Jang et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:131
https://doi.org/10.1186/s13638-019-1433-1

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Reinforcement learning-based dynamic band and channel selection in cognitive radio ad-hoc networks

Sung-Jeen Jang[1], Chul-Hee Han[2], Kwang-Eog Lee[3] and Sang-Jo Yoo[1*]

Abstract

In cognitive radio (CR) ad-hoc network, the characteristics of the frequency resources that vary with the time and geographical location need to be considered in order to efficiently use them. Environmental statistics, such as an available transmission opportunity and data rate for each channel, and the system requirements, specifically the desired data rate, can also change with the time and location. In multi-band operation, the primary time activity characteristics and the usable frequency bandwidth are different for each band. In this paper, we propose a Q-learning-based dynamic optimal band and channel selection by considering the surrounding wireless environments and system demands in order to maximize the available transmission time and capacity at the given time and geographic area. Through experiments, we can confirm that the system dynamically chooses a band and channel suitable for the required data rate and operates properly according to the desired system performance.

Keywords: Reinforcement learning, Cognitive radio, Ad-hoc network, Q-learning, Fairness

1 Introduction

As the demand for multimedia services increases, the problem of the frequency shortage continues to increase. The spectrum auction price is rising worldwide and passing on to users as a burden [1]. The Federal Communications Commission (FCC) had found that most of the spectrums are underutilized under its current fixed spectrum allocation [2]. The FCC had therefore proposed a new paradigm which provides an access to the spectrum resources not being used by the licensed user to resolve the increasing demand for the spectral access and inefficiency in use [3]. The cognitive radio (CR) technologies provide an opportunity for secondary users (SUs) to use spectrums that are not used by primary users (PUs), allowing the SUs to access the spectrum by adjusting their operational parameters [4, 5]. In relation to the application of CR, FCC adopted rules in April 2012 in [6] to allow license-exempt devices employing the TV white space database approach to access available channels in the UHF television bands. [7] presents the existing, emerging, and potential applications employing CRS capabilities and the related enabling technologies, including the impacts of CR technology on the use of spectrum from a technical perspective. The U.S. Defense Advanced Research Projects Agency (DARPA) and British defense firm BAE Systems are developing a CR IC technology for next-gen communications [8]. DARPA is developing CR technologies that maintain communications under severe jamming environment by Russian electronic warfare systems from 2011 [9]. In 2016, DARPA launched the Spectrum Collaboration Challenge (SC2) to resolve the scarcity of spectrum for DoD use and a Vanderbilt team won the round 1 [10].

The CR technology enables SUs to use free spectrum holes in radio environments that vary with a time and location. When the spectrum is used by a SU, quality of service (QoS) for both the PU and SU should be maintained by ensuring the spectrum accessibility for the SU without interfering with the service for the PU through the spectrum sensing. The SU should periodically sense the channel while using the channel and switch to another channel when the PU starts accessing the current channel. In this case, when selecting a channel, it is necessary to consider the fact that the frequency resource varies depending on the time and geographical area.

* Correspondence: sjyoo@inha.ac.kr
[1]Department of Information and Communication Engineering, Inha University, 253 YongHyun-dong, Nam-gu, Incheon, South Korea
Full list of author information is available at the end of the article

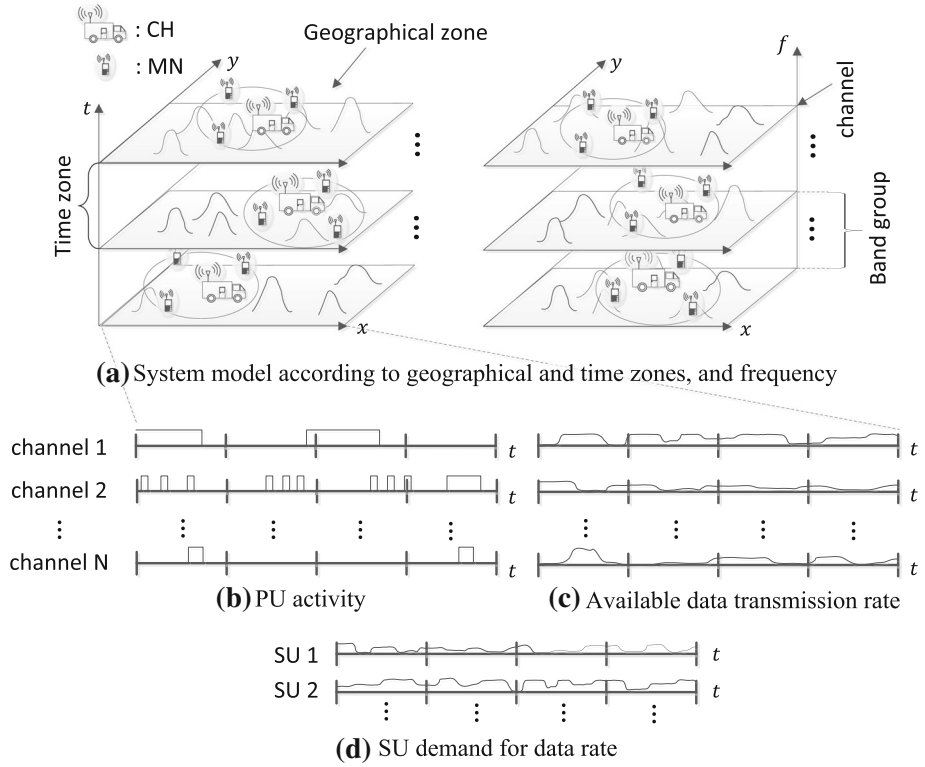


Fig. 1 Network model according to geography, time, and frequency. The cognitive radio ad-hoc network consists of a cluster head (CH) and mobile nodes (MNs) as shown in **a**. According to the time zone, band group, and each channel, the frequency resource is different. The PU activity, available data transmission rate, and SU demand for data rate vary according to the time and channel

Also, the CR system should consider the available data rate and possible channel acquisition time that can be achieved on each channel to guarantees the QoS of the SUs. Generally, depending on operating frequency bands such as HF (high frequency), VHF (very high frequency), and UHF (ultra-high frequency), a channel may have different channel bandwidths and the channel characteristic is different. Primary systems that are operating on different frequency bands also have diverse features and characteristics in terms of medium access mechanism, service types, and power requirements. Therefore, for choosing the best channel among the available frequency bands when the secondary CR network needs to move to another channel, several dynamic aspects such as primary system operation characteristics, radio channel conditions, frequency band characteristics, and secondary system requirements should be considered. We have to utilize a dynamic spectrum selection mechanism by considering the related environment and operational parameters to maximize the system performance. The channel access pattern of the PU, the requested data rate of the SU, and the available data rate and spectrum acquisition time can all vary dramatically according to environments. Therefore, the learning algorithm is required to dynamically solve these complex optimization problems.

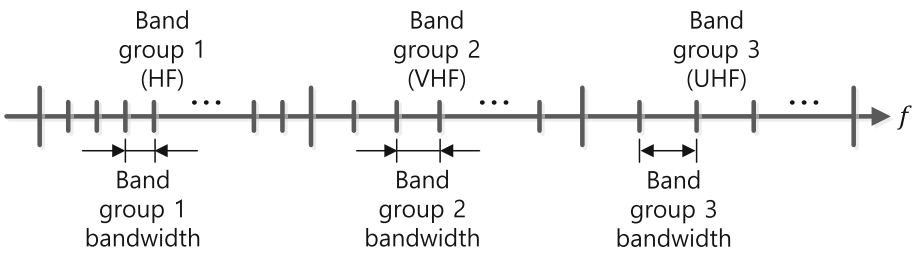


Fig. 2 Channel bandwidths for different band groups. A wireless communication system generally has a higher bandwidth and a higher data transmission rate as it goes to a higher band

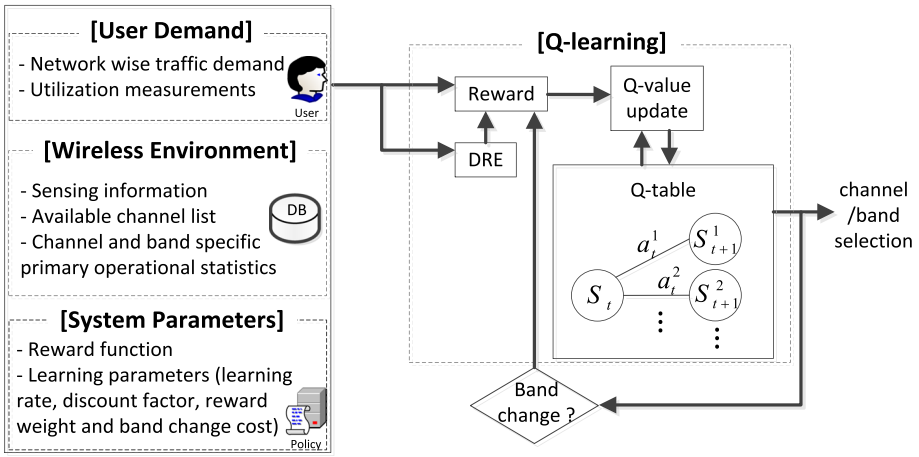


Fig. 3 Proposed system architecture. Proposed Q-learning is used to dynamically select the optimal band group and channel. As the reward function, the system considers the user demand, wireless environment and system parameters. The user demand module determines the desired data rate (DDR) of the CR ad-hoc network and measures the average utilization of the channel currently used. The wireless environment module stores the spectrum sensing results. The system parameters module is used to establish the reward function and Q-learning parameters. If the band of newly selected channel is different with the old one, the overhead for band group change is adopted to the reward function

In this paper, we propose an optimal band and channel selection mechanism in the cognitive radio ad-hoc network using the reinforcement learning. In a cluster-based CR ad-hoc network, we assumed that each member node (MN) performs a wide-band spectrum sensing periodically and reports the sensing results to the cluster head (CH) node. Based on the sensing results from the member nodes and previous channel history, the CH builds wireless channel statistic data vectors in terms of achievable data rates and average primary operational activation time (idle and busy) for each available channel of each conducted band. In addition, the CH estimates the traffic demand of the current cluster network to select a set of band and channel that provides the appropriate service to the cluster. Therefore, in CR ad-hoc networks, multiple clusters can operate in a limited area so that coexistence between ad-hoc clusters should be carefully considered in the channel selection. It is desired that if an ad-hoc cluster traffic demand is low, then the CH should select the frequency band that has relatively a narrow bandwidth (i.e., low achievable data rate). It yields the frequency band with wider bandwidth to another cluster network that needs higher traffic demand. In the proposed architecture, as a reinforcement learning, we use the Q-learning algorithm and we have designed a reward function that captures the expected consecutive operational time, affordable data rate, efficiency of spectrum utilization use, and band change overhead. In particular, the reward for channel spectrum utilization is proposed to reflect the degree of efficiency about using the supportable capacity. Using the proposed Q-learning, the CH can select an optimal band and channel that can maximize the multi-objective function of the CR network, and also, it can increase the coexistence efficiency of the overall secondary systems.

The main contributions of the proposed system architecture are as follows:

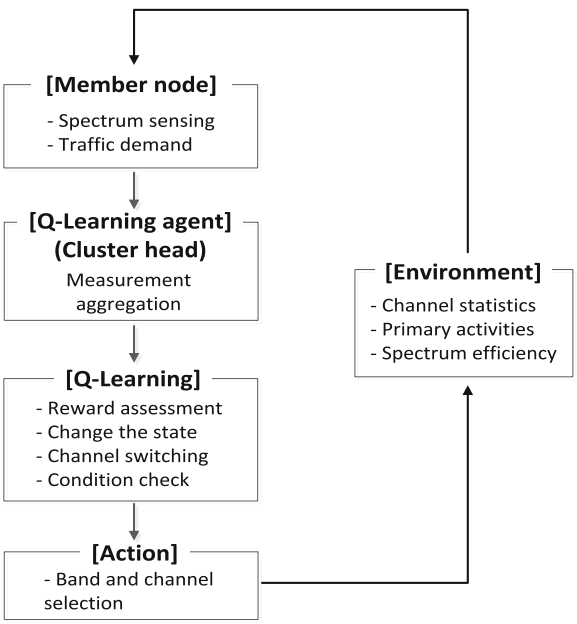


Fig. 4 Proposed Q-learning mechanism. The CH of the ad-hoc CR system is the agent of Q-learning, and the action is a selection of a tuple (band group and channel) when the PU is detected on the current band group and channel. The Q-learning agent (CH) designates the state from the information of member node and statistics of environment by the last action. From the Q-learning module, the Q-learning agent obtains the reward, change the Q-table and next action tuple

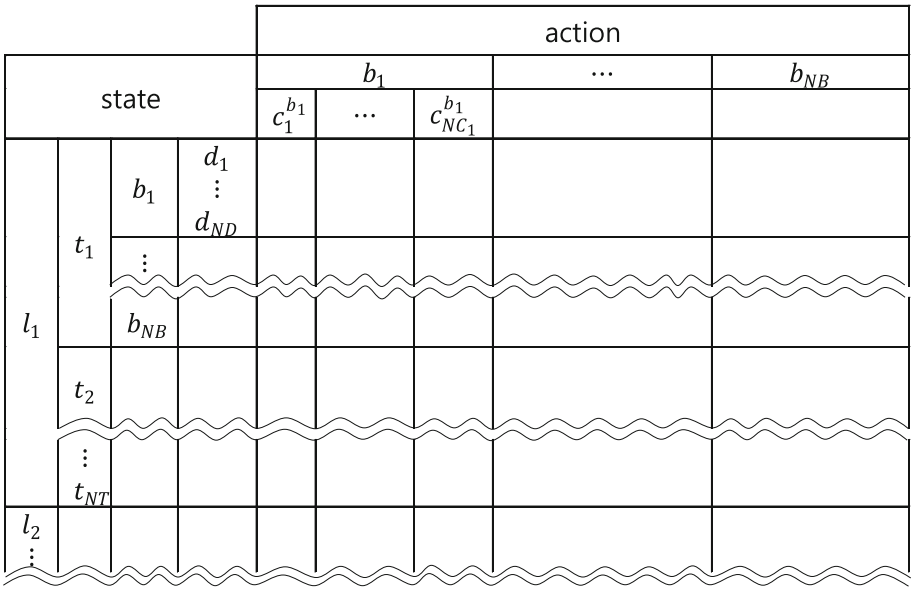


Fig. 5 Proposed Q-table structure. The column of the Q-table represents the action tuple of the band group b_q (q-th band group) and channel $c_m^{b_q}$ (m-th channel of b_q). The row of the Q-table is the state tuple of the i-th geographic location zone (l_i), j-th time zone (t_j), k-th band group (b_k), and l-th data rate efficiency level (d_l)

●We propose a new CR system architecture that maximizes the secondary user's service quality by dynamically selecting the optimal operating band and channel with consideration of the traffic demand of each CR system and the channel statistics according to the primary systems;

●We define states and actions in order to operate Q-learning considering the service state and demand of the corresponding systems, and propose a structural algorithm for it;

● We design a reward function that maximizes operating time, data rate, and channel utilization efficiency and minimizes band change overhead for secondary systems;

● The proposed system provides fairness by assigning the band and channel that are appropriate to each

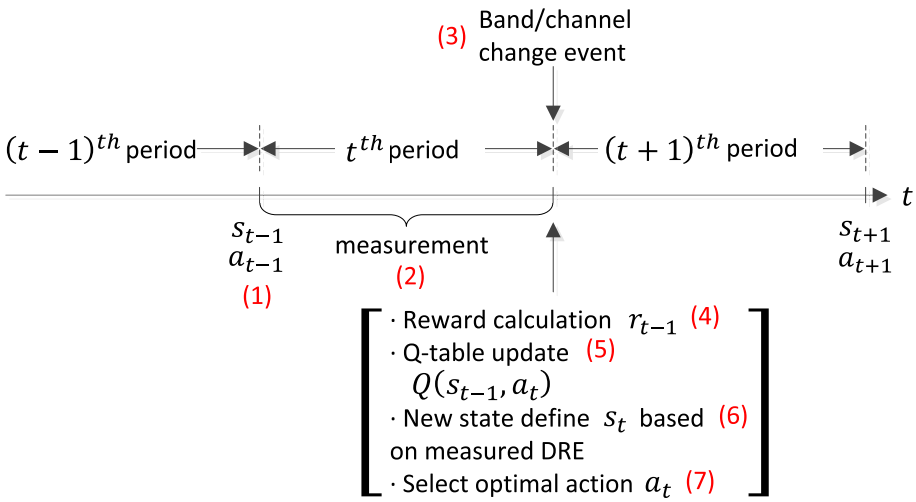


Fig. 6 Proposed procedure for Q-table update, state determination, and action selection. (1) Suppose the learning agent CH determined the state s_{t-1} and the best action a_{t-1} at the end of $(t-1)$-th time period. (2) During t-th time period, MNs and CH monitor the primary activities and channel statistics. (3) Agent CH detects the band and channel change event. (4) The CH calculates the reward r_{t-1} for the previous action a_{t-1} at state s_{t-1}. (5) The CH updates the Q-value of (s_{t-1}, a_{t-1}) in Q-table. (6) The CH determines the current state s_t based on the measured DRE during t-th time period. (7) The CH selects the optimal action a_t for the next $(t+1)$-th time period. (8) Go to step 1

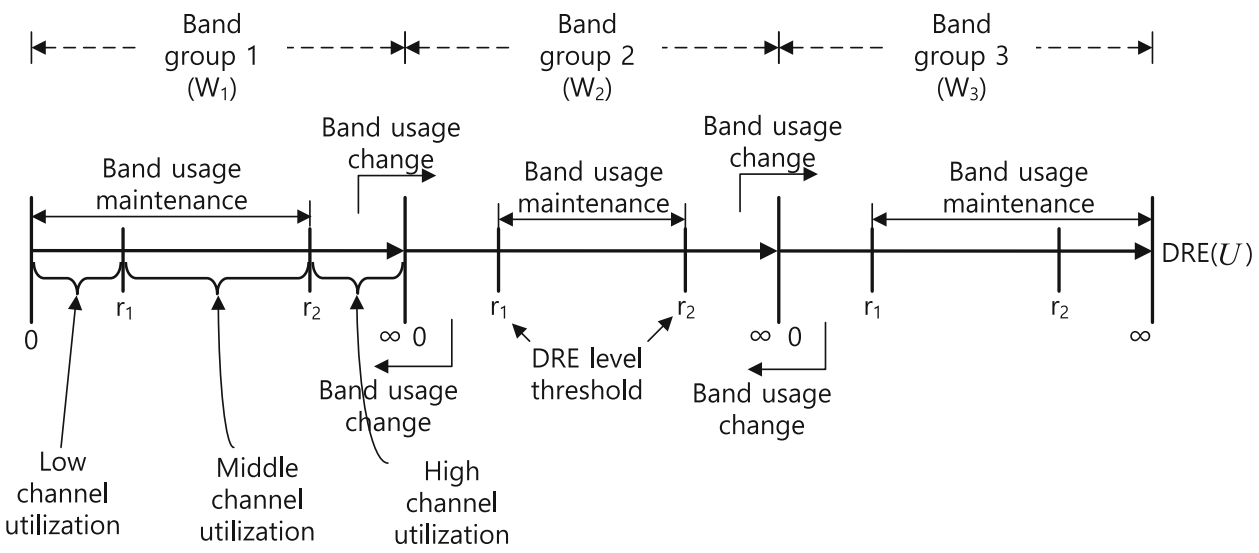


Fig. 7 Band group movement mechanism according to DRE. It is assumed the bandwidth of each channel provided by each band group follows $W_1 < W_2 < W_3$. The *x*-axis represents the DRE of each band group. The DRE is low, moderate, and high if it belongs to $[0, r_1)$, $[r_1, r_2)$, and $[r_2, \infty)$. If the DRE is low, the selection of the band need to be changed to lower one since the selected band supports too much bandwidth. However, the CH could not change the band in the range of $[0, r_1)$ in band group 1 since there is no band to move. If the DRE is high, the selection of the band need to be changed to a higher one since the selected band does not support the DDR. However, the CH could not change the band in the range of $[r_2, \infty)$ in band group 3 since there is no band to move. In the range of $[r_1, r_2)$ in each band group, the system does not need to change the band since the selected band supports adaptive bandwidth. Therefore, the whole DRE region is divided by the region of "Band usage maintenance" and "Band usage change"

secondary system based on its demand so that neighboring secondary systems coexist successfully.

The remainder of this paper is organized as follows. In Section 2, we describe related studies. In Section 3, we illustrate the system model and the tasks to be solved. In Section 4, we provide the proposed Q-learning algorithm to select the optimal operating band and channel. Section 5 contains simulation results, and conclusions are given in Section 6.

2 Related works

As the CR-based ad-hoc network is often deployed in situations where resources are insufficient, it is necessary to carefully consider the frequency resource selection. In

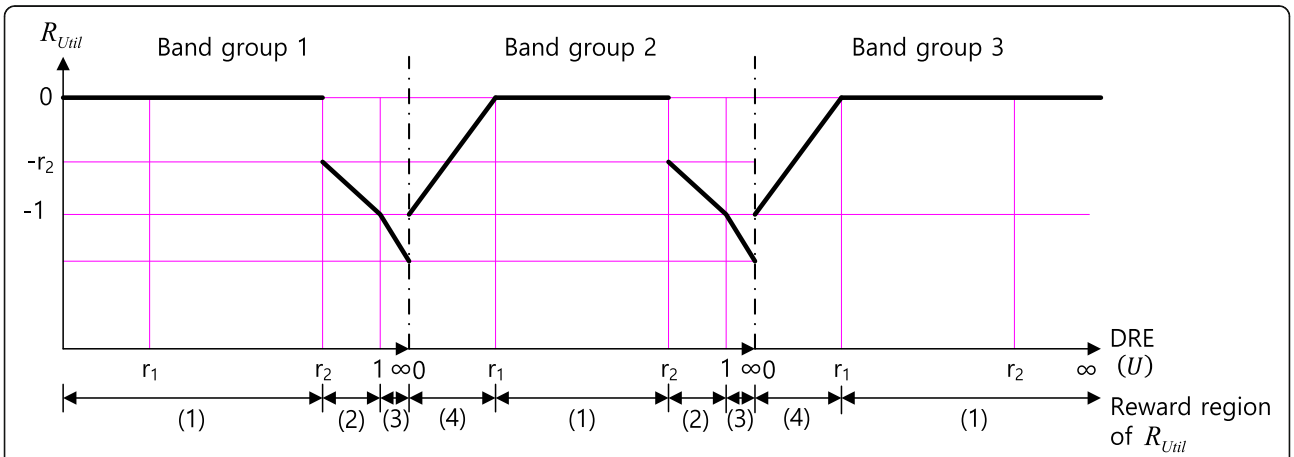


Fig. 8 Utilization efficiency reward according to DRE. To realize the mechanism in Fig. 7, the reward for channel utilization efficiency (R_{Util}) is designed as shown. The *x*-axis for each band group represents the DRE, and the *y*-axis represents R_{Util}. For the band usage maintenance range, R_{Util} is set to 0. For the band usage change range where the DRE is low, R_{Util} increases from −1 to 0 since the DRE represents better value as DRE increases. The band usage change range where the DRE is high is divided into $[r_2, 1)$ and $[1, \infty)$ to distinguish the insufficient transmission rate provided by the channel. R_{Util} represents a more rapid decrease rate in $[1, \infty)$. R_{Util} decreases from $-r_2$ to −1 in $[r_2, 1)$ range and from −1 to ∞ in $[1, \infty)$ range

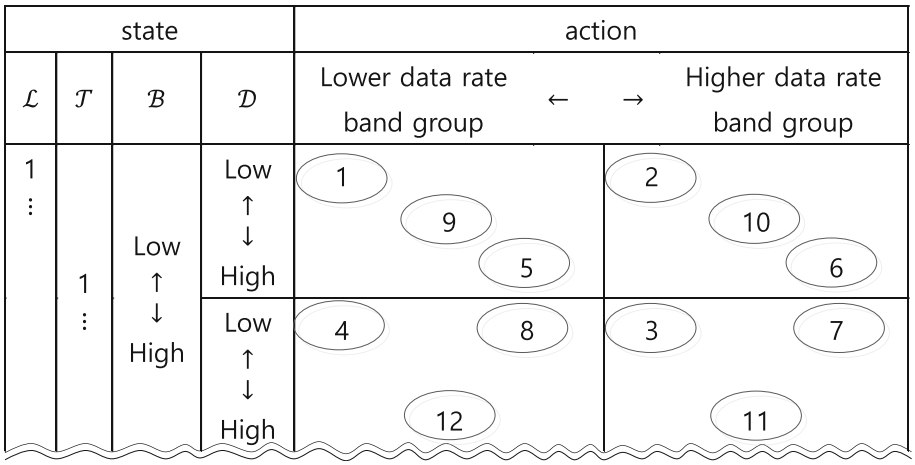


Fig. 9 Channel movement mechanism in Q-table and operating regions

this regard, studies related to the channel allocation in various fields are being conducted. Vishram et al. examined how to allocate channels using the graph coloring in the presence of homogeneous ad-hoc networks [11]. In their study, they maximized the overall performance while guaranteeing a certain grade of service to individual users with the fairness. Maghsudi and Stanczak applied the graph theory for the channel allocation in a device-to-device (D2D) environment and considered fairness by equalizing the interference for cellular users [12]. Han et al. studied channel allocation methods for maximizing the overall system performance in vehicular networks by using the submodular set function-based algorithm [13]. Li et al. investigated channel allocation methods that maximize the overall system reward using a semi-Markov decision process (SMDP) in a vehicular ad-hoc network (VANET) [14].

Other studies have considered a method of allocating channels according to either bandwidth or service characteristics. A study by Semiari et al. investigated methods of allocating a user application with dual-mode operation in the mmW and μW bands. The base station (BS) allocates a delay non-sensitive user application to the mmW while assigning a delay-sensitive user application to the μW band. Matching game theory is specifically used for channel allocation in the μW band. In non-line-of-sight (NLoS) of mmW band, the user application cannot be allocated since the wireless communication is impossible because of the frequency characteristics; therefore, channels are allocated by estimating line-of-sight (LoS) and are secured through Q-learning [15]. Liang et al. have studied a method of assigning the channel with the high transmission capacity to the vehicle-to-infrastructure (V2I) link and the channel with the high reliability to the vehicle-to-vehicle (V2V) link considering the requirements of the two types of the vehicular network links [16].

Recently, the Artificial Intelligence (AI) technology, such as machine learning, has been attracting attention in various fields [17]. Among them, the reinforcement learning is being studied in the wireless system field because it provides a solution to optimize the system parameters by learning the surrounding environment in a dynamic and complicated wireless environment [18]. The Q-learning is the representative reinforcement learning and there are also researches about using this to allocate channels in a dynamically changing environment. Asheralieva and Miyanaga studied the multi-agent reinforcement learning using rewards to maximize the signal-to-interference-plus-noise ratio (SINR) and increase the transmission capacity in D2D networks [19]. Srinivasan et al. described a way in which two BSs belonging to different operators in a cellular network can allocate channels by providing services to the nodes belonging to the all operators. They studied the reinforcement learning using the reward with the difference between quality of experience (QoE) and cost that can be obtained by providing two services [20]. Rashed et al. studied the reinforcement learning that maximizes the sum-rate of D2D users and cellular users to minimize the interference in a D2D environment [21]. Fakhfakh and Hamouda used the received SINR from the access point (AP) detected by the mobile user, QoS metrics about the channel load, and delay as the reward for choosing a WiFi over a cellular network to apply WiFi offloading and reducing the load on the cellular network [22]. Yan et al. propose a smart aggregated radio access technologies (RAT) access strategy with the aim of maximizing the long-term network throughput while meeting diverse traffic quality of service requirements by using Q-learning [23]. Maglogiannis et al. allowed the LTE system in the unlicensed band to select the appropriate muting period by using Q-learning to ensure coexistence with WiFi systems [24]. Xu et al. modeled the channel handoff

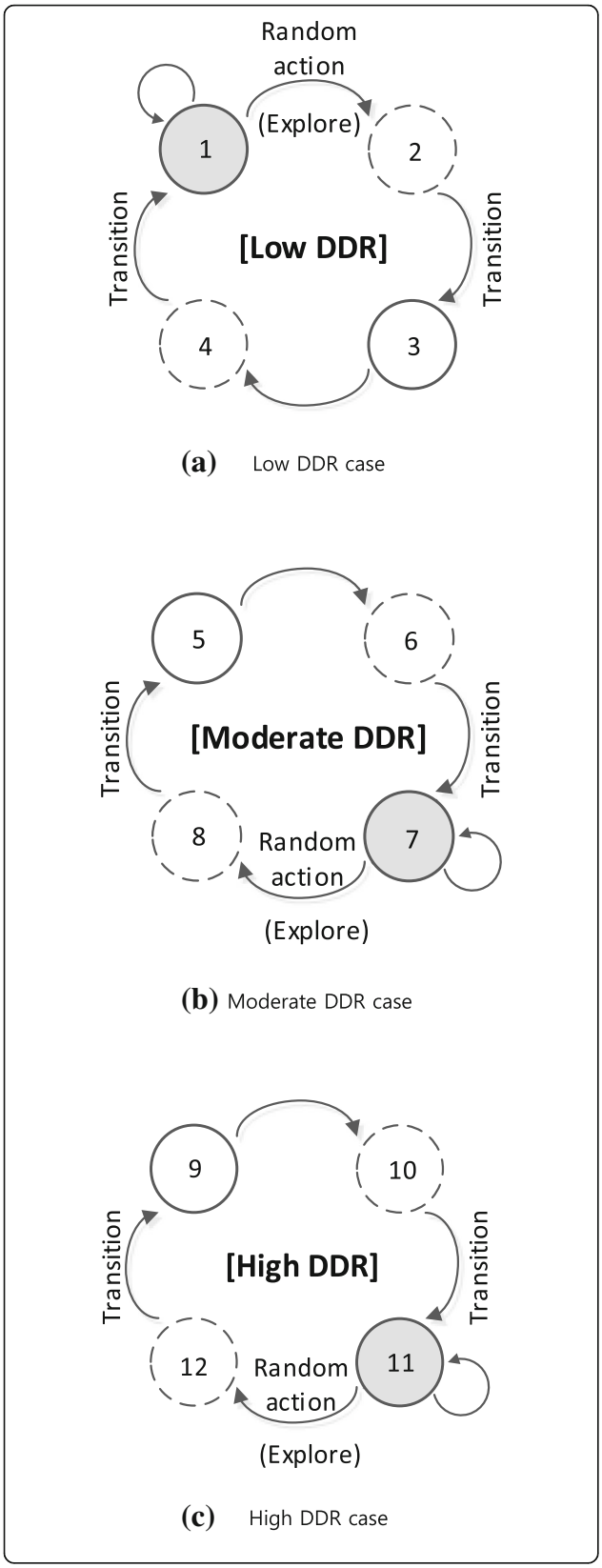


Fig. 10 Channel movement example in Q-table. The update of the Q-table represents a unique pattern according to DDR by the reward for channel utilization efficiency proposed in this paper. The channel movement example of **a** low, **b** moderate, and **c** high DDR cases in Q-table is shown. The stable domain is in gray circle in each case of DDR. Each domain changes to another one by explore or natural transition

process as a partially observable Markov decision process (POMDP) and adopted a Q-learning algorithm to find an optimal handoff strategy in a long term [25]. Jang et al. proposed Q-learning based sensing parameter (sensign time and interval) control mechanism for cognitive radio networks [26]. L. Shi et al. presented optimal resource allocation for LTE-U and WiFi coexistence network using Q-learning [27].

Various studies have been carried out about selecting channels, but most studies do not consider the fairness of the channel selection between users. Even if the fairness is taken into consideration, they just allocate resources fairly regardless of the required data rate or considered it as a central manner [28]. And the central scheme is difficult to realize the realistic implementation because of the complexity, or their scheme gave the loads to the network due to the centralized control. Some distributed resource allocation mechanisms (e.g., game theory) may also cause a loss of time or resources because the channel is selected by the interaction between the systems. The fairness of the channel usage is required in order to minimize the possibility of channel resources being unnecessarily consumed by some users and unavailable for other users who require more of them. In order to reduce the load on the system, it is necessary to consider fairness within the system itself without control message exchanges. Meanwhile, the various budgets for cognitive ad-hoc networks, such as time available to the channel, transmission speed, fairness, and bandwidth conversion cost, should be considered. Moreover, these budgets must work in concert to fit an objective function with some degree of freedom about flexible operation so that the system can be operated for various purposes without altering a predetermined objective. In this paper, the reward for spectrum utilization is designed so that fairness is taken into consideration by selecting a channel suitable for the required data transmission rate. In addition, we define a reward using weighted sums for various budgets as well as a Q-learning algorithm that can operate according to the change in weights.

3 Network model and system architecture

3.1 Network model

The system considered in this paper is the cognitive radio ad-hoc network comprised of CH and MNs as shown in Fig. 1a. The channel availabilities are different with geographic locations in accordance with the primary transmitter positions, channel gain between

Table 1 Channel parameters for band groups 1 and 2

			Operation time [min]		Supportable data rate (bps)	
			Mean (T_{op})	Variance (T_{op})	Mean (D_{rate})	Variance (D_{rate})
Channel	Band group 1	1	2.1	1	10 kbps	1
		2	4.2	2	55 kbps	3
		3	8.4	1	70 kbps	2
		4	6.3	2	85 kbps	2
		5	10.5	1	100 kbps	1
	Band group 2	6	5.2	1	0.8 Mbps	2
		7	3.8	1	1.6 Mbps	1
		8	6.7	2	2.4 Mbps	3
		9	8.1	1	3.2 Mbps	1
		10	9.5	2	4 Mbps	4

primary systems and secondary users, primary activity characteristics, and so on. The characteristics of these channels for CR ad-hoc networks can be also different in time zone and frequency band groups. Therefore, in this paper, we have considered the difference of channel characteristics according to geographical zone, time zone, band group, and frequency channel. As shown in Fig. 1b and c, the primary activities (i.e., available time to access by secondary users) and possible data transmission rates are different for each channel during the same time interval. Furthermore, the desired data rates of SUs changes according to time, as shown in Fig. 1d.

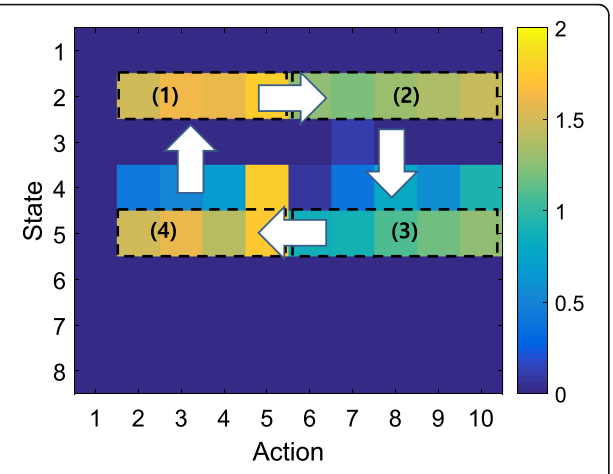
In particular, when the CR system operates cross a wide frequency range, including HF, VHF, and UHF, as shown in Fig. 2, the channel bandwidth for each band group that is defined for secondary systems can be different due to the band group-specific spectrum hole nature. In general, in HF, the spectrum holes are relatively narrow in the frequency domain because the licensed spectrum of primary systems using HF band group is also usually narrow. On the other hand, the spectrum holes of UHF are comparatively wider than that of HF. Excepting for details characteristic difference of each band group, we assume that the wider channel bandwidth is used in the higher band group frequency. Therefore, if the operating frequency of band group j (BG_j) is higher than that of band group i (BG_i), then the channel bandwidth of BG_j, W_j, is wider than W_i which is the channel bandwidth of BG_i and achievable data transmission rate (i.e., capacity) of W_j is greater than that of W_i. The greater preference for the channel is given to the band group with the higher bandwidth in the system or individual nodes. However, even though the bandwidth demanded by the secondary system can be satisfied by W_i of BG_i, if the bandwidth W_j of the higher BG_j is utilized by the secondary system, then satisfaction of the system will increase while overall spectrum resources are wasted. Because other secondary systems may exist around and their traffic demands only can be satisfied by using the bandwidth W_j of BG_j, therefore, a mechanism for adaptive allocation of band group use according to the traffic demand and bandwidth utilization efficiency of the corresponding system is required.

3.2 System architecture and problem formulation

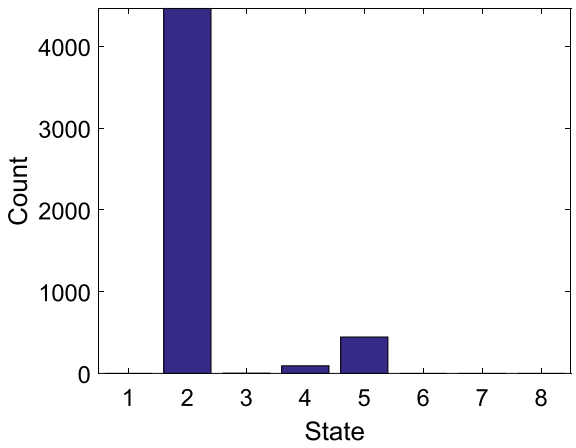
The proposed system architecture of CH is represented in Fig. 3. The Q-learning is used to dynamically select optimal band group and channel being aware of wireless environment, network user demand, and system operation parameters. The network user demand module determines the desired data rate (DDR) of the CR ad-hoc network based on each member node's traffic demand, and it also measures the average utilization of the channel currently used. The wireless environment monitoring module stores the spectrum sensing results such as average SNR (signal to noise ratio) and primary signal detection history. Using the sensing results, this module generates band- and channel-specific statistics which includes available data rate and primary idle time. The system operator can dynamically adjust the system parameter for learning using the system parameter module. The system operator can reset the reward

Table 2 Weight parameters

Weights vector (Default)	Q-learning parameters	Reward parameters	DRE parameters
$w_1 = 0.3$, $w_2 = 0.3$,	Learning rate (α) = 0.3,	overhead (η) = 0.01,	$r_1 = 1/6$,
$w_3 = 0.3$, $w_4 = 0.1$	Discount factor (γ) = 0.7	$\delta = 2$	$r_2 = 5/6$



(a) Value of Q-table



(b) Visit count of state

Fig. 11 The **a** Q-value of Q-table when DDR = 40 kbps, and **b** number of state visits. Similar to Figs. 9 and 10a, the Q-value of Q-table shows the channel movement of low DDR case. The visit count of state is high in the state of low DRE and low CBG because the stable domain is in that region

function by learning parameters for Q-learning. Based on all information, the Q-learning module determines which band group and channel can meet the data rate demand and maintain effective utilization level.

The Q-learning module changes from one channel to another when a PU appears on the current channel being used by the secondary system. The reward function is used to update the Q-table, and channel and band group selection is performed based on the current Q-table. The reward function proposed in this paper captures user demand, wireless environment, data rate efficiency (DRE), and band change overhead cost. The DRE is an evaluation metric to determine how much the ad-hoc network efficiently utilizes the data rate supported by the current channel.

In the proposed algorithm, we design the Q-learning reward function to satisfy the following criteria:

- Maximize the secondary system operational time;
- Satisfy the desired data rate of the CR ad-hoc secondary network;
- Provide the coexistence and fairness between secondary systems;
- Consider the overhead of band change for system reconfiguration;
- Guarantee operational flexibility and adaptability to meet the desired purpose.

4 Reinforcement learning for dynamic band and channel selection

4.1 Action, state, and Q-table design for Q-learning

The Q-learning is one of the model-free reinforcement learning techniques. It is able to compare the expected utility of the available actions for a given state without requiring a specific model of the environment. An agent tries to learn the optimal policy from its history of interaction with the environment, in which an agent applies a specific action at the current state and receives a response as a form of a reward from the given environment. The Q-learning eventually finds an optimal policy, in the sense that the expected value of the total reward return over all successive iterations is the achievable maximum one. The problem space consists of an agent, a set of states $\mathcal{S}$, and a set of actions per state $\mathcal{A}$. By performing an action $a \in \mathcal{A}$, the agent can move from state to state.

Figure 4 shows the Q-learning mechanism of the proposed method. The CH of the CR ad-hoc system is the agent of Q-learning, and the action of the CH is a selection of a new tuple (band group, channel) for the CR system operation when the primary signal is detected on the current band group and channel. The structure of the Q-table is expressed by rows of states and columns of actions. In this paper, the set of action $\mathcal{A}$ is given by:

$$\mathcal{A} = \mathcal{B} \times \mathcal{C}_k \tag{1}$$

where × is the Cartesian product; $\mathcal{B} = \{b_1, b_2, ..., b_{NB}\}$ expresses the set of channel band groups; NB is the number of band groups; $\mathcal{C}_k = \{c_1^{b_k}, c_2^{b_k}, ..., c_{NC_k}^{b_k}\}$ represents set of available channels in k-th band group (b_k); NC_k is the number of channels of band group b_k; and $c_j^{b_k}$ is $j - th$ channel of band group b_k.

In this paper, a multi-layered state is defined, in which it is composed of geographic location ($\mathcal{L}$), time zone ($\mathcal{T}$), channel band group ($\mathcal{B}$), and data rate efficiency level (D). The state of space in this system is defined as follows.

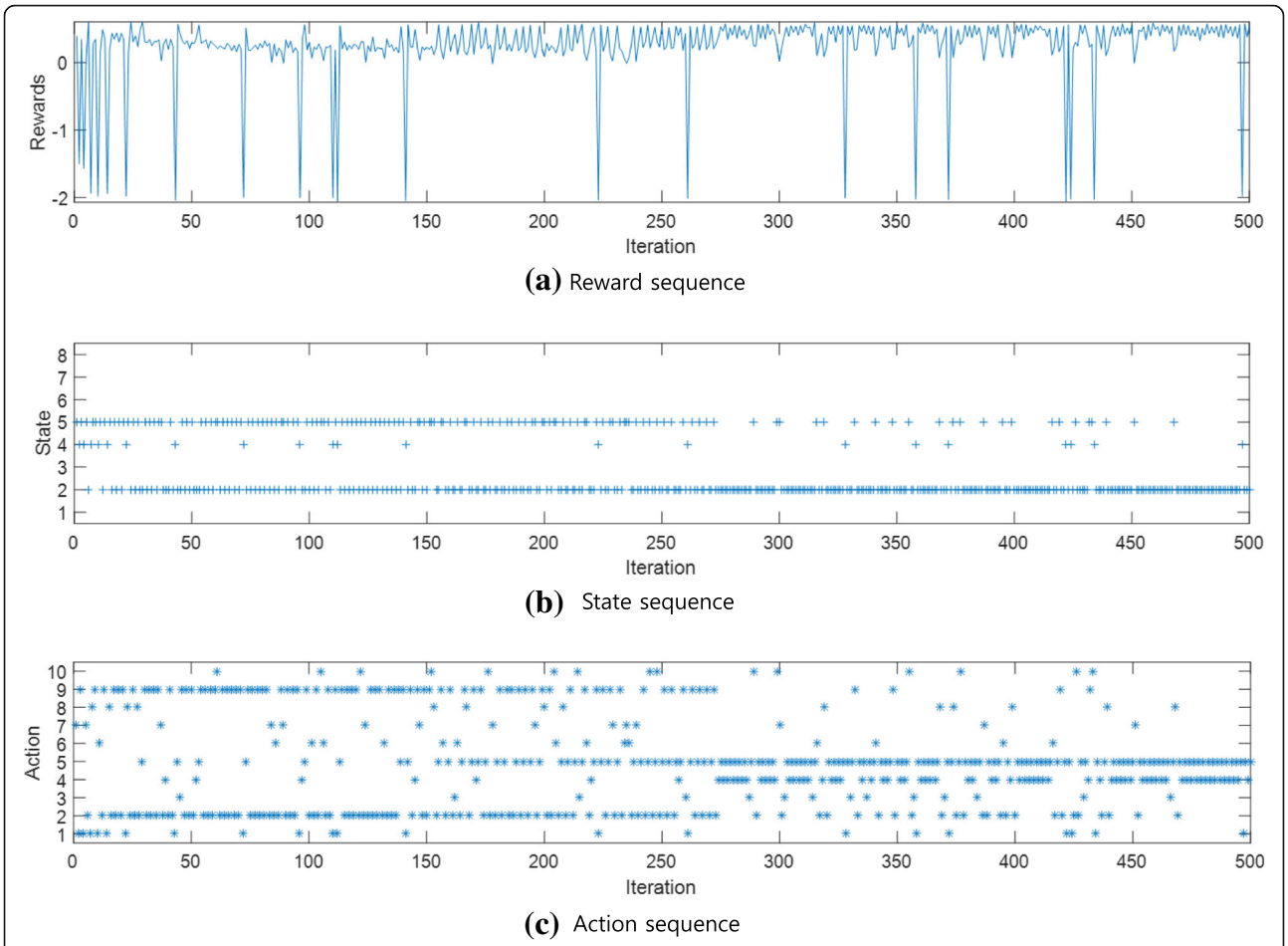


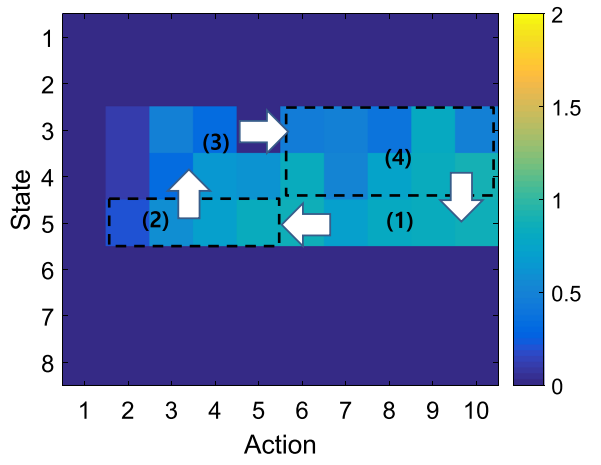
Fig. 12 The change of **a** rewards, **b** states, and **c** actions according to iteration at a low DDR of 40 kbps. The temporary low reward value is due to the random action of Q-learning exploration. The agents visits the state 2 more often than the states 4 and 5 over time as seen in Fig. 11b. As shown in Fig. 11a, the action in **c** mainly visits channel 2 or 3 and is adaptive to the DDR at the latest possible moment, even if a channel from band group 2 is selected or a channel from band group 1 offering a high data rate is selected. **c** represents how the agent selects the channels in band group 1 suitable for the DDR over time. Therefore, we can see that the agent operates according to the designed mechanism

$$\mathcal{S} = \mathcal{L} \times \mathcal{T} \times \mathcal{B} \times \mathcal{D} \tag{2}$$

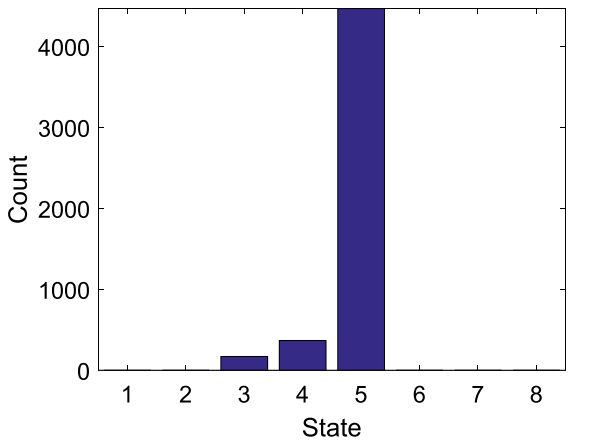
where $\mathcal{L} = \{l_1, l_2, ..., l_{NL}\}$, $\mathcal{T} = \{t_1, t_2, ..., t_{NT}\}$, and $\mathcal{B} = \{b_1, b_2, ..., b_{NB}\}$ represent the sets of geographic location zones, time zones, and band groups, respectively. NL, NT, and NB are the number of location zones, time zones, and band groups, respectively. In this paper, we have defined a new additional state dimension $\mathcal{D}$ to represent the operational state of the secondary system in terms of how much the CR system effectively utilize the given channel of the selected band group. $\mathcal{D} = \{d_1, d_2, ..., d_{ND}\}$ indicates the set of DRE levels, and ND is the predefined number of DRE levels. The DRE is the ratio of the DDR of the secondary network to the average supportable data rate of the current channel of the selected band group. Therefore, the current state is defined as a form of (l_i, t_j, b_k, d_l) tuple and it represents the current location zone, time zone, operation band group, and DRE level. For the given geolocation area and time period, the secondary CR system needs to select the next operational band group and channel whenever a channel switching is required. The current band group (CBG) and DRE capture the dynamic goodness of the selected band group and channel in terms of spectrum efficiency and support of the desired rate. The CH selects the best action for the current state using the current Q-table.

Figure 5 shows the proposed Q-table structure. At the current state (l_i, t_j, b_k, d_l), the CH selects the best action $(b_q, c_m^{b_q})$, i.e., the next band group b_q and m-th channel of b_q, which shows the maximum Q-value in the current Q-table. It needs be noted that the candidate channels of each band group as possible actions should be available channels at the current time as a result of spectrum sensing.

Figure 6 shows the procedure of the proposed Q-table update, state determination, and action selection. It is



(a) Value of Q-table



(b) Visit count of state

Fig. 13 The **a** Q-value of Q-table when DDR = 90 kbps, and **b** number of state visits. Similar to Figs. 9 and 10b, the Q-value of Q-table shows the channel movement of moderate DDR case. The visit count of a state is high in the state of low DRE and high CBG because the stable domain is in that region

assumed that the MNs transmit the average channel operation time and average supportable data rate to the CH through sensing and channel use report.

1. Suppose the learning agent CH determined the state s_{t-1} and the best action a_{t-1} at the end of (t − 1)-th time period.
2. During t-th time period, MNs and CH monitor the primary activities and channel statistics.
3. Agent CH detects the band and channel change event.
4. The CH calculates the reward r_{t-1} for the previous action a_{t-1} at state s_{t-1}.
5. The CH updates the Q-value of (s_{t-1}, a_{t-1}) in Q-table.
6. The CH determines the current state s_t based on the measured DRE during t-th time period.
7. The CH selects the optimal action a_t for the next (t + 1)-th time period.
8. Go to step 1.

The Q-learning updates the Q-value for each pair of state and action (s, a) visited through these series of processes. The Q-value reflects the value that the system can accept when selecting action a in state s.

The Q-value update of the Q-table can be represented by:

$$\mathcal{Q}(s_t, a_t) \leftarrow (1-\alpha)\mathcal{Q}(s_t, a_t) + \alpha\left\{r_t + \gamma \max_{a_{t+1}} \mathcal{Q}(s_{t+1}, a_{t+1})\right\} \quad (3)$$

where α and γ denote the learning rate and discount factor, respectively. The learning rate $\alpha \in [0, 1]$ is used as a weight to reflect the $\mathcal{Q}(s_t, a_t)$ accumulated from the past, the newly obtained reward, and the expected reward value for the next action. If α is low, it increases the weight of the past experience so the system takes an extended time to learn, but the fluctuation of the reward sequence is low. If α is high, the learning speed is increased by assigning a high weight to both the present and future values. However, an extremely high α causes instability in the system, while a fairly low α prevents the system from reaching a satisfactory reward at the desired time. The discount factor $\gamma \in [0, 1]$ is the weight for how much the Q-value of a_{t+1}, the future reward, should be reflected in the Q-value of action a_t in the Q-table of the current action and state. A high γ has a high contribution on the Q-value of the future expected reward, and a low γ weights the reward according to the current action a. That is, when Q-learning reflects the immediate reward and the future tendency in the Q-value of the action and corresponding state, γ is a weight that takes into account whether to further consider the volatility of the current action or to reflect the future value predicted from past trends of the Q-table.

If the CH only uses the updated Q-values to select actions, it may fall into local optimum. Therefore, we use ε-greedy policy to add randomness to selecting of actions that are explorative in the learning algorithm, as follows:

$$a = \begin{cases} \underset{\tilde{a}\in\mathcal{A}}{\text{argmax}}\, \mathcal{Q}(s, \tilde{a}), & \text{with probability } 1-\varepsilon \\ \text{random } a \in \mathcal{A}, & \text{with probability } \varepsilon \end{cases} \quad (4)$$

where, $\varepsilon \in [0, 1]$ is the probability of choosing a random action. If ε is high, it is more likely that new information will be added to the already accumulated information while searching for the next action; if it is low, the next action is selected using only the accumulated information. ε

starts with a specific value and lowers this value for each iteration, so that the Q-table can operate stably after a certain time. However, when the value of ε decreases continuously, a considerable amount of time is required for adapting to the changing environment by updating the Q-table. Therefore, a lower limit of ε is required.

The overhead of Q-learning can arise from the memory size for the use of Q-tables. It depends on the level of the actions (the number of channels and bands) and resolution level of states, and it increases linearly with each level. If you set the number of level too low, the system cannot use the Q-table for learning dynamic environments. Otherwise, the system takes a long time to learn the surrounding environment using the Q-table. Therefore, the selection of appropriate level is required.

4.2 Reward function design

The reward that the CR system obtained by using the selected set of (band group, channel) is composed of the system operation time, average data transmission rate, channel utilization efficiency, and overhead required for the system to change the frequency band. The reward for the action a is expressed as follows:

$$R(a) = w_1 \frac{T_{op}}{\max\left(T_{op}^{cbg}\right)} + w_2 \frac{E[D_s]}{\max\left(E\left[D_s^{cbg}\right]\right)} + w_3 R_{Util} - w_4 BC_{(a,a')} \quad (5)$$

where T_{op} is the consecutive channel operation time for the secondary system after the channel is selected, in which if a channel shows high T_{op} value, then it indicates that once the secondary system takes this channel it can use the channel relatively long time before the primary appears. $E[D_s]$ is the average supportable data rate of the selected channel. R_{Util} represents how the secondary system utilizes the channel effectively. $BC_{(a,a')}$ is the overhead for band group change. The operation time and average supportable data rate are normalized to their maximum values for the current band group. a and a' are the current action and previous action, respectively.

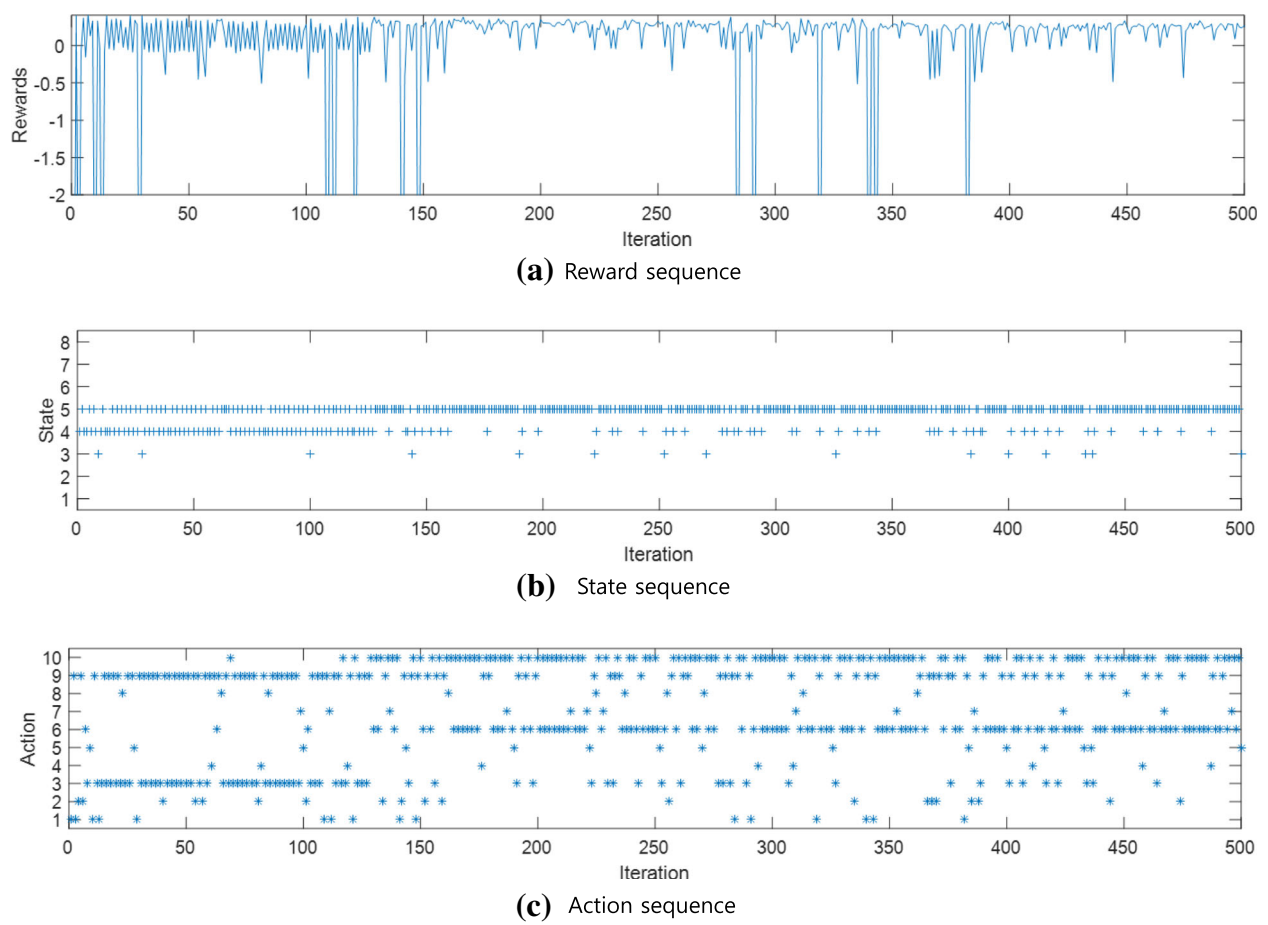
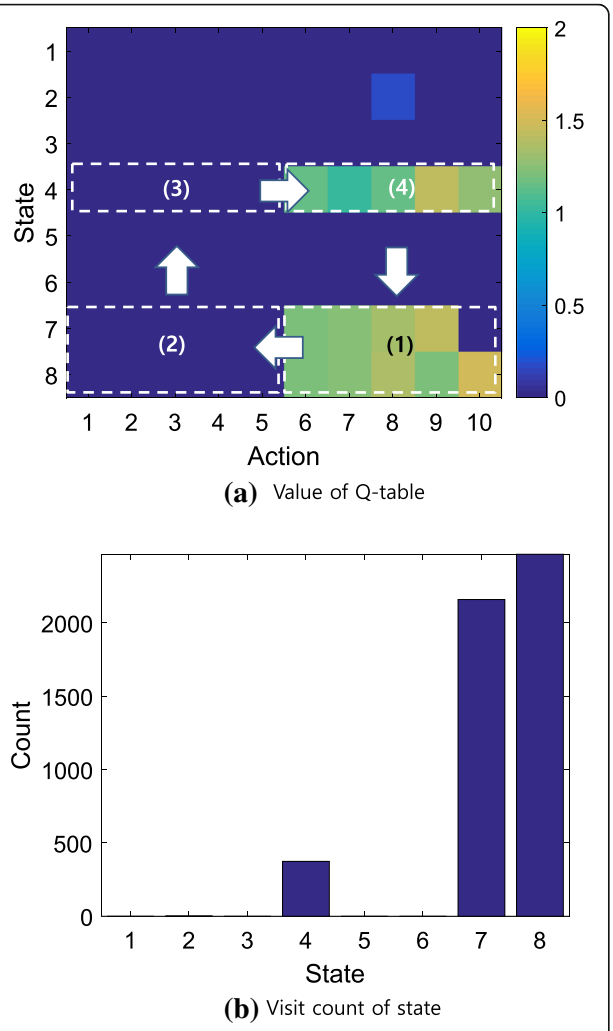


Fig. 14 Rewards, states, and actions according to iteration at DDR = 90 kbps. In **a,** the reward is stable at more than 10 iterations, and we can see that the reward is temporally low in the overall interval by random action, similarly to Fig. 12. As shown in **b** the agent mainly visits the state 5. **c** Reveals that actions in band group 2 are selected mostly



(a) Value of Q-table

(b) Visit count of state

Fig. 15 The **a** Q-value of Q-table when DDR = 3.5 Mbps, and **b** number of state visits. Similar to Figs. 9 and 10c, the Q-value of Q-table shows the channel movement of high DDR case. The visit count of state is high in the state of high DRE and high CBG because the stable domain is in that region

$max(T_{op}^{cbg})$ and $max(E[D_s^{cbg}])$ are the maximum channel operation time and maximum expected supportable data rate value from all channels in the current band group, respectively. w_i is the weight for i-th reward component and $\sum_{i=1}^{4} w_i = 1$. The first and the second term are normalized by each maximum value of the operation time and average supportable data rate in each channel group so that the relative value to the other channels can be reflected in the reward. The third term is described in (7) and serves to adjust the reward to select a channel suitable for the desired data rate. The fourth term represents the cost due to an overhead when a band group change occurs, which is described in (6). All the terms are linearly coupled to allow the system designer or user to operate the system for a specific purpose through weighting changes.

The overhead for band group change, $BC_{(a,a')}$, in (5) is to capture the required additional time and energy for reconfiguring some system operational parameters when the band group is changed. In most cases, different band groups have different channel bandwidths and wireless characteristics so that communication system may need to reconfigure radio frequency (RF) front-end, modulation method, and medium access control (MAC) layer components whenever it changes its operation band group. The overhead is represented as in (6).

$$BC_{(a,a')} = \begin{cases} \eta, & \text{channel a and a}' \text{ belong to different band groups.} \\ 0, & \text{else} \end{cases} \tag{6}$$

where η is the cost when the current channel and the previous channel belong to different band groups. In this paper, we do not consider the channel switching overhead inside the same band group.

R_{Util} of (5) is defined as a function of DRE. The DRE is defined in this paper as in (7)

$$DRE = \frac{DDR}{E[D_s]} \tag{7}$$

where DDR is the desired data rate (DDR) of the secondary CR network. To design R_{Util} function, first we considered the desired system operation in terms of band group selection depending on the current DRE value.

Figure 7 shows the example of the desired mechanism by which the channel selection is performed according to the current DRE. It is assumed that the bandwidth of each channel provided by each band group follows $W_1 < W_2 < W_3$, in which we have three band groups. The x-axis is divided by band group, and the parts represent the DRE for each band group. DRE (U) indicates that the utilization ratio of the channel is low when it belongs to $[0, r_1)$ of band group 1, and that ratio is moderate when it belongs to $[r_1, r_2)$. If U belongs to $[r_2, 1)$, it denotes a high channel utilization ratio so that some time instances of the network may not be able to meet the user traffic demand. The range of $[1, \infty)$ means the channel cannot support enough bandwidth for the system. A low channel utilization ratio means that the possible transmission rate provided by the selected channel of the current bad group is much higher than the desired CR network data rate so that most of spectrum resource is wasted after it satisfies the desired data transmission rate. It is therefore necessary to move to a channel that provides a lower bandwidth and data rate (i.e., change to the lower band group channel). Furthermore, if U shows a higher channel utilization ratio than the defined r_2, which means that the possible data transmission rate provided by the selected channel is not

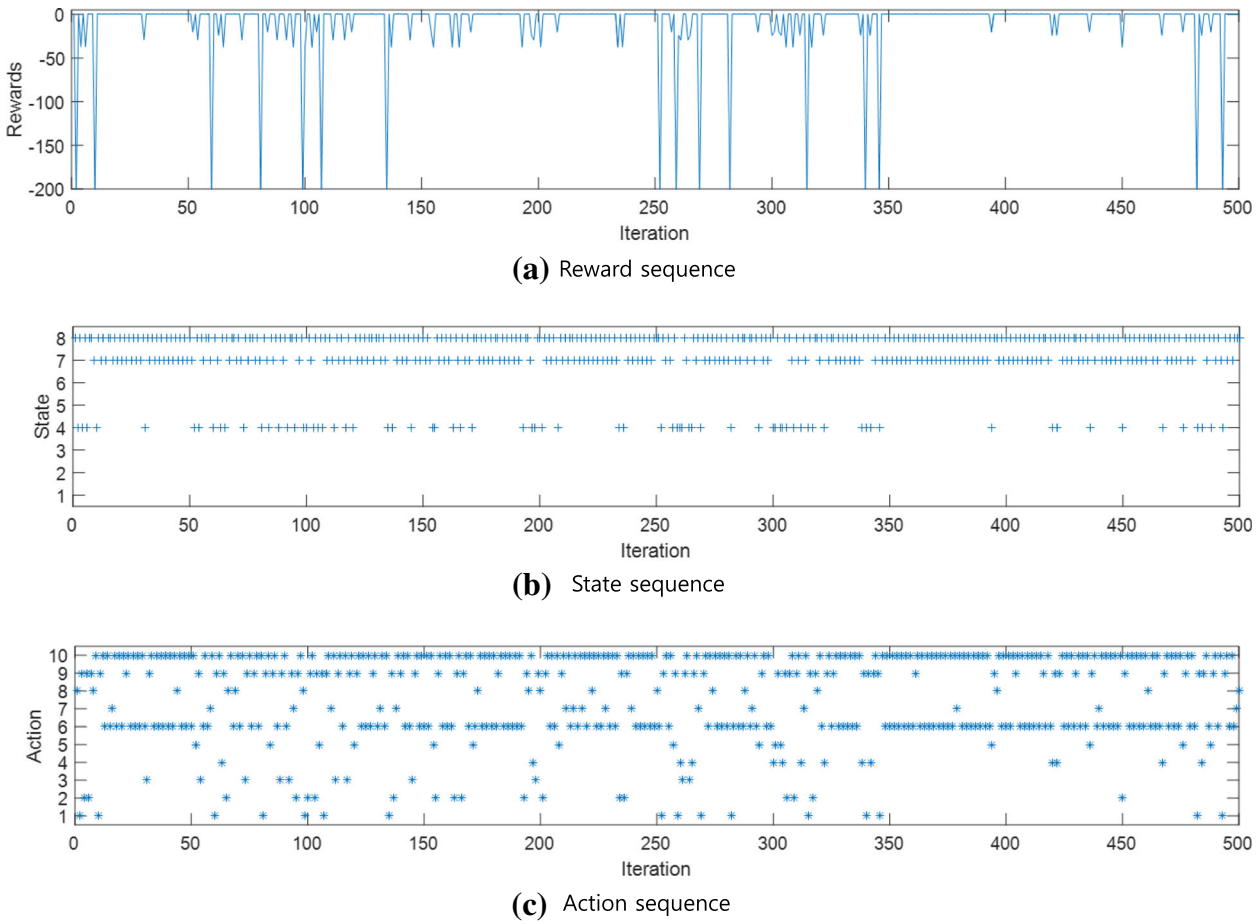


Fig. 16 Rewards, states, and actions according to iteration at DDR = 3.5 Mbps. In **a,** the reward is stable overall, while it is temporally low in the overall interval by random action, similar to Figs. 12 and 14. In **b,** the system visits state 4 to a degree, but it mainly remains in states 7 and 8. **c** shows that actions in band group 2 are selected mostly

likely to support the desired data rate of the system, then it is necessary to move to a channel band group that can provide a wider bandwidth and a higher data rate. However, in Fig. 7, even though DRE is in $[0, r_1)$ for band group 1, the secondary system does not have any band group that has narrower channel bandwidth so that it needs to keep the current band group. On the other hand, in case of band group 3, when DRE is in $[r_2, \infty)$, the system does not have any band group that has wider channel bandwidth so that it has to find other best channel in the same band group. In each band group, $[r_1, r_2)$ is the band usage maintenance interval, because the selected band group channel provides an appropriate transmission rate.

Based on the band group selection movement mechanism in Fig. 7, the proposed utilization efficiency reward function R_{Util} of Eq. (5) is shown in Fig. 8, in which we assume that there are three band groups. The x-axis for each band group represents DRE (U), and the y-axis represents R_{Util}. For the band usage maintenance range in Fig. 7, R_{Util} is set to 0 in $[r_1, r_2)$ DRE range for all band groups, which represents the medium channel utilization efficiency ratio. In $[0, r_1)$ DRE range, the selected band group channel can support much larger data rate than the desired data rate so that channel utilization efficiency is low. Therefore, as the value of DRE goes from 0 to r_1, R_{Util} increases from −1 to 0 except the in first band group 1. Any ad-hoc CR secondary systems that has its current DRE value in $[0, r_1)$ need to move to the band group channel that has narrower channel bandwidth and lower supportable data rate. It makes the secondary system to yield the current band group channel to other secondary systems that requires more data rate. For the first band group 1, there is no other narrower band group so that R_{Util} is maintained at 0 in $[0, r_1)$ DRE range. The range $[r_2, \infty)$ is divided into $[r_2, 1)$ and $[1, \infty)$ to distinguish the insufficient transmission rate provided by the channel, with R_{Util} representing a more rapid decrease rate in $[1, \infty)$ range except in the last band group 2. In $[r_2, 1)$ DRE range, the band group

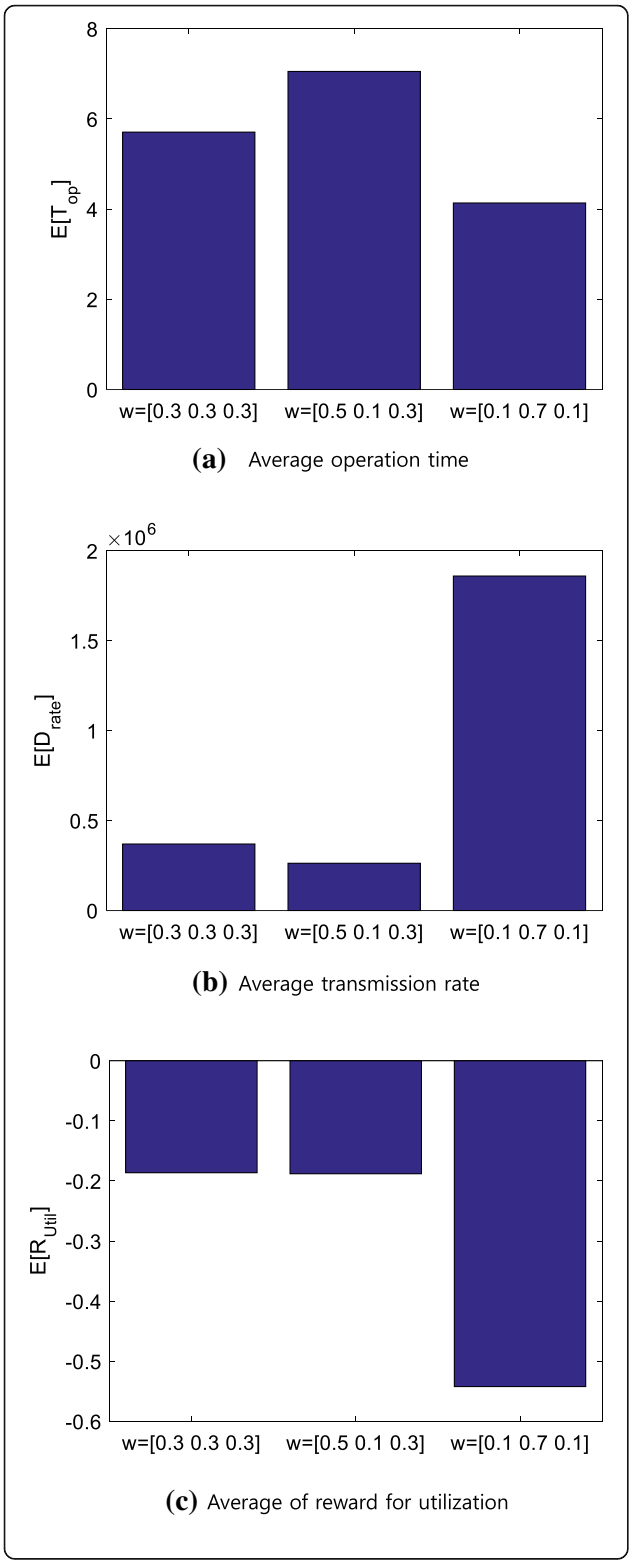


Fig. 17 **a** Average operation time, **b** average transmission rate, and **c** reward of utilization according to weight change. The average operation time, average data rate, and reward for channel utilization by changing the weight assignment for DDR to 40 kbps. Since the reward function is composed of the weighted sum of the objective functions, the Q-learning can be operated according to the desired objective function by adjusting the weight. Therefore, if the weight of the operation time is increased, the average operation time is increased, and if the weight of the data transmission rate is increased, the average transmission rate is increased. Finally, increasing the weight of reward for utilization increases the average of reward for utilization

channel supportable data rate may not be enough to guarantee the desired rate in some time instances so that R_{Util} decreases from $-r_2$ with a slope of -1 as DRE increases. In $[1, \infty)$ DRE range, the current band group channel cannot support desired data rate so that R_{Util} decreases with a slope of $-\delta$ $(\delta > 1)$. For the last band group 3, there is no other wider band group so that R_{Util} is maintained with 0 in $[r_2, \infty)$ DRE range.

Figures 9 and 10 show how this intentional mechanism is supported in the Q-table. Figure 9 shows the Q-table where the state is divided into geographic zones and time zones, and again into band groups and discrete DRE levels. The columns of the Q-table are divided into bands, which are then divided into selectable channels as possible actions. In the action shown in Fig. 9, a channel that represents a narrower band is shown as a lower data rate toward the left, and a channel that can use a wider band appears as a higher data rate toward the right. Figure 10 represents how the Q-value updating area changes in the Q-table of Fig. 9 through the example of three DDR cases. First, Fig. 10a depicts a case where the DDR is low. In Fig. 9, suppose that the secondary system is operating in action domain 1 (low CBG, low DRE) and by the Q-learning ε-greedy policy it may randomly select action domain 2 band group channel. As a result, the DRE is significantly lowered, and the system gets low R_{Util} because a high channel band provides a high data rate and it results in low reward for channel utilization efficiency. Therefore, after updating the Q-value of domain 2, the system operating area changes to domain 3 by the change of the state which represents (high CBG, low DRE). Because selecting a channel which supports a high data rate makes the DRE low, the Q-table then updates the Q-value of domain 3 and the Q-learning agent will select the best action of domain 4 because selecting a channel with a low band gives a high R_{Util}. After selecting the low-band channel, the transition to domain 1 (low CBG, low DRE) is performed. In this case, R_{Util} does not have a negative value because there is no longer a lower channel to select, as in the $[0, r_2)$ of band group 1 seen in Fig. 8, and no value is subtracted from the total reward. Therefore, in the

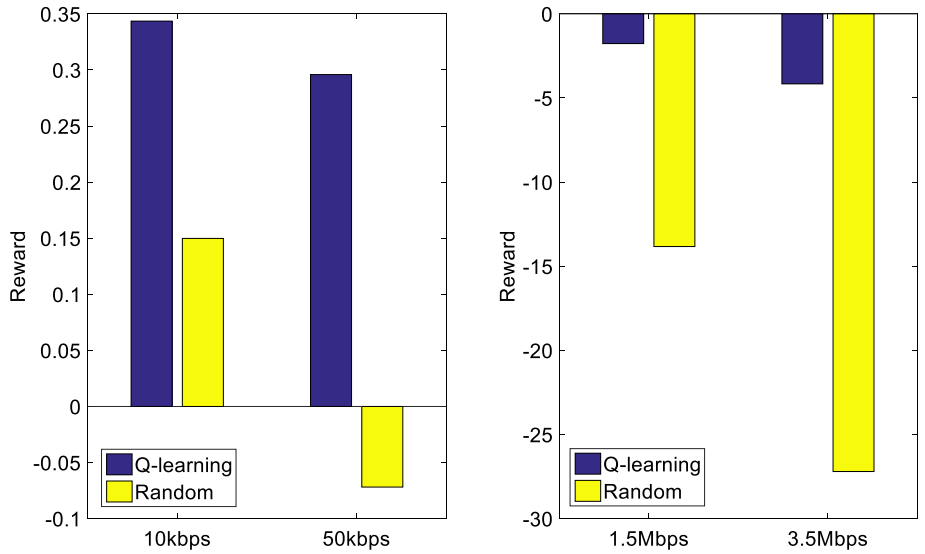


Fig. 18 Average reward comparison for Q-learning channel selection vs. random channel selection. The average reward for each DDR depending on the method of the channel selection. For all of the DDR cases, Q-learning band and channel selection has more reward value than random selection. The reward for a DDR of 1.5 or 3.5 Mbps (e.g., more than medium or high DDR case) is lower than that of 10 and 50 kbps. The ε-greedy policy in case of a high DDR causes very low reward for channel utilization R_{Util} due to the selection of a low-band channel which support insufficient data rate and these effects are accumulated in the Q-table

case of a low DDR, the preferred domain in the Q-table is not domains of 2, 3, and 4 where high band is selected by occasional ε-greedy policy for exploration but domain 1 (stable operating domain). Cases of both moderate and high DDR, as shown in b and c of Fig. 10, can be similarly explained.

5 Simulation results and discussion

The simulation environment in this paper assumes five channels for each of the two band groups, as listed in Table 1. The channels of band group 2 provides higher supportable data rates than those of the band group 1 while the operation time available for the transmission is not significantly different. We use a Gaussian distribution to determine the operation time and supportable data rate of each channel in each band group based on the mean and variance values provided in Table 1. The other simulation parameters are shown in Table 2.

In this paper, the action is defined as the selection of the channel in each band group as shown in Table 1. The index number of the action corresponds to the number of the channel in Table 1, and the total number of action is 10. We define the state as the combination of (band group, DRE) where the domain of DRE is divided as $[0, r_1)$, $[r_1, r_2)$, $[r_2, 1)$, $[1, \infty)$. The domain of DRE corresponds to each band group so the total number of state is 8.

The ε-greedy policy for action exploration is as follows:

$$p(n) = \begin{cases} p_0(0.999)^n, & p(n) > p_{low} \\ p_{low}, & p(n) \leq p_{low} \end{cases} \quad (8)$$

where $p_0 = 0.3, p_{low} = 0.1$

where n represents the iteration sequence number with time. In applying the ε-greedy policy, when the wireless environment of the system is changed, p_{low} is required to set a lower limit for a random value in order to maintain a certain degree of exploration. Otherwise, the Q-table cannot adaptively operate in the changed environment.

The overall simulation configuration starts by looking at the operation of Q-learning for each DDR, 40 kbps – 90 kbps – 3.5 Mbps, and confirming the change of average operation time and average transmission data rate according to the weight of the reward. Next, we compare the results of Q-leaning and random channel selection according to the reward, operation time, data rate, and utilization. Finally, we compare the change of DDR according to iteration for Q-learning and random channel selection.

5.1 Adaptive channel selection according to DDR

In this section, we identify our proposal adaptively selects the channel according to each DDR (e.g., low, moderate and high) as described by Figs. 9 and 10 in Section 4.2. Figure 11a shows the value of the Q-table in scaled colors when the DDR is 40 kbps, and b shows the number of visits to each state in the Q-table. The DDR of

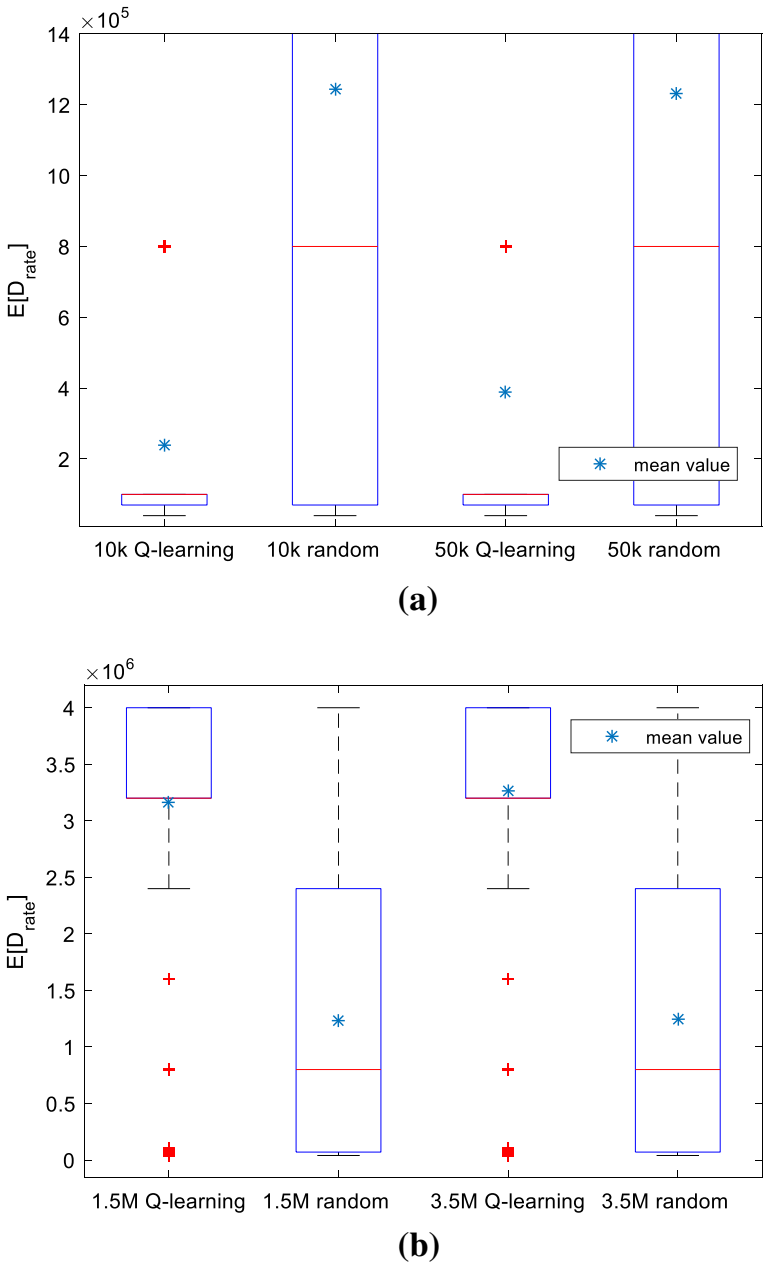


Fig. 19 a Mean value and **b** boxplot of data transmission rate for Q-learning and random channel selection. Shows the mean and boxplot of the data transmission rate for Q-learning and random channel selection for each DDR. In cases of the random channel selection, the average data transmission rate of all DDR cases is the same as the average value of the data rates for all channels in Table 1 belonging to band groups 1 and 2. The boxplots of all DDR cases for the random channel selection have a similar distribution. For the DDR of 10 kbps and 50 kbps, the mean of the Q-learning selection is lower than the random selection and the Q-learning has more narrow distribution. For the DDR of 1.5 and 3.5 Mbps, the mean of the Q-learning selection is higher than the random selection and the Q-learning has more narrow distribution

40 kbps is the low data rate comparing the data rate of channels in Table 1. Therefore, if the CH selects the channel of band group 1, the DRE belongs to almost [0, r_2) of band group 1 in Fig. 8 comparing the channels in band group 1, and R_{Util} does not give any effect on total reward. However, if the CH selects the channel of band group 2, the DRE belongs to $[0, r_1)$ of band group 2 in Fig. 8 and R_{Util} has an impact on the total reward linearly according to DRE. Figure 11a represents the process of changing the channel (action) selected by the ad-hoc CH (agent). When the CH selects the channel of band group 1, the process of updating the Q-value in the Q-table takes place in domain (1), which represents the channel selection of band group 1 and the low DRE

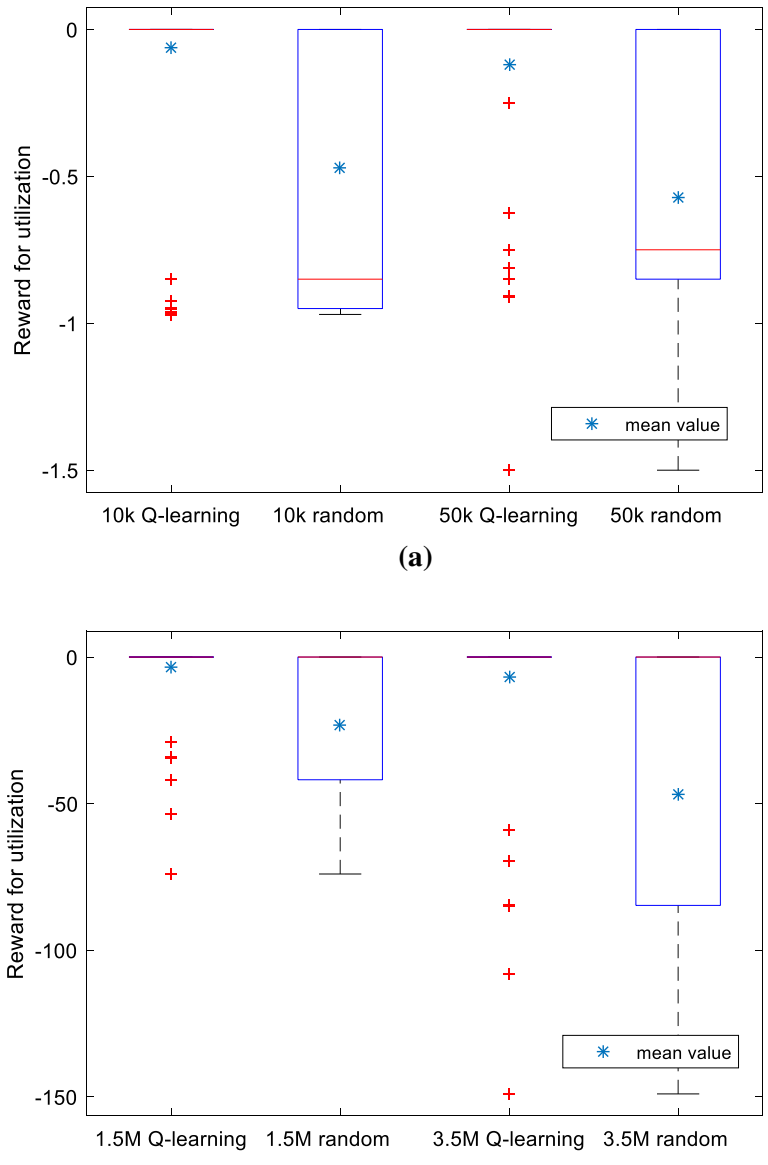


Fig. 20 The **a** mean and **b** boxplot of the reward for the channel utilization by the Q-learning and random selection at each DDR. For all DDRs, the boxplot of Q-learning has denser distribution and higher values than that of the random selection, and it has a higher average value

of band group 1. If the CH selects the channel of band group 2 by the explore policy of the Q-learning, the Q-value of the domain (1) is changed and the update process moves to the domain (2) which represents the channel selection of band group 2 maintaining the current state. Since the CH selected the channel of band group 2 that provides a high data rate for low DDR, the state is changed to the low DRE of band group 2. Thus, the update process of the Q-table moves to domain (3) and the Q-value of domain (2) changes. If CH selects channel of band group 1 in state 5, the state is maintained and update process moves to domain (4) after the Q-value change in domain (3). Finally, the state is changed to the state 2, which represents the band group 1 and low DRE, and the update process moves to domain (1) after the Q-value of domain (4) changes. The Q-value of Fig. 11a is the highest in domain (1), which represents the low DRE of band group 1 same with the low DDR case of Fig. 10a by the reward for utilization in Fig. 8. As a result, Fig. 11b represents that the number of visits in state 2 is the highest which corresponds to domain (1).

Figure 12 shows the change of rewards, states, and actions according to iteration at a low DDR of 40 kbps. The temporary low reward value is due to the random action of Q-learning exploration. The agents visit the

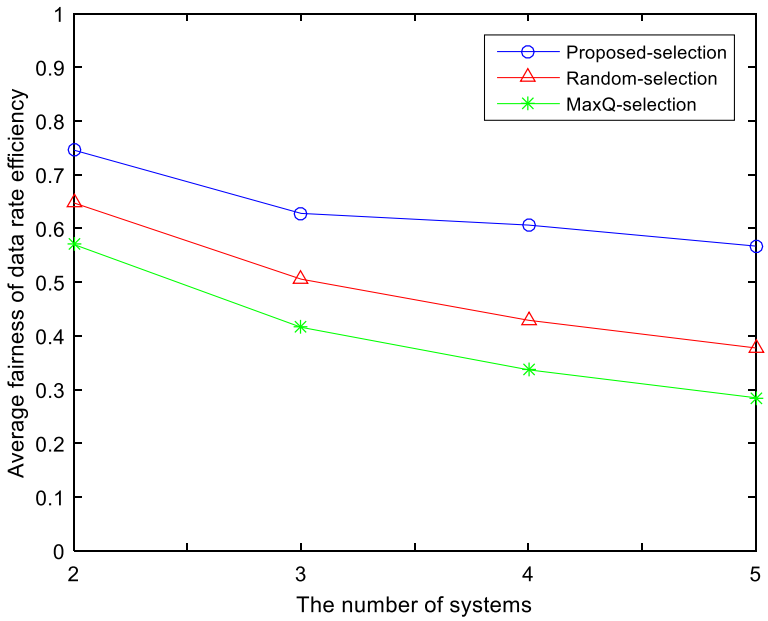


Fig. 21 The average fairness of data rate efficiency with increasing number of neighboring secondary systems

state 2 more often than the states 4 and 5 over time as seen in Fig. 11b. As shown in Fig. 11a, the action in Fig. 12c mainly visits channel 2 or 3 and is adaptive to the DDR at the latest possible moment, even if a channel from band group 2 is selected or a channel from band group 1 offering a high data rate is selected. Figure 12c represents the agent selects the channels in band group 1 suitable for the DDR over time. Therefore, we can see that the agent operates according to the designed mechanism.

Figure 13a shows the value of the Q-table in scaled colors when the DDR is 90 kbps and b shows the number of visits for each state in the Q-table. If the DDR is 90 kbps and the CH chooses a channel in the band group 1, the DRE belongs to $[r_2, \infty)$ in Fig. 8. Meanwhile, the DRE belongs to $[0, r_1)$ in Fig. 8 if a channel is chosen from the band group 2. Therefore, the supportable data rate by the channels in band group 1 is not enough in comparison with the channels in band group 2, as seen in Table 1, since the channel selection from band group

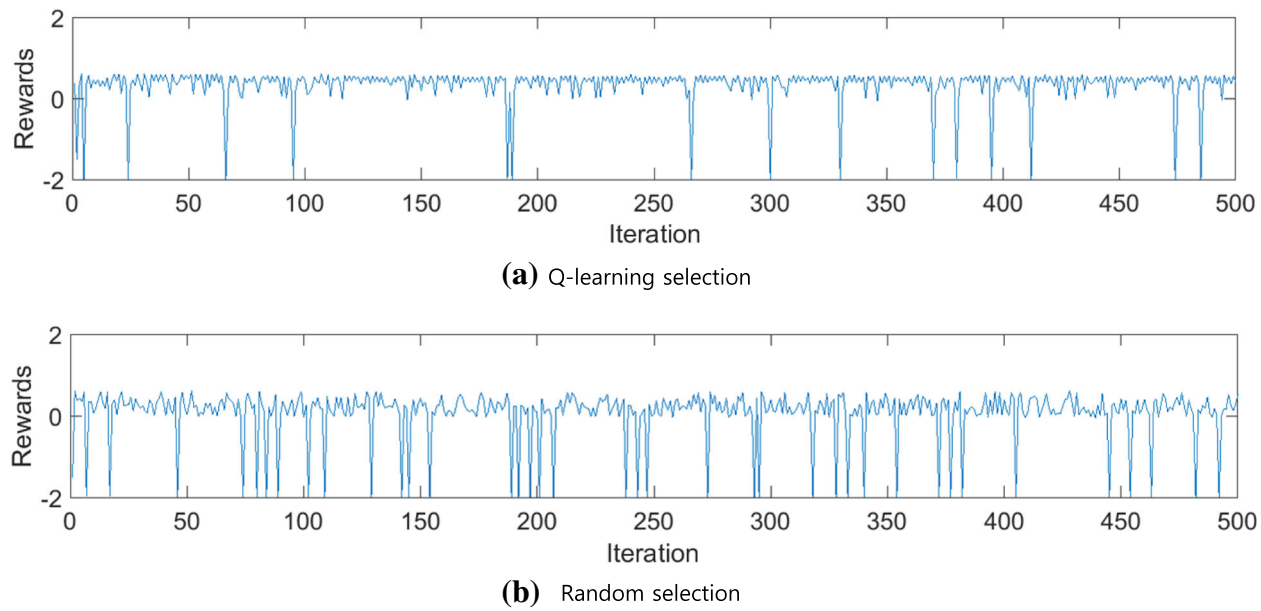


Fig. 22 Rewards for **a** Q-learning and **b** random channel selection according to iteration (DDR = 40 kbps)

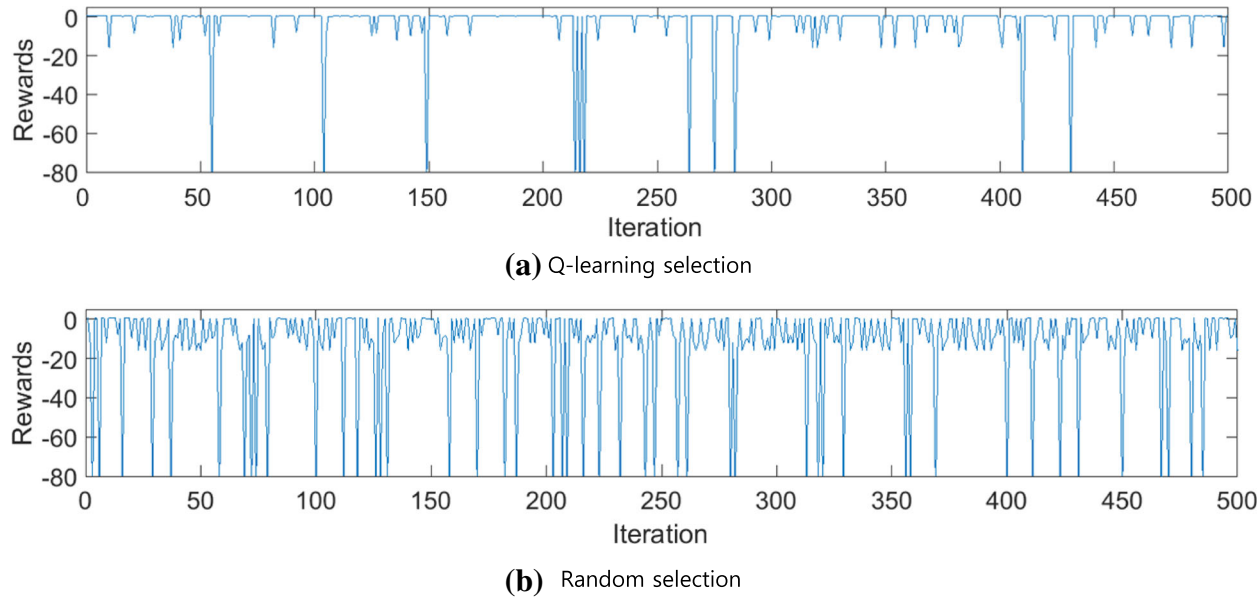


Fig. 23 Rewards for **a** Q-learning and **b** random channel selection according to iteration (DDR = 1.5 Mbps)

2 offers better R_{Util} than that of band group 1. We assume that the process of updating the Q-value starts from domain (1) which represents the channel selection of band group 2 and the DRE is low in band group 2. After selecting the channel in band group 1 by random channel selection, the Q-value of domain (1) is renewed and the update process moves to domain (2). If the CH selects channels from band group 1, the renewal process of the Q-table changes to domain (3) due to the high DRE which means the selected channel does not support a high enough transmission data rate after the renewal of domain (2). The process of updating moves to domain (3) by the change of state then transfers to domain (4) by the random or best selection. Because the channel selection in band group 2 provides more reward for utilization by Fig. 8, the Q-table in Fig. 13b has the

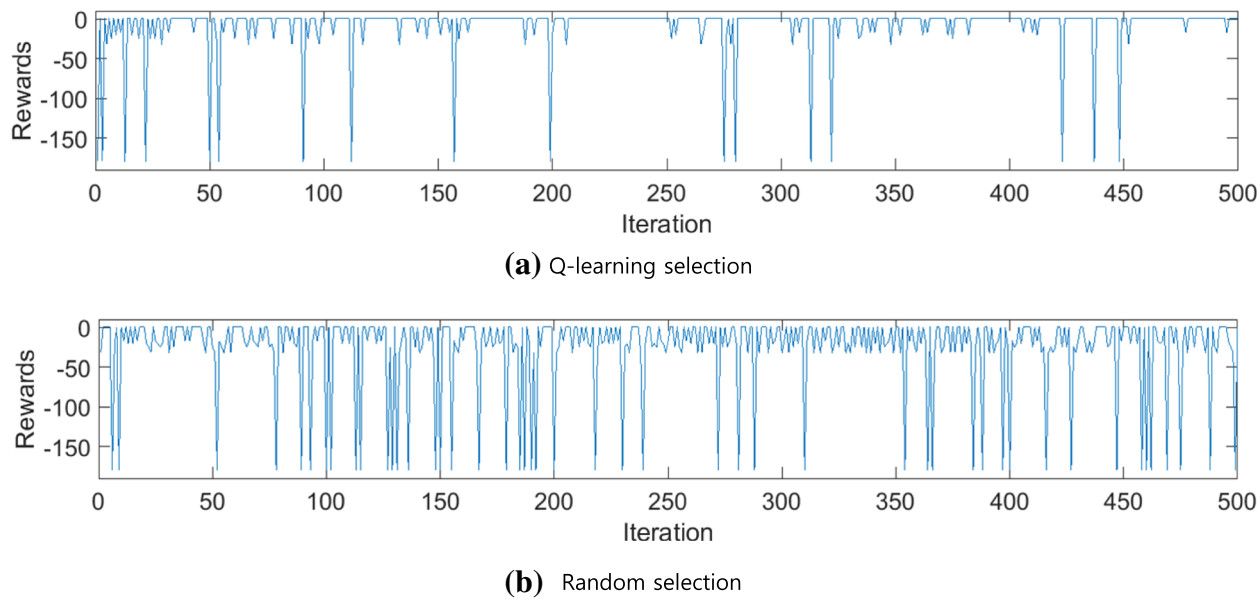
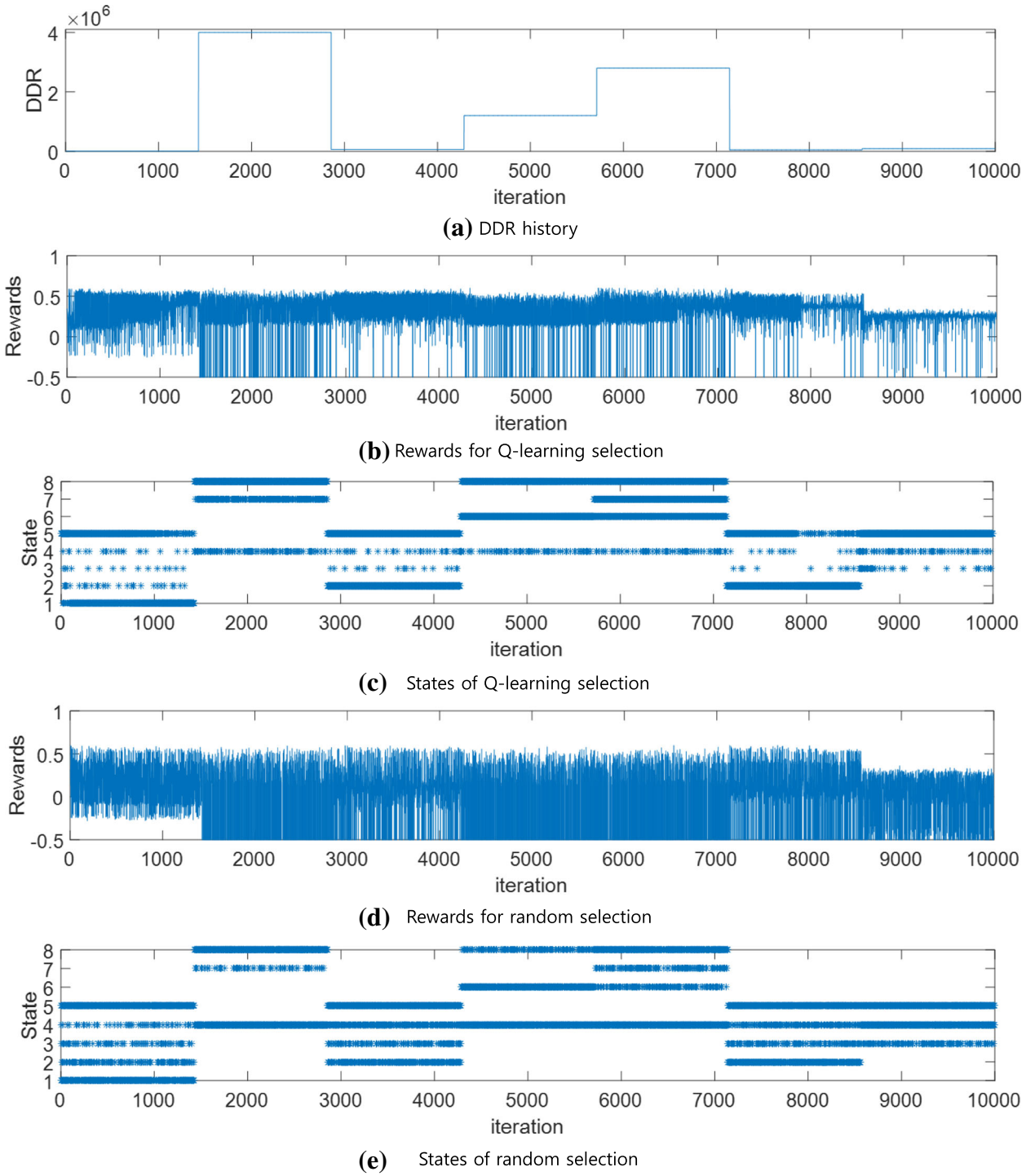


Fig. 24 Rewards for **a** Q-learning and **b** random channel selection according to iteration (DDR = 3.5 Mbps). Figures 22, 23, and 24 show the rewards for Q-learning and random band and channel selection according to the iteration for each DDR. In the case of Q-learning for all DDRs, the fluctuation decreases over time and the system operates with the intended reward design. In the random selection, there are more notches and fluctuation than Q-learning selection



(a) DDR history

(b) Rewards for Q-learning selection

(c) States of Q-learning selection

(d) Rewards for random selection

(e) States of random selection

Fig. 25 (See legend on next page.)

(See figure on previous page.)
Fig. 25 Rewards for Q-learning and random channel selection according to DDR change. Represents the rewards and visits of states according to the changes in DDR as shown in **a**. Comparing **b** and **d,** the rewards of Q-learning selection are more stable than that of the random selection. From **c** and **e,** we can see that the ad-hoc CH selects a low data rate channel and Q-learning visits the state of low DRE in band group 1 when the DDR is low. Furthermore, we can see that the Q-learning visits a state of low DRE of band group 2 when the system selects a high data rate channel by the explorer policy. When the DDR is high, the Q-learning tries to select the channel of band group 2 mainly which provides higher data rates so that the sates of band group 1 are visited less frequently. However, in random channel selection, the visits of states are distributed evenly in various DREs when the DDR is low or high. **c** and **e** shows that the visiting states of Q-learning and random channel selection are the same for a particular DDR. However, since the Q-learning channel selection tries to select a channel adaptive for the specific DDR, Q-learning mainly visits the state of the band group 1 when the DDR is low and visits the state of the band group 2 when the DDR is high. From these results, the proposed Q-learning selects the appropriate channel even if the DDR changes

highest Q-value in domain (1) which represents a low DRE for band group 2. As a result, the number of visits to state 5 belonging to the low DRE of band group 2 is the highest, as shown in Fig. 13b.

Figure 14 shows the change of rewards, states, and actions according to iteration with a medium-level DDR at 90 kbps. In Fig. 14a, the reward is stable at more than 10 iterations, and we can see that the reward is temporally low in the overall interval by random action, similarly to Fig. 12. As shown in Fig. 14b, the agent mainly visits the state 5. Figure 14c reveals that actions in band group 2 are selected mostly.

Figure 15a shows the value of the Q-table in scaled colors when the DDR is 3.5 Mbps, and b shows the number of visits to each state in the Q-table. Comparing with the supportable data rate of channels in Table 1, the DDR of 3.5 Mbps makes the DRE belong to $[1, \infty)$ of the band group 1 in Fig. 8 when the CH selects the channel from band group 1, and belongs to $[r_1, \infty)$ when a channel is selected from band group 2. However, the reward for utilization R_{Util} remains as 0 since there are no other channels to move out. As illustrated in Fig. 10c about the example of high DDR case, the same explanation can be given about Fig. 15a. At first, update process is assumed starting from domain (1) in Fig. 15a by the channel selection from band group 2. If the channel of band group 1 is selected by the explorer policy of Q-learning, the update process moves to domain (2) after the change the Q-value of domain (1). The state changes to the state 4 which represents high DRE in the band group 1 by the given DDR and the channel selection of band group 1. Therefore, after the Q-value of domain (2) is updated, the update process moves to domain (3). The update process selects the channel of band group 2 through the best or random channel selection and could be moved to domain (4), thereby the Q-value of domain (3) is updated. Finally, since the channel selection of band group 2 changes state to high state of band group 2, the update process moves to domain (1) and the Q-value of domain (4) is updated. As described for high DDR case in Fig. 10c, the CH of Fig. 15 also tends to select the channel of band group 2 and stay on the state which has high DRE of band group 2 since the channels of this band group gives no harmful effect on R_{Util}. In the Q-table of Fig. 15a, the Q-value of domain (1) showing high DRE in band group 2 is the topmost, and this is also shown in Fig. 15b as the high visit count of states 7 and 8.

Figure 16 represents the change of rewards, states, and actions according to iteration at a high level DDR of 3.5 Mbps. In Fig. 16a, the reward is stable overall, while it is temporally low in the overall interval by random action, similar to Figs. 12 and 14. In Fig. 16b, the system visits state 4 to a degree, but it mainly remains in states 7 and 8. Figure 16c shows that actions in band group 2 are selected mostly.

These results demonstrate that the proposed system can select an appropriate channel according to the DDR required by ad-hoc CR users.

5.2 Reward reconfiguration with weights

Figure 17 shows the average operation time, average data rate, and reward for channel utilization by changing the weight assignment for DDR = 40 kbps. In the reward calculation, if weights for [operation time, supportable data rate, reward for channel utilization] are assigned to [0.5, 0.1, 0.3], then it increases the importance for the operation time. As a result, it has the best average increase in operation time, as shown in Fig. 17a, and the least average of data transmission rate, as shown in Fig. 17b. This is because the system wants to reserve the highest priority for operation time and the least for data transmission rate. If weights are assigned to [0.1, 0.7, 0.1], then it increases the data transmission rate. However, it results in the lowest average operation time and reward for channel utilization because they are less important for consideration. This weight shows the highest average transmission data rate in Fig. 17b. Therefore, it is possible to operate the CR system according to the purposes of user by changing the weight assignment.

5.3 Performance comparison for the proposed Q-learning

In this section, we compare the channel selection performance between the proposed Q-learning mechanism and random selection from the available channel lists in terms of obtained rewards, average data rate of the secondary systems, and channel utilization efficiency. We

also consider the fairness of selfish channel selection without considering channel utilization efficiency.

Figure 18 shows the average reward for each DDR case. When the DDR is 10 kbps or 50 kbps, the random channel selection has a lower reward than Q-learning because the random channel selection causes a waste of channel resources and obtains the low reward for channel utilization R_{Util}. In case the DDR is 1.5 Mbps or 3.5 Mbps (e.g., more than medium or high DDR case), a channel providing a sufficient data rate is not selected adaptively by random channel selection, leading to a lower reward than Q-learning channel selection. The rewards for a DDR of 1.5 Mbps and 3.5 Mbps is lower than those of 10kbps and 50 kbps. As shown in DRE range of $[r_2, 1)$ and $[1, \infty)$ in Fig. 8, the ε-greedy policy in cases of a high DDR causes very low reward for channel utilization R_{Util} due to the selection of a low-band channel which support insufficient data rate and these effects are accumulated in the Q-table. These results show that the Q-learning channel selection adaptively selects the channel for the overall DDR.

Figures 19 and 20 show the boxplot and mean value for the data rate and reward for channel utilization resulting in Q-learning and random channel selection. The red line represents the median value, and a star denotes the mean value of the data. Figure 19 shows the mean and boxplot of the data transmission rate for Q-learning and random channel selection for each DDR. In cases of the random channel selection, the average data transmission rate of all DDR cases is the same as the average value of the data rates for all channels in Table 1 belonging to band groups 1 and 2. The boxplots of all DDR cases for the random channel selection have a similar distribution, as well. The DDR of 10 kbps and 50 kbps by Q-learning channel selection have similar distributions, and the mean for 50 kbps Q-learning is higher than that of 10 kbps, since a higher DDR attempts to choose the channel supporting higher data transmission rate. In case of a DDR for 1.5 Mbps and 3.5M bps, as in Fig. 19b, the distribution and average value of the data transmission rate by the Q-learning channel selection are higher than that of 10 kbps and 50 kbps since the channels are mainly selected from band group 2.

We can see that the Q-learning channel selection can select the channel which provides higher data transmission rate when the DDR is 3.5 Mbps than 1.5 Mbps from the mean values of each DDR case.

Figure 20 shows the mean and boxplot of the reward for the channel utilization by the Q-learning and random channel selection at each DDR. The reward for channel utilization mainly operates as a harmful value in the total reward function when the ad-hoc CH chooses an appropriate channel for its DDR. For all DDRs, the boxplot of Q-learning is denser and distributed at higher values than that of the random channel selection, and it has a higher average value since Q-learning tries to choose the channel that does not create harm in terms of R_{Util}. Outlier values of Q-learning cases are generated by random selection.

Figure 21 shows the average fairness of data rate efficiency with increasing number of neighboring secondary systems. To compare the fairness between secondary systems, two compared methods are considered in this experiment: (i) the random selection, in which the operating band and channel are selected randomly by each secondary system from its available channels and (ii) MaxQ-selection [29–32], in which each secondary system selects the channel that has the maximum supportable data rate. As we can see in Fig. 21, the proposed method provides the highest fairness because it selects the band and channel based on the desired traffic demand and current channel utilization efficiency. Therefore, if a secondary system needs relatively low data rate, then it will select the band that has a low channel bandwidth in the proposed system and it yields the bands with wider channel bandwidth to the neighbor secondary systems that require higher data rates.

Figures 22, 23, and 24 show the rewards of Q-learning and random channel selection according to an iteration for each DDR. In the case of Q-learning for all DDRs, the fluctuation decreases over time and the system operates with the intended reward design. In the random selection, there are more notches and fluctuation than Q-learning channel selection.

Figure 25 represents the rewards and visits of states according to the changes in DDR as shown in Fig. 25a. Comparing Fig. 25b and d, the rewards of Q-learning selection are more stable than those of the random selection. From Fig. 25c and e, we can see that the ad-hoc CH selects a low data rate channel and Q-learning visits the state of low DRE in band group 1 when the DDR is low. Furthermore, we can see that the Q-learning visits a state of low DRE of band group 2 when the system selects a high data rate channel by the explorer policy. When the DDR is high, the Q-learning tries to select the channel of band group 2 mainly which provides higher data rates so that the states of band group 1 are visited less frequently. However, in random channel selection, the visits of states are distributed evenly in various DREs when the DDR is low or high. Figure 25c and e show that the visiting states of Q-learning and random channel selection are the same for a particular DDR. However, since the Q-learning channel selection tries to select a channel adaptive for the specific DDR, Q-learning mainly visits the state of the band group 1 when the DDR is low and visits the state of the band group 2 when the DDR is high. From these results, the

proposed Q-learning selects the appropriate channel even if the DDR changes.

6 Conclusions

In this paper, we propose a band group and channel selection method considering the consecutive channel operation time, data transmission rate, channel utilization efficiency, and cost of the band group change for a cognitive radio ad-hoc network composed of CH and MNs. The proposed method uses the Q-learning in order to operate in a channel environment that varies dynamically according to the geographical region, time zone, band group, channel, and primary user's activity. As the core of the Q-learning operation, a Q-table and reward function consisting of an action and state are designed to consider various parameters related to the channel selected by the CR ad-hoc system. In particular, the reward for channel utilization is designed to select the appropriate band and channel so that the frequency resources are not wasted and a CR ad-hoc system can coexist with other CR systems with fair resource utilization efficiency. The simulation results represent how the proposed system selects an adaptive band and channel for the required data rate and also explain the principle of operation through the change of action and state in Q-table. It also can be confirmed that the system operates according to the intended purpose through the weight change, and the channel is selected adaptively when the required transmission rate is changed. These simulations clearly demonstrate these advantages of the proposed method.

7 Methods/experimental

The purpose of this paper is to select the band and channel for a cognitive ad-hoc system to move when the primary user appears in the channel used by the CR system and is to consider the fairness with other systems in selecting the channel. The characteristics of frequency resources such as an available transmission opportunity and data rate vary depending on the time zone, geographical location, and band group, and the activity of the primary user and the desired data rate of the secondary user are also different according to them. Therefore, considering such a complicated environment, it is necessary to select a band and a channel that can maximize the performance of the system. In this paper, the Q-learning is used to dynamically select the band and channel according to the complex surrounding environment which is time-varying. The reward function of the Q-learning is designed considering the available channel use time, data rate, utilization efficiency of the selected channel, and cost for band change. Each of the considered terms is combined with a weight sum so that the performance related to the preferred parameters can be properly realized according to the adjustment of the weights. In particular, we designed a reward for utilization in the reward function so that the CR ad-hoc system does not choose a channel that provides unnecessarily high data rate and other system has the opportunity of selecting adaptive channel which supports adaptive high data rate. The Q-table is designed so that the reward function of Q-learning works properly. The state of the Q-table is composed of time zone, geographical zone, band group, and data rate efficiency (DRE) so that the proposed Q-learning can operate well.

Experimental results in this paper had been performed using MATLAB R2015b on Intel® Core i7 3.4 GHz system. The Gaussian random function to generate the operation time and supportable data rate of each channel over time and Q-table matrix for Q-learning can be made by constructing appropriate MATLAB code.

Abbreviations

BS: Base station; CBG: Current band group; CH: Cluster head; CR: Cognitive radio; D2D: Device-to-device; DDR: Desired data rate; DRE: Data rate efficiency; FCC: Federal Communications Commission; HF: High frequency; LoS: Line-of-sight; MN: Mobile node; NLoS: Non-line-of-sight; PU: Primary user; QoE: Quality of experience; QoS: Quality of service; RF: Radio frequency; SINR: Signal-to-interference-plus-noise ratio; SMDP: Semi-Markov decision process; SNR: Signal to noise ratio; UHF: Ultra-high frequency; V2I: Vehicle-to-infrastructure; V2V: Vehicle-to-vehicle; VANET: Vehicular ad-hoc network; VHF: Very high frequency

Dataset of simulations

The simulation was performed using MATLAB in Intel Core i7 (32bit).
The operation time and supportable data rate is made of the mean and variance in Table 1 by Gaussian function using MATLAB.
The Q-table is made up of tables as defined in the paper, and it works according to the Q-table update equation.

Funding

This work was supported by a grant-in-aid of Hanwha Systems and the Agency for Defense Development (ADD) in the Republic of Korea as part of the Contract UC160007ED.

Availability of data and materials

Not applicable.

Authors' contributions

All authors contribute to the concept, the design and developments of the theory analysis and algorithm, and the simulation results in this manuscript. All authors read and approved the final manuscript.

Authors' information

- Prof. Sang-Jo Yoo, PhD (Corresponding author): Sang-Jo Yoo received the B.S. degree in electronic communication engineering from Hanyang University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, in 1990 and 2000, respectively. From 1990 to 2001, he was a Member of the Technical Staff in the Korea Telecom Research and Development Group, where he was involved in communication protocol conformance testing and network design fields. From 1994 to 1995 and from 2007 to 2008, he was a Guest Researcher with the National Institute Standards and Technology, USA. Since 2001, he has been with Inha University, where he is currently a Professor with the Information and Communication Engineering Department. His current research interests include cognitive radio network protocols, ad-hoc wireless network, MAC and routing protocol design, wireless network QoS, and wireless sensor networks.

- Mr. Sung-Jeen Jang: Sung-Jeen Jang received a B.S degree in electrical engineering from Inha University Incheon, Korea, 2007. He received his M.S. degree in Graduate School of Information Technology and Telecommunication, Inha University, Incheon Korea, 2009. Since March 2009, he has been pursuing a Ph.D. degree at the Graduate School of Information Technology and Telecommunication, Inha University, Incheon Korea. His current research interests include cognitive radio network protocols and machine learning applied wireless communications.
- Dr. Chul-Hee Han: Chulhee Han received the B.S. degree in Electronic Engineering from Chung-ang University, Korea, in 1997, and M.S. and Ph.D. degrees in Electrical and Electronic Engineering from Yonsei University, Korea, in 1999 and 2007, respectively. Currently, he is working at Hanwha Systems, Korea, as a chief engineer. He was involved in various projects including tactical mobile WiMAX system and tactical LOS PMP radio. His research interests include Tactical Broadband Communications, Combat Network Radio, and Cognitive Radio for Military Applications.
- Dr. Kwang-Eog Lee: Kwang-Eog Lee received the B.S. and M.S. degrees in electronic engineering from Kyungpook National University, Daegu, South Korea, in 1988 and 1990, respectively. He has been working in Agency for Defense Development since 1990. From 2007 to 2008, he was an exchange scientist in CERDEC (Communications-Electronics Research, Development and Engineering Center) U.S. Army. Currently, he is a principal researcher and his research interests include cognitive radio and terrestrial and satellite tactical communication.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Information and Communication Engineering, Inha University, 253 YongHyun-dong, Nam-gu, Incheon, South Korea. [2]Hanwha Systems 188, Pangyoyeok-Ro, Bundang-Gu, Seongname-Si, Gyeonggi-Do 13524, South Korea. [3]Agency for Defense Development, P.O.Box 35, Yuseong-Gu, Daejeon, South Korea.

Received: 18 April 2018 Accepted: 16 April 2019
Published online: 24 May 2019

References
1. R. Marsden, B. Soria, H.M. Ihle, *Effective Spectrum Pricing: Supporting Better Quality and more Affordable Mobile Services* (GSMA, London, UK, 2017) Tech. Rep
2. FCC, ET Docket No 03-222, Notice of Proposed Rule Making and Order. (2003)
3. FCC, ET Docket No 03-237, Notice of Proposed Rule Making and Order. (2003)
4. J. Miltola, *Cognitive radio: An Integrated Agent Architecture for Software Defined radio* (Doctor of Technology, Royal Inst. Technol. (KTH), Stockholm, 2000)
5. I.F. Akyildiz et al., Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Comp. Networks J. **50**(13), 2127–2159 (2006)
6. Unlicensed operation in the TV broadcast bands and additional spectrum for unlicensed devices below 900 MHz and in the 3 GHz band, Third Memorandum Opinion and Order, (FCC 2012), http://transition.fcc.gov/Daily_Releases/Daily_Business/2012/db0405/FCC-12-36A1.pdf. Accessed 25 Jun 2018
7. Report ITU-R M.2330–0 (11/2014) ITU-R M.2330–0 Cognitive radio Systems in the Land Mobile Service
8. DARPA Developing Cognitive radio IC Technology for Next-Gen Communications, Radar, and Electronic Warfare, (IDST 2018) www.darpa.mil/program/computational-leverage-against-surveillance-systems. Accessed 25 Jun 2018
9. DARPA, DARPA-BAA-11-61, Computational Leverage against Surveillance Systems (CLASS), (2011)
10. The Spectrum Collaboration Challenge. (DARPA, 2018), https://spectrumcollaborationchallenge.com. Accessed 25 Jun 2018
11. M. Vishram, L.C. Tong, C. Syin, A channel allocation based self-coexistence scheme for homogeneous ad-hoc networks. IEEE Wireless Commun. Lett. **4**(5), 545–548 (2015)
12. S. Maghsudi, S. Stańczak, Hybrid centralized–distributed resource allocation for device-to-device communication underlaying cellular networks. IEEE Trans. Veh. Technol. **65**(4), 2481–2495 (2016)
13. Y. Han, E. Ekici, H. Kremo, O. Altintas, Throughput-efficient channel allocation algorithms in multi-channel cognitive vehicular networks. IEEE Trans. Wirel. Commun. **16**(2), 757–770 (2017)
14. M. Li, L. Zhao, H. Liang, An SMDP-based prioritized channel allocation scheme in cognitive enabled vehicular ad hoc networks. IEEE Trans. Veh. Technol. **66**(9), 7925–7933 (2017)
15. O. Semiari, W. Saad, M. Bennis, Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations. IEEE Trans. Wirel. Commun. **16**(7), 4802–4816 (2017)
16. L. Liang, J. Kim, S.C. Jha, K. Sivanesan, G.Y. Li, Spectrum and power allocation for vehicular communications with delayed CSI feedback. IEEE Wireless Commun. Lett. **6**(4), 458–461 (2017)
17. P. Simon, *Too Big to Ignore: The Business Case for Big Data* (Wiley, (2013)
18. M.L. Littman, Reinforcement learning improves behavior from evaluative feedback. Nature **521**(7553), 445–451 (2015)
19. A. Asheralieva, Y. Miyanaga, An autonomous learning-based algorithm for Joint Channel and power level selection by D2D pairs in heterogeneous cellular networks. IEEE Trans. Commun. **64**(9), 3996–4012 (2016)
20. M. Srinivasan, V.J. Kotagi, C.S.R. Murthy, A Q-learning framework for user QoE enhanced self-organizing spectrally efficient network using a novel inter-operator proximal spectrum sharing. IEEE J. Sel. Areas Commun. **34**(11), 2887–2901 (2016)
21. K. Salma, S. Reza, G.S. Ali, Learning-based resource allocation in D2D communications with QoS and fairness considerations. Eur. Trans. Telecommun. **29**(1), 2161–3915 (2017)
22. E. Fakhfakh, S. Hamouda, Optimised Q-learning for WiFi offloading in dense cellular networks. IET Commun. **11**(15), 2380–2385 (2017)
23. M. Yan, G. Feng, J. Zhou, S. Qin, Smart multi-RAT access based on multiagent reinforcement learning. IEEE Trans. Veh. Technol. **67**(5), 4539–4551 (2018)
24. V. Maglogiannis, D. Naudts, A. Shahid, I. Moerman, A Q-learning scheme for fair coexistence between LTE and Wi-Fi in unlicensed spectrum. IEEE Access **6**, 27278–27293 (2018)
25. N. Xu, H. Zhang, F. Xu, Z. Wang, Q-learning based interference-aware channel handoff for partially observable cognitive radio ad hoc networks. Chin. J. Electron. **26**(4), 856–863 (2017)
26. S.J. Jang, S.J. Yoo, *Reinforcement learning for dynamic sensing parameter control in cognitive radio systems* (IEEE ICTC, 2017), pp. 471–474
27. Y.Y. Liu, S.J. Yoo, *Dynamic resource allocation using reinforcement learning for LTE-U and WiFi in the unlicensed spectrum* (IEEE ICUFN, 2017), pp. 471–475
28. L. Shi, S.J. Yoo, Distributed fair resource allocation for cognitive femtocell networks. Wireless Personal Commun. **93**(4), 883–902 (2017)
29. Fairness measure, (Wikipedia 2018) https://en.wikipedia.org/wiki/Fairness_measure. Accessed 25 Jun 2018
30. Q. Zhang, Q. Du, K. Zhu, *User Pairing and Channel Allocation for Full-Duplex Self-Organizing Small Cell Networks* (WCSP, Nanjing, 2017), pp. 1–6
31. M. Yan, G. Feng and S. Qin, Multi-RAT access based on multi-agent reinforcement learning. IEEE GLOBECOM, 1–6 (2017)
32. F. Zeng, H. Liu and J. Xu, Sequential channel selection for decentralized cognitive radio sensor network based on modified Q-learning algorithm. ICNC-FSKD, 657–662 (2016)