Complex & Intelligent Systems (2022) 8:1975–1986
https://doi.org/10.1007/s40747-021-00382-1

ORIGINAL ARTICLE

Dueling deep Q-networks for social awareness-aided spectrum sharing

Yonghua Wang[1] · Xueyang Li[1] · Pin Wan[1] · Le Chang[1] · Xia Deng[2]

Received: 5 November 2020 / Accepted: 15 April 2021 / Published online: 27 April 2021

Abstract
In overlapping spectrum sharing, due to the complexity of cognitive environment, it is a real challenge for a secondary user (SU) to correctly sense the usage of the spectrum in real time. To tackle this challenge, a social awareness-aided transmit power control policy for SUs is developed. First, a social network composed of a group of third-party sensing nodes that do not share the spectrum with the PU is established, which helps an SU collect the power information of the PU. Then, we design a Dueling Deep Q-Network (DQN) model to achieve efficient dynamic spectrum sharing between the PU and the SU with the power information collected in the social network. Experimental results show that the spectrum sharing success rate is higher and the comprehensive performance is improved with the sensing nodes selected by the social relationship. Moreover, compared with other deep reinforcement learning (DRL) algorithms, the performance of Dueling DQN is more stable on our targeted spectrum sharing problem.

Keywords Cognitive radio · Spectrum sharing · Social relationship · Power control

Introduction

Today's wireless networks require intelligent user demand sensing and elastic on-demand resource provisioning, as to offer high quality of service (QoS) for spectrum users [1]. As an example, Cognitive Radio Networks (CRN) rely on the sensing and rational allocation of the limited spectrum resources to allow second users (SUs) to share the spectrum without affecting the QoS of the primary users (PUs) [2,3].

Controlling the transmit power of SUs is one of the essential issues in CRN. Existing power control algorithms adjust the transmit power of SUs through multiple iterations [5], such as DCPC and DPC-ALP [6,7]. Other methods include utilizing the similarity between the graph model and wireless network model [8,9], or using the interference graph to achieve spectrum allocation with reduced interference [10,12]. On the other hand, the emergence of the novel Deep Reinforcement Learning (DRL) techniques help tackle the computation complexity of large-scale-state-and-action-space problems under dynamic environments. DRL enables agents to learn the action strategies under the guidance of an update-to-date optimal target approximation (Q value) [13,14]. DRL demonstrates its powerful control-decision capabilities in the areas of games, robotics, autonomous driving, and radio communications, etc. [15,19]. It can also apply to channel allocation and transmit power control in spectrum allocation of CRN.

To achieve the optimal power control, it is necessary to obtain the environmental information, e.g., channel state, which is dynamic in most cases. Collecting such environmental information accurately and timely requires considerable amount of resources [4]. A promising strategy is to utilize social networks. With the wide application of various social software such as WeChat, Weibo, and Facebook, an intangible relationship network has been built among users [20,21]. Revealed by existing works, these close relationships include relatives, friends, or the cooperative command relationships between subordinates, etc., and the user community

✉ Le Chang
lechang@gdut.edu.cn

Yonghua Wang
wangyonghua@gdut.edu.cn

Xueyang Li
li13897115197@163.com

Pin Wan
wanpin2@163.com

Xia Deng
gzhu_dx@gzhu.edu.cn

[1] School of Automation, Guangdong University of Technology, Guangzhou 510006, China

[2] School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China

in CRN demonstrates strong sociality [22]. A high correlation between such social relationship and data transmission rate has also been identified [23]. The social attributes of users help improve the transmission performance significantly even if there is no cooperation between users [24]. Therefore, utilizing these social attributes in spectrum allocation and power control has a far-reaching significance [25,26].

In this paper, we study the power control strategy of a second user in CRN with the aid of social networks. We assume a pair of SU and PU share the spectrum with the help of a set of third-party sensing nodes. The SU adaptively adjusts its transmit power to use the spectrum without affecting the PU, according to the information collected from the sensing nodes. Our main contributions are summarized as follows.

1. The scheme of social awareness-aided spectrum sharing. A social relationship network between users is established. The sensing nodes with more intimate social relation are chosen to assist spectrum sharing.
2. The characterization and exploration of the impact of social relationships between users on sensing the environment. A social awareness-aided spectrum sharing method is proposed.
3. A deep reinforcement learning algorithm based on Dueling Deep Q-Networks (DQN). The algorithm is designed to achieve intelligent social awareness-aided spectrum sharing, which proves to demonstrate superior performance according to our experimental results.

The rest of this paper is organized as follows. The section "Social awareness-aided spectrum sharing scheme" explains the system model, including the spectrum sharing model and the social relationship model. The section "Spectrum sharing using dueling DQN" describes in detail how to use Dueling Deep Q-Networks to achieve intelligent spectrum sharing. The section "Experimental results" presents the simulation setups and the experimental results that verify the performance of our proposed method. The section "Conclusion and future work" concludes the paper and discusses the future work.

Related work

Concerning the transmit power control, Islam et al. proposed a distributed power control strategy for beam-forming and admission control [27], aiming at minimizing the transmit power of SUs constrained by the SNR of the transmissions. Game theory has also been applied to the dynamic resource allocation of CRN, tackling the problem of "multi-person decision-making in a competitive environment", where Chen et al. achieved power allocation according to the sufficient condition of Nash equilibrium, and proposed a random power adaptive control method based on multi-agent Q-learning [28].

In applying DRL to CRN, Naparstek et al. used DRL to accomplish dynamic spectrum access (DSA) for the channel selection of multiple SUs [29]. Chang et al. combined the memory function of the recurrent neural network (RNN) with the control decision-making ability of DRL, which achieved remarkable results in DSA research [30]. From the aspect of power control, Mohammadi et al. used transfer learning to reduce the number of iterations and took advantage of DQN to adjust the power for optimizing Quality of Service (QoS) and Quality of Experience (QoE) [31]. Liu et al. input spectrum waterfall into the convolutional neural network (CNN) to extract channel state information and used the Q-function to select the optimal transmission frequency to achieve anti-interference spectrum allocation [32,33]. In [34], sensing nodes were used to perceive the environmental information to assist the SU in sharing the spectrum. Zhang et al. used the more advanced A3C algorithm in DRL to perform the power control in spectrum sharing [35]. They focused on the tuning and optimization of the A3C algorithm to reduce the dependence on gradient update in the learning process, while we adopted the Dueling DQN method.

With the increasing popularity of social software, social relationships have become an important research direction. Das et al. constructed a social network model based on online social platforms [39]. It used the formation process of the consensus in real world to build a similar social relationship network. This work served as the basis of the works utilizing social networks to improve the performance in many areas, including our work in this paper. In [36], a dynamic peer selection strategy with social awareness-aided spectrum-power trading in D2D overlaying communication was proposed. Chen et al. applied game theory to D2D networks. They considered the social relationship on the original physical relationship network (PRN) between users, established the social relationship network (SRN), and defined the social utility of users. PRN and SRN were combined to measure the overall physical–social utility and aimed to maximize the synergy between PRN and SRN [37]. In addition, social relationships such as credibility has been applied to the perception of spectrum usage [40]. They proposed an evidence-based decision fusion cooperative spectrum sensing strategy, while we proposed a DRL technology to utilize the social credibility in spectrum sharing.

Our work differs from the above in that we propose an intelligent social awareness-aided spectrum sharing strategy. The social relationship between users is used in the computation of DRL, which improves the average transmission success rate. At the same time, the intelligent control strategy, i.e., Dueling DQN, enhances the stability of spectrum sharing.

Social awareness-aided spectrum sharing scheme

In this section, we introduce our social awareness-aided spectrum sharing scheme and model.

Spectrum sharing model

Spectrum sharing with third-party sensing nodes

To achieve efficient spectrum sharing in CRN, it is necessary to obtain the environmental information in real time, which is a great challenge for such dynamic and complex systems. In many scenarios, a PU and an SU perform transmission in a non-cooperative manner, where the PU does not realize the existence of the SU [34]. In such case, the PU and SU cannot directly obtain the transmit power of each other. To tackle this problem, we adopt the idea of using third-party nodes that do not compete for spectrum resources with the PU as the *sensing nodes* following existing work [38]. These sensing nodes are also spectrum users similar to PUs and SUs. They store the information they sensed, and the SUs may collect such sensed information from them at fixed intervals. As the power information is small in size, we assume that the SU can collect it in a very short time, leaving most of the time for spectrum sharing between the PU and SU. Therefore, the collection of the power information will not add much overhead to the system, and thus, we ignore the time on collecting it in our work. The PU and SU can obtain the transmit power of each other through these sensing nodes, which facilitates the coordination of the transmit power control to ensure that the SU can access the spectrum without affecting the QoS of the PU.

Spectrum sharing with the assistance of third-party sensing nodes is achieved through the overlapping method, as illustrated in Fig. 1. We assume that user 1 is the PU with the licensed band, and others are unauthorized users. When an SU, e.g., user 6, needs to access the spectrum to transmit data, it must carefully control its transmit power to avoid impacting the QoS of the PU, user 1. We assume that the PU and SU are independent of each other, and the SU cannot sense the power adjustment strategy of the PU directly. Instead, the SU observes such power information of the PU through interacting with a group of third-party sensing users in the middle, e.g., user 2 to 7. These intermediate sensing users do not compete for the spectrum with the PU. They only sense the transmit power of the PU, and thus do not impact its QoS. With the power information of the PU obtained from these sensing users, the SU can control their transmit power accordingly.

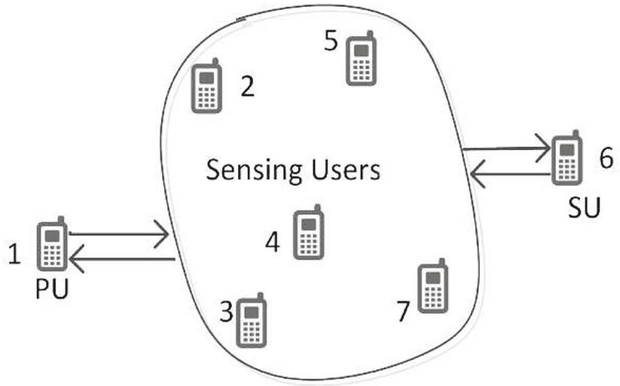


Fig. 1 Spectrum sharing with sensing nodes in cognitive radio

Power adjustment strategy of the PU

In this study, we do not require a specific power adjustment strategy of the PU. It only needs to follow a set of general rules, and the SU will try to learn it in the spectrum sharing process.

We assume that the PU updates its transmit power following a step-by-step manner. We discretize the spectrum sharing process into K time slots, and let $\mathcal{P} = \{p(k)\}$ denote the power adjustment strategy of the PU for such time slots $k = 1, 2, 3, \ldots, K$, where $p(k)$ is the transmit power of the PU at time slot k. The selection space of the transmit power of the PU is a set of discrete values $\{p_1, p_2, \ldots, p_L\}$, listed in the ascending order, i.e., $p_1 < p_2 \cdots < p_L$.

In the spectrum sharing process, the QoS of the PU and SU affects each other if they are close to each other and access the same spectrum channel. The QoS of the PU and SU are evaluated by their SNR, which is determined by both sides. The general form of the SNR of user i is calculated as

$$\mathrm{SNR}_i = \frac{h_{ii}^2 p_i}{\sum_{i \neq j} h_{ji}^2 p_j + \epsilon_i}, \tag{1}$$

where p_i and p_j denote the transmit power of user i and j, respectively. h_{ji} is the channel gain from sender j to receiver i, and ϵ_i is the received noise of user i. For example, the SNR of the PU is calculated using Eq. (1), where user i is the PU and user j is the SU. A transmission is considered successful when the SNR of sender i is higher than an SNR threshold δ_i.

Under these settings, a typical energy-saving strategy of the PU is to dynamically set the minimum transmit power at each time slot that guarantees a desirable QoS. For example, the PU can set a smaller transmit power to save energy when its previous SNR is above a certain level, or increase the transmit power to the next level (e.g., from p_l to p_{l+1}) to achieve a better QoS if its SNR in the previous time slot is below a threshold [34].

Power adjustment strategy of the SU

The power adjustment strategy of the SU is heavily dependent on that of the PU, as it is required that the SU should not impact the QoS of the PU. In this study, the SU will try to obtain some knowledge of the transmit power of the PU through third-party sensing nodes, and determine its own transmit power based on that information in the spectrum sharing process. Ideally, the optimal strategy of the SU is to maximize its own overall throughput in the process without affecting the QoS of the PU. According to Shannon theory, the throughput of the SU at time k is

$$T_{\mathrm{SU}}(k) = W \log_2(1 + \mathrm{SNR}_{\mathrm{SU}}(k)), \tag{2}$$

where W is the bandwidth. Thus, the optimization problem of the power adjustment strategy of the SU is formulated as

$$\max : \sum_{k=1}^{K} T_{\mathrm{SU}}(k) I(k) \tag{3}$$

$$\text{s.t.} : SNR_{\mathrm{PU}}(k) \geq \delta_{\mathrm{PU}}, \text{ if } p(k) > 0, \tag{4}$$

$$\forall k, I(k) = 1, \text{ if } SNR_{\mathrm{SU}}(k) \geq \delta_{\mathrm{SU}}; \text{ and } I(k) = 0, \text{ elsewise}, \tag{5}$$

$$\forall k, q(k) \in P_{\mathrm{SU}}, \tag{6}$$

where the unknowns, $q(k)$, are the transmit power of the SU at each time slot k, P_{S} is the selection space of the transmit power of the SU, and $I(k)$ is an indicator variable indicating whether the transmission of the SU is successful at time k. The first constraint, Eq. (4), guarantees that the PU will always succeed in transmission, and the second constraint, Eq. (5), takes only successful transmissions of the SU into account when calculating the overall throughput.

The solution to the problem is $q(k)$, the power selection of the SU at each time slot. $p(k)$-s represent the power of the PU, which are also unknown as $p(k)$ cannot be directly sensed by the SU and they may change dynamically following the PU's power adjustment strategy. However, $p(k)$-s are the parameters of the problem rather than the solution. Such parameters are unknown, so the problem is model-free. Therefore, in the section "Spectrum sharing using dueling DQN", we resort to DRL techniques to make the SU learn the optimal transmit power control strategy according to the sensed transmit power of the PU at each time slot.

Social awareness-aided model

With the increasing popularity of personal smart devices and online social software, social relationships among users have been identified and studied in the cognitive radio environment. In [39], the opinion information in social platforms such as Twitter is used as a reference to judge the credibility of users, and used as one of the conditions for the formation of social networks. Simpson and Sun [40] treat the credibility of the SU as a weighted averaging factor and update it with the change of the SU to improve the efficiency of spectrum sensing. In this study, we take such sociality into consideration and introduce social credibility [41] to establish a social network overlay, as shown in Fig. 2. Such social credibility is to measure the confidence level of the information collected at the sensing nodes. In real systems, the SU can obtain and maintain it in a distributed manner. It can keep records of the sensed power of each sensing node, and give each sensing node a score to represent its social credibility. The social network is denoted as $G = \{\mathcal{N}, \mathcal{E}\}$, where a set of nodes $\mathcal{N} = \{1, 2, \ldots, N\}$ represent the users, and $\mathcal{E} = \{(m, n) : \forall m, n \in \mathcal{N}\}$ is the set of edges between these nodes. Each edge (m, n) is associated with a value $\omega_{mn} \in [0, 1)$, i.e., the *social credibility*.

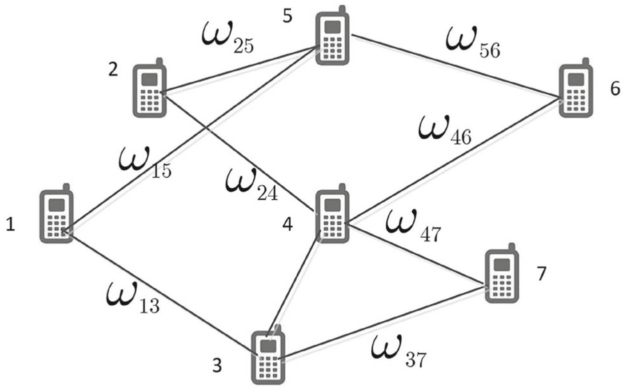


Fig. 2 The social network of cognitive users with social credibility

Given the transmit power p_n of user n, and user m as the sensing node, the sensed power of user n by user m is calculated as

$$p'_{mn} = p_n d_{mn}^{-\alpha} + \sigma_m, \tag{7}$$

where d_{mn} is the distance between the sensing node m and the transmitting node n, α is the path loss factor, and σ_m is the sensing deviation of user m. Different sensing nodes have different sensing deviations. With the social relationship network G defined above, we assume that the SU obtains the transmit power information of the PU based on the sensed power by a sensor and the corresponding social credibility between the SU and the sensor. At this time, the deviation ratio between the sensed transmit power value p' by the SU and the real transmit power value p of the PU is $\nu = \frac{p-p'}{p}$. We define ω as a decreasing function of $|\nu|$ using

$$\omega(\nu) = e^{-\frac{\nu^2}{2}}. \tag{8}$$

This becomes

$$\omega(p) = e^{-\frac{(\frac{p-p'}{p})^2}{2}} \tag{9}$$

when taking $\nu = \frac{p-p'}{p}$ into Eq. (8).

The value of p' affected by ω at time slot k is determined by the shared result at the previous time, i.e., a Markov decision process, and its expression is

$$p'(k) = \begin{cases} p(k)(1 + \sqrt{-2\ln\omega_{mj}}), & \text{if } \mathrm{SNR}_i(k-1) < \delta_i \\ p(k)(1 - \sqrt{-2\ln\omega_{mj}}), & \text{if } \mathrm{SNR}_i(k-1) \geq \delta_i \end{cases}. \tag{10}$$

According to this function, ω increases when $|\nu|$ decreases, and $|\nu|$ increases when ω decreases. When user n is sharing the authorized frequency band as a PU, we define its sensed social utility function as

$$\mathrm{Sol}_{mn} = p'_n d_{mn}^{-\alpha}. \tag{11}$$

With the spectrum sharing model and the social model above, we are ready to formulate the optimization problem of social awareness-aided spectrum sharing. When an SU controls the transmit power, it needs to find out the transmit power of the PU in real time through multiple sensing nodes to ensure that it will not affect the QoS of the PU. Since all users are included in the social network overlay, the transmit power information queried by the SU will be affected by the social credibility. Similar to [34], the transmit power information collected by the mth sensing node can be calculated by

$$Sol_m(k) = p'(k)d_{mi}^{-\alpha} + q'(k)d_{mj}^{-\alpha}, \tag{12}$$

where p' and q' represent the queried transmit power of the PU and the SU. The $p(k)$ in Eq. (2) will be approximated using $Sol_m(k)$ to complete the formulation of the social awareness-aided transmit control problem.

This problem is also NP-hard. Moreover, $p(k)$ needs to be estimated based on $\mathrm{Sol}_m(k)$ at each time slot, which is a model-free Markov decision process. Therefore, we resort to the novel reinforcement learning technique in which the SU learns the power control policy of the PU according to the transmit power information collected by the sensing nodes in Eq. (12). After repeated training, the SU can adjust its transmit power adaptively and achieve satisfactory QoS without impacting the PU.

Spectrum sharing using dueling DQN

In this section, we describe our dueling DQN model of learning the power information of the PU by the SU.

Q-learning and dueling DQN

Q-learning is a traditional reinforcement learning method that finds the optimal solution to a problem in a series of dynamic processes. Because it can estimate the expected effect after performing an action without knowing the system model in advance and support adjustment in real time, it is widely used in various decision problems. The algorithm was originally designed for a single agent that can interact with a fully observable Markov environment [42]. Then, Q-learning also demonstrated its strong optimal decision-making ability in other related fields such as multi-agent or non-Markov environment. However, Q-learning is limited to dealing with decision problems in small-scale state space. When the action-state space of the problem increases, the following problems will occur: (i) it is difficult to set up a Q table to store all possible state–action pairs when the state–action space is too complex; and (ii) as the state space continues to increase, some states will be rarely accessed again that causes poor training efficiency.

In recent years, the combination of Q-learning and deep neural networks has shown the potential of solving the control-decision problem in large-scale state space, i.e., DQN. DQN uses deep neural networks to approximate the value function instead of the Q value table in Q-learning, and obtains the optimal network parameters through multiple iterative training to solve control decision-making problems.

The dueling-DQN-based power control model

In this study, we use DQN to intelligently control the power of the SU, through further optimizing the network structure of DQN. The specific learning process is presented in Fig. 3. The transmit power of the PU is to be learned, so the PU is considered as a part of the environment, while SU is the agent that receives rewards after performing actions according to different environmental conditions. $S = \{s_1, s_2, \ldots, s_x\}$ denotes the state space, and $A = \{a_1, a_2, \ldots, a_y\}$ denotes the action space. The SU is in state $s(k) \in S$ at the kth time slot, enters a new state $s(k+1) \in S$ at the next time slot, and obtains a reward after performing $a(k) \in A$. In the process, the next state $s(k+1)$ is only relevant to $s(k)$ and $a(k)$ without aftereffect. Therefore, the power control process of the SU is a Markov decision process (MDP).

The objective of the transmit power control problem is to find a policy π with the maximum expected cumulative reward

$$G_\pi(k) = \sum_{k=0}^{\infty} \gamma^k R_{s(k)}^{a_\pi}. \tag{13}$$

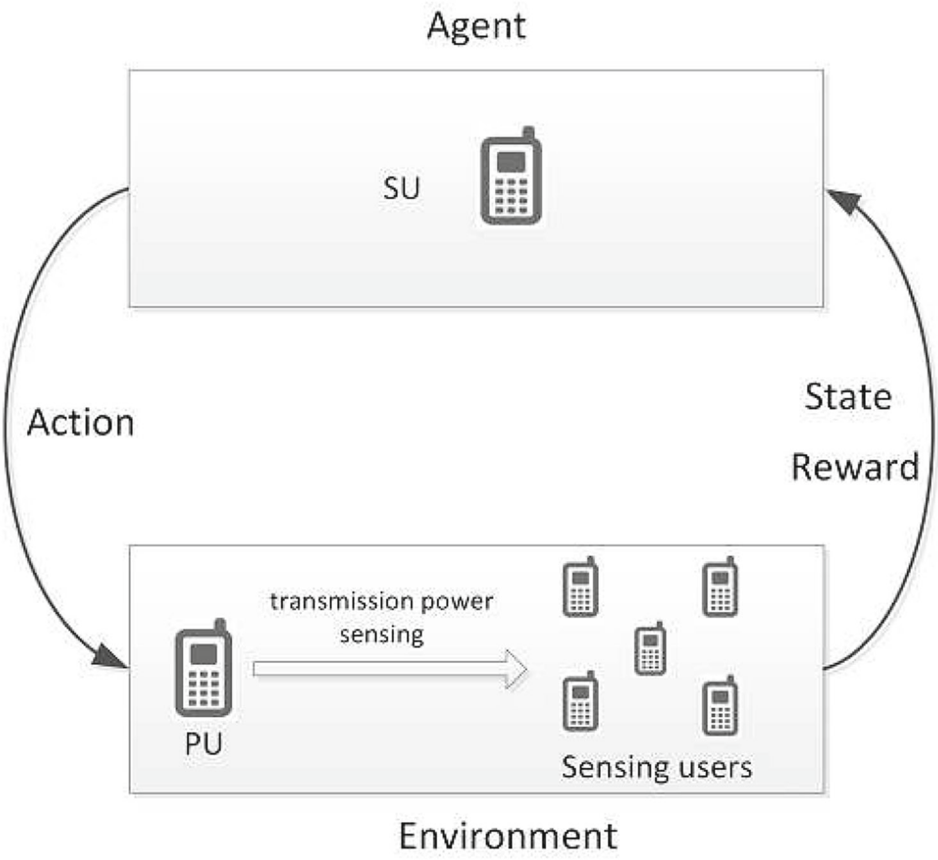


Fig. 3 The DRL learning model

The expected value of the cumulative reward is denoted by the state-value function

$$V_\pi(s) = E\left[\sum_{k=0}^{\infty} \gamma^k R_{s(k)}^{a_\pi} | s(k) = s\right]. \tag{14}$$

The optimal transmit power control policy is

$$\pi^*(s, a) = \arg\max_\pi V^\pi(s). \tag{15}$$

In DRL, the Q value [43] is calculated to find the optimal action, and its cumulative reward of taking action a at state s is

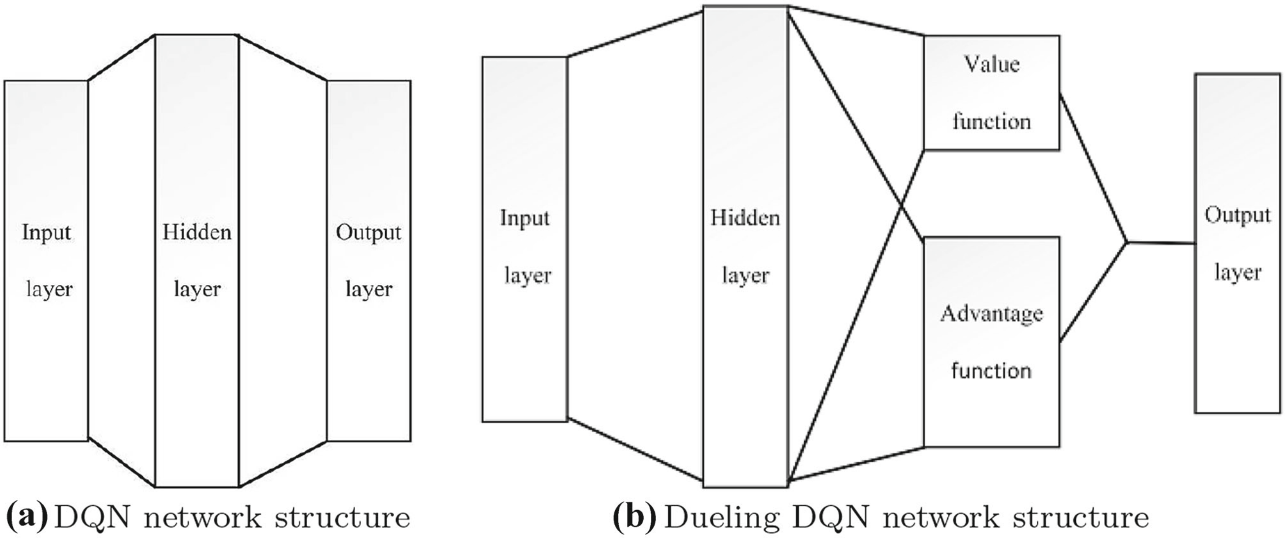
$$Q(s, a) = E[G_k | s_k = s, a_k = a], \tag{16}$$

and the optimal action is

$$a = \arg\max_a Q(s, a). \tag{17}$$

In our paper, dueling DQN is used to calculate the Q value to find the optimal solution of the transmit power control problem. With the same basic principle of DQN, the action-value function Q of Dueling DQN is approximated by neural networks (i.e., nonlinear approximation). We draw Fig. 4 to show the difference between DQN and dueling DQN. Dueling DQN does not directly calculate Q through a fully connected layer before the output layer, but divides the network into two parts. The first part calculates the value function, which is only related to the state, and has nothing to do with the action. The second part calculates the advantage function, whose value is related to both the state and action. In dueling DQN, the value function is regarded as the value of a static environment, and the advantage function is the additional value of an action. Moreover, the calculation of the Q value is related to both the environmental state and the action, but the degree of correlation is different. When the dueling DQN network is updated, instead of individually updating the Q value of an action, the Q values of all the actions in a state are adjusted to improve the network performance.

According to the section “Social awareness-aided spectrum sharing scheme”, we use the social utilities received by the SU as the environmental states. $s(k) = \{Sol_1(k), Sol_2(k), \ldots, Sol_x(k)\}$ denotes the state space at the kth time slot, which is time-varying. The SU chooses a power from $\{q_1, q_2, \ldots, q_l\}$ for transmission, so the action space is $a(k) = \{q_1(k), q_2(k), \ldots, q_l(k)\}$, which is a fixed set. The action space contains all the possible actions, and the SU selects an action from the action space at a time. The reward is defined as



(a) DQN network structure **(b)** Dueling DQN network structure

Fig. 4 The network structure comparison of DQN and dueling DQN

$$r(k) = \begin{cases} r, \text{if } SNR_i(k+1) \geq \delta_i \text{ and } \text{SNR}_j(k+1) \geq \delta_j \\ -r, \text{if } SNR_i(k+1) < \delta_i \text{ and } \text{SNR}_j(k+1) < \delta_j \\ 0, \text{otherwise} \end{cases}. \tag{18}$$

If both the PU and the SU transmit successfully, the reward is a constant r. If they both fail to transmit, the reward is $-r$. The reward is 0 for other cases.

The specific implementation process of Dueling DQN is shown in Fig. 5. We use the experience replay buffer to improve the training efficiency. When training the neural network, it is assumed that the training data are independently identically distributed. However, there is a correlation between the data collected through reinforcement learning, and some states that have appeared before will be rarely accessed. Therefore, the neural network is unstable if such data are used for training. In dueling DQN, the agent stores the data in a database, and then uses the uniform random sampling method to extract the data from the database and trains the neural network to break the correlation between the data. Finally, dueling DQN sets up the target network to deal with the TD error in the time difference algorithm alone.

Similar to DQN, dueling DQN also divides the network into the main net and the target net, with the same structure but different parameters during the parameter update process. The main net is used to update the network parameters, and the target net is used to update the Q value. At the initial moment, the main net assigns parameters to the target net. After that, the main net updates its own network parameters in real time. After a period of time, the main net assigns updated network parameters to the target net. The parameters of the two networks are updated cyclically in this way till the end of training. Such an update method can stabilize the target Q value for a period of time, thereby making the overall update of the algorithm more stable. The loss function is calculated as

$$L(\theta) = E[(\text{Target}Q - Q(s, a, \theta))^2], \tag{19}$$

$$\text{Target}Q = r + \max_{a'} Q(s', a', \theta), \tag{20}$$

where s' and a' represent the next state and action, respectively, after performing action a.

In addition, during the update process, we use the method of stochastic gradient descent (SGD) to update the parameter θ of the main net network. The specific calculation method is as follows:

$$\begin{aligned} \Delta\theta = \beta[r + \gamma \max_{a'} Q(s', a'; \theta) - Q(s, a; \theta)] \\ \nabla Q(s, a; \theta), \end{aligned} \tag{21}$$

$$\begin{aligned} \theta' \leftarrow \theta + \beta[r + \gamma \max_{a'} Q(s', a'; \theta) \\ -Q(s, a; \theta)]\nabla Q(s, a; \theta). \end{aligned} \tag{22}$$

In our social awareness-aided transmit power control strategy, the social relationship of the sensing nodes and the collected power information are used as the state input of dueling DQN, as shown in Eq. (12). According to *Definition 1* in [42], the power selection strategy can be seen as a mapping from state $s(k)$ to a probability mass function over action $a(k)$. Therefore, in our social awareness-aided approach, we set a selected probability for each sensing node to prevent it from falling into a local optimal or selecting the wrong sensing node due to malicious information. According to [44], the channel switching follows the distribution as

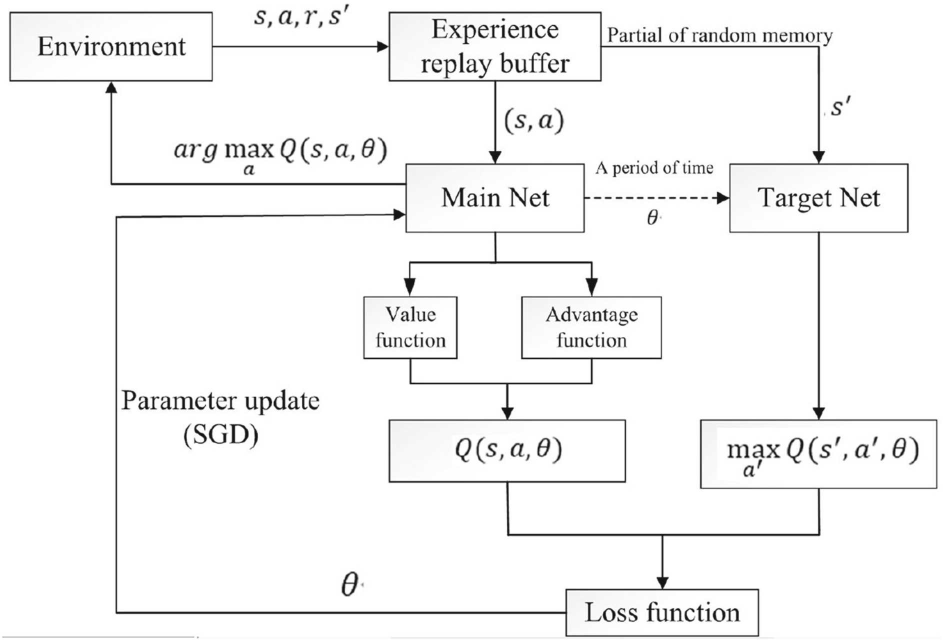


Fig. 5 The learning process of dueling DQN

$$P(k) = \frac{(1+\alpha)r_a^m(k)}{\Sigma_{n=1}^N (1+\alpha)r_a^n(k)} \quad m \in \{1, \ldots, N\}, \tag{23}$$

where α is a constant. The selecting probability is thus computed as

$$\begin{aligned} p_m(k) &= (1-\varphi)P(m(k)=m) + \frac{\varphi}{N} \\ &= (1-\varphi)\frac{(1+\alpha)r_a^m(k)}{\Sigma_{n=0}^N (1+\alpha)r_a^n(k)} + \frac{\varphi}{N} \\ &= \frac{1-\varphi}{\sum_{n=1}^N (1+\alpha)r_a^n(k)} + \frac{\theta}{C}, \end{aligned} \tag{24}$$

where φ is a constant parameter, and φ/N is a constant scaling factor. To prevent falling into the local optimal solution, the simulated annealing algorithm is introduced to select the non-optimal solution with a certain probability during channel switching. Let $\beta \in [0, 1)$ denote the simulating annealing constant, and Eq. (24) can be formulated as

$$p_m(k) = \frac{1-\varphi}{\sum_{n=1}^N (1+\alpha)e^{\beta r_a^n(k)}} + \frac{\varphi}{N}. \tag{25}$$

At a moment, when the sensing nodes are selected based on social relationship values, we allow the agent to drop the optimal choice with probability p_m, and switch to another sensing node with a lower social relationship value. This selection method improves the overall performance of the intelligent transmit power control system. The details of the algorithm are elaborated in Algorithm 1.

Algorithm 1 Intelligent transmit power control policy

```
Initialize replay memory size $D$, buffer size $O$
Initialize weight $\theta = \theta_0$
Initialize $p$,$q$
Establish the social relationship network
Compute the social utility functions for every user
Select the sensing nodes with better social utility functions
for iteration $k = 1, \ldots, K$ do
    for iteration $i = 1, \ldots, I$ do
        Input $Sol_m$ in network to compute $Q(s, a)$
        Choose $a(k) = \arg\max_a Q(s(k), a(k), \theta)$ with probability $\varepsilon$
        Store $\langle s, a, r, s' \rangle$ in $D$ following a uniform distribution in $D$
        if $D \geq O$ then
            Random extract a sample from $D$ to train network
            Compute the loss function according to Eq. (19), 20
            Update the network weights
        end if
    end for
    Re-select sensing nodes with probability $p_m$
end for
```

Experimental results

In this section, we demonstrate the superiority of our social awareness-aided power control strategy using dueling DQN through experiments.

Experiment setup

We compare our social awareness-aided Dueling DQN approach with the approach where the sensing nodes are chosen randomly. The transmit power of the PU and the SU is selected from a set of discrete values from 0.1 to 0.5. The received noise w is 0.01. SNR thresholds are set to $\delta_1 = 1.2$

and $\delta_2 = 0.7$. Channel gain is set to 0.9 and all the users are distributed in a rectangle area of $500 \times 1,000$ M^2. We do not explicitly add dynamic fading to our experiments as it is part of the environment that can be learned by the SU as well.

The dueling DQN network in the experiments has four hidden layers to approximate the optimal action-value function $Q(s, a)$. The number of neurons in the first three layers of the fully connected neuron network is set to 128, 128, and 256, respectively. The fourth layer consists of two networks with 60 neurons, which are used to compute the value function and the advantage function. The activation function is *ReLU*, in which the output less than 0 will be set to 0, or the raw output otherwise. The Adam algorithm is used to optimize the weights, θ_i. The size of the experience replay memory is $D = 1000$, the buffer size is $O = 500$, and the random explore probability is $\varepsilon = 0.8$. The training time is set to 2500 steps.

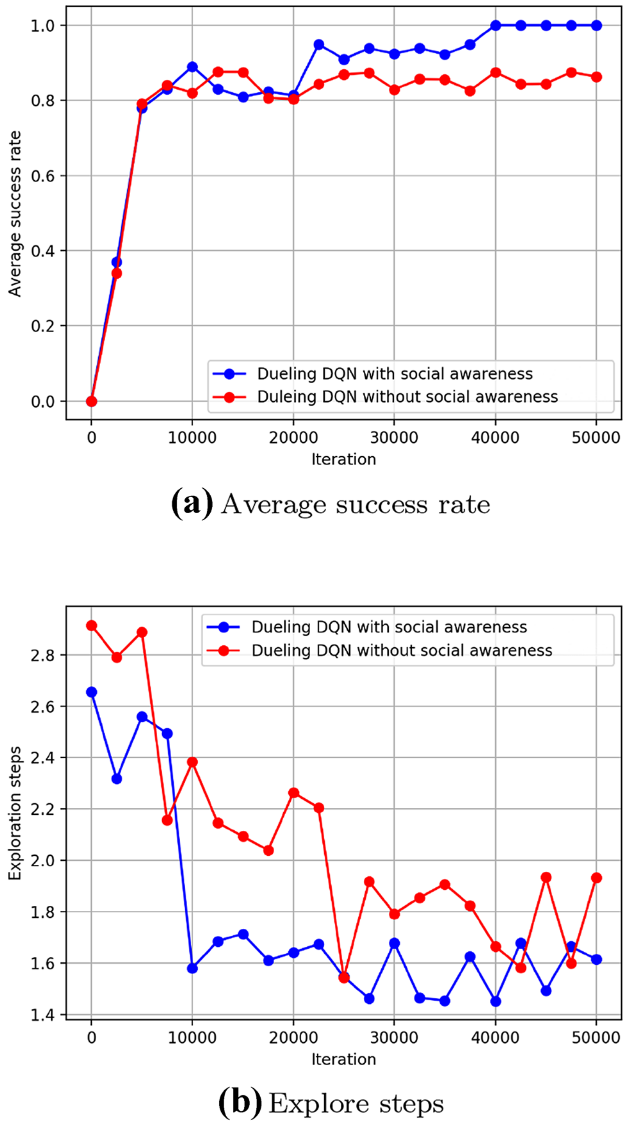
We mainly focus on two well-accepted performance metrics.

– *Average success rate*: the average number of successes over the 1000 iterations.
– *Explore rate*: the time required for the SU from start trying transmission to a successful transmission.

In our experiments, we first use dueling DQN to control the transmit power of the SU assisted by the sensing nodes which are chosen using the social utility function. After constructing the social relationship network, all sensing nodes are sorted according to social utility values, and then, users with better social relationship are selected to report power information. The experiment result is compared with the approach without social awareness. Then, we show the superiority of dueling DQN by comparing the average success rate with other popular learning algorithms.

Performance with/without social awareness

We set the number of sensing nodes to 5. The performance with and without social awareness is demonstrated in Fig. 6. The average success rate of our social awareness-aided approach gradually increases and eventually stabilizes at 1, while the approach without social awareness is not only worse but also more unstable than our approach, according to Fig. 6a. The average success rate without social awareness only fluctuates below 0.9. The training may take some time at the beginning, but after 5000 iterations, our social awareness-aided method converges, and the performance is desirable and stable thereafter. The number of exploration steps is plotted in Fig. 6b, where we see the number of exploration steps with social awareness can be eventually reduced to about 1.6, which is lower than that without social awareness in most experiments. In DRL, the agent explores different policies, which is a random process. It is possible that the strategy without social awareness outperforms the social awareness-aided approach at some steps occasionally. However, the performance of the approach without social awareness is highly unstable, and the overall performance is still worse than the social awareness-aided approach.

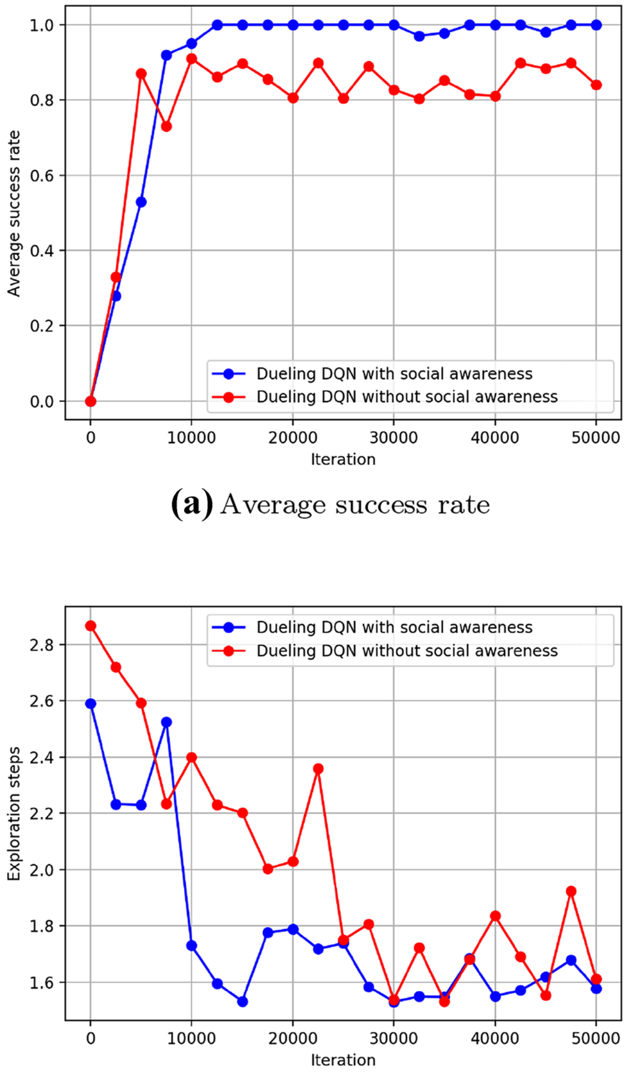


(a) Average success rate

(b) Explore steps

Fig. 6 Comparison of learning performance of five sensing nodes

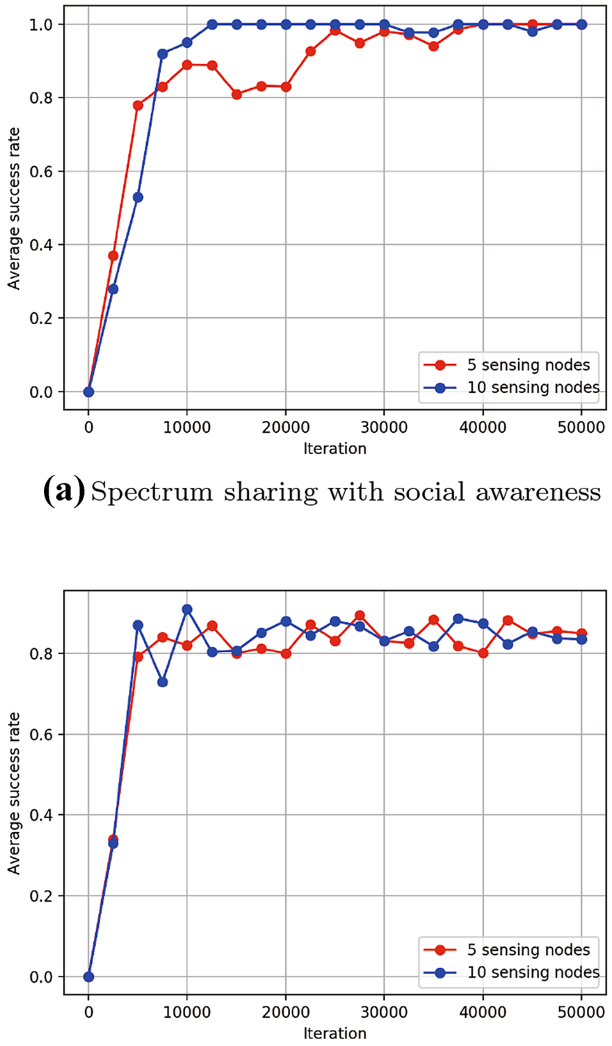
In Fig. 7, we increase the number of the sensing nodes to 10 and perform the above experiments again. We observe again that our social awareness-aided approach outperforms the one without social awareness in terms of both success rate and number of explore steps. Comparing with the performance with 5 sensing nodes, the success rate with 10 sensing nodes converges faster, as we have more candidates to select to get the power information. In terms of exploring steps, due to the increased number of the sensing nodes, the input states of Dueling DQN are also increased. Therefore, the number



(a) Average success rate

(b) Explore steps

Fig. 7 Comparison of learning performance of ten sensing nodes



(a) Spectrum sharing with social awareness

(b) Spectrum sharing without social awareness

Fig. 8 The effect of the number of sensing nodes on learning success rate

of exploration steps have not decreased, but it can still be successfully transmitted in a short time. According to the above experimental results, we can draw the conclusion: the social awareness-aided spectrum sharing can improve the transmit power control performance of the SU.

We plot Fig. 8 to gain a better understanding of the influence of adjusting the number of sensing nodes. For the social awareness-aided approach, 10 sensing nodes offer more choices, and thus, the average success rate converges faster and the performance is more stable, as suggested in Fig. 8a. In contrast, Fig. 8b shows the performance of the approach without social awareness, where the selection of the power information is random from the 10 sensing nodes. We observe that the success rates are only both floating around 0.8 for the no-social awareness approaches, i.e., little impact by the number of sensing nodes in this case. The main reason is that the sensing nodes are now chosen randomly. Therefore, only when the sensing nodes are selected based on the social relationships, the increase of the number of sensing nodes can improve the learning performance. This conclusion reflects the importance of exploring the social relationships on spectrum sharing.

Compare with other approaches

Finally, we compare the average success rate under different algorithms to verify the effectiveness and stability of our Dueling DQN approach. In the experiments, we set up ten sensing nodes as the candidates in our social awareness-aided approach. We choose Q-learning, DQN, and dueling DQN to compare with. The DQN algorithm uses fully connected neural networks with four hidden layers to approximate the

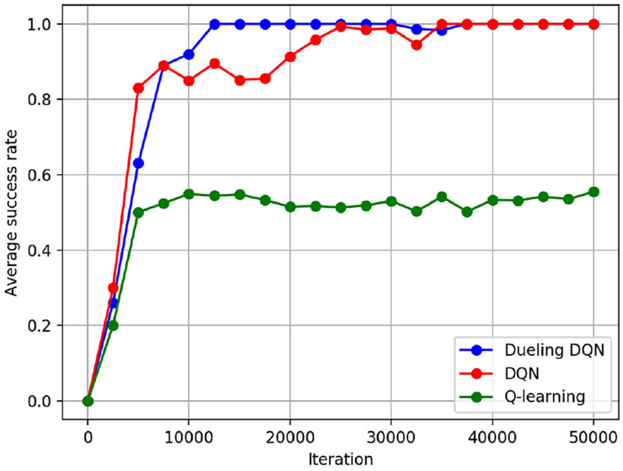


Fig. 9 Algorithm performance comparison

optimal action-value function $Q(s, a)$. The number of neuron units in the hidden layer are set to 128, 128, 256, and 256, respectively. The first three layers of the activation function is *ReLU*. The *tanh* function is employed for the fourth layer.

The comparison result is shown in Fig. 9. Due to the large scale of the state space in the process of intelligent transmit power control, Q-learning can never learn the optimal control strategy. It quickly converges to an average success rate around 0.5. Both DQN and dueling DQN belong to the DRL algorithms, which can finally learn the optimal control strategy, as the average success rates converge to 1. However, the training process of dueling DQN is accomplished by dividing the network into a value function and an advantage function, and thus, the learning is faster and more stable than DQN. Through such comparison, we can conclude that deep reinforcement learning has better learning performance with large-scale state space, and our dueling DQN has better network performance in deep reinforcement learning algorithms.

Conclusion and future work

In this paper, we studied the spectrum sharing problem. We proposed a social awareness-aided approach using dueling DQN. A social relationship network was established between users to reduce signal loss, and a social utility function was proposed to achieve intelligent transmit power control. The efficacy of our strategy was verified through experiments, which demonstrate: with our social awareness-aided approach, the average success rate of spectrum sharing converged faster and was also more stable. Compared with other DRL algorithms, our dueling DQN was more stable.

In the future, we will investigate more scenarios, including multiple PUs and SUs, varying credibility, and moving users, etc. In addition, we will study more complex transmit power adjustment strategies of the PU.

Acknowledgements This study was funded by National Natural Science Foundation of China (No. 61971147, 61702127), Hundred Young Talents Plan Project of Guangdong University of Technology (No. 220413618), Science and Technology Program of Guangzhou (No. 201804010461), China Scholarship Council (No. 201908440085, 201908440064).

Declarations

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Haykin S (2005) Cognitive radio: brain-empowered wireless communications. IEEE J Select Areas Commun 23(2):201–220
2. Mitola J III, Maguire G Jr (1999) Cognitive radio: making software radios more personal. IEEE Pers Commun 6(4):13–18
3. Mitola JI (2000) Cognitive radio: an integrated agent architecture for software defined radio dissertation. Dissert R Inst Technol Swed 294(3):66–73
4. Yang L, Wang X, Zhang JJ, Zhou M, Wang F (2020) Pedestrian choice modeling and simulation of staged evacuation strategies in Daya Bay Nuclear Power Plant. IEEE Trans Comput Soc Syst 7(3):686–695
5. Li Z, Han S, Wang X (2020) Power allocation scheme for physical-layer security of two-way untrusted relay in SCMA networks. In: Proceedings of international conference on computing, networking and communications (ICNC), pp 497–501
6. Grandhi SA, Zander J, Yates R (1994) Constrained power control. Wirel Pers Commun 1(4):257–270
7. Bambos N, Chen SC, Pottie GJ (1995) Radio link admission algorithms for wireless networks with power control and active link quality protection. In: Proceedings of IEEE international conference on computer communications (INFOCOM), Boston
8. Hoang AT, Liang Y (2008) Downlink channel assignment and power control for cognitive radio networks. IEEE Trans Wirel Commun 7(8):3106–3117
9. Hoang AT, Liang Y (2006) A two-phase channel and power allocation scheme for cognitive radio networks. In: Proceedings of IEEE international symposium on personal, indoor and mobile radio communications (PIMRC), pp 1–5

10. Behzad A, Rubin Z (2004) Multiple access protocol for power-controlled wireless access nets. IEEE Trans Mob Comput 3(4):307–316
11. Hoang AT, Liang Y (2006) Maximizing spectrum utilization of cognitive radio networks using channel allocation and power control. In: Proceedings of IEEE vehicular technology conference (VTC), pp 1–5
12. Zheng HT, Peng CY (2005) Collaboration and fairness in opportunistic spectrum access. In: Proceedings of IEEE international conference on communications (ICC), vol 5, pp 3132–3136
13. Mnih V, Kavukcuoglu K, Silver D et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
14. Mnih V, Kavukcuoglu K, Silver D et al (2013) Playing atari with deep reinforcement learning. In: Neural information processing systems, pp 201–203
15. Gu S, Holly E, Lillicrap T et al (2017) Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: Proceedings of IEEE international conference on robotics and automation (ICRA), pp 3389–3396
16. Xiong X, Wang J, Zhang F et al (2016) Combining deep reinforcement learning and safety based control for autonomous driving. arXiv:1612.00147
17. Sallab AE, Abdou M, Perot E et al (2017) Deep reinforcement learning framework for autonomous driving. Electron Imaging 19:70–76
18. Thananjeyan B, Garg A, Krishnan S et al (2017) Multilateral surgical pattern cutting in 2D orthotropic gauze with deep reinforcement learning policies for tensioning. In: Proceedings of IEEE international conference on robotics and automation (ICRA), pp 2371–2378
19. O'Shea TJ, Clancy TC (2016) Deep reinforcement learning radio control and signal detection with KeRLym, A Gym RL Agent. arXiv:1605.09221
20. Li J, Wang S, Yuan Y, Ni X, Wang F (2018) Dynamic optimization of employees work strategies in a WeChat-based evaluation system. IEEE Trans Comput Soc Syst 5(3):687–697
21. Cao J, Wang S, Li B, Wang X, Ding Z, Wang F (2020) Integrating multisourced texts in online business intelligence systems. IEEE Trans Syst Man Cybern Syst 50(5):1638–1648
22. Xu P, Hu W, Wu J, Liu W, Du B, Yang J (2019) Social trust network embedding. In: Proceedings of IEEE international conference on data mining (ICDM), pp 678–687
23. Gao W, Li Q, Zhao B, Cao G (2009) Multicasting in delay tolerant networks: a social network perspective. In: Proceedings of ACM international symposium on mobile ad hoc networking and computing (MobiHoc)
24. Chen X, Proulx B, Gong X, Zhang J (2015) Exploiting social ties for cooperative D2D communications: a mobile social networking case. IEEE/ACM Trans Netw 23(5):1471–1484
25. Ye P, Wang S, Wang F (2018) A general cognitive architecture for agent-based modeling in artificial societies. IEEE Trans Comput Soc Syst 5(1):176–185
26. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao TL, Duan W, Tsoi KKF, Wang FY (2020) Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo. IEEE Trans Comput Soc Syst 7(2):556–562
27. Islam MH, Liang Y, Hoang AT (2007) Distributed power and admission control for cognitive radio networks using antenna arrays. In: Proceedings of IEEE international symposium on new frontiers in dynamic spectrum access networks (DySPAN), pp 250–253
28. Chen X, Zhao Z, Zhang H (2013) Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks. IEEE Trans Mob Comput 12(11):2155–2166
29. Naparstek O, Cohen K (2019) Deep multi-user reinforcement learning for distributed dynamic spectrum access. IEEE Trans Wirel Commun 18(1):310–323
30. Chang H, Song H, Yi Y, Zhang J, He H, Liu L (2019) Distributive dynamic spectrum access through deep reinforcement learning: a reservoir computing based approach. IEEE Internet Things J 6(2):1938–1948
31. Shah-Mohammadi F, Kwasinski F (2018) Deep reinforcement learning approach to qoe-driven resource allocation for spectrum underlay in cognitive radio networks. In: IEEE international conference on communications workshops, pp 1–6
32. Liu X, Xu Y, Jia L, Wu Q, Anpalagan A (2018) Anti-jamming communications using spectrum waterfall: a deep reinforcement learning approach. IEEE Commun Lett 22(5):998–1001
33. Liu X, Xu Y, Cheng Y, Li Y, Zhao L, Zhang X (2018) A heterogeneous information fusion deep reinforcement learning for intelligent frequency selection of HF communication. China Commun 15(9):73–84
34. Li X, Fang J, Cheng W, Duan H, Chen Z, Li H (2018) Intelligent power control for spectrum sharing in cognitive radios: a deep reinforcement learning approach. IEEE Access 6:25463–25473
35. Zhang H, Yang N, Huangfu W, Long K, Leung VC (2020) Power control based on deep reinforcement learning for spectrum sharing. IEEE Trans Wirel Commun 19(6):25463–25473
36. Gao Y, Xiao Y, Wu M, Xiao M, Shao J (2019) Dynamic social-aware peer selection for cooperative relay management with D2D communications. IEEE Trans Commun 67(5):3124–3139
37. Chen X, Gong X, Yang L, Zhang J (2014) A social group utility maximization framework with applications in database assisted spectrum access. In: Proceedings of IEEE conference on computer communications (INFOCOM), pp 1959–1967
38. Padalkar S, Korlekar A, Pacharaney U (2016) Data gathering in wireless sensor network for energy efficiency with and without compressive sensing at sensor node. In: Proceedings of IEEE international conference on communication and signal processing (ICCSP), pp 1356–1359
39. Das R, Kamruzzaman J, Karmakar G (2019) Opinion formation in online social networks: exploiting predisposition. Interaction, and credibility. IEEE Trans Comput Soc Syst 6(4):554–566
40. Simpson O, Sun Y (2020) Efficient evidence-based decision fusion scheme for cooperative spectrum sensing in cognitive radio networks. Trans Emerg Telecommun Technol 31(4):e3901
41. Chen HB, Zhao F, Deng XF (2011) Cognitive radio spectrum sharing model based on social networks. Appl Res Comput 28(8)
42. Naparstek O, Kobi C (2018) Deep multi-user reinforcement learning for distributed dynamic spectrum access. IEEE Trans Wirel Commun 18(1):310–323
43. HWatkins CJC, Dayan P (1992) Technical note: Q-learning. Mach Learn 8(3):279–292
44. Auer P et al (1995) Gambling in a rigged casino: the adversarial multi-armed bandit problem. In: Proceedings of IEEE annual foundations of computer science, pp 322–331