Reinforcement Learning-Based Intelligent Decision-Making for Communication Parameters

Xia. Xie[1], Zheng Dou[1], and Yabin Zhang[1*]
[1] Dept. of Information and Communication Eng., Harbin Engineering University
Harbin, Heilongjiang, China
[e-mail: xx2015@hrbeu.edu.cn, douzheng@hrbeu.edu.cn, zhangyabin@hrbeu.edu.cn]
[*]Corresponding author: Yabin Zhang

Received April 23, 2021; revised August 23, 2021; revised February 17, 2022; accepted August 18, 2022; published September 30, 2022

Abstract

The core of cognitive radio is the problem concerning intelligent decision-making for communication parameters, the objective of which is to find the most appropriate parameter configuration to optimize transmission performance. The current algorithms have the disadvantages of high dependence on prior knowledge, large amount of calculation, and high complexity. We propose a new decision-making model by making full use of the interactivity of reinforcement learning (RL) and applying the Q-learning algorithm. By simplifying the decision-making process, we avoid large-scale RL, reduce complexity and improve timeliness. The proposed model is able to find the optimal waveform parameter configuration for the communication system in complex channels without prior knowledge. Moreover, this model is more flexible than previous decision-making models. The simulation results demonstrate the effectiveness of our model. The model not only exhibits better decision-making performance in the AWGN channels than the traditional method, but also make reasonable decisions in the fading channels.

Keywords: reinforcement learning, decision-making, Q-learning, cognitive radio, adaptive modulation and coding.

1. Introduction

In recent years, the rapid development of Internet of Things (IoT) and Vehicle Networks have made the requirements of communication quality even higher. At the same time, the exploration of extreme communication environments, such as polar and space, also raises higher demands on the environment adaptability of communication systems. As the current research trend of wireless communication, the concept of cognitive radio (CR) was first proposed by Joseph Mitola in 1999 [1], which is a further development based on software radio (SR) [2]. Currently, the research on CR primarily focuses on the spectrum, including spectrum sensing, dynamic spectrum allocation, and spectrum management [3,4]. However, CR should not remain in the "dynamic spectrum using radio" [5]. In our opinion, the next generation intelligent radio systems should be able to sense and extract effective information from the complex channels, then it would be able to make decisions based on the perceived results, and intelligently generate the waveform parameter configuration that is most suitable for the current channel [6]. Finally, the reliable and efficient transmission of information can be achieved. In this paper, we aim to study the intelligent decision-making of communication parameters which is the core component of the above process.

The decision-making of the waveform parameter can be considered as a further expansion and evolution of traditional adaptive modulation and coding (AMC) [7]. However, intelligent decision-making methods do not require the complex analysis and calculation of signal to noise ratio (SNR) thresholds in AMC, and are not restricted by the limited number of thresholds. RL-based decision-making can interact with the environment and does not rely on background knowledge and prior information. It has strong learning ability and flexibility. so that it can be applied to any channel condition. This paper proposes a new intelligent decision-making model by introducing the idea of RL. The complete decision-making framework is shown in **Fig. 1** [8].

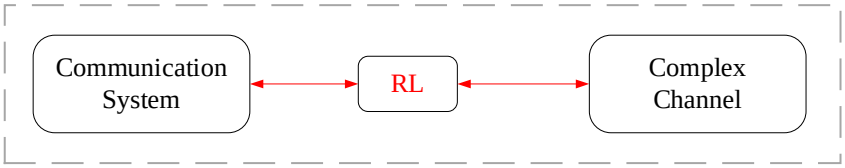

Fig. 1. The proposed decision-making model framework based on RL.

The previous decision-making models pay too much attention to the complicated analysis and modeling for channels. However, in the complex channel, for example, the bit error rate (*BER*) of different modulation and coding is difficult to be calculate. In this paper, we want to solve the problem from a new perspective. We do not care about the specific characteristics of the channel, but directly obtain the real feedback of the channel through the on-line interaction between reinforcement learning (RL) and the environment. Compared with previous decision-making models, our model has the following advantages:

•The concise state-action pairs and the modeling method avoid the large-scale problem and can converge faster.

•The model has low complexity and high efficiency, which is suitable for practical applications.

•The model is not limited by conditions and is applicable in any channel environment.

We can find the optimal parameter configuration for the communication system in any channel, and the model does not depend on any prior knowledge. In practical applications, the

RL-based method may exhibit poor performance in the initial exploration stage. In the future, we will continue to complete the model to solve this problem. We will design a classifier to store and utilize the existing decision-making results reasonably to guide future decisions.

The remainder of this paper is organized as follows: The related work about communication decision-making is presented in Section 2. Section 3 gives a brief overview of RL theory. Then we describe the communication system used in the proposed model in Section 4. Based on the above, we discuss the problems of existing models and our improvement idea, then propose a new RL-based decision-making model in Section 5. The simulation results are shown and discussed in Section 6. Finally, the paper is concluded in Section 7

2. Related Work

We will classify and introduce the decision-making methods according to their dependence on prior knowledge form high to low. The earliest expert system model is enhanced by continuous off-line simulation, and ultimately provides decision rules that can be used on-line to adapt the radio equipment to various environments [9,10]. However, the complex and varied channels in practice bring about great challenges to the design of expert systems. Since the system relies entirely on the prior knowledge, it can be difficult for it to work beyond its knowledge reserve.

The idea of optimization-based decision-making is to transform the decision-making problem into a multi-optimization problem. Based on the designed objective function that characterizes the communication performance, genetic algorithm (GA) performs an optimal search within a configurable range, and finally obtains the optimal parameters. Rieser proposed a biologically inspired cognitive radio framework based on GA, and then carried out simulation and hardware testing [11]. The related research, such as the cognitive decision engines based on ant colony optimization (ACO) [12], bacterial foraging optimization [13], etc., have mainly focused on improving the optimization performance of the algorithm. In addition, there are cognitive decision-making based on binary chaotic particle swarms (BPSO) [14], hill climbing genetic algorithm [15], and other algorithms engine. However, most of the researches in this area concentrate on the ideal additive white gaussian noise (AWGN) channel. When the real channel is complex, it is difficult for us to make accurate theoretical formulas and apply the decision engine. Moreover, the process of the optimization search requires a lot of time.

The ways of off-line learning solve the problem can be roughly divided into two categories. One method is collecting a large amount of data under the guidance of theoretical analysis and previous experience to train classifiers that can obtain output parameter configuration according to the input channel condition. The algorithms include case-based reasoning (CBR) [16], neural network (NN) [17], support vector machine (SVM) [18,19], etc. This raises a concern: when the deviation between the input channel and the historical training set is large, the output result will not be reliable. The other method is to introduce the idea of regression prediction into decision-making. Dong used neural networks to predict performance BER [20]. Dou proposed a NN-based predictive model by randomly measuring the performance of part of the parameter values [21]. Thus, the off-line learning methods require many historical cases and are not suitable for guiding decision-making individually. The learning universality and real-time performance are both poor.

On-line learning decision-making is mainly achieved by RL [22]. RL interacts with the environment, obtains environmental information and reinforcement signals, and learns based on practical experience [23]. RL is very suited to solve decision-making problems because of its inherent characteristics. RL was originally used for the spectrum to solve the problems of

spectrum congestion and insufficient spectrum utilization [24,25]. Later, RL was used for the parameter decision-making problem [26,27]. However, there are still some problems in these existing applications: they are still following the idea of optimization based decision-making and do not make full use of the interaction between the RL algorithm and the environment. We will analyze these RL-based models and propose our improved decision-making model in Section 5.

3. Background On Reinforcement Learning

In this paper, we just introduce some necessary conceptions to assist readers to easily understand the proposed model. The clearly detailed description and mathematical derivations are provided in [28].

We describe RL model as shown in **Fig. 2**.

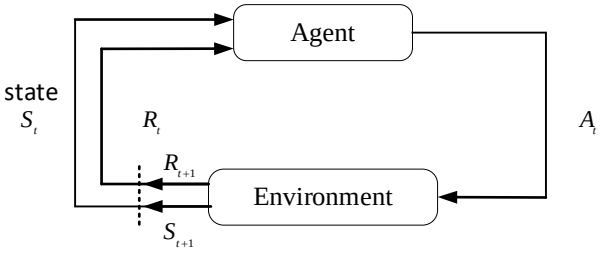

Fig. 2. A typical reinforcement learning interaction model.

At each step $s_m = 1,2,3,4,5,6$, the agent executes action A_t ,and receives observation S_t and scalar reward R_t .The environment receives action A_t, then emits observation S_{t+1} and scalar reward R_{t+1}. The agent's job is to maximize the cumulative reward. RL algorithms can be categorized into model-based and model-free according to whether the agent attempts to build a model. In model-free algorithms, the agent interacts directly with the environment to estimate and optimize the value function. The algorithms can be categorized into Monte-Carlo (MC) learning, Temporal-Difference (TD) learning, and TD(λ) learning.

Model-free algorithms learn directly from episodes of experience under the condition of having no knowledge of MDP transitions /rewards. MC methods must learn from complete episodes, which means that MC learning is not bootstrapping. The main idea of MC is to replace the value with the mean return. However, TD learning can learn from incomplete episodes, which means that it is bootstrapping. In MC learning, we update the value $V(S_t)$ towards the actual return G_t:

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right) \tag{1}$$

Where α is the updated factor. In TD learning, we update the value $V(S_t)$ towards the estimated return $R_{t+1} + \gamma V(S_{t+1})$:

$$V(S_t) \leftarrow V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right) \tag{2}$$

$R_{t+1} + \gamma V(S_{t+1})$ is called the TD target, and ($\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$) is called the TD error.

By comparing MC learning and TD learning, we find that TD can learn before knowing the final outcome and learn online after every step, but MC must wait until the end of the

episode. TD has low variance and is usually more efficient than MC. TD is also more efficient in Markov environments because it can exploit the Markov property.

Sarsa and Q-learning are two classic TD algorithms. Sarsa usually chooses a more conservative strategy compared with Q-learning. For example, in a trapped environment, Sarsa usually chooses the safest route away from traps. Q-learning, however, is bolder and more efficient, it generally chooses the fastest route. In real-time decision-making. problems, efficiency is a very important element. We choose the efficient and practical Q-learning algorithm to build the proposed model. The convergence of Q-learning can be proven.

Q-learning is a kind of off-policy learning algorithm. The core idea of off-policy learning is learning about policy π from experience samples from another policy μ:

$$V(S_t) \leftarrow V(S_t) + \alpha(\frac{\pi(A_t \mid S_t)}{\mu(A_t \mid S_t)} \cdot (R_{t+1} + \gamma V(S_{t+1})) - V(S_t)) \tag{3}$$

These methods can also re-use experience generated from old policies and learn about the optimal policy following the exploratory policies. Q-learning is one of the best algorithms to apply the off-policy idea. The key component of Q-learning is that when the agent updates the value of $Q(S_t, A_t)$, it uses the next state-action value of the target policy instead of the value of the current behavioral policy. The updated equation is given as (4):

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_t) - Q(S_t, A_t)] \tag{4}$$

In (4), the behavioral policy α is the $\varepsilon - greedy$ policy and the target policy γ is the greedy policy. $\varepsilon - greedy$ is the simplest idea for ensuring continual exploration. It makes all m actions tried with none-zero probability in any state. Under $\varepsilon - greedy$ exploration, the agent chooses the greedy action with probability $1-\varepsilon$ and chooses an action at random with probability ε.

$$\pi(a \mid s) = \begin{cases} \varepsilon / m + 1 - \varepsilon & \text{if } a^* = \arg\max_{a \in A} Q(s,a) \\ \varepsilon / m & \text{otherwise} \end{cases} \tag{5}$$

In this way, the agent can converge to the optimal state-action value function while ensuring that it sufficiently experiences reaching new states.

4. Background on Reinforcement Learning

We consider an orthogonal frequency division multiplexing (OFDM) system in our model design because of its unique efficiency and flexibility. OFDM is a special multicarrier transmission scheme that can be viewed as a modulation technique, and it is also a diversity technology. Because of the orthogonality of OFDM subcarriers, OFDM can overcome the shortcoming of low spectrum utilization in general multi-carrier modulation. Meanwhile, multi-carrier transmission can resist frequency selective fading and narrowband interference to some extent. OFDM technology has been extensively used in various communication scenarios.

Another feature of OFDM is the ease of the control of the modulation and coding scheme. We can arbitrarily configure any M-PSK or M-QAM modulation mode in the transmission process. Because the subcarriers are easily allocated and the modulation mode is easy to be changed, the OFDM system is suitable for building a parameter configurable decision-making model. In the proposed model, we presently stipulate that each subcarrier is modulated with the same modulation mode. In future research, we will make more in depth decisions on the parameter configuration of each subcarrier according to the channel and spectrum environment.

The convolutional code is also selected as the way of channel coding in our system. Additionally, a cyclic prefix of the appropriate length is inserted as the guard interval in each OFDM symbol so that the inter-symbol interference (ISI) can be eliminated. More specific details of the parameter configuration will be provided in the simulation section.

In this paper, we discuss not only the additive white Gaussian noise (AWGN) channels, but also the fading channels. We assume that the channel does not vary within one OFDM symbol. At the receiver, we do ideal estimation and equalization in the fading channels to avoid the impact of different performance of different channel estimation algorithms. In the decision-making model, the channel estimation algorithm is not the focus of research. In practical applications, we can choose the appropriate algorithm according to different communication environments and needs.

5. RL-Based Decision-Making Model

In this section, we propose a waveform parameter decision-making model with a new idea based on the analysis of the relationship between digital communication system performance and communication waveform parameters.

The decision-making problem of communication waveform parameters can be simplified as choosing a best performing parameter configuration to adapt to the current channel. Therefore, the problem can be divided in to three key parts, the relationship among them is shown in **Fig. 3**:

The parameters to be decided: Our research is based on the digital communication system. By analyzing a typical digital communication link and its modules, we find that The modulation scheme and the coding rate of channel coding have the most influence on the communication performance.

The environmental factors affecting decision-making： Various adverse factors in the Channel impact on the communication performance, and mostly include noise, interference, and fading. We can estimate these factors by channel estimation and spectrum detection.

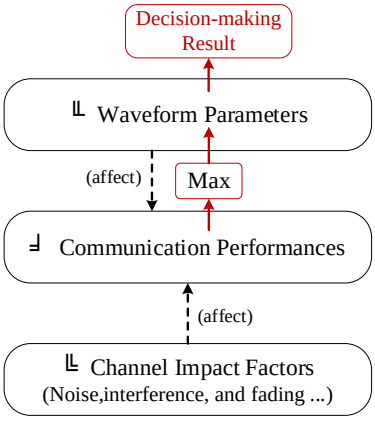

Fig. 3. The disassembly of the waveform parameter decision-making problem.

Theoretical basis for decision-making： The basis for the decision-making problem is the way to measure the communication performance. The performance of a digital communication system can be described in terms of both effectiveness and reliability. *BER* is the most intuitive parameter to indicate reliability, and the transmission rate is usually used to describe effectiveness.

Based on the above analysis, we select the modulation order m and the coding rater as the parameters to be decided in the OFDM system. Meanwhile, the OFDM Transmission rate and *BER* are the indicators for decision-making. However, effectiveness and reliability are two opposite performance indicators. In the same channel environment, when **m** and rate increased, the values of the BER and transmission rate both increase. This also denotes that when the effectiveness of the system is getting better, the reliability becomes worse. Therefore, the communication performance cannot be absolutely optimal. If the effectiveness is supposed to be improved, it will inevitably lead to a decline in the reliability. In the decision-making model, we can merely make trade-offs based on the actual demands.

At present, the research on RL-based waveform parameter decision-making still follows the idea of optimization-based decision-making, such as paper[29],[30], etc. The decision-making process of this model type is described in **Fig. 4**.

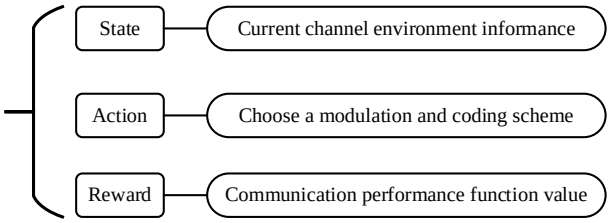

Fig. 4. Previously typical RL-based decision-making model.

In their model, the state is defined as the current channel information and the action refers to selecting one modulation and coding scheme from the optional range. The reward function still follows the multi-objective function in the optimization-based decision-making:

$$f_{\text{decision}_{\max}}(x) = \omega_1 f_{\max}(R) + \omega_2 f_{\min}(BER) + \omega_3 f_{\text{constant}}(W) \tag{6}$$

As described in (6), the goals for their system are: maximizing rate R , minimizing *BER* and keeping the bandwidth W constant. In [29], they also mentioned the problem that attempting to achieve multiple goals at the same time can cause the competition. Their approach to minimize this dispute is using weight. In every episode, they send training sequences by using the current parameter configuration， and then they can obtain the optimization function value as the reward. Applying RL to a decision-making problem in this way is a single-step problem, and it simplifies the episode to only one step. Thus, they modify the updated (4) as:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} - Q(S_t, A_t)\right] \tag{7}$$

However, RL essentially concerns a sequential decision-making problem, timing, which is a key factor in RL. Obviously, this type of application does not make full use of the interaction between the RL agent and the environment. It may need to try all the optional parameter configurations once to make a decision. When there are many optional parameter values, this kind of model will be difficult to converge and may suffer from limitations in a large-scale search.

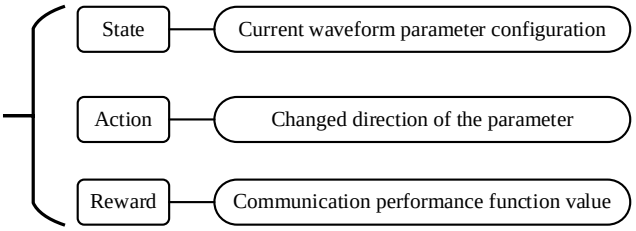

Fig. 5. A new application mode of RL in waveform parameter decision-making.

Inspired by the classical application of RL in the maze games, we design a new RL-based decision-making model, which is represented in **Fig. 5**. We want to solve the decision-making problem in a new way by making full use of the interactivity of RL.

We desire to simulate the direct process of human making decision. Alternatively, when trying to make decision for waveform parameters, we first select one parameter configuration arbitrarily. We then try to increase or decrease the parameter values and determine the direction of next action according to the communication performance after changing parameter value. The decision-making problem is analogous to the problem of finding the shortest path to the optimal parameters. It is very similar to the classical maze problem in RL. Therefore, in our model, no matter where the parameters locate initially, they will eventually arrive at the optimal position. In the context of this idea, RL model can be mapped as follows.

In our new RL-based decision-making model, we set the state *S* to the current waveform parameter configuration, including the modulation order and the coding rate. The states of the modulation mode consist of BPSK, QPSK, 8QAM, 16QAM, 32QAM, and 64QAM, which means the modulation order *m* is from 1 to 6, and we number them as $s_m = 1,2,3,4,5,6$. The state of the coding rate *c* include 1/3, 1/2, 2/3, and 1, which are numbered as $s_c = 1,2,3,4$.

The action *A* in the proposed model is the changing direction of the parameter, which means whether we increase or decrease the modulation order or the coding rate. In order to build the model, we summarize the three optional actions as adding or subtracting s_m and s_c by 1 or keeping the value unchanged. This modeling way avoids too much complexity and too long exploration process caused by the excessive actions.!!!

The most important part of the model is the reward *R*, which determines the goal of the decision-making. We set the reward as the communication performance value at the next time after the agent takes an action. As mentioned above, the *BER* and transmission rate are considered as the digital communication performance in this model. They are contradictory and cannot be improved simultaneously. Similar to (6), in the optimization-based decision-making, they want to solve this problem by introducing a set of weighting factors:

$$f(x) = \sum_{i=1}^{k} \omega_i f_i(x) \tag{8}$$

where $f_i(x)$ is the optimization sub-objective. The weighting factors can take different values for different application scenarios. However, there is no reference telling us how to choose the weight. This is difficult to measure for the multi-objective optimization.

In this case, we might as well simplify the optimization objective as obtaining the peak balance between *BER* and transmission rate. In this paper, we establish the performance equation with reference to the general throughput formula in AMC:

$$\eta = \left(1 - P_e\right) \cdot \rho_i \cdot \log_2 M_i \tag{9}$$

where ρ_i is the data rate and M_i means the number of modulation constellations. On this basis, we modify (9) to adapt to our actual modeling situation. Specifically, the reward equation suitable for our model combining the *BER* and transmission rate is shown in (10)-(12).

$$C = C_{ber} \cdot C_{\text{rate}} \tag{10}$$

$$C_{\text{rate}} = \sqrt{m \cdot \left(r_c \cdot 6\right)} \tag{11}$$

$$C_{ber} = -\log_{10}(\text{ber}) \tag{12}$$

C_{ber} and C_{rate} respectively represent the reliability and the effectiveness. *ber* is the actual bit-error-rate value, which is measured by the power of 10. *m* is the modulation order which has

the same meaning as $\log_2 M_i$ in (9) and r_c is the coding rate.

As we are only concerned about the order of the magnitude of *ber*, we take the logarithm of *ber* and adopt the opposite value as the indicator of reliability in (12). The greater the value of C_{ber} is, the greater reliability the system exhibits. In addition, we make two constraints to *ber* due to the limited precision:

•When $ber > 0.1$, we consider that the *BER* is beyond the scope of tolerance and record it as 1.

•When $ber < 10^{-6}$, we consider that the *BER* reaches the ideal state and record it as 10^{-6}.

The value of the modulation order *m* varies from 1 to 6, and the coding rate r_c ranges from 1/3 to 1. In order to balance the proportion of these two factors in the transmission rate, we expand r_c proportionally. We then multiply them together and obtain the square root of this product as C_{rate} to maintain an equilibrium with C_{ber}. The system performance *C* is jointly represented by numerically balanced C_{ber} and C_{rate}, as (10). In the practical application, (10) can be adjusted according to the actual situation. Therefore, the reward *R* can be expressed as $R = C_{t+1}$.

Following the analysis of RL algorithms in this section, we use the Q-learning algorithm to implement the proposed model. **Algorithm 1** demonstrates the procedures to build a decision-making model based on Q-learning, indicating the entire decision-making process. In the following algorithm, a_m and a_c are the optional actions in the decision-making.

Algorithm. 1 Decision-making model based on Q-learning

Initialize $Q(s_m, a_m)$ and $Q(s_c, a_c)$ arbitrarily, for all $\forall s \in S$, $\forall (s_m, s_c) \in S, \forall a_m \in A(s_m)$, $\forall a_c \in A(s_c)$.

Repeat
  (for each episode) Initialize S_m, S_c.
  Repeat
  (for each step of episode) (n_m and n_c are alternately decided, $A = A_m$ or A_c, $S = S_m$ or S_c.
  Choose A from S using policy derived from $Q(eg., \varepsilon - greedy)$.
  Take action A ,observe R, S'.
  $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_{a'} Q(S', a') - Q(S,A)]$
  $S \leftarrow S'$
  Until episode is end.
Until S are terminal

6. Simulation Results and Discussion

In this section, a set of simulation results are exhibited to analyze the performance of the proposed decision-making model for various variable configurations. We then choose the optimal variable configuration for Q-learning, and demonstrate its performance compared to that of the traditional model. The main parameter configuration as shown in **Table 1**.

Table 1. The main parameter configuration for simulation channels and the OFDM system.

Channel parameters	Channel type	AWGN/Rayleigh
	E_b / N_0	-4dB-20dB
OFDM system parameters	Available subcarriers	64
	Used subcarriers	48
	Cyclic prefix length	$1/4\times$ Data Length
	Bandwidth	10MHZ
	Modulation mode	BPSK，QPSK，16QAM，8QAM，32QAM，64QAM
	Coding mode	Convolutional code
	Coding rate	1/3，1/2，2/3，1

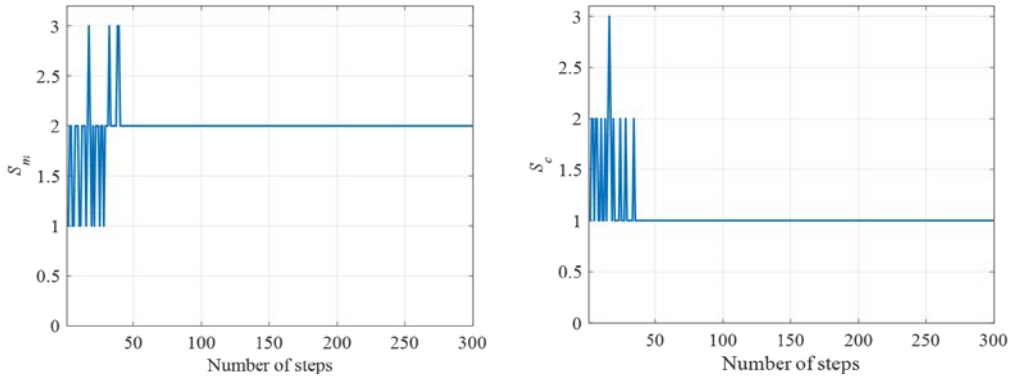

Fig. 6. The decision curves for the AWGN ($E_b / N_0 = 0\text{dB}$) channel based on Q-learning.

Fig. 6 shows that the final decision-making result for the AWGN ($E_b / N_0 = 0\text{dB}$) channel. The left graph of **Fig. 6** presents the decision curve for the modulation order. The simulation result demonstrates the modulation order gradually converges at $s_m = 2$, where the modulation mode is QPSK. Similarly, the right graph illustrates that the coding rate ultimately converges at $s_c = 1$ The small fluctuations in the decision curve are due to the exploratory character of the $\varepsilon - greedy$ policy and the closest suboptimal parameters. In order to show the decision-making process more clearly, we also take AWGN ($E_b / N_0 = 0\text{dB}$) channel as an example to present the learned Q-value tables for the decisions.

Table 2. Q-value table for the modulation-order decision in the Q-learning model

(S_m, S_c) / a_m	(1,1)	(1,2)	(1,3)	(1,4)	(2,1)
+1	**8.658**	**3.323**	0	0	5.285
—	6.004	2.335	0	0	**10.82**
-1	/	/	/	/	6.010
(S_m, S_c) / a_m	(2,2)	(2,3)	(2,4)	(3,1)	(3,2)
+1	0	0	0	0	0
—	**4.062**	0	0	5.223	0
-1	2.661	0	0	**10.55**	**3.290**

Table 3. Q-value table for the coding-rate decision in the Q-learning model

(S_m, S_c) / a_m	(1,1)	(1,2)	(1,3)	(1,4)	(2,1)
+1	2.356	0	0	0	3.320
—	**6.010**	2.335	0	0	**10.82**
-1	/	**5.409**	**2.123**	0	/
(S_m, S_c) / a_m	(2,2)	(2,3)	(2,4)	(3,1)	(3,2)
+1	0	0	0	0	0
—	0	0	0	**5.285**	0
-1	**5.568**	0	0	/	**5.223**

In **Table 2** and **Table 3**, Each state *S* corresponds to three optional actions *A*. The bold data in the table represents the most valuable one of the three actions which indicates the final decision-making result. We get to know that no matter which initial state s_m and s_c are, they will eventually reach the optimal states, $s_m = 2$ and $s_c = 1$, and this is consistent with the decision result shown in **Fig. 6**. The position with Q-value of "0" in the table indicates that this state-action pair does not appear in the exploration process, and the Q-value of other states not shown in the table is also "0". It illustrates that the RL-based model can avoid the exploration of some invalid state-action pairs and improve the decision-making efficiency. Therefore, the final decision-making result for AWGN ($E_b / N_0 = 0\text{dB}$) is (QPSK, $r_c = 1/3$).

Fig. 7 shows that the final decision-making result for the AWGN ($E_b / N_0 = 10\text{dB}$) channel is $s_m = 4$, $s_c = 3$, which denotes (16QAM, $r_c = 2/3$). Similarly, the decision-making result for the AWGN ($E_b / N_0 = 20\text{dB}$) channel is (64QAM, $r_c = 2/3$), as shown in **Fig. 8**. By analyzing the decision-making results, we can find that the modulation order and the coding rate increase with the increase of the E_b / N_0 value. This demonstrates the rationality of the results.

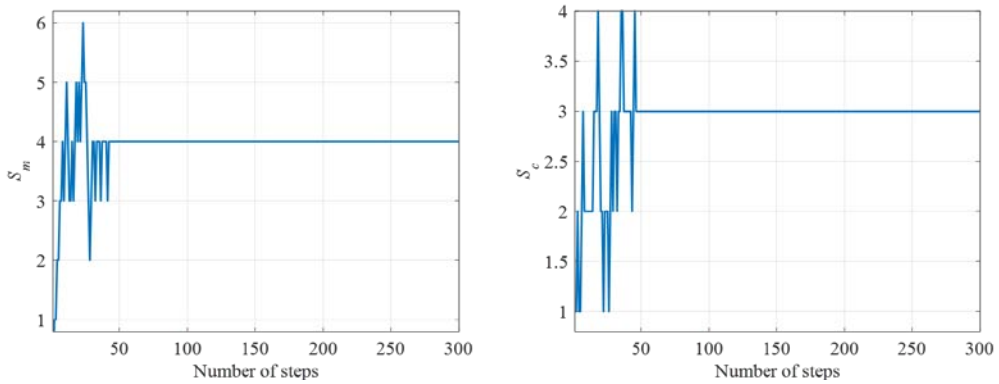

Fig. 7. The decision curves for the AWGN ($E_b / N_0 = 10\text{dB}$) channel based on Q-learning.

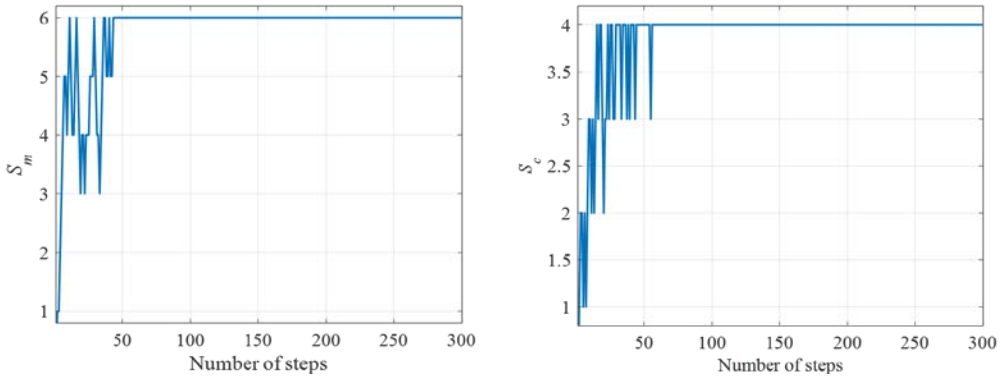

Fig. 8. The decision curves for the AWGN ($E_b / N_0 = 20\text{dB}$) channel based on Q-learning.

Firstly, taking the decision-making in AWGN ($E_b / N_0 = 20\text{dB}$) channel as an example, we compare the convergence performance of BPSO、the existing RL-based decision-making model mentioned in Section 5 and the new RL-based model proposed in this paper. The specific parameters of the communication system are also configured according to **Table 1**. The simulation results are the average of 500 Monte-Carlo experiments.

Comparing these three convergence curves in **Fig. 9**, the BPSO-based decision engine optimization process is relatively gentle and slow, and it converges after about 150 iterations. the existing RL-based decision-making model converges at about 75 times. The RL-based decision model in this paper gradually converges after exploring about 40 times. The existing RL-based decision model has a large number of actions, which leads to a large exploration space and a long time to complete the exploration. However, the state-action pairs of the new RL-based decision model is simple, avoiding the large- scale exploration problem. It makes the decision process more effective and efficient. The initial mean square error (MSE) of the convergence curve of the new RL-based decision model in **Fig. 9** is large because the model is set to start with $s_m = 1$ and $s_c = 1$ for each decision-making process.

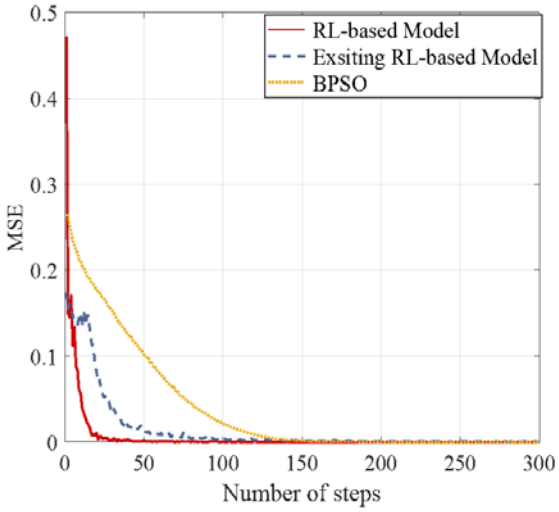

Fig. 9. Convergence performance comparison between the two decision-making models

We continue to simulate our RL-based decision-making model in the AWGN channels with various E_b / N_0. values, and compare the decision-making results with the MCS method [31].

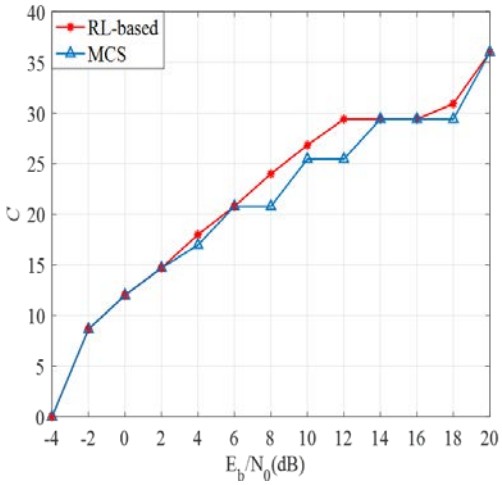

Fig. 10. Convergence performance comparison between the two decision-making models

The decision-making results we obtained in the ideal AWGN channels with different E_b / N_0, values using the RL-based model and MCS mapping respectively are shown in **Fig. 10**. From **Fig. 10**, it is evident that the performance of the RL-based model is not inferior to the MCS method. This means the proposed model can provide reasonable decision-making results. The difference in performance between the two methods is primarily due to the limited thresholds of the MCS mapping. The coarse discretization of the channel state may lead to performance loss in specific regions. Nevertheless, the RL-based model does not need prior knowledge or extensive simulations over different scenarios to obtain the E_b / N_0 thresholds.

In this subsection, we further compare the performance of the RL -based model and the MCS mapping in typical fading channels. We simulate three fading environments defined by 3GPP TS36.104 protocol in LTE standard for different application scenarios, including extended pedestrian A (EPA) channel, extended vehicular A (EVA) channel and extended typical urban (ETU) channel by using the Jakes model based Rayleigh-fading channel. In **Table 4**, we list the specific parameter values of each channel model.

Table 4. Channel parameter values including excess tap and relative power for each path.

EPA		EVA		ETU	
Delay(ns)	Power(dB)	Delay(ns)	Power(dB)	Delay(ns)	Power(dB)
0	0.0	0	0.0	0	-1.0
30	-1.0	30	-1.5	50	-1.0
70	-2.0	150	-1.4	120	-1.0
90	-3.0	310	-3.6	200	0.0
110	-8.0	370	-0.6	230	0.0
190	-17.2	710	-9.1	500	0.0
410	-20.8	1090	-7.0	1600	-3.0
		1730	-12.0	2300	-5.0
		2510	-16.9	5000	-7.0

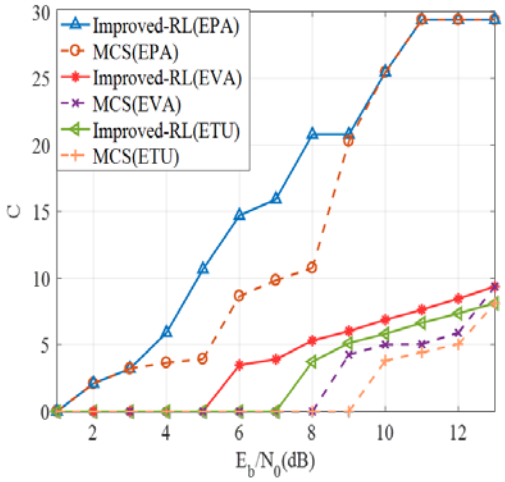

Fig. 11. Performance comparison between RL-based model and MCS.

Fig. 11 respectively present the decision-making performance curves of the RL-based model and MCS in three kind of Rayleigh- fading channels with different values. Compared to **Fig. 10**, there are apparent inconsistencies between the two decision-making methods. Moreover, from the EPA channel to the ETU and EVA channels, the fading becomes severe, and the performance of MCS is even less able to keep up with the RL-based model. The problem of the MCS mapping is clearly exposed when the Gaussian assumption does not exist. In the fading channels, the switching thresholds set on the ideal channel will no longer apply. At this time, the proposed RL-based model is still able to learn on-line from the actual environment, and exhibits excellent decision-making performance.

In order to improve decision-making efficiency and make the model more suitable for waveform parameter decision-making problems, this subsection will optimize the exploration process of RL-based decision model.

In [30], based on the RL-based waveform parameter decision model shown in **Fig. 4**, it is proposed to classify actions into "good actions" and "bad actions" according to the set performance threshold. The threshold is set to a certain percentage of the maximum performance value. At the same time, the "action rejection probability" is used to control the exploration of "good actions" and "bad actions", and the decision-making efficiency of the model can be improved by filtering out "bad actions" to a certain extent. Inspired by this idea, this section will also optimize the online decision-making process of the RL-based decision model by adding judgments of "good" and "bad" action performance.

According to the communication performance reward function constructed in the RL-based decision model, while the reward is set to $R = C_{t+1}$, the communication performance function change $\Delta C = C_{t+1} - C_t$ obtained after the action is used to determine the “good” and “bad” of the current action. "Make judgments. If $\Delta C \geq -0.1$, the action is considered to be a "good action" for the current state, then keep the action and continue the exploration process; if $\Delta C < -0.1$, then the action is considered a "bad action" for the current state. Remove the action from the range of optional actions in the current state, and re-select from the remaining actions to explore.

Comparing the performance of the RL-based decision model after adding the "action judgment mechanism" and the decision model before the improvement, the communication system and channel in the improved decision model still adopt the parameter configuration in **Table 1**.

Fig. 12-Fig. 14 compare the decision curves of the improved model and the original RL-based decision model under three different AWGN channels with different E_b / N_0 values. It can be seen from **Fig. 12-Fig. 14** that the trend of the decision curve before and after the improvement is basically similar, and the final decision result is also the same, but the exploration process of the model after adding the "action judgment mechanism" is obviously shortened, and the maximum can be shortened by 9 times, which shows that the improved decision model no longer needs to spend time on "poor performance" actions and states, can quickly find the position of the optimal waveform parameters and converge, and the exploration process is more efficient.

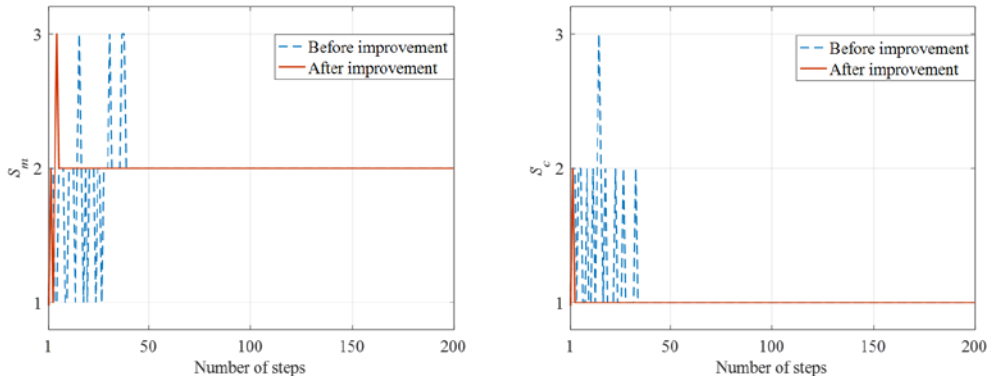

Fig. 12. Performance comparison in the AWGN($E_b / N_0 = 0\text{dB}$) channel

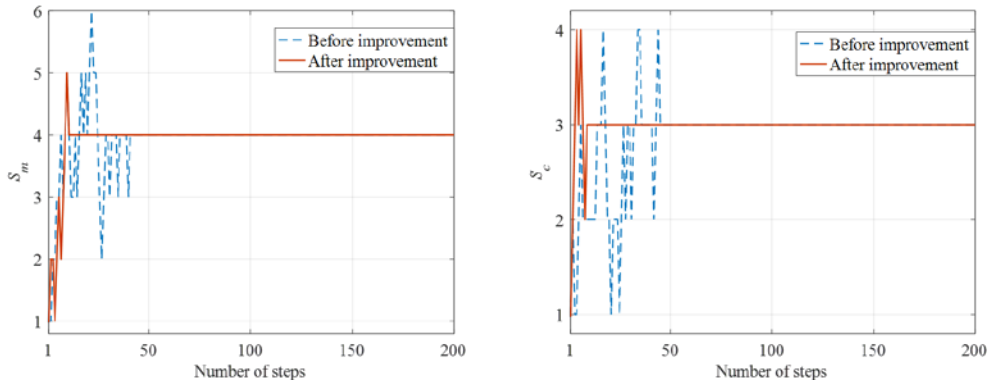

Fig. 13. Performance comparison in the AWGN($E_b / N_0 = 10\text{dB}$) channel

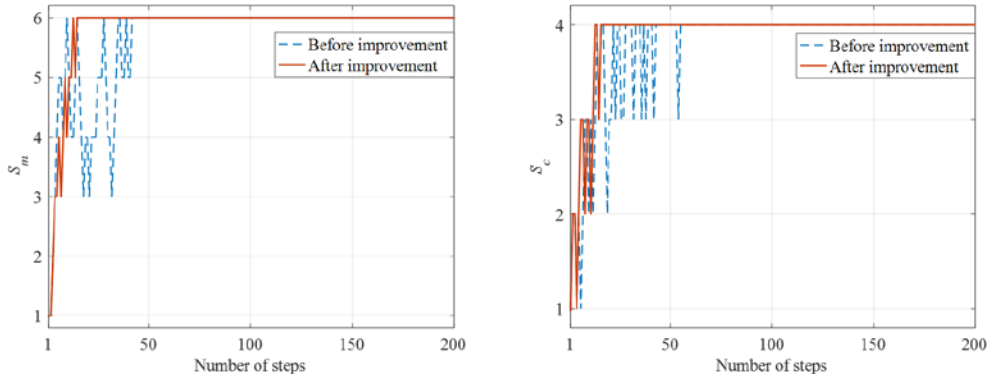

Fig. 14. Performance comparison in the AWGN($E_b / N_0 = 20\text{dB}$) channel

7. Conclusion

In this paper, we have proposed a new RL-based decision-making model for the communication waveform parameters. Simulation results have demonstrated that the configured model has the ability to search for optimal parameters in any channel and is more efficient than the existing RL-based decision-making model. Additionally, we have also compared the decision-making performance of the proposed model and the MCS mapping method in both ideal AWGN channels and fading channels. The proposed model made reasonable decisions in the ideal channels, and performed better in several typical fading channels. Moreover, we introduced an "action judgment mechanism" to optimize the decision-making model. Simulations show that the improved model can reduce the exploration process of online learning by up to 9 times while ensuring the same decision-making performance. There are still some problems need to be solved, such as the adaptability of the model to the actual communication channels. In the future research, we will perform simulation experiments and further improvements on the practical application of the model. Meanwhile, we will further apply machine learning to this system model.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 62071139).

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13-18, Aug. 1999. Article(CrossRef Link)

[2] J. Mitola, "Software radios: Survey, critical evaluation and future directions," *IEEE Aerospace and Electronic Systems Magazine*, vol. 8, no. 4, pp. 25-36, April 1993. Article(CrossRef Link)

[3] L. Zhang, M. Zhao, C. Tan, G. Li and C. Lv, "Research on Spectrum Sensing System Based on Composite Neural Network," in *Proc. of 2020 2nd International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, pp. 22-26, 2020. Article(CrossRef Link)

[4] A. Ikami, T. Hayashi and Y. Amano, "Dynamic Channel Allocation Algorithm for Spectrum Sharing between Different Radio Systems," in *Proc. of 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1-6, 2020. Article(CrossRef Link)

[5] Y. Yang, Q. Zhang, Y. Wang, T. Emoto, M. Akutagawa and S. Konaka, "Multi-strategy dynamic spectrum access in cognitive radio networks: Modeling, analysis and optimization," *China Communications*, vol. 16, no. 3, pp. 103-121, March 2019. Article(CrossRef Link)

[6] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201-220, Feb. 2005. Article(CrossRefLink)

[7] B Li, C Ju, H Yang and G Liu, "Adaptive coded modulation based on LDPC codes," in *Proc. of 2015 10th International Conference on Communications and Networking in China(China Com)*, pp. 648-651, 2015. Article(CrossRef Link).

[8] T. J. O'Shea, T. Roy and N. West, "Approximating the Void: Learning Stochastic Channel Models from Observation with Variational Generative Adversarial Networks," in *Proc. of 2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 681-686, 2019. Article(CrossRef Link)

[9] A. S. Margulies and J. Mitola, "Software defined radios: a technical challenge and a migration strategy," in *Proc. of 1998 IEEE 5th International Symposium on Spread Spectrum Techniques and Applications - Proceedings. Spread Technology to Africa (Cat. No.98TH8333)*, vol.2, pp. 551-556, 1998. Article(CrossRef Link)

[10] W. Jouini, C. Moy, and J. Palicot, "Decision making for cognitive radio equipment: analysis of the first 10 years of exploration," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 26, 2012. Article(CrossRef Link)

[11] C. J. Rieser, "Biologically inspired cognitive radio engine model utilizing distributed genetic algorithms for secure and robust wireless communications and networking," Ph.D. dissertation, Virginia Tech, 2004. Article(CrossRef Link)

[12] P. M. Pradhan and G. Panda, "Comparative performance analysis of evolutionary algorithm based parameter optimization in cognitive radio engine: A survey," *Ad Hoc Networks*, vol. 17, pp. 129–146, 2014. Article(CrossRef Link)

[13] Passino K M., "Bacterial foraging optimization," *International Journal of Swarm Intelligence Research (IJSIR)*, 1(1), 1-16, 2010. Article(CrossRef Link)

[14] Xu. H and Zhou. Z, "Cognitive radio decision engine using hybrid binary particle swarm optimization," in *Proc. of 2013 13th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 143-147, 2013. Article(CrossRef Link)

[15] Xu H, Zhou Z, "Hill-climbing genetic algorithm optimization in cognitive radio decision engine," in *Proc. of 15th IEEE International Conference on Communication Technology*, pp. 115-119, 2013.

[16] A. Amanna, D. Ali, D. G. Fitch and J. H. Reed, "Design of experiments based empirical models to support cognitive radio decision making," in *Proc. of 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1-7, 2015. Article(CrossRef Link)

[17] Y. Yang, H. Jiang, C. Liu and Z. Lan, "Research on Cognitive Radio Engine Based on Genetic Algorithm and Radial Basis Function Neural Network," in *Proc. of 2012 Spring Congress on Engineering and Technology*, pp. 1-5, 2012. Article(CrossRef Link)

[18] R. Daniels and R. W. Heath, "Online adaptive modulation and coding with support vector machines," in *Proc. of 2010 European Wireless Conference (EW)*, pp. 718-724, 2010. Article(CrossRef Link)

[19] M. Bkassiny, Y. Li and S. K. Jayaweera, "A Survey on Machine-Learning Techniques in Cognitive Radios," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1136-1159, Third Quarter 2013. Article(CrossRef Link)

[20] X Dong, Y Li, C Wu and Y Cai, "A learner based on neural network for cognitive radio," in *Proc. of 2010 IEEE 12th International Conference on Communication Technology*, pp. 893-896, 2010. Article(CrossRef Link)

[21] Z. Dou, Y. Dong, and C. Li, “Intelligent decision modeling for communication parameter selection via back propagation neural network,” in *Proc. of International Conference on Advanced Hybrid Information Processing*, Springer, pp. 465–472, 2017. Article(CrossRef Link)
[22] L. Gavrilovska, V. Atanasovski, I. Macaluso and L. A. DaSilva, "Learning and Reasoning in Cognitive Radio Networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1761-1777, Fourth Quarter 2013. Article(CrossRef Link)
[23] N. Abbas, Y. Nasser, and K. E. Ahmad, “Recent advances on artificial intelligence and learning techniques in cognitive radio networks,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 174, 2015. Article(CrossRef Link)
[24] S. Tubachi, M. Venkatesan and A. V. Kulkarni, "Predictive learning model in cognitive radio using reinforcement learning," in *Proc. of 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pp. 564-567, 2017. Article(CrossRef Link)
[25] S. Wang, H. Liu, P. H. Gomes and B. Krishnamachari, "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 257-265, June 2018. Article(CrossRef Link)
[26] R. Bruno, A. Masaracchia and A. Passarella, "Robust Adaptive Modulation and Coding (AMC) Selection in LTE Systems Using Reinforcement Learning," in *Proc. of 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pp. 1-6, 2014. Article(CrossRef Link)
[27] J. P. Leite, P. H. P. de Carvalho and R. D. Vieira, "A flexible framework based on reinforcement learning for adaptive modulation and coding in OFDM wireless systems," in *Proc. of 2012 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 809-814, 2012. Article(CrossRef Link)
[28] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054-1054, Sept. 1998. Article(CrossRef Link)
[29] P. Ferreira, R. Paffenroth, A. Wyglinski, T. M. Hackett, S. Bilén, R. Rein- hart, and D. Mortensen, “Multi-objective reinforcement learning for cognitive radio-based satellite communications,” in *Proc. of 34th AIAA International Communications Satellite Systems Conference*, p.5726, 2016.Article(CrossRef Link)
[30] S. G. Bilen, R. C. Reinhart, T. M. Hackett, R. Paffenroth, P. V. R. Ferreria, D.J. Mortensen, and A. M. Wyglinski, “Multi-objective reinforcement learning-based deep neural networks for cognitive space communications,” in *Proc. of Cognitive Communications for Aerospace Applications Workshop*, pp. 1–8, 2017.Article(CrossRef Link)
[31] C Yu, X Wen, X Lin and Z Wei, "Research on the modulation and coding scheme in LTE TDD wireless network," in *Proc. of 2009 International Conference on Industrial Mechatronics and Automation*, pp. 468-471, 2009. Article(CrossRef Link)

Xia Xie was born in ChenZhou, China, in 1998. She received her B.S. degree in electronic information engineering from Harbin Engineering University, Harbin, China, in2019. She is currently pursuing the M.S. degree in information and communication engineering at Harbin Engineering University, Harbin, China. Her research interests are in the areas of cognitive radio, intelligent radio and decision-making for communication parameters.

Zheng Dou was born in Harbin, China, in 1978. He received the B.E. degree in wireless technology and the M.E. and Ph.D. degrees from Harbin Engineering University, Harbin, China, in 2001, 2004, and 2007, respectively. He has been with Harbin Engineering University, since 2007, where he is currently a Professor. His current research interests include cognitive radio networks, intelligent communication systems, intelligent communication, and electronic jamming communication.

Yabin Zhang was born in Harbin, China, in 1978. She received the B.E. degree in wireless technology and the M.E. and Ph.D. degrees from Harbin, China, in 2001, 2004, and 2009, respectively. She has been with Harbin Engineering University, since 2004, where she is currently an Association Professor. Her current research interests include wireless interference channel and wireless waveform detection.