Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 129 (2018) 433–437

Procedia Computer Science

www.elsevier.com/locate/procedia

2017 International Conference on Identification, Information and Knowledge in the Internet of Things

Reinforcement Learning Spectrum Management Paradigm in Cognitive Radio using Novel State and Action Sets

Zhijie Yin[a], Yiming Wang[a], Cheng Wu[a,*]

[a]*School of Rail Transportation, Soochow University, 215131, Suzhou, Jiangsu, China*

Abstract

This paper proposes a reinforcement learning(RL) model for *cognitive radio*(CR). By using this model, *cognitive base station*(CBS) can preform two-step decision of channel allocation, that is, whether to switch the channel for CR users and how to select the best channel if the CBS decides to switch, to avoid excessive channel switch and improve the throughput of the unlicensed user. Also, the performance of RL spectrum management depends highly on exploration strategy. Epsilon-greedy exploration method is used to solve the balance problem of RL decision process. Simulation results show that the implementation of the epsilon-greedy in each decision step has a remarkable effect on the system performance. The proposed method is superior to traditional RL spectrum allocation scheme in improving unlicensed users' throughput and reducing channel switch.

Selection and peer-review under responsibility of the scientific committee of the 2017 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2017).

Keywords: Spectrum management; cognitive radio; reinforcement learning; exploration and exploitation dilemma;

1. Introduction

In CR community, one of the main challenge is channel allocation. Facing this problem, a large number of approaches are proposed to improve the CR users' quality of service (QoS) without interfering the licensed users which in CR community is known as primary users (PUs)[1, 3]. However, when the PU activity occurs in the channel used by the CR users, CR users should strictly follow the principle of protection of the PU to vacate the current using channel, thereby causing frequent channel switching which is quite harmful[8, 6]. Frequent channel switching not only lowers the probability of successful transmission, but also causes the long packet delay and some other costs such as high power consumption during reestablishing the link[9]. Since CR communication system is highly volatile due to the unpredictability of PU, this problem is even worse.

* Corresponding author. Tel.: +86-512-6750-1742; fax: +86-512-6750-1742.
E-mail address: cwu@suda.edu.cn

Selection and peer-review under responsibility of the scientific committee of the 2017 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2017).
10.1016/j.procs.2018.03.020

There are two key problems in the process of establishing the RL model. The first is how to build state and action sets. The second is whether to use the existing knowledge or to select best action[2, 7]. Correct implementation of exploration strategy plays an important role in improving the performance of the system.

By summarizing the above problems, this paper proposes a novel reinforcement learning model using a dual-functional Q table which contains efficient state and action sets. The ϵ-greedy method is implemented to optimize the two-step decision process in CBS, preventing the CBS from falling into local optimum.

2. CBS Spectrum Management Framework

2.1. Reinforcement Learning Based CBS

The classic RL algorithm in CBS is constructed as follows. In time slot t, the CBS perceives its current state $s_t \in S$ and the action set can be taken as A_{s_t}. The CBS chooses an action a_t from the set A_{s_t}, which will later receive a reward given by the radio environment r_t and a new state s_{t+1}. Based on these facts, the RL model in CBS must develop a policy $\pi : S \rightarrow A$ which maximizes the long-term reward $R = \sum_t \gamma r_t$ for Markov Decision Processes (MDPs), where $\gamma, 0 \leq \gamma \leq 1$ is the discounting factor which represents the impact of the future reward on the current action taken by CBS.

Considering the CBS as an agent in the CR network environment, we define the state at time slot t, denoted s_t, as $s_t = \left(\vec{ch}\right)_t$ where the $\vec{ch}$ is a channel vector, represents the licensed channels in the CBS coverage area. Assuming that there are M licensed channels scheduled by the CBS, we use the index of licensed channel to specify these channels, as $\vec{ch} = \{ch_1, ch_2, ch_3, ..., ch_M\}$. The current state represents the licensed channel CBS used to share with the CR user.

In a given time slot t and a given state $s_t = ch_i$, the action taken by CBS will directly affect the QoS of the CR user. When the CBS is serving CR user on channel ch_i, facing the active PU, the action allows the CBS to switch from its current channel and reestablish a connection with CR user on another channel, or CBS force the CR user to back off on the current channel to wait until the PU activity is off. When the PU state is idle, the CBS performs transmission with CR user, and the CBS will keep service on this channel in this slot to transmit data with CR user. Here we define action $a_t = \{\vec{k}\}_t$ where the $\vec{k}$ is a action vector, and the available action of the CBS at time slot t is $k_t \in \vec{k} = \{switch\_channel, stay\}$.

And one of the most successful of RL algorithm is Q-learning. This algorithm calculates an updates to its expected discounted reward $Q(s_t, a_t)$ using:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{1}$$

where $\alpha, 0 \leq \alpha \leq 1$ is the learning rate of the RL.

2.2. Epsilon-Greedy Exploration Strategy in Double Step Decision Process

In this paper, ϵ-greedy strategy is used to ensure that the CBS can explore all the state-action pairs in the two-step decision process and guarantee the quality of the decision at the same time. The first decision step is, in current slot t, if the sensing result of the PU activity on the current channel is busy, then in this slot CR user cannot transmit data with CBS. The CBS should take action to decide whether to switch or to stay. And the ϵ-greedy exploration chooses an action with the maximum Q with the probability ϵ_1, and otherwise chooses a random action.

The first step decision of CBS on current channel ch_i is conducted as follows :

$$\pi_1(a) = \begin{cases} arg \max\limits_{k_i} Q(ch_i, \vec{k}), & \xi < \epsilon_1, \\ random\ action\ from\ \vec{k},\ otherwise \end{cases} \tag{2}$$