Power Allocation Algorithm Based on Machine Learning for Device-to-Device Communication in Cellular Network

He Ma[1], Zhiliang Qin[2], and Ruofei Ma[1(✉)]

[1] Harbin Institute of Technology, Weihai 264209, Shandong, China
maruofei@hit.edu.cn
[2] Beiyang Electric Group Co. Ltd., Weihai, Shandong, China
qinzhiliang@beiyang.com

Abstract. With the development of the Internet, more and more mobile user equipment access to the cellular network, so the shortage of wireless spectrum resources has become increasingly prominent. Device-to-device (D2D) communication, as a key technology to solve this problem, can greatly improve the spectrum utilization rate and reduce the load of the base station. However, in the communication process of cellular users, D2D users occupying the same channel will bring complicated electromagnetic interference to them. This paper will establish a single-cell system model in which cellular users and D2D users coexist, and apply the method of power allocation to solve the problem of interference in the communication system. Then, we propose power allocation algorithm based on Q learning. Finally, the performance of the power allocation algorithm based on Q learning is analyzed and evaluated through the results of simulation experiments to verify the superiority of the algorithm over the performance of traditional power allocation algorithm.

Keywords: Device-to-device (D2D) communication · Power allocation · Q learning

1 Introduction

Device-to-device (D2D) communication technology which has been included into the development framework of a new generation of mobile communication system by the 3rd Generation Partnership Project is one of the key technologies of 5G. K. Doppler et al. put forward the concept of D2D communication in 2009 [1]. D2D communication is a technology that supports the direct communication between two terminal equipment. Because the distance between transmitting user and receiving user is relatively close, they

This work was supported partially by National Natural Science Foundation of China (Grant No. 61801144, 61971156), Shandong Provincial Natural Science Foundation, China (Grant No. ZR2019QF003, ZR2019MF035), and the Fundamental Research Funds for the Central Universities, China (Grant No. HIT.NSRIF.2019081).

S. Shi et al. (Eds.): 6GN 2021, LNICST 439, pp. 160–171, 2022.
https://doi.org/10.1007/978-3-031-04245-4_14

no longer need to relay through base station (BS) to carry out information exchange [2]. The main features of D2D communication are to save resources, improve transmission efficiency, and reduce interference [3]. When D2D users and cellular users choose to use the same spectrum resources, the spectrum utilization efficiency can be improved. But at the same time, the access of D2D users also brings complex electromagnetic interference to cellular users. Therefore, effective control of interference and reasonable allocation of resources become extremely important in D2D communication [4].

With the rapid development of computer technology, artificial intelligence and machine learning are becoming more and more closely related to our daily lives [5]. As one of research areas of artificial intelligence, machine learning technology in the robot technology, virtual personal assistant, computer games, pattern recognition, natural language processing and online transportation network has been widely applied [6]. Nowadays, the research based on machine learning has made great progress in the field of communication, which provides solutions for the management of wireless resources of D2D communication.

In [7], a Capacity Oriented Resource ALlocation (CORAL) algorithm for resource allocation in D2D communication underlaying mobile cellular network is proposed. The CORAL algorithm assumes that a D2D user can occupy the communication resources of multiple cellular links. At the same time, a Capacity-Oriented REstricted (CORE) region of the D2D user is introduced to determine the candidate cellular user set for the D2D user. And the CORAL algorithm is superior to the traditional random allocation algorithm in terms of system capacity and rate loss of all cellular users. In [8], the authors study several power allocation schemes for D2D communication, including a fixed power scheme, a fixed SNR target scheme, an (LTE) open loop fraction power control scheme, and a close loop power control scheme. In [9], an enhanced single-leader-multiple-followers Stackelberg game model is presented to investigate distributed power control strategies.

In addition, the problem of resource allocation for D2D communication is also connected with the current popular machine learning, and more solutions are obtained. These resource allocation algorithms in [10–12] all consider applying Q learning algorithm to solve the problem of resource allocation. Moreover, CART Decision Tree algorithm is also applied to research problem in [12].

In this paper, we consider the situation of D2D user multiplexing uplink of cellular user to communicate in a single-cell cellular network. The goal is to improve the performance of the communication system through resource allocation without affecting the QoS of the cellular users, and ultimately maximize the throughput of the entire system. At the same time, the machine learning method is applied to the research of resource allocation for D2D communication.

The rest of this paper is as follows. In Sect. 2, we establish a system model and formulate the problem. Section 3 briefly introduces Q learning, then we propose the power allocation algorithm based on Q learning for D2D communication. Section 4 provides the simulation results and performance analysis. Finally, we conclude in Sect. 5.

2 System Model and Problem Formulation

2.1 System Model

In this paper, we study a single-cell model as shown in Fig. 1. There is a base station (BS) fixed in the center of the cell. In its coverage area, cellular user equipment (CUE) and D2D user equipment (DUE) are randomly and evenly distributed. The number of CUE is M, denoted as $C = \{C_1, C_2, \cdots, C_M\}$. The number of DUE is N, denoted as $D = \{D_1, D_2, \cdots, D_N\}$. A D2D transmitter (DUE Tx) and a D2D receiver (DUE Rx) together make up a D2D, and they can communicate directly. It is assumed that there are a total of K orthogonal spectrum resource blocks (RB), denoted as $B = \{B_1, B_2, \cdots, B_K\}$. This paper considers the scenario of D2D user multiplexing uplink of cellular user.

In order to facilitate the discussion of future problem, we assume that only one cellular user can communicate on each resource block in the cell, and the number of cellular users is equal to the number of the resource blocks. Each D2D user can occupy only one spectrum resource block, and each resource block can be occupied by at most one D2D user. The communication of the mobile user equipment with the same communication mode is independent of each other and does not interfere with each other. It is worth noting that the transmitting power of the cellular users keeps constant throughout the communication process during the study. As this paper studies the power allocation of D2D communication, the spectrum allocation of cellular users and D2D users is fixed, that is to say, the base station will randomly allocate spectrum resource blocks to user equipment before D2D communication, and the spectrum will always be fixed during the communication process.

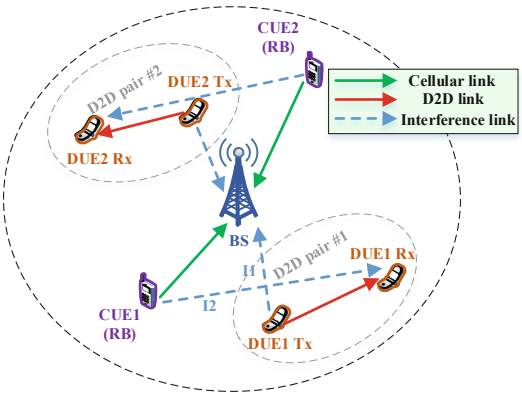

Fig. 1. System model

Due to the randomness of the location of D2D users and the communication between cellular users and base station, D2D users multiplexing the uplink resources of cellular users for communication will cause interference. According to the above definitions and assumptions of the system model, there are two main kinds of interference in the system: I_1 is the interference received by BS from DUE Tx; I_2 is the interference received by DUE Rx from CUE occupying the same resource block.

The next step is to set BS to represent the base station, C_i represents CUE i in the cell and D_j represents DUE j. Then we can analyze the channel gain between C_i and D_j, as well as between user equipment and BS.

For CUE and BS, we consider free space path loss model, and then the channel gain of C_i and BS is calculated according to the following equation:

$$G_{BS,C_i} = 10^{-PL_{BS,C_i}/10} \tag{1}$$

where PL_{BS,C_i} is the path loss between BS and C_i.

In this paper, a simple single-slope path loss model is used to calculate the channel gain between two mobile user equipment. Therefore, the channel gain between C_i and D_j is defined as:

$$G_{C_iD_j} = k \cdot d_{C_iD_j}^{-\mu} \tag{2}$$

where $d_{C_iD_j}$ is the distance between C_i and D_j, k represents gain index and μ represents path loss index.

The channel gain of D2D link can be defined as:

$$G_{D_jD_j} = k \cdot d_{D_jD_j}^{-\mu} \tag{3}$$

where $d_{D_jD_j}$ is the distance between the transmitter and receiver of D2D user.

2.2 Problem Formulation

We consider the condition that the resource blocks of the system have been well allocated. The goal of the problem of power allocation is to use the corresponding power allocation algorithm to assign the optimal transmitting power to each D2D user in the system, so as to reduce the influence of the interference existing in the above system on the communication of user equipment and maximize the throughput of the system.

Signal-to-interference-noise ratio (SINR) is an important index to measure the communication quality of link. The *SINR* of D2D user D_j, which occupies the resource block *Br* can be defined as:

$$SINR_{D_j}^r = \frac{p_{D_j}^r \cdot G_{D_jD_j}^r}{\sigma^2 + p_{C_i}^r \cdot G_{C_iD_j}^r} \tag{4}$$

where $p_{C_i}^r$ and $p_{D_j}^r$ are respectively the transmitting power of cellular user C_i and D2D user D_j, σ^2 is the noise power, and $p_{C_i}^r \cdot G_{C_iD_j}^r$ is the interference received by D_j from C_i, which shares the same resource block.

Similarly, the *SINR* of cellular user C_i occupying resource block *Br* can be defined as:

$$SINR_{C_i}^r = \frac{p_{C_i}^r \cdot G_{BS,C_i}^r}{\sigma^2 + p_{D_j}^r \cdot G_{C_iD_j}^r} \tag{5}$$

where $p^r_{D_j} \cdot G^r_{C_i D_j}$ is the interference received by C_i from D_j, which shares the same resource block.

When the resource block is not occupied by the D2D user, the cellular user occupying this resource block will not be disturbed by the D2D user. At this time, the *SINR* of the cellular user can be defined as:

$$SINR^r_{C_i} = \frac{p^r_{C_i} \cdot G^r_{BS,C_i}}{\sigma^2} \tag{6}$$

We can use Shannon equation to calculate the throughput of cellular users and D2D users to obtain the total throughput of the whole system:

$$R_D = \sum_{j=1}^{N} W \cdot \log_2(1 + SINR^r_{D_j}) \tag{7}$$

$$R_C = \sum_{i=1}^{M} W \cdot \log_2(1 + SINR^r_{C_i}) \tag{8}$$

$$R = R_D + R_C = \sum_{j=1}^{N} W \cdot \log_2(1 + SINR^r_{D_j}) + \sum_{i=1}^{M} W \cdot \log_2(1 + SINR^r_{C_i}) \tag{9}$$

where R_D and R_C are the throughput of D2D users and cellular users respectively, W is the bandwidth of the system and R is the throughput of the whole system.

Therefore, the goal of maximizing the throughput of the system:

$$\begin{aligned} &\max_{p^r_{D_j}} \{R\} \\ s.t. \quad &SINR^r_{C_i} \geq \tau_0 \\ &p_{\min} \leq p^r_{D_j} \leq p_{\max}, \forall j, r \end{aligned} \tag{10}$$

where τ_0 is the minimum *SINR* of the cellular user when it can work normally. Therefore, the constraint condition to achieve the goal of maximizing throughput is that the *SINR* of each cellular user should be higher than the threshold value. At the same time, the transmitting power of each D2D user should be kept within the allowable range.

3 Algorithm Description

3.1 Q Learning

Q learning algorithm is a model-independent reinforcement learning algorithm proposed by Watkins [13] in 1989. The algorithm does not need to know the environment model, and at the same time, it updates the policy in the continuous learning process, and finally obtains an optimal policy to solve the problem.

The value function in Q learning is defined by Q function:

$$Q_t(s, a) = E_\pi[\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i} | S_t = s, A_t = a] \tag{11}$$

where $Q_t(s, a)$ represents the expectation of the gain from taking action a when the state is s, and γ represents the discount factor.

Therefore, it can be seen that the main idea of Q learning algorithm is to build state s and action a into a two-dimensional table to store the value of Q function, that is, the Q value table. As shown in Table 1, the rows in the Q table represent the state, and the columns represent the action to be selected, and the corresponding values of $Q_t(s, a)$ between them mean the feedback given by the environment when the action a is executed under the state s.

Table 1. Q table

Q table	a_1	a_2	$\cdots$	a_N
s_1	$Q(s_1, a_1)$	$Q(s_1, a_2)$	$\cdots$	$Q(s_1, a_N)$
s_2	$Q(s_2, a_1)$	$Q(s_2, a_2)$	$\cdots$	$Q(s_2, a_N)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
s_M	$Q(s_M, a_1)$	$Q(s_M, a_2)$	$\cdots$	$Q(s_M, a_N)$

The Basic Process of Q Learning

1. A Q table is firstly created, and it can be used to store values of $Q(s, a)$. Then initialize all values of $Q(s, a)$ to 0.
2. This paper adopts $\varepsilon - greedy$ policy to choose action. When the random probability is less than ε, the action will be chosen randomly; otherwise, the current action with the highest Q value will be chosen. This algorithm can be defined by the equation:

$$A = \begin{cases} \text{Random action } a \in A, & p < \varepsilon \\ \arg\max\limits_{a} Q(s, a), & \text{others} \end{cases} \tag{12}$$

   where p is the random probability in the process of iteration.
3. The action that has been chosen is adopted.
4. The return function value is calculated through the feedback given by the environment.
5. According to the return function and Eq. (13), it can update Q table.

$$Q_{t+1}^r(s, a) = Q_t^r(s, a) + \alpha[r_{t+1} + \gamma \max_{a'} Q_t^r(s', a') - Q_t^r(s, a)] \tag{13}$$

   where $Q_t^r(s, a)$ represents the value of Q function on the resource block Br during the t-th iteration; $\alpha \in (0, 1]$ represents learning rate, which determines the degree to which the return function value affects the update of Q value in the iteration process. When the learning rate is small enough and the learning process can access all the combinations of states and actions for many times, the update and iteration are carried out according to Eq. (13), and finally Q function will converge to the optimal value Q^*. The γ is the discount factor with a value between 0 and 1, which

is used to determine the relative proportion of delayed return and current return. The larger the value of γ, the more attention is paid to long term returns. Since the final result of Q learning is Q^*, the optimal policy obtained by learning is to select the optimal action a under the state s to make Q value reach maximum.

6. The iterative process from step 2 to step 5 is repeated until the Q value converges.

3.2 Definition of Basic Components

This section solves the power allocation problem of D2D communication in the cellular network based on the Q learning algorithm. For the system model which we are considering, different user equipment occupies different resource block, so there is no interference with user equipment. Therefore, the problem of throughput optimization for each resource block in the whole system can be regarded as an independent Q learning problem. Firstly, the D2D user equipment, system throughput, and transmitting power of D2D user are related to the components of reinforcement learning, and then they are introduced one by one.

Agent. It is the performer of the action. Each D2D user is an agent in the communication system.

State. According to the related theory of reinforcement learning, it can be known that the agent will have an influence on the environment after executing the action, which makes the environment change its state after receiving the action. And in the subsequent learning process, we can get the optimal policy finally by changing actions and states to explore and adjust policy, so that we can know what actions should be taken under state s. However, in the transition process of different states, the time required for the Q value to converge to the optimal value becomes longer. Therefore, the following power allocation algorithm based on Q learning will adopt the single-state Q learning algorithm, that is, the state has no practical significance, which not only simplifies the Q table, but also shortens the convergence time of Q learning.

Action. The action performed by the agent is defined as the transmitting power of D2D user, denoted as $p \in \{p_1, p_2, \cdots, p_L\}$. There are L powers that D2D transmitter can choose.

Return Function. We can define the throughput of resource block Br as:

$$r = \begin{cases} W \cdot \log_2(1 + SINR^r_{C_i}) + W \cdot \log_2(1 + SINR^r_{D_j}), & SINR^r_{C_i} \geq \tau_0 \\ -1, & \text{others} \end{cases} \tag{14}$$

According to Eq. (14), if the $SINR$ of the cellular user occupying the resource block Br is greater than the minimum threshold τ_0, the value of return function is the sum of the throughput of all mobile user equipment occupying this resource block. Otherwise, the return function is equal to -1, which is regarded as punishment.

3.3 Steps of the Algorithm

1. Firstly, we should input K orthogonal spectrum resource blocks, M cellular users, N D2D users and L powers that can be chosen.
2. For the j-th D2D user to create and initialize the Q table, there is only a one-dimensional table to create.
3. The resource block Br that has been allocated for the D2D user is chosen. Firstly according to $\varepsilon - greedy$ policy, the agent chooses the action a and adopt the action a, that is, transmitting power is allocated to D2D user through policy. Then calculate the return function and update the Q table on the basis of the Eq. (14). Finally, this process is repeated until the Q table converges, and the corresponding optimal action in the Q table is obtained, that is, D2D user obtains the optimal transmitting power.

4 Simulation Results

This section analyzes the performance of the power allocation algorithm based on machine learning on the basis of the above system model and algorithm description. The simulation experiment is carried out on the Python3 simulation platform. In the simulation process, this paper considers a single-cell scene with a coverage radius of 500 m. D2D users share the uplink resources with cellular users, and different user equipment is distributed randomly and evenly in the cell. The number of resource blocks in the system is equal to the number of cellular users. The simulation parameters are shown in Table 2.

Table 2. Simulation parameters

Parameter	Value
Number of D2D user	1–100
Cell radius	500 m
Transmitting power of cellular user	24 dBm
Selectable powers set of D2D user	{1, 6.5, 12, 17.5, 23} dBm
Noise power density (σ^2)	−134 dBm/Hz
Gain index (k)	1
Path loss index (μ)	4
Channel gain model of cellular user	$128 + 37.6 \lg(d(\text{km}))$ dB
Minimum SINR of cellular user (τ_0)	3 dB
Distance between D2D user	50 m
Bandwidth of RB (W)	180 kHz
Learning rate (α)	0.5
Discount factor (γ)	0.9
Initial value of ε	0.9

As shown in Fig. 2, a single-cell scene with a radius of 500 m is drawn. Orange represents the base station, which is located in the center of the cell; blue represents the cellular user equipment, red represents the D2D transmitting user equipment, and green represents the receiving user equipment which is 50 m away from the transmitting user.

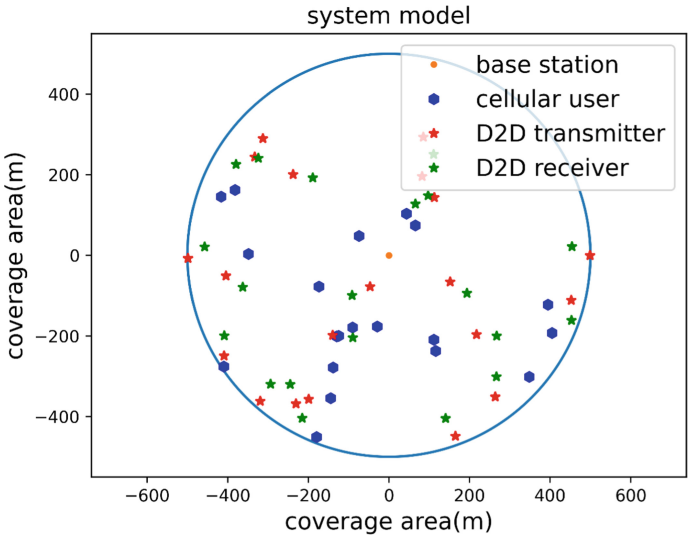

Fig. 2. System model for simulation. (Color figure online)

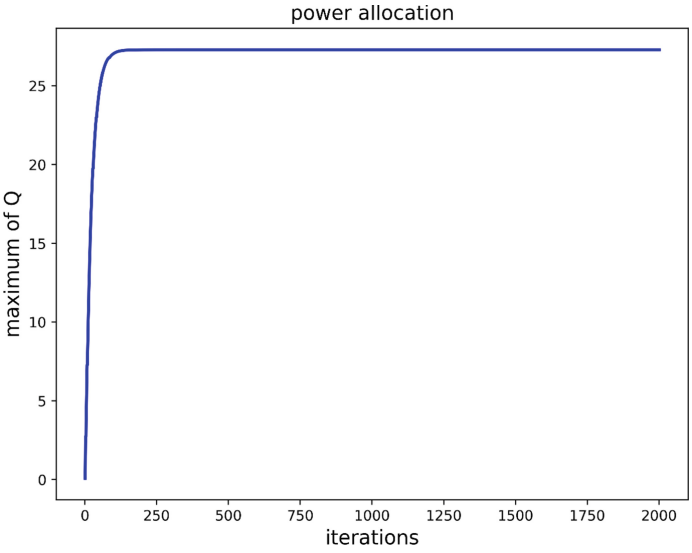

Fig. 3. Convergence of power allocation algorithm

Figure 3 shows the convergence of Q value of the power allocation algorithm based on Q learning on the same resource block. During the simulation, since this paper sets that a cellular user can be multiplexed by at most one D2D user for uplink resources, $M = 1$ and $N = 1$ are set to evaluate the convergence of the algorithm. The abscissa represents the number of iterations and the ordinate represents the maximum value of Q for each iteration. Therefore, the simulation result shows that the value of Q will eventually converge to the optimal value, and D2D user can obtain the optimal transmitting power.

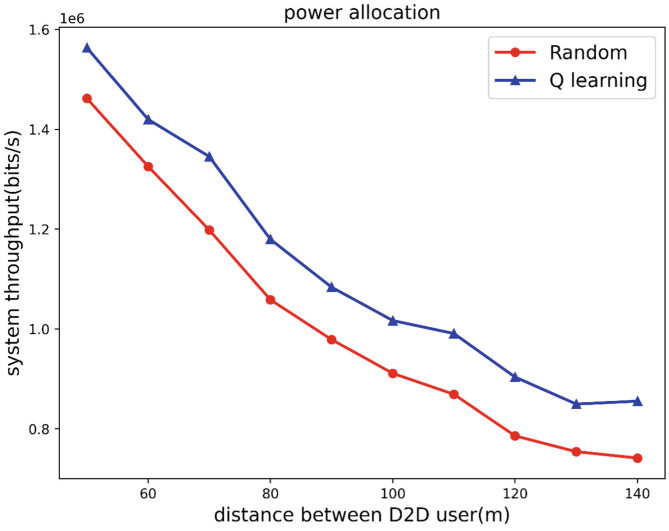

Fig. 4. Performance comparison of different power allocation algorithms

Figure 4 shows the variation of system throughput on a resource block. The abscissa represents the distance between D2D user and the ordinate represents the throughput of system. As the distance between D2D user increases, the system throughput of the two algorithms decreases gradually. The performance of the algorithm in the paper is better.

Figure 5 shows that the total throughput of the system increases significantly with the increase of number of D2D users, that is to say, the access of D2D users in the system improves the performance of the system. The performance of the power allocation algorithm based on Q learning in this paper is better than the random power allocation algorithm.

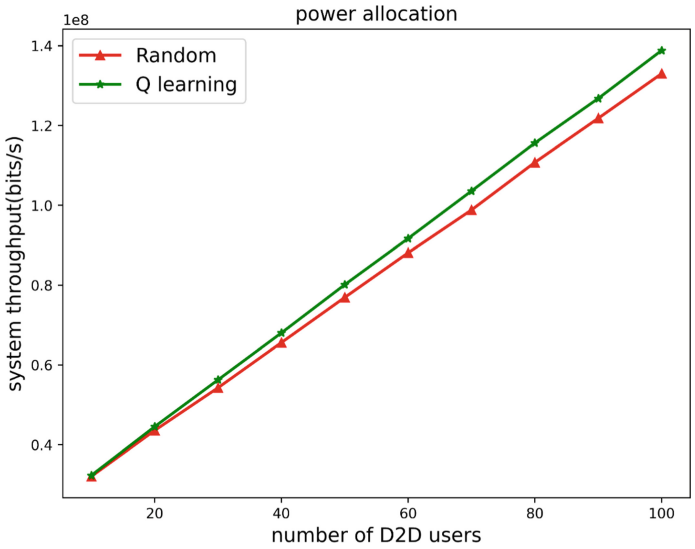

Fig. 5. Variation of system throughput with number of D2D users

5 Conclusion

The power allocation method of D2D communication in cellular system is studied in this paper. On the basis of studying and researching on the theory of D2D communication and machine learning, a power allocation algorithm for D2D communication based on Q learning is proposed. Based on the consideration of the SINR of cellular users, the algorithm aims at maximizing the system throughput. Each D2D user, as an agent, uses the $\varepsilon - greedy$ policy to select the transmitting power as the next action to be executed in the iteration process. Then the conditional throughput on the corresponding resource block is calculated, that is, the return function value, and then the Q table is updated. In the iteration, the agent will find the optimal selection policy and get the optimal action under the optimal policy. According to the simulation results, it can be observed that the power allocation algorithm based on Q learning is superior to the traditional random power allocation algorithm.

References

1. Doppler, K., et al.: Device-to-device communication as an underlay to LTE-advanced networks. IEEE Commun. Mag. **47**(12), 42–49 (2009)
2. Adnan, M.H., Zukarnain, Z.A.: Device-to-device communication in 5G environment: issues, solutions, and challenges. Symmetry **12**(11), 1762 (2020)
3. Su, L., et al.: The research of key technologies in the fifth-generation mobile communication system. In: International Industrial Informatics & Computer Engineering Conference, pp. 483–487 (2015)
4. Asadi, A., Wang, Q., Mancuso, V.: A survey on device-to-device communication in cellular networks. Commun. Surv. Tutor. **16**(4), 1801–1819 (2014)

5. Mjolsness, E., et al.: Machine learning for science: state of the art and future prospects. Science **293**(5537), 2051–2055 (2001)
6. Ray, S.: A quick review of machine learning algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 35–39
7. Cai, X., et al.: A capacity oriented resource allocation algorithm for device-to-device communication in mobile cellular networks. In: IEEE International Conference on Communications, pp. 2233–2238 (2014)
8. Xing, H., Hakola, S.: The investigation of power control schemes for a device-to-device communication integrated into ofdma cellular system. In: IEEE International Symposium on Personal Indoor & Mobile Radio Communications, pp. 1775–1780 (2010)
9. Sun, C., et al.: Distributed power control for device-to-device network using stackelberg game. In: 2014 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1344–1249 (2014)
10. Luo, Y., Shi, Z., Zhou, X., Liu, Q., Yi, Q.: Dynamic resource allocations based on q-learning for d2d communication in cellular networks. In: 2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) IEEE, pp. 19–21 (2014)
11. Nie, S., Fan, Z., Zhao, M., Gu, X., Zhang, L.: Q-learning based power control algorithm for D2D communication. In: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) IEEE, pp. 1–6 (2016)
12. Fan, Z., et al.: D2D power control based on supervised and unsupervised learning. In: 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 558–563 (2017)
13. Watkins, C., Dayan, P.: Q-learning. Mach. Learn. **8**(3–4), 279–292 (1999)