A Q-learning-Based Dynamic Spectrum Allocation Algorithm

Chunying Lv
Electronic System Engineering
Company of China
Beijing, China
e-mail: chunyinglv@163.com

Jiyang Wang
Electronic System Engineering
Company of China
Beijing, China
e-mail: jywang . nudt@gmail.com

Fei Yu
Electronic System Engineering
Company of China
Beijing, China
e-mail:YFei0210@163.com

Hui Dai
Electronic System Engineering
Company of China
Beijing, China
e-mail: daihui000@21cn.com

***Abstract*—Contraposing the status of spectrum resources is tension and spectrum utilization is low, the dynamic spectrum allocation technology is studied. Due to lack of central control entity in distributed cognitive networks, cognitive users need to select channel and power for communication intellective. Based on the intelligence of Q-learning, we proposed an improved dynamic spectrum allocation algorithm. The state space, action space and reward function of the algorithm are built in the paper, and the agents are guided to perform actions through designing the reward function. The algorithm gives fully autonomy selective to cognitive users. This strengthened the dynamic management of spectrum resources. We also did many experiments. Numerical simulation results show that the proposed algorithm can not only realize the autonomy of channel and power allocation, but also improve system throughput compared to others. It boosts spectrum utilization effectively.**

Keywords-cognitive radio; Q-learning; dynamic spectrum allocation; channel and power selection; autonomy

I. INTRODUCTION

With the rapid development of wireless business, the spectrum resources become more and more scarce, which will be a bottleneck restricts the development of wireless communication technology in future. However, the survey report about radio spectrum usage Federal Communications Commission (FCC) did in 2002 pointed out that the spectrum utilization of current allocation is very low, about 15% -85% [1]. So, there has not been shortage of spectrum resources in the true sense, the phenomenon of resources shortage and low utilization exist at the same time. As long as taking effective management manners and technical measures, those cognitive users are made to access the spectrum when they does not affect the main users, the tension phase of spectrum resources can be alleviated largely. Due to the static frequency allocation model lack of flexibility, it does not meet the need of electromagnetic spectrum management in the future. Hence, new technologies of frequency reuse must be developed in the time and space domain. Cognitive radio is put forward in this case as a spectrum resource sharing technology, its core function is to sense the electromagnetic environment, find the spectrum holes, and allocate the available frequencies to cognitive users dynamically without interfering the premise users, which improves spectrum utilization effectively.

At present, there are four kinds of spectrum allocation models: coloring of graph theory, Markov, auction bidding, and game theory [2]. But none of them involves the autonomy problem of dynamic spectrum allocation, it dose not meet the requirements of dynamic environment. In order to realize self-management of resources, the network needs self-learning ability to correct its control strategies constantly according to the actual operation. Q-learning as a model-free, teacher-free and on-line reinforcement learning algorithm is an effective way to solve such problems [3]. Recent years, the study of Q-learning has been used for cognitive radio system. Reference [4] used Q-learning to solve the data fusion problem in cooperative spectrum sensing, [5] utilized Q-learning to improve the signal detection performance for cognitive radio system, [6] applied Q-learning to limit the interference that cognitive users did to primary users in joint interference control of cognitive network, and [7] employed Q-learning solve the dynamic management of the cognitive radio system. On the basis of these studies, this paper advanced an improved Q-learning algorithm. The algorithm takes into account of the effect which frequency and power on the communication links simultaneously in the network. It realizes the spatial multiplexing of frequency effectively and achieves the purpose of making full use of spectrum resources.

II. Q-LEARNING THEORY [8]

Assume that an agent which interacts with a finite-state, time-discrete stochastic system. Let $S=\{s_1,s_2,\cdots s_n\}$ to be the state space of the system, and $A=\{a_1,a_2\cdots a_m\}$ to be the action space the agent could adopt. At some time, the agent perceives the environment state $s\in S$, chooses an action based on the current policy π effect on the environment. The environment state turns from s to the next state $s'\in S$. The immediate reward that agent gains from the environment is denoted by $r(s,a)$. Next time the agent updates policy entering a new iterative loop. The goal of learning algorithm is to learn an optimal policy for the agent to produce the largest expect cumulative reward. The cumulative reward was made as assessment function to evaluate the policy whether merits or not. It is given by:

$$V^{\pi}(s)=E\left\{\sum_{t=0}^{\infty}\gamma^t r(s_t,\pi(s_t))\middle|s_0=s\right\} \tag{1}$$

Published by Atlantis Press, Paris, France.

Where γ is the discount factor, $0<\gamma<1$. Then (1) can be expressed as:

$$V^{\pi}(s)=R(s,\pi(s))+\gamma\sum_{s'\in S}P_{s,s'}(\pi(s))V^{\pi}(s') \qquad (2)$$

Where $R(s,\pi(s))=E\{r(s,\pi(s))\}$ is the mean value of $r(s,\pi(s))$. $P_{s,s'}$ is the transition probability from state s to the next state. According to Bellman's optimality that there is at least one π^*, such that:

$$V^*(s)=V^{\pi^*}(s)=\max_{a\in A}\left[R(s,a)+\gamma\sum_{s'\in S}P_{s,s'}(a)V^*(s')\right] \qquad (3)$$

The advantage of Q-learning is to find the optimal π^* without knowing $R(s,a)$ and $P_{s,s'}(a)$ through simple iteration of Q-values. For a policy π, define a Q value as:

$$Q^{\pi}(s,a)=R(s,a)+\gamma\sum_{s'}P_{s,s'}(a)V^{\pi}(s') \qquad (4)$$

Applying the Bellman's criterion, we can find the largest $Q(s,a)$, denoted by $Q^*(s,a)$, is given by

$$Q^*(s,a)=R(s,a)+\gamma\sum_{s'}P_{ss'}(a)V^{\pi^*}(s') \qquad (5)$$

From (3) and (4), the optimal state's value and policy are

$$V^*(s)=\max_{a\in A}[Q^*(s,a)] \qquad (6)$$

$$\pi^*(s)=\arg\max_{a}Q^*(s,a) \qquad (7)$$

The Q-learning rule to update the Q-values is:

$$Q(s,a)=r(s,a)+\gamma\max_{a'}Q(s',a')) \qquad (8)$$

Where a' is the action that agent takes at state s' next time. From (8), it can be concluded that, the current Q value is consisted of the immediate reward under the current state and action, and the Q-values of subsequent states discounted by γ. Thus, the ideology of Q-learning is not to learn the optimal policy directly but optimize $Q(s,a)$ through iterations constantly. The optimal policy can be found by estimating the accumulated reward.

III. DYNAMIC SPECTRUM ALLOCATION ALGORITHM BASED ON Q-LEARNING

Because of lacking central control base station, there is no distribution body. The intelligence of Q-learning can turn the spectrum allocation problem into the problem that cognitive users choose the available channels. We make the cognitive user's channel selection module as the agent module. The interaction diagram of channel selection agent with the environment based on Q-learning is set up as shown in Fig. 1. The channel selection module selects a specific action a in the current state s, gaining the environmental feedback reward r. It observes the next state at one time, and learns the value of $Q(s,a)$, entering into the next iteration. Therefore, in the channel selection algorithm based on Q-learning, it needs to identify various elements of the algorithm, including the state space, action space, reward function, and so on. Now we will discuss them one by one.

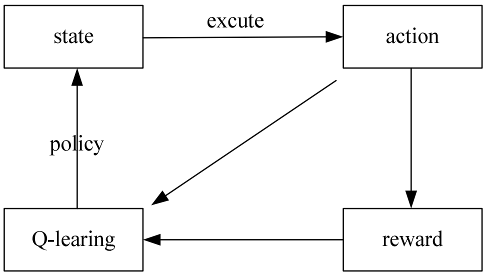

Figure 1. Channel selection module based on Q-learning

A. The State Space

The design of state space $S=\{s_1,s_2\cdots s_n\}$ is the base of agents select actions in reason. The state variable s is selected should be with the features of knowability and non-aftereffect. Knowability is the input of state must be the information that agent can extract and process. Non-aftereffect requires the previous state only affects the latter state, not flux because of the following elements [9]. The key point of Q-learning frequency selection algorithm is to consider cognitive users how to select appropriate channel and power for communication, avoiding interfere with users which are communicating. That is the problem to determine a reasonable "user-to be selected resources". Assume there are N cognitive users and C available channels in the network. The performance of transmitting equipments and receiving equipments are accordant to each cognitive user. The power values that transmitters can choose is P, and the available channels are set to be C. Any link uses channel $c_i\in C$, transmitting power $p_i\in P$ for communicating. So, the link under the current state can be expressed by $L_i=(t_i,r_i,c_i,p_i)$ using four-element array. t_i is the launching user of the link, and r_i is the receiving user. Furthermore, t_1 only communicates with r_1, t_2 with r_2, and so on. Assume there are M communication links in the network, when the channel selection algorithm is executed, the variable of cognitive network at time t can be described as $s=\left\{(t_o,r_o),(L_1,L_2\cdots L_M)\right\}_t$, $o\in\left[1,N/2\right]$. t, r is the communication pair that applies to establish link for the moment.

B. The Action Space

Q-learning-based channel selection algorithm, its action is to distribute channels and transmitting powers to the links which apply to communicate. The principle is not only make the link communicate successfully, but also does not cause interference to the link which are communicating. The action variable meet requirements above at time t can express by $a=(c_k,p_j)_t$. $c_k\in C$, $p_j\in P$ is to select available channels and transmit powers from the aggregate respectively.

C. The Design of Reward Function

There are many ways to define reward function. Typically, when the actions taken are consistent with the purpose to be achieved, the reward function can be set to a fixed value based on experience, such as in [10]. This method is too simple only considers few factors of reward function. Reference [11] defined the system throughput as reward function, [12] assessed the reward applying a specific action by evaluating the sum of the service rates of all the users on transmissions. They all achieved good results through Q-learning algorithm's iteration. We consider the algorithm's application background, putting the interfere factor between links into the design of reward function. That is, the agent selects channel at any time will take into account of the interference to the existing communication links, thereby selects the appropriate channel and power to communicate. On the basis of spectrum utilization, the reward function is expressed as:

$$r(s,a)=\sum_{n=1}^{N/2}\log_2^{(1+k\gamma_n)} \tag{9}$$

Where:

$$k=1.5\Big/\ln^{(0.2/BER_{tar})} \tag{10}$$

BER_{tar} is the threshold of error bit rate, γ_n is the ratio of signal to interference for the link n in the communication channel. And

$$\gamma_n=\frac{p_nG_{nn}}{n_0+\sum_{j=1,j\neq n}^{M}p_jG_{jn}\varepsilon(j,n)} \tag{11}$$

p_n is the transmitting power of link n, G_{nn} is the link gain between the sender and receiving end. G_{jn} is the gain between the transmitter of link j and the receiving end of link n. $G_{nn}=\omega/d_{nn}^4$, $G_{jn}=\omega/d_{jn}^4$. d_{nn} is the distance between the transmitter and receiver of link n, d_{jn} is the distance between the transmitter of link j and the receiving end of link n. ω is a fixed value. n_0 is the noise power. $\varepsilon(j,n)$ is the interference equation of user j and n, is determined by:

$$\varepsilon(j,n)=\begin{cases}1 & if\ j\ and\ n\ are\ on\ the\ same\ channel\\ 0 & other\end{cases} \tag{12}$$

D. The Searching Policy

An important aspect of Q-learning is the tradeoff between exploration and exploitation. Exploitation is to choose the best action according to (6), and exploration is to choose the action system has never been tried. The balance between exploiting and exploring depends on the estimate accuracy and the environment dynamic behavior. The searching policy of channel selection algorithm uses classical ε -greedy algorithm. The agent chooses a random action with a small probability ε, and chooses an action with a maximum Q value with probability $1-\varepsilon$.

IV. EXPERIMENTATION ANALYSIS

In order to test the effectiveness of the proposed algorithm, we use MATLAB simulation tool to simulate the algorithm's performance. The business duration of communication links in the network is modeled as Poisson process with the parameter λ. The probability of cognitive users' application is assumed to obey equality distribution. Each channel is modeled as having unit bandwidth and the same background noise power of n_0. Over the time period of interest, we assume that the channel gains are fixed. The case when application more than one never happens every time due to independent Poisson process and each link can only use one channel for communication. There are ten cognitive users in the network, including the links which has been allocated and those are requesting to communicate. The basic simulation parameters settings as shown in table 1.

TABLE I. SIMULATION PARAMETERS SETTINGS

Parameters	Settings	Parameters	Settings

Distribute region	1000(m)	Power(W)	1,3,5,7
Background noise	5×10^{-11}	ω	0.097
Available channels	3	BER_{tar}	10^{-4}
Allocated channels	2	Iteration numbers	10000
Application channels	3	λ	2

As mentioned above, the key of applying Q-learning is to construct the state space, action space and reward function. In Q-learning-based dynamic spectrum allocation algorithm, a lot of references did not consider the effects that power does to the allocation when they designed those factors. But this way is not scientific in practice. The link wants to build communication successfully not only need communication frequency but also need transmission power for the transmitter. Visibly, for making full reuse of spectrum in space, improving spectrum utilization effectively, it is important to consider the influence that power does to the algorithm. Take [10] for example, it only described channel for the state space. Fig. 2 is the convergence experiment based on Q-learning dynamic spectrum allocation algorithm. Compare the performance of the proposed algorithm with [10]. It can be seen, the collision probability is higher at starting time when the agent chooses channel and power. With the continuous learning of Q-learning algorithm, the collision probability gradually reduces and eventually reaches convergence. Obviously, the algorithm of [10] only considered channel factor has higher collision probability in the initial simulation. In our paper's simulation environment, the time required to reach convergence and the collision probability after convergence are more than the algorithm that channel and power are considered.

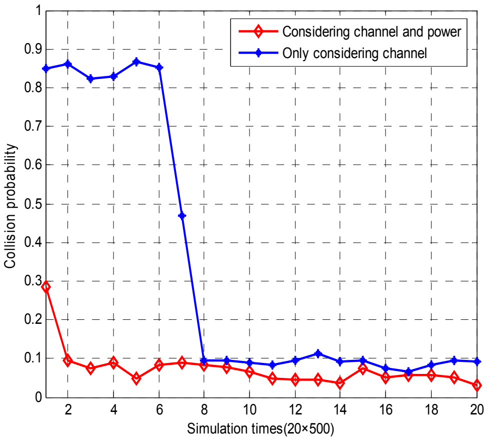

Figure 2. Convergence comparison between the two algorithms

Fig. 3 is aim at the effect of different parameters does to the convergence. The parameters of Poisson distribution are different, that is the communication duration is different. The parameter is bigger, the communication time also gets longer correspondingly. As shown in the figure, the biggest parameter has the biggest collision probability after the beginning and convergence. It shows that the longer duration of the communication, the bigger collision probability of the applying communication links. This conclusion is consistent with the practical application.

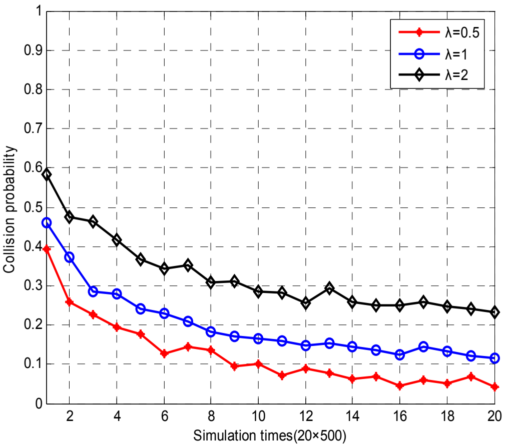

Figure 3. Convergence comparison of different parameters

Fig. 4 is average spectrum utilization of the algorithm. Compared with [12] and the static allocation algorithm, it is easy to find, the average spectrum utilization of our algorithm is significantly higher than the other two, and, the static spectrum allocation utilization is minimum.

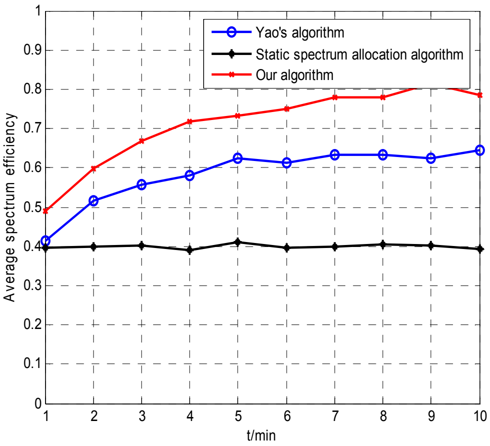

Figure 4. Average spectrum utilization

Fig. 5 simulates the users' satisfaction of each link. Our algorithm's performance is notable better than other algorithms in the article's simulation environment. The considering of interference factor in the algorithm reduces blocking probability when the users select communication frequency and power. It can meet the customer's requirements to the maximum.

Published by Atlantis Press, Paris, France.

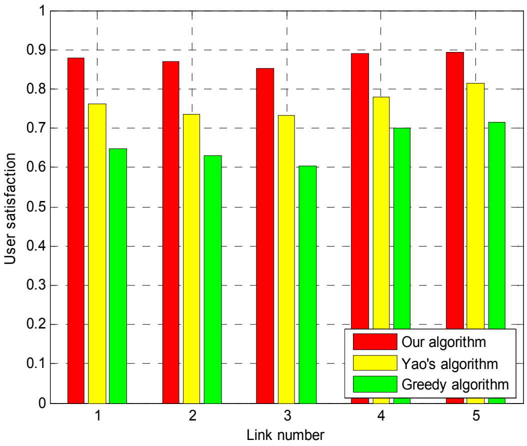

Figure 5. Users satisfaction

V. CONCLUSIONS

Q-learning as an important artificial intelligence method has been applied in many areas. The paper studies dynamic spectrum allocation based on the intelligence of Q-learning and its application. We proposed an improve dynamic spectrum allocation algorithm on the basis of Q-learning. It can be seen from the construction of state, action and reward function, the Q-learning-based dynamic spectrum allocation algorithm guides agents execute actions by designing the reward function. And, the searching policy of Q-learning algorithm wouldn't make the agent choose the action that has the biggest reward, but choose other actions at a certain probability. This also supports the dynamic management of frequency resources. The simulation experiments and analysis fully explain the validity of this algorithm. It improves the spectrum utilization effectively.

ACKNOWLEDGMENT

This paperis supported by the National Natural Science Foundation of China (NSFC) Grant #61002034 and the Natural Science Foundation Project of CQ CSTC Grant #2010BB2203.

REFERENCES

[1] Xin Xiang, “Software Radio Principle and Technology,” Xian: Xian electronic science and technology university, 2010.

[2] Jie Chen, Shaoqian Li and Chulin Liao, “The Research of Spectrum Resource Allocation Algorithm Based on Demand in Cognitive Radio Network,” Computer applications, vol. 28, A9, pp. 2188-2191, 2008.

[3] Mo Li, Youyun Xu and Junquan Hu, “A Q-Learning Based Sensing Task Selection Scheme for Cognitive Radio Networks,” International Conference on Wireless Communication & Signal Processing, Nanjing, pp. 1-5, 2009.

[4] Zhe Chen and Robert C. Qiu, “Cooperative Spectrum Sensing Using Q-Learning with Experimental Validation,” Proceedings of IEEE, pp. 405-408, 2011.

[5] Y. B. Reddy, “Detecting Primary Signals for Efficient Utilization of Spectrum Using Q-Learning,” Procedings of the Fifth International Conference on Information Technology: New Generations, Las Vegas, pp. 360-364, 2008.

[6] Ana Galindo-Serrano and L. Giupponi, “Decentralized Q-learning for Aggregated Interference Control in Completely and Partially Observable Cognitive Radio Networks,” Proccedings of the IEEE CCNC, pp. 1-6, 2010.

[7] Cheng Wu, Kaushik Chowdhury and Marco Di Felice, “Spectrum Management of Cognitive Radio Using Multi-agent Reinforcement Learning,” Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, pp. 1705-1712, 2010.

[8] Yongjing Zhang, Zhiyong Feng and Ping Zhang ping, “The Independent Joint Wireless Resource Management Algorithm Based on Q-learning,” Journal of electronics and information, vol. 30, A3, pp. 676-679, 2008.

[9] Mo Li, Youyun Xu and Yueming Cai, “A Q-Learning Based Sensing Management Algorithm for Cognitive Radio System,” Journal of electronics and information, vol. 32, A3, pp. 623-628, 2010.

[10] Kok-Lim Alvin Yau, Peter Komisarczuk and Paul D. Teal, “A Context-aware and Intelligent Dynamic Channel Selection Scheme for Cognitive Radio Networks,” Proceedings of The 4th Iternational Conference on CROWNCOM, Hannover, pp. 1-6, 2009.

[11] Kok-Lim Alvin Yau, Peter Komisarczuk, Paul D. Teal et al, “Enhancing Network Performance in Distributed Cognitive Radio Networks using Single-Agent and Multi-AgnetReinforcement Learning,” 2010 IEEE 35th Conference on Computer Networks, Denver, pp. 152-159, 2010.

[12] Yanjun Yao and Zhiyong Feng, “Centralized Channel and Power Allocation for Cognitive Radio Networks: A Q-learning Solution,” Proceedings of Future Network and Mobile Summit, Florence, pp. 1-8, 2010.

Published by Atlantis Press, Paris, France.