Resource Allocation and Power Control Policy for Device-to-Device Communication Using Multi-Agent Reinforcement Learning

Yifei Wei[1, *], Yinxiang Qu[1], Min Zhao[1], Lianping Zhang[2] and F. Richard Yu[3]

Abstract: Device-to-Device (D2D) communication is a promising technology that can reduce the burden on cellular networks while increasing network capacity. In this paper, we focus on the channel resource allocation and power control to improve the system resource utilization and network throughput. Firstly, we treat each D2D pair as an independent agent. Each agent makes decisions based on the local channel states information observed by itself. The multi-agent Reinforcement Learning (RL) algorithm is proposed for our multi-user system. We assume that the D2D pair do not possess any information on the availability and quality of the resource block to be selected, so the problem is modeled as a stochastic non-cooperative game. Hence, each agent becomes a player and they make decisions together to achieve global optimization. Thereby, the multi-agent Q-learning algorithm based on game theory is established. Secondly, in order to accelerate the convergence rate of multi-agent Q-learning, we consider a power allocation strategy based on Fuzzy C-means (FCM) algorithm. The strategy firstly groups the D2D users by FCM, and treats each group as an agent, and then performs multi-agent Q-learning algorithm to determine the power for each group of D2D users. The simulation results show that the Q-learning algorithm based on multi-agent can improve the throughput of the system. In particular, FCM can greatly speed up the convergence of the multi-agent Q-learning algorithm while improving system throughput.

Keywords: D2D communication, resource allocation, power control, multi-agent, Q-learning, fuzzy C-means.

1 Introduction

With the development of mobile networks and the Internet of Things (IoT), more and more wireless devices are being applied to our daily life and industrial production. For entertainment devices (such as mobile phones) in a cellular network, users have high requirements for large bandwidth and low latency for the increase of various high bandwidth requirement applications, such as virtual reality (VR) and augmented reality (AR).

[1] Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

[2] Alibaba Cloud Computing, Hangzhou, 311121, China.

[3] Department of Systems and Computer Engineering, Carleton University, Ottawa, K1S 5B6, Canada.

[*] Corresponding Author: Yifei Wei. Email: weiyifei@bupt.edu.cn.

Received: 13 November 2019; Accepted: 01 March 2020.

CMC.doi:10.32604/cmc.2020.09130

In order to improve the wireless bandwidth utilization and network capacity, the industry has proposed device-to-device (D2D) [Doppler, Rinne, Wijting et al. (2009); Bello and Zeadally (2016)] technology in recent years. Different from the traditional client-to-server (C/S) working mode, the D2D users can communicate directly under the control of the base station and can also indirectly transfer data to the base station through the other D2D devices. In a traditional cellular network, after a certain channel under one base station is occupied, it cannot be reused by other devices under the same base station again. However, in the D2D scenario, since the D2D pairs are relatively close together, the D2D pair can reuse the channel with very small power which will improve the frequency utilization and the network capacity. Also, as the number of D2D pairs increases, a D2D user who owns a file can directly send it to the D2D user who needs it by means of the nearby communication and achieves the traffic offload function. The short range direct D2D communication benefits the whole network with lower energy consumption, load balancing, and better quality of service (QoS) for edge users [Li, Chi, Chen et al. (2018)], and can also be widely applied to dense communication network scenarios such as traffic systems [Pan, Qin, Yi et al. (2019)].

In the traditional D2D enabled network, the base stations play an important role. For example, the D2D user pair needs to apply for a channel from the base station before establishing the D2D communication. The base station needs to consider the interference between users when allocating the channel and minimize the interference between user devices. However, the full control mode of the base station brings many problems to the system, for example, the control signaling overhead being too large and the data transmission delay being too serious.

The main contribution of our paper is as follow. Our goal is to maximize system throughput while satisfying the requirement of QoS for users. We manifest this problem as a stochastic non-cooperative game in which D2D users do not possess any previous information about the quality or availability of the selected resource block.

We put forward a self-deciding algorithm based on multi-agent Q-learning to achieve resource allocation of up-link resources in cellular communication reused by D2D users. Each D2D pair is modeled as an agent, which explores all the possible policies based on the observed channel throughput and state which are affected by the channel quality. Therefore, the problem can be described by the Markov decision process (MDP) [Sutton and Barto (1998)]. The multi-agent Q-learning process for resource blocks selection can be formulated.

Concerning the convergence complexity of the Q-learning, we also proposed a multi-agent Q-learning power control algorithm based on the Fuzzy C-mean algorithm [Parker and Hall (2014)] to accelerate convergence speed.

The structure of this paper is as follows: Sections 1 and 2 introduce the research background. Section 3 describes the system model and problem formulation. Section 4 shows the reinforcement learning algorithm based on multi-agent. The Section 5 focuses on power control based on Q-learning and fuzzy clustering algorithms. The simulation results of this paper are elaborated in Section 6. Finally, the full text is concluded in Section 7.

2 Related works

Although D2D technology can improve the communication capacity and user experience of cellular networks, it also faces many problems that traditional cellular networks do not have. The major obstacles to deploying the D2D communications are related to insufficient spectrum management and reduce interference [Asheralieva and Miyanaga (2016)]. Most existing interference management methods use a centralized approach in which resources are allocated by a central node with global channel state information (CSI) to D2D pairs [Asheralieva and Miyanaga (2016); Yang, Martin, Boukhatem et al. (2015); Kim, Kim, Bang et al. (2016); Penda, Fu and Johansson (2015)]. An et al. [An, Sun, Zhao et al. (2012)] used a proportional fairness algorithm to allocate resources to cellular users, and then used a greedy algorithm to assign D2D users multiplexing resources. Zhou et al. [Zhou, Dong, Ota et al. (2016)] concerned both energy efficiency and quality of service in LTE-A networks and proposed an algorithm that exploits the hybrid architecture of C-RAN. The distributed resource allocation problem is modeled as a non-cooperative game and proposed a centralized interference mitigation algorithm to improve the QoS performance.

At present, machine learning methods such as reinforcement learning (RL) [Chen, Li and Zhao (2016); Zhou, Lu, Wen et al. (2019); Pan, Yu, Yi et al. (2019)] are applied in the field of image processing. More and more researchers use reinforcement learning methods to solve the problem in wireless communication such as cognitive radio networks [Xu, Wu, Shen et al. (2013); Kalathil, Nayyar and Jain (2014)]. In order to reduce the computational complexity of the base station and reduce the control signaling overhead, we introduce reinforcement learning in channel resource allocation and power control. Maghsudi et al. [Maghsudi and Stanczak (2015)] modeled D2D communications as RL, demonstrated it as non-cooperative games, and D2D users explored their policies based on the experience in stochastic environments. Q-learning [Xi, Sheng, Sun et al. (2018)] is a model-free RL algorithm and can be adapted to allocate resource effectively. Maghsudi et al. [Maghsudi and Stanczak (2016)] studied the problem of resource allocation for D2D communication and achieved optimization of the scheme by adopting Q-learning. A Bayesian reinforcement learning-Based coalition formation is proposed by Asheralieva [Asheralieva (2017)] to handle the problem of resource sharing in D2D networks. The above research rarely considers the convergence time of the algorithm.

3 System model and problem formulation

3.1 System model

We consider a single-cell wireless communication system in which the D2D communication mode and the cellular communication mode coexist. In our system, it's assumed that there is a base station (BS) at the center, and the base station can be used to assist cellular users and D2D users to collaborate. In Fig. 1, the system model of this paper is demonstrated. In this paper, we use $\mathbf{N}=\{1, 2, ..., N\}$ for an array of cellular users and $\mathbf{M}=\{1, 2, ..., M\}$ for an array of D2D user pairs. At the same time, in we assume that D2D users and cellular users are evenly distributed in the cell.

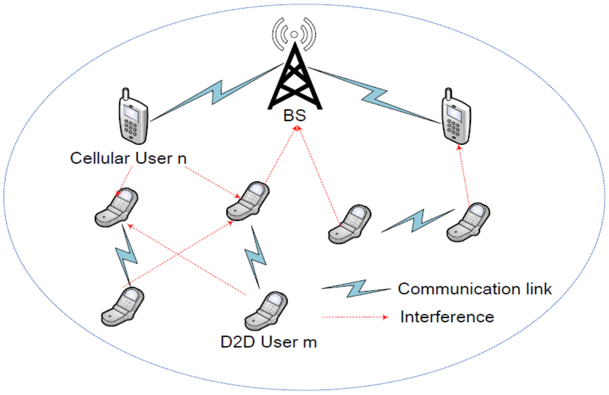

Figure 1: System model for collaboration between D2D users and cellular users

Traditionally, both the up-link resource and the down-link resource of the cellular communication mode can be used by users in the D2D communication mode. However, the fact is that in practical application scenarios, when the up-link resources are reused, the base station will be interfered; when the down-link resources are reused, the users' equipment will be interfered. Comparing with the user equipment, the base station is more capable of resisting interference. Therefore, we use a scenario in which D2D users multiplex up-link resources of the cellular users. We also assume that there is no information exchange and collaboration between D2D users, and they do not know the information of the wireless channel.

The total number of resource blocks in the system is **K**, and the resource block set is **RB**={RB$_1$, ..., RB$_K$}, with **K**={1, 2, ..., K} denoting the set of resource blocks' indexes. For ensuring the QoS of cellular users and fully using of channel resources, the total number of resource blocks is consistent with that of cellular users, that is, *K*=*N*. Each cellular user is assigned a resource block in advance. Resource blocks are orthogonal to each other to ensure no interference between cellular users. One resource block can be reused by one cellular user and multiple D2D users. For the *m*th pair of D2D users, we define a binary K-dimensional resource block selection vector, $\beta^m = [\beta_1^m, \cdots, \beta_K^m]^T$. When β_k^m is equal to 1, it indicates that the *m*th D2D user pair selects the resource block *k*, and when it equals to 0, it indicates that the resource block is not selected. It is assumed above that each D2D user can only select up to one resource block:

$$\sum_{k \in K} \beta_k^m \leq 1, \forall m \in M. \tag{1}$$

The cellular users and the D2D user pairs multiplex the uplink transmission. The signal-to-interference-plus-noise ratio (SINR) of the *m*th D2D user pair communicating over RB_k is denoted as:

$$\gamma_k^{D_m} = \frac{P_k^{D_m} \beta_k^m G_k^{D_m,m}}{P_k^n G_k^{n,m} + \sum_j^{j \neq m, \beta_k^i = 1} P_k^{D_j} G_k^{D_j,m} + \sigma^2} \tag{2}$$

where $P_k^{D_m}$ and P_k^n respectively represent the *m*th D2D pair and the *n*th cellular user transmission power communicating over the *k*th uplink resource block. $G_k^{D_m,m}$, $G_k^{n,m}$, and

$G_k^{D_{j,m}}$ respectively represent the link gain of channel over the mth D2D link, from cellular transmitter n to receiver m communicating over the kth resource block, and from transmitter j of D2D users to receiver m. σ^2 is zero-mean additive white Gaussian noise (AWGN) [Matuz, Liva, Paolini et al. (2013)] power variance.

Similarly, we can define the SINR of the nth cellular user, where $n \in N$, over the RB_k as:

$$\gamma_k^n = \frac{P_k^n G_{n,k}^{BS}}{\sum_j^{\beta_k^i=1} P_k^{D_j} G_{j,k}^{j,BS} + \sigma^2}, \forall n \in N, \forall k \in K \tag{3}$$

where $G_{n,k}^{BS}$ and $G_{j,k}^{j,BS}$ respectively indicate the link gain of channel over the RB_k from BS to cellular user n and from BS to the jth transmitter D2D user.

3.2 Problem formulation

When D2D users share the same resource block with the cellular user, interference occurs to both cellular and D2D users, which affects their communication quality. Therefore, in the allocation of resource to D2D users, the premise is to ensure the communication quality of both cellular users and D2D users. In this paper, the QoS performances of cellular users and D2D users are the constraint, and the maximum throughput of cellular users and D2D users in a cell is taken as the target. We model the problem as Eq. (4):

$$\begin{aligned} &\max \sum_{k=1}^{K} [\omega \log_2(1+\gamma_k^n) + \sum_{m}^{\beta_k^m=1} \omega \log_2(1+\gamma_k^{D_m})] \\ &s.t. \gamma_k^n \geq \tau_C, \forall n \in N, \forall k \in K, \\ &s.t. \gamma_k^{D_m} \geq \tau_D, \forall m \in N, \forall k \in K \end{aligned} \tag{4}$$

where ω is the bandwidth of each channel (in Hz), τ_C is the minimum SINR of cellular users, and τ_D is the minimum SINR of D2D users. The objective function maximizes the system throughput, and the constraints are the QoS requirements for cellular users and D2D users. In the next section, we will introduce the multi-agent RL combining with game theory based algorithm to figure out the optimal scheme of resource allocation for D2D users.

4 Reinforcement learning based resource allocation algorithm

The above optimal D2D resource allocation problem under stochastic environment can be formulated as an MDP in which the state transition probabilities and expected rewards for current states are unknown. In this section, we will introduce a multi-agent RL algorithm which combines with the game theory.

4.1 Basic definition

Reinforcement learning is a machine learning that learning what to do, and how to map situations to actions, so as to maximize a numerical reward signal. In reinforcement learning, there are several basic concepts: agent, state, action, reward function, and strategy function. The agent is the learning subject in the reinforcement learning. It will continuously observe the environment and obtain the current environmental state. The agent will receive positive or negative feedback from the environment after performing the action. Reinforcement learning is Trail-and-error. Because there is no direct guidance information, the agent should constantly interact with the environment and obtain the best strategy through trial and error. Reinforcement learning is unsupervised learning and often given in the aftermath (the last state), which leads to the question of how to assign rewards to the previous state after getting a positive or negative reward. In our system, each D2D pair can be seen as an agent. Accordingly, we formulate our problem in the following subsections.

4.1.1 State vector

For each D2D pair m, the state on RB_k at slot t can be defined as:

$$\mathbf{S}_t^{m,k} = \begin{pmatrix} C_t^k \\ D_t^{m,k} \end{pmatrix}, \forall m \in \mathbf{M}, \forall k \in \mathbf{K}, \tag{5}$$

where C_t^k denotes whether the cellular user occupying RB_k is under severe interference at slot t, defined by

$$C_t^k = \begin{cases} 1 & \gamma_k^n \geq \tau_c, \\ 0 & otherwise, \end{cases} \tag{6}$$

where τ_c is the minimum QoS that cellular users need to meet in communication. We assume that cellular users will exchange the SINR value in the communication process with the base station, and then the D2D user can obtain the information from the base station.

Similarly, $D_t^{m,k}$ is the SINR level of D2D pair m using RB_k at slot t, defined by:

$$D_t^{m,k} = \begin{cases} 1, & \gamma_k^{D_m} \geq \tau_D, \\ 0, & otherwise, \end{cases} \tag{7}$$

where τ_D is the minimum SINR meeting the requirement of D2D users' QoS performances. Providing that the D2D users report this information to BS, the D2D users obtain this value from BS.

4.1.2 Action vector

In our case, the individual actions of the players correspond to their resource block selection decisions. Hence, the action space for each D2D pair can be described by the vector:

$$\mathbf{A}^m = \{\beta^m \mid \beta_k^m \in \{0,1\}, \textstyle\sum_{k\in K} \beta_k^m \leq 1\}, \forall\, m\, \in \mathbf{M} \tag{8}$$

When β_k^m is equal to 1, it indicates that the m th D2D user pair selects the resource block RB_k, and when it equals 0, it indicates that the resource block is not selected. It is assumed above that each D2D user can only select up to one resource block.

We denote the action taken by the agent m at the slot *t* as $a_t^m \in \mathbf{A}$, and a_t^{-n} as the action vectors taken by other agents except the agent *m*:

$$\mathbf{A}^{-m} = \textstyle\prod_{k\in M\backslash\{m\}} \mathbf{A}^k \;\; and\; \mathbf{A} = \mathbf{A}^m\mathbf{A}^{-m}, \forall m \in \mathbf{M} \tag{9}$$

4.1.3 Reward function

When an agent performs an action, it affects the environment and therefore changes the environment. Therefore, the agent receives a reward signal r from the environment each time the action is performed. The function of the reward function is to perform an effect feedback i on the agent just performing the action and instruct the agent to take the next action. Thus, based on the above analysis, the reward function for agent m at slot t is indicated by the capacity on RB_k:

$$\begin{aligned}&\gamma_t^{m,k} = \omega \log_2(1+\gamma_k^n) + \omega \log_2\left(1+\gamma_k^{D_m}\right)\\ &\forall m \in \mathbf{M}, \forall n \in \mathbf{N}, \forall k \in \mathbf{K}, \forall \beta_k^m = 1\end{aligned} \tag{10}$$

Since the problem solved in this paper is to maximize the throughput of the system, and based on the problem abstraction in Section 3, we can set the learning goal as a reward function. Therefore, the expression of the above reward function is reasonable.

4.1.4 Policy and value function

Policy π is used to provide agents with rules to follow in reinforcement learning. Under the guidance of π, the agent can decide what action to perform at the next moment according to the current state. The iteration of the strategy function is in a loop, from policy evaluation to policy improvement, in turn.

Based on the well-known Bellman Equation [Li, Zhao, Sun et al. (2018)]:

$$v(s) = E[r_{t+1} + \lambda v(S_{t+1} | \, S_t = s)] \tag{11}$$

where E in the above equation refers to the expectation, while $\lambda \in (0,1)$ refers to the discount factor. $v(\cdot)$ is a function of state transitions in the given environment. We can re-write it in the form of Q-value:

$$Q^{\pi}(s,a) = E\{r_{t+1} + \lambda Q^{\pi}(s_{t+1}, a_{t+1}) | \, s_t, a_{t+1}\} \tag{12}$$

The optimal Q-value, denoted as Q^* can be expressed as:

$$Q^*(s_t, a_t) = E\left\{r_t + \lambda \max_{a_t+1} Q\ (s_{t+1}, a_{t+1}) \mid s_t, a_t\right\} \tag{13}$$

Among the lots of other methods to choose the optimal Q-value based on the action, we use the ε-greedy strategy [Wu, Wang and Yin (2019)] in this paper to take actions according to the current estimated Q-value, and it can be described as follows:

- Choose the optimal action $a^* = arg \max_{a \in A} Q(s,a)$ with the probability $1 - \varepsilon < 1$.
- Choose other action with probability $\varepsilon > 0$ at random.

In Q-learning, each $Q(s,a)$ corresponds to a corresponding Q value, and the action is selected according to the Q value during the learning process. The Q-learning algorithm is combining with two process which are learning process and take action process. Each learning process of the agent can be seen as starting from a random state, using a strategy to select actions, such as ε -greedy strategy. The ε-greedy is used to ensure that the agent can search for all possible actions and update each $Q(s,a)$. After performing the selected action, the agent observes the new state and reward, and then updates the Q value of the previous state and action according to the maximum Q value and reward of the new state. The agent will continue to select actions based on the new state until it reaches a termination state. The termination state is control by the maximize loop times of the Q value updates. In our system, multiple agents make decisions together, therefore the strategies executed by each agent are affected by the environment as well as the other agents.

4.2 Multi-agent Q-learning algorithm

Considering the situation of multiple pairs of D2D users under one base station, this paper mainly studies the scenario of multi-agent. Multi-agent system is one in which several agents attempt to maximize utility or solve tasks jointly through their interaction. Since the D2D pairs do not possess any information on the availability or quality of the channel to be selected and each D2D pair is an agent, we can now present the considered resource allocation problem as a non-cooperative game when we establish the multi-agent Q-learning algorithm, which can be demonstrated as:

$$\Gamma = \{\mathbf{M}, \{\mathbf{A}^m\},\ \{\mathbf{u}^m\}\}, \forall\, m \in \mathbf{M} \tag{14}$$

where $\mathbf{M}$ is the set of players (D2D pairs), $\mathbf{A}^m$ is the players' action space, and $\mathbf{u}^m$ [Asheralieva and Miyanaga (2016)] is the utilities of the players.

According to (14), $\mathbf{u}^m$ can be achieved by:

$$\mathbf{u}^m(s_t^m, a_t^m, \mathbf{a}_{\boldsymbol{t}}^{-\boldsymbol{m}}) = s_t^m r_t^m(a_t^m, \mathbf{a}_{\boldsymbol{t}}^{-\boldsymbol{m}}, \mathbf{G}^m, \mathbf{G}^n), \forall n \in \mathbf{N}, \forall m \in \mathbf{M} \tag{15}$$

where $\mathbf{G}^m$ and $\mathbf{G}^n$ are respectively the link gain matrix of D2D user m and cellular user n, and $r_t^m(\cdot)$ is the reward of agent m at slot t. Then each agent m selects the $\mathbf{RB}$ based on the Nash equilibrium (NE) state to maximize its utility $\mathbf{u}^m$. Also, according to Asheralieva et al. [Asheralieva and Miyanaga (2016)], there is NE. The equation state goes like:

$$\mathbf{u}^m(s_t^m, \bar{a}_t^m, \bar{\mathbf{a}}_{\boldsymbol{t}}^{-\boldsymbol{m}}) \geq \mathbf{u}^m(s_t^m, a_t^m, \bar{\mathbf{a}}_{\boldsymbol{t}}^{-\boldsymbol{m}}), \forall a_t^m \in \mathbf{A}^m \tag{16}$$

combining (13) and (16), we can get the optimal Q-value:

$$Q_t^* = \bar{Q}_t(s_t^m, a_t^m) = E\left\{u_t^m[a_t^m, \bar{\pi}^{-m}(s_t^m), s_t^m] + \lambda \max_{a_t^m \in \mathbf{A}^m} \bar{Q}_t\,(s_{t+1}, a_{t+1})\right\} \tag{17}$$

4.3 Structure

We present our multi-agent Q-learning algorithm into distributed Q-learning [Nie, Fan, Zhao et al. (2016)], which can split the large Q-value tables into m smaller Q-value tables $Q_i(s, a)(\ i = 1, 2, \cdots, m)$ so as to decrease the learning complexity. Each agent executes actions repeatedly during the learning process. Only if the newly derived Q-value is greater than the former one in the Q-value table, will the Q-value be updated. Based on Eq. (17), we can get the update rule by Eq. (18):

$$Q_{t+1} = \{\begin{matrix} (1 - a_t)Q_t(s_t, a_t) + \alpha_t\left[r_t + \lambda \max\limits_{a_{t+1}} Q(s_{t+1}, a_{t+1})\right], & if \quad s = s_t, a = a_t \\ Q(a_t, t_t), & otherwise \end{matrix} \tag{18}$$

where $\alpha_t \in (0, 1]$ is the learning rate. When the factor equals to 0, it means the agent learn nothing. In contrast, when it approaches to 1, the agent cares only about the latest information. And the discount factor is denoted as $\lambda \in (0, 1]$, which indicates the influence of rewards in the future. When λ equals to 0, the agent only considers about the present reward. When the factor of λ approaching 1, it makes the agent influenced by the long-term reward to a great extent.

First, each agent initializes its Q-table, and we select the initial state $s \in \mathbf{S}$ randomly. Then we do the following iterations until the Q-value reach convergence. First, the next action a_t using the λ-greedy policy is selected and the agent obtains the immediate reward r_t after executing the action, and then the agent observes the next state s_{t+1} based on a_t andr_t. Finally, the Q-table is updated according to Eq. (18).

5 Power control problem in D2D communication

The power control problem in D2D communication can also be solved by the Q-learning algorithm. The objective function is to maximize the system throughput. Similar to Eq. (4), we get Eq. (19).

$$\max \sum_{k=1}^{K} \left[\omega \log_2(1+\gamma_k^n) + \sum_{m}^{\beta_k^m=1} \omega \log_2\left(1+\gamma_k^{D_m}\right)\right]$$

$$s.t. \gamma_k^n \geq \tau_C, \forall n \in N, \forall k \in K, \qquad (19)$$

$$s.t. 0 \leq p_k^m \leq P_{max}, \forall m \in N, \forall k \in K.$$

where P_{max} is the maximum transmission power of D2D user. The objective function is to maximize the system throughput. The constraint of this problem is the minimum QoS requirement of the cellular user, that is, the communication service quality of the cellular user needs to be guaranteed while realizing the objective function. In order to speed up the convergence of the algorithm, we will introduce the multi-agent RL algorithm combined with the Fuzzy C-means (FCM) algorithm to get the optimal power control scheme for D2D users.

5.1 D2D communication power control based on fuzzy C-Means and Q-learning algorithm

The FCM algorithm can obtain the degree of membership of each point in the sample for the center of the group by optimizing the objective function, so that each sample point in the sample can be automatically grouped by the class attribute of each sample point. This algorithm is one of the important technical means in unsupervised machine learning. It can more objectively and accurately describe the uncertainty of sample category attributes and can effectively group the analysis of sample category attributes.

The input of the FCM algorithm is a data set to be grouped, in which each data contains a certain number of features, and the output is a matrix of c rows and n columns, where c is the number of clusters after clustering, and n is the data set. The number of data, through the matrix can show the results of the cluster. For example, a column in a matrix indicates the degree to which an element belongs to each cluster. The largest value of the data in the column indicates that the element has the highest degree of membership to the class, and thus can be classified into the class.

Firstly, based on the FCM algorithm, M D2D user pairs can be divided into C D2D user groups, denoted as G. The attributes of the D2D pairs in each user group should have great similarities in order to reduce interference between users within the group and increase system throughput. In order to approximate the actual situation, the attribute set of the D2D user is defined as {location, data to be sent, rate requirement, bit error rate requirement, SINR requirement, maximum tolerable waiting time delay}, that is, each D2D user pair has a 6 attribute (H=6). The location refers to the location of the D2D user. The data to be sent is the amount of data that the D2D user will send. The bit error rate requirement refers to the bit error of the D2D user. The SINR requirement refers to the SINR requirement of the D2D user. Maximum tolerable waiting time refers to the maximum tolerable delay requirement of the D2D user. Assume that the six eigenvalues of the two D2D users in each D2D pair are the same. Therefore, the matrix **X** of all D2D pair attributes can be expressed as:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1H} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MH} \end{bmatrix} \tag{20}$$

where x_{ij} represents the jth attribute of the ith D2D user pair, $\mathrm{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_M]^{\boldsymbol{T}}$, $\mathbf{X}_i$ represents the i th row in the matrix and $\mathbf{X}_i = [x_{i1}, x_{i2}, \cdots, x_{iH}]$.

When grouping, the following two conditions should also be met: 1) each group contains at least one D2D pair; 2) each D2D pair can belong to only one group. The minimum distance threshold between D2D user groups is set to ρ. Assuming that each pair of D2D users is in a fixed position without moving, the distance between any two D2D pairs can be defined as:

$$d_{i,j} = \left|\left|X_i - X_j\right|\right| = (X_i - X_j)^T \cdot \left(X_i - X_j\right), \forall i, j \in \mathbf{N} \tag{21}$$

Then, the membership function $z_{im} = z_{G_i}(\mathbf{X}_m)$ is defined to indicate the degree to which the m th D2D user pair belongs to the group G_i, which is defined as:

$$z_{im} = \frac{d_{im}^2}{\sum_{c=1}^{C} d_{cm}^2}, \forall m \in \mathbf{M} \tag{22}$$

where z_{im} is the ratio of the degree of $\mathbf{X}_m$-to-group G_t membership to the degree of subordination of all other user groups except G_i, where $z_{im} \in [0,1]$. When z_{im} is close to 1, it indicates that the *m*th D2D user pair is more subordinate to the group G_i and is closer to the cluster center.

In order to avoid generating a trivial solution, the distance from each user object to the cluster center is measured by the square of its membership, and the objective function is obtained. The expression is as follows:

$$J_\mu(\mathbf{Z}, \mathbf{W}) = \sum_{m=1}^{M} \sum_{i=1}^{C} \frac{1}{(z_{im}^{\mu})(d_{im}^2)}, \mu \in [1, \infty) \tag{23}$$

where $\sum_{i=1}^{C} z_{im} = 1$, $\mathbf{Z} = [z_{im}]_{c \times M}$, $\mathbf{W} = [w_i | i = 1, 2, \cdots, C]$ denotes the clustering center, and μ is a fuzzy factor, which determines the weighting index of the ambiguity of the clustering result.

The FCM is an iterative solution process that minimizes the objective function $J_\mu(\mathbf{Z}, \mathbf{W})$:

$$\min | J_\mu(\mathbf{Z}, \mathbf{W}) | = \min\left\{\sum_{m=1}^{M} \sum_{i=1}^{C} (z_{im}^{-\mu})(d_{im}^{-2})\right\} \tag{24}$$

Under the constraint of $\sum_{i=1}^{C} z_{im} = 1$, the extremum of Eq. (24) can be solved by Lagrangian multiplication under constraints.

$$F = \sum_{i=1}^{C} z_{im}^{-\mu}\left(d_{im}^{-2} + \lambda\left(\sum_{i=1}^{C} z_{im} - 1\right)\right), \forall m \in \mathbf{M} \tag{25}$$

Solve the z and w that minimize F.

$$z_{im} = \frac{1}{\sum_{j=1}^{C}\left(\frac{d_{im}}{d_{jm}}\right)^{\frac{2}{\mu+1}}}, \forall m \in \mathbf{M} \tag{26}$$

$$\mathbf{w}_m = \frac{1}{\sum_{i=1}^{C}(z_{im}^{-\mu})(d_{im}^{-3})}\sum_{i=1}^{C}\{(z_{im}^{-\mu})\left(d_{im}^{-3}\right)\mathbf{X}_m\} \tag{27}$$

After the user grouping is completed, the RL algorithm model of the multi-agent will be established. In the system model, each D2D user group grouped by the FCM is regarded as an agent. The agent is represented by user group c and $1 < c < C$.

The action taken by each agent consists of a set of levels of transmit power, defined as:

$$\mathbf{A} = \{a_1^k, a_2^k, \cdots, a_L^k\}, \forall k \in K \tag{28}$$

where L is the number of the action.

For the D2D user group c, s_t^c is the state at time slot t and is one of the element in the state space S_C. Therefore, it can be defined as:

$$s_t^c = \begin{cases} 1, & \gamma_k^n(t) \geq \tau_D, \\ 0, & otherwise, \end{cases} \tag{29}$$

It is assumed that the cellular user will send the value of the SINR to the BS, and the D2D user obtains the information from the BS.

The reward function of agent c at time slot t is represented by the system capacity on resource block RB_k:

$$r_k^c(t) = \omega \log_2\left(1 + \gamma_k^n(t)\right) + \omega \sum_{c_i \in c} \log_2\left(1 + \gamma_k^{c_i}(t)\right), \forall n \in \mathbf{N}, \forall k \in \mathbf{K} \tag{30}$$

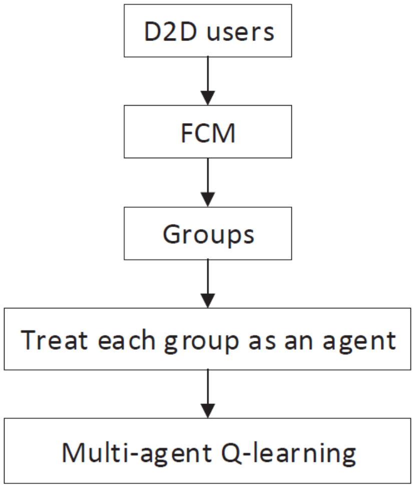

Figure 2: The flowchart of the multi-agent Q-learning based on the FCM

In the selection of the best Q-value strategy based on action, this section still selects the ε-greedy strategy to take action based on the currently estimated Q value. The update equation of the Q value is as same as Eq. (18). The flowchart of the multi-agent Q-meters. The distance between each D2D pair is 50 meters. There are 12 D2D pairs and 20 cellular users (i.e., M=12 and N=20) that are evenly distributed within the BS coverage. The users' equipment is placed outdoors and stationary (assuming a typical urban environment). The number of RBs is the same as that of cellular users (i.e., K=20). and each resource block RB_k corresponds to a channel with ω=180 kHz. For guaranteeing the QoS of cellular users, the minimum SINR requirement is τ_C=0.5 dB, while that of D2D users is τ_D=0 dB. We set the transmit power of the D2D user to 1 mW and the transmit power of the cellular user to 200 mW. The reason for this is that D2D users have relatively close communication distances, and relatively small transmit power can satisfy the D2D pairs QoS requirements. At the same time, this can also reduce the interference of D2D pairs communication to cellular users. The values of the basic parameters in Q-learning are: the learning rate α is set to 0.5, and the discount factor λ is set to 0.8. See Tab. 1.

Table 1: Parameters of simulation

Parameter	Value
the minimum SINR of cellular users τ_C	0.5 dB
the minimum SINR of D2D users τ_D	0 dB
learning rate α	0.5
discount factor γ	0.8
M	12
N	20
K	20
transmission power of cellular users	200 mW
transmission power of D2D users	1 mW
P_{max}	256 dBm
L	20
bandwidth of resource block	180 kHZ
Noise Power/RB	-116 dBm
coverage radius of base station	500 m
distance of D2D pairs	50 m
path loss model between BS and users	16.2+38.9lg(d(km))(dB)
path loss model between users	29+41.2lg(d(km))(dB)

In this paper, the multi-agent based Q-learning algorithm is abbreviated as MAQ. The following three algorithms are used to compare with the proposed resource allocation algorithm. An ε-greedy algorithm [Wu, Wang and Yin (2019)] (ε-GA): the selection action only considering the current maximum utility value, that is, whenever there is service arrival, as long as the system resources can satisfy the service requirement, the system will allocate the corresponding resources to the user, otherwise do not allocate. ε-greedy means that even if the resource can meet the quality of service requirements, there is still a probability that ε will not allocate resources to it (with ε=0.1). The ε -greedy algorithm for resource equalization (ε-GRE) is introduced to divide the available resources of the system into two parts equally for cellular network users and D2D users, and to allocate resources according to greedy algorithms in their respective resource sets. The optimal centralized policy (OCP): The BS uses global CSI to assign channels to D2D users to maximize overall network throughput under SINR constraints which represents the best channel selection strategy.

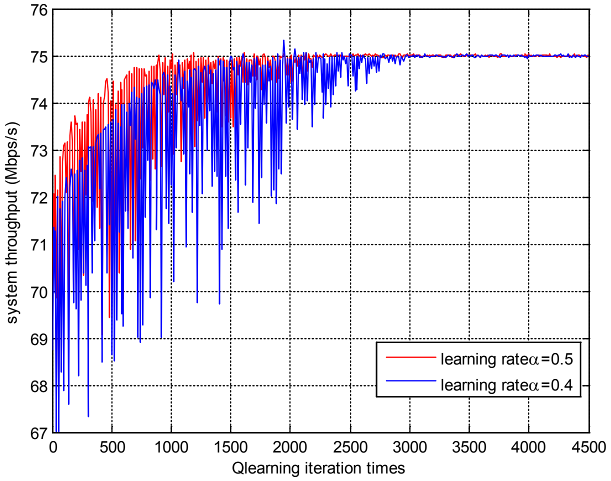

Figure 3: Convergence of algorithm with different learning rates

Fig. 3 presents the convergence of the proposed algorithm under different learning rates. Two learning rates are set for comparison: α=0.4 and α=0.5. When the learning rate is small, that is, α=0.4, the convergence speed is slower, and the convergence is gradually achieved after more than 2500 learning iterations. However, when we improve the learning rate, that is, α=0.5, we can see from the simulation results that the convergence speed has been significantly improved, and approximately 2,000 iterations of learning can be achieved. In reinforcement learning, the rate of learning influences the convergence rate of Q-value, but the Q-value achieved by the final convergence under different learning rates is the same.

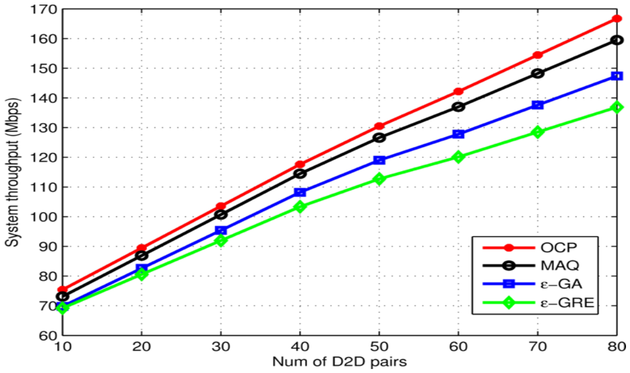

Figure 4: Changes in system throughput with different D2D user pairs

Fig. 4 shows the system throughput as the number of D2D users changes as the iteration of the system reaches a stable state. The performance of the MAQ algorithm was compared with the other three algorithms described above. First of all, we can conclude that the total system throughput shows a trend of synchronous growth as the number of D2D users increases. In addition, according to the trend of different curves in Fig. 4, it can be seen that the performance of the proposed algorithm is closest to the OCP, which is the ideal resource allocation algorithm.

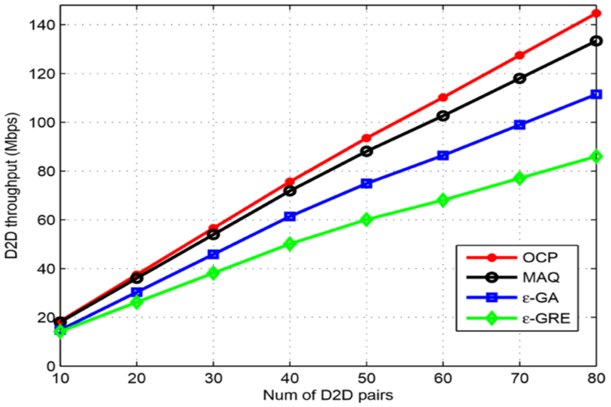

Figure 5: Changes in D2D throughput with different D2D user pairs

Fig. 5 presents how the D2D user's throughput changes with the number of D2D users after the algorithm reaches convergence. Similar to the trend of system throughput, the throughput of D2D users also tends to increase with the increase of the number of D2D users. At the same time, Fig. 5 confirms the superiority of the proposed algorithm. The closest match to the performance of the OCP algorithm is the proposed MAQ algorithm. Similarly, the ε-GA algorithm takes the next place, and the ε-GRE algorithm is the worst.

The shortcoming of the two greedy algorithm is that the optimal solution is obtained step by step, and there is no global consideration, so it is easy to fall into the local optimum, and the global optimal cannot be obtained.

6.2 Results of power control problem in D2D communication

In this section, the FCM combined with multi-agent Q-learning is abbreviated as FC-MAQ. The following three algorithms are used to compare with the proposed algorithm: i) A multi-agent Q learning MAQ algorithm that does not use FCM, where each D2D pair is treated as an agent in the algorithm without pre-grouping them; ii) Optimal Centralized Policy (OCP); iii) The open loop power control algorithm (hereinafter referred to as the OP algorithm) is the most commonly used power control method in D2D communication. Open loop power control refers to a scheme that directly adjusts the transmit power of a D2D user without feedback information.

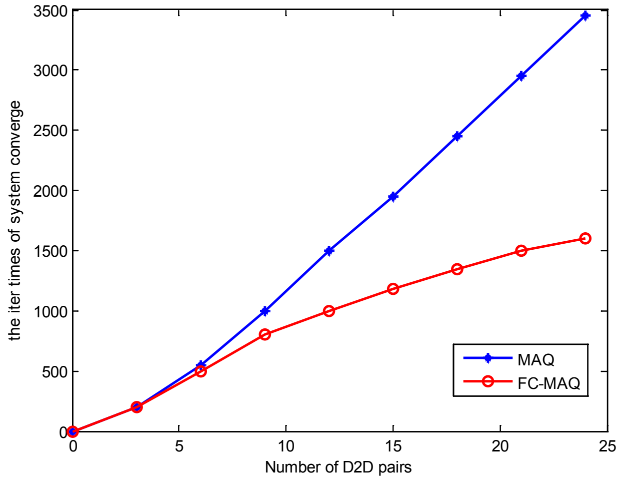

Figure 6: Convergence time with different D2D user pairs

Fig. 6 compares the number of iterations when the FC-MAQ algorithm and the MAQ algorithm system reach their optimal solution under different D2D pairs, and the learning rates of both are set to 0.5. It can be seen that FC-MAQ effectively reduces the number of system convergence compared with the MAQ algorithm, and as the number of D2D increases, the number of iterations of the FC-MAQ algorithm is much smaller than the MAQ algorithm due to the limitation of the maximum number of clusters. The FC-MAQ algorithm has this significant advantage in reducing system complexity.

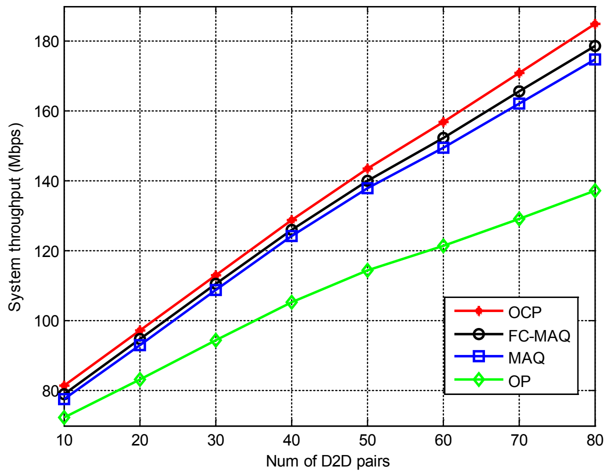

Figure 7: Changes in D2D throughput with different D2D user pairs

Fig. 7 shows the trend of the total system throughput under the different algorithm. The proposed FC-MAQ algorithm is compared with the ideal case OCP algorithm, MAQ algorithm and OP algorithm. It can be concluded that as the number of D2D users increases, the total system throughput shows a trend of simultaneous growth. It can be seen that the performance of the proposed FC-MAQ algorithm is closest to the ideal OCP algorithm, the performance of the MAQ algorithm is slightly lower, and the performance of the OP algorithm is the worst. Combined with the simulation results of Figs. 6 and 7, we can see that the proposed algorithm outperforms other algorithms in convergence time and network throughput.

7 Conclusion

In this paper, we provide a multi-agent Q-learning method to improve the throughput of D2D systems. Firstly, channel resource utilization is improved by the channel resource allocation method based on the multi-agent Q-learning algorithm. Secondly, in order to solve the problem of slow convergence based on Q-learning method in D2D communication system, we introduce FCM based on multi-agent Q-learning algorithm to improve the convergence speed of power control based on Q-learning. The experimental results show that the multi-agent Q-learning algorithm can improve the throughput of the system, and the FCM can speed up the convergence of the algorithm and reduce the computational complexity.

Funding Statement: This work was supported by the National Natural Science Foundation of China (61871058), and Key Special Project in Intergovernmental International Scientific and Technological Innovation Cooperation of National Key Research and Development Program (2017YFE0118600).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References:

An, R.; Sun, J.; Zhao, S.; Shao, S. (2012): Resource allocation scheme for device-to-device communication underlying lte downlink network. *International Conference on Wireless Communications and Signal Processing*, pp. 1-5.

Asheralieva, A. (2017): Bayesian reinforcement learning-based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5016-5032.

Asheralieva, A.; Miyanaga, Y. (2016): An autonomous learning-based algorithm for joint channel and power level selection by D2D pairs in heterogeneous cellular networks. *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3996-4012.

Asheralieva, A.; Miyanaga, Y. (2016): Dynamic buffer status-based control for LTE-A network with underlay D2D communication. *IEEE Transactions on Communications*, vol. 64, no. 3, pp. 1342-1355.

Asheralieva, A.; Miyanaga, Y. (2016): Qos-oriented mode, spectrum, and power allocation for D2D communication underlaying LTE-A network. *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9787-9800.

Bello, O.; Zeadally, S. (2016): Intelligent device-to-device communication in the internet of things. *IEEE Systems Journal*, vol. 10, no. 3, pp. 1172-1182.

Chen, H.; Li, X.; Zhao, F. (2016): A reinforcement learning-based sleep scheduling algorithm for desired area coverage in solar-powered wireless sensor networks. *IEEE Sensors Journal*, vol. 16, no. 8, pp. 2763-2774.

Doppler, K.; Rinne, M.; Wijting, C.; Ribeiro, C. B. et al. (2009): Device-to-device communication as an underlay to LTE-advanced networks. *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42-49.

Kalathil, D.; Nayyar, N.; Jain, R. (2014): Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331-2345.

Kim, J.; Kim, S.; Bang, J.; Hong, D. (2016): Adaptive mode selection in D2D communications considering the bursty traffic model. *IEEE Communications Letters*, vol. 20, no. 4, pp. 712-715.

Li, R.; Zhao, Z.; Sun, Q.; I, C.; Yang, C. et al. (2018): Deep reinforcement learning for resource management in network slicing. *IEEE Access*, vol. 6, pp. 74429-74441.

Li, Y.; Chi, K.; Chen, H.; Wang, Z.; Zhu, Y. (2018): Narrow band Internet of Things systems with opportunistic d2d communication. *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1474-1484.

Maghsudi, S.; Stanczak, S. (2015): Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309-1322.

Maghsudi, S.; Stanczak, S. (2016): Hybrid centralized-distributed resource allocation or device-to-device communication underlaying cellular networks. *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2481-2495.

Matuz, B.; Liva, G.; Paolini, E.; Chiani, M.; Bauch, G. (2013): Low-rate non-binary IDPC codes for coherent and blockwise non-coherent AWGN channels. *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4096-4107.

Nie, S.; Fan, Z.; Zhao, M.; Gu, X.; Zhang, L. (2016): Q-learning based power control algorithm for D2D communication. *IEEE 27th Annual International Symposium on Personal*, *Indoor and Mobile Radio Communications*, pp. 1-6.

Pan, Z.; Qin, H.; Yi, X.; Zheng, Y.; Khan, A. (2019): Low complexity versatile video coding for traffic surveillance system. *International Journal of Sensor Networks*, vol. 30, no. 2, pp. 116-125.

Pan, Z.; Yu, W.; Yi, X.; Khan, A.; Yuan, F. et al. (2019): Recent progress on generative adversarial networks (GANs): a survey. *IEEE Access*, vol. 7, pp. 36322-36333.

Parker, J. K.; Hall, L. O. (2014): Accelerating fuzzy-c means using an estimated subsample size. *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 5, pp. 1229-1244.

Penda, D. D.; Fu, L.; Johansson, M. (2015): Mode selection for energy efficient D2D communications in dynamic TDD systems. *IEEE International Conference on Communications*, pp. 5404-5409.

Sutton, R. S.; Barto, A. G. (1998): Reinforcement learning: an introduction. *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054-1054.

Wu, C.; Wang, Y.; Yin, Z. (2019): Realizing railway cognitive radio: a reinforcement base-station multi-agent model. *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1452-1467.

Xi, X.; Sheng, V. S.; Sun, B.; Wang, L.; Hu, F. (2018): An empirical comparison on multi-target regression learning. *Computers*, *Materials and Continua*, vol. 56, no. 2, pp. 185-198.

Xu, Y.; Wu, Q.; Shen, L.; Wang, J.; Anpalagan, A. (2013): Opportunistic spectrum access with spatial reuse: graphical game and uncoupled learning solutions. *IEEE Transactions on Wireless Communications*, vol. 12, no. 10, pp. 4814-4826.

Yang, K; Martin. S; Boukhatem, L.; Wu, J.; Bu, X. (2015): Energy-efficient resource allocation for device-to-device communications overlaying lte networks. *IEEE 82nd Vehicular Technology Conference*, pp. 1-6.

Zhou, F.; Lu, G.; Wen, M.; Liang, Y.; Chu, Z. et al. (2019): Dynamic spectrum management via machine learning: state of the art, taxonomy, challenges, and open research issues. *IEEE Network*, vol. 33, no. 4, pp. 54-62.

Zhou, Z.; Dong, M.; Ota, K.; Wang, G.; Yang, L. T. (2016): Energy-efficient resource allocation for D2D communications underlaying cloud-ran-based LTE-A networks. *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 428-438.