4\eta\left|h_{m,n}^k\right|^2[(\sigma_{m,n}^2+I_{m,n}^k)\xi_{m,n}^{k(l)}-\dfrac{\left|h_{m,n}^k\right|^2}{\ln 2}]}-\eta(\sigma_{m,n}^2+I_{m,n}^k)-\left|h_{m,n}^k\right|^2\xi_{m,n}^{k(l)}}{2\eta L_{m,n}D_{m,n}^{k(l)}\left|h_{m,n}^k\right|^2} \end{cases} \tag{15}$$

其中，$\xi_{m,n}^{k(l)}=\mu_m^{(l)}-\eta P_m^{\max}+\eta(\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{K}L_{m,i}D_{m,i}^{j(l)}p_{m,i}^{j(l)}-L_{m,n}D_{m,n}^{k(l)}p_{m,n}^{k(l)})$。

此外，拉格朗日乘子迭代方程式为

$$\mu_m^{(l+1)}=\max(0,\mu_m^{(l)}-\eta(P_m^{\max}-\sum_{n=1}^{N}\sum_{k=1}^{K}L_{m,n}D_{m,n}^{k(l)}p_{m,n}^{k(l)})) \tag{16}$$

4) 将移动用户与基站的接入关系 $L_{m,n}$ 和移动用户干扰信息 $I_{m,n}^k$ 作为输入，各基站内频率分配 $D_{m,n}^k$、功率分配 $p_{i,n}^k$ 和拉格朗日乘子 μ_m 根据式(15)和式(16)，依次迭代，形成如下数据流。

$$\{L_{m,n},I_{m,n}^k\}\to D_{m,n}^{k(1)}\to p_{m,n}^{k(1)}\to\mu_m^{(1)}\to D_{m,n}^{k(2)}\to p_{m,n}^{k(2)}\to\mu_m^{(2)}\to\cdots\to D_{m,n}^{k(l)}\to p_{m,n}^{k(l)}\to\mu_m^{(l)}\to\cdots\to D_{m,n}^{k(L)}\to p_{m,n}^{k(L)}$$

5) 根据迭代更新数据流来构造 DNN，如图 3 所示。DNN 包括输入层、频率分配层、功率分配层、乘子层和输出层，深度取决于频率分配 $D_{m,n}^{k}$、功率分配 $p_{i,n}^{k}$ 和拉格朗日乘子 μ_m 的迭代更新次数。DNN 中频率分配层和功率分配层的权值参数为信道增益 $h_{m,n}^{k}$ 和噪声 $\sigma_{m,n}^{2}$；非线性转换函数分别为频率分配 $D_{m,n}^{k}$、功率分配 $p_{i,n}^{k}$ 和拉格朗日乘子 μ_m 的迭代更新方程式。

6) 初始化 DNN 的权值参数，即将信道增益 $h_{m,n}^{k}$ 初始化为瑞利分布，将噪声 $\sigma_{m,n}^{2}$ 初始化为高斯白噪声。

7) 在时刻 t，将观测到的蜂窝网用户接入信息 $L_{m,n}$ 和干扰信息 $I_{m,n}^{k}$ 作为 DNN 的输入，设定阈值 θ_D、θ_p 和最大迭代更新次数 Q_1，经过 DNN 的前向传输后，当 $\left|D_{m,n}^{k(l+1)}-D_{m,n}^{k(l)}\right|<\theta_D$ 且 $\left|p_{m,n}^{k(l+1)}-p_{m,n}^{k(l)}\right|<\theta_p$ 时，或迭代更新次数达到最大迭代更新次数 Q_1 时，在输出层输出一组数值，每一个数值对应一种频谱分配方案和功率分配方案，从输出的数值中寻找出最大数值，并将最大数值 $\max Q(L_{m,n},I_{m,n}^{k},D_{m,n}^{k},p_{m,n}^{k})$ 所对应的频率分配方案 $D_{m,n}^{k}$ 和功率分配方案 $p_{m,n}^{k}$ 作为时刻 t 的资源分配策略。

3.2.2 *反向训练过程*

构造误差函数来训练 DNN 是反向训练过程的核心，主要分成以下 5 个步骤。

1) 在执行频率分配方案 $D_{m,n}^{k}$ 和功率分配方案 $p_{m,n}^{k}$ 后，观测系统能量效率，将能量效率作为奖惩值，即

$$r_t=H=\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{B_{m,n}^{k}\mathrm{lb}\left(1+\dfrac{L_{m,n}D_{m,n}^{k}p_{m,n}^{k}\left|h_{m,n}^{k}\right|^{2}}{\sigma_{m,n}^{2}+I_{m,n}^{k}}\right)}{p_{m,n}^{k}} \tag{17}$$

2) 观测蜂窝网用户接入信息 $L_{m,n}^{(')}$ 和干扰信息 $I_{m,n}^{k(')}$，重新输入 DNN，经过 DNN 前向传输后，采取同样方法，从输出层输出的数值中寻找最大数值，将最大数值 $\max Q(L_{m,n}^{(')},I_{m,n}^{k(')},D_{m,n}^{k(')},p_{m,n}^{k(')})$ 所对应的频谱分配方案 $D_{m,n}^{k(')}$ 和功率分配方案 $p_{m,n}^{k(')}$ 作为时刻 $t+1$ 的资源分配策略。需要注意的是，在资源分配策略形成过程中用户接入信息被认为是固定不变的信息，即时刻 $t+1$ 观测到的用户接入信息 $L_{m,n}^{(')}$ 与时刻 t 观测到的用户接入信息 $L_{m,n}$ 相同。

3) 依据式(9)，构建如式(18)所示的误差函数。

$$E=r_t+\gamma\max Q_k(L_{m,n}^{(')},I_{m,n}^{k(')},D_{m,n}^{k(')},p_{m,n}^{k(')})-Q_k(L_{m,n},I_{m,n}^{k},D_{m,n}^{k},p_{m,n}^{k}) \tag{18}$$

其中，折扣因子 $\gamma\in[0,1]$ 决定了资源分配策略偏重程度，若采用反向传播算法使用损失函数 E 趋于最小化，当 $\gamma\to 0$，神经网络当前时刻输出的动作状态值函数 $Q_k(L_{m,n},I_{m,n}^{k},D_{m,n}^{k},p_{m,n}^{k})$ 趋近于奖惩值 r_t，即资源分配策略偏重于优化系统能量效率；当 $\gamma\to 1$，奖惩值 r_t 和神经网络下一时刻输出的动作状态值函数 $\max Q_k(L_{m,n}^{(')},I_{m,n}^{k(')},D_{m,n}^{k(')},p_{m,n}^{k(')})$ 占有同样的比重，此时资源分配策略综合优化系统能量效率和传输速率。

4) 设定阈值 θ_E，将损失函数值 E 与阈值 θ_E 进行比较。若损失函数值 $E\geqslant\theta_E$，则执行 5)，否则，将选定的频谱分配方案 $D_{m,n}^{k}$ 和功率分配方案 $p_{m,n}^{k}$ 作为最优资源管理策略，完成蜂窝网资源分配。

5) 采用反向传播算法，使损失函数值 E 趋于最小化，沿着损失函数梯度下降方向逐层修正信道增益 $h_{m,n}^{k}$ 和噪声 $\sigma_{m,n}^{2}$，若 DNN 的权值参数更新次数达到限定的最大次数 Q_2，则将获得的频谱分配方案 $D_{m,n}^{k}$ 和功率分配方案 $p_{m,n}^{k}$ 作为最优资源分配策略，完成蜂窝网资源分配，否则，修正好 DNN 的权值后，继续执行 DNN 的前向传输操作。

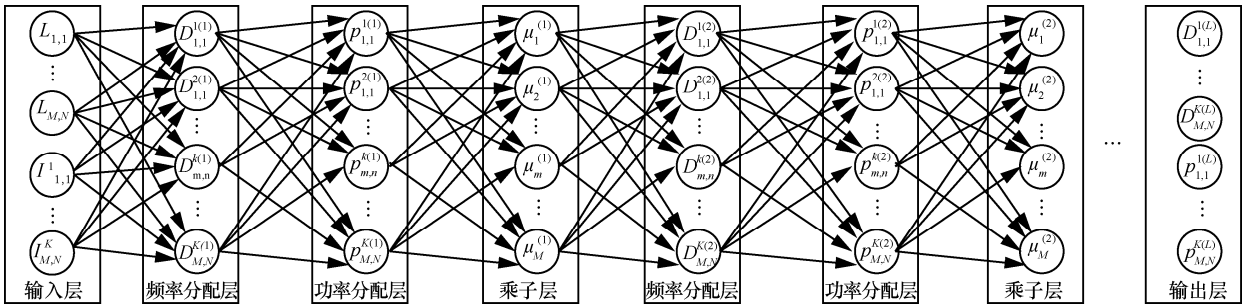


图 3　DNN 的基本架构

以图 3 中神经元 $p_{m,n}^{k(l)}$ 为例，该神经元包含 2 个输入，即 $\mu_m^{(l-1)}$ 和 $D_{m,n}^{k(l)}$，该神经元的输出作为神经元 $\mu_m^{(l)}$ 和 $D_{m,n}^{k(l+1)}$ 的输入。对于神经元 $p_{m,n}^{k(l)}$，误差函数关于权值修正的梯度为

$$\begin{cases}\dfrac{\partial E}{\partial h_{m,n}^{k(l)}}=\dfrac{\partial E}{\partial p_{m,n}^{k(l)}}\dfrac{\partial p_{m,n}^{k(l)}}{\partial h_{m,n}^{k(l)}}\\ \dfrac{\partial E}{\partial \sigma_{m,n}^{2(l)}}=\dfrac{\partial E}{\partial p_{m,n}^{k(l)}}\dfrac{\partial p_{m,n}^{k(l)}}{\partial \sigma_{m,n}^{2(l)}}\end{cases} \tag{19}$$

其中，$\dfrac{\partial p_{m,n}^{k(l)}}{\partial h_{m,n}^{k(l)}}$ 通过式(15)求得，$\dfrac{\partial E}{\partial p_{m,n}^{k(l)}}$ 通过式(20)计算。

$$\frac{\partial E}{\partial p_{m,n}^{k(l)}}=\frac{\partial E}{\partial \mu_m^{(l)}}\frac{\partial \mu_m^{(l)}}{\partial p_{m,n}^{k(l)}}+\frac{\partial E}{\partial D_{m,n}^{k(l+1)}}\frac{\partial D_{m,n}^{k(l+1)}}{\partial p_{m,n}^{k(l)}} \tag{20}$$

其中，$\dfrac{\partial E}{\partial \mu_m^{(l)}}$ 和 $\dfrac{\partial E}{\partial D_{m,,n}^{k(l+1)}}$ 通过式(17)和式(18)求得，$\dfrac{\partial D_{m,n}^{k(l)}}{\partial p_{m,n}^{k(l)}}$ 和 $\dfrac{\partial \mu_m^{(l)}}{\partial p_{m,n}^{k(l)}}$ 通过式(15)和式(16)求得。

求得误差函数关于权值修正的梯度后，利用式(21)更新 DNN 的权值 $h_{m,n}^{k(l)}$ 和 $\sigma_{m,n}^{2(l)}$。

$$\begin{cases}h_{m,n}^{k(l)}\leftarrow h_{m,n}^{k(l)}-\lambda\dfrac{\partial E}{\partial h_{m,n}^{k(l)}}\\ \sigma_{m,n}^{2(l)}\leftarrow \sigma_{m,n}^{2(l)}-\lambda\dfrac{\partial E}{\partial \sigma_{m,n}^{2(l)}}\end{cases} \tag{21}$$

其中，λ 为学习速率。

4 仿真与分析

本文分别仿真分析了折扣因子对蜂窝网资源分配策略、基于深度强化学习的资源分配算法的收敛性和性能的影响，采用蒙特卡洛方法重复执行 1 000 次，然后对结果取平均值。在每一次算法执行过程中，蜂窝用户均随机分布在系统中，仿真参数如表 1 所示。

表 1　　仿真参数

参数	取值
微基站/个	3
小区半径/m	200
微基站的最大发射功率/dBm	38
载波频率/GHz	2.0
子信道带宽/kHz	180
可用信道数/个	0~8
移动用户数/个	10

首先，分析折扣因子对资源分配策略的影响。将可用子载波数设为 4，图 4 仿真了折扣因子在[0,1]内的取值情况，显示了折扣因子对蜂窝网资源分配策略的影响情况，当折扣因子取值为 0 时，资源分配策略偏重于奖惩值，即偏重优化能量效率，此时获得的能量效率最高，传输速率最低。随着折扣因子增大，误差函数 E 中，动作状态值函数占有比重越来越大，资源分配策略所获得的传输速率越来越高，能量效率越来越低。当折扣因子取值为 1 时，系统获得的传输速率最高，能量效率最低。因此，在仿真过程中，可以根据资源分配策略的偏重程度来合理设置折扣因子。

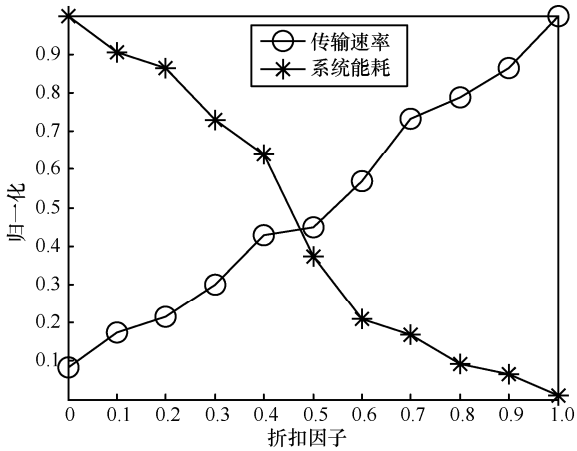


图 4　折扣因子对资源分配策略的影响

其次，分析算法收敛性。将可用子载波数设为 4，算法运算速度取决于 DNN 深度和反向训练 DNN 的次数。设定阈值 $\theta_D=\theta_p=0.01$，图 5 显示了 DNN 的深度。当 DNN 的深度为 6 时，差值 $\left|D_{m,n}^{k(6)}-D_{m,n}^{k(5)}\right|<0.01$ 且 $\left|p_{m,n}^{k(6)}-p_{m,n}^{k(5)}\right|<0.01$，DNN 输出频率分配方案和功率分配方案。设定阈值

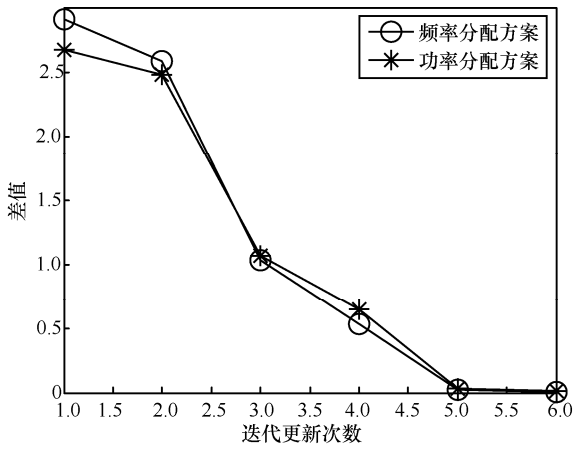


图 5　DNN 的深度

$\theta_E = 0.001$，图 6 显示了反向训练 DNN 的次数。当反向训练次数达 5 次时，$E < 0.001$，反向训练过程结束，输出最优的频率分配方案和功率分配方案。

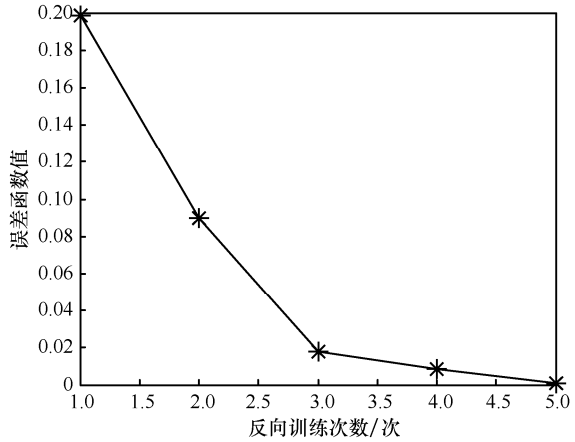


图 6　DNN 的反向训练次数

最后，分析算法性能。通过改变子信道数，将本文提出的算法分别从传输速率和能量效率两方面与随机分配算法、贪婪算法进行比较。图 7 和图 8 分别给出了传输速率和能量效率比较结果。可以看出，当折扣因子为 1 时，本文提出算法得到的资源分配策略偏重于优化传输速率，系统获得的传输速率接近于贪婪算法，但是获得的能量效率高于贪婪算法；虽然获得的能量效率低于随机分配算法，但是传输速率高于随机分配算法。当折扣因子为 0 时，本文提出算法得到的资源分配策略偏重于优化能量效率，即奖惩值，虽然系统获得的传输速率相对较低，但是系统获得的能量效率高于贪婪算法和随机分配算法。

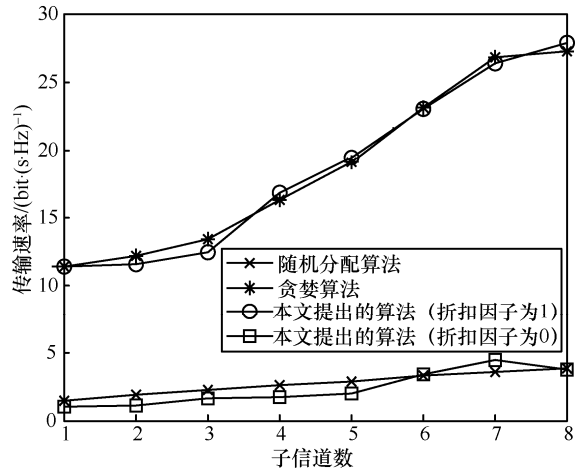


图 7　传输速率

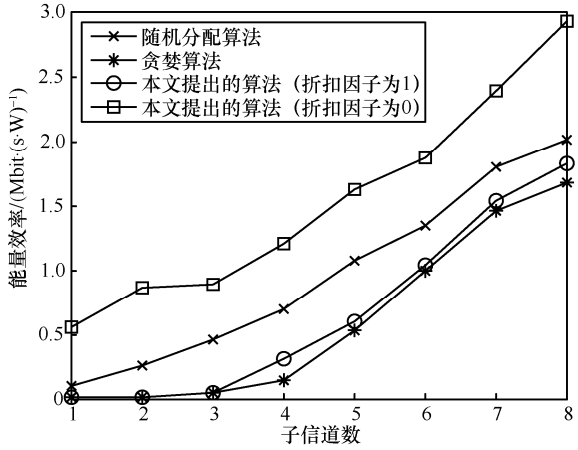


图 8　能量效率

5　结束语

为了提高蜂窝网传输速率的同时，尽可能地增大能量效率，本文讨论了蜂窝网中的资源分配问题，提出了一种基于深度强化学习的蜂窝网资源分配算法，该算法包括前向传输和反向训练 2 个过程。在前向传输过程中，主要构建 DNN，以最优化传输速率；在反向训练过程中，将能量效率作为奖惩值，采用 Q-learning 机制来构建误差函数，反向训练 DNN 中的权值参数。仿真结果显示，本文提出的算法可以通过设置折扣因子，自主选择资源分配策略的偏重程度，收敛速度快，在传输速率和系统能耗优化方面都明显优于其他算法，有效地解决了蜂窝网资源分配多目标优化问题。

参考文献：

[1] HUANG J, YIN Y, ZHAO Y, et al. A game-theoretic resource allocation approach for intercell device-to-device communications in cellular networks[J]. IEEE Transactions on Emerging Topics in Computing, 2016, 4(4): 475-486.

[2] WANG J, CHOU S. Secure strategy proof ascending-price spectrum auction[C]//IEEE Symposium on Privacy-Aware Computing. 2017: 96-106.

[3] YANG T, ZHANG R, CHENG X, et al. Graph coloring based resource sharing scheme(GCRS) for D2D communications underlaying full-duplex cellular networks[J]. IEEE Transactions on Vehicular Technology, 2017, 66(8): 7506-7517.

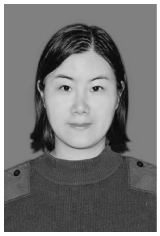
[4] TAKSHI H, DOĞAN G, ARSLAN H. Joint optimization of device to device resource and power allocation based on genetic algorithm[J]. IEEE Access, 2018, 6: 21173-21183.

[5] CHALLITA U, DONG L, SAAD W. Proactive resource management for ITE in unlicensed spectrum: a deep learning perspective[J]. IEEE Transactions on Wireless Communications, 2018, 17(7): 4674-4689.

[6] LEE W. Resource allocation for multi-channel underlay cognitive radio network based on deep neural network[J]. IEEE Communications Letters, 2018, 22(9): 1942-1945.

[7] 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.
[8] 赵慧, 张学, 刘明, 等. 实现无线传输能量效率最大化的功率控制新方法[J]. 计算机应用, 2013, 33(2): 365-368.
ZHAO H, ZHANG X, LIU M, et al. New power control scheme with maximum energy efficiency in wireless transmission[J]. Journal of Computer Application, 2013, 33(2):365-368.
[9] GAO X Z, HAN H C, YANG K, et al. Energy efficiency optimization for D2D communications based on SCA and GP method[J]. China Communications, 2017, 14(3): 66-74.
[10] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Massachusetts: MIT Press, 2017.
[11] 焦李成, 杨进, 杨淑媛, 等. 深度学习、优化与识别[M]. 北京：清华大学出版社, 2017.
JIAO L C, ZHAO J, YANG S Y, et al. Deep learning, optimization and recognition [M]. Beijing: Tsinghua University Press, 2017.

[作者简介]



廖晓闽（1984− ），女，江西德兴人，西安电子科技大学博士生，国防科技大学信息通信学院副教授，主要研究方向为频谱管控、隐蔽通信。

严少虎（1976− ），男，四川绵竹人，博士，中国电子科技集团公司第二十九研究所高级工程师，主要研究方向为频谱管控、体系集成。

石嘉（1987− ），男，陕西西安人，博士，西安电子科技大学副教授，主要研究方向为无线系统资源分配、毫米波通信、隐蔽通信等。

谭震宇（1987− ），男，广西玉林人，西安电子科技大学博士生，主要研究方向为无线频谱管理。

赵钟灵（1995− ），男，河北张家口人，西安电子科技大学博士生，主要研究方向为频谱资源管理。

李赞（1975− ），女，陕西西安人，西安电子科技大学教授、博士生导师，主要研究方向为隐蔽通信、频谱管控。