An Improved Dynamic Spectrum Access Algorithm Based on Reinforcement Learning

Chen Zhong[1], Chutong Ye[2], Chenyu Wu[2(✉)], and Ao Zhan[2]

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[2] School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
jerry916@zstu.edu.cn

Abstract. This paper proposes an improved dynamic spectrum access algorithm based on reinforcement Learning in cognitive radio networks. Q-learning algorithm is used as the core to update the optimal strategy for the established Markov decision process according to specific scenarios, and Q-table is updated iteratively to improve the learning rate. In order to verify the effectiveness of the proposed algorithm, we construct the mathematical model and the simulation environment. The simulation results validate the effectiveness of the proposed algorithm, which can effectively improve the system throughput under the condition that not affect primary users' communication. The proposed algorithm can quickly adjust the corresponding reward value and strategy in the iterative process of reinforcement learning training, so as to quickly converge to the optimal strategy, and the training results are consistent with the expected results.

Keywords: Dynamic spectrum access · Reinforcement learning · Q-Learning · Cognitive radio networks

1 Introduction

At a time when the number of mobile subscribers worldwide is growing, the available licensed spectrum is limited. In this regard, the new spectrum management scheme can not only meet the communication needs of the new generation of mobile communication technology with high bandwidth and low delay, but also solve the problem of scarce spectrum resources and the risks and even crises that spectrum management may face in the future [1]. Dynamic spectrum access in cognitive radio technology is the key entry point [2]. At the beginning of the

Supported by the Fundamental Research Funds of Zhejiang Sci-Tech University under grant 2021Q029.

X. Jiang (Ed.): MLICOM 2022, LNICST 481, pp. 13–25, 2023.
https://doi.org/10.1007/978-3-031-30237-4_2

21st century, the theory of dynamic spectrum access has been perfected after more than ten years of research, and the traditional dynamic spectrum access technology model has gradually taken shape [5]. At present, dynamic spectrum access technology based on reinforcement learning theory is the main direction to solve the current resource limitation problem, and it is also the inevitable trend of future spectrum allocation strategy, which is also the significance of this paper [3,4].

In [6,7], the authors combined Deep Q-Network (DQN) with dynamic spectrum access technology and verified that the system throughput and the network utilization were significantly improved. For the combination of reinforcement learning theory and traditional single-agent learning, domestic researchers put forward a new multi-agent distributed learning and centralized strategy theory [8–10]. Based on the DQN algorithm, the optimization of the algorithm and the improvement of the strategy by domestic researchers have further improved the communication effect of dynamic spectrum access [11–14].

This paper mainly studied the design and implementation of dynamic spectrum access algorithm based on reinforcement learning. By setting up training environment and adjusting training strategy and return, the q-table corresponding to channel state of single cognitive user in random spectrum environment close to the real environment was obtained, and the q-table obtained by training was taken as the core of the system. A dynamic spectrum access system algorithm based on Q-learning is constructed.

1.1 Dynamic Spectrum Access Technology

Dynamic spectrum access technology does not apply to primary users who are authorized to use a fixed frequency band by the spectrum resource management party, but to secondary users who are not authorized to use this frequency band, which is also called unauthorized users. After resource allocation, the users who can legally preferentially use this frequency band are called the authorized users of this frequency band, namely primary users (PUs). Unauthorized users who are not authorized are usually called cognitive users (CUs) or secondary users (SUs). cognitive users can carry out unauthorized communication when the authorized frequency band produces spectrum holes. Therefore, the biggest challenge faced by dynamic spectrum access technology is how to reduce the interference caused by unauthorized communication to the original authorized frequency band of primary users when the cognitive user accesses the unauthorized frequency band.

As shown in Fig. 1, after each cycle, the secondary system senses the environment and captures the corresponding frequency band information to learn about the current environment status and use these parameters to select and plan access policies. The basic principle of machine learning is that the agent obtains the current environmental information in the external environment, makes judgments and decisions according to the obtained information, and stores the environmental information and decisions in the knowledge base for updating, so as to achieve the effect of interacting with the environment, feeding back to the knowledge base and making decisions.

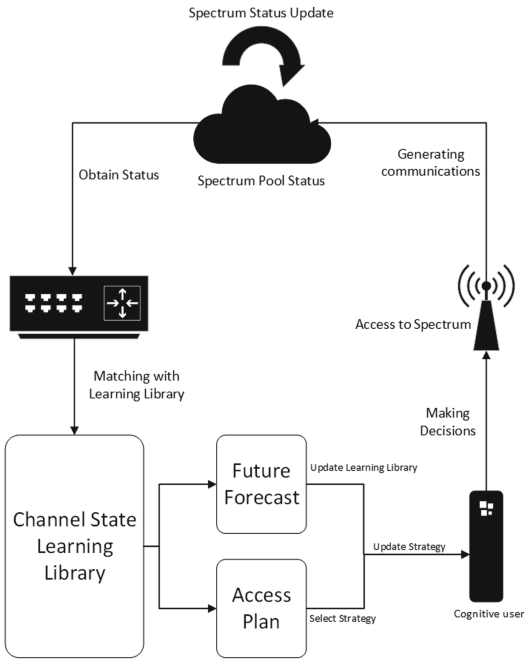


Fig. 1. Dynamic spectrum access System concept diagram.

1.2 Reinforcement Learning

As the most widely used and important branch of machine learning, reinforcement learning is inspired by psychology. Its main content is to focus on how CUs take different strategies and take different actions in different environments according to different environmental states, so as to obtain the highest possible returns.

Reinforcement learning was first born in psychology, and in principle, it is an algorithm evolved according to the adaptability of humans or other animals to the environment. In reinforcement learning theory, an interaction between an CU and the environment will generally affect the current environment, and this influence is the feedback obtained by the CU. How to judge whether the interaction is appropriate requires judgment based on the feedback of the environment, which is the common "reward". Generally speaking, the reward value representing the feedback is positive or negative, and the greater the absolute value of the reward, the greater the impact of the interaction. After multiple interactions with the environment, the rewards obtained from each interaction will be recorded in the learning module of the CU, which will be continuously updated, accumulated and optimized. Finally, after completing a certain degree of learning, the reward values of interactions under different environmental states will be formed in the learning module. In theory, the higher the reward value, the higher the probability of choosing an action in such an environmental state, so as to select the optimal and avoid mistakes [15].

As shown in Fig. 2, the model structure of reinforcement learning is divided into CU and environment in terms of objects, which are the largest components of reinforcement learning. The environment itself will constantly update its state with time or other conditions, and the CU will make the decision with the greatest theoretical return based on the experience in the learning library according to the observation of the environmental state, and interact with the environment. After the interaction is completed, the environment will give feedback to the interaction of the CU, and the CU will collect the feedback status and update the learning library again according to the feedback, so as to achieve a learning cycle.

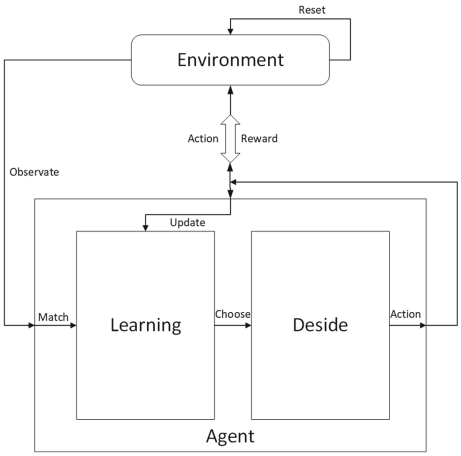


Fig. 2. Structure diagram of reinforcement learning model.

According to the above, reinforcement learning is essentially a process of multiple interactions to achieve the learning effect. In a single reinforcement learning process, there are five steps to complete an interaction:

- Step 1: According to the current environment, obtain the environmental state of the current CU, and extract various parameters required by this reinforcement learning from the obtained state.
- Step 2: The state parameters obtained in step 1 are combined into a set, which is consistent with the state set type set in the learning library. The state set is matched with the states in the learning library to select the strategy for this state, which contains the actions to be carried out by the CU in the current state.
- Step 3: After selecting the actions to interact with the environment according to the strategy, the CU starts to interact with the environment. For the environment, if an CU performs an action, it will have an impact on the environment, that is, the result and feedback under the action will be generated. At this point, the environment status, the environment feedback, will be updated.

– Step 4: After the environment status and feedback are updated, the ai obtains the environment status and feedback again. If there is an interaction with the environment, the feedback is extracted and recorded, and if there is no interaction with the environment, only the updated environment state is retrieved.
– Step 5: Update the feedback to the learning library in the form of rewards, that is, give a "positive" and "negative" feedback result to the interaction. The reward update after each interaction makes the environment-action strategy in the learning library constantly updated, so that the standard can be referred to when the environment state is obtained next time, so as to achieve the learning effect.

In addition, after introducing the basic principle of reinforcement learning, the classification of reinforcement learning is also the key content of reinforcement learning. For reinforcement learning, it can be divided into reinforcement learning with model and reinforcement learning without model from the perspective of environment model. The biggest difference between the two is whether an CU needs to interact with the environment to explore the environment state.

Generally speaking, model-free reinforcement learning is more widely used at present. Compared with model-free reinforcement learning, modeling reinforcement learning has the advantages of higher generalization ability and stronger sampling efficiency. Once the environment model is established, the CU can even break away from the actual model to some extent and train in the environment model, and then put into the actual scene after learning. Common Model reinforcement learning algorithms include World Model, Alpha Zero and so on. Model free of intensive study, although from the learning efficiency will be slightly lower than there are models of reinforcement learning, but it has the best performance steps, namely when the strategy and the environment interaction gradually achieve convergence condition, no model than reinforcement learning with a model of reinforcement learning as a result, the strategic choice, more outstanding, especially the application in the actual situation, Model-free reinforcement learning performs better in avoiding compounding problems. Under the classification of model-free reinforcement learning, it can also be divided into two categories, namely, reinforcement learning based on value optimization and reinforcement learning based on strategy optimization. For the difference of the two, reinforcement learning based on value optimization is usually applied to discrete environment, according to the value of every action is updated, learning, and determine the next action, and reinforcement learning based on strategy optimization is often applied to the continuous action space, and the state after discretization, action dimension higher environment, in this premise, to reduce the learning cost, Figuring out the probability of the next action to choose the best strategy is reinforcement learning based on strategy optimization. Generally speaking, the more common value-based reinforcement Learning includes Q-learning, SARSA, DQN, etc., while the strategy-based reinforcement Learning includes PPO and TRPO. The detailed classification of reinforcement learning is shown in Fig. 3.

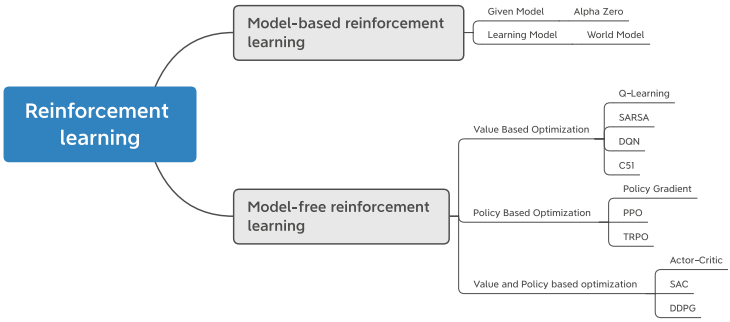


Fig. 3. Types of reinforcement learning algorithms.

2 Dynamic Spectrum Access Algorithm Based on Reinforcement Learning

2.1 Reinforcement Learning Algorithm

The main principle of reinforcement learning is to keep trying in the process of interaction with the environment, and then conduct the next interaction according to the reward, so as to form the learning of the environment. Model-free reinforcement learning can also be divided into value optimization based reinforcement learning and strategy optimization based reinforcement learning. The more common algorithms in model-free reinforcement Learning based on value optimization include Q-learning, DQN and SARSA(State-Action-Reward-State-Action). Among them, Q-learning is the most classical and earliest model, and it has good effectiveness and coverage while being less difficult to implement.

The purpose of Q-learning is to obtain more suitable solutions to the problems corresponding to different models established by Markov decision process under different environmental conditions. In markov decision making, the environment models are discretized in time, and the states are discretized into finite dimensional tuples. Generally speaking, Markov decision process model is expressed as M, and expressed as $M = (S, A, T, R,)$, where the components of M are also called "quintuples", which are:

- S: $S = \{S_1, S_2, S_3, \cdots, S_n\}$ represents the set of environment states.
- A: $A = \{A_1, A_2, A_3, \cdots, A_n\}$ represents the action set that CU can select.
- R: R represents the reward obtained by the CU after executing the action.
- T: $T : S \times A \rightarrow [0, 1]$ represents the state transition probability between the environment and the CU.
- π: π represents the result and is usually expressed as the optimal policy.

Compared with Markov decision process model, the most important thing of Q-learning is the addition of Q-table, that is, Q Table. Generally, in Q-learning

algorithm, the interaction results of state S and action A are expressed and stored in the form of matrix. The interaction result, i.e., reward R, is the performance and reward degree of each action under different finite discrete states. For each state, the return values of different actions are calculated and then become a new quantization standard, namely Q-value, which is stored in Q-Table. The Q-value calculation formula is as follows:

$$\pi Q'(S_t, a) = (1 - \alpha) * Q(S_t, a) + \alpha * [R + \gamma * Q_{\max}(S_{t+1},)] \quad (1)$$

In addition to the quintuple described above, there are several key parameters in Formula (1), which are:

- α: Learning rate of Q-learning, that is, the updating degree of Q-value after each Learning.
- γ: discount factor, which takes into account expected future return in addition to current return after performing an action.

In addition, parameters such as greed rate θ and limit index also play a key role. Under the influence of greed rate θ, the action selection of an CU will not only cover the most current situation, but also cover the overall action as far as possible in different environments, so as to find the optimal solution among all solutions and avoid falling into local optimum [16].

According to the characteristics of Q-learning, the selection of actions and strategies in Q-learning is based on values, namely Q-value, which is stored in a list named Q-table. A Q-table generally has multiple dimensions, and each dimension divides states in a discrete form. In this paper, Q-table has a total of four dimensions, including state three dimension and action one dimension. The state three dimensions are the internal influencing factors of frequency band: bandwidth and signal to noise ratio, and the external influencing factors: interference rate. One dimension of action is the corresponding effect of two actions "communication" and "non-communication" made by the CU, and the communication and non-communication are directly regarded as the action set of the CU.

For the primary user in the unauthorized frequency band, the interference rate should be as low as possible. As shown in Table 1, when the interference rate is higher than 40%, the reward obtained by CU is negative and positively correlated with the interference rate. If the CU chooses to stop communication knowing that the interference rate of the unauthorized frequency band is greater than 40%, the reward obtained will not become 0. When the interference rate is lower than 40%, the CU chooses to communicate with the unauthorized frequency band to obtain the reward and the interference rate is negatively correlated. If the interference rate is lower than 40%, the reward of the CU ending the communication with the unauthorized frequency band will be smaller than the reward of generating the communication.

Table 1. CU performs the main rules of action reward.

Interference rate of unauthorized frequency band (%)	Communication reward	Non-communication reward
0–40	Positive	Negative
40–100	Negative	Zero

Meanwhile, the q-value update rule is optimized. After optimization, the updated q-value of the action will be updated not only in the action in the environment state, but also in the action selected by the initial environment Value. That is, since the interference rate and noise ratio in the training environment are affected by the previous state, and in the system, access is only determined by the initial Value of the system state before access, q-value will be updated in the initial value and present value positions.

2.2 Dynamic Spectrum Access Algorithm

As shown in Fig. 4, the whole dynamic spectrum access structure is divided into two parts: dynamic spectrum access module and the corresponding reinforcement learning module. The reinforcement Learning module contains the learning environment required by reinforcement Learning, the core algorithm Q-learning and the corresponding Q-table. Dynamic spectrum access module includes DSA algorithm and DSA environment. The DSA environment includes authorized frequency band status and unauthorized frequency band status. The CU obtains the status of authorized frequency band and unauthorized frequency band, and according to the Learning results of Q-learning, accesses the unauthorized frequency band for communication on the premise of not affecting the primary user as much as possible.

In the system, we assume that the communication quality is low when the interference rate of authorized frequency band exceeds 70% and the SNR is less than 16 dB; the communication quality is extremely low and the communication is considered invalid when the interference rate exceeds 85% and the SNR is less than 10 dB. When the communication quality is low, if the status of the authorized frequency band is not ideal, and the status of the unauthorized frequency band does not meet the accessibility requirements, the CU will not forcibly access the unauthorized frequency band, and will continue to use the authorized frequency band for communication.

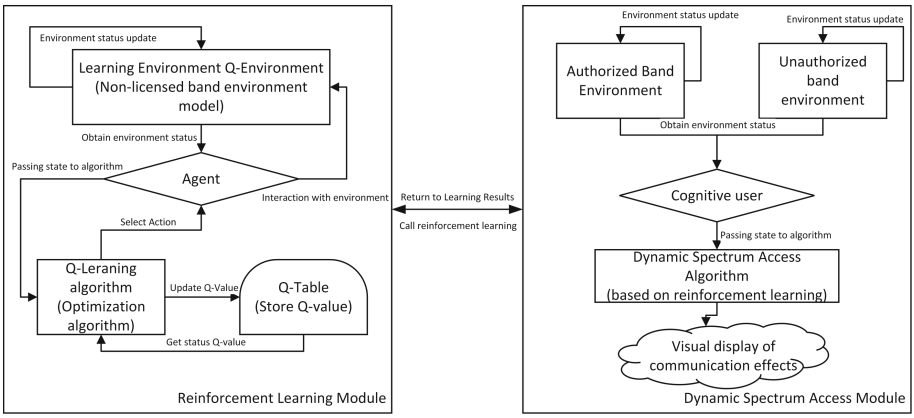


Fig. 4. Environment model parameter structure.

3 Simulation Analysis and Results

3.1 Simulation Environment Setting

To simplify the channel, the main elements included in the environment model will be set as interference rate, noise ratio and channel bandwidth, as well as the parameters used to reflect the communication quality: average interference rate, SNR and communication throughput.

The authorized band environment is based on the realistic domestic three major operators mobile, Unicom or telecom civil wireless communication operator bands, and the static spectrum allocation strategy is used, and the band environment parameters are shown in Table 2.

Table 2. Authorized band environmental parameters.

Authorized Band No	Band (MHz)	Authorized band name	Bandwidth
0	1745–1765/1840–1860	China Unicom FDD-LTE	42.5
1	2575–2635	China Mobile 4G TD-LTE	45.0
2	1765–1780/1860–1875	China Telecom FDD-LTE	40.0

The unlicensed band environment is based on a realistic domestic unlicensed wireless communication frequency bands selected from ten bands with different bandwidths, and the band environment parameters are shown in Table 3.

The initial interference rate is set as a Gaussian distributed random number with mean 0.5 and variance 0.17, and the mean interference rate at a later moment is a Gaussian distributed random with the same variance of 0.17 as the interference rate at the previous moment, and is re-randomized when it is greater than 1 or less than 0. The noise rate is set to a Gaussian-distributed

Table 3. Unauthorized band environmental parameters.

Unauthorized Band No.	Band (MHz)	Unauthorized band name	Bandwidth
0	885–890/930–935	Railroad Communication	10.0
1	1400–1427	Earth Satellite Exploration	27.0
2	1965–1980/2155–2170	Unallocated FDD	30.0
3	1626–1660	Maritime Satellite Communication	34.0
4	2655–2690	Unallocated TD-LTE	35.0
5	1660–1710	Meteorological satellite communication	50.0
6	2400–2483.5	Unallocated ISM	83.5
7	2025–2110	Fixed station communication I	85.0
8	2200–2300	Fixed station communication II	100.0
9	3400–3600	TDD BAND42	110.0

random number with a mean value of the median of the communication quality midpoint and a variance of 0.08 at each moment.

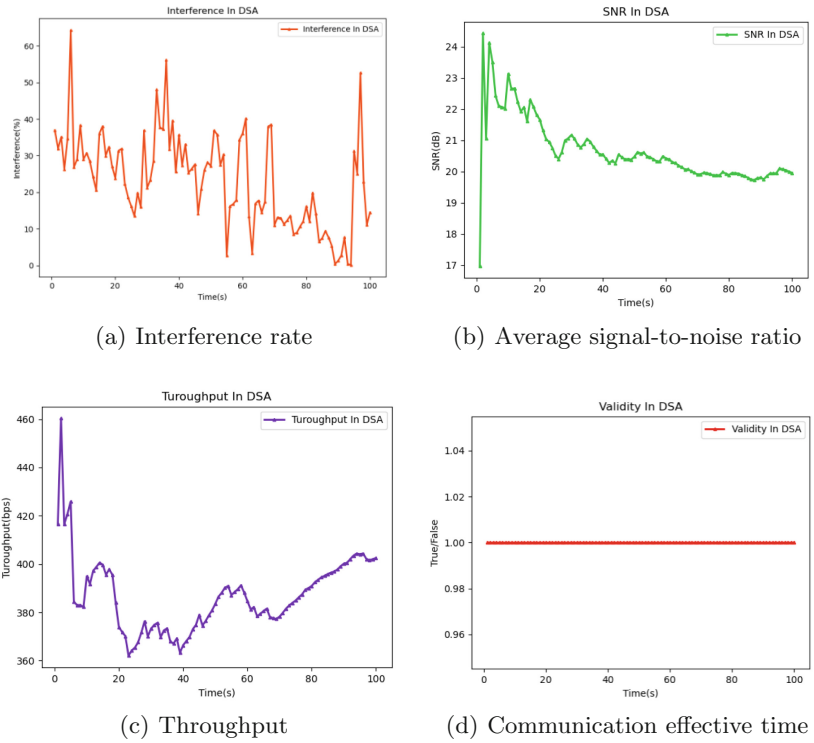
For non-authorized frequency band, set the initial interference rate for the mean value of 0.3, variance of 0.08 Gaussian distribution of random numbers, the latter moment interference rate mean value of the previous moment interference rate variance of the same 0.08 Gaussian distribution of random, re-random in greater than 1 or less than 0. The noise rate as above, but set its mean value at each moment for through wireless communication standard signal-to-noise ratio communication quality median, variance of 0.02 Gaussian-distributed random number.

3.2 Analysis of Simulation Results

In the simulation, DSA represents the proposed dynamic spectrum access system for communication based on reinforcement learning, while Tradition represents the communication via traditional static spectrum allocation.

As shown in Fig. 5(a), the communication interference rates are all lower than 70%, which is in line with the expected communication interference rate target. The large jump in the interference rate in the communication is because when the interference rate of the authorized band is high, if the interference rate is higher than 70% at the next moment, the band will be selected from the unauthorized band for communication. As shown in Fig. 5(b), the average signal-to-noise ratio of this system gradually converges to 19.26 dB, which indicates a relatively excellent communication quality according to the Chinese mobile communication signal-to-noise ratio standard. As shown in Fig. 5(c), the communication throughput of this system gradually stabilizes at 400 bps, and as shown in Fig. 5(d), there is no invalid communication time in the communication process, which ensures the continuity of communication.

Figure 6 gives the performance comparison between DSA and Tradition. As shown in Fig. 6(a), it can be seen that the average interference rate using tra-

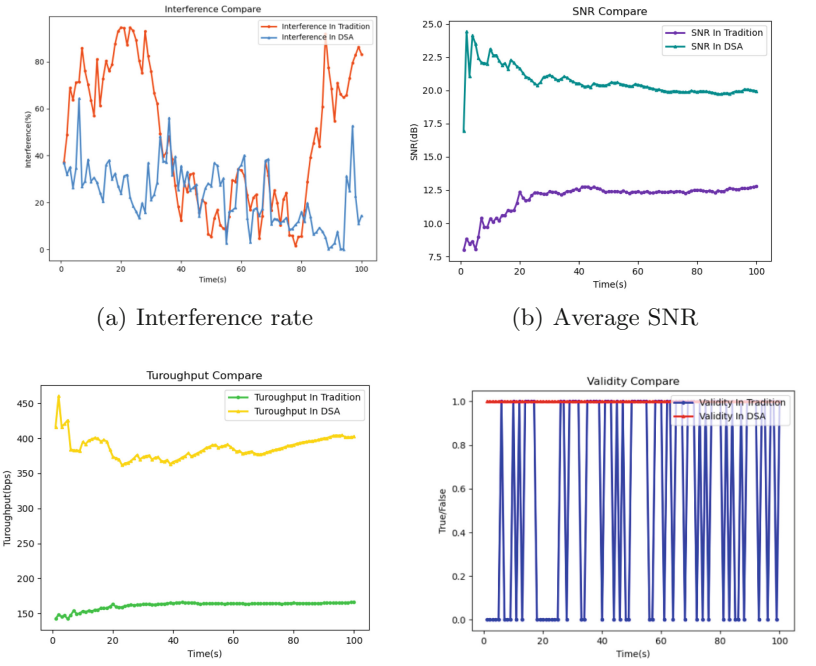


(a) Interference rate (b) Average signal-to-noise ratio

(c) Throughput (d) Communication effective time

Fig. 5. Performance of the proposed DSA algorithm based on reinforcement learning

ditional static spectrum allocation communication is 47.159%, while the average interference rate of dynamic spectrum access algorithm communication is 23.275%, which proves the effectiveness of the algorithm in improving the interference rate. Figure 6(b) shows the signal-to-noise ratio comparison graph. According to the domestic mobile communication standard, 12.8 dB signal-to-noise ratio is near the midpoint of communication quality, and 19.9 dB signal-to-noise ratio has been the midpoint of good communication quality. It can be seen from the graph that the signal-to-noise ratio of this system has been greatly improved compared with the traditional method, which proves that the algorithm can effectively improve the signal-to-noise ratio. As shown in Fig. 6(c), it can be seen that since the CU can access to the unauthorized band for communication when the quality of the authorized band is not satisfactory, it effectively improves the communication throughput of the algorithm, which is almost 2–3 times of the traditional method. As can be seen from Fig. 6(d), during the communication process, there is no invalid communication under this system, while the communication efficiency under the static spectrum allocation strategy is only 58%, so the proposed DSA algorithm can effectively improve the communication continuity.



(a) Interference rate (b) Average SNR

(c) Throughput (d) Communication effective time

Fig. 6. Performance comparison between DSA and Tradition

4 Conclusion

Aiming at the classical dynamic spectrum access technology algorithm based on Q-learning algorithm, this paper realizes and optimizes the original algorithm, and builds a simulation system environment model on the premise that the default hardware can obtain the real frequency state, and verifies the effectiveness of the system algorithm in the form of simulation. The overall system effect is about 90% similar to the expected effect, and the communication quality is tested, including interference rate, signal-to-noise ratio, etc., and the effectiveness of the algorithm is verified. Finally, the communication effect is visually compared with that of the traditional static spectrum allocation algorithm, which implies that the proposed DSA algorithm can quickly adjust the corresponding reward value and strategy in the iterative process of reinforcement learning training, and also achieve better performance.

References

1. Li, C., Jian, S.: Research on dynamic spectrum allocation in high mobile wireless communication network. Comput. Integr. Manuf. Syst. **37**(10): 124–127+146 (2020)
2. Song, T., Zhang, D., Chen, Z., Hang, X.: Spectrum allocation of cognitive wireless sensor network based on residual energy. Chin. J. Sens. Actuators **32**(12), 1875–1880 (2019)
3. Mekuria, F., Mfupe, L.: Spectrum sharing for unlicensed 5G networks. In: IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–5 (2019)
4. Oyewobi, S.S., Hancke, G.P., Abu-Mahfouz, A.M., et al.: An effective spectrum handoff based on reinforcement learning for target channel selection in the industrial Internet of Things. Sensors **19**(6), 1395–1416 (2019)
5. Hu, F., Chen, B., Zhu, K.: Full spectrum sharing in cognitive radio networks toward 5G: a survey. IEEE Access **6**, 15754–15776 (2018)
6. Nguyen, H.Q., Nguyen, B.T., Dong, T.Q., et al.: Deep Q-learning with multiband sensing for dynamic spectrum access. In: Proceedings of International Symposium on Dynamic Spectrum Access Networks, pp. 1–5. IEEE, Seoul (2018)
7. Han, Z., Lei, T., Lu, Z., et al.: Artificial intelligence-based handoff management for dense WLAN: a deep reinforcement learning approach. IEEE Access **7**, 31688–31701 (2019)
8. Zhang, L., Liang, Y., et al.: 6G visions: mobile ultra-broadband, super internet-of-things, and artificial intelligence. China Commun. **16**(8), 1–14 (2019)
9. Luong, N.C., Hoang, D.T., Gong, S., et al.: Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun. Surv. Tutorials **21**(4), 3133–3174 (2019)
10. Karmakar, R., Chattopadhyay, S., Chakraborty, S.: Dynamic link adaptation in IEEE 802.11AC: a distributed learning based approach. In: IEEE 41st Conference on Local Computer Networks (LCN), pp. 87–94. IEEE, Dubai (2016)
11. Xiao-hua, W.: Dynamic spectrum access technology for cognitive wireless communication. Inf. Technol. **44**(11), 137–141 (2020)
12. Zhou, X., Chen, Y., Zhang, Y., He, P.: Multi-channel access dynamic spectrum allocation under hybrid spectrum sharing mode. Commun. Technol. **54**(11), 2518–2526 (2021)
13. Le, T., Tao, L., Zhang, Yu., Pengzhi, Q.: Dynamic spectrum allocation method based on multi-agent reinforcement learning. J. Terahertz Sci. Electron. Inf. Technol. **19**(04), 573–580 (2021)
14. Zhang, Y., Zhou, Y.: Dynamic spectrum access algorithm based on Q-learning. Natural Science Journal of Hainan University (2018)
15. Liu, Q., et al.: A survey on deep reinforcement learning. Chin. J. Comput. **41**(01), 1–27 (2018)
16. Bin, M., Haibo, C., Chao, Z.: Network selection algorithm based on improved deep Q-learning. J. Electron. Inf. Technol. **44**(1), 346–353 (2022). https://doi.org/10.11999/JEIT200930