文章编号:1007-5321(2022)06-0122-05 **DOI**:10.13190/j.jbupt.2022-039

交替方向乘子法与深度强化学习的资源分配

郭兴康, 孙 君
(南京邮电大学 通信与信息工程学院, 南京 210003)

摘要: 针对有限信道状态信息下密集型网络资源分配的问题,提出将交替方向乘子法与深度强化学习算法相结合的模型驱动学习框架。区别于数据驱动框架,利用所提框架能够根据具体问题进行一对一建模。建模内容包括基站选择、功率和子载波分配,并用交替方向乘子法进行交替优化;用深度强化学习算法优化权重,求解目标函数,提高了算法的性能;利用有效信道状态信息而非多余信息,降低了通信开销;加强对最低用户服务质量的要求,在保证用户具有良好体验的情况下使小区的频谱效率最大化。仿真结果表明,在较少的迭代次数下,利用所提框架可使小区用户的频谱效率收敛,达到最大值。

关 键 词: 密集型网络; 模型驱动; 资源分配; 深度强化学习; 交替方向乘子法
中图分类号: TN929.5 **文献标志码**: A

Resource Allocation Based on Alternating Direction Multiplier Method and Deep Reinforcement Learning Algorithm

GUO Xingkang, SUN Jun
(School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to optimize the resource allocation of dense network under limited channel state information, a model-driven learning framework is proposed by combining with alternating direction multiplier method and deep reinforcement learning algorithm. This framework differs from data-driven ones, and it enables one-to-one modeling for specific problems. Modeling includes four aspects: (i) alternately optimizing base station selection, power, and subcarrier allocation by using alternating direction multiplier method; (ii) using deep reinforcement learning algorithm to optimize weights, solve target functions, and improve performance of the system; (iii) using effective channel state information instead of redundant information to reduce overhead on communication; (iv) adding constraints on users' quality of service requirements to maximize cell spectral efficiency while ensuring that users have a good experience. The simulation results show that the proposed framework can make the spectrum efficiency of cell users converge and reach the maximum value with a small number of iterations.

Key words: dense network; model-driven; resource allocation; deep reinforcement learning; alternating direction multiplier method

随着移动设备数量的增长和物联网的出现,密集型网络应运而生,Zhao 等[1]利用具有不同传输功率和覆盖范围的微基站、毫微基站对现有的蜂窝网络进行了优化。一般而言,资源分配需要全局信道

收稿日期: 2022-03-09
作者简介: 郭兴康(1998—), 男, 硕士生。
通信作者: 孙 君(1980—), 女, 副研究员, 邮箱: sunjun@njupt.edu.cn。

状态信息和用户的信干扰噪声比。但是全局信道状态信息往往会超出通信系统的链路容量。为了解决通信系统中链路容量有限的问题，大多数研究者使用局部有限信道状态信息发射器进行信息交换，并将加权的信干扰噪声比最大化。在多输入多输出通道[2-3]中，最大化方法即实现纳什均衡；Bistritz 等[4]考虑了一种次优方法，在频率选择信道下，仅需根据每个用户一定数量的最佳信道信息，即可将最优资源分配给用户，从而减少各信道的反馈信息。在服务质量的约束下，Bonsu 等[5]采用拉格朗日乘子法进行功率分配来优化能量效率。Dinh 等[6]采用增广拉格朗日算法使无线接入网络的能量效率最大化。此外，还可使用博弈论方法[7]、线性规划方法和马尔可夫逼近策略的资源分配方案。对于上述方案，若使用常规算法设计发射功率或资源分配方案，会降低系统的吞吐量且不能满足用户最低服务质量的要求。目前大多学者利用大量数据训练的数据驱动学习模型[8]，而数据驱动的深度学习方法使神经网络拓扑缺乏理论上的解释。因此，Xu 等[9]提出了模型驱动的深度学习方法，使神经网络可预测。Liao 等[10]在密集型网络中利用有限信道状态信息并使用模型驱动的深度学习方法有效地分配不同类型的无线资源，但没有考虑用户的最低服务质量要求。

笔者主要研究了在有效的信道状态信息下联合优化密集型网络中基站选择、资源分配等问题，并提出基于模型驱动的深度强化学习框架，根据有效的信道状态信息，考虑用户的最低服务质量要求并进行资源分配。使用该框架可以减轻时间和样本消耗，并降低算法的时间复杂度。

1　系统模型与问题建模

笔者所提模型的应用场景为下行链路，小区具有 3 种基站的密集型异构蜂窝网络。图 1 所示为密集型网络模型。每个基站的传输功率和覆盖半径不同，3 种基站分别为宏基站、微基站和毫微基站。假设小区内共有 N 个互相正交的子载波、M 个基站和 K 个用户，$b_{m,k}$ 为基站 m 选择用户 k，c_k^n 为用户 k 选择子载波 n。定义基站、用户和子载波之间的相互选择关系为 $b_{m,k}c_k^n$，若其值为 1，表示三者之间相关联；若值为 0，则不关联。

每种基站的发射功率 p 分为 2 个等级，即 $\{0, p_{\max}/N\}$，$p_{\max}$ 为基站的最大功率，不同基站的最大功率也不同。不同用户在不同基站下可能会分配

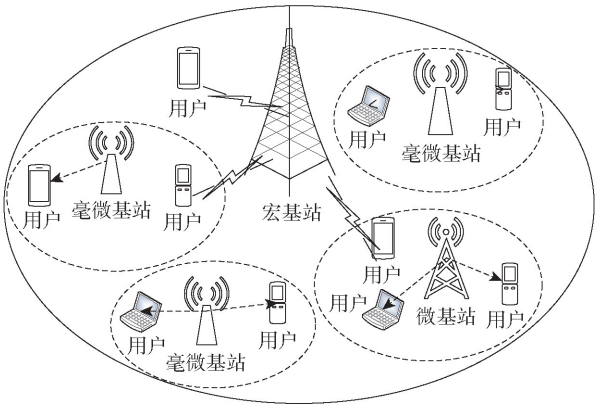

图 1　密集型网络模型

相同的子载波，用户 k 在基站 m 和子载波 n 下受到的干扰为

$$I_{m,k}^n = \sum_{m' \neq m}^{M} \sum_{k'=1}^{K} b_{m',k'} c_{k'}^n p g_{m',k}^n \tag{1}$$

其中：m' 为不同于基站 m 的其余基站，k' 为不同于用户 k 的其余用户，$g_{m,k}^n$ 为在子载波 n 上基站 m 到用户 k 的信道增益。信号干扰噪声比为

$$S_{m,k}^n = \frac{b_{m,k} c_k^n p g_{m,k}^n}{\sigma_{m,k} + I_{m,k}^n} \tag{2}$$

其中：$\sigma_{m,k}$ 为加性高斯白噪声，独立同分布并且服从瑞利衰落。

一般情况下，最大化频谱效率为

$$\max_{\{b_{m,k}, c_k^n\}} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{lb}(1 + S_{m,k}^n)$$

约束条件中用户的最低服务质量要求为 $\sum_{m=1}^{M} \sum_{n=1}^{N} S_{m,k}^n \geqslant \Omega_k, \forall k \in K$，$\Omega_k$ 为用户 k 的最低服务质量要求。

2　问题求解与算法实现

为了利用学习算法寻找最优解，设计了基于交替方向乘子优化算法的模型驱动深度强化学习框架。

2.1　交替方向乘子法

利用交替方向乘子法更改最大化频谱效率和约束条件，对应的增广拉格朗日函数为

$$F(b_{m,k}, c_k^n, \alpha_k, \mu) = -\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{k=1}^{K} \mathrm{lb}\left(1 + \frac{S_{m,k}^n}{p}\right) + \frac{1}{2\mu} \sum_{k=1}^{K} \left\{ \left[\max\left(0, \alpha_k - \mu\left(\sum_{m=1}^{M} \frac{S_{m,k}^n}{p} - \Omega_k\right)\right)\right]^2 - \alpha_k^2 \right\} \tag{3}$$

其中:α_k 为拉格朗日乘子,μ 为惩罚因子。若使频谱效率最大化,用式(3)分别对 $b_{m,k}$ 和 c_k^n 求偏导,即

$$\left.\begin{aligned}\frac{\partial F(b_{m,k},c_k^n,\alpha_k,\mu)}{\partial b_{m,k}}=0\\ \frac{\partial F(b_{m,k},c_k^n,\alpha_k,\mu)}{\partial c_k^n}=0\end{aligned}\right\} \tag{4}$$

为了简化计算,定义4个中间变量,即

$$\rho_{m,k}^n=\frac{\sigma_{m,k}}{p}+\sum_{m'\neq m}^{M}\sum_{k'=1}^{K}b_{m',k'}c_{k'}^n g_{m',k}^n \tag{5}$$

$$\omega_{m,k}^n=\sum_{m'=1}^{M}\sum_{n'=1}^{N}\frac{b_{m',k}c_k^{n'}g_{m',k}^{n'}}{\rho_{m',k}^{n'}}-\frac{b_{m,k}c_k^n g_{m,k}^n}{\rho_{m,k}^n} \tag{6}$$

$$\psi_{m,k}^n=\mu\omega_{m,k}^n-\mu\Omega_k-\alpha_k \tag{7}$$

$$v_{m,k}^n=\mu+\psi_{m,k}^n \tag{8}$$

最终得到2个动作策略,即基站如何选择用户和用户如何选择子载波:

$$b_{m,k}=\frac{\rho_{m,k}^n\sqrt{(v_{m,k}^n)^2-4\mu\left(\psi_{m,k}^n-\frac{1}{\ln 2}\right)}-\rho_{m,k}^n v_{m,k}^n}{2\mu g_{m,k}^n c_k^n} \tag{9}$$

$$c_k^n=\frac{\rho_{m,k}^n\sqrt{(v_{m,k}^n)^2-4\mu\left(\psi_{m,k}^n-\frac{1}{\ln 2}\right)}-\rho_{m,k}^n v_{m,k}^n}{2\mu g_{m,k}^n b_{m,k}} \tag{10}$$

2.2 深度强化学习算法

深度强化学习算法包含状态集、动作集、奖励函数和策略4个要素。

1)状态集。包含状态$\{s_0,s_1,\cdots,s_t\}$,在此模型中,状态就是用户的干扰 $I_{m,k}^n$,此状态会随着动作 $b_{m,k}$ 和 c_k^n 的改变而发生改变。

2)动作集。根据现有状态和策略执行动作,动作为选择合适的基站 $b_{m,k}$ 和子载波 c_k^n。

3)奖励函数。深度强化学习算法可使长期奖励值最大化,而且每个用户都应该满足最低服务质量的要求。若满足要求,则对应的奖励值为+1;否则为-1。

4)策略。策略是深度强化学习算法的最终目的,可以根据当前状态执行下一个动作。

基于交替方向乘子法的深度神经网络包括输入层、隐藏层和输出层,其结构如图2所示。

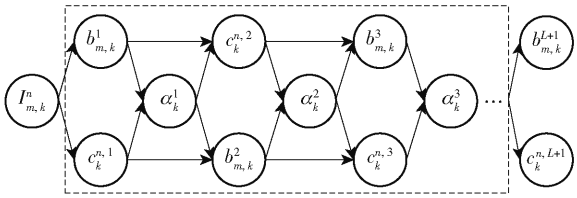

图2 基于交替方向乘子法的深度神经网络结构

输入层:第1层不需要经过任何计算,只是将初始状态传递给下一层。

隐藏层:除第1层和最后一层,第 l 层包括动作值 $b_{m,k}^l$,$c_k^{n,l}$ 和拉格朗日乘子 α_k^l,隐藏层包含 L 次迭代且单节点 α_k^{l+1} 没有权重,输入为上一层的动作 $b_{m,k}^l$,$c_k^{n,l}$,输出为 $b_{m,k}^{l+1}$,$c_k^{n,l+1}$,拉格朗日乘子节点的非线性变换定义为

$$\varphi_{\alpha_m}^{l+1}=\max\left\{0,\alpha_k^{l+1}-\mu\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{b_{m,k}c_k^n g_{m,k}^n}{\rho_{m,k}^n}-\Omega_k\right)\right\} \tag{11}$$

2个动作值节点的非线性变换由式(9)和式(10)得到。

输出层:最后一层的输出结果就是最终的资源分配策略,当且仅当$|b_{m,k}^{L+1}-b_{m,k}^L|\to 0$ 且 $|c_k^{n,L+1}-c_k^{n,L}|\to 0$ 或者迭代次数达到最大值才会输出结果。

算法1为模型驱动框架的学习阶段。只有当动作值 $b_{m,k}^l$ 和 $c_k^{n,l}$ 都收敛时,该学习阶段才结束。

算法1 模型驱动框架的学习阶段

1 更新现有的环境初始状态 $I_{m,k}^n$。
2 初始化权重 $\theta=\{g_{m,k}^{n,l},\sigma_{m,k}^l\}_{l=1}^{L+1}$,初始动作值 $b_{m,k}^0$,$c_k^{n,0}$ 和拉格朗日乘子 α_k^0。
3 设置隐藏层动作收敛阈值 ξ_b,ξ_c 和最大迭代次数 L,最低服务质量要求 Ω_k 和惩罚参数 μ。
4 for $l=(0,1,\cdots,L)$ do
5 根据深度神经网络结构计算 $b_{m,k}^{l+1}$,$c_k^{n,l+1}$ 和 α_k^{l+1}。
6 if $|b_{m,k}^{l+1}-b_{m,k}^l|<\xi_b$ and $|c_k^{n,l+1}-c_k^{n,l}|<\xi_c$ then
7 输出最终结果$\{b_{m,k}^{l+1},c_k^{n,l+1}\}$。
8 end if
9 end for

除了算法1的学习阶段,所提框架还包括算法2的训练阶段。学习阶段是前向传播,训练阶段则是梯度下降反向传播。

根据Q学习算法,Q 值的更新公式为

$$Q_{k+1}(s_t,a_t)=Q_k(s_t,a_t)+\beta[r_{t+1}+\gamma\max_{a'\in A}Q_k(s_{t+1},a')-Q_k(s_t,a_t)] \tag{12}$$

其中:s_t 为状态,a_t 为动作,β 为学习率,γ 为折扣因

子，r_{t+1}为奖励值，$\max_{a'\in A}Q_k(s_{t+1},a')$为下一个状态所有动作的最大 Q 值。若 γ 值为 0，则利用学习算法将短期效益最大化，从长远角度着，总体效益不是最大的。

在式(12)中，当 $r_{t+1}+\gamma\max_{a'\in A}Q_k(s_{t+1},a')-Q_k(s_t,a_t)\to 0$时，$Q_k(s_t,a_t)$接近 $Q_{k+1}(s_t,a_t)$，因此损失函数为

$$E=\frac{1}{2}\left[r_{t+1}+\gamma\max_{a'\in A}Q_k(s_{t+1},a')-Q_k(s_t,a_t)\right]^2 \tag{13}$$

根据损失函数分别对 $g_{m,k}^{n,l}$，$\sigma_{m,k}^{l}$求偏导：

$$\left.\begin{aligned}\frac{\partial E}{\partial g_{m,k}^{n,l}}&=\frac{\partial E}{\partial b_{m,k}^{l}}\frac{\partial b_{m,k}^{l}}{\partial g_{m,k}^{n,l}}+\frac{\partial E}{\partial c_k^{n,l}}\frac{\partial c_k^{n,l}}{\partial g_{m,k}^{n,l}}\\ \frac{\partial E}{\partial \sigma_{m,k}^{l}}&=\frac{\partial E}{\partial b_{m,k}^{l}}\frac{\partial b_{m,k}^{l}}{\partial \sigma_{m,k}^{l}}+\frac{\partial E}{\partial c_k^{n,l}}\frac{\partial c_k^{n,l}}{\partial \sigma_{m,k}^{l}}\end{aligned}\right\} \tag{14}$$

为了最小化损失函数，使用梯度下降法更新神经网络的权重，即

$$\left.\begin{aligned}g_{m,k}^{n,l}&=g_{m,k}^{n,l}-\lambda_g\frac{\partial E}{\partial g_{m,k}^{n,l}}\\ \sigma_{m,k}^{l}&=\sigma_{m,k}^{l}-\lambda_\sigma\frac{\partial E}{\partial \sigma_{m,k}^{l}}\end{aligned}\right\} \tag{15}$$

其中：λ_g 为增益学习率，λ_σ 为噪音学习率。

算法 2 中需要用到算法 1 进行多次迭代输出的资源分配结果。当且仅当损失函数值小于阈值时，训练阶段才算完成。

算法 2　模型驱动框架训练阶段

1　设置折扣因子 γ、损失函数阈值 ξ_E 和最大轮询次数 T。
2　for $t=(0,1,\cdots,T)$ do
3　　计算算法 1 输出的动作值$\{b_{m,k},c_k^n\}$。
4　　更新下一状态 $I_{m,k}^{n,t+1}$以及奖励值 r_{t+1}。
5　　计算损失函数 E。
6　　if $E\geqslant\xi_E$ then
7　　使用梯度下降法更新权重 θ 后执行算法 1。
8　　else
9　　输出最终的动作值$\{b_{m,k},c_k^n\}$。
10　　end if
11　end for

3　仿真分析

通过仿真实验验证了基于交替方向乘子法模型驱动框架的性能。假设小区的基站数为 3，用户数为 4，基站随机分布在基站信号覆盖的范围内，并以小速度在小区内随机移动；互相正交的子载波数量为 5；所有频带上每种基站的总功率限制分别为 38，36，35 dBm；小区的半径为 200 m；信道带宽 B 为 180 kHz；噪声功率谱密度 D_0 为 −174 dBm/Hz；μ 为 2；γ 为 0.6；收敛阈值 ξ_b,ξ_c,ξ_E 分别为 0.01，0.01，0.001。

图 3 所示为基于交替方向乘子算法模型驱动框架的性能。当$|b_{m,k}^{L+1}-b_{m,k}^{L}|<\xi_b$ 且$|c_k^{n,L+1}-c_k^{n,L}|<\xi_c$时，该模型驱动框架会输出最终的资源分配策略。设小区 1～4 分别有 3，6，9，50 个基站和 4，8，12，80 个用户以及 5，10，15，100 个子载波。由图 3 可知，该模型驱动框架动作值收敛的速度非常快，小区 1 收敛时的迭代次数小于小区 2 和小区 3 收敛时的迭代次数，而最大的小区 4 也可在几十次迭代内收敛。

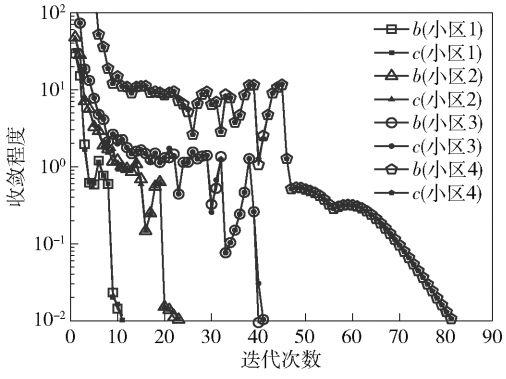

图 3　基于交替方向乘子算法模型驱动框架的性能

图 4 所示为不同学习率对算法性能的影响。在不同的学习率下模型驱动框架的收敛性能不同。利用所提框架采用交替方向乘子法优化并使用梯度下降方法训练神经网络，从而使损失函数在适合的学习率下收敛更快。由图 4 可见，轮询次数大于 30 时损失函数基本趋于平稳。

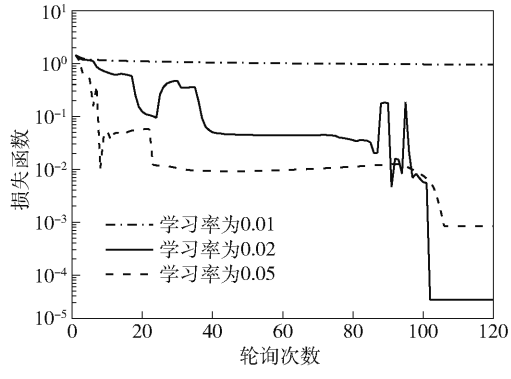

图 4　不同学习率对算法性能的影响

图 5 所示为不同用户数量对小区频谱效率的影

响。将所提框架与其他文献中多智能体的深度强化学习算法、随机分配算法和Q学习算法进行了对比。由图5可知,模型驱动框架的总体性能优于其他算法,且该小区的频谱效率会随着用户数量的增加而增加。

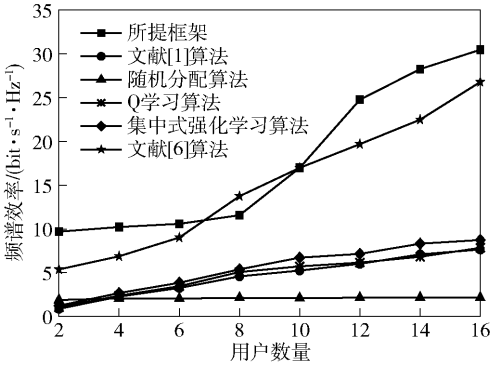

图5 不同用户数量对小区频谱效率的影响

4 结束语

笔者提出了一个改进的模型驱动深度学习框架,利用功率的离散化简化了目标函数并增加了用户服务质量要求的约束,可用于密集型网络中的基站选择、子载波分配和功率控制。在不超过最大功率的前提下,确保每个用户都能达到最低服务质量的要求。此外,使用了有效的状态信道信息,产生的通信开销很小。仿真结果表明,所提框架具有快速收敛的能力,并且优于其他现有资源分配算法。

参考文献:

[1] ZHAO N, LIANG Y C, NIVATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.

[2] BJORNSON E, ZAKHOUR R, GESBERT D, et al. Cooperative multicell precoding: rate region characterization and distributed strategies with instantaneous and statistical CSI[J]. IEEE Transactions on Signal Processing, 2010, 58(8): 4298-4310.

[3] KIM Y, YANG H J. Sum-rate maximization of multicell MISO networks with limited information exchange[J]. IEEE Transactions on Vehicular Technology, 2020, 69(7): 7247-7263.

[4] BISTRITZ I, LESHEM A. Asymptotically optimal resource block allocation with limited feedback[J]. IEEE Transactions on Wireless Communications, 2018, 18(1): 34-46.

[5] BONSU K A, ZHOU WW, PAN S, et al. Optimal power allocation with limited feedback of channel state information in multi-user MIMO systems[J]. China Communications, 2020, 17(2): 163-175.

[6] DINH T H L, KANEKO M, FUKUDA E H, et al. Energy efficient resource allocation optimization in fog radio access networks with outdated channel knowledge[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(1): 146-159.

[7] BAYAT S, LOUIE R H Y, HAN Z, et al. Distributed user association and femtocell allocation in heterogeneous wireless networks[J]. IEEE Transactions on Communications, 2014, 62(8): 3027-3043.

[8] ZHOU Y B, FADLULLAH Z M, MAO B M, et al. A deep-learning-based radio resource assignment technique for 5G ultra dense networks[J]. IEEE Network, 2018, 32(6): 28-34.

[9] XU Z, SUN J. Model-driven deep-learning[J]. National Science Review, 2018, 5(1): 22-24.

[10] LIAO X, SHI J, LI Z, et al. A model-driven deep reinforcement learning heuristic algorithm for resource allocation in dense cellular networks[J]. IEEE Transactions on Vehicular Technology, 2019, 69(1): 983-997.