Hindawi
Complexity
Volume 2020, Article ID 1628023, 11 pages
https://doi.org/10.1155/2020/1628023

Research Article

Joint Channel Allocation and Power Control Based on Long Short-Term Memory Deep Q Network in Cognitive Radio Networks

Zifeng Ye,[1,2] Yonghua Wang ,[1] and Pin Wan[1,3]

[1]*School of Automation, Guangdong University of Technology, Guangzhou 510006, China*
[2]*School of Electronics and Communication Engineering, Sun Yat-Sen University, Guangzhou 510006, China*
[3]*Hubei Key Laboratory of Intelligent Wireless Communications, South-Central University for Nationalities, Wuhan 430074, China*

Correspondence should be addressed to Yonghua Wang; sjzwyh@163.com

Received 18 April 2020; Accepted 19 May 2020; Published 11 June 2020

Academic Editor: Shuping He

Efficient spectrum resource management in cognitive radio networks (CRNs) is a promising method that improves the utilization of spectrum resource. In particular, the power control and channel allocation are of top priorities in spectrum resource management. Nevertheless, the joint design of power control and channel allocation is an NP-hard problem and the research is still in the preliminary stage. In this paper, we propose a novel joint approach based on long short-term memory deep Q network (LSTM-DQN). Our objective is to obtain the channel allocation schemes of the access points (APs) and the power control strategies of the secondary users (SUs). Specifically, the received signal strength information (RSSI) collected by the microbase stations is used as the input of LSTM-DQN. In this way, the collection of RSSI can be shared between users. After the training is completed, the APs are capable of selecting channels with small interference while the SUs may access the authorized channels in an underlay operation mode without knowing any knowledge about the primary users (PUs). Experimental results show that the channels are allocated to the APs with a lower probability of collision. Moreover, the SUs can adjust their power control strategies quickly to avoid the harmful interference to the PUs when the environment parameters change randomly. Consequently, the overall performance of CRNs and the utilization of spectrum resources are improved significantly compared to existing popular solutions.

1. Introduction

Cognitive radio networks (CRNs), also known as cognitive wireless networks (CWNs), are formed when cognitive radio devices are organically connected through cognitive base stations. Spectrum resource management is one of the basic tasks of CRNs, which aims to achieve high unitization of the spectrum resource through dividing it into a group of channels or resource blocks and designing proper management strategies. Faced with the increasing demand for mobile data capacity, channel allocation and power control play a key role in spectrum resource management [1, 2].

Spectrum resource management is to determine the most suitable channels for secondary users (SUs) without affecting the communication of primary users (PUs), based on the analysis of available channels. Currently, optimization and game theory have been widely used in spectrum management. In [3], spectrum sharing was made according to interference temperature and radio frequency (RF) power per unit of bandwidth measured in the receiving antenna. The optimal solution can be obtained by particle swarm optimization (PSO) algorithm, if the objective function was convex. In addition, simulated annealing (SA) is applied to prevent falling into suboptimal solutions. Three improved algorithms of PSO, namely, binary PSO, sociocognitive PSO, and derivation zero algorithm were proposed and the throughput of SU links was compared under the interference constraints in [4]. The spectrum access algorithm, proposed

in [5], improved the throughput and spectrum sensing ability of the network system by formulating a Lagrange dual optimization problem and derived the optimal power allocation strategy and target detection probability. In the research of spectrum resource management based on game theory, the core idea is to obtain the equilibrium of optimal distribution of spectrum resources among SUs. In [6], the double auction model from microeconomic theory was used in TV band transactions between TV broadcasting companies and wireless regional area network (WRAN) service providers. For WRAN service providers, spectrum bidding and pricing problems were formulated as a noncooperative game model and obtained the Nash equilibrium. Tehrani and Uysal [7] proposed a sealed bid first-price auction model, aiming to maximize the revenue of service provider and the satisfaction of SUs under incomplete spectrum sensing conditions. Tan et al. [8] considered cooperative and noncooperative spectrum access schemes based on threshold policy. Experimental results showed that, in noncooperative cases, the optimal scheme met the Nash equilibrium.

Existing work using the optimal control or game theory often assumes that users in the wireless networks have obtained the complete environmental state information. However, such information is difficult, if not impossible to obtain in complex and dynamic scenarios, so in many cases, a solution has to be given based on partial environmental information. Inspired by the emerging artificial intelligence, reinforcement learning and neural network provide us a new tool to tackle challenges in CRNs [9–12]. Deep reinforcement learning (DRL) has used the model free feature of reinforcement learning (RL) and the ability of deep learning (DL) to process data in spectrum resource management. The potential advantages of applying DRL to spectrum resource management are threefold. First, the optimal solution for decision-making problems can be obtained through trial and error, and the cycle of manual spectrum planning is greatly reduced. So, CRNs can learn and obtain efficient spectrum resource management solutions. Second, it is possible to simulate the complex real-loop scenario that is difficult to model mathematically and constantly accumulate new experiences to adapt to various extreme situations. Third, real-time effective monitoring of dynamic environment, mining the potentially important data and information, and improving the performance of CRNs can be achieved. These advantages boost a few research works [13–17]. For instance, Wan and Cohen [14] proposed a distributed dynamic spectrum access algorithm based on deep multiuser reinforcement learning, aiming at maximizing network utility in multichannel wireless networks. At each time slot, each SU mapped its current state into the spectrum access action by using the trained deep Q network (DQN). Experimental results showed that, in some observable environments, SUs were able to learn out good control strategies to ensure network performance without using online acknowledgement (ACK) signals. Liu et al. [16] adopted a multiagent DQN technology, which further optimized the learning process by combining the DQN algorithm with transfer learning so that SUs of the new access network could obtain more experience and knowledge.

In spite of the aforementioned research work, spectrum resource management based on DRL is still in its infancy stage. Existing results revealed that the state information of the channels has a high degree of self-correlation [18, 19]. However, this property may have a considerable time interval from the current state. There is still a large gap in the study of this problem. Considering the extraordinary network structure of long short-term memory, it is possible to explore such self-correlation and make a better estimate of the state of the channels. Motivated by the limitations of the current state-of-the-art and the joint design problem of channel allocation and power control for spectrum resource management, this paper proposes a long short-term memory deep Q network- (LSTM-DQN-) based joint channel allocation and power control algorithm, which helps to achieve spectrum utilization flexibility by sharing the received signal strength information (RSSI) among users. Additionally, we consider that PUs may have multiple alternative power control strategies rather than a single strategy and choose the appropriate one dynamically according to the changing environment. The evaluations show that the adjacent access points (APs) access available channels without conflict, whereas SUs maximize the power control strategies to avoid harmful interference to PUs.

The remainder of this paper is organized as follows. Section 2 introduces the system model and formulates the problem to be solved. The implementation of the proposed algorithm is discussed in Section 3. Section 4 describes the simulation experiments and result analysis, and finally, the conclusion and future work are presented in Section 5.

2. Preliminaries

2.1. System Model. The channel allocation problem is raised due to huge number of wireless devices accessing limited spectrum space. In such problem, there is no one-to-one connection between channels and APs. The main challenges are adjacent channel interference (ACI) and co-channel interference (CCI). For the joint optimization of channel allocation and power control, it is necessary to consider not only the transmit power of primary and secondary users but also the selection of channels at different access points and their possible conflicts to each other.

The system model we focus in this paper is shown in Figure 1. There are 5 APs deployed in the scenario, and each AP serves several primary and secondary users distributed randomly within its communication range. We allow overlapping between APs. For instance, the service range of AP1 and AP2 overlap with each other, and so do AP3 and AP4. In contrast, AP5 is independent of others. Within the service range of each AP, the PUs always transmit data on their authorized channels, whereas SUs are only allowed to access channels without affecting the communication of PUs. The base station in the middle is mainly responsible for the communication of PUs. Meanwhile, microcells assist SUs to control the transmit power. These microcells collect the RSSI of primary and secondary users, package the collected information into packets occupying a few bytes,

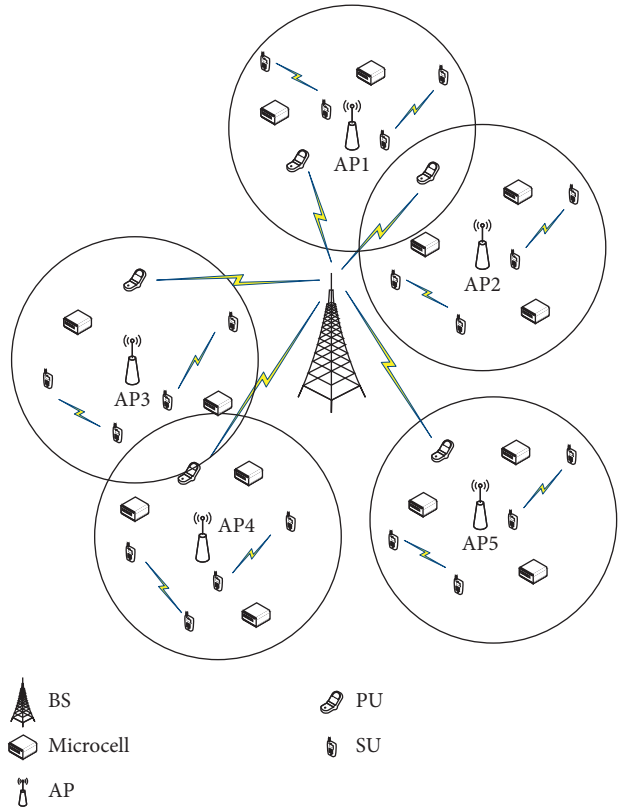

Figure 1: The system model of CRNs.

and then send them to SUs through a dedicated control channel. It is assumed that each PU adjusts the transmitting power according to its own control strategy and always transmits data on its authorized channel. Both PUs and SUs are ignorant of others' power control strategy. To be more specific, PUs are never concerned about the existence of SUs. Therefore, SUs need to learn appropriate transmit power strategies through utilizing the RSSI, as to accomplish their own transmission tasks.

2.2. Problem Formulation. In the joint optimization of channel allocation and power control, the first thing to determine is whether to allow the same channel to be selected between different APs. In this paper, this is not allowed, i.e., we consider the case of no channel conflicts. Based on such assumption, the transmit power and control strategies of primary and secondary users are then determined. Table 1 specifies the symbols used in this paper.

The set of APs is denoted as $\mathscr{P}$, and the set of available channels is $\mathscr{C}$. Each AP can only use one channel. The channel matrix is ρ: $\mathscr{C} \times \mathscr{P} \longrightarrow [0,1]$ in which each element is defined by

$$\rho(c,p) = \begin{cases} 1, & \text{if AP } p \text{ occupies channel } c, \\ 0, & \text{otherwises,} \end{cases} \tag{1}$$

where $c \in \{1,2,\ldots,|\mathscr{C}|\}$, $p \in \{1,2,\ldots,|\mathscr{P}|\}$.

Accordingly, we define $\Omega_{|\mathscr{P}|\times|\mathscr{P}|}$ as the interference matrix, and each element is defined by the following formula:

$$\Omega_{p,q} = \begin{cases} 1, & \text{adjacent AP } p, q \text{ occupy the same channel,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In order to measure the service quality, the SINR of primary and secondary users need to be defined. We assume that the users are able to communicate only if the relevant adjacent APs access the channel successfully. Let the SINR of PU i in AP p at time t be written as follows:

$$\mathrm{SINR}_{i,p}(t) = \frac{\left(1-\Omega_{p,q}\right)h_{ii}^{p}(t)P_{i,p}(t)}{\sum_{j}h_{ji}^{p}(t)P_{j,p}(t)+\delta_{i,p}(t)}. \tag{3}$$

Similarly, the SINR of SU j in AP p at time t is

$$\mathrm{SINR}_{j,p}(t) = \frac{\left(1-\Omega_{p,q}\right)h_{jj}^{p}(t)P_{j,p}(t)}{h_{ij}^{p}(t)P_{i,p}(t)+\sum_{j\neq k}h_{kj}^{p}(t)P_{k,p}(t)+\delta_{j,p}(t)}. \tag{4}$$

In multichannel scenarios, both the available channels and the channel gain change with time. Therefore, the problem becomes dynamic, and thus more complicated. The throughput of a single SU j in AP p at time t is

$$T_{j,p}(t) = W\log_2\left(1+\mathrm{SINR}_{j,p}(t)\right). \tag{5}$$

The objective is to maximize the total throughput of all SUs, which is denoted as follows:

Table 1: Notation of definitions.

Symbol	Definition
$P_{i,p}(t)$	The transmit power of the ith PU in AP p at time t
$P_{j,p}(t)$	The transmit power of the jth SU in AP p at time t
$h_{ii}^{p}(t)$	The channel gain of AP p from the ith PU transmitter to the ith PU receiver at time t
$h_{jj}^{p}(t)$	The channel gain of AP p from the jth SU transmitter to the jth SU receiver at time t
$h_{ji}^{p}(t)$	The channel gain of AP p from the jth SU transmitter to the ith PU receiver at time t
$h_{ij}^{p}(t)$	The channel gain of AP p from the ith PU transmitter to the jth SU receiver at time t
$h_{kj}^{p}(t)$	The channel gain of AP p from the kth SU transmitter to the jth SU receiver (j is not equal to k) at time t
$\delta_{i,p}(t)$	The noise power received by the ith PU of AP p at time t
$\delta_{j,p}(t)$	The noise power received by the jth SU of AP p at time t
$\mu_{i,p}$	The SINR threshold required by the PU
$\mu_{j,p}$	The SINR threshold required by the SU
M_p	The number of PUs in the area served by AP p
N_p	The number of SUs in the area served by AP p
$d_{il,p}(t)$	The distance from the ith PU in AP p to the microcells at time t
$d_{jl,p}(t)$	The distance from the jth SU in AP p to the microcells at time t

$$\begin{aligned} \max \quad & \sum_{p}\sum_{j=1}^{N} T_{j,p}(t) \\ \text{s.t.} \quad & \text{(I)SINR}_{i,p}(t) \geq \mu_{i,p}, \quad \forall i,t \\ & \text{(II)SINR}_{j,p}(t) \geq \mu_{j,p}, \quad \exists j,t \\ & \text{(III)} P_{i,p}(t) \geq \sum_{j} P_{j,p}(t). \end{aligned} \tag{6}$$

3. Deep Reinforcement Learning-Based Framework

Due to the widespread application of CRNs, the network structure is becoming more and more complex. It is difficult to establish a corresponding mathematical model to simulate a highly complex network environment. The model-free RL can effectively solve this problem. In recent years, DRL has shown excellent ability in dealing with complex problems and data operations. Therefore, this paper focuses on the application of DRL in spectrum resource management, especially the joint optimization of power control and channel allocation to improve the robustness and adaptability of CRNs.

3.1. Description of RL. The model-free learning is one type of method through continuous interaction with the virtual environment in RL. In general, RL constructs the problem as a Markov decision process (MDP). At every moment t, the agent can observe the current state of the environment $s \in S$ and then select an action $a \in A$. After the action is executed, the environment state is transitioned with a certain probability $P_{ss'}(a)$ to a new state $s' \in S$. Meanwhile, the environment will feed back a reward value $r \in R$ to the agent. The schematic diagram is shown in Figure 2. In a word, RL aims to find the best strategy by maximizing the cumulative reward value through a limited number of steps [9].

Using RL to solve the joint design problem in CRNs, an array (S, A, R) should be defined in advance, where S represents the set of environmental states, A is the set of SU actions, and R: $S \times A \longrightarrow \Re$ denotes the reward obtained when taking the next action in the current state.

3.1.1. State Space. There are 5 APs deployed in the network environment, with several primary and secondary users around each AP. The SUs can only obtain incomplete environmental information at APs to implement their transmission tasks. Assuming that L microcells are responsible for collecting the RSSI of primary and secondary users in the service area of each AP, a total of 5 L microcells are distributed in the whole network environment. We adopt a discretized-time model. According to the nonfree space propagation [20], the RSSI collected by the microcells in the area served by the AP p at time slot t is denoted by the following equation:

$$\mathbf{S_p}(t) = \left[s_{1,p}(t), s_{2,p}(t), \ldots, s_{L,p}(t)\right]^T, \tag{7}$$

where $s_{l,p}(t)$ is defined by

$$s_{l,p}(t) = \sum_{i=1}^{M_p} P_{i,p}(t)\left[\frac{d_{il,p}(t)}{d_0(t)}\right]^{-\tau} + \sum_{j=1}^{N_p} P_{j,p}(t)\left[\frac{d_{jl,p}(t)}{d_0(t)}\right]^{-\tau} + \Delta(t). \tag{8}$$

Therefore, the RSSI of these 5 APs is integrated and used as the input layer of LSTM-DQN, namely,

$$\mathbf{Input}(t) = \left[\mathbf{S}_1(t), \mathbf{S}_2(t), \mathbf{S}_3(t), \mathbf{S}_4(t), \mathbf{S}_5(t)\right]. \tag{9}$$

3.1.2. Action Space. We add the set of SU transmit power into the action space, and the action of all SUs in AP p at time t is

$$\mathbf{A_p}(t) = \left[P_{1,p}(t), P_{2,p}(t), \ldots, P_{N,p}(t)\right]^T, \tag{10}$$

where $P_{j,p}(t)$ represents the transmit power of the SU j in AP p.

Therefore, the action value of all APs in the whole network environment is

$$\mathbf{Action}(t) = \left[\mathbf{A}_1(t), \mathbf{A}_2(t), \mathbf{A}_3(t), \mathbf{A}_4(t), \mathbf{A}_5(t)\right]. \tag{11}$$

3.1.3. Reward Function. For the problem of channel allocation and power control, it is firstly necessary to consider

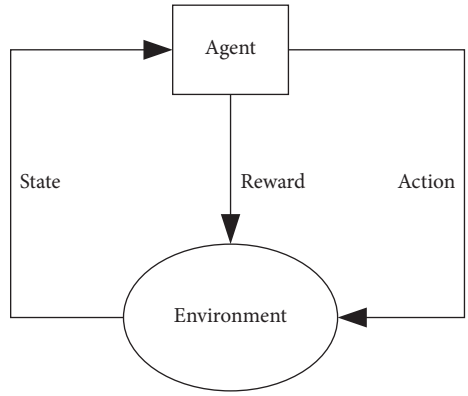

Figure 2: The interaction model of RL.

that the channels are selected by APs without conflict. Specifically, APs 1 and 2 choose different channels, 3 and 4 choose different channels, and 5 can choose any channel. Only after the APs successfully select the channels can the users perform data transmission. It should be considered that both primary and secondary users in each AP meet the service quality requirements and do not exceed the threshold. According to the constraint conditions, the reward at AP p is defined by the following equation:

$$R_p(t) = \begin{cases} \sum_j \mathrm{SINR}_{j,p}, & I_{11}, \\ -\sum_i \mathrm{SINR}_{i,p}, & I_{22}, \\ -\left(\sum_i \mathrm{SINR}_{i,p} + \sum_j \mathrm{SINR}_{j,p}\right), & \text{otherwise}, \end{cases} \tag{12}$$

where the constraints are given as follows: I_{11}: AP p access the available channel, $\forall \mathrm{SINR}_i \geq \mu_i$, $\exists \mathrm{SINR}_j \geq \mu_j$ and $\forall P_i \geq \sum_j P_j$ and I_{22}: AP p accesses the available channel, $\mathrm{SINR}_i \leq \mu_i$, $\quad i \in \{1, 2, \ldots, N\}$.

The reward function of the whole network system is

$$R(t) = \frac{\sum_p R_p(t)}{|\mathscr{P}|}, \tag{13}$$

which represents the mean value of rewards obtained by all APs.

3.2. Power Control Strategy of PUs. We consider that the PUs can adjust their transmit power according to the specified control strategy and always transmit data on the authorized channels. The typical power control strategy proposed in [21] is

$$P_i(k+1) = D\left(\frac{\mu_i P_i(k)}{\mathrm{SINR}_i(k)}\right), \tag{14}$$

where the value of $D(x)$ is no less than the minimum value of x according to the predefined range of the discretization threshold.

We also adopt the more intelligent strategy proposed in [22] as follows:

$$P_i(t+1) = \begin{cases} P_i(t) + \Delta P, & \mathrm{SINR}_i(t) \leq \mu_i \text{ and } \mathrm{SINR}_i'(t) \geq \mu_i, \\ P_i(t) - \Delta P, & \mathrm{SINR}_i(t) \geq \mu_i \text{ and } \mathrm{SINR}_i'(t) \geq \mu_i, \\ P_i(t), & \text{otherwise}, \end{cases} \tag{15}$$

where $\mathrm{SINR}_i'(t) = h_{ii}(t)P_i(t+1)/[\sum_{j\neq 1} h_{ji}(t)P_i(t) + \delta_i(t)]$, which represents the SINR of the PU i at the predicted time $t+1$.

When a PU conducts the intelligent control strategy of equation (15), according to the current SINR at time t and the predicted SINR at time $t+1$, it only needs to adjust its own transmit power only once. Therefore, the advantage of this intelligent strategy lies in that it can reduce the extra energy consumption caused by frequent power switching. At the same time, it comprehensively considers the trend estimation to determine whether the PU should adjust its transmit power and has the ability of spectrum prediction.

In order to cope with the complexity of network environment, PUs may have multiple alternative power control strategies rather than a single strategy and choose the appropriate one according to the actual situation. Equation (14) is denoted as power control strategy 1 of the PU, and equation (15) is strategy 2. We will discuss and analyse these strategies in detail in the experiments in Section 5.

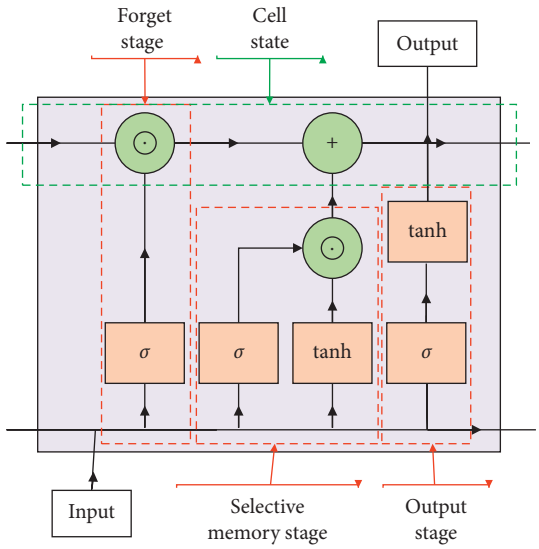

Figure 3: The unit of LSTM.

3.3. LSTM-DQN-Based Joint Channel Allocation and Power Control Algorithm. LSTM is a special recurrent neural network (RNN) [23]. As shown in Figure 3, the unit of LSTM mainly includes the forget stage, selective memory stage, and output stage, which is realized through the forget gate, input gate, and output gate, respectively. The core of LSTM is to control the cell state through these three interactive gate states. It can catch the important but implicit knowledge for a long time and discard the unnecessary message. Therefore, it shows excellent performance in solving the problem of gradient disappearance or gradient explosion in the process of long sequence training.

On one hand, it is verified that the state information of the channels has a high degree of self-correlation, which may have a considerably long time interval from the current state [24]. On the other hand, there is great potential to improve the probability of successfully access the channels owing to the unique network structure of LSTM because LSTM can effectively capture valuable knowledge that is not obvious. To track the implicit correlation over a long period of time, we combine LSTM with DQN (as shown in Figure 4) to integrate the collected partial known information and obtain better control strategies through offline learning. Once the training phase is completed, the users only need to communicate with the central unit by slightly adjusting the weight of the neural network. At each moment, the APs select the available channels and the SUs choose the optimal transmit power according to the trained DQN. The specific algorithm is shown in Algorithm 1.

4. Performance Evaluation

In this section, we evaluate the performance of our proposed algorithm through simulation-based experiments.

4.1. Experiment Settings. In our simulated scenarios, there is a circular area with a radius of 1,000 m. 3 available channels are provided for 5 APs. AP 1 has overlap with AP 2, and AP 3 has overlap with AP 4. AP 5 is independent of others. There are 10 microcells in the service range of each AP, where 1 PU and 2 SUs contend for accessing the spectrum resources. Thus, the whole network environment includes one base station, 50 microcells, 5 PUs, and 10 SUs. Specifically, the transmission power range of the PU is $\{0.0, 5.0, 10.0, \ldots 30.0\}$ mW, and the transmit power range of SU is $\{0.0, 1.0, 2.0, \ldots 12.0\}$ mW. The white noise is 0.1 mW. The SINR thresholds for primary and secondary users are 1.0 dB and 0.5 dB, respectively. According to the path loss rule of nonfree space, the channel model is now considered as the 2-ray ground reflection model of wireless propagation, and the channel gain expression is

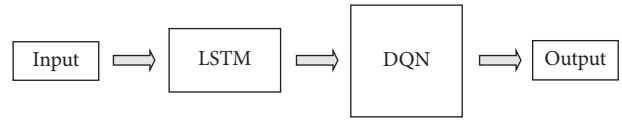

Figure 4: The structure of LSTM-DQN.

$$g = \frac{G_t G_r h_t^2 h_r^2}{d^\tau}, \tag{16}$$

where path loss index $\tau = 4$, G_t and G_r are the gain of the transmitter and receiver, respectively, and h_t and h_r are the heights of the transmit and receive antennas, respectively [20]. In order to simulate the complex change of the environment, the number of each iteration is now set to 40,000. Furthermore, the position of primary and secondary users in the environment as well as the channel gain are randomly initialized every 10,000 iterations.

The LSTM-DQN is constructed with 5 hidden layers. The first hidden layer is the LSTM layer, and the middle 4 hidden layers are the full connection layer. The number of neurons in the full connection layer is 256, 128, 128, and 256, respectively. The activation function of the second, third, and fourth hidden layers adopt ReLUs function, and the activation function of the fifth hidden layer is tanh function. Besides, Adam algorithm is used to update the weight of the neural network. The size of the training samples is set to 128. The initial exploration probability of greedy algorithm is 0.8 and linearly decreases to 0 with the number of iterations. Moreover, the memory bank has a capacity of 1,000, whereas training is not started until the capacity reaches 500 or more.

For the dynamic and the complexity of the application environment, we consider the PUs take different power control strategies. One case is in which the PUs take single control strategy 2. Another one is that each time the environmental parameters are updated, the power control strategy of 1 or 2 is chosen randomly by PUs. The proposed joint algorithm based on LSTM-DQN will be compared with two benchmark algorithms: the original DQN-based algorithm and priority memory combined with DQN- (PM-DQN-) based algorithm.

4.2. Simulation Results. Figure 5 shows the loss function of different algorithms when the PUs adopt control strategy 2,

(1) Initialization: the capacity O of memory D, the transmit power of PU and SU is $P_{i,p}(t), P_{j,p}(t)$ respectively, the channel interference matrix $\Omega_{|\mathscr{P}|\times|\mathscr{P}|}$, LSTM-estimates LSTM-DQN Q weight $\theta = \theta_0$, targets LSTM-DQN $\widehat{Q}$ weight $\widehat{\theta} = \theta_0$
(2) **For** episode = 1 to E **do**
(3) According to the initial state **Input**(0), SUs randomly select actions **Action**(0) with ε probability, otherwise choose actions **Action**(0) = $\max_a Q(\textbf{Input}(0), a; \theta)$ with $1 - \varepsilon$ probability
(4) **For** $t = 1$ to T **do**
(5) The PUs update the transmit power according to their own power control strategies
(6) SUs select actions **Action**(t) with ε_t probability, otherwise select the action **Action**$(t) = \max_a Q(\textbf{Input}(t), a; \theta)$
(7) Obtain rewards $R(t)$ and the next state **Input**$(t+1)$
(8) Save empirical data $d(t) \equiv \{\textbf{Input}(t), \textbf{Action}(t), R(t), \textbf{Input}(t+1)\}$ to memory D
(9) **If** $t > O/2$ **then**
(10) Select training sample $d(l)$ randomly from D
(11) Calculate $\widehat{Q}(l) = R(l) + \gamma \max_{a'} \widehat{Q}(\textbf{Input}(l+1), a'; \widehat{\theta})$
(12) Use the gradient descent method to minimize the loss function $[\widehat{Q}(l) - (\textbf{Input}(l+1), a'; \widehat{\theta})]^2$ and update parameters θ
(13) **End If**
(14) **End For**
(15) Reset environment parameters randomly
(16) **End For**

Algorithm 1: The joint design algorithm of LSTM-DQN.

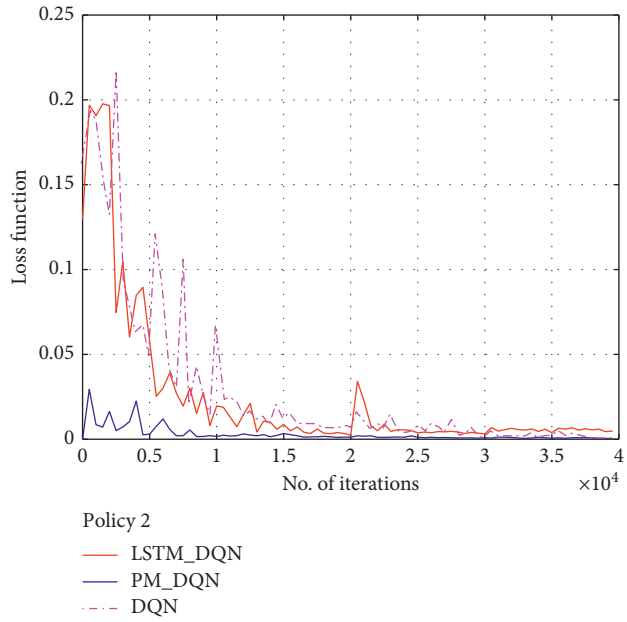

Figure 5: Relationship between the number of iterations and loss function (policy 2).

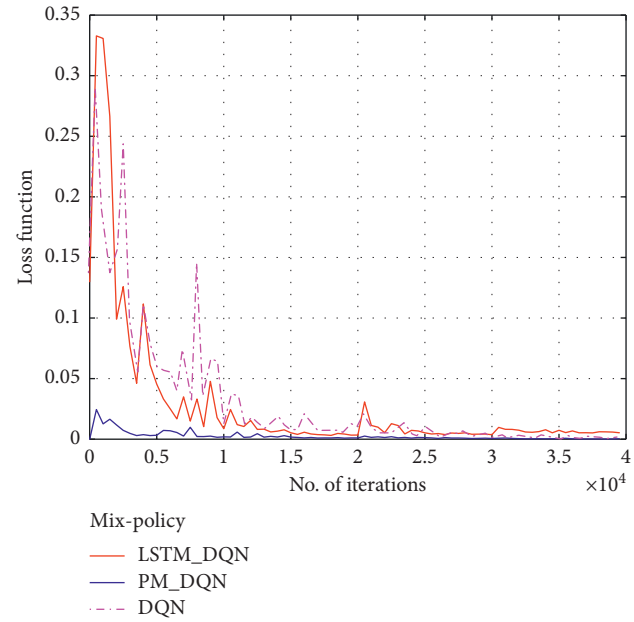

Figure 6: Relationship between the number of iterations and loss function (mix-policy).

and Figure 6 plots the loss function when the PUs employ mixed control strategies. It can be seen that all of algorithms meet convergence after iterative learning. Our LSTM-DQN algorithm has a large instantaneous fluctuation when the environmental parameters change, which is slightly better than the benchmark. On the other hand, the algorithm based on PM-DQN has less fluctuation. This is because the PM greatly accelerates the convergence rate of the loss function by cutting off the correlation, whereas the LSTM needs to correlate the past experience so that the loss function does not converge to the minimum value quickly. Nevertheless, it is meaningful for the joint problem of channel allocation and power control without Markov property. We will explain from other aspects below.

Figures 7 and 8 describe the comparison of the cumulative rewards when the PUs adopt a single and mixed control strategies, respectively. It can be seen from the results that the reward of the benchmark algorithm is always decreasing, whereas the cumulative rewards of our LSTM-DQN and the algorithm based on PM-DQN are relatively stable. Moreover, the reward of LSTM-DQN is higher. It is worth noting that the cumulative reward of LSTM-DQN is close to or slightly higher than the horizontal line of 0, which indicates that the channel allocation and power control

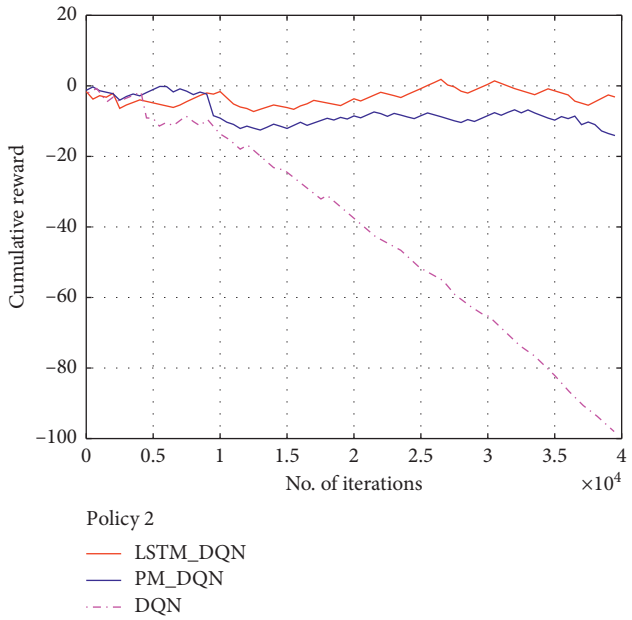

Figure 7: Relationship between the number of iterations and reward function (policy 2).

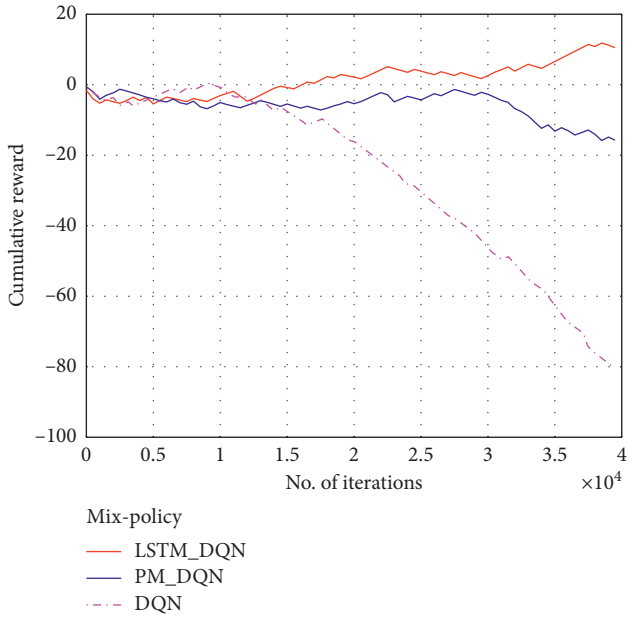

Figure 8: Relationship between the number of iterations and reward function (mix-policy).

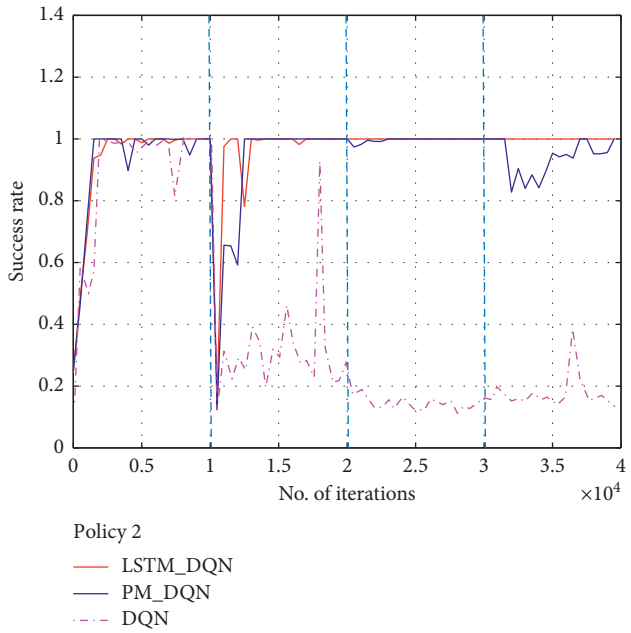

Figure 9: Relationship between the number of iterations and success rate (policy 2).

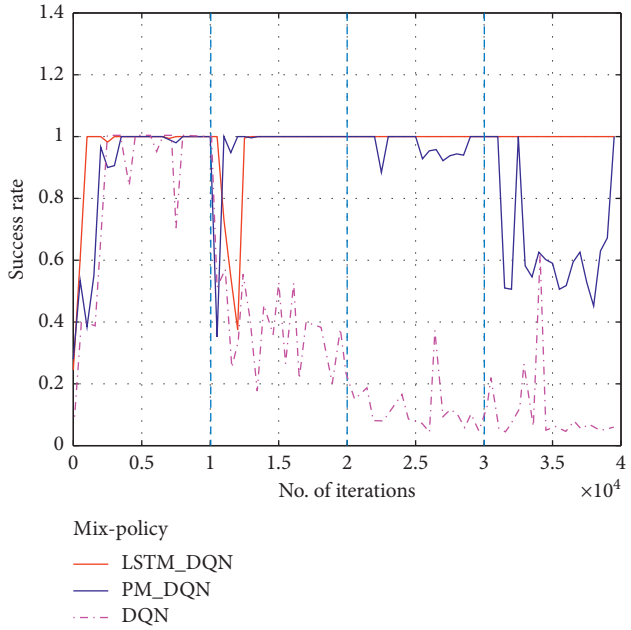

Figure 10: Relationship between the number of iterations and success rate (mix-policy).

scheme still have room for further improvement in the future work.

Figures 9 and 10 are evaluated in terms of the switching success rate. Once the user is able to access the channel and successfully complete the transmission task within 20 switches, it is deemed to a successful experience. It can be concluded from the simulation results that our LSTM-DQN can ensure the maximum success rate and adjust the strategy rapidly when the environment parameters are updated randomly. Moreover, when the PU adopts the mixed strategy, the proposed algorithm can still show excellent robustness and desirable generalization ability.

Figures 11 and 12 depict the comparison of handover steps. We observe that regardless of the control strategies adopted by the PUs, and the proposed algorithm guarantees that the optimal strategy can be found after an average of one

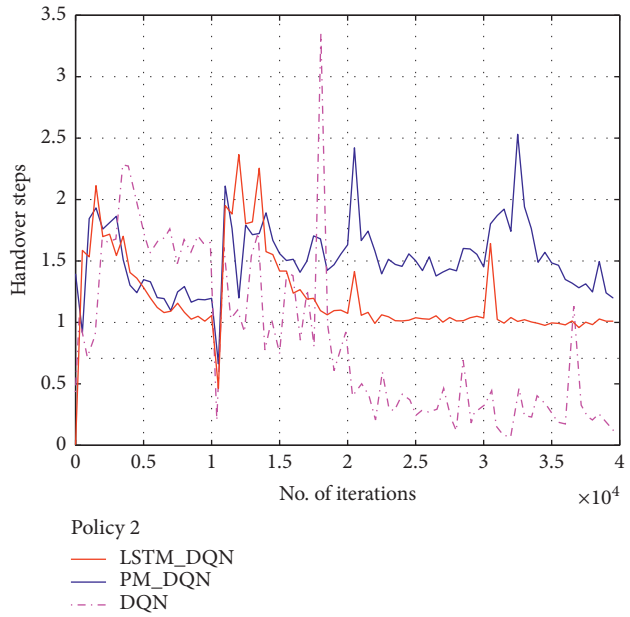

Figure 11: Relationship between the number of iterations and handover steps (policy 2).

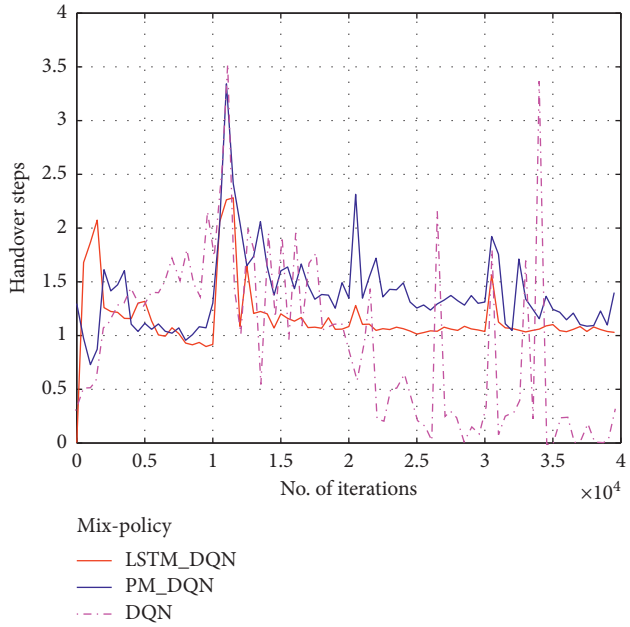

Figure 12: Relationship between the number of iterations and handover steps (mix-policy).

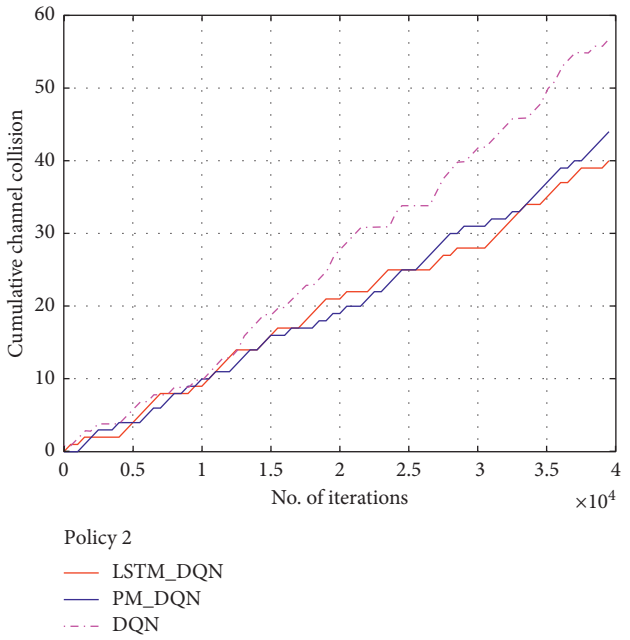

Figure 13: Relationship between the number of iterations and channel collision (policy 2).

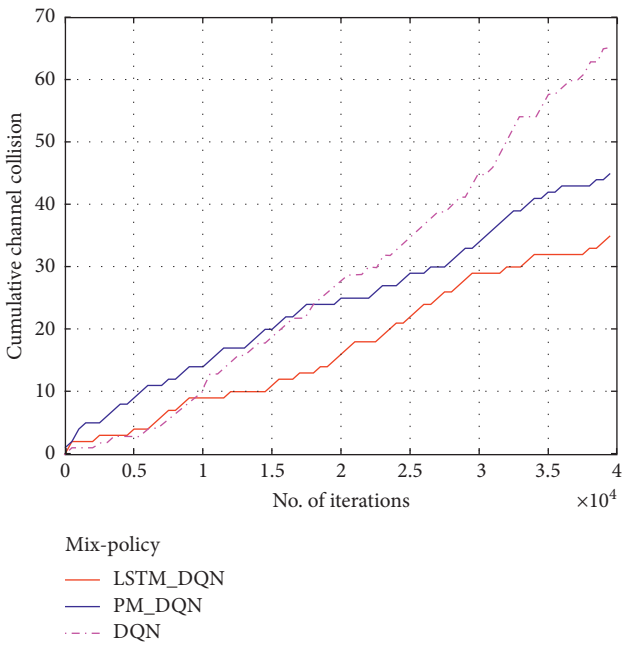

Figure 14: Relationship between the number of iterations and channel collision (mix-policy).

handover. It helps reduce the energy consumption and greatly improve the sensitivity of the users, which can react to the change of the real-time environment more quickly. Moreover, when the environmental parameters update, the proposed algorithm shows the anti-interference performance and generalization ability.

We then analyse the channel cumulative conflicts shown in Figures 13 and 14. When the PUs take the single control strategy, the proposed algorithm and the algorithm based on PM-DQN perform closely. In the situation that PUs employ the mixed strategy, LSTM-DQN-based algorithm can further reduce channel conflict. It shows that the proposed algorithm has a good potential in dealing with complex conditions.

5. Conclusion and Future Work

Aiming at the joint design problem of channel allocation and power control in CRNs, this paper proposed a novel

algorithm based on LSTM-DQN. We analysed the feasibility and implementation process of the proposed algorithm. Through simulation-based experiments, the advantages of LSTM-DQN-based algorithm were discussed and illustrated from the aspects of loss function, reward function, success rate, handover steps, and channel cumulative conflict. Specially, our proposed method outperformed other two DQN-based competitors.

Our future work will involve using real data to verify the feasibility of the algorithm. Moreover, various factors of the environment, e.g., mobility of users, can be taken into account, as to further study the large-scale spectrum resource management problems.

Data Availability

The data used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, 12 months after publication of this article, will be considered by the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant no. 61971147), Special Funds from Central Finance to support the development of local universities (Grant nos. 400170044 and 400180004), Foundation of National & Local Joint Engineering Research Center of Intelligent Manufacturing Cyber-Physical Systems, and Guangdong Provincial Key Laboratory of Cyber-Physical Systems (Grant no. 008).

References

[1] J. Chapin and W. Lehr, "Cognitive radios for dynamic spectrum access - the path to market success for dynamic spectrum access technology," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 96–103, 2007.

[2] S. Srinivasa and S. Jafar, "Cognitive radios for dynamic spectrum access-the throughput potential of cognitive radio: a theoretical perspective," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 73–79, 2007.

[3] M. Tang, C. Long, and X. Guan, "Nonconvex dynamic spectrum allocation for cognitive radio networks via particle swarm optimization and simulated annealing," *Computer Networks*, vol. 56, no. 11, pp. 2690–2699, 2012.

[4] A. Martínez-Vargas and Á. G. Andrade, "Comparing particle swarm optimization variants for a cognitive radio network," *Applied Soft Computing*, vol. 13, no. 2, pp. 1222–1234, 2013.

[5] S. Stotas and A. Nallanathan, "On the throughput and spectrum sensing enhancement of opportunistic spectrum access cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 97–107, 2012.

[6] D. Niyato, E. Hossain, and Z. Zhu Han, "Dynamic spectrum access in IEEE 802.22-based cognitive wireless networks: a game theoretic model for competitive spectrum bidding and pricing," *IEEE Wireless Communications*, vol. 16, no. 2, pp. 16–23, 2009.

[7] M. N. Tehrani and M. Uysal, "Auction based spectrum trading for cognitive radio networks," *IEEE Communications Letters*, vol. 17, no. 6, pp. 1168–1171, 2013.

[8] S.-S. Tan, J. Zeidler, and B. Rao, "Opportunistic spectrum access for cognitive radio networks with multiple secondary users," *IEEE Transactions on Wireless Communications*, vol. 12, no. 12, pp. 6214–6227, 2013.

[9] R. S. Sutton and A. G. Barto, "Reinforcement learning," *Journal of Cognitive Neuroscience*, vol. 11, no. 1, pp. 126–134, 1999.

[10] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[11] Y. Rusu, K. G. Vamvoudakis, and H. Modares, "Safe reinforcement learning for dynamical games," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 9, pp. 2706–3726, 2020.

[12] Y. Wang, Z. Ye, and P. Wan, "A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks," *Artificial Intelligence Review*, vol. 51, no. 3, pp. 493–506, 2019.

[13] Y. Zhao, S. Zhang, Y. Zhang, P. Wan, and S. Wang, "A cooperative spectrum sensing method based on signal decomposition and *k*-medoids algorithm," *International Journal of Sensor Networks*, vol. 29, no. 3, pp. 171–180, 2019.

[14] O. Wan and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310–323, 2019.

[15] S. He, M. Zhang, H. Fang, F. Liu, X. Luan, and Z. Ding, "Reinforcement learning and adaptive optimization of a class of Markov jump systems with completely unknown dynamic information," *Neural Computing and Applications*, 2019.

[16] J. Liu, V. Koivunen, S. R. Kulkarni et al., "Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks,," in *Proceedings of the 2011 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pp. 642–646, Aachen, Germany, May 2011.

[17] C. Wang, H. Fang, and S. He, "Adaptive optimal controller design for a class of LDI-based neural network systems with input time-delays," *Neurocomputing*, vol. 385, pp. 292–299, 2020.

[18] D. Willkomm, S. Machiraju, J. Bolot et al., "Primary users in cellular networks: a large-scale measurement study,," in *Proceedings of the 2008 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pp. 1–11, Chicago, IL, USA, October 2008.

[19] X. Xing, T. Jing, W. Cheng, Y. Huo, and X. Cheng, "Spectrum prediction in cognitive radio networks," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 90–96, 2013.

[20] T. S. Huo, *Wireless Communications: Principles and Practice*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2002.

[21] S. A. Grandhi, J. Zander, and R. Yates, "Constrained power control," *Wireless Personal Communications*, vol. 1, no. 4, pp. 257–270, 1994.

[22] X. Li, J. Fang, W. Cheng et al., "Intelligent power control for spectrum sharing in cognitive radios: a deep reinforcement learning approach," *IEEE Access*, vol. 6, no. 25, pp. 463–473, 2018.

[23] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proceedings of the 2015 AAAI Fall Symposium Series*, pp. 29–37, Arlington, Virginia, 2015.

[24] M. Malajner, K. Benkic, P. Planinsic et al., "The accuracy of propagation models for distance measurement between WSN node," in *Proceedings of the 2009 16th International Conference on Systems, Signals and Image Processing*, pp. 1–4, Chalkida, Greece, June 2009.