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条件下的快速配对和功率分配问题。该文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，该文考虑多用户 NOMA 不完美连续干扰消除(Serial Interference Cancellation, 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，该方案能够有效的提升 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　**文献标识码**：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

Abstract: In order to solve the fast pairing and power allocation problem of Non-Orthogonal Multiple Access (NOMA) under imperfect serial interference cancellation conditions. The paper proposes a deep reinforcement learning-based user pairing and power optimization scheme for NOMA. First, the paper considers the scenario of imperfect Serial Interference Cancellation (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 the optimal exhaustive search algorithm, and has the advantages of real-time and adapting to the dynamic change of the number of users.

Key words: Non Orthogonal Multiple Access(NOMA); User pairing; Power Allocation; Imperfect successive interference cancellation; Deep reinforcement learning

1 引言

第六代移动通信网络（6G）是实现智能信息社会的关键推动技术，它的用户体验通信速率和连接密度将达到第五代移动通信网络（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，NOMA[1]采用的是功率域的复

[1] NOMA 分为功率域 NOMA 和码域 NOMA，本文主要研究功率域的 NOMA。

基金项目：国家自然科学基金(62101595，62271501)，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)

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有很多，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 技术通常无法完美的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 配对方案不能一味的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]和文献[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的研究。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]提出一种这种方案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 Networks, 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。指针网络使用长短期记忆网络(Long Short Term Memory)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)作为编码器改进了指针网络，提出动态指针网络(Dynamic Pointer Network, DPN)，在不影响指针网络性能的情况下，节省了近 60%的训练时间。文献[17]和文献[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。由于指针网络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指针网络和 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 的网络用户配对的问题，本文提出了基于改进指针网络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，使用改进的指针网络选择配对用户。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，本文对 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。

2 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 $u \in \boldsymbol{U} = \{1, 2, \ldots, N$ ，在 OMA 的情况下用户的信噪比(Signal Noise Ratio, SNR)为

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

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

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

g **为瑞利衰落信道系数**，β_0 表示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

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

其中 0.5 表示 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 $i \in \{0,1,..., N/2\}$ 表示第 i 对用户对，每个用户对的两个用户为 $s_i, w_i \in \boldsymbol{U}$ ，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

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

其中 $\tilde{\gamma}_{s_i}$ 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，根据公式(4)可以观察到，当不完美系数 $\beta \neq 0$ 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。

根据公式(1)和公式(4)可以得到

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

其中，$\gamma_{s_i}, \gamma_{w_i}$ 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

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

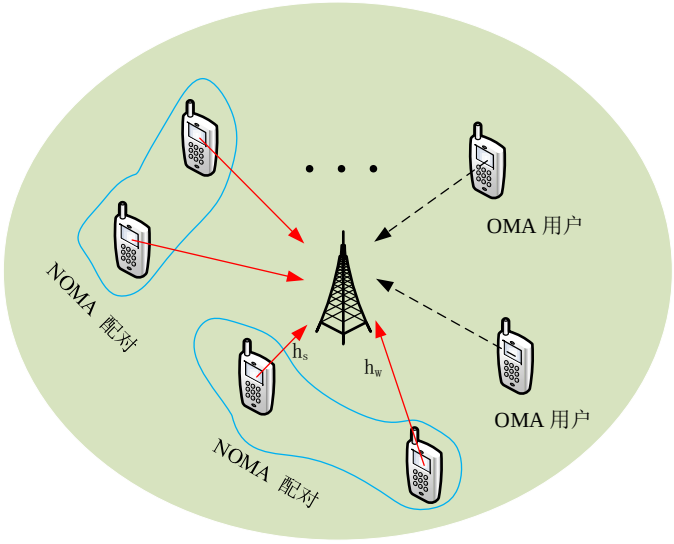


图 1 系统模型图

Fig. 1 System model

对于每个用户对，NOMA 方式的 ASR 表示为 $R_i^{NO}(\alpha) = \tilde{R}_{s_i}(\alpha) + \tilde{R}_{w_i}(\alpha)$，OMA 方式的 ASR 表示为 $R_i^O = R_{s_i} + R_{w_i}$，$R_{s_i}, R_{w_i}$ 为对应用户采用 OMA 接入的可达速率。定义向量 $\boldsymbol{y} = [y_1, y_2 ..., y_{N/2}]$，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

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

根据公式(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^N 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

$$\begin{aligned} \text{P0}: \quad & \max_{\alpha, y_i, \pi} R_{sum}(\alpha) \\ s.t. \quad & 0 \le \alpha \le 1 \qquad (a) \\ & y_i \in \{0,1\} \qquad (b) \\ & |h_i^s| > |h_i^w| \qquad (c) \end{aligned} \tag{8}$$

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(a)表示 NOMA 用户功率分配因子的约束，(b)表示用户的 NOMA 和 OMA 两种接入方式，(c)表示强用户信道增益必须大于弱用户。

3 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$\tilde{R}_{s_i} \ge R_{s_i}$。根据公式(5)和公式(6)得

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

再根据公式(9)可以解得

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

公式(10)表明当功率分配因子$\alpha > \alpha_L$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$\tilde{R}_{w_i} \ge R_{w_i}$，同样可以得到类似的表达

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

再根据(11)可以得到功率分配因子的上界

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

公式(12)表明当功率分配因子$\alpha < \alpha_U$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，$\alpha_U > \alpha_L$都能得到满足；但对于不完美的 SIC，$\alpha_U > \alpha_L$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$\tilde{R}_{s_i} \ge R_{s_i}, \tilde{R}_{w_i} \ge R_{w_i}$。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$\alpha_U > \alpha_L$根据公式(10)和公式(12)可以得到

$$\beta < \frac{(\gamma_{w_i}-\sqrt{1+\gamma_{w_i}}+1)-\alpha_L\gamma_{w_i}}{\gamma_{s_i}(\sqrt{1+\gamma_{w_i}}-1)} \triangleq \beta_L \tag{13}$$

公式(13)表示当$\beta < \beta_L$时$\alpha_U > \alpha_L$成立，此时一定存在$\alpha^* \in [\alpha_L, \alpha_U]$使得$\tilde{R}_{s_i} \ge R_{s_i}, \tilde{R}_{w_i} \ge R_{w_i}$同时满

足，因此可以认为公式(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 $\beta = \beta_{th}$ 为固定值，这由设备本身决定。观察公式(13)和公式(10)，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 $\beta_{th} < \beta_L$ 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 $\alpha^* = \alpha_U$。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^N 是在 $\alpha_U > \alpha_L$ 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。

4 用户配对算法

上一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 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，但这种关系是需要一定时间训练得到的，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。也就说 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，本文之间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。

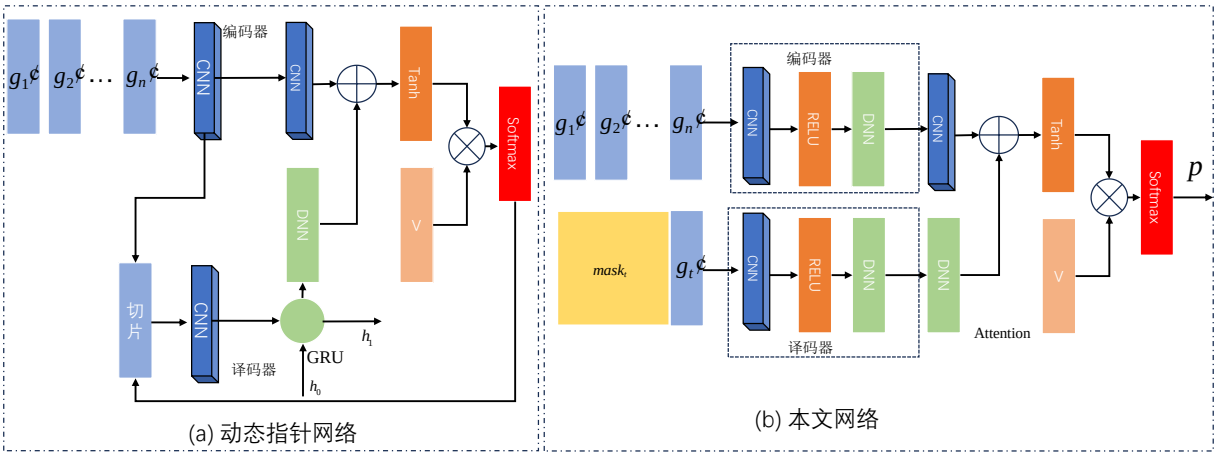


图 2 网络对比图

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

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

其中 $n = 1, 2, ..., N$，γ_k 表示第 n 个用户的 SNR。

图 2(b)中 Attention 机制采用了 MLP 的相关度计算方法，具体计算方法为

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

公式(15)中 $\boldsymbol{V} \in \mathbb{R}^{1\times d}, \boldsymbol{W}_{ref}, \boldsymbol{W}_q \in \mathbb{R}^{d\times d}, u \in \mathbb{R}^{1\times N}$，$\boldsymbol{W}_{ref}, \boldsymbol{W}_q$ 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。$\boldsymbol{q} \in \mathbb{R}^{d\times 1}$ 为译码器的输出，$\boldsymbol{ref} \in \mathbb{R}^{d\times N}$ 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

$$u_n = \begin{cases} u_n = u_n, mask_{t,n} = 0 \\ u_n = -\infty, mask_{t,n} = 1 \end{cases} \tag{16}$$

则配对用户的概率分布为

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

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

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

其中 $\boldsymbol{g}_0 = \boldsymbol{q}$，$\boldsymbol{W}_{ref}^{\,l}, \boldsymbol{W}_q^{\,l}, \boldsymbol{V}^l$ 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 $\boldsymbol{\Gamma} = \{\gamma_1, .., \gamma_n, ., \gamma_N\}$。定义 $\boldsymbol{mask}_t \in \mathbb{R}^{1\times N}$ 向量表示 t 时刻已配对的用户，$mask_{t,n} = 1$ 表示第 n 个用户已配对，$mask_{t,n} = 0$ 表示第 n 个用户未配对。$mask_t$ 有两个作用，一是指示当前环境的用户配对情况，二是防止用户多次被选择。令初始状态 $\boldsymbol{S_1} = \{\gamma_1, .., \gamma_n, ., \gamma_N, \gamma_1, \boldsymbol{mask_1}\}$，$\boldsymbol{S_1}$ 中最后一个元素 γ_1 表示当前配对的用户为第一个用户。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， $\boldsymbol{S_t}=\{\gamma_1,..,\gamma_n,..,\gamma_N,\gamma_t,\boldsymbol{mask_t}\}t=0,1,..N/.$。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 $a_t \in \{0,...,N-1\}$ 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 $r_t=(1-y_i)R_i^O+y_iR_i^{NO}(\alpha)$，一个回合的总奖励计算方式如公式(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 $\boldsymbol{S_1}=\{\gamma_1,..,\gamma_n,.,\gamma_N,\gamma_t,\boldsymbol{mask_t}\}$；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(13)计算 β_L，若 $\beta_L>\beta$ 则采用 NOMA 接入。设图 3 例子中的 $\beta=0.01$，则第一对用户和第二对用户采用 NOMA 接入，第三对用户采用 OMA 接入。确定配对用户和接入方式后利用公式(10)和公式(12)计算功率取数值 $\alpha^*=\alpha_U$ 作为 NOMA 用户的发射功率。

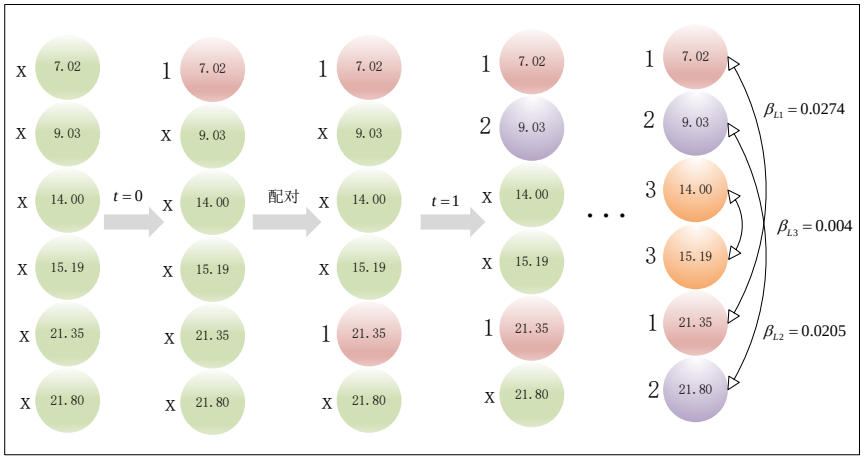


图 3 配对流程图

Fig. 3 Pairing process diagram

本文采用的强化学习算法训练所提网络，具体算法结合了 Rollout Baseline[21]和近端梯度优化算法(Proximal Policy Optimization, PPO)。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)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 的方法替换 PPO 算法中的 Critic 具体如图 4 所示，且将优势函数更换为

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

其中 R_{sum}^A 为当前回合 Actor 获得的总奖励， R_{sum}^C 为 Rollout Baseline 方法的总奖励，若 $adv<0$ 则说明

当前 Actor 表现不如 Critic，Actor 应该像 Critic 学习，反之 $adv>0$ 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(19)计算优势函数，此时 PPO 算法 Actor 的损失函数可以表示为

$$J_\theta = -\min\left(\frac{p_{\theta now}}{p_\theta}, clip\left(\frac{p_{\theta now}}{p_\theta}, 1-蝌1+\right)\right)*adv-\xi entropy \tag{20}$$

其中

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

表示当新策略与旧策略的相识程度，$entropy$ 为概率分布 $p_{\theta now}$ 的熵，ξ 为熵的权重系数。Critic 按 Rollout Baseline 方法更新，即当优势函数均值 $\overline{adv}>0$ 时对优势函数进行单边 T 检验，若 T 检验的 $pvalue$ 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。

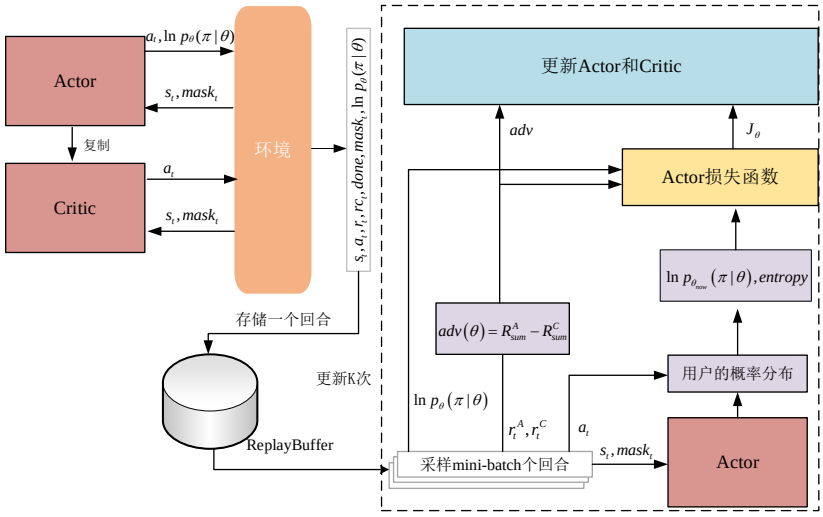


图 4 RPPO 训练算法框图

Fig. 4 RPPO Training Algorithm Block Diagram

所提配对方法的具体训练流程如表 1 所示

表 1 RPPO 训练算法

Tab. 1 RPPO Training algorithms

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

1: **for** episode=1,...,E **do**

2: **while** not done

3: 执行 Actor 网络得到可选用户的概率分布

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

5: Actor 环境执行动作选择用户进行配对，根据(10)、(12)和(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

6: 执行 Critic 网络得到可选用户的概率分布

7: 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(10)、(12)和(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(19)计算优势函数 adv
for k= $1,...,K$
随机采样 mini-batch 个回合的数据
随机采样的数据带入 Actor 得到当前参数下的对数概率和熵 $\ln p_{\theta_{now}}(\pi|\theta)$， $entropy$
根据(20)计算损失函数
Adam 优化器更新 actor 网络参数 θ
end for
if $mean(adv) > 0$ && OneSidedPairedTTest(adv,0)>0.05 **then**
$\theta_b = \theta$
end if
end for

5 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 $\sigma^2 = -100$ 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 $\beta_{th} = 0.001$ 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 GPU 仅在训练阶段使用，测试阶段采用 CPU。

图 5 为所提算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所提方案收敛后性能明显比 DQN 要好；DPN+RolloutBaseline 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，但相比于本文所提算法收敛速度要慢，因此本文所提方案能有效提高网络收敛速度和用户 ASR。

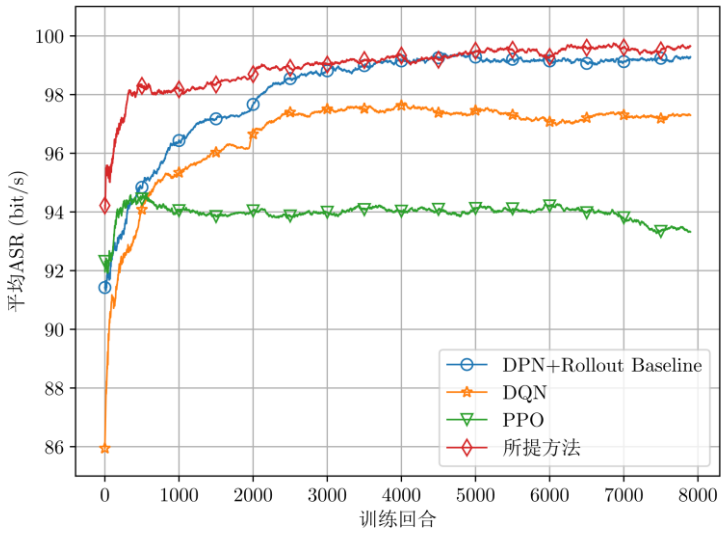
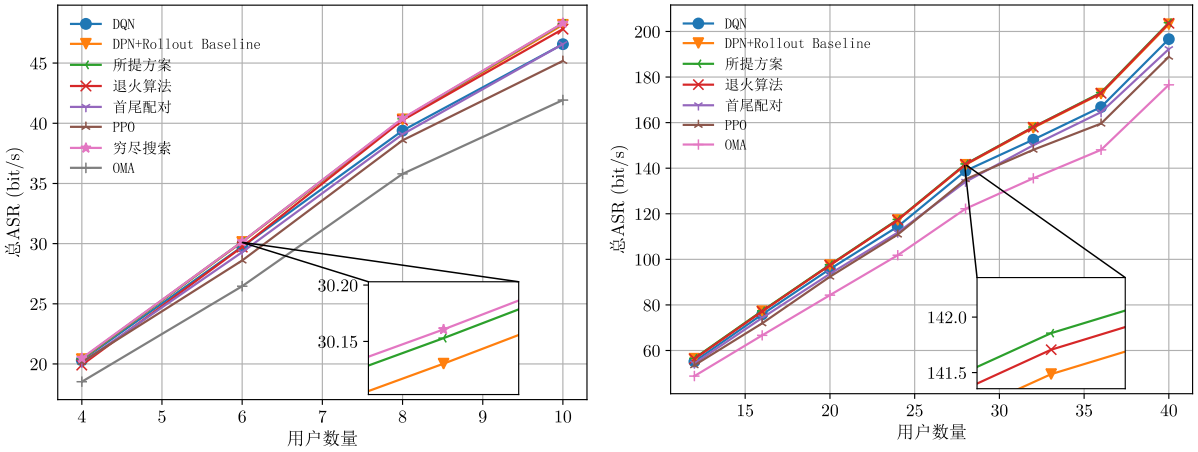


图 5 学习曲线图

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 $\alpha^* = \alpha_U$。对于所有方案来说，随着用户的数量的增加总 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，本文所提方案明显优于首尾配对；与推测的一样由于文献[14]所提的 DQN 算法优于只能在组内配对，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 算法和本文所提的方案十分接近穷搜算法，本文所提的方案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+RolloutBaseline 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，而所提方案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+RolloutBaseline 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，而本文所提方案仍然优于退火算法。



(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. 5 Relationship between User Count and Total ASR

虽然 DPN 和本文所提的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+RolloutBaseline 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，所提方案与退火算法相对误差增大，但所提方案相比 DPN+RolloutBaseline 算法误差要小，因此所提算法相比 DPN+RolloutBaseline 算法更稳定，用户数量泛化能力更好。

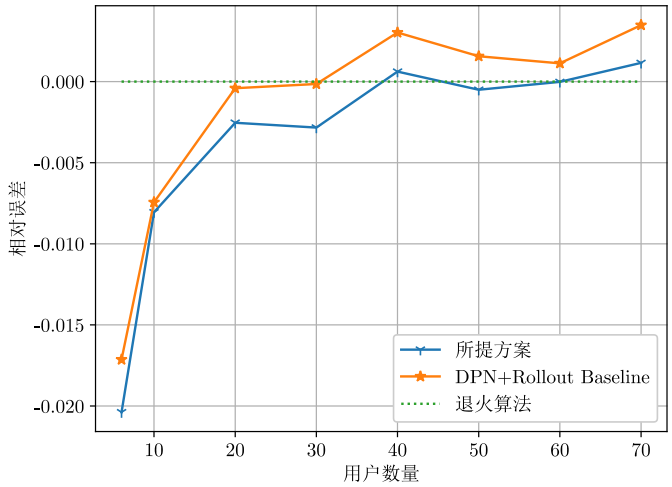


图 7 随着用户数量增加相对误差对比

Fig. 6 Comparison of relative errors as the number of users increases

图 8 给出本文提出的算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 的问题网络执行速度较慢，而所提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%。

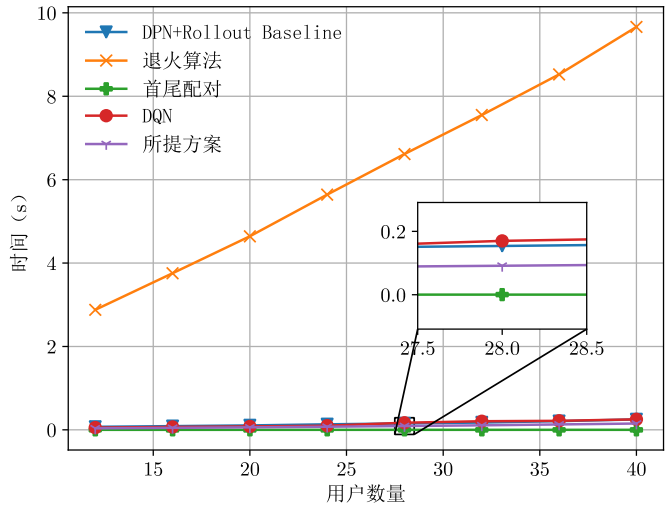


图 8 算法时间对比图

Fig. 7 Comparison of Algorithm Time

图 9 为不完美系数变化对比图，从图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 $\beta_{th}=0.001$ 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，而所提方案由于在线学习的优势能很快的得到较好的性能。而当不完美系数变化较大时，如 $\beta_{th}=0.008$ 时可以看到所提方案需要更多的训练回合才能达到接近退火算法的性能。

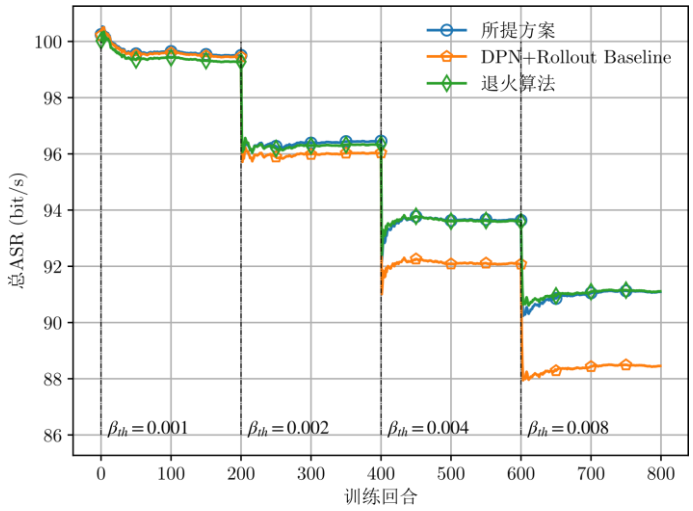


图 9 不完美系数变化对比图

Fig. 8 Comparison of Imperfect Coefficient Changes

6 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.
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.

[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, 2022(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]//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]. 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]//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[C]//Cortes C, Lawrence N, Lee D, et al. Advances in Neural Information Processing Systems: Vol. 28. Montreal, Canada: Curran Associates, Inc., 2015.

[17] Bello I, Pham H, Le Q V, et al. Neural combinatorial optimization with reinforcement learning[C]//International Conference on Learning Representations. Toulon, France, 2017.

[18] Nazari M, Oroojlooy A, Snyder L, et al. Reinforcement Learning for Solving the Vehicle Routing Problem[C]//Bengio S, Wallach H, Larochelle

H, et al. Conference and Workshop on Neural Information Processing Systems: Vol. 31. Montreal, Canada: Curran Associates, Inc., 2018.

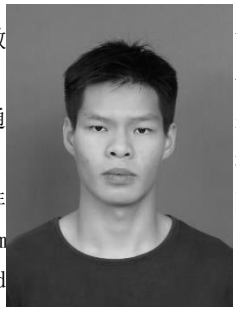
[19] Ding Z, Peng M, Poor H V. Cooperative Non-Orthogonal Multiple Access in 5G Systems[J]. IEEE Communications Letters, 2015, 19(8): 1462-1465.

[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.

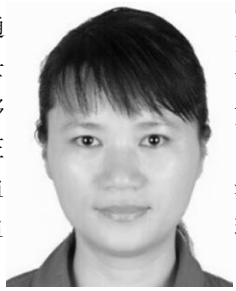
[21] Kool W, Hoof H van, Welling M. Attention, Learn to Solve Routing Problems![C]//International Conference on Learning Representations. New Orleans, America, 2019.

作者简介：

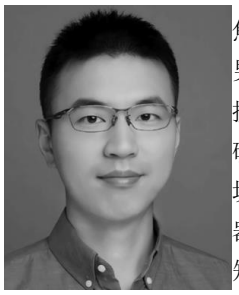
李国鑫（1990-），男，博士，副教授，硕士生导师，研究方向：协作通信、认知无线电、无线携能通信和非正交多址技术。E-mail:guoxin@aeu.edu.cn



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。