Reinforcement Learning-Based Interference Avoidance in Cognitive Radio Networks

Jinhua Chen, Luyong Zhang*, Ao Wang

School of Information and Communication Engineering, Beijing University of Posts and Telecommunication, Beijing, 100876, China

*Corresponding author’s e-mail: lyzhang@bupt.edu.cn

Abstract. Cognitive radio is an important technology developed by the next generation of wireless networks to effectively use limited spectrum resources and meet the rapidly growing demand for wireless applications and services. However, reliability is very important but not well-resolved problem in cognitive wireless networks. In this paper, we focus on the algorithm of interference avoidance in the presence of secondary user with traditional power control capabilities. Using the reinforcement learning algorithm based on proximal policy optimization, and adjusting its own transmit power, secondary user achieves its own quality of service requirements. Through experiments, it is proved that the algorithm used in this paper can not only deal with different secondary user power control strategies, but also show good stability in the changes of wireless environment noise and small amount data of the environment.

1. Introduction

Cognitive radio (CR) [1] [2] is a technology that effectively solves the problem of spectrum shortage, is increasingly used in wireless communications [3]. In cognitive wireless networks, it allows secondary user (SU) to opportunistically access spectrum "holes" that are not occupied by primary users (PU) through spectrum detection. However, although spectrum detection allows SU to detect "holes" in the spectrum, and uses unused spectrum to improve spectrum utilization, SU may face the problem of interference at the same time, other SU for example. Therefore, the existence of other SU may affect the communication service quality of the SU.

SUs can perform opportunistic access transmission by detecting spectrum unused channels by performing a spectrum sensing process. As the number of users increases, we need to adjust the transmit power between SUs to avoid mutual interference [4] among users. For example, two SUs performing communication transmission in the same channel need to adjust their own transmit power according to the transmit power of the other party to reduce the other party ‘s interference with their own communication process. In this article, we consider a situation where in the same area: There are two types of SU networks, one is a network of SUs with traditional power adjustment functions, and the other is an intelligent power adjustment node combined with reinforcement learning algorithms. The two types of networks need to adjust the transmit power to achieve their respective communication services quality.

Inspired by the results of reinforcement learning in dynamic programming problems, researchers have begun to seek solutions to wireless communication problems based on reinforcement learning. The authors of [5] and [6] studied Q-learning and state-action-reward-state-action (Sarsa) reinforcement learning in power control, respectively. The on-off state of the base station is controlled by a deep Q network (DQN), with the aim of improving the energy efficiency in [7]. The author of [8] introduced...
the allocation of computing resources and proposed an optimization strategy based on semi-MDP to schedule cloud computing resources to improve the practicability of the system. The author in [9] proposed a primary and secondary user power adjustment mode based on DQN. However, the proposal in [9] has an obvious disadvantage: when the noise in the environment is relatively large, especially when the data in the environment is little, the algorithm based on DQN will be greatly affected, and compared with the proximal policy optimization (PPO) used in this paper, the stability of the system model is not as good as PPO.

2. System model

2.1. Network model

This paper discusses the coexistence of two kinds of cognitive wireless networks. In this system, based on the overlay cognitive mode, the network composed of traditional SU and the intelligent SU network are deployed in the same geographical area, as shown in figure 1:

![Figure 1 Network model](image)

As shown in figure 1, the SU network can obtain parameters of the environment by receiving signal that are periodically reported by detectors randomly deployed in the environment. In our system model, the nodes in the SU network access the channel through Orthogonal Frequency Division Multiplexing Access (OFDMA), and we suppose that only one pair of communication pairs can access one channel in the same time slot, but it may be affected by traditional SU, so we focus on discussing the intelligent SU achieves its own data transmission scenario by accessing the communication channel being used by the traditional SU node.

2.2. Traditional SU working mode

The traditional SU node will adaptively adjust its own transmit power through wireless communication conditions to meet its SINR requirements.

The following describes two ways in which traditional SU interferes with network power control. Power control methods can be divided into fixed step size control and adaptive step size control [10]. According to our research scenarios, we consider that the power control schemes of traditional secondary networks are divided into the above two types, which can be specifically described as adaptive step size power control (ASPC) and fixed step size power control (FSPC). ASPC can be described by the following equation:

$$p_{k}(k + 1) = D \left( \frac{\eta_{k} p_{k}(k)}{\text{sinr}_{k}(k)} \right)$$

Where $k$ represents the $k$th time slot, $p_{k}$ represents the transmit power in the $k$th time slot, $\text{sinr}_{k}(k)$ represents the signal-to-interference and noise ratio of the node in the network in the $k$th time slot, and
\(\eta_1\) represents the normal communication between the nodes Minimum signal-to-interference and noise ratio. \(D\) is a mapping function that can map \(\left(\frac{\eta_1 p_i}{\sin r_i(k)}\right)\) to a series of discrete transmit powers:

\[
p_i = \{p_1^1, p_1^2, p_1^3, ..., p_1^l\}
\]

(2)

FSPC \[9\] can also be described by equation (1), Where \(p^i_j < p^i_j, i < j\). Assuming that the transmit power \(p^i_j(k) = p^i_j\) in the kth slot, its power in the next slot is defined by the following equation:

\[
p^i_j(k + 1) = \begin{cases} 
p^i_j + 1 & \text{if } p^i_j \leq \mu \leq p^i_j + 1 \text{ and } i + 1 \leq l \\
p^i_j - 1 & \text{if } p^i_j - 1 \leq \mu \leq p^i_j \text{ and } i - 1 \leq l \\
p^i_j & \text{else}
\end{cases}
\]

(3)

Where \(\mu = \left(\frac{\eta_1 p_i(k)}{\sin r_i(k)}\right)\).

3. Problem formulation

3.1. Communication service quality requirements

The determination of the communication quality \[11\] on each communication link is determined by its own SINR. The SINR of each receiver can be expressed as:

\[
\text{SINR} = \frac{P_{1,1} |h_{1,1}|^2}{\sum P_{2,2} |h_{2,2}|^2 + \sigma^2}
\]

(4)

Among them, 1,2 respectively indicate cognitive nodes and interfering nodes that communicate with them at the same time. Therefore, certain conditions must be met to achieve service quality:

\[
\text{SINR} \geq \eta
\]

(5)

Where \(\eta\) represents the minimum SINR requirement by the receivers of different networks to receive signals normally.

3.2. Detector received signal strength

In the environment, sensors are deployed to measure the received signal strength. In a unit time slot, the signal strength received by the nth sensor can be expressed as follows \[9\]:

\[
p^i_n = p_1 g_{1,n} + p_2 g_{2,n}
\]

(6)

Among them, \(p_1, p_2\) represent the transmitting power of the interference node and the cognitive node, and \(g_{1,n}, g_{2,n}\) represents the channel gain from the transmitting node to the nth sensor. For the free space wireless propagation model \[12\], \(g_{1,n}, g_{2,n}\) can be:

\[
g_{1,n} = \left(\frac{\lambda}{4\pi d_{1,n}}\right)^2, \ g_{2,n} = \left(\frac{\lambda}{4\pi d_{2,n}}\right)^2
\]

(7)

Where \(\lambda\) is the signal wavelength, and \(d_{1,n}, d_{2,n}\) respectively represent the distance from the transmitting end to the nth sensor. Considering other factors affecting communication in the wireless communication environment, such as shadow fading, estimation errors, etc., these random factors are modelled as zero-mean Gaussian random variables, and the variance \(\sigma^2\) is designed as\[9\]:
\[ \sigma^2 = \frac{P_{\tau} g_{1,n} + P_{\tau} g_{2,n}}{n_{\text{noise}}} \]  

(8)

3.3. Reinforcement learning algorithm PPO

PPO [13] trains \( \theta \) by \( \pi_\theta \) sampling, and \( \theta' \) is fixed so the sampling data is reused. PPO proposed a new objective function that can be updated in small batches in multiple training steps, which solves the problem of difficult step determination in the Policy Gradient algorithm.

According to the techniques of the important sample, the gradient estimation can be transformed into:

\[ \nabla_{\theta} J(\theta) = E_{\pi_\theta} \left[ \frac{\pi_\theta(s, a)}{\pi'_\theta(s, a)} R(s, a) \nabla_{\theta} \log \pi_\theta(s, a) \right] \]

(9)

\[ \nabla_{\theta} J(\theta) = E_{\pi_\theta} \left[ R(s, a) \nabla_{\theta} \log \pi_\theta(s, a) \right] \]

(10)

The PPO algorithm used in this paper uses the clip function [13] to fix it within a certain range, which can play a restrictive role:

\[ \nabla_{\theta} J(\theta) = E_{\pi_\theta} \left[ \min \left\{ \frac{\pi_\theta(s, a)}{\pi'_\theta(s, a)} R(s, a) \nabla_{\theta} \log \pi_\theta(s, a), 1 - \varepsilon, 1 + \varepsilon \right\} R(s, a) \nabla_{\theta} \log \pi_\theta(s, a) \right] \]

(11)

Figure 2 Reinforcement learning framework

Therefore, combined with the reinforcement learning framework, as shown in figure 2, where the environmental state \( s_t \in S \), \( S \) is the set of signals received by the detector at regular intervals, each \( s_t \) represents two types of networks in a time slot State parameters of the network after power adjustment.

And \( a_t \in A \), \( A \) is the set of power level. In this paper, this set is discrete. The reward function \( r_t \) is set that the nodes of the two types of networks can meet their SINR requirements, as follows:

\[ r_t = \begin{cases} 
0.01 & \text{if } SINR_{\text{traditional SU}} \geq \eta_{\text{traditional SU}} \text{ and } SINR_{SU} \geq \eta_{SU} \\
0 & \text{else}
\end{cases} \]

(12)

4. Simulation parameter setting and result analysis

This paper uses a single-threaded PPO algorithm. In the neural network module, the PPO algorithm contains of three layers of neural networks, and the number of neurons in each layer is 256. The learning rate of the network is 0.00001 to stimulate neural network training. The discount factor is 0.8, the batchsize is 512, and the \( \epsilon \) of the clip function is set to 0.1.
The power selection range for both types of networks is \( \{0.05, 0.1, ..., 0.4\} \) (in Watt). with the SINR requirements of the two types of networks are set to \( \eta_1 = 1.2 \), \( \eta_2 = 0.7 \) by default. For simplicity, the channel gain from the sender to the receiver is set to 1. The noise power at the receiver is set to 0.01W. Three detectors express as \( n_{\text{sensor}} = 3 \) are randomly deployed in the network environment in the range of \([100, 300]\) (in meters). As for random factors, the calculation parameter of detector error noise variance is \( p_r = 0.05 \) (in Watt). According to different simulation requirements, different \( n_{\text{noise}} \in \{3, 6\} \) are set to set different ambient random noise levels.

After training, each round of iteration starts 1,000 independent test experiments; each test randomly initializes the node states of the two networks, and the maximum time slot is limited to 10 times, that is, if the system does not reach stability within 10 time slots The status determines that the current power adjustment has failed. The difference between [9] is that the number of transition steps is the number of adjustments required before reaching success.

For DQN and ACER, the number of neurons in each layer is 256, 256, and 512, respectively. The remaining parameter settings are the same as the PPO algorithm used in this paper.

As shown in figure 4, the comparison of the success rate of the same type of algorithm PPO and ACER when the traditional secondary network uses ASPC can only adjust the power of the SU network to achieve the stability of the two types of network transmission service quality and the required adjustment steps. Number of performance. We can see that under the simple power control strategy, the system can achieve stability using different algorithms, and the PPO converges faster with ACER and the success rate is higher than ACER during the adjustment process. At the same time, we It can be inferred that the average throughput of the entire cognitive network system is relatively high.

![Graph](image)

(a) Success Rate  
(b) Transfer steps

Figure 3 Performance in different algorithms when traditional SU uses ASPC with different random factor

As shown in figure 5, the DQN or ACER in the same environment did not converge after the same training 200,000 times, and showed relatively large fluctuations, which is quite unfavorable for wireless communication scenarios; and, from reaching the SINR of the two types of networks, the success rate of the PPO algorithm reaches 100% in a stable trend. We can also see that when traditional SU apply FSPC, the intelligent SU need to take more training times to converge. Meanwhile, the number of transfer steps required is basically lower than the number of transfer steps required by the DQN or ACER algorithm. Using the PPO algorithm, compared with other reinforcement learning algorithms, nodes of intelligent cognitive networks can adjust their own transmit power with strong robustness to ensure the quality of communication services even in difficult conditions.
5. Conclusion

This paper proposes a situation in which a network composed of traditional nodes with power control capabilities and an intelligent SU network coexist. It provides opportunistic access to unused spectrum not yet used by PU. In this system, both types of networks need to achieve their own quality of communication service, which is measured by SINR in this paper. According to the change of the transmit power of the traditional SU, the nodes of the intelligent SU network adjusted their own transmit power using the PPO algorithm based on reinforcement learning to achieve the equilibrium state of the two types of networks. Moreover, the PPO algorithm we used in the model can achieve good results, when traditional SU using different power control strategies and the obtaining network environment parameter information is relatively small as well as containing larger random error. Compared to the other two algorithms: DQN and ACER, PPO shows good stability, which is in line with the requirements of wireless communication conditions, and intelligent SU network adjusts its own power to meet the quality of communication service. It also maintains normal communication on the secondary network and maintains the level of network utility.

References

[1] Mitola J I, Maguire G Q. (1999) Cognitive radio: making software radios more personal[J]. IEEE Personal Communications, 6(4):0-18.

[2] Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, et al. (2006) Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey[J]. Computer Networks, 50(13):2127-2159.

[3] Zhang L Y, Yang Y J, Chen J H. (2018) Cognitive Communication in Link-Layer Evaluation Based Cellular-Vehicular Networks[J]. IEEE Access, 6: 75293-75301.

[4] Zhang L Y, Yao H P, Zhou Z, et al. (2010) A novel ultra-wide band signal generation scheme based on carrier interference and dynamics suppression[J]. EURASIP Journal on Wireless Communications and Networking, 2010: 1-6.

[5] Xiao L, Li Y, Dai C, Dai H, and Poor H. V., (2018) “Reinforcement learning based NOMA power allocation in the presence of smart jamming,” IEEE Transactions on Vehicular Technology, vol. 67, no. 4, pp. 3377–3389.

[6] Ortiz A., Al-Shatri H., Li X., T. Weber, and A. Klein, (2017) “Reinforcement learning for energy harvesting decode-and-forward two-hop communications,” IEEE Transactions on Green Communications and Networking, vol. 1, no. 3, pp. 309–319.

[7] Li H, Gao H, Lv T, and Lu Y. (2018) “Deep q-learning based dynamic resource allocation for self-powered ultra-dense networks,” in 2018 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6, IEEE.

[8] H. Liu, S. Liu, and K. Zheng. (2018) “A reinforcement learning-based resource allocation scheme for cloud robotics,” IEEE Access, vol. 6, pp. 17215–17222.

[9] Li X J, Fang J, Cheng W, et al. (2018) Intelligent Power Control for Spectrum Sharing in Cognitive Radios: A Deep Reinforcement Learning Approach[J]. IEEE Access, 6:25463-25473.

[10] Li F M, Xu W J, Liu X H, et al. (2008) Power Control Technology of Wireless Sensor Networks
[11] Zhang L Y, Yang Y J, Li X, et al. (2019) Effective capacity in cognitive radio networks with relay and primary user emulator[J]. China Communications, 16(11): 130-145.

[12] T. S. Rappaport. (2002) Principles and Practice. Wireless Communications. Englewood Cliffs, NJ, USA: Prentice-Hall.

[13] Schulman, John, Wolski, Filip, Dhariwal, Prafulla, et al. (2017) Proximal Policy Optimization Algorithms[J].