A New Opportunistic Spectrum Access Approach in Unslotted Primary Systems

Kai Zhang, Ou Li, Baiwei Yang, Yang Liu
Department of Communication Engineering
Information Science and Technology Institute
Zhengzhou, China
meetkai@126.com, zzliou@126.com, youngbw@163.com, liuyang0925@sohu.com

***Abstract*—To solve the channel selection issue in opportunistic spectrum access, a new Q-Learning based algorithm for channel selection is proposed. The algorithm can lead secondary user to select channels with maximum cumulative reward, and maximize secondary user throughput. A Boltzmann learning rule is adopted to achieve well tradeoff between channel exploration and exploitation. From the simulation results, compared with random selection algorithm, the algorithm does not require prior knowledge or prediction models of the channel environment, yet can still select the optimal channel adaptively, improve the secondary user capability and attain to the convergence in short time.**

Keywords-cognitive radio; opportunistic spectrum access; learning; channel selection

I. INTRODUCTION

Opportunistic spectrum access (OSA) [1], which mainly builds on the cognitive radio technology [2], has been regarded as a promising solution to lessen the spectrum scarcity problem and hence has drawn great attention. The overall design objective of OSA is to provide sufficient benefit to secondary users while protecting spectrum licensees from interference. One of the important and difficult problems in OSA system is that how to select the best available channel to achieve efficient spectrum exploitation. The channel selection strategy should not only primary users from harmful interference, but also satisfy secondary users requirement, and increase spectrum utilization.

There have been many significant developments in the past few years on OSA. To optimize spectrum access while considering physical layer spectrum sensing and primary user's traffic statistics, a decision-theoretic approach based on partially observable Markov decision process (POMDP) is proposed in[3],which can optimize secondary user's performance, accommodate spectrum sensing error, and protect primary users from harmful interference. An extension of [3] is presented in [4-6]. A primary-prioritized Markov approach for dynamic spectrum access is proposed in [7], which models the interactions between the primary users and the secondary users as continuous-time Markov chains. However, there are still several unsolved problems for OSA systems. First most existing work is based on the assumption that the secondary users have full prior knowledge and prediction models of the environment's dynamics and behaviors. Secondly, the environment is required to be static during the convergence of the algorithms. However, the assumptions are impossible in practice [8], because obtaining the prior knowledge consumes enormous network resources, e.g., time, power, and bandwidth, and may not be feasible in some scenarios; moreover, acquiring information about other users also leads to heavy communication overhead. Therefore, we should call for innovative techniques that can achieve good performances by learning directly from interaction with the environment and without needing models of such environments.

Q-Learning [9] is a model-free, teacher-free, online reinforcement learning algorithm, which can solve the above problems. In this paper, we apply Q-Learning algorithm to channel selection strategy in OSA system. We consider an OSA system with the following characteristics: (1)the channel is realistic in practice, (2)the spectrum holes are time-varying, (3)no need to know the channel availability statistics. Moreover the secondary users learn from interaction with the environment using a Boltzmann rule to optimize the channel selection.

This paper is organized as follows. In Section 2, we state the OSA system model and discuss its assumption. In Section 3, we formulate the Q-Learning algorithm for channel selection. In Section 4, we evaluate the proposed approach. Finally, we conclude the paper in Section 5.

II. SYSTEM MODEL

Assume that there are N channels available for transmissions by the primary and secondary users, each with bandwidth $B_i(i = 1,\cdots,N)$. Limited by its hardware constraints and energy supply, a secondary user can only tries to access one of the N channels opportunistically. The primary system is not slotted; primary users can access the channel at any time. The occupancy of each channel by a primary user evolves independently according to a continuous-time Markov chain with idle and busy state. The holding times are exponentially distributed with parameters λ_i^{-1} for the idle and μ_i^{-1} for the busy state, respectively. We stress the primary system is not slotted; primary users can access the channel at any time. Note that the primary traffic load η_i on band B_i can be expressed as $\mu_i^{-1}/(\lambda_i^{-1}+\mu_i^{-1})$.

The secondary user employs a slotted communication protocol. At the beginning of each slot, a secondary user with data to transmit chooses one of the N channels according to the channel selection strategy, and uses the sensing outcome to decide if and in which channel to transmit. At the end of each slot, secondary user formulates the reward to update the

Published by Atlantis Press, Paris, France.

learning algorithm for channel selection. The basic slot structure is illustrated in Figure 1

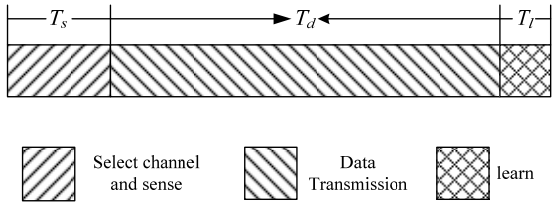

Figure 1 the slot structure

III. A Q-LEARNING BASED CHANNEL SELECTION ALGORITHM

A. Q-Learning

Q-Learning [9] is typically formalized in the context of Markov Decision Process (MDP), which is a finite-state, discrete-time, stochastic dynamical system. Let S be the set of the environment state, $S=\{s_1, s_2,..., s_n\}$ and A be a set of discrete actions, $A=\{a_1, a_2,..., a_n\}$ and r be the system reward. The objective of the learner is then to find an optimal policy $\pi^*(s) \in A$ for each s, which maximizes the cumulative measure of the reward $r(s,a)$ received over time. The total expected discounted return over an infinite time horizon, is given by

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

where $0 \ll \gamma < 1$ is the discount factor. According to Bellman optimal rule, the maximal equation (1) can be expressed as

$$\begin{aligned} V^*(s) &= V^{\pi^*}(s) = \max_{\pi} V^{\pi}(s) \\ &= \max_{a \in A}[R(s,a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^*(s')] \end{aligned} \tag{2}$$

where $R(s,a)$ is the mean value of $r(s_t,a_t)$, $P_{s,s'}(a)$ is the probability from state s to s' when executing action a.

Q-Learning does not need predict the environment model, yet can find the optimal strategy by simple iteration without prior knowledge of $R(s,a)$.

For a policy π, define a Q value (state-action value) as

$$Q^{\pi}(s,a) = R(s,a) + \gamma \sum_{s' \in S} P_{s,s'}(a) V^{\pi}(s') \tag{3}$$

which is the expected discounted cost for executing action at state s and then following policyπ thereafter. From equation (2) and (3), we can obtain

$$V^*(s) = \max_{a \in A} Q^*(s,a) \tag{4}$$

$$\pi^*(s) = \arg\max_{a} Q^*(s,a) \tag{5}$$

$Q^*(s,a)$ can be obtained through the following recursive manner

$$Q_{t+1}(s,a) = (1-\alpha) Q_t(s,a) + \alpha(r_t + \gamma \max_{a'} Q_t(s',a')) \tag{6}$$

where $\alpha = 1/(1+visit(s,a)), \alpha \in (0,1]$ is the learning rate, $visit(s,a)$ is the visited times of state-action (s,a) pair. It has been shown [9] that if Q-value of each admissible (s,a) pair is visited infinitely often, and if the rate is decreased to zero in a suitable wan, then as $t \to \infty$, $Q_t(s,a)$ converges to $Q^*(s,a)$ with probability 1.

B. A Q-learning based channel selection algorithm

We employ Q-Learning to channel selection for OSA system, the channel selection and access structure is shown in Figure 2. As it noted, secondary user select a channel according to the channel selection strategy, and observe the wireless spectrum environment and decide whether or not to access the channel. In this manner, secondary user can obtain a reward and achieve interaction with the environment. We formulate the OSA as a finite Markov decision process (MDP), which consist of state set S, action set A, state transition function δ and reward function r:

State set: S consists of m state $\{s_1, s_2,..., s_m\}$. The secondary user is said to be in state s_i when it is using band B_i at the current time, i.e., no primary users are using band B_i.

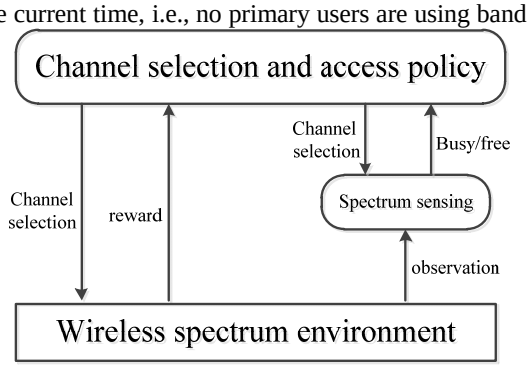

Figure 2 channel selection and access structure

Action set: A consists of m actions, $A=\{a_1, a_2,..., a_m\}$. The secondary user can switch to state s_i when executing action a_i.

Transition function: $\delta: S \times A \to S$ is the transition function, specifying the next state the system enters provided its current state and the action to be performed, $\delta(s_j,a_k)=s_k$.

Reward function: the design of reward function r is based on the system performance. The objective is to obtain the maximal throughput when using the available spectrum. The reward at the time slot j is defined as:

$$r(j) = \frac{T_d}{T_s + T_d} I_i \Phi_{T_d,j}(t) B_i \tag{7}$$

where I_i indicates whether the channel i is idle. $\Phi_{Td,j}(t)$ is the idle time of the transmission time T_d at the time slot j on channel i, B_i is the channel bandwidth.

C. Boltzmann learning rule

The obvious strategy for secondary user to select channel is always to choose the action with the maximal Q-value. However, it exists risks using the above strategy, because secondary user may limit into early training actions with high Q-value so as not to explore other higher actions. In practice, the convergence theory requires that each state-action pair happens infinitely. Obviously, if secondary user always selects actions with the current maximal Q-value, it would not guarantee the infinity. In order to balance between "exploitation" and "exploration", we employ the Boltzmann learning rule [10] and select channels with probability

$$P(a_i / s) = \frac{\exp[\frac{\hat{Q}(s,a_i)}{T}]}{\sum_j \exp[\frac{\hat{Q}(s,a_j)}{T}]} \quad (8)$$

where T is a positive parameter called the temperature. High temperatures cause the actions to be all (nearly) equally probable, and make secondary user randomly select all channel. Low temperatures cause a greater difference in selection probability for action s that differ in their value estimates, and lead secondary user to select the actions with higher Q-value (reward).

According to the set of state, action, reward, and experiment strategy, our proposed Q-Learning based channel selection algorithm can lead secondary user to selection the channels that would maximize the system throughput. Moreover, the exploration strategy can guarantee that secondary user cannot always use the channels with maximal reward, but select channels with probability. This scheme guarantees that secondary user can cumulate experiences on different channels and select the optimal channel adaptively, improve the secondary user capability.

IV. Performance Evaluation

In this section, we study the proposed Q-Learning scheme by evaluating and comparing its performance with random-access scheme which select channels randomly. We employ Monte Carlo experiment and each point in the figure is the mean value of 1000 formulation at the same parameters. The algorithm of the channel selection is set as: the channel number m=7, channel bandwidth B_i = 200 kHz, the slot length T_l=100×10^{-3}s, sensing time T_s=5×10^{-3}s, and transmission time T_d=95×10^{-3}s, the discount factor γ=0.9.

Throughout this section, we characterize the primary user traffic system load by $\varphi = (1/m)\sum_{i=1}^{m} \eta_i$ (which is denoted as fai in figures) and $Cov=\sigma/\varphi$, which, respectively, denote the average and the coefficient of variation of PU traffic load across all channels, where φ denotes the standard deviation of the traffic loads. The advantage of Q-Learning lies in its capability to converge to an approximately optimal behavior without needing prior knowledge of the PUs' traffic behavior.

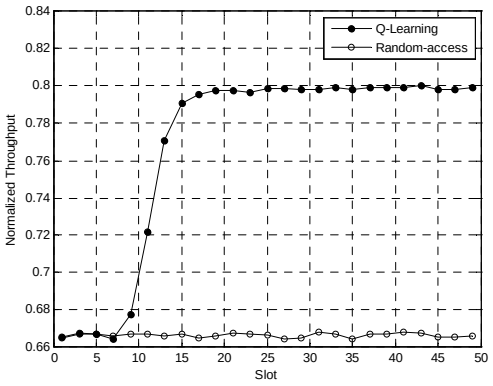

Figure 3 Performance of the proposed Q-Learning algorithm and random access algorithm (φ=0.5, Cov=0.5)

Figure 3 plots the total throughput, which is normalized with respect to the maximal achievable throughput, which the secondary user achieves as a result of using our Q-Learning and the random-access schemes. In the simulation scenario, the φ and Cov is all set to0.5. Simulation results show that our proposed Q-Learning based channel selection algorithm always outperform the random-access algorithm by learning from the experience and without prior knowledge of the environment as the slot time increasing. Furthermore, the proposed algorithm gradually converges to a stable throughput in a short time.

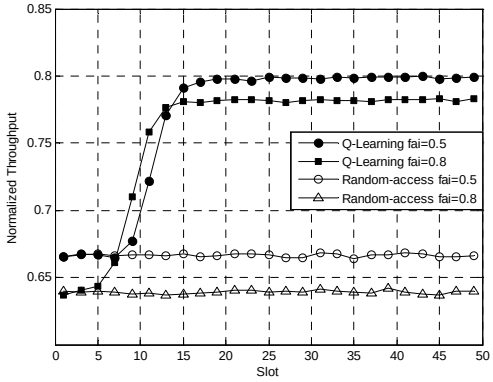

Figure 4 Performance of two different algorithms under different PU traffic load.

Figure 4 plots the total throughput that secondary user achieves under our proposed Q-Learning and the random-access schemes for two different PU load variations: φ=0.5, and φ=0.8, Cov is set to 0.5. As expected, the lower the φ, the more larger the achievable throughput under both algorithm.

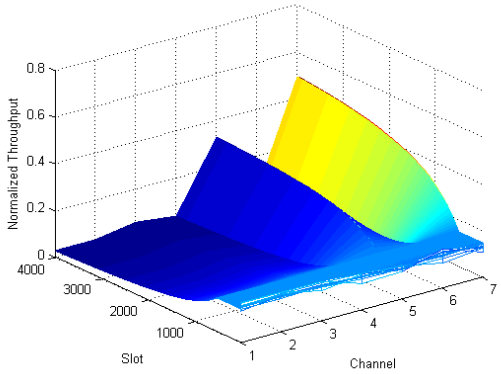

Figure 5 Performance and process of the proposed channel selection algorithm

To further illustrate the efficiency of the proposed Q-Learning channel selection algorithm, we plot in Figure 5 the channel selection process as the time increasing. In this simulation scenario, the slot number is set to 4000, φ=0.5, and Cov=0.5, the PU traffic load from channel 1 to 7 is set to [0.90, 0.88, 0.45, 0.44, 0.23, 0.43, 0.21]. As shown in Figure 5, at the beginning, the secondary user is in the exploration state, and randomly access different channels in order to obtain more reward. As the time increasing, the percentage

of selecting the optimal channel 5 and 7 is gradually increasing; the channel selection algorithm is gradually transferring from the exploration state to the exploitation state. The percentage of selecting the optimal channels of all the channels is nearly 80%. These obtained results show that the proposed Q-Learning algorithm can efficiently learn from history and adapt to future channel selection through online learning, hence attain better system performance.

V. CONCLUSION

We have presented in this paper a Q-Learning based channel selection algorithm to solve the channel selection issue for OSA system without prior knowledge and prediction models of the environment. A Boltzmann learning rule is adopted to achieve well tradeoff between channel exploration and exploitation. Simulation results show that the proposed algorithm can achieve good channel selection for OSA system by learning from experience and without prior knowledge and prediction models of the environment.

REFERENCES

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *Selected Areas in Communications, IEEE Journal on,* vol. 23, pp. 201-220, 2005.

[2] Z. Qing and B. M. Sadler, "A Survey of Dynamic Spectrum Access," *Signal Processing Magazine, IEEE,* vol. 24, pp. 79-89, 2007.

[3] Q. Zhao, L. Tong, and A. Swami, "Decentralized cognitive mac for dynamic spectrum access," in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, 2005, pp. 224-232.

[4] C. Yunxia, Z. Qing, and A. Swami, "Joint Design and Separation Principle for Opportunistic Spectrum Access in the Presence of Sensing Errors," *Information Theory, IEEE Transactions on,* vol. 54, pp. 2053-2071, 2008.

[5] S. H. A. Ahmad, M. Y. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of Myopic Sensing in Multichannel Opportunistic Access," *Ieee Transactions on Information Theory,* vol. 55, pp. 4040-4050, Sep 2009.

[6] Y. X. Chen, Q. Zhao, and A. Swami, "Distributed Spectrum Sensing and Access in Cognitive Radio Networks With Energy Constraint," *Ieee Transactions on Signal Processing,* vol. 57, pp. 783-797, Feb 2009.

[7] B. B. Wang, Z. Ji, K. J. R. Liu, and T. C. Clancy, "Primary-Prioritized Markov Approach for Dynamic Spectrum Allocation," *Ieee Transactions on Wireless Communications,* vol. 8, pp. 1854-1865, Apr 2009.

[8] Y. H. Xu, J. L. Wang, Q. H. Wu, A. Anpalagan, and Y. D. Yao, "Opportunistic Spectrum Access in Unknown Dynamic Environment: A Game-Theoretic Stochastic Learning Solution," *Ieee Transactions on Wireless Communications,* vol. 11, pp. 1380-1391, Apr 2012.

[9] A. Gosavi, "Reinforcement Learning: A Tutorial Survey and Recent Advances," *Informs Journal on Computing,* vol. 21, pp. 178-192, Spr 2009.

[10] A. Weissensteiner, "A Q-Learning Approach to Derive Optimal Consumption and Investment Strategies," *Ieee Transactions on Neural Networks,* vol. 20, pp. 1234-1243, Aug 2009.

.