ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 6,Nov. 2025,pp. 1477－1489
DOI:10. 16337/j. 1004-9037. 2025. 06. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于指针网络深度强化学习NOMA用户配对和功率分配方案

李国鑫，甘　麒，陈　瑾，焦雨涛，王海超，贺　兴

（中国人民解放军陆军工程大学通信工程学院，南京 210007）

摘　要：为了解决非正交多址接入（Non-orthogonal multiple access，NOMA）在不完美串行干扰消除（Serial interference cancellation，SIC）条件下的快速配对和功率分配问题，提出一种基于深度强化学习的NOMA用户配对和功率优化方案。首先，考虑多用户NOMA不完美SIC的场景，以用户配对和用户发射功率分配因子为优化变量构建最大化系统可达通信速率的优化问题。分析了不完美SIC条件下用户使用NOMA配对的条件，并推出该条件下最大可达速率的用户功率分配。其次，将用户配对问题当作组合优化的问题，基于实时性的要求使用改进的指针网络设计了一种新型用户配对方案。仿真结果表明，该方案能够有效提升NOMA系统的可达速率，达到了最优的穷搜算法的99.8%，并具有实时性和适应用户数量动态变化的优势。

关键词：非正交多址接入；用户配对；功率分配；不完美串行干扰消除；深度强化学习

中图分类号：TN926.1　　**文献标志码：**A

NOMA User Pairing and Power Allocation Scheme of Deep Reinforcement Learning Based on Pointer Network

LI Guoxin，GAN Qi，CHEN Jin，JIAO Yutao，WANG Haichao，HE Xing

（College of Communications Engineering，Army Engineering University of PLA，Nanjing 210007，China）

Abstract：To solve the fast pairing and power allocation problem of non-orthogonal multiple access（NOMA）under imperfect serial interference cancellation（SIC）conditions，a deep reinforcement learning-based user pairing and power optimization scheme is proposed. First，this paper considers the scenario of imperfect SIC for multiuser NOMA，and constructs an optimization problem to maximize the system reachable communication rate with user pairing and user transmit power allocation factor as optimization variables. The condition of user pairing using NOMA under the imperfect SIC condition is analyzed，and the user power allocation for the maximum reachable rate under this condition is introduced. Second，the user pairing problem is treated as a combinatorial optimization problem，and a novel user pairing scheme is designed based on the real-time requirement using an improved pointer network. Simulation results show that this scheme can effectively improve the reachable rate of the NOMA system to 99.8% of that of the optimal exhaustive search algorithm. It achieves real-time performance and adapts to the dynamic change of

基金项目：国家自然科学基金（62471489，62271501）；国家资助博士后研究人员计划（GZB20240996）；江苏省自然科学基金（BK20240200）；江苏省前沿引领技术基础研究重大项目（BK20212001）。
收稿日期：2024-08-21；**修订日期：**2024-12-19

the number of users.

Key words: non-orthogonal multiple access (NOMA); user pairing; power allocation; imperfect serial interference cancellation (SIC); deep reinforcement learning

引　言

6G是实现智能信息社会的关键技术,它的用户体验通信速率和连接密度将达到5G的10倍[1]。要达到如此高的通信速率和连接密度需要一种不同于5G的多址方式。非正交多址接入(Non-orthogonal multiple access, NOMA)被认为是6G网络中解决高连接密度的关键技术之一[2]。与传统正交多址(Orthogonal multiple access, OMA)相比,NOMA①采用的是功率域的复用,用户能够同时同频发送信号。由于信道和发射功率的差异,接收端可以采取串行干扰消除(Serial interference cancellation, SIC)技术解调目标信号,从而达到多址接入的效果,显著提高频谱利用率[3]。考虑到接收端需要解调非目标用户信号,在NOMA网络中,接收端信号处理复杂度会随着非正交多址接入的用户数量增多而急剧增大。因此,在大规模用户接入网络中,一种平衡有效性和复杂度的方案是将多个用户两两配对,每对用户基于NOMA技术接入同一信道资源,不同用户对之间基于OMA方式接入不同信道资源。在这种大规模用户的NOMA网络中用户配对策略和用户的功率分配方案是重要的研究问题[4]。

有关用户配对的研究很多,首尾配对是一种经典的用户配对方法,它能保证配对用户总体信道差异的最大化。因此,在理想SIC情况下首尾配对为两用户配对的最优方案[4]。但对于密集用户的场景,首尾配对存在近用户问题,即选择信道增益相近的用户进行配对造成用户之间干扰较大而无法解调。此时,采用首尾配对的网络可达速率(Achievable sum rate, ASR)可能小于基于OMA接入方案的ASR[5]。针对这个问题,文献[6]提出了均匀信道增益差分配对方案,将用户按信道增益顺序排列并从中间分为两组,每组将最大的用户按顺序配对。这样使得配对用户之间信道差异变得均匀,避免了近用户问题,但该方案无法发挥用户配对的最佳性能。除了近用户问题外,信道估计的准确性和即时性、硬件精度和干扰等问题,SIC技术通常无法完美地消除用户带来的干扰[7-8]。由于存在近用户和不完美SIC的问题,NOMA配对方案不能一味地追求信道的最大差异,需要穷尽搜索$n!$种配对可能才能得到配对的最优解。一些方案以最小信道差异或最小信噪比作为约束基于启发式迭代搜索的方法进行用户配对。文献[9,10]分别提出了基于模拟退火算法和匈牙利算法进行用户配对,都获得较好的ASR性能。一些研究将不完美SIC建模为类似滤波器的固定比例消除,文献[11]研究了固定比例消除的不完美SIC下行NOMA的配对条件和自适应配对方案。但是,上述基于迭代的算法存在计算效率低和延时过高的问题[12]。为了克服这些问题,基于机器学习的配对方案被广泛研究。文献[13]提出一种根据信道条件和相对位置使用户自适应配对或不配对的算法,并使用深度神经网络(Deep neural network, DNN)去拟合该算法。这种深度学习拟合的方式需要大量数据训练才能获得较好的效果,因此这种方法需要花费大量的时间准备数据标签。而且,这种基于DNN的网络结构只能处理用户数量与训练用户数量相同的情况。文献[14]提出一种无需数据标签的基于深度Q网络(Deep Q-network, DQN)的用户配对方案。具体方式为,用户根据信道增益大小分为两组,使用深度Q网络从两组内选择用户优化系统的ASR。这种方法只能在两组之间选择用户配对,组内用户无法配对,在信道差异较小的情况下可能无法获得较好的性能。综上所述,在NOMA用户配对问题上,传统迭代搜索算法需要时间过久,优化结果可能由于信道变化而失效。基于DNN的深度学习和深度强化学习无法适应用户数量的动态变化。文献[15]同样表明,目前大多数NOMA配对相关的研究都没有解决信道状态的动态变化和用

①NOMA分为功率域NOMA和码域NOMA,本文主要研究功率域的NOMA。

户数量的动态变化。因此,本文致力于:(1)在考虑不完美SIC的情况下,提出一种兼顾实时性和性能的用户配对算法;(2)所提用户算法能够动态适应不同用户数量的变化。

本文将NOMA配对看作组合优化的问题。文献[16]提出一种指针网络(Pointer network, PN)的深度学习拟合方案解决组合优化问题,获得了较好的效果且该网络具有较好的泛化能力,但仍然存在需要大量时间准备数据标签的问题。文献[17]结合指针网络和强化学习解决了需要数据标签的问题,使用优势演员-评论家(Advanced actor-critic, A2C)算法训练网络。PN使用长短期记忆网络(Long short term memory, LSTM)作为编码器,LSTM的记忆功能适合处理具有相关性的序列,例如文本序列。对于组合优化问题来说序列之间往往是独立的,不需要传递序列之间的相关信息,即序列输入的顺序不影响最后的结果。LSTM需要额外学习这一点,因此收敛速度较慢。文献[18]使用卷积神经网络(Convolutional neural network, CNN)作为编码器改进了PN,提出动态指针网络(Dynamic PN, DPN),在不影响PN性能的情况下,节省了近60%的训练时间。文献[17,18]主要使用离线的深度强化学习,难以适应用户的动态特性,因此需要一种在线学习的方法来适应用户的动态特性。由于PN在解决组合优化问题上有较好的性能,且可以泛化到不同用户的场景。因此,本文在文献[18]的基础上使用改进的PN和A2C算法提出一种在线的深度强化学习NOMA配对方案。

基于上述内容,本文主要贡献如下:

(1) 针对不完美SIC的网络用户配对问题,提出了基于改进PN的深度强化学习方法来解决用户配对。以用户的SNR作为网络模型的输入,使用改进的PN选择配对用户。有效避免配对用户的信道差异过大和过小带来的影响,并且本文所使用的网络具有很好的泛化能力,能够适应不同用户数量的通信场景。

(2) 对于NOMA用户的功率分配问题,在考虑公平性的角度上,对NOMA配对的两用户可达速率都不小于OMA的情况下推导了功率分配的上下界,确定了最大ASR的功率分配,并在此基础上得出NOMA的配对条件。

1 系统模型

如图1所示,本文考虑多用户上行NOMA传输网络。该系统由N个用户和一个基站(Base station, BS)组成。假设每个用户到BS的信道状态信息(Channel state information, CSI)可以在BS处获取,且CSI在一段时间内不变。用户的集合表示为$u\in \boldsymbol{U}=\{1,2,\cdots,N\}$,在OMA的情况下用户的信噪比(Signal noise ratio, SNR)为

$$\gamma_u=\frac{P|h_u|^2}{\sigma^2} \tag{1}$$

式中:P为用户的最大发射功率,σ^2为高斯白噪声功率,h_u为用户信道增益,表达式为

$$h_u=g\beta_0 d_u^{-\lambda} \tag{2}$$

式中:g为瑞利衰落信道系数,β_0为单位距离下的路径损耗,d_u为用户u到BS的距离,λ为路径损耗指数。此时OMA用户的可达通信速率可以表示为

$$R_u=0.5B\log_2(1+\gamma_u) \tag{3}$$

式中:0.5为OMA复用的损耗,B为信道带宽。

在两用户NOMA情况下,将N用户分为$N/2$个用户对,其中$i\in\{0,1,\cdots,N/2\}$表示第i对用户对,每个用户对

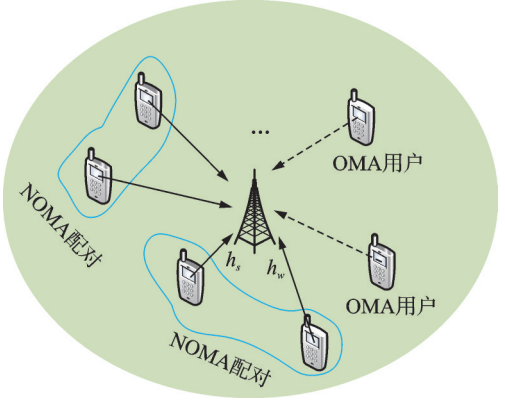


图1 系统模型图

Fig.1 System model

的两个用户为$s_i\in\boldsymbol{U}$,$w_i\in\boldsymbol{U}$,信道增益较大的用户的信道增益为h_{s_i};信道增益较小的用户信道增益为h_{w_i},即$|h_{s_i}|>|h_{w_i}|$。第i对用户对中强用户与弱用户的NOMA接入信干噪比(Signal to interference plus noise ratio, SINR)可以分别表示为

$$\tilde{\gamma}_{s_i}(\alpha)=\frac{\alpha P|h_{s_i}|^2}{\sigma^2+(1-\alpha)P|h_{w_i}|^2},\qquad \tilde{\gamma}_{w_i}(\alpha)=\frac{(1-\alpha)P|h_{w_i}|^2}{\sigma^2+\alpha\beta P|h_{s_i}|^2} \tag{4}$$

式中:$\tilde{\gamma}_{s_i}$为强用户的SINR,$\tilde{\gamma}_{w_i}$为弱用户的SINR,$\alpha\in[0,1]$表示用户的发射功率分配比例,$\beta\in[0,1]$表示不完美系数。对于完美SIC的NOMA网络来说,信道之间的差异越大,NOMA配对的增益越大[19]。但在上行NOMA网络中,根据式(4)可以观察到,当不完美系数$\beta\neq0$时信道差异越大,$\tilde{\gamma}_{w_i}$越小,即弱用户的可达速率越小。信道差异增大虽然会使强用户的SINR增大,但同样会使得弱用户的SINR下降,因此需要对用户的信道差异进行权衡。

根据式(1,4)可以得到

$$\tilde{\gamma}_{s_i}(\alpha)=\frac{\alpha\gamma_{s_i}}{1+(1-\alpha)\gamma_{w_i}},\qquad \tilde{\gamma}_{w_i}(\alpha)=\frac{(1-\alpha)\gamma_{w_i}}{1+\beta\alpha\gamma_{s_i}} \tag{5}$$

式中:γ_{s_i},γ_{w_i}分别表示与$\tilde{\gamma}_{s_i}$,$\tilde{\gamma}_{w_i}$对应用户的OMA接入方式的信噪比。因此,NOMA两用户的可达通信速率可以表示为

$$\tilde{R}_{s_i}(\alpha)=B\log_2(1+\tilde{\gamma}_{s_i}(\alpha)),\qquad \tilde{R}_{w_i}(\alpha)=B\log_2(1+\tilde{\gamma}_{w_i}(\alpha)) \tag{6}$$

对于每个用户对,NOMA方式的ASR表示为$R_i^{\mathrm{NO}}(\alpha)=\tilde{R}_{s_i}(\alpha)+\tilde{R}_{w_i}(\alpha)$,OMA方式的ASR表示为$R_i^{\mathrm{O}}=R_{s_i}+R_{w_i}$,$R_{s_i}$,$R_{w_i}$为对应用户采用OMA接入的可达速率。定义向量$\boldsymbol{y}=[y_1,y_2,\cdots,y_{N/2}]$,当$y_i=0$时表示用户对$i$使用OMA传输方式;当$y_i=1$时表示用户对$i$使用NOMA传输方式。此时$N$个用户的总ASR表示为

$$R_{\mathrm{sum}}(\alpha)=\sum_{i=1}^{N/2}(1-y_i)R_i^{\mathrm{O}}+y_iR_i^{\mathrm{NO}}(\alpha) \tag{7}$$

根据式(7)可知,系统总ASR主要取决于OMA或NOMA的选择(即y_i的取值);以及对应的R_i^{O},R_i^{NO}的大小。这两者与NOMA的用户的信道增益以及功率分配是密切相关的,因此本文构建的优化问题如下

$$\mathrm{P0}:\max_{\alpha,y_i,\pi}R_{\mathrm{sum}}(\alpha) \tag{8a}$$

$$\text{s.t.}\quad 0\leqslant\alpha\leqslant1 \tag{8b}$$

$$y_i\in\{0,1\} \tag{8c}$$

$$|h_i^s|>|h_i^w| \tag{8d}$$

在P0中通过优化用户的发射功率分配因子α、用户接入方式y_i和用户配对策略π来使得系统总ASR最大,其中式8(b)表示NOMA用户功率分配因子的约束,式8(c)表示用户的NOMA和OMA两种接入方式,式8(d)表示强用户信道增益必须大于弱用户。

2 给定用户下的接入方式选择和功率分配

由于不同的功率分配系数和不完美系数的存在,NOMA用户的可达速率可能无法保证优于OMA用户。本节将推导NOMA用户速率大于OMA用户的条件,并确定最优功率分配系数和采用NOMA接入的条件。

对于强用户要NOMA可达速率大于OMA,即$\tilde{R}_{s_i} \geqslant R_{s_i}$。根据式(5,6)得

$$\log_2\left(1+\frac{\alpha\gamma_{s_i}}{1+(1-\alpha)\gamma_{w_i}}\right)>\frac{1}{2}\log_2(1+\gamma_{s_i}) \tag{9}$$

再根据式(9)可以解得

$$\alpha>\frac{(1+\gamma_{w_i})(\sqrt{1+\gamma_{s_i}}-1)}{\gamma_s+\gamma_{w_i}(\sqrt{1+\gamma_{s_i}}-1)}\triangleq\alpha_{\mathrm{L}} \tag{10}$$

式(10)表明当功率分配因子$\alpha>\alpha_{\mathrm{L}}$时,强用户采用NOMA传输的速率大于OMA传输方式的速率,因此α_{L}为功率分配因子的下界。这与完美SIC的NOMA相同,且随着α的增大,NOMA用户的速率会提升。但α不能太大,因为过大的α会使得弱用户的通信速率下降。因此,下一步将推导α的上界。

推导α的上界,弱用户需要满足$\tilde{R}_{w_i} \geqslant R_{w_i}$,同样可以得到类似的表达

$$\log_2\left(1+\frac{(1-\alpha)\gamma_{w_i}}{1+\beta\alpha\gamma_{s_i}}\right)>\frac{1}{2}\log_2(1+\gamma_{w_i}) \tag{11}$$

再根据式(11)可以得到功率分配因子的上界

$$\alpha<\frac{\gamma_{w_i}-\sqrt{1+\gamma_{w_i}}+1}{\gamma_{w_i}+\beta\gamma_{s_i}(\sqrt{1+\gamma_{w_i}}-1)}\triangleq\alpha_{\mathrm{U}} \tag{12}$$

式(12)表明当功率分配因子$\alpha<\alpha_{\mathrm{U}}$时,NOMA弱用户的速率大于对应用户采用OMA的速率。对于完美SIC的情况,$\alpha_{\mathrm{U}}>\alpha_{\mathrm{L}}$都能得到满足;但对于不完美的SIC,$\alpha_{\mathrm{U}}>\alpha_{\mathrm{L}}$不一定会满足,即单个NOMA用户的速率不一定大于OMA用户[11]。这种情况下,无论α取何值都无法同时满足两个用户采用NOMA的速率都大于采用OMA,即$\tilde{R}_{s_i} \geqslant R_{s_i}, \tilde{R}_{w_i} \geqslant R_{w_i}$。此时,从用户公平性的角度出发,若用户不想牺牲自己的速率则不会采用NOMA。因此,本文认为只有两个用户采用NOMA速率都大于采用OMA时,才会采用NOMA接入,其余情况仍然使用传统OMA的传输方式。

令$\alpha_{\mathrm{U}}>\alpha_{\mathrm{L}}$,根据式(10,12)可以得到

$$\beta<\frac{(\gamma_{w_i}-\sqrt{1+\gamma_{w_i}}+1)-\alpha_{\mathrm{L}}\gamma_{w_i}}{\gamma_{s_i}(\sqrt{1+\gamma_{w_i}}-1)}\triangleq\beta_{\mathrm{L}} \tag{13}$$

式(13)表示当$\beta<\beta_{\mathrm{L}}$时$\alpha_{\mathrm{U}}>\alpha_{\mathrm{L}}$成立,此时一定存在$\alpha^*\in[\alpha_{\mathrm{L}},\alpha_{\mathrm{U}}]$使得$\tilde{R}_{s_i} \geqslant R_{s_i}, \tilde{R}_{w_i} \geqslant R_{w_i}$同时满足,因此可以认为式(13)为使用NOMA传输方式的条件。在本文的系统中,假设BS无法消除,干扰的比例$\beta=\beta_{\mathrm{th}}$为固定值,这由设备本身决定。观察式(10,13),可以发现β_{L}只与两用户的信噪比有关,若最大发射功率P固定,则β_{L}只与用户信道有关。也就是说,获取用户的CSI后可以计算出β_{L},若$\beta_{\mathrm{th}}<\beta_{\mathrm{L}}$,则认为用户可以采用NOMA接入方式,否则用户不进行配对采用OMA。根据此条件,可以使配对的用户信道差异在合适的范围内,信道差异过大或过小都无法满足NOMA传输的条件。

与完美SIC情况一样,R_i^{NO}是功率分配因子α的增函数。为了使得ASR最大且保证弱用户的正常通信,本文中NOMA用户功率分配因子取值为$\alpha^*=\alpha_{\mathrm{U}}$。此时,强用户的速率$\tilde{R}_{s_i}>R_{s_i}$,弱用户速率$\tilde{R}_{w_i}=R_{w_i}$且两用户的和速率$R_i^{\mathrm{N}}$是在$\alpha_{\mathrm{U}}>\alpha_{\mathrm{L}}$条件下的最大值。OMA用户的传输功率则采用最大发射功率P。

3 用户配对算法

第2节推导了功率分配因子的选取,以及NOMA配对的成立条件。本节主要利用NOMA配对条件提出一种基于深度强化学习的用户配算法,将用户的信噪比输入深度神经网络得到每个用户的配对

对象，并利用强化学习算法优化该网络，使得配对尽可能满足NOMA配对条件，同时最大化用户的ASR。

由于指针网络在解决组合优化问题上具有较好的泛化能力，本文继续沿用指针网络的结构，并在DPN的基础上进行了改进。图2给出了本文采用的网络与DPN的对比，图2(a)为DPN网络结构，DPN将用户信息序列进行卷积编码后，将其编码信息的切片输入到译码器并由门控循环单元(Gated recurrent unit, GRU)网络记忆当前选择用户的信息。这样使得网络在选取当前用户时可以考虑用户序列前后的关系。GRU虽然可以记忆用户之间的关系，但这种关系需要一定时间训练得到，而且在强化学习中使用GRU会引起架构上的变化，即每次执行网络需要将上一次GRU网络的隐藏状态反馈到当前网络，如图2(a)中的h_1，这样降低了网络的收敛速度和效率。因此DPN只输入一个用户信息序列，中间网络相当于一个整体，依靠GRU学习用户之间的关系，中间不涉及外部状态的变化，直接输出整个组合的结果。因此，DPN是一个离线强化学习的架构。

图2(b)中本文将DPN的译码器换为了CNN与DNN的组合，避免了采用GRU。对于记忆已选择用户的问题，图2(b)在所提网络输入信息中加入一个动态变量表示当前选择的用户，同时使用mask机制表示已选择的用户。相比于DPN利用GRU通过学习来记忆用户，本文之前给出已选择用户和当前选择用户。避免了GRU的额外学习，提升了网络收敛速度。本文网络不仅输入用户信息序列，每次配对还将当前用户信息和已选择用户作为输入，这样每次只输出当前选取的结果。因此，所提网络是在线学习的架构。

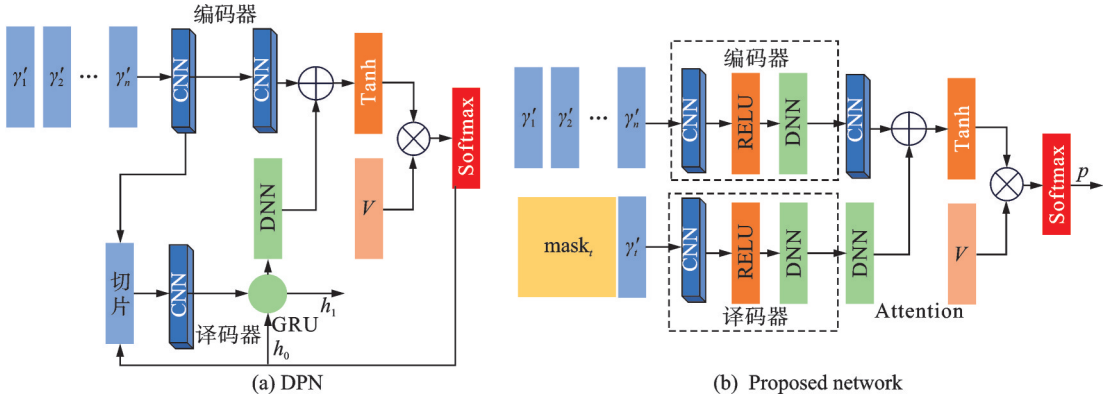


图2 网络对比图

Fig.2 Network comparison

图2中网络输入γ'_n为用户信噪比序列的规范化值[20]。具体表示为

$$\gamma'_n=\frac{\log_{10}\gamma_n-E\left[\log_{10}\gamma_n\right]}{\sqrt{E\left[\left(\log_{10}(\gamma_n)-E\left[\log_{10}(\gamma_n)\right]\right)^2\right]}} \tag{14}$$

式中：$n=1,2,\cdots,N$，γ_n表示第n个用户的SNR。

图2(b)中Attention机制采用了多层感知器(Multi-layer perceptron, MLP)的相关度计算方法，具体计算方法为

$$\boldsymbol{u}=\boldsymbol{V}^{\mathrm{T}}\tanh\left(\boldsymbol{W}_{\mathrm{ref}}\cdot\boldsymbol{O}_{\mathrm{ref}}+\boldsymbol{W}_q\cdot\boldsymbol{q}\right) \tag{15}$$

式中：$\boldsymbol{V}\in\mathbf{R}^{1\times d}$；$\boldsymbol{W}_{\mathrm{ref}},\boldsymbol{W}_q\in\mathbf{R}^{d\times d}$；$\boldsymbol{u}\in\mathbf{R}^{1\times N}$；$\boldsymbol{W}_{\mathrm{ref}},\boldsymbol{W}_q$为全连接层参数，$\boldsymbol{V}$为可学习的参数，$d$为网络的隐藏单元数量。$\boldsymbol{q}\in\mathbf{R}^{d\times 1}$为译码器的输出，$\boldsymbol{O}_{\mathrm{ref}}\in\mathbf{R}^{d\times N}$为编码器的输出。为了防止用户被多次选中，利

用mask对相关度$\boldsymbol{u}$修正

$$u_n=\begin{cases}u_n & M_{\text{mask}_t,n}=0\\ -\infty & M_{\text{mask}_t,n}=1\end{cases} \tag{16}$$

则配对用户的概率分布为

$$p(\boldsymbol{O}_{\text{ref}},\boldsymbol{q};\ \boldsymbol{W}_{\text{ref}},\boldsymbol{W}_q,\boldsymbol{V})=\text{Softmax}(\boldsymbol{u}) \tag{17}$$

每次配对以概率分布p进行采样来选择用户。多次Attention可以进一步提升网络的性能[18],具体表示为

$$\boldsymbol{g}_l(\boldsymbol{O}_{\text{ref}},\boldsymbol{g}_{l-1};\ \boldsymbol{W}_{\text{ref}}^l,\boldsymbol{W}_q^l,\boldsymbol{V}^l)=\sum_{i=1}^{N}\boldsymbol{O}_{\text{ref}_i}p_i \tag{18}$$

式中:$\boldsymbol{g}_0=\boldsymbol{q}$,$\boldsymbol{W}_{\text{ref}}^l$、$\boldsymbol{W}_q^l$、$\boldsymbol{V}^l$为第$l$次Attention的全连接参数,取最后一次Attention的概率分布进行采样。

本文对用户配对构建的强化学习环境、状态和奖励如下。

环境和状态:在本文中环境状态定义为所有用户的信噪比序列和当前需要配对用户的信噪比的组合。首先将用户的信噪比按升序排列为$\boldsymbol{\Gamma}=\{\gamma_1,\gamma_2,\cdots,\gamma_n,\cdots,\gamma_N\}$。定义$\boldsymbol{M}_{\text{mask}_t}\in\mathbf{R}^{1\times N}$向量表示$t$时刻已配对的用户,$M_{\text{mask}_t,n}=1$表示第$n$个用户已配对,$M_{\text{mask}_t,n}=0$表示第$n$个用户未配对。$\boldsymbol{M}_{\text{mask}_t}$有两个作用:(1)指示当前环境的用户配对情况;(2)防止用户多次被选择。令初始状态$\boldsymbol{S}_1=[\gamma_1,\gamma_2\cdots,\gamma_n,\cdots,\gamma_N,\gamma_1,\boldsymbol{M}_{\text{mask}_1}]$,$\boldsymbol{S}_1$中最后一个元素$\gamma_1$表示当前配对的用户为第1个用户。当智能体与环境发生交互后,环境状态会发生改变,环境会在当前未配对的用户中选择信噪比最小的用户作为当前需要配对的用户。因此,一次完整的配对过程包含$N/2$个状态,$\boldsymbol{S}_t=\{\gamma_1,\gamma_2,\cdots,\gamma_n,\cdots,\gamma_N,\gamma_t,\boldsymbol{M}_{\text{mask}_t}\}$,$t=0,1,\cdots,N/2$。由于配对用户仅与当前需要配对的用户有关,所以符合马尔可夫决策过程。

动作:动作$a_t\in\{0,1,\cdots,N-1\}$为选择相应的用户与需要配对的用户进行配对。

奖励:优化目标为最大化网络的ASR,因此每个动作的奖励设置为两个配对用户的速率之和。具体为$r_t=(1-y_i)R_i^{\text{O}}+y_iR_i^{\text{NO}}(\alpha)$,一个回合的总奖励计算方式如式(7)所示。

图3为6用户配对示意图,首先对用户信噪比进行升序排列,然后选择未配对的最小信噪比用户作为配对的弱用户,并初始化强化学习环境状态,即环境状态为$\boldsymbol{S}_1=\{\gamma_1,\gamma_2,\cdots,\gamma_n,\cdots,\gamma_N,\gamma_t,\boldsymbol{M}_{\text{mask}_t}\}$;将状态输入所提网络中进行配对,网络输出所有用户的概率分布,根据概率分布采样得到与最小信噪比用户配对的强用户。如图3所示$t=0$时刻选择用户1作为弱用户进行配对,通过所提网络得到对应的强用户为用户5。当前配对完成后再选取其余未配对的最小信噪比用户,直到$t=N/2$且所有用户完成配对。其中,每对用户完成配对后根据式(13)计算β_{L},若$\beta_{\text{L}}>\beta$则采用NOMA接入。设图3例子中的$\beta=0.01$,则第1对用户和第2对用户采用NOMA接入,第3对用户采用OMA接入。确定配对用户和接入方式后利用式(10,12)计算功率,取数值$\alpha^*=\alpha_{\text{U}}$作为NOMA用户的发射功率。

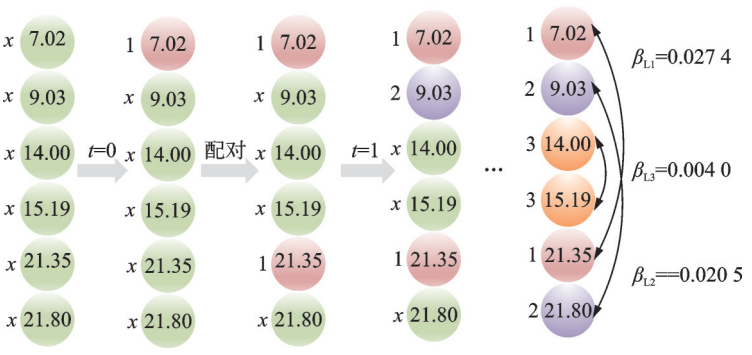


图3 配对流程图

Fig.3 Pairing process diagram

本文采用强化学习算法训练所提网络,具体算法结合了Rollout

baseline[21]和近端梯度优化(Proximal policy optimization, PPO)算法。PPO算法是一种高效的基于策略梯度的在线强化学习算法,算法基于演员-评论家(Actor-critic, AC)框架,利用重要性采样提高数据的利用率。Rollout baseline采用一个与Actor相同结构的网络,且该网络参数为历史奖励最高的参数,去评价当前回合Actor的表现。Rollout baseline方法需要一个回合内的总奖励去评价Actor,即智能体需要完整执行一回合才能训练一次,类似蒙特卡罗方法,这是由于在组合优化问题中的奖励通常需要一个完整的组合才能得出,因此这样的方法更适合组合优化问题。而PPO算法训练是以每个时间步为单位,这样虽然数据利用变高但不利于组合优化问题的收敛。

因此本文采用Rollout baseline方法替换PPO算法中的Critic,具体如图4所示,且将优势函数更换为

$$\mathrm{adv}(\theta)=R^{\mathrm{A}}_{\mathrm{sum}}-R^{\mathrm{C}}_{\mathrm{sum}} \tag{19}$$

式中:$R^{\mathrm{A}}_{\mathrm{sum}}$为当前回合Actor获得的总奖励,$R^{\mathrm{C}}_{\mathrm{sum}}$为Rollout Baseline方法的总奖励,若$\mathrm{adv}(\theta)<0$则说明当前Actor表现不如Critic,Actor应该向Critic学习,反之$\mathrm{adv}(\theta)>0$说明Actor学习方向是对的,则加强该方向的学习。Actor采用图2(b)的网络模型,Critic网络由Actor复制得到。与传统PPO不同,本文中Actor-critic都需要与环境交互。将交互得到的经验按回合保存到缓冲区中,根据式(19)计算优势函数,此时PPO算法Actor的损失函数可以表示为

$$J_\theta=-\min\left(\frac{p_{\theta_{\mathrm{now}}}}{p_\theta},\mathrm{clip}\left(\frac{p_{\theta_{\mathrm{now}}}}{p_\theta},1-\epsilon,1+\epsilon\right)\right)\times\mathrm{adv}(\theta)-\xi E_{\mathrm{e}} \tag{20}$$

式中

$$\frac{p_{\theta_{\mathrm{now}}}}{p_\theta}=\exp\left(\ln p_{\theta_{\mathrm{now}}}(\pi|\theta)-\ln p_\theta(\pi|\theta)\right) \tag{21}$$

表示当新策略与旧策略的相识程度,E_{e}为概率分布$p_{\theta_{\mathrm{now}}}$的熵,$\xi$为熵的权重系数。Critic按Rollout baseline方法更新,即当优势函数均值$\overline{\mathrm{adv}(\theta)}>0$时对优势函数进行单边T检验,若T检验的值大于0.05时表明此时Actor奖励显著优于Critic。则将Actor参数复制到Critic中,否则不更新Critic。

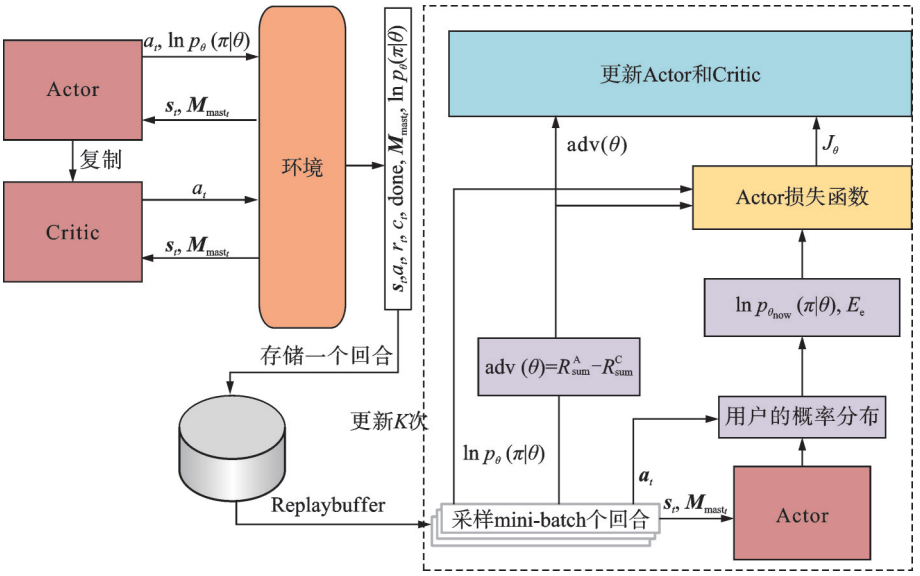


图4 RPPO训练算法框图

Fig.4 RPPO training algorithm block diagram

所提配对方法的具体训练流程如算法1所示。

算法1 Rollout baseline PPO

初始化:最大训练回合E;Actor网络参数θ;Critic网络参数θ_b;环境状态S_t;mimi-batch大小

for episode=1,2,…,E do

while not done

执行Actor网络得到可选用户的概率分布,

根据概率分布采样选择用户和得到对数概率$\ln p_\theta(\pi|\theta)$,

Actor环境执行动作选择用户进行配对,根据式(10,12,13)分别计算接入方式和功率分配因子给出奖励和下一状态,

执行Critic网络得到可选用户的概率分布,

贪婪采样(选择概率最大的用户)选择用户和计算对数概率$\ln p_\theta(\pi|\theta_b)$,

Critic环境执行动作选择用户进行配对,根据式(10,12,13)分别计算接入方式和功率分配因子给出奖励和下一状态,

分别保存Actor和Critic的状态、动作,对数概率和奖励到缓冲区中

if 所有用户配对完成 then

done = True

end if

根据式(19)计算优势函数adv(θ)

for k=1,2,…,K

随机采样mini-batch个回合的数据,

随机采样的数据代入Actor得到当前参数下的对数概率和熵$\ln p_{\theta_{new}}(\pi|\theta)$,$E_e$,

根据式(20)计算损失函数,

Adam优化器更新actor网络参数θ,

end for

if mean(adv(θ))>0&.单边T检验的值小于0.05 then

$\theta_b=\theta$

end if

end for

4 仿真分析

本节对所提方法进行仿真并给出仿真结果,分析算法性能。假设N=20个用户随机分布在半径为80 m的区域内,用户最大发射功率P=10 dBm,噪声功率σ^2=−100 dBm,信道路径损耗指数为4,当用户离BS太近时损耗固定为−20 dB,用户带宽为B=1,不完美系数的阈值β_{th}=0.001即−30 dB,初始学习率为$1e^{-3}$,学习率衰减率为0.95。训练阶段依照概率分布采样选取用户,测试阶段采用贪婪的方式选取用户。本文仿真环境为Pytorch 1.9.1+CUDA 11.2。其中图形处理器(Graphics processing unit,GPU)仅在训练阶段使用,测试阶段采用CPU。

图5为本文方法与其他深度强化学习算法的的学习曲线对比，可以看出原始PPO算法对求解组合优化问题效果较差，而DQN算法[14]中没有采用mask机制而是采用奖励约束防止用户被多次选中配对，所以初始奖励较低，从最后收敛结果可以看出本文方法收敛后性能明显比DQN要好；DPN+Rollout baseline算法为文献[18]中的DPN网络和Rollout baseline方法的结合，可以看出该方法同样可以得到较好的效果，但相比于本文方法收敛速度要慢，因此本文方法能有效提高网络收敛速度和用户ASR。

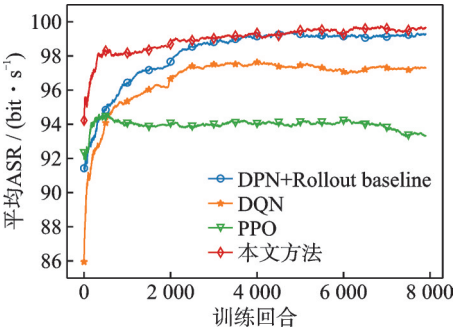


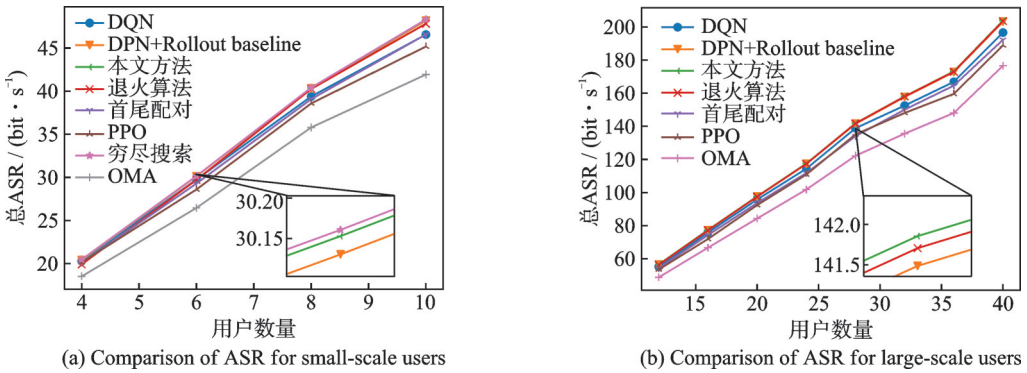
图5 学习曲线图

Fig.5 Learning curve

图6(a)对比了不同用户配对方案随着用户数量的变化对总ASR的影响。OMA用户均采用最大发射功率P，NOMA用户的功率分配取值为$\alpha^{*}=\alpha_{\mathrm{U}}$。对于所有方案，随着用户数量的增加总ASR呈线性增长。可以看出全部采用OMA方案的总ASR最低，这表明即使在不完美的SIC条件下NOMA接入方式还是能获得一定的增益。配对方法上除了对比深度强化学习方法外还对比了传统基于迭代搜索的退火算法[9]和基于规则的首尾配对算法[4]。

随着用户数量的变化，本文方法明显优于首尾配对。与推测的一样，由于文献[14]所提的DQN算法只能在组内配对，在用户数量较多时，用户信道差异较小，算法性能较低。退火算法、AC算法和本文方法十分接近穷搜算法，本文方法性能小幅度优于退火算法。由于穷搜算法在用户数量较多时需要时间难以接受，而退火算法和DPN+Rollout baseline算法性能接近穷搜，所以下文以退火算法作为基准算法对比。

图6(b)增加用户数量，对比了不同算法的总ASR的变化趋势，值得注意的是文献[14]中DQN算法采用的DNN网络结构无法适应不同用户的情况，为了对比性能，本文训练了多个DQN进行对比，而本文方法测试使用的是同一次训练得到的模型，即使用户数量超出了训练使用的用户数量仍能获得较好的性能，表明本文所用的模型可以适应不同用户数量的变化且拥有较好的泛化能力。这点在实际使用中十分重要，因为通常只能使用有限的用户去训练模型，但实际使用中用户数量往往不固定。另外，可以看出DPN+Rollout baseline算法虽然仍能获得较好的性能但随着用户数量的增加其性能不如退火算法，而本文方法仍然优于退火算法。



(a) Comparison of ASR for small-scale users (b) Comparison of ASR for large-scale users

图6 用户数量与总ASR关系图

Fig.6 Relationship between user count and total ASR

虽然DPN和本文网络都可以解决用户动态变化的问题,但是随着用户数量的变化其性能也会随之变化。图7为以退火算法为基准,随着用户数量的增加DPN+Rollout baseline算法和所提方案与退火算法的相对误差对比。可以看出随着用户数量的增加,本文方法与退火算法相对误差增大,但本文方法相比DPN+Rollout baseline算法误差要小,因此本文方法相比DPN+Rollout baseline算法更稳定,用户数量泛化能力更好。

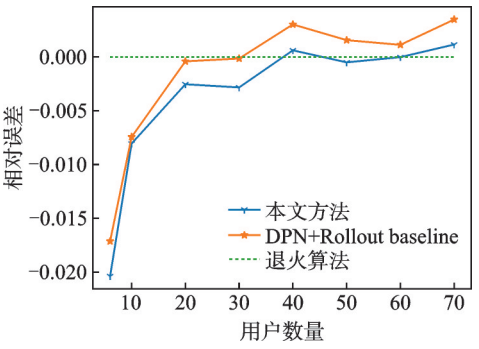


图7 随着用户数量增加相对误差对比

Fig.7 Comparison of relative errors with the increase of number of users

图8给出了本文方法和退火算法在用户数量增加的情况下所用时间的对比。最优算法需要穷尽搜索$n!$种可能的排列,耗时过久,因此本文不对比最优算法的运行时间。退火算法虽然也能得到近似的最优解,但退火算法需要多次迭代可能造成通信时延过长的问题。本文提出的算法时间上远远小于退火算法,且随着用户数量增加算法耗时增长速度也远远小于退火算法。另外,DPN因为采用GRU网络执行速度较慢,而本文网络能有效提高用户配对速度,所提网络推理时间相比DPN提升约40%。

图9为不完美系数变化对比图,从图9中可以发现在不完美系数变大时,总ASR呈下降趋势。当不完美系数$\beta_{\text{th}}=0.001$时,所提方案、DPN和退火算法性能接近。由于DPN为离线学习方案,在不完美系数变大时该方案ASR下降幅度更大,无法得到较好的性能,而本文方法由于在线学习的优势能很快得到较好的性能。而当不完美系数变化较大时,如$\beta_{\text{th}}=0.008$时可以看到本文方法需要更多的训练回合才能达到接近退火算法的性能。

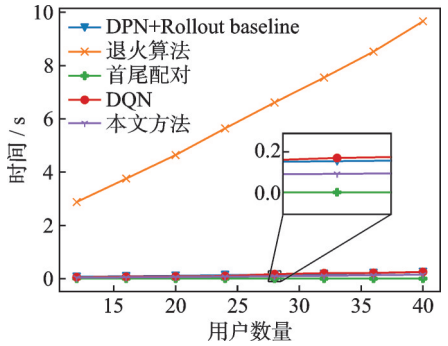


图8 算法时间对比图

Fig.8 Comparison of algorithm time

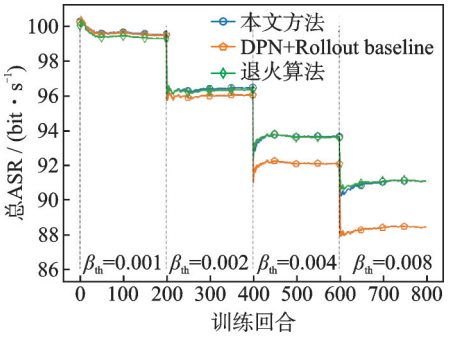


图9 不完美系数变化对比图

Fig.9 Comparison of imperfect coefficient changes

5 结束语

本文在不完美SIC的上行NOMA网络的通信场景下,提出一种NOMA功率分配方案以及NOMA配对方案优化用户的ASR。对于NOMA功率分配,本文推导了功率分配边界和不完美系数的边界并得到NOMA接入条件;提出了在确保NOMA弱用户的通信速率的条件下的用户功率分配方案。对于NOMA配对,本文提出一种基于深度强化学习的NOMA配对方法。使用一种改进的指针网络作为配对的网络模型,并使用Rollout baseline PPO方法训练该模型,将用户的信噪比作为模型的输入,得到用户的配对方案。仿真结果表明了本文所提算法在不完美SIC的NOMA网络中能够在较低的时间内提

高系统的总ASR,而且所提模型有着较好的泛化能力可以适应不同用户的情况,利用在线学习的方法使用环境的动态变化。虽然本文采用的是上行网络,但本文所提的配对方法同样适用于下行网络。本文仅考虑了两用户NOMA配对的情况,这也是目前主要研究的场景,未来将进一步考虑不固定数量的NOMA用户分组方案。

参 考 文 献:

[1] ZHANG Z, XIAO Y, MA Z, et al. 6G Wireless networks: Vision, requirements, architecture, and key technologies[J]. IEEE Vehicular Technology Magazine, 2019, 14(3): 28-41.

[2] LIU Y, ZHANG S, MU X, et al. Evolution of NOMA toward next generation multiple access (NGMA) for 6G[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(4): 1037-1071.

[3] 瞿儒枫, 王鸿. 协作IRS辅助的CoMP-NOMA网络传输方案[J]. 数据采集与处理, 2024, 39(5): 1287-1296.
QU Rufeng, WANG Hong. A transmission scheme for cooperative-IRS-aided CoMP-NOMA networks[J]. Journal of Data Acquisition and Processing, 2024, 39(5): 1287-1296.

[4] ZHU L, ZHANG J, XIAO Z, et al. Optimal user pairing for downlink non-orthogonal multiple access (NOMA)[J]. IEEE Wireless Communications Letters, 2019, 8(2): 328-331.

[5] DING Z, FAN P, POOR H V. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions[J]. IEEE Transactions on Vehicular Technology, 2016, 65(8): 6010-6023.

[6] SHAHAB M B, IRFAN M, KADER M F, et al. User pairing schemes for capacity maximization in non-orthogonal multiple access systems: User pairing schemes in NOMA[J]. Wireless Communications and Mobile Computing, 2016, 16(17): 2884-2894.

[7] 殷志远, 陈瑾, 李国鑫, 等. 延迟CSI反馈下的协作NOMA系统用户选择方法[J]. 陆军工程大学学报, 2022, 1(2): 21-28.
YIN Zhiyuan, CHEN Jin, LI Guoxin, et al. A user selection scheme for cooperative NOMA system with delayed CSI feedback [J]. Journal of Army Engineering University of PLA, 2022, 1(2): 21-28

[8] DO D T, NGUYEN T T T. Impacts of imperfect SIC and imperfect hardware in performance analysis on AF non-orthogonal multiple access network[J]. Telecommunication Systems, 2019, 72(4): 579-593.

[9] ZHANG X, WANG J, WANG J, et al. A novel user pairing in downlink non-orthogonal multiple access[C]//Proceedings of 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). Valencia, Spain: IEEE, 2018: 1-5.

[10] DINH P, ARFAOUI M A, SHARAFEDDINE S, et al. A low-complexity framework for joint user pairing and power control for cooperative NOMA in 5G and beyond cellular networks[J]. IEEE Transactions on Communications, 2020, 68(11): 6737-6749.

[11] MOUNI N S, KUMAR A, UPADHYAY P K. Adaptive user pairing for NOMA systems with imperfect SIC[J].Proceedings of IEEE Wireless Communications Letters, 2021, 10(7): 1547-1551.

[12] KIM D, JUNG H, LEE I H. User selection and power allocation scheme with sinr-based deep learning for downlink NOMA [J]. IEEE Transactions on Vehicular Technology, 2023, 72(7): 8972-8986.

[13] PERDANA R H Y, NGUYEN T V, AN B. Adaptive user pairing in multi-irs-aided massive MIMO-NOMA networks: Spectral efficiency maximization and deep learning design[J]. IEEE Transactions on Communications, 2023, 71(7): 4377-4390.

[14] JIANG F, GU Z, SUN C, et al. Dynamic user pairing and power allocation for NOMA with deep reinforcement learning[C]//Proceedings of 2021 IEEE Wireless Communications and Networking Conference (WCNC). Nanjing, China: IEEE, 2021: 1-6.

[15] TRABELSI N, CHAARI FOURATI L, CHEN C S. Interference management in 5G and beyond networks: A comprehensive survey[J]. Computer Networks, 2024, 239: 110159.

[16] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[EB/OL].(2015-01-01). https://papers.nips.cc/paper_files/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html.

[17] BELLO I, PHAM H, LE Q V, et al. Neural combinatorial optimization with reinforcement learning[C]//Proceedings of International Conference on Learning Representations. Toulon, France: [s.n.], 2017.

[18] NAZARI M, OROOJLOOY A, SNYDER L, et al. Reinforcement learning for solving the vehicle routing problem[C]// Proceedings of Neural Information Processing Systems. Montreal, Canada: Curran Associates, 2018.

[19] DING Z, PENG M, POOR H V. Cooperative non-orthogonal multiple access in 5G systems[J]. IEEE Communications Letters, 2015, 19(8): 1462-1465.

[20] 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.

[21] KOOL W, HOOF H VAN , WELLING M. Attention, learn to solve routing problems![C]//Proceedings of International Conference on Learning Representations. New Orleans, USA: [s.n.], 2019.

作者简介:

李国鑫(1990-),男,博士,副教授,硕士生导师,研究方向:协作通信、认知无线电、无线携能通信和非正交多址技术,E-mail:guoxin@aeu.edu.cn。

甘麒(1999-),**通信作者**,男,硕士生,研究方向:移动边缘计算、非正交多址技术,E-mail:ganwuqi@foxmail.com。

陈瑾(1971-),女,博士,教授,博士生导师,研究方向:认知无线电、分布式优化算法和数字信号处理。

焦雨涛(1992-),男,博士,副教授,硕士生导师,研究方向:移动区块链、去中心化机器学习和群智感知。

王海超(1991-),男,博士,副教授,硕士生导师,研究方向:无线通信、无人机网络和隐蔽通信。

贺兴(1984-),男,硕士生,研究方向:无线通信对抗、智能干扰。

(编辑:陈珺)