Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 2731485, 8 pages
https://doi.org/10.1155/2019/2731485

Research Article

A Cognitive Relay Network Throughput Optimization Algorithm Based on Deep Reinforcement Learning

Shaojiang Liu,[1] Kejing Hu,[2] Weichuan Ni,[1] Zhiming Xu,[3] Feng Wang,[3] and Zhiping Wan[3]

[1]*Department of Equipment and Laboratory Management, Xinhua College of Sun Yat-Sen University, Guangzhou, China*
[2]*Academic Affairs Office, Guangzhou Pearl-River Vocational College of Technology, Guangzhou, China*
[3]*Department of Information Science, Xinhua College of Sun Yat-Sen University, Guangzhou, China*

Correspondence should be addressed to Zhiping Wan; wzp888@xhsysu.edu.cn

Received 10 January 2019; Revised 14 April 2019; Accepted 2 July 2019; Published 15 July 2019

Academic Editor: Mauro Femminella

In cognitive relay networks, the cognitive user opportunistically accesses the authorized spectrum segment of the primary user and simultaneously serves as the data relay node of the primary user while sharing the spectrum resource of the primary user. This not only improves the utilization efficiency of the network spectrum resources but also improves the throughput of the primary users. However, if the primary user randomly selects the relay node, there is no guarantee for an optimal throughput. Moreover, the system power consumption may increase. In order to improve the throughput of cognitive relay network and optimize system utility, this paper proposes a cognitive relay network throughput optimization algorithm based on deep reinforcement learning. For the system model of cognitive relay networks, the Markov decision process is used to describe the channel transition probability of the system model in the paper. The algorithm proposes a cooperative wireless network cooperative relay strategy, analyzes the system outage probability under different transmission modes, and optimizes the system throughput by minimizing the outage probability. Then, the maximum utility optimization strategy based on deep reinforcement learning is proposed to maximize the system utility revenue by selecting the optimal behavior. The experimental results show that the proposed algorithm has a good effect in improving system throughput and optimizing system energy efficiency.

1. Introduction

Cognitive radio technology allows unauthorized cognitive users to access the spectrum resources of authorized primary users, which is an important technical means to solve the scarcity of spectrum resources. Cognitive radio technology allows cognitive users to share licensed spectrum, effectively increasing the utilization of wireless spectrum resources. In the spectrum-sharing cognitive network, the cognitive user accesses the idle spectrum without affecting the normal communication of the primary user and assists the primary user in data transmission, which can improve the throughput of the primary user [1–3].

Reinforcement learning (RL) is an important machine learning method that has many applications in areas such as intelligent control of robots and analysis and prediction. The reinforcement learning algorithm can learn the mapping from environmental state to behavior, which makes the behavior selected by the agent can obtain the biggest reward of the environment [4, 5]. However, reinforcement learning often faces dimensional disasters when the state of the system is large. In order to solve this problem, people combine the perception of deep learning with the decision-making ability of reinforcement learning to propose a deep reinforcement learning algorithm [6–8]. In this paper, we take advantage of the deep reinforcement learning algorithm that solves the input problems of high-dimensional data and learn the optimal strategy, to maximize the system utility of cognitive wireless networks.

In the research of throughput optimization of cognitive wireless networks, many researchers have proposed different research methods. Literature [9] proposed a cognitive radio maximization throughput algorithm based on wireless spectrum sensing. By reducing the perceived time of the

secondary user, the transmission time was increased and the achievable throughput is maximized. In addition, the spectrum utilization rate was improved and the throughput was improved by reducing the false positive probability of spectrum sensing. Literature [10] proposed a throughput and outage probability algorithm for cognitive DF relay networks based on wireless energy harvesting. The algorithm uses wireless energy harvesting strategy based on cognitive radio constraints to evaluate the throughput performance and outage probability of secondary users (SU) in the relay network and studies the impact of different system parameters on cognitive networks, such as power allocation ratio, main transmitter power, and allowable interference threshold of the primary user (PU) receiver to SU Impact of throughput and interrupt performance. Literature [11] proposed a cognitive radio network based on probability relay. A random service cooperation strategy with probabilistic relay was adopted, so that the secondary user served the primary queue or the primary user with a specific service probability to relay the packet queue, controlled the transmission delay according to the QoS constraints of the application and the system, and improved the system throughput. Literature [12] proposed a cognitive relay network throughput optimization algorithm using an improved time exchange protocol. The paper discussed a relay-based dual-hop relay protocol for the underlying cognitive radio. In this scheme, energy was transmitted in small packets, where the number of packets to be used is estimated and fed back by the relay node, and throughput and average energy optimization effects can be achieved. Literature [13] proposed a novel interference alignment scheme based on antenna selection in Cognitive Radio Networks. In this scheme, an efficient IA-AS algorithm based on discrete stochastic optimization was proposed to improve the computational performance, and a scheme called the channel state information (CSI) filtering was proposed to weaken the influence of the imperfect CSI. The experimental results show that the scheme improves the average rate of SUs and rate of PU. Literature [14] proposed a secure primary system transmission scheme coordinated by secondary networks; the paper used two schemes to improve the total data rate of the secondary network while guaranteeing the secrecy rate of PU, and the principle of interference alignment was employed to eliminate interference from PU and other SUs at each secondary receiver. From the experimental results, the scheme could effectively improve the sum rate of SUs.

The contributions of this paper are as follows: (1) propose a cooperative relay transmission mode of cognitive wireless networks, and the system interruption probability analysis is carried out for different transmission methods; (2) use the deep reinforcement learning algorithm to learn and explore the state transition information of the system when the state transition probability is unknown; (3) combining system throughput and power optimization problems, a system utility function is proposed to maximize the benefits by selecting the optimal behavior by deep reinforcement learning algorithm; (4) in the experimental scheme, it is proved that the proposed algorithm has better performance in improving system throughput and system utility than cognitive wireless network algorithm based on reinforcement learning or energy overhead minimization.

The paper is structured as follows: Section 2 introduces the system model of cognitive wireless networks studied in this paper. Section 3 discusses the cooperative relay mode of the system and analyzes the outage probability for the data transmission mode. Section 4 discusses the maximum utility optimization strategy of cognitive wireless networks based on deep reinforcement learning. Section 5 introduces the experimental scheme and the experimental results and comparative analysis of each algorithm.

2. System Model

In the cognitive wireless network model of Figure 1, a primary user transmitter (PT), a primary user receiver (PR), a secondary user receiver (SR), and a cognitive relay user (CR) are included. It is assumed that there are a total of M frequency domain channels in the cognitive network, all channels are independent Rayleigh fading channels, and the channel remains unchanged during the transmission of one data frame. Each PT has a dedicated licensed band for transmitting signals to PR, and each cognitive user has an antenna that transmits and receives signals. During the operation of the network, PT will select the relay node to forward the data packet to PR or directly send the data to PR. In the process of cooperative PT relaying data, CR can also transmit data that needs to be transmitted to the secondary user receiver. There is a certain degree of mutual interference between PT and CR during data transmission. Suppose all noise n_o satisfies Gaussian white noise, the mean is 0, the variance is N_0, and the channel gain is $h_{a,b} = \lambda |d_{a,b}|^{-\alpha}$, λ indicating the gain factor and α representing the path loss exponent.

In the system model, the channel state of each frame does not change, and the channel state changes occur between adjacent frames. In this paper, the finite state Markov chain method is used to establish the channel state transition probability. Since the channels are independent Rayleigh fading channels, the signal-to-noise ratio obeys the Rayleigh distribution. It is assumed in this paper that S_i is used to represent the signal-to-noise ratio of channel i. The probability density function is

$$p(S_i) = \frac{1}{\overline{S}} \exp\left(-\frac{S_i}{\overline{S}}\right) \tag{1}$$

$\overline{S}$ indicates the SNR average. Then the transition probability of the channel is

$$p(S_i, S_{i+1}) = \frac{f\sqrt{2\pi S_i/\overline{S}} \exp\left(-S_i/\overline{S}\right) T}{\int_{S_i}^{S_{i+1}} p(S_i)\, dS_i} \tag{2}$$

f indicates the maximum Doppler shift and T represents the time frame.

3. Cognitive Wireless Network Cooperative Relay Algorithm

3.1. Collaborative Relay Mode. The cognitive wireless network model mentioned in this paper mainly uses the amplifying

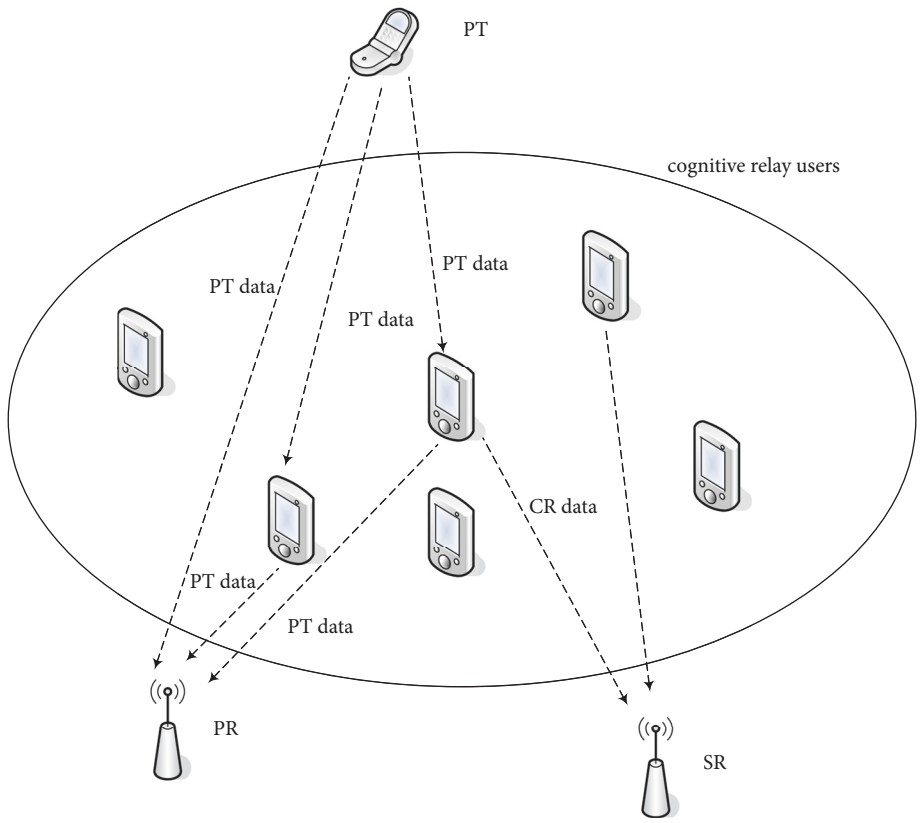

Figure 1: System model diagram.

and forwarding protocol to make the PT find the appropriate CR node for cooperative transmission [15, 16]. The transmission process mainly includes three stages: the first stage PT sends data to the selected CR node and PR node. The second stage is such that the CR node processes the PT data and sends it to the PR. The third stage is such that the CR node sends its own data to the SR.

Supposing that, in the time slot t, PT sends data $x_{PU}(t)$ to CR_i, the received data by CR_i is

$$x_{PC_i}(t) = h_{PT,C_i}\sqrt{P_T}x_{PU}(t) + n_0 \tag{3}$$

where h_{PT,C_i} represents the channel gain from PT to CR_i. After the CR_i receives the data, it amplifies the received data. If the amplification factor is γ_i, the amplified data is $\gamma_i x_{PC_i}(t)$. PT sends data $x_{PU}(t)$ to PR. For PR, the received data is

$$x_{TR}(t) = h_{PT,PR}\sqrt{P_T}x_{PU}(t) + n_0 \tag{4}$$

In the time slot $t+1$, the PR receives all the data relayed from the CR node and combines all the data, including the data from the relay and the data directly transmitted by the PT. Let the set of CR nodes as the relay node be Q, and get

$$x_{PR}(t+1) = \varepsilon\left(\sum_{i\in Q} h_{C_i,PR}\sqrt{P_{CR_i}}\gamma_i x_{PC_i}(t) + n_0\right) + (1-\varepsilon)\,x_{TR}(t) \tag{5}$$

where $h_{C_i,PR}$ represents the channel gain from CR_i to PR, P_{CR_i} represents the transmit power of CR_i, and ε is the weighting factor of the relay data.

CR_i sends its own data to the SR. For SR, the received data is

$$x_{SR}(t+1) = h_{CR_i,SR}\sqrt{P_{CR_i}}x_{CR_i}(t) + n_0 \tag{6}$$

$h_{CR_i,SR}$ represents the channel gain from CR_i to SR, and $x_{CR_i}(t)$ is the data transmitted by CR_i.

3.2. Outage Probability Analysis. According to the amplified forwarding protocol used, the direct data transmission rate between the primary transmitter PT and the primary receiver PR is

$$v_{PT,PR} = \log_2\left(1 + \frac{P_T}{P_o n_0}\right) \tag{7}$$

P_o represents the unit noise power, using v_{PT} to indicate the data transmission rate of the PT. The probability of $v_{PT} > v_{PT,PR}$ is the probability of interruption of data transmission between PT and PR, namely,

$$\rho_{PT,PR} = \Pr\{v_{PT} > v_{PT,PR}\} = 1 - e^{-(2^{v_{PT}}-1)p_o n_0/p_T} \tag{8}$$

The data transmission rate at which the PT transmits data to the PR through the relay node CR_i is

$$v_{PT,CR_i,SR} = \log_2\left(1 + \frac{P_T}{\omega P_o n_0}\right) \cdot \log_2\left(\frac{P_{CR_i}}{(1-\omega)\sqrt{P_o P'_{T,CR_i} n_0 + 1}}\right) \tag{9}$$

P'_{T,CR_i} represents the interference noise power of PT to CR_i, which is related to the transmission power of PT. $P'_{T,CR_i} = \delta P_T$, and δ are interference factors.

Therefore, the probability of the PT transmitting data to the PR through the relay node CR_i is

$$\begin{aligned}\rho_{PT,CR_i,SR} &= \Pr\left\{v_{PT} > v_{PT,CR_i,SR}\right\} \\ &= 1 - e^{-(2^{v_{PT,CR_i,SR}} p_o n_0/P_T)\sqrt{\omega P_{CR_i}/2P'_{T,CR_i}}}\end{aligned} \tag{10}$$

The data transfer rate at which the CR_i transmits data to the SR is

$$v_{CR_i,SR} = \log_2\left(\frac{P_{CR_i}}{\sqrt{P_o P'_{T,CR_i} n_0 + 1}}\right) \tag{11}$$

Therefore, the probability of interruption of the relay node CR_i transmitting data to the SR is

$$\begin{aligned}\rho_{CR_i,SR} &= \Pr\left\{v_{CR_i} > v_{CR_i,SR}\right\} \\ &= 1 - e^{-(2^{v_{CR_i,SR}}-1)\sqrt{P_o P'_{T,CR_i}} n_0/(P_{CR_i}+1)}\end{aligned} \tag{12}$$

4. Maximum Utility Optimization Based on Deep Reinforcement Learning

Deep reinforcement learning is an algorithm proposed on the basis of reinforcement learning algorithm combined with deep convolutional neural network. Reinforcement learning interacts with the environment through trial-and-error mechanism, learns the optimal strategy by maximizing the accumulated reward, and solves the input problem of high-dimensional data by combining convolutional neural network.

4.1. Analysis of MDP Problems. Assume that, in the current network, the state object of the system only considers the channel state of the node. In the current state, the system switches to the next state by selecting the behavior of the relay node by the PT, so the selection scheduling problem of the relay node can be modeled as a Markov decision process (MDP).

Suppose there are N cognitive nodes in the network. The channel state of the PT node changes when a PT node selects one of the cognitive nodes as a relay node. The Markov decision process consists of a quadruple $M = (K, A, \rho, R)$; the elements of the quaternion represent the state set, action set, state transition probability, and return function, respectively. In the model of this paper, let the action set of the PT in the time slot t be $A = \{a_{t,1}, a_{t,2}, \cdots, a_{t,N}\}$, $a_{t,i}$ $(i \in N)$ denoting that the PT node selects CR_i as the relay node in the t-slot. The set of channel states of the PT node is defined as $K = \{k_1, \cdots k_N\}$, and the state transition probability that the system state k_i shifts to k_{i+1} after taking the action $a_{t,i}$ can be obtained:

$$\begin{aligned}p(k_i, k_{i+1}) &= \frac{f\sqrt{2\pi a_{t,i}/\bar{S}}\exp\left(-a_{t,i}/\bar{S}\right)T}{\int_{a_{t,i}}^{a_{t,i+1}} p(a_{t,i})\,da_{t,i}} \\ &= \frac{f\sqrt{2\pi S_i/\bar{S}}\exp\left(-S_i/\bar{S}\right)T}{\int_{S_i}^{S_{i+1}} p(S_i)\,dS_i}\end{aligned} \tag{13}$$

In the MDP problem, for the rewards R obtained by the agent after performing a certain behavior, we mainly study the throughput optimization problem of the cognitive wireless network in the model of this paper. Therefore, for the reward of agent, we mainly consider the maximization of network throughput. The network throughput is the number of packets successfully received by the PR and SR per unit time. Therefore, in order to maximize network throughput, it is necessary to minimize the outage probability. After the previous analysis, the network's outage probability includes $\rho_{PT,PR}$, $\rho_{PT,CR_i,SR}$, and $\rho_{CR_i,SR}$, where $\rho_{PT,PR}$ does not change because of the next behavior of the system. Suppose that in the time slot t, the current system selects the behavior $a_{t,i}$ under state k_i; then, the rewards that the system can obtain are

$$\begin{aligned}r_i &= \max R(k_i, a_{t,i}) \\ &= \frac{1}{\min\left(\alpha\rho_{PT,PR} + \beta\rho_{PT,CR_i,SR} + \chi\sum_{i\in N}\rho_{CR_i,SR}\right)}\end{aligned} \tag{14}$$

The coefficients α, β, and χ are weighting coefficients, which are related to the data volume of the three transmission methods.

Since the network throughput is related to the system's transmit power, in order to improve the throughput, the system needs to pay more energy. Therefore, considering the energy-saving problem of the system, this paper defines a system utility function EB_i, which considers both throughput and energy consumption in terms of system performance. The expression of the system utility function EB_i is

$$EB_i = \frac{r_i}{\psi\left(P_T + \sum_{i\in N} P_{CR_i}\right)} \tag{15}$$

4.2. Deep Reinforcement Learning Algorithm. In the reinforcement learning algorithm, we not only pay attention to the return of the current behavior but also consider the long-term impact of the next n steps. In this paper, r_i is used to indicate the reward (or punishment) obtained by the current system after selecting the behavior $a_{t,i}$ under the state k_i; then, the income of the next n steps is expressed as

$$V^{\pi}(k) = E_{\pi}\left[\sum_{i=0}^{n}\gamma^i r_i \mid k_0 = k\right] \tag{16}$$

where k_0 is the initial state of the system, and $\gamma \in (0, 1)$ is the discount factor. In $V^{\pi}(k)$, policy π and initial state k_0 are given by the user, while initial action a is determined by policy π and state k, that is $a = \pi(k)$. In the reinforcement algorithm, in order to evaluate the system state and behavior, the $Q(k, a)$ value is used to represent the state action value function:

$$Q_i(k_i, a_i) = p(k_i, k_{i+1})[r_i + \gamma V^{\pi}(k_{i+1})] \tag{17}$$

When the system state is large, iterative update is required in order to obtain better state behavior data. The iterative formula is used to realize the optimization learning of state action value function:

$$\begin{aligned} Q_{i+1}(k_i, a_i) &= Q_i(k_i, a_i) + \upsilon_i \delta_i \\ \delta_i &= r_i + \gamma \cdot \max Q_i(k_{i+1}, a_{i+1}) - Q_i(k_i, a_i) \end{aligned} \tag{18}$$

The discount factor γ represents the impact of future earnings on current behavior, and υ_i represents the learning rate. The ultimate goal of this paper to enhance learning is to maximize system utility, which maximizes EB_i. Therefore, using the system utility function EB_i of the cognitive relay network instead of r_i, we can get optimized learning function for system utility:

$$\begin{aligned} Q_{i+1}(k_i, a_i) &= (1 - a_i) Q_i(k_i, a_i) \\ &\quad + \upsilon_i (EB_{\mathrm{i}} + \gamma \cdot \max Q_i(k_{i+1}, a_{i+1})) \end{aligned} \tag{19}$$

If the number of cognitive nodes is too large, the traditional reinforcement learning algorithm is difficult to achieve fast convergence of behavior selection when the system chooses relay nodes. Because the deep neural network has the characteristics of good generalization ability, so this paper uses the deep neural network method to establish the mapping relationship between state and behavior based on the traditional reinforcement learning algorithm. The specific implementation is as follows.

In the deep learning method adopted in this paper, first define the Q function, that is the state action value function; we use $Q(k_i, a_i)$ to define the return obtained after taking action a_i under state k_i:

$$Q(k_i, a_i) = E[r_i + \gamma Q(k_{i+1}, a_{i+1})] \tag{20}$$

As can be seen from the above equation, the definition of the Q function is a recursive expression. In the actual situation, if it is in the case of a large state dimension, we cannot recursively calculate the value of Q every step, so we use a convolutional neural network to simulate the Q function. We use K neurons in the convolutional neural network input layer, which are represented as input vector $Input = [l_1, l_2 \cdots, l_K]$, representing the channel state of the network. The output layer neurons are $Output = [a_i]$, indicating that the PT node selects channel i for data transmission. The model structure of the convolutional neural network is shown in Figure 2.

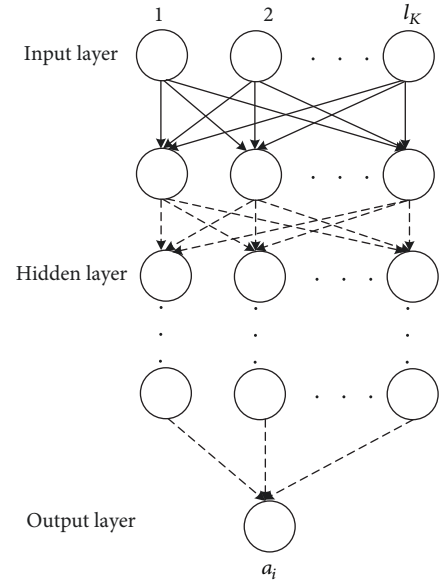

Figure 2: Model structure of convolutional neural network.

The loss function uses the softmax cross entropy loss function to get

$$Loss = \arg\min \frac{1}{2} \sum_i \left(y_i - \log\left(y_i'\right)\right) + h \sum_{i=1}^{S_l} \left(c_i^l\right)^2 \tag{21}$$

h is the weight parameter of the weight matrix, c_i^l represents the l-th weight matrix, and s_l represents the number of neuron nodes in the hidden layer. y_i is the state action value function, that is,

$$y_i = E[r_i + \gamma Q(k_{i+1}, a_{i+1})] \tag{22}$$

The weighting value is then updated using the gradient descent method until the optimal value is reached, at which point the maximum system utility max EB_{i} can be obtained. The update rules of weight and deviation vector are as follows:

$$w(k+1) = w(k) - \beta\left(\frac{\partial Loss}{\partial w}\right) \tag{23}$$

$$b(k+1) = b(k) - \beta\left(\frac{\partial Loss}{\partial b}\right) \tag{24}$$

5. Experimental Results and Analysis

In the simulation experiment, the number of cognitive nodes in the simulated cognitive wireless network scenario is $N = 100$, the data relay mode adopts the AF protocol, the Doppler frequency shift value is $f = 40$Hz, the frame length is $T = 5 \times 10^{-3}s$, the noise power is $P_o = 1 \times 10^{-3}$W, the interference noise power is $P'_{T,CR_i} = 1 \times 10^{-2}$W, the learning rate $a_k = 1 \times 10^{-2}$, the discount factor is $\gamma = 0.8$, the training error precision is less than 2×10^{-5}, and all channels are independent Rayleigh fading channels. The channel remains unchanged during the transmission of one data frame. The

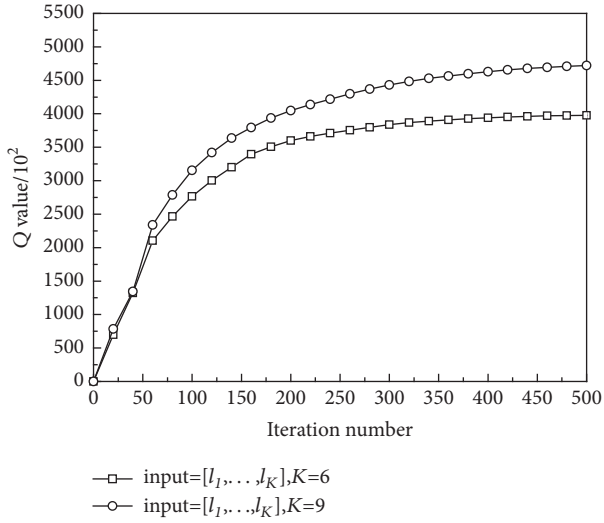

Figure 3: The Q value of convolutional network under different input conditions.

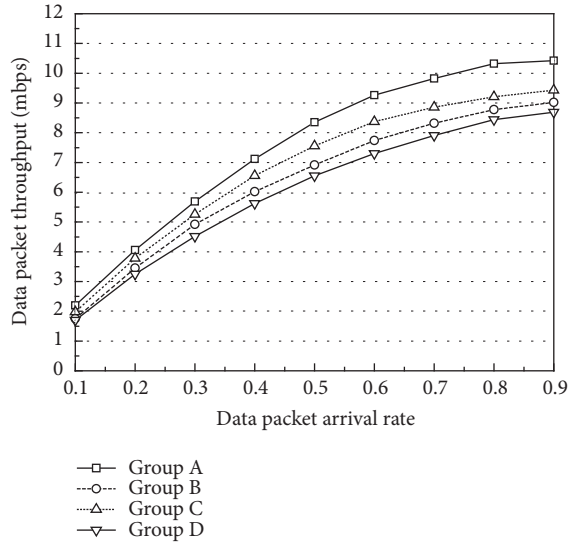

Figure 4: Data packet throughput.

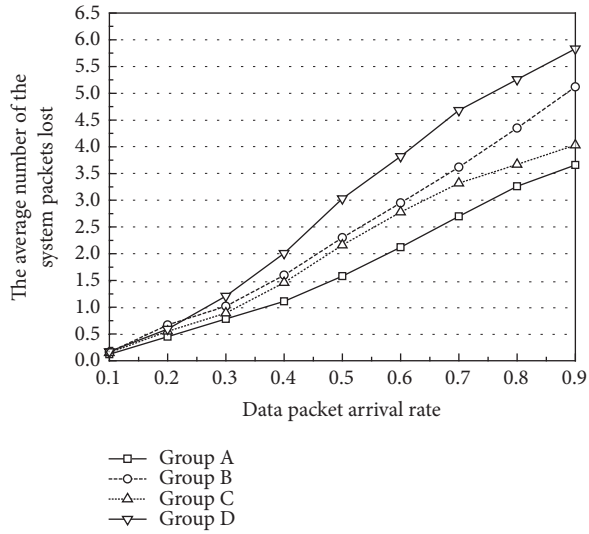

Figure 5: The average number of the system packets lost.

communication performance and computing power of all cognitive nodes are the same.

Figure 3 shows the Q value curve of different input conditions $Input = [l_1, l_2 \cdots, l_6]$, $Input = [l_1, l_2 \cdots, l_9]$. Under the same conditions of other parameters, the Q value of the system under different input conditions gradually converges to different values, indicating the convergence of the algorithm.

In order to verify the performance of the algorithm, four sets of experiments were set up in the experiment for comparative analysis. The first group was group A, which is the algorithm of this paper. The second group is group B; on the basis of the algorithm in this paper, the deep reinforcement learning algorithm is not used, and only the cooperative relay mode is reserved. The third group is group C; on the basis of the algorithm in this paper, the deep reinforcement learning algorithm is replaced by the reinforcement learning algorithm. The fourth group D is a cooperative relay method with minimum energy overhead. The minimum energy cost means that the system selects the cognitive node with the smallest transmit power as the cooperative relay node.

The data packet throughput of the four sets of experiments is shown in Figure 4 under different data packet arrival rates. As can be seen from Figure 4, system data packet throughput is increasing gradually with the increase of data packet arrival rate. The experimental results show that the data packet throughput of group A is more than the data packet throughput of other groups. This is because group A based on system outage probability analysis, using deep reinforcement learning algorithm for behavior selection to optimize system throughput gain, so it performs better in throughput optimization than other groups. If the deep learning algorithm of group A is replaced by the reinforcement learning algorithm, the experimental results are as shown in group C. The throughput of group C is slightly lower than that of group A, but the performance of group C performed better than group B and group D.

Under different data packet arrival rates, the average number of lost packets in the four groups of experiments is shown in Figure 5. It can be seen from the results of Figure 5 that the average number of packet loss of the system also increases with the increase of data packet arrival rate. The increase in the arrival rate of the data packet means that the number of packets transmitted by the system per unit time increases. Since the system has a certain probability of interruption, therefore, the more data packets are transmitted, the greater the number of lost packets. Among them, it can be seen from the performance of group A that the average number of packet loss in the system of this algorithm is the least, and the average packet loss of system in group C is better than that in group B and group D. The average number of

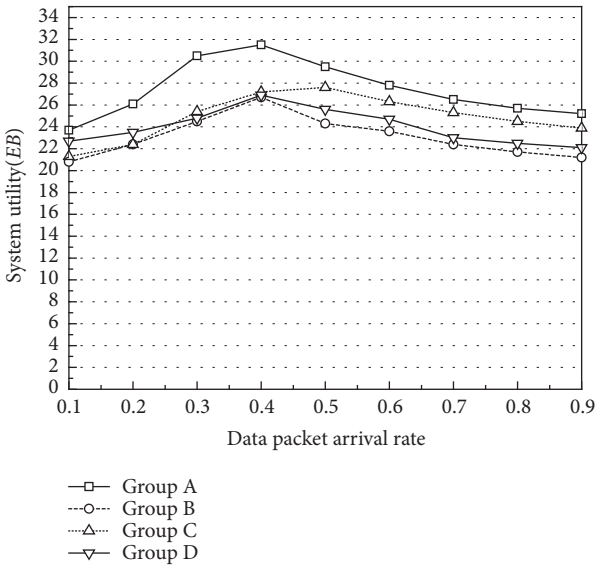

Figure 6: System utility.

packets lost in group D is the largest, because the D group uses the cooperative relay method with minimum energy overhead; only the energy consumption is considered when selecting the relay node, so the overall outage probability of the system may be greater.

The system utility of the four sets of experiments is shown in Figure 6 under different data packet arrival rates. System utility, the function in equation (15), is determined by both throughput and system emission energy. It can be seen from the results in Figure 6, with the increase of data packet arrival rate, the system utility first grows then falls and gradually becomes stable. As the data packet arrival rate begins to increase, system throughput increases, although the system transmission energy consumption also increases, but the throughput increases more, so the total system utility increases. When the data packet arrival rate is increased to a certain extent, due to the limited buffer space of the system, there is not enough space to cache more data, so data loss may occur, and the throughput increase rate decreases, but the system emission energy consumption continues to rise at this time; therefore, the system utility began to decline gradually. Comparing with the experimental results of other groups, the system utility of this paper algorithm has the largest result (group A).

6. Conclusions

In order to improve the throughput of cognitive relay network and optimize system utility, a cognitive relay network throughput optimization algorithm based on deep reinforcement learning is proposed in this paper. In this algorithm, the cooperative relay mode of cognitive relay network is proposed, and the outage probability of the system is analyzed according to the transmission process. The algorithm adopts the deep reinforcement learning method in the selection of relay nodes of cognitive networks and learns the optimal strategy by interacting with the environment to maximize the accumulation of system utility benefits. In the simulation experiment, the experimental results are compared and analyzed by the system throughput, the average number of lost packets, and the system utility. In our algorithm comparison, the proposed one shows better performance than the others, such as improving the throughput of cognitive relay networks and optimizing system utility.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the 2017 School-Level Scientific Research Startup Fund in Xinhua College of Sun Yat-Sen University (Project No. 2017YB005) and the 2017 School-Level Scientific Research Startup Fund in Xinhua College of Sun Yat-Sen University (Project No. 2017YB001)

References

[1] M. E. Ahmed, D. I. Kim, J. Y. Kim, and Y. Shin, "Energy-arrival-aware detection threshold in wireless-powered cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9201–9213, 2017.

[2] J. V. Van Hecke, P. D. Del Fiorentino, V. Lottici, F. Giannetti, L. Vandendorpe, and M. Moeneclaey, "Distributed dynamic resource allocation for cooperative cognitive radio networks with multi-antenna relay selection," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1236–1249, 2017.

[3] T. P. Do, I. Song, and Y. H. Kim, "Simultaneous wireless transfer of power and information in a decode-and-forward two-way relaying network," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1579–1592, 2017.

[4] Z. Chen, T. Lin, and C. Wu, "Decentralized learning-based relay assignment for cooperative communications," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 2, pp. 813–826, 2016.

[5] F. L. D. Silva, R. Glatt, and A. H. R. Costa, "MOO-MDP: an object-oriented representation for cooperative multiagent reinforcement learning," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–13, 2018.

[6] L. P. Tuyen, N. A. Vien, A. Layek et al., "Deep hierarchical reinforcement learning algorithm in partially observable markov decision processes," *IEEE Access*, no. 99, pp. 49089–49102, 2018.

[7] L. Shao, D. Wu, and X. Li, "Learning deep and wide: a spectral method for learning deep networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2303–2308, 2014.

[8] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: a deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2018.

[9] A. W. Ahmad, H. Yang, and C. Lee, "Maximizing throughput with wireless spectrum sensing network assisted cognitive radios," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 195794, 10 pages, 2015.

[10] B. Prasad, A. U. G. Sankararao, S. D. Roy, and S. Kundu, "Throughput and outage of a wireless energy harvesting based cognitive relay network," in *Proceedings of the 5th International Conference on Advances in Computing, Communications and Informatics (ICACCI '16)*, pp. 2009–2014, IEEE, September 2016.

[11] M. Ashour, A. A. El-Sherif, T. ElBatt, and A. Mohamed, "Cognitive radio networks with probabilistic relaying: stable throughput and delay tradeoffs," *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 4002–4014, 2015.

[12] K. Janghel and S. Prakriya, "Throughput of underlay cognitive energy harvesting relay networks with an improved time-switching protocol," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 1, pp. 66–81, 2018.

[13] X. Li, N. Zhao, Y. Sun, and F. R. Yu, "Interference alignment based on antenna selection with imperfect channel state information in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5497–5511, 2016.

[14] Y. Cao, N. Zhao, F. R. Yu et al., "Optimization or alignment: secure primary transmission assisted by secondary networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 4, pp. 905–917, 2018.

[15] S. Zhang, B. Di, L. Song, and Y. Li, "Sub-channel and power allocation for non-orthogonal multiple access relay networks with amplify-and-forward protocol," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2249–2261, 2017.

[16] C. Cai and R. Qiu, "Energy-efficient cooperative two-hop amplify-and-forward relay protocol in cognitive radio networks," *IET Communications*, vol. 10, no. 16, pp. 2135–2142, 2016.