Journal of Computers Vol. 34 No. 5, October 2023, pp. 251-264
doi: 10.53106/199115992023103405020

The Study on Cognitive Radio Spectrum Allocation Based on Tabu-Q Learning

Xiquan Zhang[1*], Jianwu Dang[1], Shuxu Zhao[1], Shuyang Li[2]

[1] School of Electronic information and Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

1139507130@qq.com

[2] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China

Received 29 September 2022; Revised 19 January 2023; Accepted 13 March 2023

Abstract. With the increasing number of mobile communication devices, the problem of insufficient spectrum resources has emerged. In addition, traditional spectrum allocation model also exacerbates the problem by under utilizing idle spectrum. We propose a dynamic spectrum scheme based on Tabu-Q learning. Firstly, the spectrum allocation problem is formulated as a continuous Markov decision process (MDP), premised on the power constraints of primary users (PUs) and secondary users (SUs) are met. Tabu-Q learning is applied to adjust optimization strategy, so as to maximize the total transmission rate of users. Secondly, the idea of cooperative learning is added in the scheme to improve the convergence speed of algorithm. That is, new users are allowed to learn the experience of old users, so as to improve the speed of spectrum sensing and save the execution time of algorithm. Finally, The mean opinion score (MOS) is used to measure different traffic. The simulation shows that when the number of users is consistent, Tabu-Q learning can improve the transmission bit rate by about 13% compared with Q learning, and keep MOS above the acceptable level (MOS>3). In summary, the scheme proposed in this paper can effectively improve the utilization of idle spectrum.

Keywords: cognitive radio, spectrum allocation, q-learning, MOS, tabu search

1 Introduction

In recent years, with the rapid development of 5G technology, the Internet is accelerating its evolution to the Internet of Things (IoT) [1]. IoT is based on the concept of the Internet of Everything, and is called by scientists as the third wave of the information industry after computers and the Internet [2]. According to Ericsson's report, by 2025, advanced technology will realize a complete sensory Internet, and it is expected to realize the ability to communicate ideas digitally by 2030 [3]. However, advanced technologies require huge spectrum resources to support, and the existing available spectrum cannot meet the high-speed and low-latency requirements of mobile devices for bandwidth. In addition, according to the investigation of the US Federal Communications Commission, the spectrum utilization rate of many devices that have allocated resources is less than 25% [4], which means that the limited spectrum resources are not fully utilized.

In response to the above problems, the concept of cognitive radio technology [5] was proposed. Users can detect the communication environment to know the network status of the current scene [6]. The spectrum sensing process is shown in Fig. 1. When the licensed frequency band of PU is idle, cognitive radio technology can assist the SU to use frequency band indirectly. And this process will not affect the normal use of PU. Therefore, cognitive radio can effectively improve the utilization of idle spectrum.

In cognitive radio systems, spectrum allocation is an important link in the cycle of cognitive networks. Since reinforcement learning can perform offline learning through the time difference method and reduce the use of data, it is often used to solve the spectrum allocation problem. Applying reinforcement learning to the spectrum allocation process can enable end users to independently complete environmental awareness and learn channel selection strategies [7]. To achieve the purpose of improving the performance of cognitive network system. As a classic algorithm in reinforcement learning, Q-learning is widely used to solve spectrum allocation problems [8]. However, adopting Q-learning alone leads to excessive computational complexity. The reason for this phe-

* Corresponding Author

nomenon is that when Q table is updated, the criterion for selecting an action is to obtain best reward value at the next moment. Therefore, repeatedly performing actions increases the complexity of algorithm. This paper introduces tabu search algorithm to optimize malpractice. The feature of tabu search is to use the flexible "memory" function to mark learned actions as tabu objects, and reduce the selection of unnecessary actions to reduce the complexity of algorithm [9]. So far, this paper proposes a Tabu-Q learning algorithm to optimize the spectrum allocation problem.

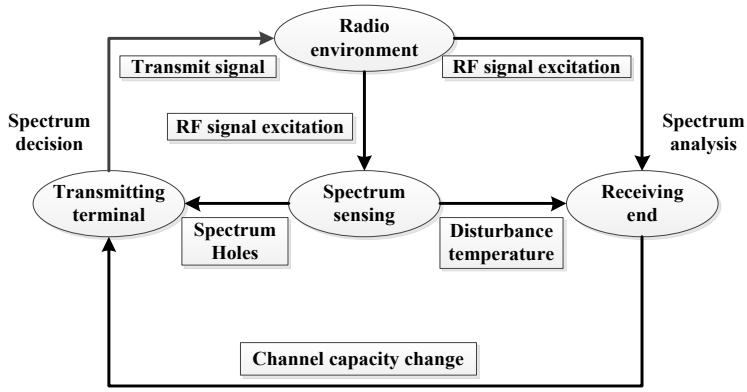

Fig. 1. The process of spectrum sensing

In addition, since user experience is an indispensable link in the development of today's communication systems, the quality of experience (QoE) of end users is becoming a persuasive indicator. QoE can be understood as a user-oriented evaluation standard for services, which puts end users at the center of decision-making. Commonly used indicators for evaluating user experience quality are MOS, standard deviation of opinion score and net promoter score [10]. Among the above indicators, MOS is the most widely used QoE indicator, which is mostly used to evaluate traffic management and resource allocation strategies [11]. In the research field of this paper, some existing works only consider the single traffic demand of users. In order to study the influence of different traffic on cooperative learning between users, this paper uses the characteristics that MOS can be measured in two dimensions of video and data. Using MOS to centrally measure different traffic (data or video). Therefore, this paper will use MOS as a measure for the cooperative learning scheme based on Tabu-Q learning.

The main contributions of this paper are as follows:

In order to improve the utilization of idle spectrum, this paper proposes a dynamic spectrum access scheme based on Tabu-Q learning. The scheme is divided into two parts:

1) In cognitive radio systems, the spectrum allocation problem is usually formulated as a continuous MDP. When Q-learning is used alone to solve the MDP problem, the calculation complexity of the algorithm will be too high. This is because when updating Q-table, the basis for selecting an action is always to target best reward value at the next moment. As a result, repeated actions are performed continuously, which increases the complexity of the algorithm. Therefore, we propose Tabu-Q learning algorithm based on the idea of tabu. Tabu-Q learning marks the learned actions as tabu objects, and then stores them in the Tabu-Q table. Actions that have already been learned are no longer executed, which significantly reduces computational complexity. Finally, cognitive user finds the optimal channel access by searching Tabu-Q table.

2) In order to improve the convergence speed of algorithm, the idea of cooperative learning is added to scheme. When user performs Tabu-Q learning to complete the spectrum sensing, and store the experience value behind Tabu-Q table. Newly joined users are allowed to learn from the experience of old users, which can improve user perception speed and save algorithm execution time. In addition, this paper also studies the exchange of experience between users carrying different traffic. MOS is used to uniformly measure different traffic (data or video), and a comprehensive MOS formula is derived. Simulation results show that the dynamic spectrum access scheme based on Tabu-Q learning is effective.

The rest of the paper is organized as follows; In Section 2, we describe related work on spectrum allocation; In Section 3, we introduce communication model and MOS model; In Section 4, we detail the main contribution of this paper - the cooperative learning scheme based on Tabu-Q learning; Section 5 presents the simulation results of our method; While section 6 presents conclusions and future directions.

2 Related Work

With the deepening of theoretical research, scholars have proposed a variety of solutions for the spectrum allocation problem of cognitive radio networks. Such as: swarm intelligence algorithm, game theory and reinforcement learning. Chen et al. [12] under the premise of ensuring that the PUs and SUs do not exceed the constraints, a hybrid algorithm combining genetics and particle swarm optimization is proposed, which effectively reduces the user's transmit power. Huang J et al. used the auction mechanism to build a channel allocation model, and selected the current optimal channel through price bidding. The pricing scheme proposed by this model effectively improved PU's income [13]. The above algorithms have effectively improved the efficiency of spectrum allocation. However, the premise of the above implementation is that perfect channel state information in the network needs to be obtained, which is difficult to achieve in an actual network.

Since reinforcement learning uses relatively little data in the network, it is more suitable for application in dynamically changing resource scheduling problems. Therefore, the following describes the use of reinforcement learning algorithms to solve spectrum allocation problem. Ding et al. proposed a power control channel switching strategy using reinforcement learning. SUs learns the current optimal channel switching scheme through reinforcement learning, and selectively switches channels according to the channel status of PU, effectively reducing the channel switching time. ground power consumption [14]. To solve the problem of long channel state sensing time, Cao et al. used reinforcement learning and Bayes algorithm to predict the time of channel idle state, which effectively reduced the power consumption for channel sensing detection [15]. As one of the classical reinforcement learning, Q-learning is widely used in spectrum allocation problems. Chen et al. extended single Q-learning to multi-learner Q-learning, and proposed a guess-based multi-learner Q-learning algorithm [16]. Under certain known constraints, the algorithm can achieve fast convergence. In the literature [17], the author combined Q-learning and SARSA algorithm and proposed a resource allocation scheme for multi-agent reinforcement learning. This scheme can effectively reduce the problem of base station aggregation interference in cognitive networks. The above literature reduces transmission power consumption by using improved reinforcement learning, but does not take into account the end user's demand for transmission quality.

Based on this, consider the quality of experience for end user. A resource allocation architecture with automatic quality of experience assessment is proposed in Dudin et al. [18], which builds on advances in affective computing and sensing to efficiently utilize scarce spectrum resources. Although the above algorithm has achieved relatively good results in the spectrum allocation problem. However, with the increasing shortage of spectrum and the continuous improvement of end users' requirements for delay, more factors need to be considered in spectrum allocation. For example: how to strengthen cooperation between end users.

3 System Model

In this section, a spectrum sharing communication scenario is introduced, consisting of a primary network (PN) and a secondary network (SN), as shown in Fig. 2. It is assumed that the PN shares a spectrum with the SN at the same time, and the PN consists of primary base stations (PBS) and PUs, the SN includes secondary base stations (SBS) and SUs. Considering the interference problem between SUs and PUs, it is assumed that SUs can access the idle spectrum under the condition that the Signal to Noise Ratio (SINR) of PN is not lower than a pre-designed threshold. Since PUs has been authorized to access the channel, in order not to interfere with PUs, SUs accessing the network needs to adjust the transmission parameters dynamically. Therefore, both the primary link and the secondary link in the network adopt the Adaptive Modulation and Coding (AMC) scheme to transmit information [19].

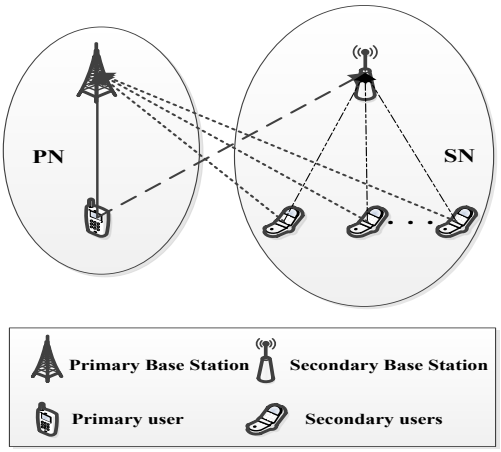

Fig. 2. Spectrum allocation system model

The secondary channel is assumed to be a quasi-static channel with white Gaussian noise. The primary channel adopts the scheme of AMC technology, and transmission power is a fixed value. Users needs to infer the channel state information through active learning while estimating the channel gain [20]. The SINR of PBS and SBS are represented as $SINR^{(P)}$ and $SINR_i^{(s)}$, respectively, which can be written as:

$$SINR^{(P)} = \frac{G_0^{(p)} P_0}{\sigma^2 + \sum_{j=1}^{N} G_j^{(p)} P_j} \quad . \tag{1}$$

$$SINR_i^{(s)} = \frac{G_i^{(s)} P_i}{\sigma^2 + G_0^{(s)} P_0 + \sum_{j \neq i} G_j^{(s)} P_j} \quad . \tag{2}$$

P_0 is the transmit power of PUs, and P_j is the transmit power of SUs. $G_0^{(P)}$ is the channel gain between PUs and PBS, $G_j^{(P)}$ is the channel gain between SUs and PBS, $G_j^{(s)}$ is the channel gain between SUs and SBS, and σ^2 is the noise power.

3.1 MOS Model

In the communication network reference index, the quality of service (QoS) is often used to quantitatively describe the performance of communication network. Such as: delay, packet loss rate, throughput, bit error rate, etc. With the trend of future IoT demands tending to be user-centric, the frequency of using QoE as an indicator for evaluating user subjective satisfaction is increasing. The European Telecommunication Standards Institute defines QoE as "the objective and subjective performance measurement of users when using information and communication technology services or products" [21]. In addition, in order to facilitate the optimization and management of QoE, the International Telecommunication Union adopts the MOS when users score the service subjectively and experimentally as the evaluation standard of QoE [22]. In this section, two different types of traffic, data and video, are represented by MOS, and MOS formula for comprehensive data and video is given. The two MOS models are described below.

MOS Model of Data: It evaluates the transmission rate parameters in the network. Referring to the work done in [23-24], the MOS for data traffic is calculated as follows.

$$Q_{date} = a \log_{10}(br_i^{(s)}(1 - p_{e2e})) \quad . \tag{3}$$

Among them, Q_{date} represents the MOS of data traffic, P_{e2e} and $r_i^{(s)}$ represent the end-to-end data packet loss

rate and data transmission bit rate, respectively. Parameters *a* and *b* are constants within a specific interval, which can be set using the maximum and minimum perceived data quality of the end user [25]. The data packet loss rate is defined as the difference between user's transmission rate and effective reception rate. If the packet loss rate is zero, it means that end users experience the best MOS quality. At this time, the MOS value is an ideal value of 5, and vice versa. The relationship between user subjective experience quality and MOS value is shown in Table 1.

Table 1. Relationship between MOS value and user experience

MOS	Quality of experience	Perceived quality impairent
5	very good	barely feel
4	good	Injured but not angry
3	generally	slightly annoyed
2	not very good	angry
1	very bad	very annoyed

MOS Model of Video: It is the user's analysis of subjective feelings such as video freezes. After research in [26], it was found that the relationship between MOS and objective measurement of image distortion is not a typical linear relationship. In a further study by Hanhart et al. [27], the peak signal-to-noise ratio (PSNR) can be used to measure image distortion. PSNR is an important indicator for video quality evaluation, which can objectively reflect the performance of video coding. The following is the relationship between MOS and PSNR:

$$Q_{video} = \frac{c}{1+\exp(d(PSNR-\mathrm{g}))} \quad . \tag{4}$$

Where Q_{video} represents the MOS value of the video, *c*, *d* and *g* are the parameters of the logic function. Meanwhile, in the [28], the average value of different resolutions is taken for multiple MPEG-4 encoded videos, then the relationship between PSNR and transmission rate can be expressed as:

$$PSNR = i\log r_i^{(s)} + p \quad . \tag{5}$$

In the above formula, $r_i^{(s)}$ is the bit rate of video transmission, *i* and *p* are constants.

Comprehensive MOS Model: As a quality evaluation indicator, MOS can allow different types of data traffic to be integrated. We integrate the end user's data traffic and real-time video requirements into a model, which is approximated by mathematical formulas as:

$$\frac{1}{N}(\sum_{i=1}^{U} Q_{data} + \sum_{i=U+1}^{U} Q_{video}) \quad . \tag{6}$$

Among them, *U* is the amount of data traffic transmitted by SUs, and $N - U$ is the amount of real-time video transmitted by remaining users.

3.2 Problem Description

Users in PN and SN need to consider the SINR threshold to determine their own actions, in order to meet the goals of dynamic spectrum access and QoE. The SINR of PUs and SUs are limited as follows:

$$\begin{aligned} &Find: \{P_0, P_i, \forall i\} \\ &s.t: \begin{cases} SINR^{(p)} \geq \beta_0 \\ SINR_i^{(s)} \geq \beta_i, i = 1, \cdots, N \end{cases} \end{aligned} \quad . \tag{7}$$

β_0 and β_1 represent the SINR thresholds of PUs and the *i-th* SUs, respectively. Assuming that SINR constraints satisfy the equality condition, SINR constraints for PUs and SUs can be expressed as:

$$\begin{aligned} &Find: \{P_0, P_i, \forall i\} \\ &s.t: \begin{cases} SINR^{(p)} = \beta_0 \\ SINR_i^{(s)} = \beta_i, i = 1, \cdots, N \end{cases} \end{aligned} . \tag{8}$$

SUs can obtain corresponding $r_i^{(s)}$ by adjusting β_i. Meanwhile, the transmission power of SUs can be expressed as [29]:

$$P_i = \frac{\Psi_i(\sigma^2 + G_0^{(s)} P_0)}{G_i^{(s)}(1 - \sum_{j=1}^{N} \Psi_j)}, i = 1, \ldots, N \ . \tag{9}$$

Where $\Psi_i = (1 + \frac{1}{\beta_i})^{-1}$, at the same time, to obtain effective power allocation, the above formula needs to meet the following conditions:

$$1 - \sum_{j=1}^{N} \Psi_j > 0 \ . \tag{10}$$

After substituting formula (9) into formula (7) to replace the power of SUs, the SINR constraint in formula (7) can be rewritten as:

$$\sum_{j=1}^{N} \alpha_j \Psi_j \le 1 \ , \tag{11}$$

$$\alpha_j = \frac{G_j^{(p)}(\sigma^2 + G_0^{(s)} P_0)}{G_j^{(s)}(G_0^{(p)} P_0 / \beta_0 - \sigma^2)} + 1 \ . \tag{12}$$

Because β_0 is assumed to be constant, β_i needs to be adjusted on each SU. According to the system settings in [30], the relationship between transmission bit rate and SUs threshold SINR is written as:

$$r_i^{(s)} = W \log_2 M(\beta_i) \ . \tag{13}$$

$M(\beta_i) = (1 + k\beta_i)$ represents the number of bits per modulation symbol. $k = \frac{1.5}{-\ln(5BER)}$ generally takes a constant, which is determined by the tolerance of the maximum transmission error rate.

4 Dynamic Spectrum Allocation Based on Tabu-Q Learning

In cognitive radio networks, SUs must ensure that the interference level to authorized users is below a threshold while ensuring opportunistic access to the spectrum. The dynamic spectrum allocation problem can be viewed as a continuous Markov decision process, after which reinforcement learning is used to achieve the goal of maximizing the expected benefit of SUs.

4.1 Allocation Strategy for Reinforcement Learning

Reinforcement learning methods are composed of a set of states, actions, and reward functions. The principle of the algorithm is that agent selects an action according to the state of existing experience, and obtains corresponding reward from environment based on the effectiveness of decision. The policy of updating SUs has a large impact due to the reward patterns of different state-action combinations. In this paper, Q-learning [31],

a classical algorithm in reinforcement learning, is chosen to solve the problem of spectrum resource allocation. In Q-learning, the agent selects a set of strategies from the existing state according to the current environment, and then stores the return value in the Q-table. Each SU chooses an optimal strategy to maximize the MOS under the premise of ensuring that Equation (7) is satisfied. The interference between SUs is $S_t = (I_t, L_t)$, I_t and L_t respectively defined as follows:

$$I_t = \begin{cases} 0, if \sum_{i=1}^{N} \Psi_i(\beta_t^{(i)}) < 1 \\ 1, otherwise, \end{cases} , \tag{14}$$

$$L_t = \begin{cases} 0, if \sum_{i=1}^{N} \alpha_i \Psi_i(\beta_t^{(i)}) \le 1 \\ 1, otherwise. \end{cases} . \tag{15}$$

The reward function is defined as following function:

$$R_t^{(i)}(a_t, s_t) = \begin{cases} N, if I_{t+1} + L_{t+1} > 0 \\ Q^{(i)}{}_{MOS}, I_{t+1} + L_{t+1} = 0 \end{cases} . \tag{16}$$

Where A is a constant smaller than the reward of any other policy. When the strategy is $I_{t+1} + L_{t+1} > 0$, the interference condition formula (11) is violated, indicating that the adopted strategy is unsuccessful. It is assumed that SUs cannot predict the impact of each other's actions on the overall environment, and the states of SUs are considered as part of the surrounding environment. SUs make actions according to the revenue cycle, and finally obtain the optimal policy. The maximum reward can be expressed as:

$$V_i(s, \pi) = \sum_{t=0}^{\infty} \Upsilon^t E(R_t^{(i)} \mid \pi, s_0 = s), i = 1, 2, \ldots, N . \tag{17}$$

Where π represents the local policy in the learning process, and S_0 is the initial state. According to the principle of optimality in [32], it can be assumed that the strategies adopted are all optimal. After taking the best action, the above formula can be further expressed as:

$$V_i^*(s, \pi^*) = \max_a [R(s, a) + \Upsilon \sum_{s'} p(s' \mid s, a) V_i(s', \pi^*)] . \tag{18}$$

4.2 Tabu-Q Learning Algorithm

Combined with the above analysis, SUs uses Q-learning algorithm to complete the perception of environment, make corresponding actions and finally get rewards. After the above-mentioned one process is completed, the corresponding value is stored in the Q table. The Q table will contain the optimal power allocation information under the current environmental state, after continuous looping and learning. However, the computational complexity of using Q-learning alone is high, and we opted for the Tabu search algorithm for optimization [33]. Marking the learned actions as Tabu objects can avoid the selection of repetitive actions, thus reducing the algorithm complexity. Meanwhile, SBS provides user with the optimal power allocation factor by looking up Tabu-Q table, thereby improving the transmission rate of SUs. The formula for the Q-learning algorithm is as follows:

$$Q(s, a) \leftarrow Q(s, a) + a[r + \gamma \max_{a' \in A} Q(s', a') - Q(s, a)] . \tag{19}$$

Where $Q(s, a)$ represents the estimated utility of learner choosing action a in state$s$. The above formula is updated to Tabu-Q table by combining Tabu table and Q table:

$$Q(s, a) = p(s_{t+1} \mid s_t, a_t) \times \{(1 - \beta) Q(s_{t-1}, a_{t-1}) + \beta[r_t + \gamma \max Q(S_{t+1}, a_{t+1})]\} , \tag{20}$$

$$p(s_{t+1} \mid s_t, a_t) = \sum_{s_{t+1} \in s} P_s(s_{t+1} \mid s_t, a_t) . \tag{21}$$

Where $p(s_{t+1}|\ s_t,\ a_t)$ is the state transition probability density function, which represents the transition probability that state changes from s_t to s_{t+1} after learner takes an action. $\beta \in [0,1]$ represents the learning rate of algorithm update in each time step, and $\gamma \in [0,1]$ is the discount factor. In addition, this section approximates $V_i^*(s, \pi^*)$ in formula (18) with Tabu-Q table function, and the update formula is as follows:

$$Q_{t+1}^{\ i}(s, a_t^{\ i}) = \sum_{s_{t+1} \in s} P_s(s_{t+1} \mid s_t, a_t) \times \{(1-\beta) Q_t^{\ i}(s_{t-1}, a_{t-1}^{\ i}) + \beta[R_t^{\ i}(s, a_t^{\ i}) + \gamma Q_t^{\ i*}(s_{t+1}^{\ `})]\} \quad . \tag{22}$$

The Q value of SUs is equal to the maximum value of this stage, then the equation is $Q_t^{i*}(s^`) = \max_b Q_t^i(b, s^`)$. Therefore, the problem of confirming the transmit power and the corresponding bit rate is replaced by finding the optimal transmit power ($\hat{\beta}$). The optimal cognitive network spectrum usage problem is formulated as:

$$\begin{aligned} &\{(\hat{\beta}_i)\} = \arg\max_{\beta_i} \sum_{i=1}^{N} Q_{i(MOS)}(\beta_i) \\ &s.t \begin{cases} \sum_{i=1}^{N} \Psi_i(\beta_i) \le 1 - \varepsilon \\ \sum_{j=1}^{N} \alpha_j \Psi_j(\beta_j) \le 1 \end{cases} \end{aligned} \quad . \tag{23}$$

Mark executed action a_t as a Tabu object, so that it cannot be executed in the following loop, and then update reward value and Tabu object to the Tabu-Q table together. Add video and data to secondary user traffic requests, so as to measure the effect of comprehensive MOS on different traffic evaluations. The pseudocode of the algorithm is shown in Algorithm 1.

Algorithm 1. Tabu-Q learning algorithm

Initialization: Tabu-Q table $Q_0 = 0$ for all the SUs. Determine the maximum of iterations t_{max} and parameters β, γ

```
for time t < t_max do
    for all SU_i, i = 1, ..., N do
    Observe the current state S_t, select the action
        a_t^(i) = arg max a_t^(i) Q(s_t^(i), a_t^(i)) according to the Tabu-Q table
    Determine whether a_t^(i) is a Tabu object, if so, return to the previous line
    If not, continue executing the statement
    Update the state a_{t+1}^(i) by formula (14) and (15)
    Update the reward R_t^(i) by formula (22)
    end for
end for
```

4.3 Cooperative Learning Pattern

As mentioned in the previous section, after completing a Tabu-Q learning, Tabu-Q table should contain reward value and Tabu objects for this time. According to Wang et al. in [34], data in the Tabu-Q table can also reflect the connection between individual SU's own state and overall network system environment. When a SU has learned surrounding environment, the value of network parameters of environment does not change much. If SU has to re-perceive surrounding environment every time, and ignore environmental information obtained by other secondary users in the system, this will lead to low utilization of the environmental information. Therefore, this paper introduces the Docition radio, which imparts environmental information stored in the Tabu-Q table to new users through experienced old users. It reduces the time for new users to learn environment, thereby effectively improving information utilization [35]. It is assumed that the way new user joins network is the low bit rate control channel. Then Q-table combining Docition radio and Tabu search is written as:

$$Q_c = \frac{1}{N} \sum_{i=1}^{N} Q^{(i)} \quad . \tag{24}$$

5 Experimental Results and Analysis

To verify the effectiveness of Tabu-Q learning algorithm in allocating idle spectrum to SUs, we conduct the following experimental simulations. A two-dimensional coordinate system unit model is designed, which can directly reflect the location information of base stations and users, as shown in Fig. 3. The PBS is located at the coordinate origin, and one PU is distributed within a radius of 200m from the origin. The SBS and SUs are respectively distributed within a unit circle with a radius of 500m and 1400m from the origin. In the simulation process, it is assumed that up to 25 SUs can be connected to a PN. Meanwhile, the linear distance between two adjacent secondary users is set to be no more than 500m in the experiment so that study the teaching ability of Docition radio. The simulation parameters used in the experiment are shown in Table 2.

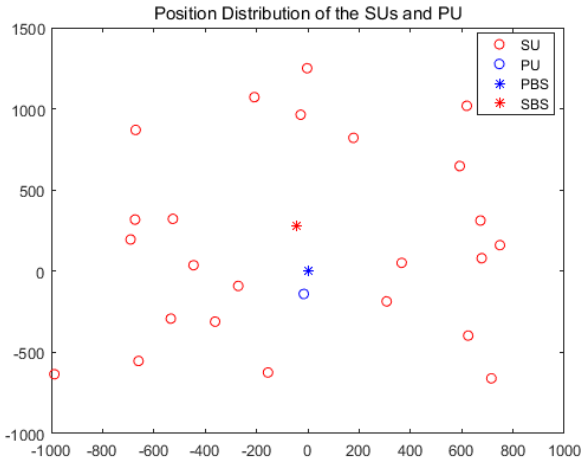

Fig. 3. Location distribution of base stations and users

In order to verify the validity and applicability of Tabu-Q learning algorithm proposed in this paper, the algorithm was compared with Q-learning algorithm and the improved QLPSOES [36] algorithm based on Q-learning. QLPSOES is an algorithm that combines particle swarm optimization with Q-learning. Its technical contribution lies in the fact that particles can share the "behavioral experience" of the optimal particle through Q table, which improves the convergence speed of Q table. MOS measures the different demands (video or data traffic) of newly added users, and judges the impact of delays caused by different data switching on end users. When the average MOS score is greater than 3, it is an acceptable range for end-user quality perception [15]. Average bit rate refers to the number of bits transmitted per second [37]. The higher the bit rate, the more data is transmitted in the network. The simulation results are shown in Fig. 4 and Fig. 5.

Table 2. Simulation parameters for testing algorithm performance

Indicator name	Specific value
PU's bandwidth	10MHz
Target SINR of PU	10dB
Transmit power of PU	10nW
Gaussian noise power of PU	1nW
Logical parameter *c*	6.6431
Logical parameter *d*	-0.1344
Logical parameter *g*	30.4264
Constant *i*	10.4
Constant *p*	-28.7221
Path loss factor	2.8
Learning efficiency of SU	0.1
Discount factor of SU	0.4
SINR set of SU	{-5, -3, -1, 1, 3, 5, 7, 9, 11, 15}dB

Fig. 4 shows the MOS of three different spectrum allocation algorithms. In order to verify the influence of Tabu-Q learning algorithm proposed in this paper on the spectrum allocation process, all three algorithms use the independent learning scheme. That is, newly joined users independently execute Tabu-Q learning algorithm, QLPSOE algorithm and Q-learning algorithm, regardless of the learning experience of other users. As can be seen from Fig. 4, with the number of secondary users in the network gradually increasing, the score of MOS shows a downward trend. The reason is that as the number of new users increases, the interference constraints between users also increase. Each secondary user will converge to a lower SINR value in order to satisfy the constraints in the current network. It can also be seen from Fig. 4 that before the number of users is less than 21, compared with the other two algorithms, the MOS score of Tabu-Q learning algorithm is the largest, which shows that the improved algorithm can effectively guarantee the spectrum allocation efficiency. And when the maximum number of users is set to 25, MOS score is greater than 3.5, which meets the quality requirements of end users for the transmission rate.

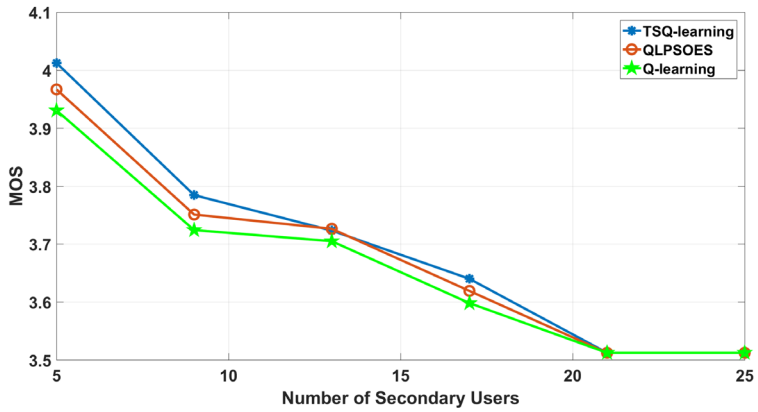

Fig. 4. MOS score curves of different allocation algorithms

Fig. 5 compares the average Bitrate of three different spectrum allocation algorithms. In general, the higher the value of average bit rate, the clearer the quality of transmitted video. It can be seen in Fig. 5 that when the number of users is less than 21, the transmission performance of Tabu-Q learning algorithm has certain advantages. At the same time, based on actual data estimation, in the range of 5-25 secondary users, using Tabu-Q learning can increase the average bit rate of data transmission by about 13% compared with Q learning. This shows the rationality and effectiveness of Tabu-Q learning in algorithm design. On the whole, when the number of secondary users does not exceed the set value, Tabu-Q learning can still maintain relatively stable video transmission quality, which meets the needs of most scenarios.

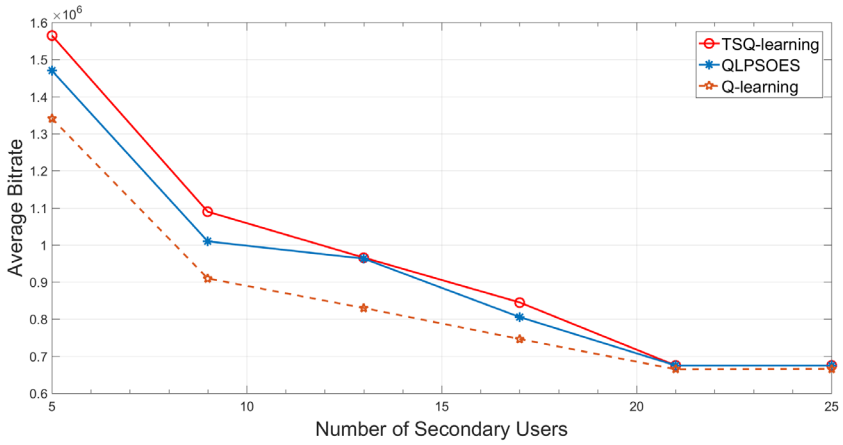

Fig. 5. Average Bitrate Curves for Different Allocation Algorithms

The above experiment is to verify the performance of Tabu-Q learning algorithm alone, and the simulation results show that algorithm has certain advantages in the optimization effect. In order to further study the impact of the entire process of users performing reinforcement learning and then imparting previous experience to new users on spectrum allocation, the following experiments are designed.

Scheme 1 is the cooperative learning scheme. Specifically: old user executes Tabu-Q learning algorithm, and stores the experience in the Tabu-Q table. After the new user joins system, which can learn the experience of old user to select a channel through formula (24). Scheme 2 is the independent study scheme. Specifically: new users independently execute Tabu-Q learning algorithm without considering the learning experience of old users. Scheme 3 and 4 are a further extension of scheme 1, and classify the traffic demand types of new users. Scheme 3 is a collaborative learning scheme for similar users. Specifically: new users only learn the experience selection channel of users with the same traffic type. The fourth scheme is the collaborative learning scheme for dissimilar users. Newly joined users only learn to select channels with experience of users with different traffic types. The simulation results are shown in Fig. 6 to Fig. 8 below.

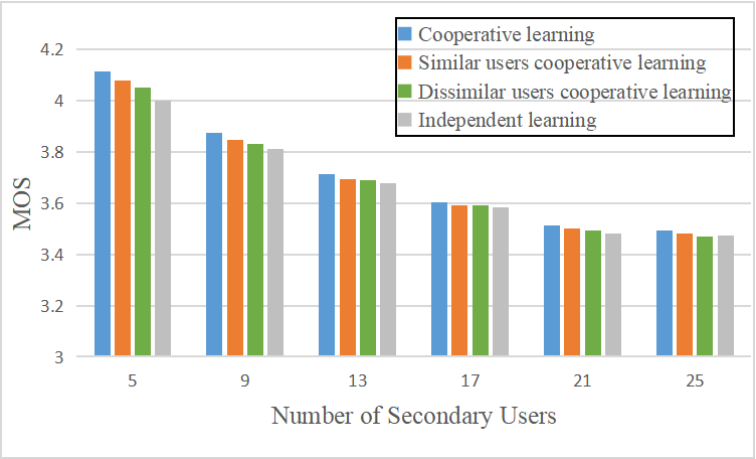

Fig. 6. MOS scores for four spectrum allocation schemes

It can be seen from Fig. 6 that the cooperative learning scheme that integrates all users is the scheme that maintains the highest MOS score. The MOS score of scheme that only learns the user experience of same traffic type is slightly higher than the remaining two allocation schemes, and the scheme of completely independent learning is the scheme with the lowest MOS score. This shows that teaching new users can improve the overall user experience, and the MOS-based dynamic spectrum access scheme is effective.

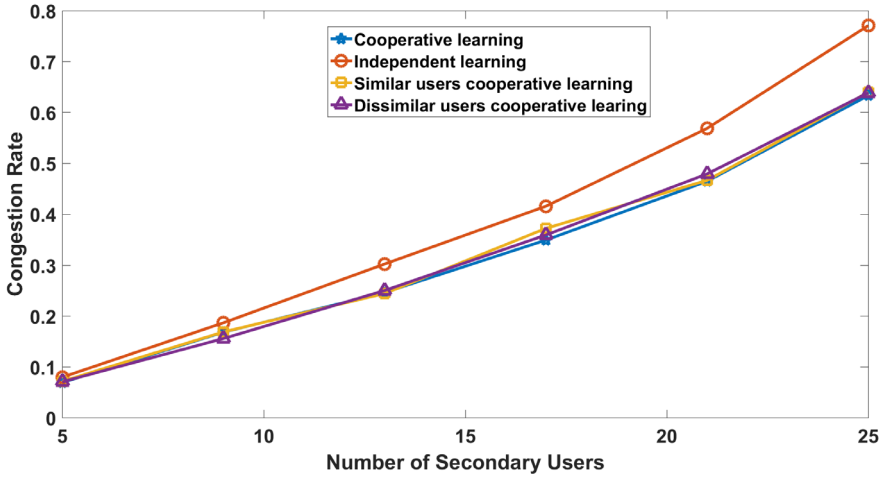

Fig. 7. Congestion rate curves of four spectrum allocation schemes

It can be seen from Fig. 7 that when the number of secondary users continues to increase, the overall network congestion rate increases, which is in line with real network scenarios. When the number of users is higher than 21, the rate of increase in congestion rate is significantly accelerated. This is because as the number of users increases, the interference between users intensifies, which greatly reduces the communication rate of users, resulting in serious congestion in the system. In addition, in the range of 5-25 secondary users, the congestion rate of the same traffic or different traffic schemes based on cooperative learning is significantly lower than that of independent learning schemes in the same scenario. This indicates that the cooperative learning scheme can host more secondary users to access the network.

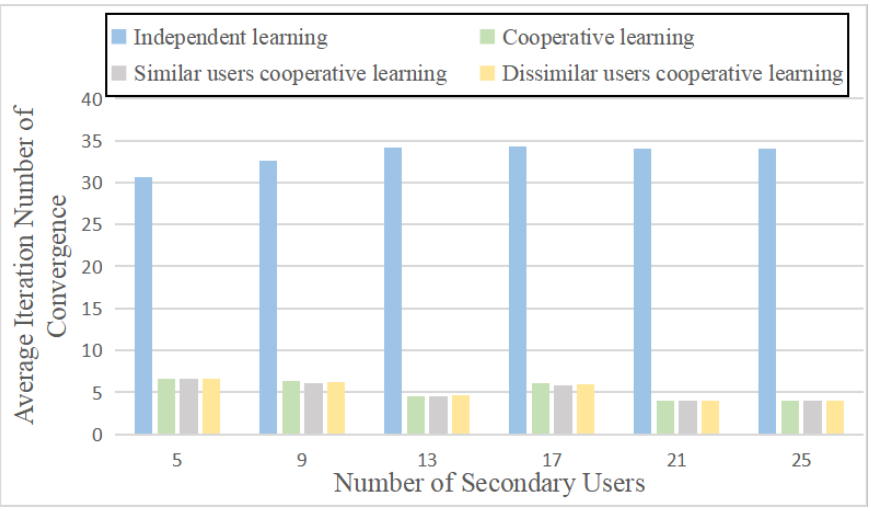

Fig. 8. The average number of iterations of four spectrum allocation schemes

Fig. 8 shows the number of iterations required for the four schemes to achieve convergence. According to statistics, in the range of 5-25 users, compared with the independent learning scheme, the number of iterations required for convergence of the scheme based on cooperative learning is reduced by about 60%. Moreover, the three schemes based on cooperative learning can converge quickly at the initial stage of iteration, and the convergence tends to be stable as the number of users increases. This shows that the cooperative learning scheme can effectively reduce the execution time of algorithm, thereby improving the communication performance of system.

6 Conclusion

In this paper, considering that multiple SUs generate spectrum access requirements, Tabu-Q learning is used to mark repeated actions as Tabu objects, and to speed up SU' perception of idle spectrum, so as to maximize the user's transmission rate. Meanwhile, the MOS is used for unified integration of different types of traffic (data or video), and a comprehensive MOS formula is derived. Finally, the idea of teaching in Doction Radio is introduced into the dynamic spectrum scheme based on Tabu-Q learning, and the spectrum access rate of users is improved through the way that experienced old users teach experience to new users according to Tabu-Q table. The simulation results show that Tabu-Q learning can effectively reduce the rate of decline of the average MOS value, and enhance the perception ability of new users with different needs to the current network environment, so that the spare spectrum resources can be better utilized.

This study also has certain limitations, mainly due to the lack of full consideration of the reward mechanism. In view of the selfishness of nodes, we need to design a more complete incentive mechanism to promote the process of teaching new users for old users with experience. In future research, game theory and neural network algorithms are considered to solve the problem of spectrum allocation, so as to better improve the utilization of spare spectrum. In addition, facing the problem of excessive number of unlicensed band users in the cell, we plan

to build a memetic algorithm framework that integrates Q learning and intelligent optimization algorithms. Its idea is to build a cooperative alliance between user nodes to communicate unoccupied spectrum information to each other, thereby speeding up the rate of spectrum sensing.

7 Acknowledgement

We would like to express our appreciation to the support by the National Natural Science Foundation of China (11461038), and the Supported by Key R&D Projects in Gansu Province (20YF8GA123).

References

[1] E. Borgia, R. Bruno, M. Conti, D. Mascitti, A. Passarella, Mobile edge clouds for information-centric IoT services, in: Proc. IEEE Symposium on Computers and Communication (ISCC), 2016.
[2] M. Ghobaei-Arani, A. Souri, A. Rahmanian, Resource management approaches in fog computing: a comprehensive review, Journal of Grid Computing 18(1)(2020) 1-42.
[3] R. Q. Hu, Y. Qian (Eds.), Heterogeneous Cellular Networks, Wiley Telecom, 2013 (pp. 1-25).
[4] Y. Chen, H.-S. Oh, A survey of measurement-based spectrum occupancy modeling for cognitive radios, IEEE Communications surveys & tutorials 18(1)(2016) 848-859.
[5] W. Zhang, C.-X. Wang, X. Ge, Y. Chen, Enhanced 5G cognitive radio networks based on spectrum sharing and spectrum aggregation, IEEE Transactions on Communications 66(12)(2018) 6304-6316.
[6] M. Usman, D. Har, I. Koo, Energy-efficient infrastructure sensor network for Ad-Hoc cognitive radio network, IEEE Sensors Journal 16(8)(2016) 2775-2787.
[7] Y. Zhou, F. Zhou, Y. Wu, R.Q. Hu, Y. Wang, Subcarrier assignment schemes based on Q-learning in wideband cognitive radio networks, IEEE Transactions on Vehicular Technology 69(1)(2020) 1168-1172.
[8] L. Tong, T. Liang, Y. Zhang, P.-Z. Qian, Dynamic spectrum allocation method based on multi-Agent reinforcement learning, Journal of Terahertz Science and Electronic Information 19(4)(2021) 573-580.
[9] F. Glover, Tabu Search: a tutorial, Interfaces 20(4)(1990) 74-94.
[10] R. Mahmud, S.N. Srirama, K. Ramamohanarao, R. Buyya, Quality of experience (QoE)-aware placement of applications in fog computing environments, Journal of Parallel and Distributed Computing 132(2019) 190-203.
[11] X. Cao, L. Liu, Y. Cheng, X. Shen, Towards energy-efficient wireless networking in the big data era: a survey, IEEE Communications Surveys & Tutorials 20(1)(2018) 303-332.
[12] H. Wang, F. Jiang, M. Zhou, Power allocation of cognitive radio system based on genetic particle swarm optimization, Journal of Jilin University (Engineering and Technology Edition) 49(4)(2019) 1363-1368.
[13] S.-C. Lin, K.-C. Chen, Improving spectrum efficiency via in-Network computations in cognitive radio sensor networks, IEEE Transactions on Wireless Communications 13(3)(2014) 1222-1234.
[14] H. Ding, X. Li, Y. Ma, Y. Fang, Energy-efficient channel switching in cognitive radio networks: a Reinforcement Learning Approach, IEEE Transactions on Vehicular Technology 69(10)(2020) 12359-12362.
[15] H. Cao, H. Tian, J. Cai, A.-S. Alfa, S. Huang, Dynamic load-balancing spectrum decision for heterogeneous services provisioning in multi-channel cognitive radio networks, IEEE Transactions on Wireless Communications 16(9)(2017) 5911-5924.
[16] X. Chen, Z. Zhao, H. Zhang, Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks, IEEE Transactions on Mobile Computing 12(11)(2013) 2155-2166.
[17] A. Kaur, K. Kumar, Energy-efficient resource allocation in cognitive radio networks under cooperative multi-agent model-free reinforcement learning schemes, IEEE Transactions on Network and Service Management 17(3)(2020) 1337-1348.
[18] B. Dudin, N.-A. Ali, A. Radwan, A.-E. M. Taha, Resource allocation with automated QoE assessment in 5G/B5G wireless systems, IEEE Network 33(4)(2019) 76-81.
[19] E. Ekrem, S. Ulukus, Secrecy in cooperative relay broadcast channels, IEEE Transactions on Information Theory (57)(1)(2011) 137-155.
[20] R. Zhang, On active learning and supervised transmission of spectrum sharing based cognitive radios by exploiting hidden primary radio feedback, IEEE Transactions on Communications 53(10)(2010) 2960-2970.
[21] K.-U. R. Laghari, On quality of experience (QoE) for multimedia services in communication ecosystem, [dissertation] France: Institut National des Télécommunications, Télécom SudParis, 2012.
[22] S. Pezzulli, M.-G. Martini, N. Barman, Estimation of quality scores from subjective tests-beyond subjects' MOS, IEEE Transactions on Multimedia (23)(2021) 2505-2519.
[23] F. Shah-Mohammadi, A. Kwasinski, Deep reinforcement learning approach to QoE-driven resource allocation for spectrum underlay in cognitive radio networks, in: Proc. IEEE International Conference on Communications Workshops

(ICC Workshops), 2018.
[24] F.-S. Mohammadi, A. Kwasinski, QoE-Driven integrated heterogeneous traffic resource allocation based on cooperative learning for 5G cognitive radio networks, in: Proc. IEEE 5G World Forum (5GWF), 2018.
[25] M. Nasimi, F. Hashim, A. Sali, R.K.Z. Sahbudin, QoE-driven cross-layer downlink scheduling for heterogeneous traffics over 4G networks, Wireless Personal Communications (96)(3)(2017) 4755-4780.
[26] S. Shi, X. Zhang, S. Wang, R. Xiong, S. Ma, Study on subjective quality assessment of Screen Content Images, in: Proc. Picture Coding Symposium (PCS), 2015.
[27] A. Ahar, M. Pereira, T. Birnbaum, A. Pinheiro, P. Schelkens, Validation of dynamic subjective quality assessment methodology for holographic coding solutions, in: Proc. 13th International Conference on Quality of Multimedia Experience (QoMEX), 2021.
[28] N. Gera, D.-D. Fatta, Determinants of consumer's buying behaviour for digital products in trade fair, International Journal of Business Excellence 22(4)(2020) 542-563.
[29] X. Mi, L. Xiao, M. Zhao, X. Xu, J. Wang, Statistical QoS-driven resource allocation and source adaptation for D2D communications underlaying OFDMA-Based cellular networks, IEEE Access (5)(2017) 3981-3999.
[30] K.-M. Humadi, W.-P. Zhu, W. Ajib, Performance analysis of adaptive modulation for millimeter wave cellular systems, in: Proc. IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020.
[31] C. Watkins, P. Dayan, Technical note: Q-Learning, Machine Learning (8)(3-4)(1992) 279-292.
[32] Y. Huang, W. Lu, Online parallel optimization approach to courier routing problems, in: Proc. IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 2020.
[33] A. Chaghari, M.-R. Feizi-Derakhshi, M.-A. Balafar, Fuzzy clustering based on Forest optimization algorithm, Journal of King Saud university-computer and information sciences 30(1)(2018) 25-32.
[34] A. Ali, S. Tariq, M. Iqbal, L. Feng, I. Raza, M.H. Siddiqi, A.K. Bashir, Adaptive bitrate video transmission over cognitive radio networks using cross layer routing approach, IEEE Transactions on Cognitive Communications and Networking 6(3)(2020) 935-945.
[35] Q. Zhao, D. Grace, T. Clarke, Transfer learning and cooperation management: balancing the quality of service and information exchange overhead in cognitive radio networks, Transactions on Emerging Telecommunications Technologies 26(2)(2015) 290-301.
[36] Y. Luo, J. Liu, R. Hu, D. Zhang, G. Bu, Particle Swarm Optimization Combined with Q-learning of Experience Sharing Strategy, Journal of Frontiers of Computer Science and Technology 16(9)(2022) 2151-2162.
[37] C. Chen, K. Liu, C. Dong, H. Zhou, An adaptive bit rate algorithm model based on wireless network, Journal of Beijing University of Posts and Telecommunications 45(5)(2022) 115-120.