User Clustering Scheme for Downlink Hybrid NOMA Systems Based on Genetic Algorithm

HANLIANG YOU, ZHIWEN PAN, NAN LIU, AND XIAOHU YOU

1National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
2Purple Mountain Laboratories, Nanjing 211100, China
Corresponding author: Zhiwen Pan (pzw@seu.edu.cn)

This work was supported in part by the National Key Research and Development Project under Grant 2018YFB1802402, and in part by the European Union’s Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie Grant (TESTBED2 Project) under Agreement 872172.

ABSTRACT Non-orthogonal multiple access (NOMA) is considered to be a promising technology for improving bandwidth utilization efficiency and reducing power consumption. In this paper, we consider the optimization of user clustering in the NOMA scenario, where the goal is to maximize the system total throughput under minimum rate constraints. Different from most existing literatures, there is no limit on the number of users in each cluster. Simulation results show that the proposed scheme can significantly reduce computational complexity and have a better performance compared with schemes based on the other heuristic algorithms and random user clustering with greedy strategy.

INDEX TERMS Non-orthogonal multiple access (NOMA), user clustering, genetic algorithm (GA), power allocation.

I. INTRODUCTION

With the dramatic increase in demand for wireless traffic, non-orthogonal multiple access (NOMA) technology is considered a better alternative to orthogonal multiple access (OMA) in next-generation mobile communication system [1], [2]. Compared with OMA, NOMA system where different users can simultaneously utilize the same resource has a significant improvement in achievable rate and energy efficiency, which can effectively solve the congestion problem in the massive device access scenario [3], [4].

NOMA can be categorized into power-domain (PD) NOMA [5]–[7], and code-domain (CD) NOMA [8]. In this paper, we only consider PD-NOMA. PD-NOMA allows multiple users to communicate simultaneously on the same subcarrier. At receivers, successive interference cancellation (SIC) is applied [9]. However, in multi-user NOMA scenarios, SIC is limited by device complexity and error propagation. Therefore, the concept of hybrid NOMA has been proposed [10]. In hybrid NOMA system, users are divided into many clusters, the resources of each cluster are orthogonal to each other, and different users in the same cluster are allocated different power and share the same resources within the cluster. Therefore, the performance of the hybrid NOMA system is closely related to the user clustering strategy.

User clustering in NOMA system attracted much research enthusiasm recently. In [11], the authors solve the user clustering problem under the perspective of achieving rate fairness by using sequential convex approximation method and find the local optimal solution. In [12], the authors propose two user clustering methods based on two heuristic neighborhood search algorithms, which can give an approximate optimal solution. Zhu et al. [13] propose an optimal user clustering scheme of the even user scenario. They first prove the optimality in the scenarios where there are only 2 and 4 users, and then extend the results to the cases with any even number of users.

Joint optimization of user clustering and other related issues, such as power allocation, are investigated in some literatures. Sun et al. [14] incorporate NOMA into full-duplex systems to solve the joint power and subcarrier allocation problem via monotonic optimization and successive convex approximation. Ding et al. [15] compare the impact of user clustering on system performance under two different power allocation strategies. In [16], the authors use reinforcement learning to solve the power allocation problem in...
a PD-NOMA system to maximize total energy efficiency. Zeng et al. [17] propose a joint user clustering and power allocation scheme in hybrid NOMA system based on swap matching theory, and give a suboptimal version with low complexity. Shao et al. [18] apply some of the results of [13] to the Internet of Things (IoT) NOMA scenario, then propose a power allocation scheme within each cluster and a dynamic joint optimization algorithm. However, most of these studies are conducted with only two users in each cluster. It is pointed out in [19] that multi-user NOMA still has considerable performance improvement, but few works focus on user clustering in this scenario. A suboptimal solution to the user clustering problem in multi-user NOMA scenario is proposed in [20] and Tsai and Wei [21] extend it to two scenarios with minimum service rate constraint and dynamic service rate constraint, respectively. However, both studies require that the number of users in each cluster is the same, which is not practical.

In this paper, we focus on the user clustering optimization to maximize the system total throughput in multi-user hybrid NOMA scenario, where minimum rate constraints are introduced to relatively ensure fairness among users. There is no limit on the number of users in each cluster, i.e., each cluster can contain one or more users, and the number of users in different clusters is not required to be the same. Genetic algorithm (GA) is employed to solve this problem [22], [23]. Firstly, a user clustering algorithm based on GA is proposed. Then we prove that for a specific scenario, the optimal user clustering result can be given directly. Finally, a power allocation strategy that can satisfy the minimum rate constraint is presented.

Simulation results show that the complexity of the proposed scheme is much lower than that of exhaustive search. It can approach an approximate optimal solution for any given number of users and clusters, the performance of the proposed algorithm in system total throughput is very close to the optimal result, which is superior to two heuristic algorithms in [12] and the random clustering algorithm based on greedy strategy.

The rest of this paper is organized as follows. In Section II, the system model and the optimization problem of hybrid NOMA system is given. In Section III, a heuristic algorithm based GA is proposed to solve the user clustering problem under minimum data rate constraints. Then we prove that in the case where there is two or one user in each cluster, the optimal result of above problem can be given directly without iterative calculations. The simulation results and the conclusion of this paper are presented in section IV and section V, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION
A. SYSTEM MODEL
Consider a downlink NOMA system as shown in Figure 1. There are $M$ users randomly distributed in a service area $D$, base station (BS) is at the center of $D$. The BS and users are all equipped with single antenna. $M$ users are divided into $K$ clusters, and the number of users in each cluster is not necessarily the same. $C_k$ denotes the set of users in the $k$-th cluster, and $|C_k|$ denotes the number of users in this cluster. Users in the same cluster can share the same resources, while resources occupied by users in different clusters are orthogonal to each other. In the $k$-th cluster, the received signal at user-$m$ is

$$y_m = \sum_{i=1}^{[C_k]} h_{mi} s^k_i + n_m$$

(1)

where $h_m$ is the channel gain between the BS and user-$m$, which is assumed to follow complex-normal distribution $CN(0, d_m^{-v})$, where $v$ is the large scale fading path loss exponent and $d_m$ is the distance between the BS and user-$m$. $p^k_i$ and $s^k_i$ represent the power and transmitted signal of the $i$-th user in this cluster, respectively. And $n_m$ is the additive white Gaussian noise (AWGN) with the two-sided power spectral density $N_0/2$. Additionally, assume that the total transmit power of the BS is $P_t$ and the power resource allocated by each cluster is equal, hence we have

$$P_t/K \geq p^k_1 + p^k_2 + \cdots + p^k_{|C_k|}$$

(2)

Under the premise that there is no relative movement between the users and the BS, we assume that the users’ channel state will not change for a period of time and the BS can obtain all users’ channel state information without error. Without loss of generality, assume $|h_1|^2 \geq |h_2|^2 \geq \cdots \geq |h_M|^2$. According to [24], in the same cluster, the user with better channel condition will always be allocated less power than the user with poor channel condition. The throughput of user-$m$ is given by (3).

$$R_m = \frac{B}{K} \log_2\left(1 + \frac{P_m h_m}{\sum_s p_s h_{ms} + B n_s K}\right)$$

(3)

where $s \in C_k$ and $|h_s| \geq |h_m|$, $B$ denotes the total bandwidth of this system. The total throughput of the system can be calculated by (4).

$$R_{sum} = \sum_{m=1}^{M} R_m$$

(4)
B. PROBLEM FORMULATION

The user clustering problem in downlink hybrid NOMA system is presented as follows:

\[
\begin{align*}
\max_{x_{i,j}, p_i} & \quad \sum_{j=1}^{K} \sum_{i=1}^{M} x_{i,j} \frac{B}{K} \log_2 \left( 1 + \frac{p_i h_i}{\text{Inter} + \frac{B}{K} n_0} \right) \\
\text{s.t.} & \quad C1: p_i \geq 0 \quad i = 1, 2, \ldots, M \\
& \quad C2: \sum_{i=1}^{M} p_i \leq P_t \\
& \quad C3: x_{i,j} (p_i - p_j) \geq 0 \quad l < i, j = 1, 2, \ldots, K \\
& \quad C4: R_i \geq R_{i,\text{OMA}} \quad i = 1, 2, \ldots, M \\
& \quad C5: \sum_{j=1}^{K} x_{i,j} = 1 \quad i = 1, 2, \ldots, M \\
& \quad C6: x_{i,j} \in \{0, 1\} \quad i = 1, 2, \ldots, M \quad j = 1, 2, \ldots, K
\end{align*}
\]

(5)

where \(\text{Inter} = \sum_{i=1}^{K} x_{i,j} \sum_{i=1}^{M} x_{i,j} p_i h_i\), \(x_{i,j}\) is a binary variable representing the affiliation of user and cluster.

\[
x_{i,j} = \begin{cases} 
1, & \text{user } i \text{ in cluster } j \\
0, & \text{otherwise}
\end{cases}
\]

(6)

Constraint C1 is to ensure that each user’s power is a nonnegative number. Constraint C2 is to ensure that the sum of power does not exceed the upper limit of the BS power. Constraint C3 follows the basics of the NOMA system, which allocates more power to users with worse channel conditions. Constraint C4 is the minimum rate constraint for each user to relatively ensure fairness, \(R_{i,\text{OMA}}\) is the throughput of user-\(i\) in OMA system. Constraint C5 is to ensure that each user can only belong to one cluster.

Problem (5) is a mixed integer nonlinear programming (MINLP) problem. Due to the existence of the parameter \(x_{i,j}\), this problem is a non-convex problem. The complexity of finding the optimal solution directly is unacceptable. To solve the problem, a heuristic algorithm based on GA is proposed.

III. USER CLUSTERING SCHEME BASED ON GA

In this section, GA is employed to solve Problem (5), and the execution of GA needs to repeatedly perform power allocation in each cluster. Firstly, a GA based user clustering scheme is proposed under the assumption that users’ power is known. Then, a power allocation strategy that satisfies the minimum power constraint is found by linear programming within each cluster.

When the power allocation strategy is determined, Problem (5) is degenerated as follows:

\[
\begin{align*}
\max_{x_{i,j}} & \quad \sum_{j=1}^{K} \sum_{i=1}^{M} x_{i,j} \frac{B}{K} \log_2 \left( 1 + \frac{p_i h_i}{\text{Inter} + \frac{B}{K} n_0} \right) \\
\text{s.t.} & \quad C1: \sum_{j=1}^{K} x_{i,j} = 1 \quad i = 1, 2, \ldots, M \\
& \quad C2: x_{i,j} \in \{0, 1\} \quad i = 1, 2, \ldots, M \quad j = 1, 2, \ldots, K
\end{align*}
\]

(7)

Due to the existence of the parameter \(x_{i,j}\), Problem (7) is still a non-convex MINLP problem. Hence, GA is employed to solve it to find the best performing user clustering scheme for any given \(M\) and \(K\). GA mainly include bio-inspired operators such as mutation, crossover and selection. The flow chart of GA is shown in Figure 2. The evolution usually begins with a population of randomly generated individuals, which is an iterative process, and the population produced by each iteration is called a generation. Individuals with better adaptability are stochastically selected with greater probability and then form a new generation through recombination and mutation. The new generation of candidate solutions is then used in the next iteration of the algorithm. In each generation, the fitness of every individual in the population is evaluated; in this paper the fitness function is the value of the objective function in Problem (7).

Before the execution of GA, the first thing is to choose an appropriate encoding strategy. The encoding strategy maps specific clusters into abstract sequences for calculation and directly affects the subsequent operator selection and algorithm convergence speed. Here, we assume that \(M\) users are sorted in descending order of channel gain, the user with the largest channel gain is denoted by \(u_1\), the user with the second largest channel gain is denoted by \(u_2\), and so on. Then arrange these users randomly, and there are a total of \(M!\) different combinations, each combination maps a specific user clustering as (8), as shown at the bottom of the next page, where \(u_i^k\) is the user assigned to the \(i\)-th position in the clustering, and we have \(\sum_{i=1}^{K} |C_i| = M\). Following the concept of genetics, \(I^k\) can also be called the genotype of the \(k\)-th individual in the current population. The arrangement order in \(I^k\) will change on subsequent crossover and mutation, then \(I^k\) will map a new kind of user clustering.
After having a specific encoding strategy, we then randomly generate a set of potential solutions, which is called the initial population $G_0$. The dimension of $G_0$ is a hyper-parameter, which is also called population size. Another hyper-parameter that needs to be set empirically is the maximum number of generations, which represents the number of iterative calculations. In each generation, we use roulette wheel selection to bring individuals who perform well on the fitness function into the next generation. And we introduce an elitist selection mechanism to guarantee that the best individual in the current population is not lost. In this way, each generation of population will perform better and better on the fitness function. But the best performing individual is always the same, which does not help us find the optimal solution, so we need crossover and mutation operations.

The crossover in GA follows the concept of chromosome intersection in genetics, which changes the encoding structure of two parent individuals to generate child individuals, and the child individuals will map a new user clustering. Here we adopt a partial matching crossover strategy. Randomly select two individuals $I^m$ and $I^n$ as parent individuals in the current population, and then select a certain length in their encodings as the matching part as (9), as shown at the bottom of the page, where the part in bold (including the users between the first and last users in bold) is the matching part, and then we exchange the users that are not in the matching part to obtain the encodings of the child individuals $I^m_{new}$ and $I^n_{new}$ as (10), as shown at the bottom of the page. When users repeat in child individuals, keep the matching part unchanged and adjust the repeated users which are not in the matching part according to the one-to-one correspondence in matching part until there is no repetition. In the process of generating child individuals, there is a given probability that the mutation will occur. Specifically, randomly select two positions in the encoding sequence and exchange users in these two positions, then we can obtain an individual with a new encoding sequence. The probability of mutation cannot be too large, but in order to prevent the algorithm from frequently falling into the local optimization, the probability of mutation will be increased when the current result does not change for a period of time. The elitist selection mechanism mentioned above can protect the algorithm from degenerating to a random search due to a high probability of mutation.

The proposed GA based algorithm will stop when satisfying the given accuracy requirements or reaching the maximum number of iterations. Then, the individual with the largest fitness function value in the last population will be decoded, and the clustering result obtained by decoding is considered to be an approximate optimal solution of Problem (5). The detailed algorithm is described in Algorithm 1.

**Algorithm 1 User Clustering Based on GA**

1. Initialize population size and set termination condition;
2. Randomly generate the initial population $G_0$, and set the generation counter $t=0$;
3. **repeat**
   4. Calculate fitness function value for the individuals in current population;
   5. Keep the best individual directly to the next generation, then perform the selection operation;
   6. Perform crossover and mutation operations;
   7. Update the generation counter $t=t+1$;
8. **until** the termination condition is reached
9. Decode the best individual in the last population as the solution to the user clustering problem.

By Algorithm 1, we can obtain an approximate optimal user clustering result for any given $M$ and $K$. Limited by computational complexity, the scenario with a small number of users in each cluster has certain application value. We have already known that in the 2-user hybrid NOMA scenario, the optimal user clustering result is determined [13], [18]. Based on this, we find that in the hybrid NOMA scenario with two or one user in each cluster, the optimal clustering scheme is also determined. In this case, we have $K \leq M \leq 2K$, and Problem (7) can be transferred to the following one:

$$\max_{x_{ij}} \sum_{j=1}^{K} \sum_{i=1}^{M} x_{ij} B \log_2 (1 + \frac{p_i h_j}{\eta B n_0})$$

subject to:

C1: $\sum_{j=1}^{K} x_{ij} = 1 \quad i = 1, 2, \ldots, M$

C2: $x_{ij} \in \{0, 1\} \quad i = 1, 2, \ldots, M \quad j = 1, 2, \ldots, K$

C3: $\sum_{i=1}^{M} x_{ij} \leq 2 \quad j = 1, 2, \ldots, K$

**Theorem 1:** The optimal user clustering result is to make the user with better channel condition monopolize a cluster

$$I^K = [(u^k_i, \cdots, u^k_i_{|C_1|}), (u^k_i_{|C_1|+1}, \cdots, u^k_{i|C_1|+|C_2|}), \cdots, (u^k_{|M|-|C_K|}, \cdots, u^k_{M})]$$

(8)

$$I^m = [(u^m_i, \cdots, u^m_i_{|C_1|}), (u^m_i_{|C_1|+1}, \cdots, u^m_{i|C_1|+|C_2|}), \cdots, (u^m_{|M|-|C_K|}, \cdots, u^m_{M})]$$

(9)

$$I^n = [(u^n_i, \cdots, u^n_i_{|C_1|}), (u^n_i_{|C_1|+1}, \cdots, u^n_{i|C_1|+|C_2|}), \cdots, (u^n_{|M|-|C_K|}, \cdots, u^n_{M})]$$

$$I^m_{new} = [(u^m_i, \cdots, u^m_i_{|C_1|}), (u^m_i_{|C_1|+1}, \cdots, u^m_{i|C_1|+|C_2|}), \cdots, (u^m_{|M|-|C_K|}, \cdots, u^m_{M})]$$

(10)
as much as possible, and then cluster the remaining users according to the optimal user clustering result of 2-user hybrid NOMA.

Proof: Consider a system with only three users, and we have \( h_1 > h_2 > h_3 \). \( R_{1+2,3} \) denotes the system total throughput when user-1 belongs to the first cluster and the other two users share the second cluster. If we can prove:

\[
R_{1+2,3} > R_{2+1,3} > R_{3+1,2}
\]  
(12)

then we can extend it to multiple users to prove Theorem 1.

First we prove the first inequality in (12), which can be rewritten as:

\[
\log_2 \left( \frac{1 + \alpha p h_1}{n_1} \right) + \log_2 \left( \frac{1 + \alpha p h_2}{n_0} \right) + \log_2 \left( \frac{1 + \alpha p h_3}{n_0} \right) > \log_2 \left( \frac{1 + \alpha p h_1}{n_0} \right) + \log_2 \left( \frac{1 + \alpha p h_2}{n_0} \right) + \log_2 \left( \frac{1 + \alpha p h_3}{n_0} \right)
\]  
(13)

where \( p \) is the total power of each cluster, \( \alpha \) the power allocation index in the cluster which have two users. And we assume the bandwidth of each cluster is 1, after simplification we have:

\[
\log_2 \left( \frac{1 + \alpha p h_1}{n_0} \right) + \log_2 \left( \frac{1 + \alpha p h_2}{n_0} \right) + \log_2 \left( \frac{1 + \alpha p h_3}{n_0} \right) > 0
\]  
(14)

According to the characteristics of logarithmic function, the proof of (14) can be transformed into the solution of Problem (15). If the minimum value of the objective function in Problem (15) is greater than or equal to 0, then it can be proven that above inequality is always true.

\[
\min_{\{h_i\}} \quad (n_0 + h_1)(n_0 + \beta h_2)(n_0 + \beta h_3)
\]

\[
-(n_0 + \beta h_1)(n_0 + h_2)(n_0 + \beta h_3)
\]

s.t.

\[
h_1 - h_2 > 0
\]

\[
h_3 - h_2 > 0
\]

\[
h_1 > 0
\]

\[
h_2 > 0
\]

\[
h_3 > 0
\]

Here we use the Lagrangian multiplier method and KKT conditions to obtain the Lagrangian function and the unconstrained optimization problem corresponding to Problem (15), where \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are Lagrangian multipliers.

\[
L = (n_0 + h_1)(n_0 + \beta h_2)(n_0 + \beta h_3) - (n_0 + \beta h_1)(n_0 + h_2)(n_0 + \beta h_3) + \lambda_1(h_2 - h_1) + \lambda_2(h_3 - h_2) - \lambda_3 h_3
\]

\[
- \beta^2 n_0 h_3 + \beta n_0 h_3 - \beta n_0^2 + n_0^2 = \lambda_1
\]

\[
\beta^2 n_0 h_3 - \beta^2 n_0 h_3 + \beta n_0^2 - n_0^2 + \lambda_1 = \lambda_2
\]

\[
\beta^2 n_0 h_2 - \beta^2 n_0 h_1 + \beta n_0 h_1 - \beta n_0 h_2 + \lambda_2 = \lambda_3
\]

\[
\lambda_1(h_2 - h_1) = 0
\]

\[
\lambda_2(h_3 - h_2) = 0
\]

\[
\lambda_3 h_3 = 0
\]

\[
\lambda_1, \lambda_2, \lambda_3 \geq 0
\]

(16)

By solving the equation set in (17), we find that the points satisfying the KKT constraints are greater than or equal to 0 in the objective function of Problem (15), which means that (13) and (14) are always true. Thus the proof of the first inequality in (12) is completed and the second inequality can be proved by a similar method. Then we extend the results to the scenario with multiple users, that is, to make the users with good channel conditions monopolize one cluster as much as possible, and then arrange the remaining users to 2-user clusters. The system throughput obtained by such a user clustering scheme is always optimal.

Next, a power allocation strategy that satisfies the fairness constraints is introduced. Since the power allocation is not affected by users in other clusters, without loss of generality, we only focus on the problem within cluster-\( k \). Then Problem (5) is degenerated as follows:

\[
\max_{p_i, i \in C_k} \frac{B}{K} \sum_{i=1}^{M} x_i k B \log_2 \left( 1 + \frac{p_i h_i}{\sum_{s=1}^{i-1} x_s k p_s h_i + \frac{B}{K} n_0} \right)
\]

s.t. \( C_1: p_i \geq 0 \), \( i = 1, 2, \ldots, M \)

\( C_2: \sum_{i=1}^{M} x_i k p_i \leq P_k \)

\( C_3: x_i k h_i(p_i - p_l) \geq 0 \), \( l < i \), \( j = 1, 2, \ldots, K \)

\( C_4: R_i \geq R_i,OMA \), \( i = 1, 2, \ldots, M \), \( i \in C_k \)

(18)

If there is no constraint \( C_4 \), the total throughput can be maximized by directly allocating all power to the user with the best channel condition. Constraint \( C_4 \) prevents this extreme situation from occurring. Therefore, constraint \( C_4 \) is key in solving the problem. For user-\( m \) in cluster-\( k \), constraint \( C_4 \) is

\[
\frac{B}{K} \log_2 \left( 1 + \frac{p_m h_m}{\sum_{s=1}^{m-1} x_s k p_s h_m + \frac{B}{K} n_0} \right) \geq R_m,OMA
\]

(19)

After some simple inequality transformations, we have

\[
p_m \geq \frac{2(KR_m,OMA)}{B} - 1 \sum_{s=1}^{m-1} x_s k p_s + \frac{n_0 B}{h_m K}
\]

(20)

Algorithm 2 Power Allocation Within Each Cluster

1: Sort and number users in one cluster according to the channel gain: \( h_1 > h_2 > \cdots > h_{|C_k|} \);
2: for \( i = 1 \) to \( |C_k| \) do
3: Compute \( p_i \) using (20);
4: end for
5: if \( \sum_{i=1}^{C_k} p_i > P_k/K \) then
6: Raise error: Cannot meet minimum rate constraints.
7: else
8: Compute \( \Delta p = P_k/K - \sum_{i=1}^{C_k} p_i \);
9: Update \( p_i = p_i + \Delta p \);
10: Accept \( \{p_i\} \) as optimal result and end the algorithm;
11: end if

It can be seen that the lower limit of \( p_m \) is a linear function of \( p_s \), which are power of users with better channel conditions in the same cluster. Then Problem (18) becomes a linear programming problem, and we can easily find its solution. The detailed algorithm is described in Algorithm 2.
IV. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed user clustering scheme in the hybrid NOMA system. Locations of the users are randomly and uniformly distributed in the coverage area of the BS. Important simulation parameters are given in Table 1.

First, we evaluated the complexity of the proposed algorithm. The complexity of GA based algorithm depends on the genetic operators, the representation of the individuals and the population, and obviously on the fitness function. Average running time is employed as complexity comparison metric to show the efficiency advantage of proposed scheme compared with exhaustive search (denoted as ES) in TABLE 2. Without loss of generality, it is assumed that the number of users in each cluster is three. In ES, since the number of calculations required increases exponentially with the increase of $M$, the time required to find the optimal solution will also increase exponentially. In addition, the time required for the proposed scheme in the table is conservative. If the requirement of optimal solution is not strict, the time required to find the suboptimal solution will be shorter.

In Table 2, we assume that the number of users in each cluster is equal, but there are many different groupings for any given $M$ and $K$. Figure 3 shows the differences in system total throughput between different groupings when the optimal user clustering result is obtained. It is worth noting that due to the existence of the minimum rate constraints, too many users cannot occupy one cluster at the same time. For example, in Figure 3, when $M=10$ and $K=7$, there is only one feasible grouping, while there are two feasible groupings when $M=10$ and $K=4$. In addition, it can be seen that under the same value of $M$, the smaller value of $K$, the greater total throughput of the system. This is because a smaller value of $K$ means more users participating in communication in the form of NOMA, this performance gains come at the cost of increased complexity of device and algorithm.

In Figure 4, we compare the performance of the proposed user clustering scheme with that of utilizing random user clustering based on greedy strategy, the power allocation strategy used by the two schemes is the same. We can see that when $K = 5$ remains unchanged, the total throughput obtained by the proposed GA-based scheme proposed in this paper is very close to the optimal result. Moreover, it can be seen that compared with OMA system, NOMA system has an obvious improvement in terms of the total throughput. As the number of users in the system increases, this improvement will continue to increase.

Figure 5 compares the performance between the proposed scheme and the two user clustering algorithms in [12], the hill
clustering based algorithm (denoted as HC) and the simulated annealing based algorithm (denoted as SA). We can see that the performance of the proposed GA based scheme is the best, the simulated annealing based algorithm is the second, and the hill climbing based algorithm has the worst performance. The main reason is that the user clustering in hybrid NOMA system is a MINLP problem. Compared with the other two algorithms, the random search ability and the parallel computing ability of GA are more helpful for finding a better result. In addition, as the number of clusters $K$ increases, the total throughput of the system will decrease under the same value of $M$, which is consistent with the content shown in Figure 3.

V. CONCLUSION

In this paper, we investigate the user clustering problem under minimum rate constraints to maximize the system total throughput in the downlink hybrid NOMA system. Considering that user clustering is a MINLP problem, a GA-based heuristic algorithm is proposed to solve it. Simulation results under various scenarios show that the scheme in this paper can find an approximate optimal solution in a short time for any given number of users and clusters. In addition, we have found that in the scenario where there are two or one user in a cluster, the optimal user clustering result can be directly given without iterative calculations. In the future, the optimal solution with a lower computational complexity will continue to be explored and the NOMA user clustering problem in more specific NOMA scenarios will be studied.

REFERENCES

[1] Y. Chen, A. Bayesteh, Y. Wu, B. Ren, S. Kang, S. Sun, Q. Xiong, C. Qian, B. Yu, Z. Ding, S. Wang, S. Han, X. Hou, H. Lin, R. Visoz, and R. Razavi, “Toward the standardization of non-orthogonal multiple access for next generation wireless networks,” IEEE Commun. Mag., vol. 56, no. 3, pp. 19–27, Mar. 2018.

[2] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, “On the performance gain of NOMA over OMA in uplink communication systems,” IEEE Trans. Commun., vol. 68, no. 1, pp. 536–568, Jan. 2020.

[3] M. ShirvaniNemhaddad, M. Dohler, and S. J. Johnson, “Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations,” IEEE Commun. Mag., vol. 55, no. 9, pp. 55–61, Sep. 2017.

[4] L. Song, Y. Li, Z. Ding, and H. V. Poor, “Resource management in non-orthogonal multiple access networks for 5G and beyond,” IEEE Netw., vol. 31, no. 4, pp. 8–14, Jul. 2017.

[5] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” IEEE Commun. Surveys Tuts., vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[6] Y. Pan, C. Pan, Z. Yang, and M. Chen, “Resource allocation for D2D communications underlaying a NOMA-based cellular network,” IEEE Wireless Commun. Lett., vol. 7, no. 1, pp. 130–133, Feb. 2018.

[7] Y. Cui, Z. Qin, F. Cui, G. Y. Li, and J. A. McCann, “Modulation and multiple access for 5G networks,” IEEE Commun. Surveys Tuts., vol. 20, no. 1, pp. 629–646, 1st Quart., 2018.

[8] J. van de Beek and B. M. Popovic, “Multiple access with low-density signatures,” in Proc. IEEE Global Telecommun. Conf. (GLOBECOM), Nov. 2009, pp. 1–6.

[9] P. Patel and J. Holtzman, “Analysis of a simple successive interference cancellation scheme in a DS/CDMA system,” IEEE J. Sel. Areas Commun., vol. 12, no. 5, pp. 796–807, Jun. 1994.
ZHIVEN PAN (Member, IEEE) has been with the National Mobile Communications Research Laboratory, Southeast University, as an Associate Professor, since 2000, and a Professor, since 2004. From 2000 to 2001, he was involved in the research and standardization of 3G, and since 2002, he has been involved in the investigations on key technologies for IMT-A and 5G. He has published over 50 articles recently, and holds over 50 patents. His research interests include self-organizing networks, wireless networking, and radio transmission technology for wireless communications.

NAN LIU (Member, IEEE) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2001, and the Ph.D. degree in electrical and computer engineering from the University of Maryland at College Park, College Park, MD, USA, in 2007. From 2007 to 2008, she was a Postdoctoral Scholar with the Wireless Systems Lab, Department of Electrical Engineering, Stanford University. In 2009, she became a Professor with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing, China. Her research interests include network information theory for wireless networks, SON algorithms for next-generation cellular networks, and energy-efficient communications. She is an Associate Editor of China Communications.

XIAOHU YOU (Fellow, IEEE) has been with the National Mobile Communications Research Laboratory, Southeast University, since 1990, where he was a Professor and a Chair Professor of the Cheung Kong Scholars Program, and served as the Director of the Laboratory. He has contributed over 50 IEEE journal articles and two books in the areas of adaptive signal processing, neural networks, and their applications to communication systems, and holds over 80 patents. His research interests include mobile communication systems, signal processing, and its applications. His current research interests include wireless and mobile communication systems, and modern digital signal processing. He was a Premier Foundation Investigator of the China National Science Foundation. From 1999 to 2002, he was a Principal Expert of the C3G Project, responsible for organizing China’s 3G Mobile Communications Research and Development activities. From 2001 to 2006, he was a Principal Expert of the National 863 FuTURE Project. He has been the Section C hair of the IEEE Nanjing Section, since 2010. He served as the General Co-Chair for the IEEE WCNC 2014.