




# A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks

Yonghua Wang<sup>1,2</sup>  · Zifeng Ye<sup>1</sup> · Pin Wan<sup>1</sup> · Jiajun Zhao<sup>1</sup>

Published online: 22 June 2018  
© Springer Nature B.V. 2018

## Abstract

Cognitive radio is an emerging technology that is considered to be an evolution for software device radio in which cognition and decision-making components are included. The main function of cognitive radio is to exploit “spectrum holes” or “white spaces” to address the challenge of the low utilization of radio resources. Dynamic spectrum allocation, whose significant functions are to ensure that cognitive users access the available frequency and bandwidth to communicate in an opportunistic manner and to minimize the interference between primary and secondary users, is a key mechanism in cognitive radio networks. Reinforcement learning, which rapidly analyzes the amount of data in a model-free manner, dramatically facilitates the performance of dynamic spectrum allocation in real application scenarios. This paper presents a survey on the state-of-the-art spectrum allocation algorithms based on reinforcement learning techniques in cognitive radio networks. The advantages and disadvantages of each algorithm are analyzed in their specific practical application scenarios. Finally, we discuss open issues in dynamic spectrum allocation that can be topics of future research.

**Keywords** Reinforcement learning · Spectrum allocation · Cognitive radio networks

---

✉ Yonghua Wang  
sjzwyh@163.com; wangyonghua@gdut.edu.cn

Zifeng Ye  
littlecommayezifeng@gmail.com

Pin Wan  
wanpin2@163.com

Jiajun Zhao  
1243588305@qq.com

<sup>1</sup> School of Automation, Guangdong University of Technology, Guangzhou 510006, China

<sup>2</sup> State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

## 1 Introduction

The concept of cognitive radio (CR) was first proposed by Mitola and Maguire (1999) and Iii (2000) to exploit unused portions of the spectrum. He stated that CR can utilize intelligent computing of wireless personal digital devices and related networks to detect users' communication requirements. According to the definition given by the Federal Communications Commission (FCC) (Iii 2000), CR is capable of interacting with the communication environment and changing its own transmission parameters based on interactive information.

In a time-varying environment, a good framework for adaptive network configuration based on CR should address the following issues (Wang et al. 2016):

- (1) How to properly configure the transmission parameters when the network model and environmental observability are limited?
- (2) How to coordinate the distribution of transmission equipment with limited resources?
- (3) How to ensure the convergence of the network under the condition of conflict between transmission equipment?

Dynamic spectrum allocation (DSA), which is defined as a technique for determining the optimal mapping between the available licensed channels and cognitive radios to achieve optimal performance, is one of the basic mechanisms to control interference in cognitive radio networks (CRNs) (Ahmed et al. 2014). The objectives of DSA in CRNs are:

- (1) To assign the available channels to cognitive users to achieve efficient spectrum utilization.
- (2) To reduce the interference among cognitive users.
- (3) To minimize the interference between licensed users and cognitive users.

In recent years, researchers have systematically studied and published reviews on spectrum allocation (SA) concerning the technical aspects of CR (Ahmed et al. 2014; Marinho and Monteiro 2012; Tragos et al. 2013; Bkassiny et al. 2013; Zhang et al. 2013; Ahmad et al. 2015; Wang et al. 2016; Yau et al. 2012; Al-Rawi et al. 2015). Interestingly, more and more researchers have recently proposed a number of different approaches, such as game theory (Yang et al. 2010), fuzzy logic (Le and Ly 2008), graph theory (Zhao et al. 2008; Chen et al. 2011), linear or non-linear programming (Shu and Krunz 2010; Ru et al. 2017; Salameh 2011), heuristics (Yu et al. 2010), markov random field (Anifantis et al. 2012) and evolutionary algorithms (Cheng and Jiang 2011), to solve specific problems and especially to explore the application of reinforcement learning (RL) in SA (Fig. 1).

The RL technique is a major category of machine learning that enables agents without any experience to continuously learn by trial-and-error and to maximize the reward function or obtain the optimal strategy. There have been several practical applications of RL, including Atari, AlphaGo and AlphaZero.

RL can be used to solve large-scale (having large state and action spaces) problems with complex Markov decision process (MDP). Two main features of RL are:

*Trial-and-error* The early RL technology was a typical trial-and-error learning system, in which agents implement actions to explore the environment without any prior knowledge. During the trial-and-error process, the key technical issue is the balance between exploration and development. The representative and groundbreaking model-free RL algorithms are temporal difference (TD) learning and Q-learning (Sutton and Barto 1998). Current RL methodology developed from these algorithms.

*Delayed rewards* Feedback signals are received by the agents from the environment after taking actions. RL tasks can be categorized into two types according to whether the decision-making tasks are sequential (Qadir 2016). Non-sequential tasks emphasize the consideration

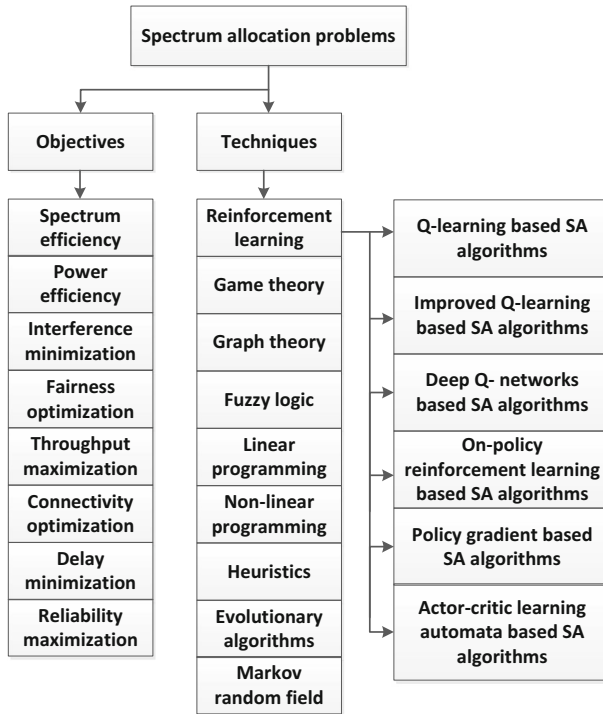


Fig. 1 Taxonomy of SA problems

of current rewards. In other words, the goal is to maximize the rewards from state to action immediately. By contrast, sequential tasks are more challenging and aim to maximize the long-term rewards from a series of actions.

There are some advantages of RL for SA (Wang et al. 2016):

- (1) Effective monitoring of the real-time dynamic environment in a friendly manner.
- (2) Continuous learning of new knowledge to adapt to various extreme environments.
- (3) Easy simulation of a complex environment that is difficult to model accurately.
- (4) Access to important data and information that are difficult to find and improved DSA performance.
- (5) Promotion of breakthroughs in heterogeneous networks of information exchange to improve network performance.

In contrast to the previously mentioned surveys, our investigation focuses on different RL methods in SA.

## 2 System model

Clearly, the proposed algorithms are designed to solve specific problems [such as throughput and signal-to-interference-plus-noise ratio (SINR)] in SA. Various methods, including estimation technique, game theory, evolutionary computation, fuzzy logic, MDP, pricing theory, theory of social science and RL, have been applied to build all types of mathematical models of SA (Li et al. 2010). Our review focuses on the research and application of RL in SA.

Therefore, RL is not only the method used to construct the models but also the solution to various challenges.

SA technology can be divided into different categories according to various standards. Considering the method of access to spectrum resources, the technology can generally be classified as spectrum underlay or spectrum overlay. The network topology structure can be divided into two types: distributed and centralized spectrum allocation. Furthermore, the cooperation mode can be classified as non-cooperative or cooperative spectrum allocation (Zhao and Sadler 2007).

(1) Spectrum underlay and spectrum overlay

Spectrum resource sharing between licensed users and cognitive users can be divided into two types: spectrum underlay and spectrum overlay. Cognitive users have different restrictions in the SA process; therefore, there are two different types of spectrum resource sharing.

*Spectrum underlay* Ultra-wide band (UWB) is generally used as a spread spectrum technology from which cognitive users can access and use the same frequency spectrum as that of licensed users. Therefore, the frequency can be covered completely.

*Spectrum overlay* The spectrum resources may be accessed opportunistically by cognitive users in the absence of licensed users. Moreover, if licensed users occupy their own spectrum bandwidth, cognitive users should trade off immediately.

(2) Distributed and centralized spectrum allocation

*Distributed spectrum allocation* Each cognitive user needs to detect whether there are licensed users in the real-time environment. Then, they detect the information according to the cooperation between cognitive users and implement SA in combination with the strategy of spectrum resource sharing.

*Centralized spectrum allocation* The central controller coordinates and manages cognitive users. Therefore, the central controller can allocate idle spectrum to cognitive users (Hossain and Bhargava 2007).

(3) Non-cooperative and cooperative spectrum allocation

*Non-cooperative spectrum allocation* The pattern of non-cooperative allocation of spectrum resources can be described as follows. A single cognitive user observes the surrounding environment and then makes spectrum decisions, which do not recognize the mutual information exchange between users.

*Cooperative spectrum allocation* Cooperative spectrum resource allocation can occur not only between cognitive users and licensed users but also between cognitive users. Efficient sharing of spectrum resources is achieved through information exchange; thus, it can increase the utilization of spectrum resources and avoid interference throughout the whole system.

Nie and Haykin (1997) are the pioneers in applying RL to SA problems. In this paper, the authors developed a mathematical model for DSA problems based on the Q-learning algorithm, which is the best-known algorithm among RL techniques. Here, we present a brief description because all types of RL methods are developed based on this model.

Nie and Haykin (1997) regarded the mobile communication system as a discrete-time event system. To take advantage of the learning scheme, it is necessary to develop the DSA as a dynamic programming problem or, equivalently, to determine the system state  $x$ , the behavior  $a$ , the related cost  $r$ , and the next state  $y$ .

*State* It is assumed that there are  $N$  cells and  $M$  available channels in the mobile communication system.  $x_t = (i, A(i))_t$  is the definition of the state  $x_t$  at time  $t$ , where  $i \in \{1, 2, \dots, N\}$  is called the cell index, indicating that a call arrival event or departure event occurs for cell  $i$ .  $A(i) \in \{1, 2, \dots, M\}$  represents the number of available channels in cell  $i$  at time  $t$ , which depends on the channel usage in cell  $i$  and its interfering cells  $I(i)$ . To obtain  $A(i)$ , the channel state of cell  $q$  is defined as an  $M$ -dimensional vector with each component  $u_{qk}$  defined

as

$$u_{qk} = \begin{cases} 1, & \text{if channel } k \text{ is in use in cell } q \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where  $q = 1, 2, \dots, N, k = 1, 2, \dots, M$ .

Furthermore, an available vector  $s_q = \{0, 1\}^M$  is formed as  $s_q \in (s_{q1}, s_{q1}, \dots, s_{qM})$ , with each component  $s_{qk}$  defined as

$$s_{qk} = \begin{cases} 0, & \text{if channel } k \text{ is available for use in cell } q \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

where  $q = 1, 2, \dots, N, k = 1, 2, \dots, M$ .

We know that the channel state of cell  $i$  and its interfering cells  $j \in I(i)$  and the availability vector  $s_i$  can easily be constructed. The corresponding elements are  $s_{ik} = \max\{u_{qk} | q \in \bar{I}(i)\}, k = 1, 2, \dots, M, \bar{I}(i) = I(i) \cup i$ . From  $s_{ik}$ , it is easy to obtain  $A(i) = \sum_{k=1}^M \bar{s}_{ik}$ , where  $\bar{s}_{ik}$  is the logical negation of  $s_{ik}$ .

**Actions** Action  $a$  is defined by assigning a channel  $k$  from the available channel  $A(i)$  to the current call request in cell  $i$ , that is,  $a = k, k \in \{1, 2, \dots, M\}$  and  $s_{ik} = 0$ .

**Cost** Cost  $r(x, a)$  is used to represent the immediate cost of taking action  $a$  in state  $x$ , that is,  $r(x, k) = n_1(k)r_1 + n_2(k)r_2 + n_3(k)r_3$ , where  $n_1(k)$  is the number of compact cells in reference to cell  $i$  in which channel  $k$  is being used.  $n_2(k)$  is the number of cochannel cells, and  $n_3(k)$  is the number of other cochannel cells currently using channel  $k$ .

Many methods can be used to define the cost function depending on which parameters are the main consideration.

**Next state** State transitions from  $x_t$  to  $x_{t+1}$  are determined by two random events (call arrivals and call departures).

### 3 RL-based SA algorithms

#### 3.1 Q-learning based SA algorithms

Among RL techniques, the Q-learning algorithm is the most widely used. The algorithm constantly updates the values in the Q table that represent states and then determines how to select an action according to the current state. Q-learning is an off-policy algorithm because the Q table updates values based not only on current experience but also according to past or even others' experience.

Lv et al. (2013) proposed an algorithm based on Q-learning to improve the performance of DSA. Because of the characteristics of the Q-learning search strategy, cognitive users do not always choose the channel with the largest values but instead select another channel with a given probability to explore the environment for the optimal long-term rewards. The algorithm not only achieves the autonomy of channel and power allocation but also improves the data throughput and channel efficiency.

Yang and Grace (2011) presented a random picking distributed SA scheme exploiting RL in multicast terrestrial communication systems that can significantly improve the power adjustment by limiting the reassignment and dropping rates. However, this improvement is achieved at the expense of a higher blocking rate.

Alsarhan and Agarwal (2011) derived a resource assignment scheme to help primary users (PUs) optimally allocate the offered spectrum among different classes of secondary

users (SUs) while maximizing the PUs' rewards. The Q-learning algorithm is utilized to extract the optimal control policy and to manage the spectrum dynamically. The cognitive wireless mesh network is able to support the additional SUs' traffic while guaranteeing the PUs' QoS.

The radio resource management problem in public femtocell networks has been fully studied in Li et al. (2011). Using an ingenious resource allocation technique, that is, the Q-learning based algorithm, multiple femtocells can be configured optimally in real time.

In Teng et al. (2010), an auction-based Q-learning (QL-BA) algorithm was proposed for a primary user (PU) and multiuser (OPMS) scenario. The PUs in the model know all the information of the SUs, but the SUs know only their own information. The PUs and SUs can transmit information through the public channel. The packet loss rate, allocation efficiency and transmission rate performance are greatly improved.

However, in Teng et al. (2010), the learning strategies of primary and secondary users are the same, which is not fair for SUs who know only their own information. Therefore, in Teng et al. (2013), a Q-learning algorithm based on a double auction (QL-DA) is proposed. This algorithm reflects the users' selfishness in practical applications.

### 3.2 Improved Q-learning-based SA algorithms

Although Q-learning has been applied to SA, the common problem with this algorithm is that it involves careless consideration of the exact information and channel transmission conditions of user behavior. Considering the coexistence of sensor nodes and other wireless functional devices that may share the spectrum, Faganello et al. (2013) developed three improved Q-learning algorithms for SA in distributed dynamic industrial CRNs. The first algorithm, called Q-learning+, uses accurate channel occupancy information to learn to improve the channel allocation decisions. The second algorithm, named Q-noise, evaluates channel transmission conditions by analyzing the signal-to-noise ratio. The third algorithm, Q-Noise+, integrates Q-learning+ and Q-noise into a single framework to consider both accurate channel occupancy information and channel transmission conditions. Experiments show that the performance of the three algorithms is improved compared with that of the traditional Q-learning algorithms. However, the improved algorithms require accurate channel-specific information and therefore increase the overhead of the channels.

Feng et al. (2009) proposed a heterogeneous network self-optimization algorithm (HNSA) that achieved self-optimization of a dynamic network based on an enhanced Q-learning method. This algorithm regards the intelligent self-optimizing controller (ISOC) of each radio access network as an independent agent to transform the adaptation problem in the reconfigurable systems into a multiagent RL problem. This method reduces the system blocking rate and the complexity in DSA.

Focusing on 5G mobile communications, an anti-jamming MIMO NOMA power allocation and downlink transmission scheme was proposed in Xiao et al. (2017). Without knowing the channel and jamming environment, the NOMA system is modeled as a zero-sum game in which two mechanisms are built innovatively. On one hand, the Dyna Architecture, which can be regarded as constructing an experience replay, accelerates the speed of obtaining the optimal strategy. On the other hand, the hotbooting technique, which uses a fast Q-learning algorithm to obtain the initialization value, improves the convergence compared to that of the zero initialization. As a result, the system performance is greatly improved.

The algorithms described thus far are all Q-learning algorithms or their variants, that is, minor improvements made to Q-learning algorithms. Therefore, we present a simple summary

in Table 1. In the following sections, we will describe the various algorithms proposed for the combination of RL and other methodologies.

### 3.3 Deep Q-networks-based SA algorithms

In recent years, neural networks have received considerable attention. Mnih et al. (2013) combined the advantages of Q-learning with neural networks and proposed deep Q-networks (DQN) in 2013. The biggest advantage of DQN is the use of neural networks, rather than Q table, to estimate the state values, in which all states and actions must be stored. Facing a large amount of data, DQN can overcome the limitations of computer hardware and greatly improve data access and update speed.

However, before Mnih et al. (2013), an algorithm similar to DQN was proposed and successfully applied in DSA (Yi and Hong 2012). The algorithm replaces the Q table with a multilayer forward-propagation neural network by combining Q-learning and back propagation, which reduces the external signal interference and improves the network performance. The prerequisite for application is that the historical activity of the PU is not included in channel selection, and each node must be accessed to exchange information on the available channel. Thus, this method increases the communication overhead.

Han et al. (2017) presented a two-dimensional anti-jamming communication scheme. In the absence of an interference model and radio channel model, an SU can avoid the heavy interference region and select the frequency hopping pattern by using the DQN algorithm. Specifically, SUs use Q-learning to obtain an optimal anti-jamming communication strategy and a deep convolutional neural network to accelerate the learning speed. The results show that the proposed method achieves a faster convergence rate, higher SINR, lower anti-interference cost and better performance of SUs against cooperative jammers compared with those of Q-learning alone.

### 3.4 On-policy RL-based SA algorithms

The Q-learning algorithms mentioned above are off-policy algorithms. In other words, the next state and action are uncertain when the algorithm is updated. By contrast, the most representative on-policy algorithm is the state-action-reward-state-action (SARSA) algorithm (Sutton and Barto 1998). The SARSA algorithm represent the whole cycle in a path. The next state and action are the state and action it actually takes and updates when the algorithm has been identified. If Q-learning is a greedy and brave algorithm, SARSA is a relatively conservative and timid algorithm. Because the SARSA algorithm is more sensitive to trial-and-error rate, in practical applications, if the damage is a major consideration, the SARSA algorithm can achieve better performance during training.

It is impossible to obtain an accurate MDP model for actual complex situations. Lilith and Dogancay (2005) used the readily available local environment information for learning users. Reducing the state space leads to reduced memory requirements and cost without significant performance loss. In addition, a simplified SARSA algorithm for spectrum redistribution using only local information was simulated. The new call and switching blocking probability is almost the same as the simplified SARSA using system-level incentive information, which indicates the feasibility of the distributed RL method.

**Table 1** Summary of RL-based SA algorithms with minor improvements

Method	Execution model	Network type	Number of radio	Channel model	Characteristic and limitations
Q-learning based algorithm	Distributed	CRN	Multiple	N/A	Improves the data throughput and channel efficiency; enhances the power adjustment
Q-learning based algorithm	N/A	Cooperative CRNs	Single	N/A	Increases SINR and enhances the anti-jamming performance
Q-learning based algorithm	Decentralized	Cognitive wireless mesh networks	Multiple	Overlapping	Ensures PU's QoS and maximizes the PU's profiles; supports additional SUs traffic
Q-learning based algorithm	Distributed	Public femtocell networks	Multiple	Orthogonal	Eliminates the interference between femtocells and maximizes system capacity
Q-learning based auction algorithm	Centralized	MPMS-CRN	Multiple	Orthogonal	Improves the packet loss rate, allocation efficiency and transmission rate performance
Q-learning based double auction algorithm	Decentralized	MPMS-CRN	Multiple	Orthogonal	Reflects the users' selfishness
Q-learning +	Distributed	Industrial networks	Multiple	N/A	Accurate historic behavior of the channel; considers the historic weight
Q-noise +	Distributed	Industrial networks	Multiple	N/A	Considers the channel's quality for transmission and noise weight
Q-noise	Distributed	Industrial networks	Multiple	N/A	Accurate historic behavior of the channel while considering the channel's quality for transmission
Fast Q-learning	Centralized	Wireless networks	Multiple	Non-orthogonal	Significantly increases the sum data rates of users



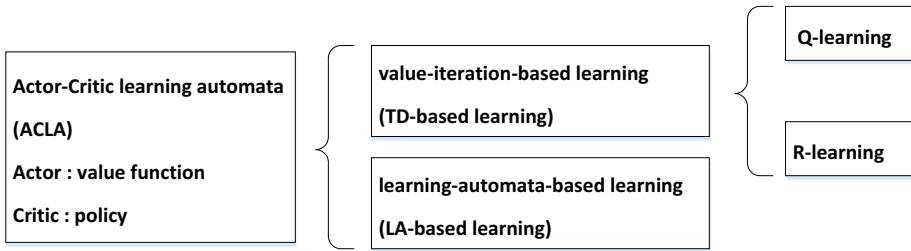


Fig. 2 Relationship among ACLA, TD-based learning and LA-based learning

### 3.5 Policy-gradient-based SA algorithms

In contrast to the value-iteration algorithms of Q-learning and SARSA, the policy gradient algorithm is based on policy iterations. Its output is not the value of the action but the specific action, which can be selected on a continuously distributed action.

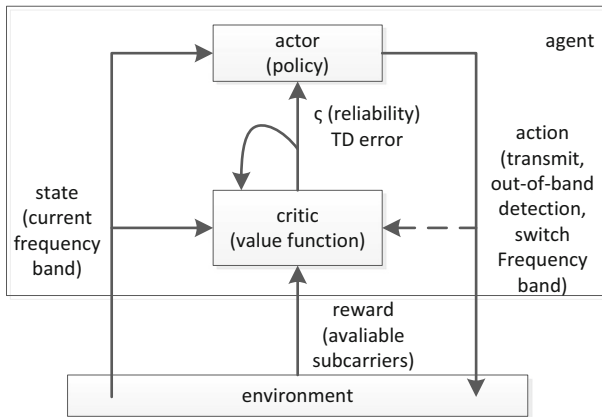
To solve the problem of DSA in CRNs, SUs’ channel access process was built as a restricted partially observable Markov decision process (POMDP) in Levorato et al. (2012). In the restricted POMDP, the reward function is used to collect SUs’ instantaneous rewards, and the cost function reflects the instantaneous cost of the PU due to channel interference from SUs. By using the Lagrangian multiplier method to convert the original constrained POMDP into an unconstrained POMDP, the learning algorithm based on the policy gradient gains the local optimal strategy.

In Tanwongvarl and Chantaraskul (2015), the authors migrated the true policy gradient algorithm, POMDP algorithm and episodic-reinforcement algorithm to a cognitive radio awareness network. The experimental results show that the true policy gradient algorithm is superior to the other two algorithms in terms of accuracy and data end-to-end delay performance.

Most RL algorithms are applied to discrete random problems, such as the usage of industrial scientific medical (ISM) band spectrum. Observation of the channel environment can determine the system environment limits and possible criteria. The next step is to calculate all possible solutions based on MDP, which analyzes the behavior of the parameters for future predictions before setting the final parameters. In short, the learning process is the most appropriate strategy for the current environment. The system then sets the initial channel as the current channel that the sensor network is using to calculate the next corresponding channel based on the obtained policy.

### 3.6 Actor-critic learning-automata-based SA algorithms

This learning algorithm, similar to Q-learning, is called the critic-only algorithm because of the power-based iterations and the lack of an explicit policy function. However, it is difficult for a critic-only algorithm to converge. The gradient learning algorithm based on learning automata (LA) allows the subject to directly learn a fixed randomization strategy, so LA can be considered to be an actor-only algorithm that does not use any form of stored value functions. This method provides a new algorithm, the actor-critic LA (ACLA) algorithm, which is based on the iterative algorithm, to study the state (action) value of MDP and the policy based on the LA algorithm (Fig. 2). The advantage of the algorithm is that it can perform updates in a single step, making it faster than the traditional policy gradient.



**Fig. 3** Interaction of the secondary user with the environment and the applied actor-critic structure (Berthold et al. 2008)

In Berthold et al. (2008), the authors proposed a multiband cognitive radio spectrum detection algorithm based on actor-critic LA. The structure of the algorithm (Fig. 3) was developed in by integrating a cross-layer optimization framework, and it is applicable to environments with dynamic spectrum resource availability.

We summarize the above methods in Table 2.

## 4 Challenges and open issues

In the previous sections, we presented an overview and a brief description of the proposed RL algorithms. Nevertheless, several open issues remain.

### *CRN problems*

- (1) *Operating frequency* For SA problems, the operating frequency is one of the most important parameters or indicators. The most notable feature of cognitive users is that they can change their operating frequency to another available operating frequency based on their location or the signal quality (such as whether the PU is busy). In other words, SUs can switch to the most suitable frequency. This characteristic needs to be considered.
- (2) *Transmission power* Transmission power is another important indicator in SA. High transmission powers of SUs are undesirable because they rapidly deplete the SUs' battery and cause more interference between the SUs and PU. To this end, the dynamic transmission power is configured within the allowable power range. The power control should also consider the actual channel conditions and the QoS requirements of the ongoing SUs. Under the power control, the transmission power is reduced to an acceptable minimum. Accordingly, the corresponding interference is reduced, which allows more users to share the available spectrum.
- (3) *Channel model* Most of the proposed SA algorithms use a Rayleigh fading channel model, and few consider other models (such as the Nakagami channel model). The existing algorithms should be tested for different channel models, and new algorithms should be proposed. For different application scenarios, CR entity communication affects the characteristics of wireless channels, whether urban or rural, indoor or out-

**Table 2** Summary of RL-based SA algorithms with major improvements

Method	Execution model	Network type	Number of radio	Channel model	Characteristic and limitations
Q-learning based network self-optimization algorithm	Distributed	Wireless local area network	Multiple	N/A	Reduces the system blocking rate; improves network revenue; obtains much lower complexity and better performance
Similar deep Q networks based algorithm	Distributed	Cognitive wireless local area network	N/A	Non-overlapping	Reduces external signal interference but increases communication overhead
Deep Q networks based algorithm	Distributed	Anti-jamming communication CRNs	Multiple	N/A	Achieves faster convergence rate, higher SINR, lower anti-interference cost and better performance of SUs against cooperative jammers
Price search algorithm	Distributed	Cognitive interference networks	Single	N/A	Improves the convergence rate; captures the long-term effect of interference
True policy gradient algorithm	Distributed	Cognitive wireless sensor networks	Multiple	N/A	High accuracy and data end-to-end delay performance
POMDP algorithm	Distributed	Cognitive wireless sensor networks	Multiple	N/A	Low accuracy and data end-to-end delay performance
Episodic-reinforcement algorithm	Distributed	Cognitive wireless sensor networks	Multiple	N/A	Moderate accuracy and data end-to-end delay performance
Actor-critic learning-automata-based algorithm	Distributed	Multi-band cognitive radio scenario	Multiple	N/A	Easy to integrate into a cross-layer optimization framework

door. Changing the adopted channel model will affect the analysis and the optimality of the algorithm performance.

- (4) *Research on distributed and mobile scenarios* The research on distributed resource allocation for bottom-level CRNs is limited compared to that on centralized algorithms. For increasingly complex communication systems, distributed solutions will have to prevail to satisfy user mobility and high QoS standards. Therefore, research in this area is useful and encouraging.
- (5) *Experimental testbeds* It is important to test these techniques in practical environments to effectively evaluate the performance of CRN technology. CR testbeds provide an effective way for researchers to accurately evaluate their results in real situations. A good simulation environment not only enhances the efficiency of the study but also substantially reduces costs. This problem has not been widely considered in the existing literature on underlying CRNs, as existing simulators lack a sophisticated CR module.

#### *Reinforcement learning problems*

(1) *Seamless integration of RL and other approaches:* Most of the above works consider only RL algorithms to solve radio resource management problems. To achieve better resource allocation or more flexible solutions, the advantages of RL models and other alternatives (such as Markov chains, auction models, and genetic algorithms) can be combined. For example, game-based analysis applies only to the agent's learning dynamics, while environmental dynamics are not explicitly considered (Busoniu et al. 2008). Therefore, it is promising to solve challenging problems by combining multiple state-of-the-art algorithms, such as the combination of RL and game theory.

(2) *Improve the applicability* DSA not only involves finding the central frequency but also selecting the frequency and optimal bandwidth the SU needs to access in a flexible and real-time manner. Some emerging RL algorithms have not been applied in CRNs and remain an open area for future research. Furthermore, although Teng et al. (2013) proposed an RL algorithm for the participation of selfish users in SA learning schemes, the challenge of how they can benefit from cooperation with other SUs is worth further investigation.

(3) *Consider large-scale and dense allocation* At present, many RL algorithms are only applied to some single DSA problems. However, the state and action space in reality is large-scale and continuous, so the current methods are not applicable. Therefore, to strengthen the methods to solve various practical problems, further research is required to improve the generalizability of the current algorithms. For example, in practical scenarios, cognitive users are large scale, and it is impractical to use traditional Q-learning with a Q table to store information. Therefore, we can combine neural networks with RL to enhance the ability to handle large amounts of data.

## 5 Conclusion

DSA is a key design issue for cognitive radio technology. In this paper, we classify the existing RL algorithms in CRNs by presenting a thematic taxonomy and a survey of state-of-the-art SA algorithms. There are six components, including Q-learning, improved Q-learning, deep Q-networks, on-policy RL, policy gradient and actor-critic learning automata. The critical aspects of the current SA algorithms are also analyzed to determine their strengths and weaknesses. Finally, we discuss several open issues and challenges that have not been fully investigated and that could be the basis for future work in this area.

**Acknowledgements** This work was supported in part by special funds from the central finance to support the development of local universities under No. 400170044, the project supported by the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences under Grant No. 20180106, the science and technology program of Guangdong Province under Grant No. 2016B090918031, the degree and graduate education reform project of Guangdong Province under Grant No. 2016JGXM\_MS\_26, the foundation of key laboratory of machine intelligence and advanced computing of the Ministry of Education under Grant No. MSC-201706A and the higher education quality projects of Guangdong Province and Guangdong University of Technology.

## References

- Ahmad A, Ahmad S, Rehmani MH, Hassan NU (2015) A survey on radio resource allocation in cognitive radio sensor networks. *IEEE Commun Surv Tutor* 17(2):888–917
- Ahmed E, Gani A, Abolfazli S, Yao LJ, Khan SU (2014) Channel assignment algorithms in cognitive radio networks: taxonomy, open issues, and challenges. *IEEE Commun Surv Tutor* 18(1):795–823
- Al-Rawi HA, Ng MA, Yau KLA (2015) Application of reinforcement learning to routing in distributed wireless networks: a review. *Artif Intell Rev* 43(3):381–416
- Alsarhan A, Agarwal A (2011) Profit optimization in multi-service cognitive mesh network using machine learning. *EURASIP J Wirel Commun Netw* 1:36
- Anifantis E, Karyotis V, Papavassiliou S (2012) A markov random field framework for channel assignment in cognitive radio networks. In: 2012 IEEE international conference on pervasive computing and communications workshops (PERCOM workshops). IEEE, pp 770–775
- Berthold U, Fu F, Van M der Schaar, Jondral FK (2008) Detection of spectral resources in cognitive radios using reinforcement learning. In: *New frontiers in dynamic spectrum access networks*, pp 1–5
- Bkassiny M, Li Y, Jayaweera SK (2013) A survey on machine-learning techniques in cognitive radios. *IEEE Commun Surv Tutor* 15(3):1136–1159
- Busoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. *IEEE Trans Syst Man Cybern Part C Appl Rev* 38(2):156–172
- Chen S, Huang Y, Namuduri K (2011) A factor graph based dynamic spectrum allocation approach for cognitive network. In: *Wireless communications and networking conference (WCNC), 2011 IEEE*. IEEE, pp 850–855
- Cheng X, Jiang M (2011) Cognitive radio spectrum assignment based on artificial bee colony algorithm. In: 2011 IEEE 13th international conference on communication technology (ICCT). IEEE, pp 161–164
- Faganello LR, Kunst R, Both CB, Granville LZ (2013) Improving reinforcement learning algorithms for dynamic spectrum allocation in cognitive sensor networks. In: *Wireless communications and networking conference*, pp 35–40
- Feng Z, Liang L, Tan L, Zhang P (2009) Q-learning based heterogenous network self-optimization for reconfigurable network with CPC assistance. *Sci China (Ser F)* 52(12):2360–2368
- Han G, Xiao L, Poor HV (2017) Two-dimensional anti-jamming communication based on deep reinforcement learning. In: *IEEE international conference on acoustics, speech and signal processing*, pp 2087–2091
- Hossain E, Bhargava V (2007) *Cognitive wireless communication networks*. Springer, New York
- Iii JM (2000) *Cognitive radio: an integrated agent architecture for software defined radio*, Ph.D. dissertation. ResearchGate 6(4):13–18
- Le HST, Ly HD (2008) Opportunistic spectrum access using fuzzy logic for cognitive radio networks. In: *Second international conference on communications and electronics, ICCE 2008*. IEEE, pp 240–245
- Levorato M, Firouzbadi S, Goldsmith A (2012) A learning framework for cognitive interference networks with partial and noisy observations. *IEEE Trans Wirel Commun* 11(9):3101–3111
- Li H, Zhu G, Liang Z, Chen Y (2010) A survey on distributed opportunity spectrum access in cognitive network. In: *International conference on wireless communications networking and mobile computing*, pp 1–4
- Li Y, Feng Z, Chen S, Chen Y, Xu D, Zhang P, Zhang Q (2011) Radio resource management for public femtocell networks. *EURASIP J Wirel Commun Netw* 1:181
- Lilith N, Dogancay K (2005) Distributed reduced-state sarsa algorithm for dynamic channel allocation in cellular networks featuring traffic mobility. *IEEE Int Conf Commun* 2:860–865
- Lv C, Wang J, Yu F, Dai H (2013) A Q-learning-based dynamic spectrum allocation algorithm. *ICCSEE-13*
- Marinho J, Monteiro E (2012) Cognitive radio: survey on communication protocols, spectrum decision issues, and future research directions. *Wirel Netw* 18(2):147–164

- Mitola J, Maguire GQ (1999) Cognitive radio: making software radios more personal. *IEEE Pers Commun* 6(4):13–18
- Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. *Comput Sci*
- Nie J, Haykin S (1997) A Q-learning-based dynamic channel assignment technique for mobile communication systems. *IEEE Trans Veh Technol* 48(5):1676–1687
- Qadir J (2016) Artificial intelligence based cognitive routing for cognitive radio networks. *Artif Intell Rev* 45(1):25–96
- Ru M, Yin S, Qu Z (2017) Power and spectrum allocation in d2d networks based on coloring and chaos genetic algorithm. *Procedia Comput Sci* 107:183–189
- Salameh HAB (2011) Throughput-oriented channel assignment for opportunistic spectrum access networks. *Math Comput Modell* 53(11–12):2108–2118
- Shu T, Krunz M (2010) Exploiting microscopic spectrum opportunities in cognitive radio networks via coordinated channel access. *IEEE Trans Mob Comput* 9(11):1522–1534
- Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, vol 1. MIT Press, Cambridge
- Tanwongvarl C, Chantaraskul S (2015) Performance comparison of learning techniques for intelligent channel assignment in cognitive wireless sensor networks. In: Seventh international conference on ubiquitous and future networks, pp 503–507
- Teng Y, Zhang Y, Niu F, Dai C (2010) Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks. In: Vehicular technology conference fall, pp 1–5
- Teng Y, Yu FR, Han K, Wei Y, Zhang Y (2013) Reinforcement-learning-based double auction design for dynamic spectrum access in cognitive radio networks. *Wirel Pers Commun* 69(2):771–791
- Tragos EZ, Zeadally S, Fragkiadakis AG, Siris VA (2013) Spectrum assignment in cognitive radio networks: a comprehensive survey. *IEEE Commun Surv Tutor* 15(3):1108–1135
- Wang W, Kwasinski A, Niyato D, Han Z (2016) A survey on applications of model-free strategy learning in cognitive wireless networks. *IEEE Commun Surv Tutor* 18(3):1717–1757
- Xiao L, Li Y, Dai C, Dai H, Poor HV (2017) Reinforcement learning-based NOMA power allocation in the presence of smart jamming. *IEEE Trans Veh Technol* 67:3377–3389
- Yang M, Grace D (2011) Cognitive radio with reinforcement learning applied to multicast downlink transmission with power adjustment. *Wirel Pers Commun* 57(1):73–87
- Yang R, Ye F, et al (2010) Non-cooperative spectrum allocation based on game theory in cognitive radio networks. In: 2010 IEEE fifth international conference on bio-inspired computing: theories and applications (BIC-TA). IEEE, pp 1134–1137
- Yau KLA, Komisarczuk P, Teal PD (2012) Reinforcement learning for context awareness and intelligence in wireless networks: review, new features and open issues. *J Netw Comput Appl* 35(1):253–267
- Yi L, Hong J (2012) Q-learning for dynamic channel assignment in cognitive wireless local area network with fibre-connected distributed antennas. *J China Univer Posts Telecommun* 19(4):80–85
- Yu L, Liu C, Liu Z, Hu W (2010) Heuristic spectrum assignment algorithm in distributed cognitive networks. In: 2010 6th International conference on wireless communications networking and mobile computing (WiCOM). IEEE, pp 1–5
- Zhang Y, Lee C, Niyato D, Wang P (2013) Auction approaches for resource allocation in wireless systems: a survey. *IEEE Commun Surv Tutor* 15(3):1020–1041
- Zhao C, Zou M, Shen B, Kim B, Kwak K (2008) Cooperative spectrum allocation in centralized cognitive networks using bipartite matching. In: Global telecommunications conference, IEEE GLOBECOM 2008. IEEE, pp 1–6
- Zhao Q, Sadler BM (2007) A survey of dynamic spectrum access. *IEEE Signal Process Mag* 24(3):79–89