Reinforcement Learning Enabled Cooperative Spectrum Sensing in Cognitive Radio Networks

Wenli Ning, Xiaoyan Huang, Kun Yang, Fan Wu, and Supeng Leng

Abstract: **In cognitive radio (CR) networks, fast and accurate spectrum sensing plays a fundamental role in achieving high spectral efficiency. In this paper, a reinforcement learning (RL) enabled cooperative spectrum sensing scheme is proposed for the secondary users (SUs) to determine the scanning order of channels and select the partner for cooperative spectrum sensing. By applying Q-learning approach, each SU learns the occupancy pattern of the primary channels thus forming a dynamic scanning preference list, so as to reduce the scanning overhead and access delay. To improve the detection efficiency in dynamic environment, a discounted upper confidence bound (D-UCB) based cooperation partner selection algorithm is devised wherein each SU learns the time varying detection probability of its neighbors, and selects the one with the potentially highest detection probability as the cooperation partner. Simulation results demonstrate that the proposed cooperative spectrum sensing scheme achieves significant performance gain over various reference algorithms in terms of scanning overhead, access delay, and detection efficiency.**

Index Terms: **Cooperative sensing, multi-armed bandit, Q-learning, reinforcement learning, spectrum sensing.**

I. INTRODUCTION

IN wireless networks, inefficient and fixed spectrum usage mode results in low utilization of spectrum resources. Cognitive radio (CR) technology is envisaged to solve this problem by exploiting the existing wireless spectrum opportunistically [1], [2]. In CR networks, secondary users (SUs) can opportunistically transmit in the vacant portions of the spectrum already assigned to the licensed primary users (PUs). Before transmitting, SUs are required to sense the available channels which are not occupied by PUs so as to minimize the interference caused to the PUs. In order to maximize the throughput of CR network, SUs needs to efficiently identify and exploit the spectrum holes of the primary network. Thus, fast and accurate spectrum sensing is crucial to the performance of both primary and CR networks.

Note that the detection accuracy of a single SU is susceptible to fading and shadowing effects, which may bring about missed detection and false alarm. Though the capability of detecting weak signals can be improved by equipped with more sensitive receiver, it will increase the implementation complexity and the associated hardware cost. Furthermore, taking the mobility of SUs into account, the detection ability of SUs changes dynamically and is unknown to each other. On the other hand, energy detection is a common method in local detection due to its simplicity, which is incapable of detecting multiple channels simultaneously. When there is a demand, an SU needs to detect the licensed channels sequentially until it finds an available channel or fails, which may lead to high access delay and scanning overhead.

Cooperative spectrum sensing technology [3], [4] has been widely used in CR networks to overcome the performance degradation of spectrum sensing due to multi-path fading and shadowing while without increasing the implementation cost of SUs. Meanwhile, being a powerful tool in process control, reinforcement learning [5] technique has been widely applied in a wide range of areas. Motivated by the related works, a cooperative spectrum sensing scheme based on RL is proposed in this paper, so as to improve the performance of spectrum sensing in dynamic CR networks. In the proposed scheme, each SU is an agent who learns the behaviors of channels and neighbors, and then takes action to improve the detection efficiency and reduce the scanning overhead and access delay. Our contributions can be summarized as follows:

- We propose a channel selection algorithm based on Q-learning to determine the scanning order of the channels, so as to reduce the scanning overhead and access delay. Specifically, each SU learns the occupancy pattern of the primary channels, and updates a dynamic scanning preference list of the channels based on the predicted channel status. A novel reward function is devised to improve the accuracy of channel status prediction during the learning process.
- We propose a cooperation partner selection algorithm based on discounted upper confidence bound (D-UCB) algorithm to improve the detection efficiency. Distinct from the static network scenario with fixed detection ability of each user, we assume that the detection ability of each SU dynamically changes due to the mobility of users and the time variation of wireless channels, which is more applicable to the practical network scenarios. In this case, each SU learns the time varying detection probability of its neighbors, and selects the one with the potentially highest detection probability as the partner for cooperative spectrum sensing.
- Numerical results show the proposed RL enabled cooperative spectrum sensing scheme achieves less number of attempts, higher detection probability, and lower call block

Manuscript received by April 1, 2019; approved for publication by Jang-Won Lee, Division II, September 18, 2019.

This work is supported by the National Natural Science Foundation of China under Grant No.61601083, the National Key R&D Program of China under Grant No.2018YFC0807101, and the Science and Technology Program of Sichuan Province, China under Grant No.2019YFH0007.

W. Ning, X. Huang, F. Wu, and S. Leng are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, email: {201721010505, xyhuang, wufan, spleng}@uestc.edu.cn.

K. Yang is with School of Computer Science and Electronic Engineering, University of Essex, email: kunyang@essex.ac.uk.

X. Huang is the corresponding author.

Digital Object Identifier 10.1109/JCN.2019.000052

1229-2370/19/$10.00 © 2019 KICS

rate compared with the reference algorithms, thus reducing the scanning overhead and access delay while improving the detection efficiency.

The remainder of this paper is organized as follows: Section II addresses the related works. Section III describes the system model. Section IV elaborates the proposed RL enabled cooperative spectrum sensing scheme. Section V evaluates the performance of the proposed scheme. Finally Section VI concludes the paper.

II. RELATED WORK

Cooperative spectrum sensing technology has attracted significant attention over the last few years as a promising technology to address spectrum detection issues for CR networks. In [6], the authors proposed a simple quantization-based multibit cooperative sensing method to address the soft decision fusion strategy in a limited bandwidth of control channel, which achieves a tradeoff between the sensing performance and the control channel overhead. In [7], all the CR nodes participating in cooperation have been distinguished based on their reliability which depends on the past decisions of corresponding node and of the central node. Based on this, the authors proposed a reliability based weighted algorithm to improve the spectrum sensing performance mainly in low SNR regions of the SUs. In [8], the authors applied adaptive threshold to each cooperation node to enhance the sensing performance of energy detection scheme for low SNR region. The authors in [9] pointed out that when a SU with the highest detection ability cooperates with others, the detection ability of cooperation is likely lower than employing directly the local decision of this SU as the final decision result. However, it's difficult to acquire other SUs' detection ability in a dynamic situation. In [10], the authors adopted support vector machine (SVM) to group users for cooperative sensing. The resulting user group which participates in cooperative sensing procedures is safe, less redundant, or the optimized user group, leading to better performance in terms of security, energy consumption, and sensing efficiency.

In RL framework, the action-taking agent interacts with the external environment through reward mechanisms, and adjusts its action according to the reward values obtained in the environment. The aim of the agent is to learn the optimal action to maximize the total reward. Recently, Q-learning [11], one of the RL algorithms, has been used to model the behaviors of the SUs in CR networks. In [12], SU is modeled by Q-learning mechanism to learn other SUs' behaviors and select the independent users under correlated shadowing for cooperation to improve detection efficiency. However, the detection ability of SUs is not considered when selecting partners for cooperative sensing, which may result in poor detection efficiency. In order to alleviate scanning overhead and access delay, authors in [13] adopted Q-learning technique to estimate status of channels based on the history of channel usage, so that each SU can select the most likely idle channel to detect and access accordingly. However, the proposed algorithm may not follow up with the dynamic changes of channel status due to the separation of exploration and exploitation stage. In addition, the local detection by a single user may also cause detection inefficiency. The authors in [14] proposed a two-stage learning approach in CR networks, including a reinforcement learning approach for channel selection and a Bayesian approach to determine the transmission duration. In [15], the authors studied the scheduling strategy for different buffers on multiple channels by using Q-learning and deep learning, aiming at maximizing the system throughput. In [16], the authors formulated the distributed throughput maximization problem as a non-cooperative game, and designed a stochastic learning automata-based algorithm to find Nash equilibrium point. In [17], the authors compared the performance of different machine learning approaches in terms of spectrum classification accuracy and computational time.

Multi-armed bandit (MAB) [18] is another RL algorithm, has been used to guide SUs to make selection decisions in CR networks. In [19], authors formulated the online sequential channel sensing and accessing problem as a sequencing multi-armed bandit problem to improve the throughput. In [20], the authors rewrote information gain of the conventional greedy method as the reward of the multi-armed bandit, and introduced multi-armed bandit into adaptive boolean compressive sensing. Experimental results indicate that the proposed method outperforms the conventional greedy method. In [21], the authors modeled the channel selection problem as MAB problem, and applied ϵ-greedy algorithm to select a channel with lowest utilization ratio. The results show that the proposed scheme improves the performance of the communication systems by efficient sensing of the channels with lower utilization rate. In [22], the authors considered the dynamic spectrum access problem and formulated the problem as a restless multi-armed bandit problem with a time varying constraint on the set of arms that can be activated at each given time.

The aforementioned works applied learning algorithms to channel selection or user cooperation, and achieved preferable performance. However, none of them studied the joint design of channel selection and user cooperation based on machine learning. Motivated by the existing works, we are focused on the cooperative spectrum sensing scheme design based on RL, including channel selection and cooperation partner selection, aiming at improving the performance of spectrum sensing in dynamic CR networks.

III. SYSTEM MODEL

We consider a CR network as shown in Fig. 1, consisting of K randomly distributed SUs. The CR network coexists with the primary network, and there are L primary channels that can be accessed by the SUs opportunistically. Let $SU_k, 1 \leq k \leq K$ represent secondary user k, and $\mathcal{N}_k$ denote the set of indexes of the neighbors of SU_k. Let $c_i, 1 \leq i \leq L$ denote primary channel i. Different from the static network scenario in [13], we consider a dynamic distributed CR network, wherein SUs move randomly in the network. In this case, we assume that the detection capability of each SU is time varying due to its mobility and the effects of fading and shadowing, and is unknown to other SUs.

When there is a demand at SU_k, SU_k attempts to find an idle primary channel to access. Specifically, SU_k needs to scan all the primary channels sequentially until successfully finds an

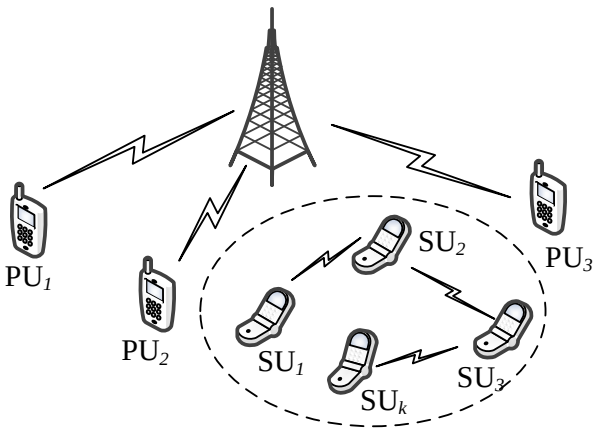

Fig. 1. Cognitive radio network.

available channel or fails. To speed up this procedure thus reducing scanning overhead and access delay, proper scanning order of the primary channels plays a vital role. On the other hand, cooperative spectrum sensing is a promising technology to effectively combat shadowing and multipath fading, which are the main factors affecting the detection accuracy of a single user. Thus, SU_k can select one or more partners among its neighbors to perform cooperative spectrum sensing, in order to improve the detection efficiency. More importantly, the partnership among SUs for cooperative spectrum sensing should be adapted to the varying network environment, so as to maximize the achievable cooperation gain. How to choose the proper partner is one of the key challenges for cooperative sensing in dynamic CR networks.

Therefore, the essential motivation of this work is to address two key issues in dynamic distributed CR networks, i.e., 1) what kind of scanning order of the primary channels should be adopted by the SU so that it can quickly access an available channel? 2) How to choose the proper partner for cooperative sensing to improve the detection efficiency when the detection capabilities of the other SUs are unknown? To this end, we focus on the distributed and self-learning channel selection and cooperation partner selection strategies in dynamic CR networks, aiming at improving the performance in terms of the scanning overhead, access delay, and detection efficiency.

IV. COOPERATIVE SPECTRUM SENSING SCHEME DESIGN

A. *Q-learning based Channel Selection Algorithm*

Q-learning is a widely used reinforcement learning algorithm. In Q-learning model, an agent in state $s \in S$ interacts with the environment by taking an action $a \in A$ and then obtains a reward $r(s,a)$ subsequently. Based on reward $r(s,a)$, the agent updates Q-value $Q(s,a)$ and transits to state s'. The agent learns from the state-action-reward history. Q-value $Q(s,a)$ is updated by:

Table 1. Main variables.

Variables	Description
K	Number of SUs
$\mathcal{N}_K$	Set of indexes of neighbors of SU_k
L	Number of primary channels
c_i	Primary channel i
$\hat{p}_t^k(j)$	Estimated detection probability of SU_k time t
$s^j(c_i)$	Status of channel c_i maintained by SU_j
$W_t^j(c_i)$	Aging weight of $s^j(c_i)$
$r^k(s_t, c_i)$	Reward of SU_k selecting channel c_i at time t
$Q^k(s_t, c_i)$	Q-value of channel c_i maintained by SU_k
α	Learning rate in Q-learning algorithm
R_t	Reward in D-UCB algorithm
γ	Discount factor in D-UCB algorithm
a_t^k	Cooperation partner selected by SU_k at time t

$$Q(s,a) \leftarrow (1-\alpha)\,Q(s,a) + \alpha\left\{r(s,a) + \beta \max_{b\in A}\left[Q(s',b)\right]\right\}, \quad (1)$$

where α is the learning rate, $0 \le \alpha \le 1$. With α closer to 0, the agent learns fewer from instant rewards and concentrates more on the history. β is the discount factor, $0 \le \beta \le 1$, denoting the attenuation of the rewards in future.

In the considered CR network, SUs can only access the vacant primary channels opportunistically. Thus, the first question is: What kind of scanning order of the primary channels should be adopted by the SU so that it can quickly access an available channel? Different from the traditional sequential scanning strategy, which may lead to high scanning overhead and access delay, we leverage Q-learning approach to guide the SUs to determine the scanning order of the primary channels, so as to reduce the scanning overhead and access delay. In the proposed Q-learning model for channel selection, state s_t indicates the status of the primary channels, i.e., whether the primary channels are occupied by the PUs at time t. SUs act as the agents to learn the occupancy pattern of the primary channels, so that they can access the vacant primary channels opportunistically without affecting the transmission of the PUs. When a call arises at SU_k, it takes an action by scanning a particular primary channel, e.g. channel c_i, and gets a real-valued reward evaluating the choice of the action, which is given by

$$r^k(s_t,c_i) = \begin{cases} 1 - \sum\limits_{j\in\mathcal{N}_k\cup\{k\}} \frac{\left(1-s^j(c_i)\right)*W_t^j(c_i)*\hat{p}_t^k(j)}{|\mathcal{N}_k|+1}, \\ \text{if} \quad s^k(c_i) = 1 \\ \\ -\sum\limits_{j\in\mathcal{N}_k\cup\{k\}} \frac{\left(1-s^j(c_i)\right)*W_t^j(c_i)*\hat{p}_t^k(j)}{|\mathcal{N}_k|+1}, \\ \text{if} \quad s^k(c_i) = 0, \end{cases} \quad (2)$$

where $\mathcal{N}_k$ represents the set of indexes of the neighbors of SU_k. $s^j(c_i)$ is the status of channel c_i maintained by neighbor SU_j, where $s^j(c_i) = 1$ if channel c_i was detected being idle and SU_j accessed it successfully, otherwise $s^j(c_i) = 0$. $W_t^j(c_i) = exp\left(-\left(t - t^j(c_i)\right)\right)$ is the aging weight of $s^j(c_i)$, where t is the current time, $t^j(c_i)$ is the time when SU_k obtained $s^j(c_i)$. $\hat{p}_t^k(j)$ is the detection weight of SU_j estimated by SU_k at time t,

representing the estimated detection probability of SU_j, which is given by (5).

Distinct from the work in [13], the devised reward function in (2) takes into consideration not only the timeliness of the information from the neighbors but also the detection ability of the neighbors, so that SUs can acquire the changes in channel status more efficiently, leading to a better estimation of the channel status. According to (2), the reward $r^k(s_t, c_i)$ is calculated in different ways depending on the value of channel status $s^k(c_i)$. Precisely, SU_k will get a positive reward if $s^k(c_i) = 1$ indicating it is able to access channel c_i successfully, otherwise a negative reward. To accelerate the learning procedure with a better view of the recent usage of the primary channels, the computation of reward is based on the channel status maintained by both SU_k and its neighbors, weighted by the user-specific aging and detection weights. It is worth mentioning that each SU maintains the detection weights of its neighbors locally, which are acquired by the proposed cooperation partner selection algorithm in the subsequent section.

After obtaining the reward, SU_k updates the corresponding Q-value by:

$$\begin{aligned} Q^k(s_{t+1}, c_i) = &(1-\alpha)\cdot Q^k(s_t, c_i) \\ &+ \alpha \cdot \left\{ r^k(s_t, c_i) - exp(-\tau \cdot m) \right\}, \end{aligned} \tag{3}$$

where τ is a constant system parameter, $0 \leq \tau \leq 1$. m represents that SU_k scans channel c_i at the mth attempt to find an idle channel. The augmented term $\exp(-\tau \cdot \mathrm{m})$ applies a bias to the channel with large value of m, thus increasing the probability that the channel with large value of m will be scanned preferentially. In this way, SUs can find the available primary channel more quickly, thereby reducing the number of attempts. Note that parameter τ controls the curvature of the function $\exp(-\tau \cdot \mathrm{m})$, thereby controlling the degree of difference in the value of the term $\exp(-\tau \cdot \mathrm{m})$ with respect to different m. The role of the augmented term in updating the Q-value depends on τ, and the optimal value of τ may vary with the network settings. To focus on the effect of m on the Q-value, we set τ equal to 1 in the performance evaluations as presented in Section V. As the learning process proceeds, SU_k keeps updating the Q-values of the primary channels. The larger the corresponding Q-value, the more likely the channel being available. Consequently, a scanning preference list of the channels is formed in decreasing order of the Q-values, which is adjusted dynamically as the Q-values are updated.

As mentioned before, when there is a demand at SU_k, it needs to choose a channel to scan according to a certain strategy. In Q-learning model, only the Q-value related to the selected channel is updated. Consequently, if some channels are not selected for a period of time, the corresponding Q-values cannot be updated promptly, such that these Q-values fails to accurately characterize the status of the channels. On the other hand, due to the time-varying property and randomness of the occupancy pattern of the primary channels, the channels which were occupied in the past period may become idle recently. In this case, selecting these channels is not only helpful for exploring other possible options, but also for updating the related Q-values promptly. Therefore, different from the work in [13] which always exploits the best strategy acquired so far while neglecting the exploration of other possible choices, we adopt ϵ-greedy strategy [11] as the action select selection strategy in order to strike a balance between exploitation and exploration. To be specific, SU_k selects the channel with the highest priority in the scanning preference list with a probability of $1-\epsilon$ in exploitation stage, whereas randomly selects a channel with a probability of ϵ in exploration stage. In this case, parameter ϵ, $0 \leq \epsilon \leq 1$, controls the degree of exploration versus exploitation. For small ϵ, SU_k is mainly focused on utilizing the best channel selection that has been performed so far to reduce scanning overhead and access delay. On the contrary, for large ϵ, SU_k is more inclined to look for other possible channels to improve the efficiency.

B. D-UCB based Cooperation Partner Selection Algorithm

Cooperative spectrum sensing is an effective way to overcome the shortcomings of single node sensing. It should be noted that if a SU cooperates with a partner with poorer detection ability, the partner may degrade the detection performance [9]. In order to improve the detection efficiency, each SU should select the ones with stronger detection ability as partners. But the more partners participant in cooperation, the higher overhead and complexity. Therefore, we propose that each SU selects the neighbor with the strongest detection ability as the cooperation partner to detect the primary channel of interest, so as to strike a balance between efficiency and overhead. However, due to the mobility of users and the time variation of wireless channels, the detection ability of each SU may change dynamically and is unknown to other SUs in the distributed network scenario. In this case, each SU needs to estimate the detection ability of its neighbors, so that it can select the proper cooperation partner to improve the detection efficiency. To this end, we model the partner selection algorithm as a D-MAB problem, wherein each SU estimates the detection probability of its neighbors in a dynamic situation, and learns about the cooperation partner selection strategies to maximize the detection efficiency.

MAB problem is one kind of RL problem wherein a gambler (agent) plays a slot machine (arm). At time t, by pulling arm $a \in A$, the agent gets the reward $R_t = 1$ with a probability of $p(a)$, otherwise $R_t = 0$. The agent learns to choose the most optimal arm among the available arms so as to maximize the total reward. In the cooperative spectrum sensing scenario considered in this paper, each SU acts as the learning agent, and its neighbors which are candidates for cooperation partner act as the arms of the bandit. At time t, SU_k chooses action $a = f$ by pulling arm f, representing that it selects SU_f as its cooperation partner. The resulting reward R_t relies on whether the selected cooperation partner can detect the channel correctly, which is given by

$$R_t = \begin{cases} 1, & if\ the\ detection\ result\ is\ right \\ 0, & otherwise. \end{cases} \tag{4}$$

Let $p_t^k(f)$ denote the expected reward of SU_k choosing SU_f as cooperation partner. According to (4), it can be seen that $p_t^k(f)$ actually represents the detection probability of SU_f at time t.

To address the cooperation partner selection in dynamic CR networks scenario, we model it as a D-MAB problem. Distinct

from MAB problem, D-MAB problem assumes the reward distribution is time varying, which is more suitable for the practical network environment. To be specific, in exploration stage, SU selects each neighbor fairly and estimates the detection probability of each neighbor based on the obtained reward. In this way, SU can get a better estimation of the detection probability at the cost of losing the opportunity to select the best neighbor. In exploitation stage, SU selects the potentially best neighbor as its cooperation partner based on the estimated detection probability to maximize the reward. In this case, it lacks exploration of other possible options. Consequently, the exploration-exploitation tradeoff is crucial for solving the D-MAB problem.

The discounted upper confidence bound (D-UCB) algorithm [23], [24] utilizes a discounted rate to address the exploration-exploitation tradeoff for the D-MAB problem. Based on D-UCB algorithm, we devised a cooperation partner selection algorithm for dynamic CR networks. Since the detection probability of SUs varies with time, the recent rewards play a more important role than the previous rewards in the estimation of detection probability. We can use a discount factor to give different weight to the reward obtained at different time. Thus, for SU_k, the estimated detection probability $\hat{p}_t^k(f)$ of neighbor SU_f at time t is updated by

$$\hat{p}_t^k(f) = \sum\nolimits_{s=1}^{t} \gamma^{t-s} R_s I_{\{a_s^k=f\}}, \tag{5}$$

where γ is the discount factor, $0 < \gamma < 1$. a_s^k is the action selection strategy, and $I_{\{a_s^k=f\}}$ is an indicator function, where $I_{\{a_s^k=f\}} = 1$ if SU_k selects SU_f as the cooperation partner at time s, i.e., a_s^k=f, otherwise $I_{\{a_s^k=f\}} = 0$.

To maximize the total reward, the partner selection strategy should take into account both the estimated detection probability and the exploration degree of the neighbors. At time t, SU_k selects the cooperation partner by the following rule:

$$a_t^k = \underset{f\in\mathcal{N}_k\cup\{k\}}{\operatorname{argmax}} \left[\hat{p}_t^k(f) + c\sqrt{\frac{\log n_t^k}{n_t^k(f)}}\right], \tag{6}$$

$$n_t^k(f) = \sum\nolimits_{s=1}^{t} \gamma^{t-s} I_{\{a_s^k=f\}}, \tag{7}$$

$$n_t^k = \sum_{f\in\mathcal{N}_k\cup\{k\}} n_t^k(f), \tag{8}$$

where $n_t^k(f)$ denotes the discounted number of times that SU_f is selected as the partner by SU_k up to time t. In (6), the first term is the estimated detection probability of SU_f, which is given by in (5). The second term represents the exploration degree of SU_f, and it is inversely proportional to the relative number of times that SU_f is selected as the cooperation partner. Accordingly, the higher the detection probability of SU_f, the more likely it will be selected as the partner. At the meanwhile, the fewer times SU_f was selected as the partner in the past, the larger the value of the second term, thus the more likely SU_f will be selected. The benefit of introducing the second term is that the cooperation partner selection strategy can fully explore all the possible options. c is a system parameter, which controls the degree of exploration versus exploitation. If c is set properly, a good balance between exploration and exploitation can be achieved. Note that if the estimated detection probability of SU_k is higher than that of all its neighbors, then SU_k will detect the selected primary channel by itself. Besides, if $n_t^k(f) = 0$, SU_f will be chosen firstly.

Table 2. RL enabled cooperative spectrum sensing scheme.

Input : The set of SUs, $W_{t-1}^k(c_i)$ and $Q^k(s_{t-1}, c_i)$ for all k and all c_i, $\hat{p}_{t-1}(f)$ for all k and $f \in \mathcal{N}_k \cup \{k\}$.
for each SU_k **do**
  if (a demand appears) **then**
    success = 0;attempt = 0;
    repeat
      Select a channel c_i by ϵ-greedy strategy based on the scanning preference list of channels;
      Select a cooperation partner SU_f with (6) - (8);
      if (SU_f detects correctly) **then**
        $R_t = 1$;
      else
        $R_t = 0$;
      end
      Update $\hat{p}_t^k(f)$ with (5);
      Calculate $W_t^j(c_i)$ for all $j \in \mathcal{N}_k \cup \{k\}$;
      if (SU_k accesses c_i correctly) **then**
        $s^k(c_i) = 1$;
        $r^k(s_t, c_i) = 1 - \sum_{j\in\mathcal{N}_k\cup\{k\}} \frac{\left(1-s^j(c_i)\right)*W_t^j(c_i)*\hat{p}_t^k(j)}{|\mathcal{N}_k|+1}$;
        success $= 1$;
      else
        $s^k(c_i) = 0$;
        $r^k(s_t, c_i) = -\sum_{j\in\mathcal{N}_k\cup\{k\}} \frac{\left(1-s^j(c_i)\right)*W_t^j(c_i)*\hat{p}_t^k(j)}{|\mathcal{N}_k|+1}$;
      end
      Update $Q^k(s_t, c_i)$ with (3);
      ++attempt;
    until success $= 1$ || attempt $=$ M
    if (success $= 0$)
      Declare call blocked;
    end
    Broadcast $s^k(c_i)$ to neighbors;
  end
end

It's worth mentioning that the proposed partner selection algorithm is not limited to the scenario of selecting one partner, but also can be extended to the scenario of selecting multiple partners. Let N_{coop} be the size of cooperation cluster for spectrum sensing, i.e., the number of SUs participating in channel detection. In case that multiple partners are considered, i.e., the case of $N_{coop} > 1$, SU_k can apply the selection rule in (6) to select the first N_{coop} neighbors (itself may be included) as the cooperation partners.

C. RL enabled Cooperative Spectrum Sensing Scheme

In summary, the proposed RL enabled cooperative spectrum sensing scheme consists of the aforementioned Q-learning based channel selection algorithm and the D-UCB based cooperation partner selection algorithm, as presented in Table 2. Particularly,

with the proposed Q-learning based channel selection algorithm, each SU learns the occupancy pattern of the primary channels, so that a scanning preference list can be formed accordingly to determine the detection order of the primary channels. Meanwhile, with the proposed D-UCB based cooperation partner selection algorithm, each SU estimates the detection probability of its neighbors in a dynamic situation, and selects the neighbor with the potentially highest detection probability as the partner to perform the cooperative spectrum sensing.

In practical implementation, the proposed RL enabled cooperative spectrum sensing scheme is performed in a distributed manner without any central controller. Specifically, it is executed at each SU individually, and only depends on the locally maintained information and limited information interaction with neighbors. In the dynamic distributed CR network considered in this paper, SUs exchange the management messages via the dedicated common control channel, whereas send and receive the data packets by opportunistically occupying the primary channels.

Taking SU_k for example, it maintains the status and Q-value of each primary channel thus the scanning preference list, as well as the estimated detection probability of itself and its neighbors. Moreover, it also collects the channel status maintained by its neighbors, which is received via the dedicated control channel. When there is a demand at SU_k, it selects a primary channel to detect and attempt to access by ϵ-greedy strategy based on the scanning preference list. Moreover, by applying the selection rule in (6), SU_k chooses SU_f as the cooperation partner, and sends a notification message to inform SU_f to detect the selected primary channel. Once SU_f finishes the detection, it reports the result to SU_k. Then, SU_k makes the final decision based on the received detection result. In case that multiple cooperation partners are selected, SU_k will send a notification message to each partner. The partners detect the selected primary channel individually, and report the results to SU_k. Then, SU_k fuses the detection results of the partners based on a certain fusion rule, such as the majority rule or the weighted rule [25], to make the final decision. If the final decision is that the primary channel is idle, SU_k tries to access the channel. Based on the detection and access result, SU_k updates the corresponding channel status, so that it can calculate the resulting reward with (2), and update the corresponding Q-value with (3) and the estimated detection probability with (5), respectively. If the selected primary channel is detected busy or SU_k fails to access it due to detection error, SU_k will reselect a channel among the residual channels and try it again by repeating the above procedures, until it successfully finds an idle channel. In case SU_k fails with a maximum number of attempts, the call is blocked. Finally, SU_k updates the scanning preference list based on the updated Q-values, and broadcasts the updated channel status to its neighbors via the dedicated control channel.

The time complexity of proposed channel selection and cooperative partner selection algorithms are $O(L)$ and $O(N)$, respectively. In the worst case, the time complexity for one call is $O(M(L+N))$. Here, M is the maximum number of attempts for each call before declaring a call block. L is the total number of the channels in primary network. N is the number of the neighbors.

Remark: It should be noted that the dynamic distributed CR network considered in this work is a non-stationary environment, wherein the occupancy pattern of the primary channels is presumed to be fixed, whereas the detection capability of each SU is presumed to change over time due to the mobility and the time variation of wireless channels. The effectiveness of the proposed algorithms in such a non-stationary environment depends on the same assumption as in [26], [27], i.e., we assume that the environment changes slowly enough such that on-line RL algorithms can be employed to keep track of the changes. Specifically, we assume the low mobility of the SUs, hence the changes in detection probability of the SUs are not frequent, such that each SU can learn the detection probability of its neighbors before they have changed. Under this assumption, the non-stationary environment in large time-scale can be viewed as consisting of a series of stationary environment contexts in small time-scale. And we focus on the solutions in the small time-scale environment context in this work.

In order to broaden the applicability of the proposed algorithms in non-stationary environments, the issues with assumption relaxation and developing RL algorithms for non-stationary environment models need to be addressed. Recently, this concern has led to diverse research efforts, ranging from hidden-mode MDP based varying environment modeling [26], to model-based method for detecting changes in environment models [27], to context detection based RL algorithm [28] and its extension [29], to repeated update Q-learning [30], and so on. Motivated by these prior works, we will investigate an extension to the proposed algorithms as a direct solution for non-stationary environment in our future research. One possible solution is to combine the change detection design of environment contexts with the algorithms proposed in this work. Besides, developing the proper learning rate adaption mechanism to speed up convergence and the issue with computational efficiency in non-stationary environment are the important topics along this line that deserve further research.

V. NUMERICAL RESULT

In this section, we evaluate the performance of the proposed cooperative spectrum sensing scheme. The performance metrics include average number of attempts, average call block rate, and average detection probability. Specifically, the average number of attempts represents the average number of times that a SU has tried for a successful access to the primary channel. The average call block rate means the ratio of the number of the call blocked due to access failure to the total number of the call requests generated at a SU. The average detection probability indicates the probability of correctly detecting the status of the channels.

A. Simulation Setup

It is assumed that time is discrete with fixed time unit. In the primary network, there are 10 channels, and the PU usage rate of each channel varies from 40% to 90% [13]. The call holding time is assumed to follow exponential distribution with mean value $\mu = 4$ time units for the PUs. In the CR network under consideration, there are 10 SUs, each of which has 4 neighbors. The traffic is generated following Poisson process with mean

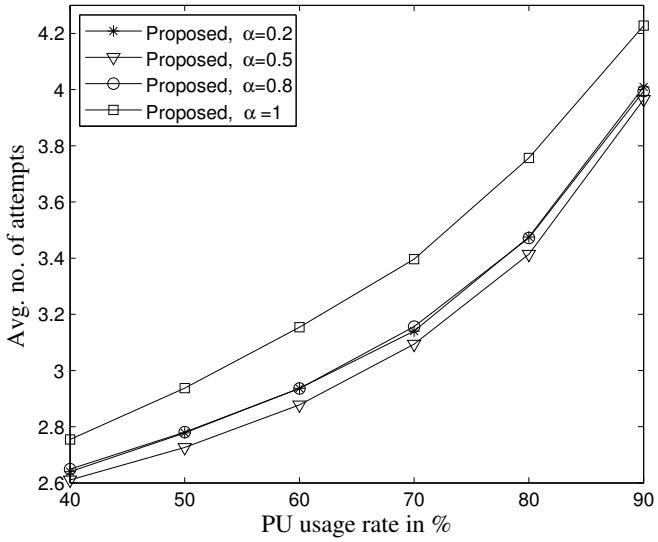

Fig. 2. Average number of attempts for different α.

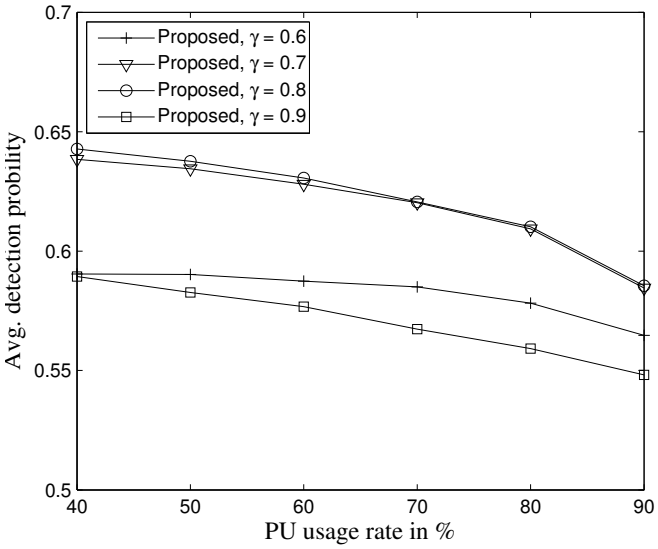

Fig. 3. Average detection probability for different γ.

arrival rate $\lambda = 2$ per time unit. The detection probability of each SU changes randomly every 10 time units with a range from 0 to 1, due to its mobility and the effects of fading and shadowing. It is assumed that the maximum number of attempts for a call is 5. In case that a SU fails to access a channel after 5 attempts, its call will be abandoned and announced blocked. In the following simulation, ϵ in ϵ-greedy strategy is set to 0.3. τ in (3) is set to 1. c in (6) is set to 0.8.

B. Effect of System Parameters on the Performance of the Proposed Algorithm

In the proposed channel selection algorithm, parameter α is the learning rate, which determines the accuracy of the prediction of channel status thus the effectiveness of the resulting scanning preference list. It should be noted that the optimal value of α may vary with the network settings. Fig. 2 shows the average number of attempts versus PU usage rate for different α. It can be seen from Fig. 2 that the different learning rate α leads to the different average number of attempts for a successful channel access. When α is set to 0.5, the proposed scheme achieves the least average number of attempts in the considered network scenario. It reveals that the agent can not obtain good learning result if it is only focused on the instant rewards or the history. Learning from both the instant rewards and the history can better perceive the changes in the channel status and obtain a more accurate scanning preference list consequently. Thus, SUs can find an idle channel more quickly so as to reduce the scanning overhead and access delay.

In the proposed cooperation partner selection algorithm, parameter γ controls the weight of the recent reward when learning the detection probability of the neighbors, thereby determining the validity of the selection of cooperation partner. Fig. 3 shows the average detection probability versus PU usage rate for different γ. It can be observed from Fig. 3 that different discount factor γ achieves different average detection probability. Overall, γ=0.8 is more suitable for the considered network scenario. This stems from the fact that the detection probability of each

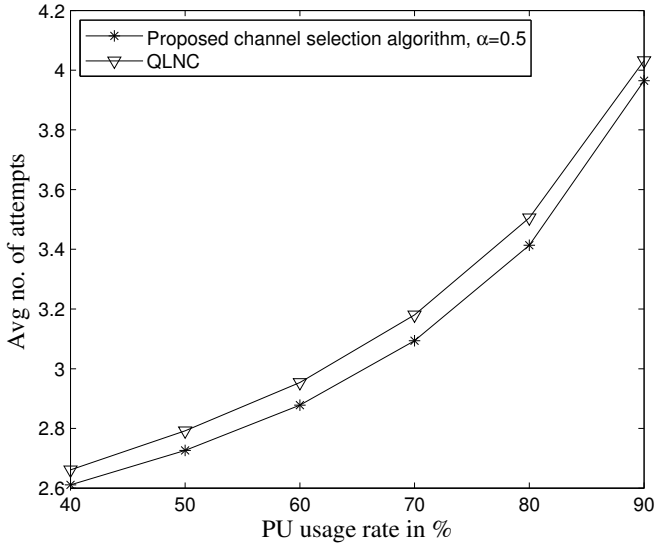

Fig. 4. Average number of attempts for different algorithms.

SU varies randomly every 10 time units in the simulation settings, that means a relatively rapid change in the reward distribution. In this case, γ taking a value close to 1 helps to estimate the dynamic detection probability of the neighbors more precisely. Therefore, each SU is more likely to choose the potential best neighbor as the cooperation partner, so as to improve the detection efficiency.

C. Performance Gain of the Proposed Scheme with Respect to the Baseline

To evaluate the performance of the proposed cooperative spectrum sensing scheme, we compare it with other two algorithms. The first reference algorithm is the one proposed in [13], denoted as "QLNC", wherein SUs select channel based on a Q-Learning approach without consideration of the cooperation among SUs. The second reference algorithm is denoted as "QLKN", which incorporates the "QLNC" algorithm with the K/N rule [31] to perform cooperative spectrum sensing. Accord-

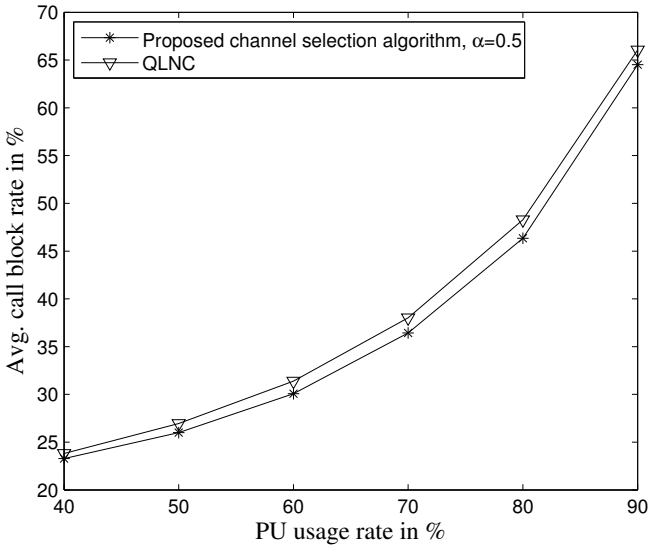

Fig. 5. Average call block rate for different algorithms.

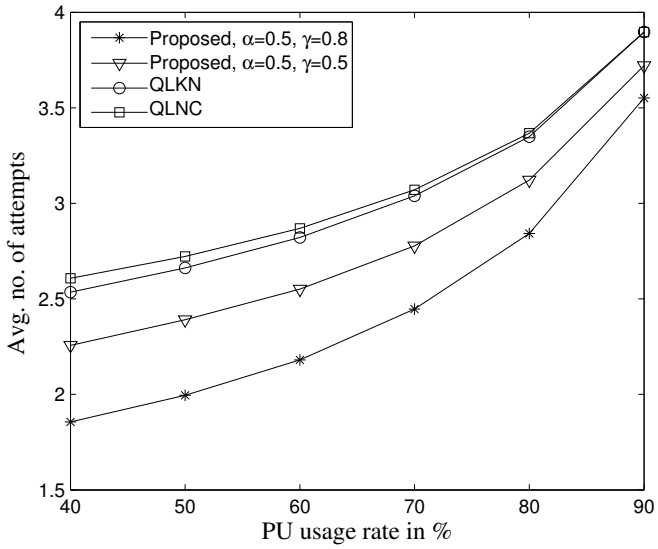

Fig. 6. Average number of attempts for different algorithms.

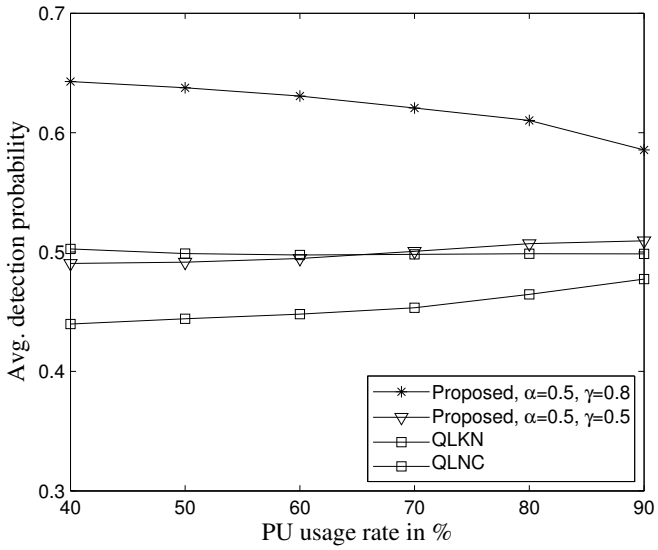

Fig. 7. Average detection probability for different algorithms.

ing to the results presented in Section B, parameters α and γ are set to be 0.5 and 0.8 in the following comparison, respectively.

Firstly, we compare the proposed channel selection algorithm with the "QLNC" algorithm in the case that the two algorithms combine with local detection rather than cooperative sensing. Figs. 4 and 5 show the average number of attempts and average call block rate versus PU usage rate for different algorithms respectively. It can be seen from Figs. 4 and 5 that both average number of attempts and average call block rate increase with PU usage rate. Intuitively, with the increase of PU usage rate, the possibility of choosing a busy channel in exploration stage increases accordingly, thus SUs need to try more times to access an idle channel even probably drop the call finally. More importantly, as shown in Figs. 4 and 5, the proposed channel selection algorithm outperforms the "QLNC" algorithm in all the cases. The reason is twofold. First, it is attributed to the devised reward function in (2), which takes into consideration not only the timeliness of the information from the neighbors but also the detection ability of the neighbors, so that SUs can acquire the changes in channel status more efficiently, leading to a better estimation of the channel status. In contrast, the "QLNC" algorithm only considers the timeliness of information from the neighbors. Second, it is owing to the ϵ-greedy strategy, so that SUs can choose the potentially best channel meanwhile exploring other possible options. As a result, with the proposed channel selection algorithm, SUs can reduce the number of attempts and the call block rate, thus reducing the scanning overhead and access delay.

Then, we compare the performance of the proposed cooperative spectrum sensing scheme with the "QLNC" and "QLKN" algorithms in terms of average number of attempts, average detection probability, and average call block rate, as shown in Figs. 6–8, respectively. Fig. 6 shows the average number of attempts versus PU usage rate for different algorithms. It can be seen that the average number of attempts increases with PU usage rate in all algorithms. Given a PU usage rate, the number of attempts required to find an idle channel mainly relies on the accuracy of the prediction of channel status. The more accurate the prediction, the less number of attempts. As described in the analysis of Fig. 4, the proposed scheme is more sensitive to the changes in channel status, thus obtaining more accurate knowledge about the channel status compared to the "QLNC" and "QLKN" algorithms. Consequently, the proposed scheme achieves the least average number of attempts, so as to reduce the scanning overhead and access delay. Moreover, the proposed scheme introduces cooperative spectrum detecting in addition to the Q-learning based channel selection, whereas the "QLNC" algorithm does not consider any cooperation among SUs when detecting channels. As a result, the superiority of the proposed scheme over the "QLNC" algorithm is increased significantly compared to the results in Fig. 4. Additionally, since the proposed algorithms with different parameters achieve different performance as shown in Figs. 2 and 3, there is a gap between the case with $\alpha = 0.5$ and $\gamma = 0.8$ and the case with $\alpha = 0.5$ and $\gamma = 0.5$.

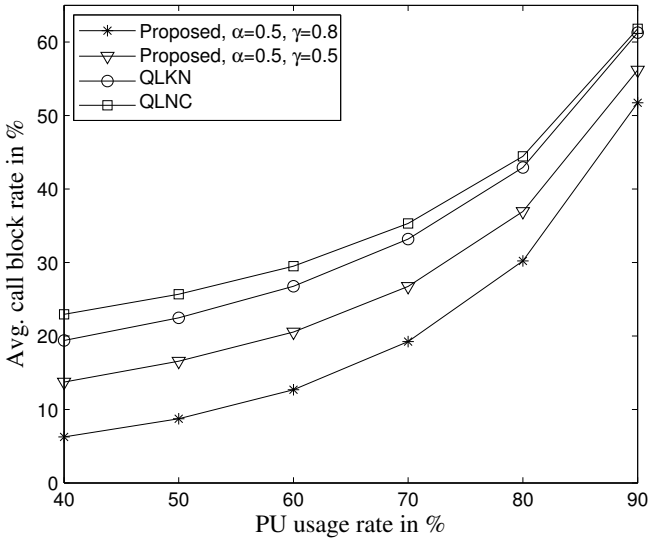

Fig. 8. Average call block rate for different algorithms.

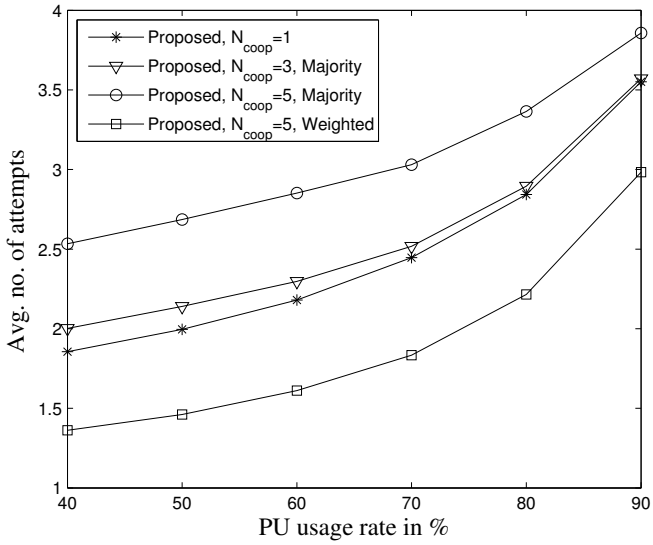

Fig. 9. Average number of attempts for different N_{coop} and fusion rules.

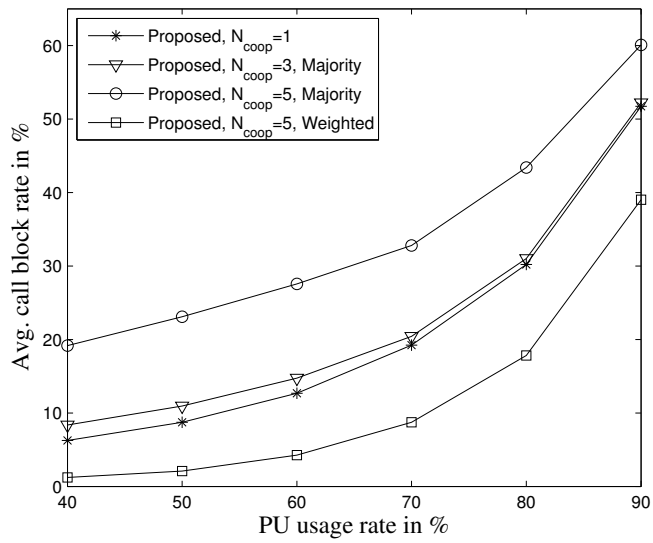

Fig. 10. Average call block rate for different N_{coop} and fusion rules.

Fig. 7 shows the average detection probability versus PU usage rate for different algorithms. It can be seen from Fig. 7 the proposed scheme with $\alpha = 0.5$ and $\gamma = 0.8$ outperforms the other two reference algorithms evidently. On one side, only local detection rather than cooperative detection is utilized in the "QLNC" algorithm, while cooperative detection is employed in both the proposed and "QLKN" algorithms. Therefore, the "QLNC" algorithm attains the lowest detection probability. On the other side, in the proposed scheme, each SU learns the dynamic detection probabilities of its neighbors, and selects the neighbor with the highest estimated detection probability as the partner to detect the channel cooperatively, whereas the "QLKN" algorithm fuses the detection results of the neighbors based on the "K out of N" rule to make the final decision. As a result, the proposed scheme can significantly improve the detection efficiency, while the detection efficiency of the "QLKN" algorithm may be degraded due to the poor detection capabilities of some neighbors.

Fig. 8 shows the average call block rate versus PU usage rate for different algorithms. Consistent with the results in Fig. 6, the average call block rate increases with the PU usage rate in all algorithms, and the proposed cooperative spectrum sensing scheme attains the lowest average call block rate compared to the "QLNC" and "QLKN" algorithms. With the increase of the PU usage rate, the number of available channels decrease, thus resulting in the increase of the average call block rate. Owing to the advantages in the number of attempts and detection probability as shown in Figs. 6 and 7, the proposed scheme is significantly superior to the reference algorithms in terms of call block rate, so that the proposed scheme is able to provide better service guarantees for SUs.

D. Effect of the Size of Cooperation Cluster on the Performance of the Proposed Scheme

In the previous simulation experiments, we are focused one the case of $N_{coop} = 1$. Next, we evaluate the effect of N_{coop} on the performance of the proposed cooperative spectrum sensing scheme. In this experiment, when there is a demand at SU_k, it selects the first N_{coop} neighbors as the cooperation partners according to (6)–(8). The selected N_{coop} partners detect the channel of interest individually, and report the results to SU_k. SU_k fuses the received detection results to make the final decision based on the majority rule or the weighted rule [25]. When fusing based on the weighted rule, the detection results are weighted based on the corresponding estimated detection probability.

Figs. 9–11 show the average number of attempts, average call block rate, and average detection probability versus PU usage rate for different N_{coop} and fusion rules, respectively. Note that the case of $N_{coop} = 1$ can be considered as a special case of fusing the detection result of one partner based on the majority rule. As shown in Figs. 9–11, the performance of the proposed cooperative spectrum sensing scheme decreases with the increase of the size of cooperation cluster when fusing based on the majority rule. This is due to fact that the partner with poor

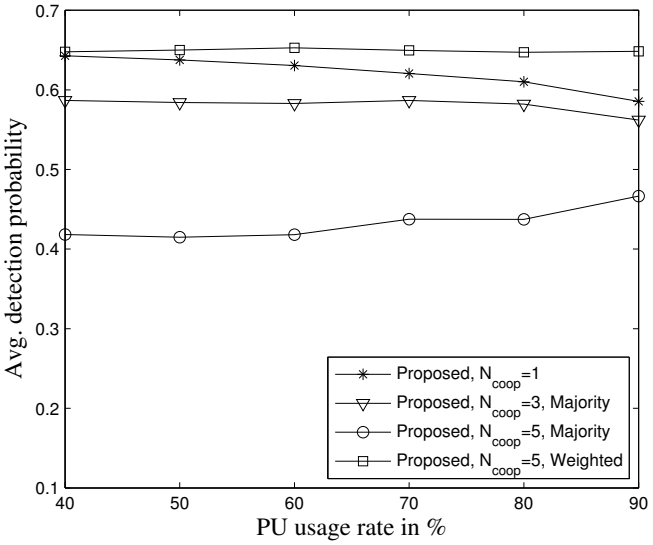

Fig. 11. Average detection probability for different N_{coop} and fusion rules.

detection capability will degrade the accuracy of the final decision, thus lowering the detection efficiency of the cooperation cluster accordingly. On the other hand, with the consideration of the difference in detection capabilities of the different partners, weighting the detection results based on the detection probability can improve the accuracy of the final decision, thereby achieving higher cooperation gain. Therefore, the performance obtained by the weighted rule is superior to that obtained by the majority rule. It reveals that the performance of the proposed scheme can be further improved by introducing more partners and adopting appropriate fusion rule, at the cost of an increase in overhead.

VI. CONCLUSION

In this paper, we proposed a RL enabled cooperative spectrum sensing scheme, consisting of the Q-learning based channel selection algorithm and the D-UCB based cooperation partner selection algorithm. With the proposed scheme, SUs learn the occupancy pattern of the primary channels to select a proper channel to access, meanwhile SUs also learn the dynamic detection probability of the neighbors to choose the cooperation partner. Simulation results demonstrate that the proposed scheme achieves less number of attempts, higher detection probability, and lower call block rate compared with the reference algorithms, thus reducing the scanning overhead and access delay while improving the detection efficiency.

REFERENCES

[1] B. Wang and K. J. R. Liu, "Advances in cognitive radio networks: A survey," *IEEE J. Sel. Topics Signal Process.,* vol. 5, no. 1, pp. 5–23, May. 2011.

[2] S. Haykin, D. J. Thomson, and J. H, "Spectrum sensing for cognitive radio," *Proc. IEEE,* vol. 97, no. 5, pp. 849–877, May. 2009.

[3] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Elsevier Science Publishers B. V.,* 2011.

[4] S. M. Mishra, A. Sahai, and R. W. Brodersen, "Cooperative sensing among cognitive radios," in *Proc. IEEE ICC,* June 2006, pp. 1658–1663.

[5] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS J. Comput.* vol. 21, no. 2, pp. 178–192, 2009.

[6] Y. H. Fu, Z. M. He, and F. Yang, "A simple quantization-based multibit cooperative spectrum sensing for cognitive radio networks," in *Proc. IEEE ICCWAMTIP,* Feb. 2018, pp. 220–223.

[7] M. Gupta and G. Yerma, "Improved weighted cooperative spectrum sensing algorithm based on reliability in cognitive radio networks," in *Proc. IEEE RTEICT.,* Jan. 2017, pp. 609–612.

[8] M. Gupta and G. Yerma, "Cooperative spectrum sensing for cognitive radio based on adaptive threshold," in *Proc. IEEE CICT,* Aug. 2016, pp. 444–448.

[9] Y. Zheng, X. Xie X, and L. Yang, "Cooperative spectrum sensing based on SNR comparison in fusion center for cognitive radio," in *Proc. IEEE ICACC,* May. 2009, pp. 212–216.

[10] Z. Li, W. Wu, X. Liu, and P. Qi, "Improved cooperative spectrum sensing model based on machine learning for cognitive radio networks," *IET Commun.,* vol. 12, no.19, pp. 2485–2492, Nov. 2018.

[11] Watkins, J. C. H. Christopher, and P. Dayan, "Q-learning," *Machine Learning,* Springer, vol. 8, pp. 279–292, May 1992.

[12] B. F. Lo and I. F. Akyildiz, "Reinforcement learning-based cooperative sensing in cognitive radio ad hoc networks," in *Proc. IEEE PIMRC,* Sept. 2010, pp. 2244–2249.

[13] A, Das, S. C. Ghosh, and N. Das, "Q-Learning based cooperative spectrum mobility in cognitive radio networks," in *Proc. IEEE LCN,* Oct. 2017, pp. 502–505.

[14] V. Raj *et al.*, "Spectrum access in cognitive radio using a two-stage reinforcement learning approach," *IEEE J. Sel. Topics Signal Proc.,* vol. 12, no. 1, pp. 20–34, Feb. 2018.

[15] J. Zhu *et al.*, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of things," *IEEE Internet Things J.,* vol. 5, no. 4, pp. 2375–2385, Aug. 2018.

[16] H. J. Cao and J. Cai, "Distributed opportunistic spectrum access in an unknown and dynamic environment A stochastic learning approach," *IEEE Trans. Veh. Technol.,* vol. 67, no. 5, pp. 4454–4465, May. 2018.

[17] F. Azmat, Y. Chen, and N. Stocks, "Analysis of spectrum occupancy using machine learning algorithms," *IEEE Trans. Veh. Technol.,* vol. 65, no. 9, pp. 6853–6860, Sept. 2016.

[18] S. Vakili *et al.*, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE J. Sel. Topics Signal Proc.,* vol. 7, no. 5, pp. 759–767, Oct. 2013.

[19] B. Li, *et al.,* "Almost optimal dynamically-ordered channel sensing and accessing for cognitive net-works," *IEEE Tran. Mobile Comput.,* vol. 13, no. 10, p. 1, 2014.

[20] Y. Kawaguchi, et al., "Adaptive boolean compressive sensing by using multi-armed bandit," in *Proc. IEEE ICASSP,* Mar. 2016, pp. 3261–3265.

[21] T. Kato, et al., "Application of multi-armed bandit algorithms for channel sensing in cognitive radio," in *Proc. IEEE APCCAS,* Jan. 2013, pp. 503–506.

[22] K. Cohen, et al., "Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access," in *Proc. IEEE ACSSC,* Apr. 2015, pp. 1575–1578.

[23] M. Niimi and T. Ito, "Budget-limited multi-armed bandit problem with dynamic rewards and proposed algorithms," in *Proc. IEEE IIAI-AAI,* 2015, pp. 540–545.

[24] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," *J. Machine Learning Research,* Jan. 2014.

[25] A. Ali and W. Hamouda, "Advances on spectrum sensing for cognitive radio networks: Theory and applications," *IEEE Commun. Surveys Tuts.,* vol. 19, no. 2, pp. 1277–1304, 2017.

[26] S. P. M. Choi, D. Yeung, and N. L. Zhang, "Hidden-mode Markov decision processes for nonstationary sequential decision making," *Sequence Learning*, Springer, 2001, pp. 264–287.

[27] T. Banerjee, M. Liu, and J. P. How, "Quickest change detection approach to optimal control in markov decision processes with model changes," in *Proc. IEEE ACC*, May 2017, pp. 399–405.

[28] B. C. da Silva, *et al.*, "Dealing with non-stationary environments using context detection," in *Proc. ACM ICML*, 2006, pp. 217–224.

[29] E. Hadoux, A. Beynier, and P. Weng, "Sequential decision-making under non-stationary environments via sequential change-point detection," *LMCE*, Sept. 2014.

[30] S. Abdallah and M. Kaisers,"Addressing environment non-stationarity by repeating q-learning updates," *J. Machine Learning Research,* vol. 17, no. 46, pp.1–31, 2016.

[31] Q. Qin *et al.*, "A study of data fusion and decision algorithms based on cooperative spectrum sensing," in *Proc. IEEE FSKD,* Aug. 2009, pp. 14–16.

Wenli Ning received her B.Sc. degree on Communication Engineering from Chongqing University, in 2017. She is now a graduate student at the University of Electronic Science and Technology of China. Her research focuses on spectrum sensing and wireless networking.

Xiaoyan Huang received her Ph.D. degree from the University of Electronic Science and Technology of China, China, in 2012. She is currently an Associate Professor with the School of Information & Communication Engineering, University of Electronic Science and Technology of China. Her research interests are generally in wireless communications and networking, with a focus on cross-layer design and optimization for broadband wireless networks.

Kun Yang received his Ph.D. from the Department of Electronic & Electrical Engineering of University College London (UCL), UK, and MSc and BSc from the Computer Science Department of Jilin University, China. He is currently a Chair Professor in the School of Computer Science & Electronic Engineering, University of Essex, leading the Network Convergence Laboratory (NCL), UK. He is also an Affiliated Professor at UESTC, China. Before joining in University of Essex at 2003, he worked at UCL on several European Union (EU) research projects for several years. His main research interests include wireless networks, future Internet technology and network virtualization, mobile cloud computing and networking. He manages research projects funded by various sources such as UK EPSRC, EU FP7/H2020 and industries. He has published 100+ journal papers. He serves on the editorial boards of both IEEE and non-IEEE journals. He is a Senior Member of IEEE (since 2008) and a Fellow of IET (since 2009).

Fan Wu received his Ph.D. degree from the University of Electronic Science and Technology of China, China, in 2015. He is currently an Associate Professor with the School of Information & Communication Engineering, University of Electronic Science and Technology of China. His research interests include broadband wireless access networks, vehicular networks, and wireless sensor networks.

Supeng Leng received his Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2005. He is currently a Professor in the School of Information & Communication Engineering, University of Electronic Science and Technology of China. His research focuses on resource, spectrum, energy, routing and networking in wireless sensor networks, broadband wireless access networks, smart grid, and vehicular networks. He published over 100 research papers in recent years. He serves as an organizing committee chair and TPC member for many international conferences, as well as a reviewer for over 10 international research journals.