

Dynamic spectrum access based on double deep Q-network and convolution neural network

Guangjin Fang, Bin Shen, Hong Zhang, Taiping Cui

School of Communication and Information Engineering(SCIE)

Chongqing University of Posts and Telecommunications (CQUP), Chongqing 400-065, P. R. China

s190131164@stu.cqupt.edu.cn; shenbin@cqupt.edu.cn; s190101180@stu.cqupt.edu.cn; cuitaiping@cqupt.edu.cn

Abstract—As a new technology, deep reinforcement learning (DRL) has been applied in many fields of wireless communication in recent years. In this paper, we propose a centralized dynamic spectrum access (DSA) scheme based on DRL and convolutional neural network (CNN). The traditional distributed DSA scheme encounters two important issues, namely, interference suppression for primary users (PU) and interference coordination for secondary users (SU). In order to cope with these two issues, a centralized DSA scheme implemented with deep convolution Q-network is proposed in this paper. Operating with the proposed centralized deep convolution Q-network spectrum access (CDCQSA) algorithm, a central unit (CU) uses DRL to unify the spectrum access actions in the network, allowing multiple SU to share spectrum resources with multiple PU without requiring any prior information. Simulation results show that this proposed scheme improves the channel utilization and reduces the conflict probability between PU and SU while avoiding the collisions between SU as well.

Keywords—Dynamic Spectrum Access; Deep Reinforce Learning; Double Deep Q-Network; Convolutional Neural Network

I. INTRODUCTION

To solve the contradictory problem of spectrum scarcity for unlicensed spectrum users and low spectrum utilization of the licensed frequency band (LFB), cognitive radio (CR) has been proposed to extract secondary spectrum access opportunities for the secondary users (SU) when the primary users (PU) are not temporarily operating over the LFB. The CR users, a.k.a. the SU, is essentially a group of users that are always dynamically monitoring the LFB status and striving to seize every granularity of chance to acquire the spectrum resource when it is sporadically available over the LFB. The spectrum access mechanism that the SU assesses to exploit the spectrum opportunity over the LFB is usually viewed as dynamic spectrum access (DSA), where ideally the SU and PU would coexist in a heterogeneous and orthogonal manner.

With the purpose of guaranteeing no or minimum interference between the PU and SU, recently there have been some different ideas and work contributed in combining DSA with deep reinforcement learning (DRL). The work [1,2] focused on studying DSA based on reinforcement learning (RL) theory under a game theoretical model, but the performance is limited when the prior knowledge of the radio environment is insufficiently available. In [3,4], the authors investigated more on RL theory itself, and the idea of combining RL with DSA gradually shifts from

Q-learning to deep Q-learning network (DQN). For the study of combining DSA with DRL to solve practical application problems, the authors in [5] investigated effective spectrum sharing between different types of users in the context of the industrial Internet of Things, and proposed an improved deep Q-network with each base-station in the network serving as a single agent to manage spectrum resources separately. This scheme improves the spectrum utilization to some extent, but it still cannot solve the interference coordination problem among the homogeneous group of SUs. The distributed DSA problem with only partial observation was investigated in [6] and a deep recurrent neural network was proposed to solve the access policy problem with multiple independent LFB channels and multiple PU. It was claimed that this network can eventually learn the spectrum access policy with a low conflict rate and high spectrum utilization. However, this method actually focused only on the problem of insufficient observation data, and neither does it distinguish the PU and SU in the network and nor take the perspective of regarding the PU and SU as heterogeneous users. Therefore, it cannot fully support interference coordination among the SU and the SU-to-PU interference suppression as well. The authors in [7] investigated the distributed DSA problem under the condition of spectrum sensing errors and proposed a special recurrent neural network, which is integrated with the DRL technique to optimize the mutual interference problem between different SUs. This method not only outperforms the traditional myopic policy access scheme in terms of spectrum utilization but also surpasses the Q-Learning algorithm in terms of convergence speed when the number of states is high. However, the proposed reservoirs computing (RC)-based neural network is complex in design, more difficult to train, and still cannot avoid the conflict problem among the SU.

According to the existing literature, the combination of DRL and DSA is a potentially effective solution to the problem of interference-free spectrum sharing among the SU and PU over the LFB. The authors in [8] pointed out that most of the current research focuses on solving the distributed DSA problem, while it is difficult to avoid the conflict problem among SUs in distributed DSA, and since the learning activity must be done by each SU individually, the hardware requirement for each SU device and the cost increases accordingly. In this context, bearing in mind the difficulties and drawbacks of distributed DSA methods

implemented via DRL, we propose a new type of DRL-based centralized DSA model, in which the SU are able to avoid the inter-SU conflict problem by action screening without any prior knowledge in a centralized DSA network, and use online learning mechanisms to enable the central unit (CU) to learn an efficient spectrum allocation strategy to improve spectrum utilization. Compared with the work in [9,10], the proposed algorithm uses CNN for feature extraction of data to solve the challenging classification problem of wireless channels. Simulation results show that the proposed scheme in this paper improves the channel utilization and avoids conflicts between different SUs; while reducing the conflict rate between the PU and SU. These technical merits make the proposed algorithm a practical candidate DSA solution.

The rest of this paper is organized as follows. In Section II, the system model is described. Section III presents the proposed centralized deep convolution Q-network-based DSA. In Section IV, the performance evaluation results of the proposed algorithm are given. Section V concludes the paper.

II. SYSTEM MODEL

The system model concerned in this paper is based on a centralized cooperative spectrum access approach, where the SU gain access to the LFB channel in interweave mode. In this model, a central unit (CU) collects the local spectrum observations from the SUs in the network for processing and analysis to ascertain the global spectrum occupancy status. Under the centralized mechanism, spectrum analysis and access scheduling of all SUs in the network are under the control of the CU, so as to coordinate the conflict problem between SUs. The interweave mode requires that SU and PU are not permitted to transmit over the same LFB channel simultaneously, and one assumption is thus made that if the PU and SU coincidentally access to the same LFB channel at the same time, both of them will fail in maintaining successful transmission.

A. DataSet Acquisition

We consider a DSA network in which there are N PU transmitters (PUTs) and each PUT has a corresponding PU receiver (PUR). In this paper, we assume that each PU occupies a channel independently and does not interfere with each other. The working state s_n of the n -th PU, $n = 1, 2, \dots, N$, can be marked as 'on' or 'off', denoted by 1 or 0 respectively. Accordingly, there are a total of 2^N joint transmitting states of the N PUTs in the network. The geographical area of the DSA network is divided into G equal-area grids. At time t , a spectrum data matrix $\mathbf{Y}_t \in \mathbb{R}^{N \times G}$ can be constructed in the CU based on the energy value of the received signal in frequency bands in the G grids. The n -th channel signal received at the g -th grid can be expressed as $z_{n,g}(i)$

$$z_{n,g}(i) = \sqrt{P_{TX,n} s_n} h_{n,g} X_n(i) + N_{n,g}(i) \quad (1)$$

where i is the time index for signal samples, $P_{TX,n}$ represents the transmit power of the n -th PUT, $X_n(i)$ is the i -th transmitted signal of the n -th PUT, $N_{n,g}(i)$ is the complex additive Gaussian white noise with zero mean and variance σ_0^2 corresponding to the SU receiver at the g -th grid, and $h_{n,g}$ refers to the channel coefficient from the n -th PUT to the g -th grid.

The spectral observation obtained by the SU in the g -th grid over the n -th channel is:

$$Z_{n,g} = \sum_{i=1}^{2\omega_n\tau} |z_{n,g}(i)|^2, \quad g = 1, 2, \dots, G \quad (2)$$

where ω_n is the bandwidth of the n -th channel and τ is the sensing interval within each spectrum analysis period. The aggregate spectrum data matrix $\mathbf{Y}_t$ can be expressed as:

$$\mathbf{Y}_t = \begin{pmatrix} Z_{1,1}^{(t)} & \dots & Z_{1,G}^{(t)} \\ \vdots & Z_{n,g}^{(t)} & \vdots \\ Z_{N,1}^{(t)} & \dots & Z_{N,G}^{(t)} \end{pmatrix} \quad (3)$$

The dataset acquired over a relatively long time is denoted as D_T , $D_T = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_T\}$. It is assumed that each SU has one transmitter (SU transmitter, SUT) and only one receiver (SU receiver, SUR). We further consider that there are M SUs in the DSA network, each applying for transmitting over one LFB-channel under the control of the CU. The system model is shown in Figure 1.

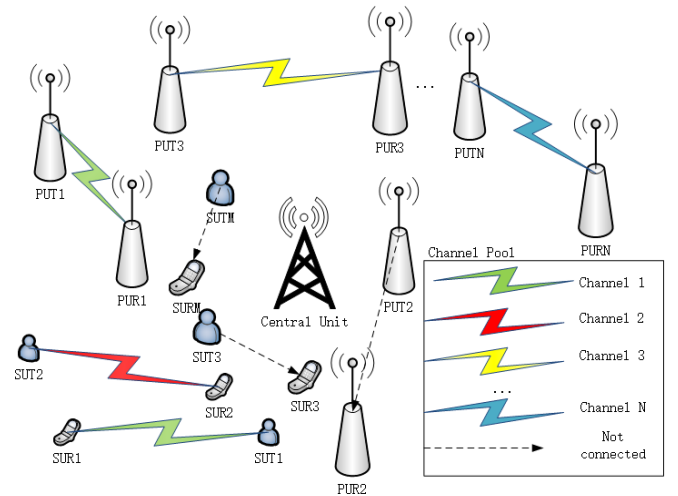


Fig. 1. System model

B. Centralized Reinforcement Learning Model

In this subsection, we explain the indispensable elements of the RL model, i.e., the sets of actions $\mathbf{A}$, the state set $\mathbf{S}$, and the reward R calculation methodology, and the strategy π .

The state set $\mathbf{S}$ is involved with 2^N different joint PUT states, $\mathbf{S} = \{s^0, s^1, \dots, s^{2^N-1}\}$. The PUTs' joint transmit state s_t holds that $s_t \in \mathbf{S}$. Usually, after performing spectrum sensing on $\mathbf{Y}_t$, we can obtain the estimated state $\hat{s}_t$. However, in order to improve efficiency, we bypass the spectrum sensing

stage and directly feed the spectrum dataset for double deep Q-network (DDQN) training. It is already known that $\mathbf{s}^t = [s_1^t, s_2^t, \dots, s_N^t]$, where s_n^t indicates the working state of the n -th PUT at time t . When the n -th PUT is transmitting, $s_n^t = 1$, otherwise $s_n^t = 0$.

The action set $\mathbf{A}$ is composed of $(N+1)^M$ different actions. It is defined that $\mathbf{A} = \{\mathbf{a}^0, \mathbf{a}^1, \dots, \mathbf{a}^{(N+1)^M-1}\}$, where $\mathbf{a}^\ell = [a_1, a_2, \dots, a_M]$ stands for the actions of the M SUTs with $a_m \in \{0, 1, 2, \dots, N\}$ indicating the a_m -th LFB channel is selected by the m -th SUT for spectrum access. Note that $a_m = 0$ means that the m -th SUT chooses to stand idle and select no LFB-channel for transmission. However, since there are conflicting actions of different SUTs in $\mathbf{A}$, i.e., different SUTs transmit over the same LFB-channel at the same time, the CT carries out an action-screening procedure, and the set of screened actions is denoted by $\mathbf{A}_c$. Algorithm 1 shows the pseudo-code of the action filtering algorithm, where some unnecessary and conflicting actions are removed from $\mathbf{A}$.

Algorithm 1 Action Filtering Algorithm

Input: $\mathbf{A}$: Original action set

Output: $\mathbf{A}_c$: Filtered action set

- 1: Initialize $\mathbf{A}_c$ as an empty set
 - 2: **for** i in $\mathbf{A}$ **do**
 - 3: If no duplicate elements of i other than 0 are found,
 add i in $\mathbf{A}_c$
 - 4: **end for**
-

The action at time t is denoted as $\mathbf{a}_t$, $\mathbf{a}_t = [a_1^t, a_2^t, \dots, a_M^t]$, $\mathbf{a}_t \in \mathbf{A}_c$, where $a_m^t \in \{0, 1, 2, \dots, N\}$. As mentioned before, $a_m^t = n$ indicates that the m -th SUT utilize the n -th LFB-channel at time t , whereas when $a_m^t = 0$, it means that no access action is performed.

The reward R is determined by the state $\mathbf{s}$ and action $\mathbf{a}$. The reward at time t is denoted by R_t and calculated as:

$$R_t = R(\mathbf{s}_t, \mathbf{a}_t) = R_S \cdot S_{UC} + R_N \cdot S_{NO} + R_F \cdot S_F \quad (4)$$

where R_S indicates the reward for successful access, S_{UC} indicates the number of SUs successfully gained access to the LFB, R_N denotes the no-access award, S_{NO} is the number of idle SUs, R_F denotes the access-failed reward, and S_F is the number of SUs with access failure.

The problem we want to solve is how to find the optimal LFB-channel access strategy π . π maximizes channel utilization and avoids inter-SU conflicts. Assuming that each user has packets to send at any given moment, the channel utilization r_{cu} is the occupancy of N channels in a test period T_{test} .

$$r_{cu} = \frac{\sum_{t=1}^{T_{test}} N_{cu}^t}{N * T_{test}} \quad (5)$$

where N_{cu}^t indicates the number of occupied channels at time t .

III. CENTRALIZED DEEP CONVOLUTION Q-NETWORK SPECTRUM ACCESS

It is well known that the Q-learning algorithm develops access policies by iterating out the Q-table and has good performance when the set of states has a limited cardinality. When the number of PUTs increases, the number of states will grow exponentially, leading to slow training speed and dimensional explosion problems. In contrast, the DQN scheme does not need to develop Q-table and obtain the access policy by training the deep neural network, but it will bring the overestimation problem. The CDCQSA scheme proposed in this paper selects the DDQN technique, introduces target-network and eval-network to solve the overestimation problem. By selecting CNN, the proposed algorithm is good at handling matrix data analysis, as the network to be trained.

The spectrum data matrix $\mathbf{Y}_t$ is preprocessed to form the input $\mathbf{F}_t$ of the network, and then the action $\mathbf{a}_t$ is randomly selected according to the greedy strategy to obtain the reward R_t of the environmental feedback. In the next DSA period, the $\mathbf{F}_{t+1}$ is obtained through data preprocessing, according to $\mathbf{Y}_{t+1}$. The data $(\mathbf{F}_t, \mathbf{a}_t, R_{t+1}, \mathbf{F}_{t+1})$ are stored in the experience pool D . Then, the whole network is trained by batching the training data and assigning the values of the target network to the valuation network every C steps. In this way, the finally obtained DDQN network can converge to grasp the optimal LFB-channel access action and serve the new spectrum observation data obtained online. The iterative reward is updated as:

$$y_j = \begin{cases} R_{j+1} & j+1 = T \\ R_{j+1} + \gamma \max_{\mathbf{a}'} (\phi_{j+1}, \mathbf{a}'; \omega^-) & \text{otherwise} \end{cases} \quad (6)$$

The proposed CDCQSA algorithm is described in algorithm 2.

IV. SIMULATIONS

In this section, the algorithms mentioned above, namely the Q-learning algorithm, DQN algorithm, DQSA algorithm, CDCQSA algorithm, and Random Allocation (RA) algorithm, are simulated and compared in detail.

A. Simulation Settings

(1) Dataset: The table of parameters used in simulation is shown in Table 1.

(2) CDCQSA: The table of hyperparameters used for the CDCQSA is shown in Table 2.

B. Experimental Results and Discussions

We first compare the convergence speed and performance of the four algorithms under the 5 dB SNR. From Fig. 2, we can see that the channel utilization of the algorithm proposed in this paper reaches about 91%, which is much better than the other algorithms, while it cannot reach 100% utilization due to the greedy strategy adopted in selecting the action.

Since a centralized network is used in this paper and the CU schedules all SUs, the conflict problem among SUs can be completely avoided. Fig. 3 shows the average conflict rate

Algorithm 2 CDCQSA**Input:** $\mathbf{Y}_t, \mathbf{a}_t$ **Output:** $Q(\mathbf{Y}_t, \mathbf{a}_t)$

- 1: Initialize experience pool D and set the capacity to D_{ds}
- 2: Initialize eval-network Q with random weights ω
- 3: Initialize target-network with random weights ω^-
- 4: **for** $episode = 1 : epoch$ **do**
- 5: initialize sequence $\mathbf{Y}_t$ and preprocess sequence
- 6: $\mathbf{F}_1 = \mathbf{F}(\mathbf{Y}_1)$
- 7: **for** $t = 1 : T - 1$ **do**
- 8: With probability ϵ select a random action $\mathbf{a}_t$
 otherwise select $\mathbf{a}_t = \text{argmax}_a Q(\mathbf{F}(\mathbf{Y}_t), \mathbf{a}; \omega)$
- 9: Execute action $\mathbf{a}_t$ and observe the reward R_t and sequence $\mathbf{Y}_t$
- 10: Set $\mathbf{Y}_{t+1} = \mathbf{Y}_t$, preprocess $\mathbf{F}_{t+1} = \mathbf{F}(\mathbf{Y}_{t+1})$
- 11: Store transition $(\mathbf{F}, \mathbf{a}_t, R_t, \mathbf{F}_{t+1})$ in D
- 12: Sample random minibatch of transitions $(\mathbf{F}_j, \mathbf{a}_j, R_j, \mathbf{F}_{j+1})$ from D
- 13: According to Equation (5) obtain y_j , $Q(\mathbf{F}_j, \mathbf{a}_j; \omega)$
- 14: Perform a gradient descent step on $(y_j - Q(\mathbf{F}_j, \mathbf{a}_j; \omega))^2$ with respect to the network parameters ω
- 15: Every C steps reset $\omega^- = \omega$
- 16: **end for**
- 17: **end for**
- 18: Entering a new observation matrix, the CU makes the optimal access decision based on the trained DDQN

TABLE I
PARAMETERS OF SYSTEM MODEL

Parameters	Values
Number of PUs	3
Number of SUs	5
Bandwidth/MHz	5
PUT power/W	10
Sensing time/us	100
Pathloss exponential coefficient	4
Shadow fading/dB	2

and training time of the PU and SUs corresponding to different algorithms in a DSA network under 5 dB SNR. It can be seen from Fig. 3 that the average conflict rate is lower than other algorithms when CDCQSA converges.

Fig. 4 shows the conflict probability between SUs of different algorithms, from which it can be seen that CDCQSA solves the conflict problem between SUs compared to other algorithms.

TABLE II
HYPERPARAMETERS OF CDCQSA

Hyperparameters	Values
D_{ds}	2000
epoch	3000
ϵ	0.1
γ	0.9
Input layer size	100*3
Output layer size	1*1

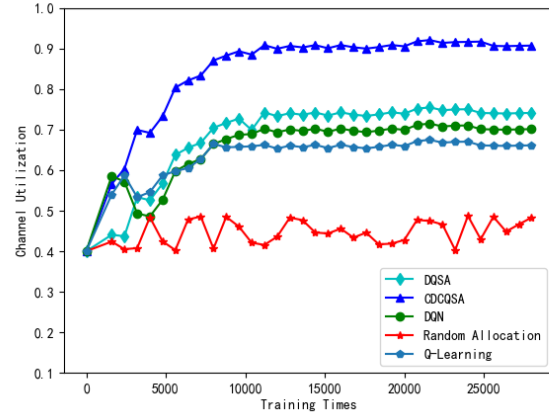


Fig. 2. The convergence rate and channel utilization of different algorithms with SNR=5dB

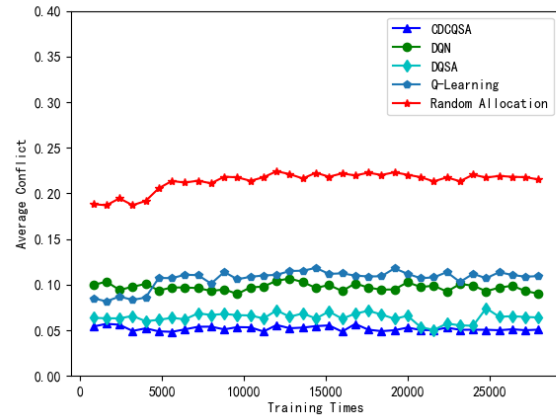


Fig. 3. The average collision rate when the five algorithms converge (PU and SU)

V. CONCLUSION

The spectrum access problem under a centralized DSA network is investigated in this paper. The problem of improving the spectrum utilization is modeled as a reinforcement learning process. In this paper, we propose the CDCQSA scheme for spectrum access, which optimizes the conflict problem among different SUs through the action screening

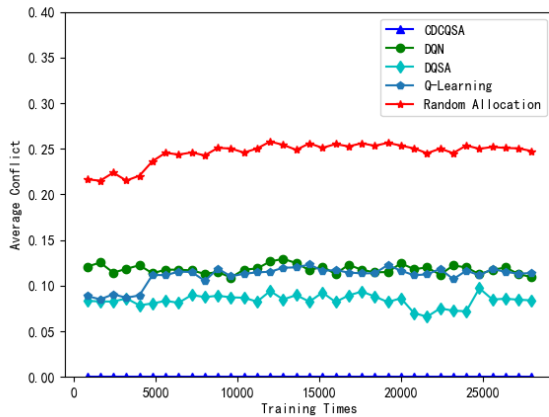


Fig. 4. The average collision rate when the five algorithms converge (SU and SU)

mechanism of the CU on the one hand, and achieves a major breakthrough in spectrum utilization improvement on the other hand. According to the simulation results, it is proved that our proposed CDCQSA scheme can improve the channel utilization and effectively reduce the PU-SU conflict rate even in the absence of a priori knowledge of PU signal.

ACKNOWLEDGEMENT

This work was supported by The National Natural Science Foundation of China under Grant 62071078.

REFERENCES

- [1] N. Morozs, D. Grace and T. Clarke, "Distributed Q-learning based dynamic spectrum access in high capacity density cognitive cellular systems using secondary LTE spectrum sharing," *2014 International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2014, pp. 462-467.
- [2] F. Shah-Mohammadi and A. Kwasinski, "Deep Reinforcement Learning Approach to QoE-Driven Resource Allocation for Spectrum Underlay in Cognitive Radio Networks," *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1-6.
- [3] O. Naparstek and K. Cohen, "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 310-323, Jan. 2019.
- [4] N. Morozs, T. Clarke and D. Grace, "A Novel Adaptive Call Admission Control Scheme for Distributed Reinforcement Learning Based Dynamic Spectrum Access in Cellular Networks," *The Tenth International Symposium on Wireless Communication Systems*, 2013, pp. 1-5.
- [5] N. Zhu, X. Xu, S. Han and S. Lv, "Sleep-Scheduling and Joint Computation-Communication Resource Allocation in MEC Networks for 5G IIoT," *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1-7.
- [6] H. Chang, H. Song, Y. Yi, J. Zhang, H. He and L. Liu, "Distributive Dynamic Spectrum Access Through Deep Reinforcement Learning: A Reservoir Computing-Based Approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1938-1948, April 2019.
- [7] Y. Xu, J. Yu and R. M. Buehrer, "The Application of Deep Reinforcement Learning to Distributed Spectrum Access in Dynamic Heterogeneous Environments With Partial Observations," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4494-4506, July 2020.
- [8] YC.Liang, "Smart Wireless Communication Technology Research Overview," *Journal of Communication*, vol. 6, no. 2, pp. 1-17, 2020.
- [9] N. Zhu, X. Xu, S. Han and S. Lv, "Sleep-Scheduling and Joint Computation-Communication Resource Allocation in MEC Networks for 5G IIoT," *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1-7.

- [10] Q. Cong and W. Lang, "Deep Multi-User Reinforcement Learning for Centralized Dynamic Multichannel Access," *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, 2021, pp. 824-827.
- [11] M. M. Rana, "An adaptive channel estimation technique for OFDM based cognitive radio systems," *14th International Conference on Computer and Information Technology (ICCIT 2011)*, 2011, pp. 315-320.
- [12] K. M. Thilina, K. W. Choi, N. Saquib and E. Hossain, "Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks," in *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2209-2221, November 2013.
- [13] M. M. Rana, "Power control algorithm for cognitive radio systems," *14th International Conference on Computer and Information Technology (ICCIT 2011)*, 2011, pp. 6-11.
- [14] A. Agarwal, S. Dubey, M. A. Khan, R. Gangopadhyay and S. Debnath, "Learning based primary user activity prediction in cognitive radio networks for efficient dynamic spectrum access," *14th International Conference on Computer and Information Technology (ICCIT 2011)*, 2016, pp. 1-5.
- [15] M. M. Rana, "Performance of sub-carrier mapping in single carrier FDMA systems under radio mobile channels," *2010 13th International Conference on Computer and Information Technology (ICCIT)*, 2010, pp. 175-180.
- [16] O. P. Awe, Z. Zhu and S. Lambotharan, "Eigenvalue and Support Vector Machine Techniques for Spectrum Sensing in Cognitive Radio Networks," *2013 Conference on Technologies and Applications of Artificial Intelligence*, 2013, pp. 223-227.



Guangjin Fang(1997-) is a master's student from Chongqing University of Posts and Telecommunications, with research interests in cognitive radio, machine learning, and spectrum access.