Distributed Dynamic Spectrum Access for D2D Communications Underlying Cellular Networks Using Deep Reinforcement Learning

Zhifeng Jiang[1], Liang Han[1,2(✉)], and Xiaocheng Wang[1,2]

[1] College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
hanliang@tjnu.edu.cn

[2] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

Abstract. In this paper, we investigate a deep Q-network (DQN)-based method for applying a dynamic spectrum access model to device-to-device (D2D) communications underlying cellular networks. Dynamic spectrum access (DSA) devices have two critical concerns, namely avoiding interference to primary users (PUs) and interference coordination with other secondary users (SUs). We consider that the issues faced by DSA users are also applicable to the D2D communication underlying cellular network. Therefore, we propose a distributed dynamic spectrum access scheme based on deep reinforcement learning (DRL). It enables each D2D user to learn a reliable spectrum access policy through imperfect spectrum sensing without knowledge of system prior information, avoiding collisions with cellular users and other D2D users and maximizing system throughput. Finally, the simulation results demonstrate the effectiveness of our proposed dynamic spectrum access scheme.

Keywords: Device-to-device (D2D) communication · Distributed dynamic spectrum access (DSA) · Deep reinforcement learning (DRL)

1 Introduction

Radio spectrum resources are an essential resource. According to a white paper published by Cisco on global mobile data traffic forecasts for 2017–2022, global mobile data traffic will grow sevenfold between 2017 and 2022 [1]. However, related studies have revealed a phenomenon that many spectrum resources are not used effectively [2,3]. D2D communication technology is considered a feasible solution to the problem of poor spectrum resources, with the advantages of improved spectrum efficiency and reduced communication delays [4,5]. Furthermore, considering the limitations of traditional static spectrum allocation policies, dynamic spectrum access techniques have also been proposed to improve spectrum efficiency [6]. In D2D communication technology, cellular users are

Q. Liang et al. (Eds.): AIC 2022, LNEE 871, pp. 340–348, 2023.
https://doi.org/10.1007/978-981-99-1256-8_40

subject to severe interference when D2D users share the same spectrum as cellular users and strong interference between D2D users can also seriously affect the quality of communication [4,7,8]. Similarly, dynamic spectrum access also faces two fundamental problems, namely interference coordination between DSA users and interference suppression for primary users [9].

Previous research has proposed a number of schemes for spectrum allocation between D2D users and cellular users [10,11]. These studies investigated the reuse of cellular user resources by D2D communication users in a non-orthogonal spectrum allocation. In [10], the authors proposed a distributed Q-learning-based spectrum allocation scheme to maximize D2D user system throughput and maintain the QoS requirements for cellular users. In [11], the authors proposed a distributed DRL-based spectrum allocation scheme to address the issue of interference and resource allocation between D2D and cellular users, with the aim of maximizing system throughput. However, little existing research has considered the use of distributed spectrum access schemes to avoid conflicts between D2D communication users and cellular users as well as other D2D users.

In this paper, we consider an uplink scenario of a D2D underlying cellular communication network, and to address the collision problem between DUEs and CUEs, we propose a DRL-based distributed dynamic spectrum access scheme. In particular, we introduce the concept of a "reusable area" [12], where D2D users can choose the number of reusable CUEs based on the range of "reusable areas". According to the DRL theory, we enable each agent to learn the optimal access policy only through imperfect spectrum sensing without knowing the system a priori information, increasing the system throughput while avoiding collisions with other DUEs and CUEs.

2 System Model

We consider a dynamic spectrum access scenario in the uplink of a D2D underlying cellular network. As shown in Fig. 1, the system model includes N cellular users (CUEs) denoted by $\mathbb{N} = \{1, 2, \ldots, N\}$ and K pairs of D2D users denoted by $\mathbb{K} = \{1, 2, \ldots, K\}$, each D2D pair consists of a set of transmitters (DTx) and receivers (DRx). d_{ii} denotes the distance between DTx and DRx, d_{jk} denotes the distance between the CUEs and the BS, and d_{ik} denotes the distance between DTx and the BS. We assume that the system has N channels and that each CUE transmits on a unique channel, thus avoiding interference between CUEs.

2.1 Channel Model

We adopt the WINNER II channel model to calculate the path loss generated by the signal propagation in space [15], which is described as a distance-dependent function

$$PL\left(d, f_c\right) = \overline{PL} + B\log_{10}(d[m]) + C\log_{10}\left(\frac{f_c\left(GH_z\right)}{5}\right), \tag{1}$$

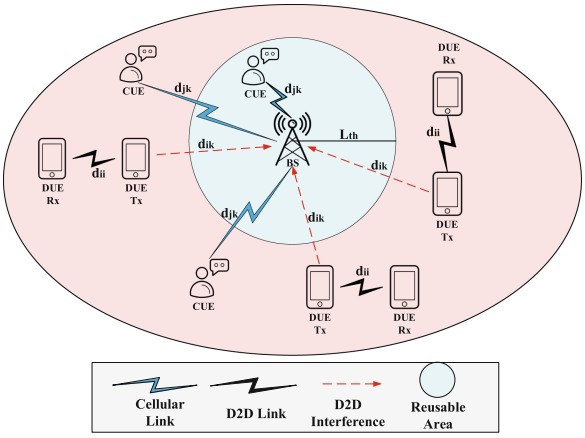

Fig. 1. Uplink scenario for D2D communications underlying cellular networks.

where f_c denotes the carrier frequency, $\overline{PL}$, B and C denote the unit distance loss reference, the path loss exponent and the path loss frequency dependence, respectively. For simplicity, we assume the existence of a strong line of sight (LOS) path between the signal transmission links. Therefore, our model can use the Rician channel model for channel modeling [13], which can be expressed as

$$h = \sqrt{\frac{\kappa}{\kappa+1}}\sigma e^{j\theta} + \sqrt{\frac{1}{\kappa+1}}CN\left(0,\sigma^2\right) \tag{2}$$

where $\sigma^2 = 10^{-\frac{\overline{PL}+B\cdot\log_{10}(d[m])+C\cdot\log_{10}\left(\frac{f_c[GHz]}{5}\right)}{10}}$ is determined by path loss, κ means the $\kappa - factor$, defined as the power ratio of the LOS component to the scattering component, θ denotes the phase and takes the value of a uniform distribution between 0 and 1, $CN(\cdot)$ denotes a circularly symmetric complex Gaussian random variable.

2.2 Uplink Signal Model

For the uplink scenario, when DUEs and CUEs transmit in the same time slot, DUEs can cause harmful interference to CUEs. Hence, the instantaneous signal to interference plus noise ratio (SINR) received by the BS from the CUEs can be expressed as

$$SINR_j = \frac{P_c \cdot |h_{jk}|^2}{P_d \cdot |h_{ik}|^2 + B \cdot N_0} \tag{3}$$

where P_c and P_d represent the transmit power of the CUE and the DTx, respectively. $|h_{jk}|^2$ denotes the channel gain of the cellular link, and $|h_{ik}|^2$ denotes the channel gain from the i^{th} D2D transmitter to the BS, , which can be derived according to (2). B and N_0 represent the channel bandwidth and noise spectral density, respectively. Furthermore, we assume in this model that each channel can be used by at most one D2D pair.

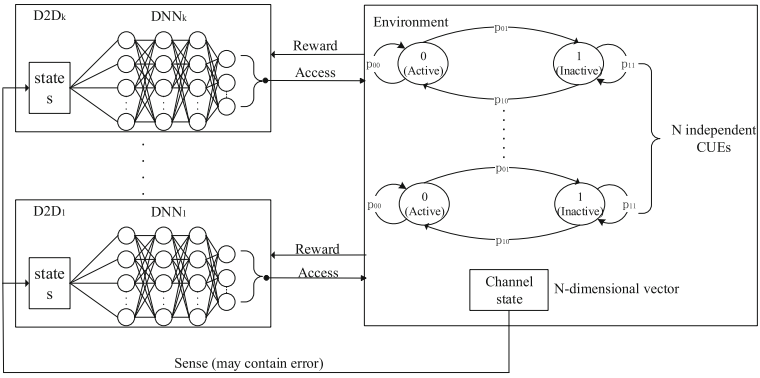

Fig. 2. Dynamic spectrum access framework based on D2D underlying cellular networks.

3 DRL-based Dynamic Spectrum Access Scheme

3.1 Deep Reinforcement Learning

A reinforcement learning model contains three components: possible states in the environment, possible actions that the agent may take based on a policy π, and a feedback reward function that the agent receives after making an action. These three components are defined as s_t, a_t, and r_{t+1}. The goal of the agent is to learn an optimal policy π^* to maximize the cumulative discount reward $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+1+i}$, where $\gamma \in [0, 1]$ represents the discount factor. Q-values are updated with the following rules:

$$\begin{aligned} Q\left(s_t, a_t\right) = & Q\left(s_t, a_t\right) + \\ & \alpha\left(r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right), \end{aligned} \tag{4}$$

where $\alpha \in (0, 1]$ is the learning rate. Furthermore, the policy function π is updated by means of the ϵ-greedy algorithm.

DRL uses deep neural network (DNN) to approximate Q values (DQN), i.e. $Q(s_t, a_t; \boldsymbol{\theta}) \approx Q(s_t, a_t)$, where $\boldsymbol{\theta}$ is the network weights. In DQN, the TD algorithm is mostly used to calculate the loss function,

$$Loss(\theta) = E[(y_t - Q(s_t, a_t; \boldsymbol{\theta}))^2] \tag{5}$$

where y_t represents the target Q-value and is defined as

$$y_t = r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a; \boldsymbol{\theta}\right) \tag{6}$$

After that, the agent can minimize the loss function by the gradient descent algorithm as follows:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} + \alpha E\left[\left(y_t - Q\left(s_t, a_t; \boldsymbol{\theta}\right)\right) \nabla Q\left(s_t, a_t; \boldsymbol{\theta}\right)\right] \tag{7}$$

where α is the learning rate.

3.2 Uplink Dynamic Spectrum Access Framework

In our proposed scheme, as shown in Fig. 2, we have designed the distributed dynamic spectrum access as a deep reinforcement learning model. In Sect. 2, we show that the system model consists of N channels and K D2D users, where each channel is occupied by a CUE. Therefore, we describe the channel states as *Active* and *Inactive*, denoted by 0 and 1 respectively, where *Active* indicates that the channel is occupied by a CUE and *Inactive* indicates that the channel is not occupied by a CUE. In addition, the detailed definitions of "state", "action" and "reward" are as follows.

State: At the beginning of each time slot, each D2D pair will sense the channel state in the environment, which may contain errors. Subsequently, the agent uses the sensed results as input data for the neural network to be trained. Therefore, the state space $\mathbf{S}^k(t)$ of each D2D pair is defined as $\mathbf{S}^k(t) = [s_1^k(t), \ldots, s_n^k(t)]$, where $\mathbf{S}^k(t)$ denotes an N-dimensional vector, k denotes the k^{th} D2D pair, n denotes the number of channels and $s_n^k(t)$ denotes the state of the channel (*Active* or *Inactive*).

Action: The agent decides whether to access and which wireless channel to access based on the spectrum access policy. Hence, the action space A can be defined as $A \in \{0, 1, \ldots, N\}$, where $a_t = 0$ means that the agent does not access the channel and $a_t = N$ means that the agent accesses the n^{th} channel.

Rewards: According to the situations that the agent may face after making an action choice, the following reward function setting scheme is developed.

1. The D2D pair collides with the CUE. This indicates that the D2D pair accesses the channel where the CUE is located when the cellular link resources cannot be reused. In Sect. 2, we mention the concept of warning signals. Therefore, we give a penalty value of -4 as the result of a warning signal being received by the agent. For convenience, we define this case as $\mathbb{C}$.
2. A collision between D2D users. This case indicates that different D2D users are accessing the same channel. We set the reward value for this case to 0 and define this case as $\mathbb{D}$.
3. The D2D pairs do not access any channel. We set the reward value for this case to 1. Similarly, we define this case as $\mathbb{I}$.
4. The D2D pair successfully accesses the channel. The reward value for this case should be set to the maximum. We considered the normalized $S\hat{I}\hat{N}R_j$ and applied it to our reward function setting, described as $1 + \log_2(1 + S\hat{I}\hat{N}R_j)$. We define this case as $\mathbb{S}$.

In summary, the reward function for the k^{th} D2D pair on the n^{th} channel can be described as

$$r_{t+1}^k = \begin{cases} -4, & \text{Case } \mathbb{C} \\ 0, & \text{Case } \mathbb{D} \\ 1, & \text{Case } \mathbb{I} \\ 1 + \log_2(1 + S\hat{I}\hat{N}R_j), & \text{Case } \mathbb{S} \end{cases} \tag{8}$$

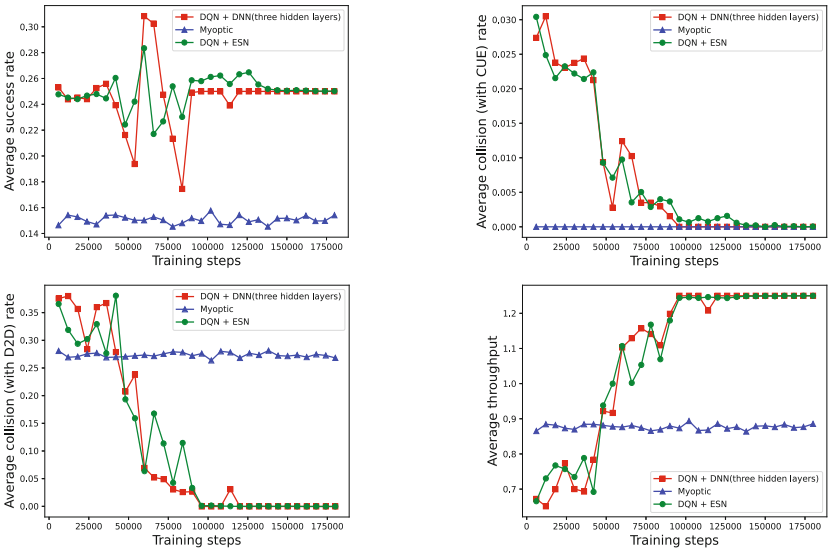

Fig. 3. Performance evaluation in non-orthogonal scenarios ($P = \frac{1}{2}R$).

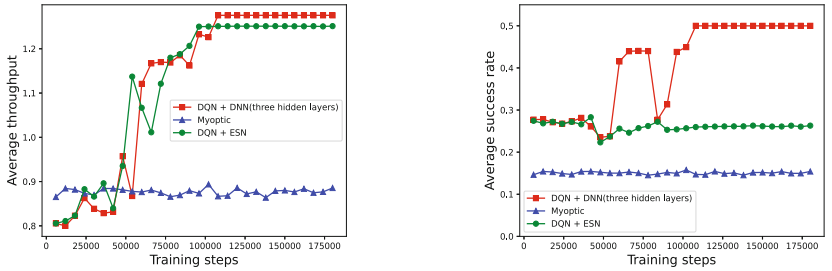

Fig. 4. Performance evaluation in non-orthogonal scenarios ($P = R$).

4 Performance Evaluation

In this section, we evaluate the performance of the algorithmic framework proposed in the scheme. Specifically, we compare the algorithm used in the scheme with the Myopic algorithm [14] based on a priori information about the system and the DQN+ESN [13] to verify the performance.

In our scheme, we consider a cellular cell scenario with a radius of 100m. The locations of the D2D pair and the CUE in this cell are randomly generated and we specify that the communication distance between the transmitter and the receiver of the D2D pair is randomly generated in the range of 20 m and 40 m. Therefore, the location of the CUE may fall within the "reusable area" that we have defined. We express the percentage of the reusable area in the cell by $P = \frac{\pi \cdot L_{th}^2}{\pi \cdot R^2}$, where the choice of the threshold distance L_{th} is determined by the value of P. The specific simulation parameters are shown in Table 1.

Table 1. Simulation parameters

Parameter	Value
Cell radius	100 m
Carrier frequency	2 GHz
Bandwidth	2 MHz
D2D transmit power P_d	23 dBm
CUE transmit power P_C	23 dBm
Noise power density N_0	-174 dBm/Hz
Path loss reference $\overline{PL}$	41
Path loss exponent B	22.7
Path loss frequency dependencer C	20
κ factor	8
Learning rate α	0.01
Discount factor γ	0.5
Exploration ϵ	$0.7 \longrightarrow 0$
Spectrum sensing error probability	$(0, 0.2)$

In this scenario, we set the number of D2D users to 4 and the number of CUEs to 2. The positions of the CUEs are randomly generated, and by calculating the distances from the CUEs to the BS, we can obtain the distances to be 95.75 m and 30.36 m. Therefore, we first consider the case of $P = \frac{1}{2}R$, after which we can calculate the corresponding threshold distance L_{th} of $50\sqrt{2}$ m. According to the range of reusable zones corresponding to the threshold distance, the D2D pair can reuse one CUE resource. The simulation results are shown in Fig. 3, where the performance obtained using the DQN-based method is significantly better than using the Myopic algorithm. In particular, the Myopic algorithm selects the action with the greatest immediate reward and therefore it performs well in avoiding collisions with the CUE. However, simulation results show that using the DQN-based method after training also maximizes throughput while achieving collision avoidance. In this scenario, our scheme performs approximately the same as DQN+ESN. Additionally, we consider the non-orthogonal access scenario when $P = R$. In this scenario, the reusable area covers the entire cellular cell. Therefore, all CUEs in the cell can be reused by the D2D pair. The simulation result is shown in Fig. 4, which shows that our scheme achieves the theoretical maximum success rate and has better throughput.

5 Conclusion

In this paper, we investigated the case of dynamic spectrum access in the uplink of a D2D underlying cellular network. Specifically, we proposed a distributed dynamic spectrum access scheme under imperfect spectrum sensing conditions,

which aims to allow D2D users to learn an optimal spectrum access policy to maximize throughput without knowing a priori information. Besides, we introduced the concept of a reusable area, where the D2D user can choose the number of reusable CUEs based on the coverage of that area. Simulation results show that our scheme can avoid collisions while maximizing the system throughput.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (62001327, 61701345), Natural Science Foundation of Tianjin (18JCZDJC31900).

References

1. Forecast, G., et al.: Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. Update **2017**, 2022 (2019)
2. Yin, S., Chen, D., Zhang, Q., Liu, M., Li, S.: Mining spectrum usage data: a large-scale spectrum measurement study. IEEE Trans. Mob. Comput. **11**(6), 1033–1046 (2012)
3. McHenry, M.A., Tenhula, P.A., McCloskey, D., Roberson, D.A., Hood, C.S.: Chicago spectrum occupancy measurements & analysis and a long-term studies proposal. In: Proceedings of the first International Workshop on Technology and Policy for Accessing Spectrum, August 2006
4. Asadi, A., Wang, Q., Mancuso, V.: A survey on device-to-device communication in cellular networks. IEEE Commun. Surv. Tutorials **16**(4), 1801–1819 (2014)
5. Ansari, R.I., Chrysostomou, C., Hassan, S.A., Guizani, M., Mumtaz, S., Rodriguez, J., Rodrigues, J.J.: 5g d2d networks: techniques, challenges, and future prospects. IEEE Syst. J. **12**(4), 3970–3984 (2018)
6. Kolodzy, P., Avoidance, I.: Spectrum policy task force. Federal Commun. Comm. Washington, DC, Rep. ET Docket **40**(4), 147–158 (2002)
7. Shah, S.W.H., Mian, A.N., Crowcroft, J.: Statistical qos guarantees for licensed-unlicensed spectrum interoperable d2d communication. IEEE Access **8**, 27 277–27 290 (2020)
8. Kai, Y., Wang, J., Zhu, H., Wang, J.: Resource allocation and performance analysis of cellular-assisted OFDMA device-to-device communications. IEEE Trans. Wirel. Commun. **18**(1), 416–431 (2019)
9. Song, H., Liu, L., Ashdown, J., Yi, Y.: A deep reinforcement learning framework for spectrum management in dynamic spectrum access. IEEE Internet Things J. 8(14), 11 208–11 218 (2021)
10. Zia, K., Javed, N., Sial, M.N., Ahmed, S., Pirzada, A.A., Pervez, F.: A distributed multi-agent RL-based autonomous spectrum allocation scheme in d2d enabled multi-tier hetnets. IEEE Access **7**, 6733–6745 (2019)
11. Gong, P.-Y., Wang, C.-H, Sheu, J.-P., Yang, D.-N.: Distributed drl-based resource allocation for multicast d2d communications. In: 2021 IEEE Global Communications Conference (GLOBECOM), pp. 01–06. December 2021
12. Huang, J., Yang, Y., He, G., Xiao, Y., Liu, J.: Deep reinforcement learning-based dynamic spectrum access for d2d communication underlay cellular networks. IEEE Commun. Lett. **25**(8), 2614–2618 (2021)
13. Chang, H.-H., Song, H., Yi, Y., Zhang, J., He, H., Liu, L.: Distributive dynamic spectrum access through deep reinforcement learning: a reservoir computing-based approach. IEEE Internet Things J. **6**(2), 1938–1948 (2019)

14. Zhao, Q., Krishnamachari, B., Liu, K.: On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. IEEE Trans. Wirel. Commun. **7**(12), 5431–5440 (2008)
15. Meinilä, J., Kyösti, P., Jämsä, T., Hentilä, L.: Winner ii channel models. In: Radio Technologies and Concepts for IMT-Advanced, February 2008