ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.02.004

Research on Multi-user Dynamic Spectrum Allocation Strategy Using Reinforcement Learning in Unknown Environments

Shilong Cao [1], Fei Lin [2 +] and Xianzhi Jin [3]

[1] School of Electrical Engineering and Automation, Qilu University of Technology (Shandong Academy of Sciences) Jinan, China

[2] School of Electronic and Information Engineering (Department of Physics), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[3] School of Electrical Engineering and Automation, Qilu University of Technology (Shandong Academy of Sciences) Jinan, China

Abstract. With the advent of 5g era, the demand of spectrum expands, which leads to the shortage of spectrum resources. Because of the low utilization of spectrum resources, it is very important to find efficient spectrum allocation strategy for wireless communication. Cognitive Radio technology will become the key to solving this problem [1]. This paper proposes a multi-user system model that conforms to the actual 5G communication situation, and uses the reinforcement learning Deep-Q-Network (DQN) algorithm to study its dynamic spectrum allocation problem. The simulation results show that the algorithm under this model can converge quickly and improve the efficiency of spectrum resource utilization.

Keywords: Cognitive Radio, multi-user, 5G, DQN, dynamic spectrum allocation.

1. Introduction

Frequency spectrum has become an important issue restricting the development of mobile communication systems. In 5G, there are two direct methods to solve this problem: release the allocated frequency for use by the 5G system; the other is to use a higher millimeter wave frequency band for communication. However, the above two methods still have limitations. The Federal Communications Commission (FCC) has researched and pointed out that a large number of allocated spectrum resources are idle to a large extent in time and space. The average utilization rate of spectrum at any time and any place does not exceed 5% [2]. To this end, the industry is looking for a third path-optimization and utilization of spectrum. To achieve this goal is the Cognitive Radio technology proposed by Dr. Joseph Mitola in 1999 [3].

Cognitive radio systems usually have two basic users: licensed user (LU) and cognitive user (CU). The sharing of spectrum resources between LU and CU is the core idea of cognitive radio: The CU can improve the spectrum utilization by sensing the surrounding radio environment and opportunistically accessing the spectrum without causing interference to the LU. This technology realizes the access of multiple frequency bands through dynamic spectrum allocation technology and makes full use of idle spectrum.

[4] proposed a simple auction mechanism based on cognitive radio networks, but the author only considers a relatively simple situation in which each channel can be allocated to at most one user, It does not consider the situation when there are multiple users per channel.

In [5], the author proposes a reinforcement learning scheme that uses two alternative update rules to determine the sensing order of available channels and compares them with two existing channel selection schemes. But in this paper, LU and CU are regarded as independent agents to operate, and the mutual influence between LU and CU is not considered.[6] uses Universal Software Radio Peripheral (USRP) equipment to study radio frequency communications and used Q-learning to analyze incoming signal and adjust gain on the radio by using LabView to control USRP. However, to realize the conclusions of this

[+] Corresponding author. Tel: +86 13405411066
E-mail address: linfei@qlu.edu.cn

paper, a large amount of channel prior knowledge is required, which will be a challenge in the future complex wireless communication environment.

In practical communication, it is necessary to consider not only the interaction between multi-user, but also the complex channel state without prior knowledge. Therefore, this article uses Deep-Q-Network (DQN) to solve the spectrum allocation problem of multi-user in 5G multi-user unknown environment. The content of this article is arranged as follows: firstly, the background knowledge part introduces the concept of reinforcement learning (RL) and analyzes the advantages of DQN; secondly, it introduces the two major technologies in the system model of this article: V2X and D2D; thirdly, this paper designs a multi-user system which accords with the actual communication situation of 5G, and analyzes the performance requirements of different users; finally, the algorithm of this paper is analyzed, and the communication system model proposed in this paper is realized by simulation. The simulation results show that the proposed DQN algorithm can quickly converge and improve spectrum utilization.

2. Background Knowledge

This section introduces the background knowledge related to the research content of this article.

2.1. Reinforcement Learning

RL is a learning method that can understand the environment and behavior. It is very suitable for cognitive radio networks (CRN) where the network conditions in the future cannot be predicted RL is to study the effect of self-learning and adaptive agents on environment. Its goal is to maximize the return on actions taken. Agent only learns to improve performance by observing the state changes in its operating environment and the reward feedback received after taking measures [7], as shown in Fig. 1.

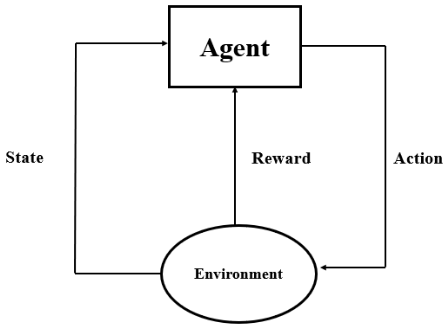


Fig. 1: Schematic diagram of interaction between agent and environment

Q-Learning algorithm is a popular model-free reinforcement learning technology, which has been used in various resource allocation schemes in CR. In [8], a channel spectrum access algorithm based on online synchronous Q-Learning is proposed to actively avoid Channel congestion in the cognitive radio network.

The working principle of the Q-Learning algorithm is to use a table to store each state, and the Q value of each action in this state, and gradually optimize the transmission parameters according to the rewards obtained through interaction with the environment. However, if the state space and the number of actions is large, the Q-Learning algorithm will be affected by the slow learning speed, which will reduce the anti-interference performance. In [9], Han studied a frequency domain anti-jamming communication game, and proposed a two-dimensional anti-jamming system based on the DQN algorithm. Using the deep convolutional neural network (CNN), the anti-jamming system based on DQN can solve the limitation of high-dimensional Q-Learning and accelerate the learning rate.

Therefore, this paper can solve this defect by establishing a DQN. DQN is an algorithm that combines neural network and Q-Learning [10]. Its basic idea is to use neural network to learn Q value and make Bellman formula, that is, a large number of parameters in a complex environment are outputted with a small amount of Q value through neural network, and then through Q-Learning optimization. DQN not only has the advantage of Q-Learning autonomously in environmental learning, but also uses neural networks to solve the problem of large action state space parameters that are difficult to converge. This algorithm can accelerate Q by using artificial neural networks to approximate the Q-value function at the core Convergence in learning.

In machine learning, neural networks are often used to deal with a large number of parameters. Therefore, states and actions can be regarded as the input of the neural network, and then the Q-value of the action can be obtained after the neural network analysis, so that there is no need to record in the table. It is to directly use the neural network to generate the Q-value. Then according to the principle of Q-Learning, the action with the maximum value is directly selected as the next action to be done. The probability of selecting the optimal action in this simulation is 0.9, that is, there is a probability of 0.1 to randomly select other actions. This is to prevent local the occurrence of the optimal solution situation. In general, the neural network receives various external parameter information, and finally selects actions through reinforcement learning, that is, DQN is a combination of neural network and Q-Learning.

The simulation structure in this paper is implemented by two neural networks, in which the target network is used to predict the Q-value of the target, and it will not update the parameters in time. Evaluation network is used to predict the Q-value of the evaluation. This neural network has the latest neural network parameters.

2.2. V2X and D2D

V2X is the direct connection communication technology for connecting vehicles and everything, intelligent networked cars. Simply put, models equipped with this system can automatically select the best driving route through the analysis of real-time traffic information in the automatic driving mode, thereby greatly alleviating traffic jams. V2X includes: V2V: vehicle to vehicle, V2P: vehicle to people, V2I: vehicle to infrastructure and so on. As shown in Fig. 2.

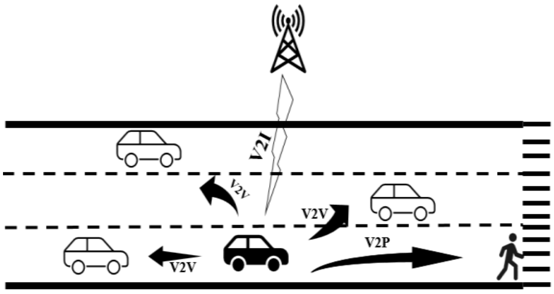


Fig. 2: Schematic diagram of V2X

Spectrum is the key resource of intelligent connected vehicle, it’s a challenge to apply 5G to V2X. At present, countries all over the world have issued policies to divide the 5.9GHz frequency band to V2X to meet its requirements of low delay and high speed, thus greatly improving its security.

Similar to the concept of Machine to Machine (M2M) in the Internet of Things, D2D aims to enable user communication devices within a certain distance to communicate directly [11] to reduce the load on the serving base station. The frequency band requirements are generally in the Industrial Scientific Medical Band (ISM) frequency band. At present, countries divide the 2.4GHz frequency band into the ISM frequency band. There is no license or fee to apply these frequency bands, and only a certain transmission power (generally less than 1w) is required. And don't cause interference to other frequency bands.

In 5G, the application scenarios of D2D communication can be divided into:

- Relay transmission: Relay users help users with poor edge signals to communicate with the base station and improve the coverage of the base station.
- Local business data transmission: location-based advertising, marketing, maps, dating services, etc.
- Emergency communication: emergency communication when the base station is damaged in a disaster.
- Smart home: mobile terminals become the control center of the home Internet of Things.

3. System Model

This section introduces the system model and spectrum allocation strategy proposed in this article.

ITU-R (International Telecommunication Union Radiocommunication Department) defines three usage scenarios for 5G: eMBB(Enhance Mobile Broadband), uRLLC (Ultra Reliable Low Latency Communications) and mMTC (Massive Machine Type Communications) [12]. The actual application of the three application scenarios and the challenges encountered in their implementation are shown in Table 1.

For V2X, its most concerned performance indicator is millisecond-level latency, followed by capacity. For mobile users, the most important thing is the reliability of channel data transmission, followed by capacity. For D2D, the most important thing is capacity, followed by reliability and other performance indicators.

Table 1: Three different usage scenarios of 5G

Three different usage scenarios of 5G	**eMBB**	**Practical application:** 4K/8K HD video; AR/VR; 3D holography, etc.
		Challenge: QoE，Data rate，Capacity.
	uRLLC	**Practical application:** Industrial manufacturing; Telemedicine; V2X, etc.
		Challenge:E2E latency, Data rate，QoE
	mMTC	**Practical application:** Smart home; smart city; D2D, etc.
		Challenge: Capacity, Massive number of connections, Data rate.

The system model of this paper is to apply DQN to the band allocation strategy in the application scenarios of V2X, mobile user data transmission and D2D relay transmission in 5G Internet of things environment to improve 5G spectrum utilization and communication system reliability. The system model is shown in Fig. 3. Suppose that in the V2X scene, street lights and billboards are regarded as the relay end in the D2D relay transmission application. You can choose whether to relay according to the channel conditions; regard the users in the car and the roadside as the relay end. As mobile users, their network experience priority is higher than D2D.

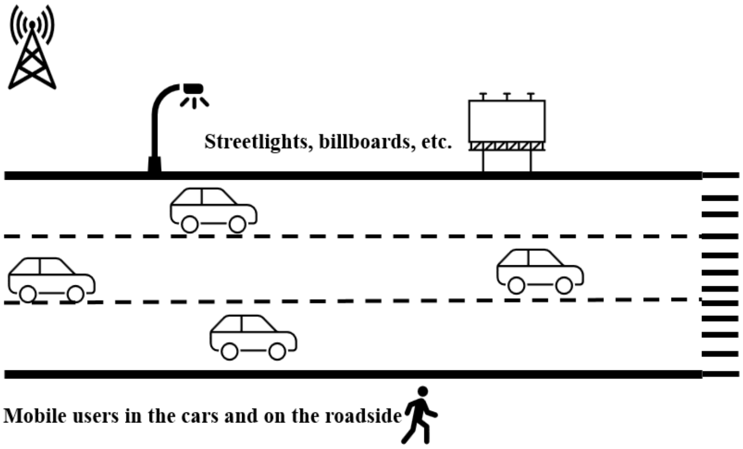

Streetlights, billboards, etc.
Mobile users in the cars and on the roadside

Fig. 3: System model

For the three types of users, the size of the frequency band is given priority, the bandwidth of the frequency band is secondly considered, and the most suitable channel is selected for access. Assume that the frequency band F_i and the corresponding bandwidth are shown in Table 2.

The communication system model in this paper makes the following assumptions for the optimal and suboptimal channels of the three users:

For V2X, the millisecond delay is the guarantee of vehicle safety, so the optimal frequency band is F_4. The larger the bandwidth, the more data can be transmitted. Therefore, if the F_4 is crowded in a certain environment and the channel quality is poor, the V2X sub-optimal channel can choose F_1

For mobile users, the reliability of data transmission is given priority, so the optimal frequency band is F_2. If the mobile user moves to an area with more users, the sub-optimal channel can be F_1.

For D2D relay transmission, lots of relay terminals require larger bandwidth. However, due to the short interval between terminals, F_1 is the best choice. If F_1 is occupied, F_3 is selected as the suboptimal channel of D2D is also a good choice.

Table 2: Frequency band and bandwidth allocation diagram of this system model

F_i	Spectrum name	bandwidth
F_1	Millimeter wave(mm wave)	100 MHz

F_2	N41, N78of 5G
	N41(2496-2690MHz)
	N78(3300-3800MHz)
		N41:194MHz;
		N78: 500MHz

F_3	ISM	50 MHz
F_4	Less than 6GHz	10 MHz

4. Multi-user Dynamic Spectrum Allocation Strategy in Different Unknown Environments

For a brief explanation, this paper only discusses the case of two channels. Firstly, this paper assumes that there are two channels, Ch1 and Ch2, and then there are two kinds of environment position 1 and position 2, and there is no prior knowledge about the channel for each user.

Secondly, in order to indicate whether the channel is occupied and the two channels are independent, two randomly generated numbers are used to represent the idle state of the two channels respectively. According to the research on spectrum utilization, it is known that the allocated spectrum is idle about 70% to 80% of the time, so this paper assumes that the probability of channel occupied is 30%.

Advance agreement: for Environment1 and environment2, the optimal channel of Environment1 is Ch1, and the suboptimal channel is Ch2. The optimal channel of environment2 is Ch2, and the suboptimal channel is Ch1.

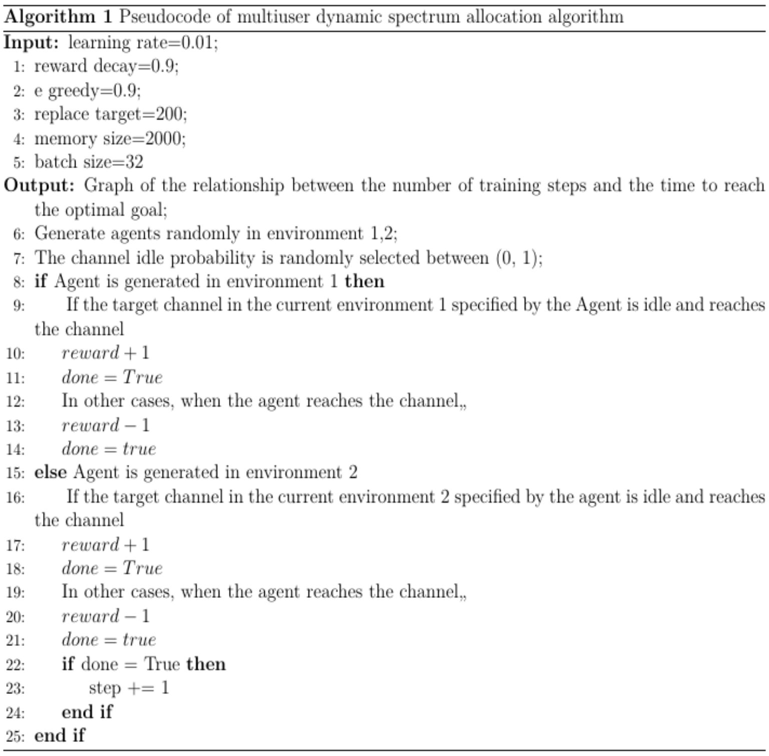
We only used two kinds of channels in the actual exercise, because regardless of two kinds of channels or multiple channels, only the optimal and sub-optimal solutions are considered for the Agent, and the third optimal solution is not considered. The pseudo code of the algorithm in this article is shown in Fig. 4.

When done = True, the next training will be performed. For the agent in environment2 there are similar reward achievement conditions. It should be noted that the optimal channel in environment2 is different from that in environment 1, that is, different environments correspond to different optimal channels.

When done = True, the next training will be performed. For the agent in environment2 there are similar reward achievement conditions. It should be noted that the optimal channel in environment2 is different from that in environment 1, that is, different environments correspond to different optimal channels.

5. Simulation Results

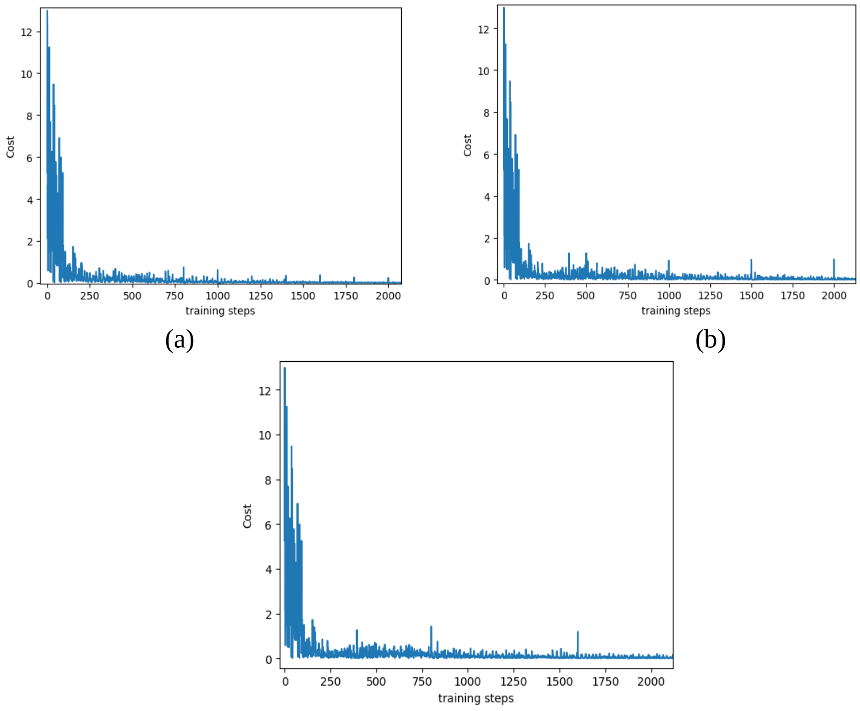
In the communication environment shown in Figure 3, the three types of communication modes are regarded as three types of users, and priority is given to them. Set V2X as the highest priority, which is determined by vehicle safety. It is beyond doubt that only a high-quality channel environment can obtain a low latency of milliseconds, thereby ensuring the safety of the vehicle. Secondly, this article considers the 5G experience of mobile users. For mobile users, under the premise of ensuring security, ultra-clear video look and feel, distortion-free voice calls and high-definition picture browsing depend on the reliability of the channel. Only with higher reliability can the user experience be further improved. Finally, consider the communication performance of D2D relay transmission. If ISM is available, select access. If there is no suitable frequency band, the V2X channel usage will be guaranteed first, and the user experience of mobile users will be considered second.



```
Algorithm 1 Pseudocode of multiuser dynamic spectrum allocation algorithm
Input: learning rate=0.01;
  reward decay=0.9;
  e greedy=0.9;
  replace target=200;
  memory size=2000;
  batch size=32
Output: Graph of the relationship between the number of training steps and the time to reach
    the optimal goal;
  Generate agents randomly in environment 1,2;
  The channel idle probability is randomly selected between (0, 1);
  if Agent is generated in environment 1 then
      If the target channel in the current environment 1 specified by the Agent is idle and reaches
    the channel
      $reward + 1$
      $done = True$
      In other cases, when the agent reaches the channel,,
      $reward - 1$
      $done = true$
  else Agent is generated in environment 2
      If the target channel in the current environment 2 specified by the agent is idle and reaches
    the channel
      $reward + 1$
      $done = True$
      In other cases, when the agent reaches the channel,,
      $reward - 1$
      $done = true$
      if done = True then
          step += 1
      end if
  end if
```

Fig. 4: The pseudo code of the algorithm

The simulation analyzed when the reward discount factor is 0.9, the probability of selecting the optimal action is 0.9, the batch size is 32, the learning rate is 0.01, 0.03, 0.06, and the replacement target is 200, 500, 800, respectively (that is, every 200 steps, 500 steps, 800 steps to update the neural network parameters), the user reaches the current optimal channel learning time curve in the unknown environment.



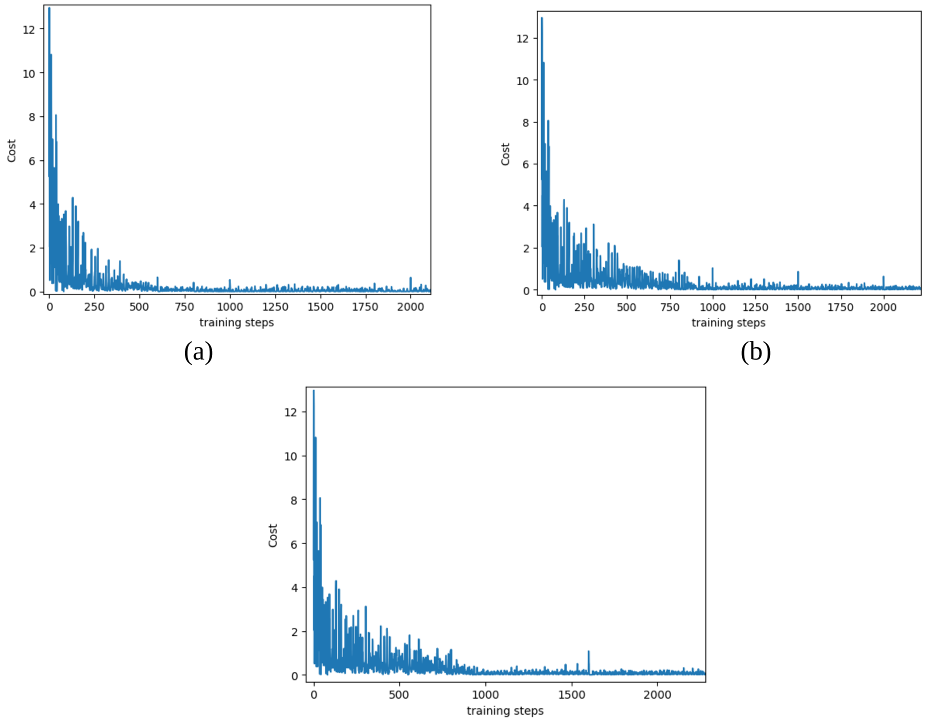
(a)

(b)

(c)

Fig. 5: (a) Learning rate=0.01, replace target=200 (b) Learning rate=0.01, replace target=500, (c) Learning rate=0.01, replace target=800

From a horizontal perspective, when the learning rate is 0.01, convergence can be gradually reached after about 1000 iterations. From a longitudinal point of view, the change of the replacement target has little effect on the simulation convergence. After about 1000 iterations, it takes about 0.2s for the user to match the spectrum at the fastest.



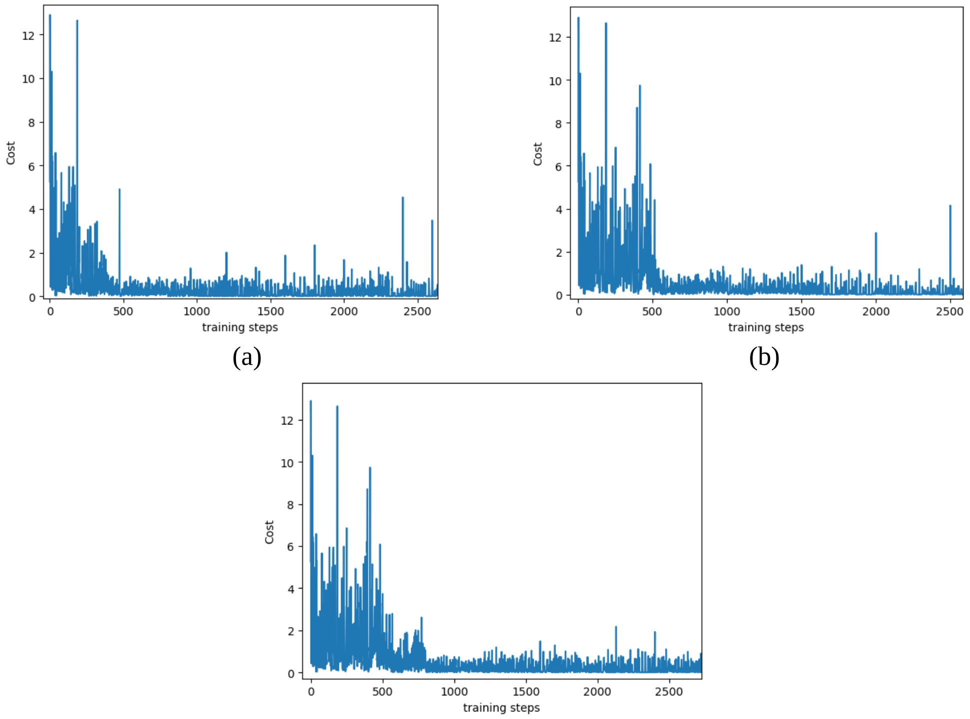
(a)

(b)

(c)

Fig. 6: (a) Learning rate=0.03, replace target=200; (b) Learning rate=0.03, replace target=500; (c) Learning rate=0.03, replace target=800

When the learning rate is 0.03, comparing Fig. 6(b) and Fig. 5(b), it can be seen that the cost convergence rate is faster than when the learning rate is 0.01. After about 800 iterations, it takes about 0.2s to match the user with the spectrum.



(a)

(b)

(c)

Fig. 7: (a) Learning rate=0.06, replace target=200; (b) Learning rate=0.06, replace target=500; (c) Learning rate=0.06, replace target=800

Similarly, it can be seen from Fig. 7 that when the learning rate is 0.06, although only about 500 iterations can converge, the cost convergence effect is not good because the learning rate is too large. In order to better understand the changes in the learning curve, the simulation results of the first 2500 rounds are shown here. Obviously, after 2500 iterative learning at this learning rate, it takes about 1s to match the user with the spectrum. According to analysis, this should be due to excessive oscillation caused by the high learning rate.

In summary, comparing Fig. 5 with Fig. 6, although Fig.5 has a fast convergence speed, it has the disadvantage of high learning cost due to the low learning rate. Compared with Fig.6 and Fig.7, Fig. 7 has a higher learning rate, so it is easy to overfit and cause slower convergence. Overall, the best learning rate is 0.03.

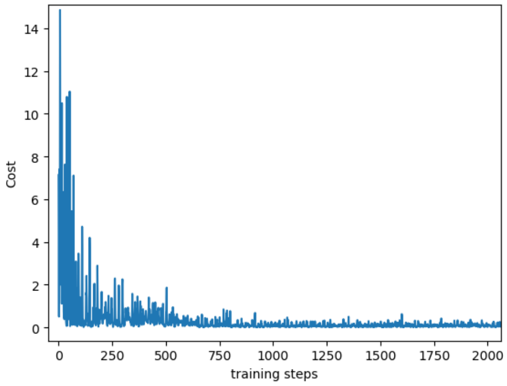


Fig. 8: Learning rate=0.03, replace target=800, the channel occupancy probability = 0.1

This paper also considers the influence of channel occupancy probability on the convergence speed of the algorithm. When the learning rate is 0.03, the replacement target is 800, and the channel occupancy probability is 0.1, that is, the channel has more time to be idle, and the cost is shown in Fig. 8.

Comparing Fig.8 with Fig.6 (c), it can be seen that the longer the channel idle time, the shorter time it will take for the agent to find the optimal channel. This is indeed the case.

Compared with other algorithms, DQN is suitable for the situation that the state parameters are very large and the prior information is very little. This is what other supervised learning algorithms can't do. Because Q-learning of unsupervised learning can't process so many parameters because it stores data in tables, DQN has great advantages in dealing with various complex communication models in 5G.

6. Conclusion

First, this article proposes a DQN algorithm suitable for complex communication scenarios: if the user’s communication scenario changes, the user’s current optimal channel will also change. Without any prior information about the channel, the user finds the current optimal channel through independent learning. This article also sets the idle rate and occupancy rate of the channel, which is in line with the actual communication scenario.

Secondly, this article proposes a 5G communication system model that combines V2X, D2D and mobile users, and considers the priority between the three to achieve dynamic spectrum matching of the three in different situations. The conclusion shows that the algorithm proposed in this paper has advantages in complex communication scenarios in 5G.

It is foreseeable that with the development of 5G, the model proposed in this article can be extended to more complex situations, which has great practical value. In fact, due to the limitations of people's imagination, what 5G will eventually turn into our communication world is still unknown. In the future, there will be more wireless communication challenges waiting for us to solve.

7. Acknowledgements

This paper is supported by project of Shandong province post graduate education innovation program (SDYY16032). Project of Shandong Province Higher Educational Science and Technology Program (Grant No. J18KA345).

8. References

[1] W. Wang, A. Kwasinski, D. Niyato, and Z. Han, "A survey on applications of model-free strategy learning in cognitive wireless networks," IEEE Communications Surveys Tutorials, vol. 18, no. 3, pp. 1717–1757, Mar. 2016.

[2] TANDRA R, SAHAI A. Fundamental limits on detection in low SNR under noise uncertainty [C]//2005 International Conference on Wireless Networks, Communications and Mobile Computing. IEEE, 2005(1): 464-469. DOI: 10.1109/WIRLES.2005.1549453.

[3] Mitola J, Maquire G J. Cognitive radios: making software radios more personal. IEEE Personal Communications, 1999, 6 (4)

[4] R. Mochaourab, B. Holfeld and T. Wirth. Distributed Channel Assignment in Cognitive Radio Networks: Stable Matching and Walrasian Equilibrium[J]. IEEE Transactions on Wireless Communications, 2015,14(7):3924-3936.

[5] A. V. KordaliandP. G. Cottis, "A reinforcement-learning based cognitive scheme for opportunistic spectrum access," Wireless Personal Communications, vol. 86, no. 2, Mar. 2016.

[6] Pitcha Rungsawang ; Amnach Khawne ,The Implementation of Spectrum Sensing and Spectrum Allocation on Cognitive Radio, ICACT2017 February 19 ~ 22, 2017, PyeongChang, Korea〕

[7] R. Karmakar, S. Chattopadhyay, and S. Chakraborty, "Dynamic link adaptation in ieee 802.11ac: A distributed learning based approach," in IEEE Conference on Local Computer Networks (LCN), Dubai, United Arab Emirates, Nov. 2016, pp. 87–94.

[8] Slimeni F, Scheers B, Chtourou Z, et al. Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm[C]. International Conference on Military Communications and Information Systems. IEEE, 2015:1-7.

[9] Han G, Xiao L, Poor H V. Two-dimensional anti jamming communication based on deep reinforcement learning[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017:2087-2091.

[10] 〔X. Wan, G. Sheng, Y. Li, L. Xiao, and X. Du, "Reinforcement learning based mobile offloading for cloud-based malware detection," in IEEE Global Communications Conference, Singapore, Dec. 2017.〕

[11] IMT-2020(5G) Promotion Group. 5G wireless technology architecture white paper [R]. 2015.

[12] International Telecommunications Union (ITU), Recommendation ITU-R M.2083., "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," Sept. 2015.