ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse211952020.pdf

https://doi.org/10.30534/ijatcse/2020/211952020

Cognitive Radio Networks with Reinforcement Learning Algorithms for Spectrum Allocation: A Survey

Dr. Anusha Marouthu1, Dr. V. Srikanth2, S. Sandeep3, M. Jeevan Babu4, Dr. D.Haritha5

1 Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dt, AP
2, 5 Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur dt, AP
3Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur dt, AP
4Assistant Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Namburu,Guntur dt, AP

ABSTRACT

Cognitive radio is a rising innovation which includes an advancement of software defined radio in which intelligence and decision making are incorporated. The principle capacity of cognitive radio is to misuse "spectrum-holes" or "white-spaces" to address the challenge of the low use of radio-resources. Dynamic spectrum-allocation, whose huge capacities are to guarantee that secondary users access the accessible frequency and data transmission to impart in an opportunistic way and to limit the impedance among licensed and unlicensed users, is a key system in cognitive radio networks (CRN). Reinforcement Learning (RL) which quickly investigates the measure of information without a model significantly encourages the exhibition of dynamic spectrum allocation in real-time application circumstances. This paper presents a study on the best spectrum-allocation algorithms in view of RL methods in CRN. The preferences and hindrances of every algorithm are investigated in their particular applications. At long last, we talk about open issues in dynamic spectrum allocation that can be subjects of future research.

Key words: Cognitive radios, reinforcement learning, spectrum-allocation

1. INTRODUCTION

The term CR was first coined by Mitola III in a seminar at royal institute of technology in Stockholm in 1998 and published in an article in 1999 for utilizing unused spectrum holes. He expressed that CR can use canny registering of remote individual computerized gadgets and related systems to recognize users' requirements for communication. The Federal Communications Commission (FCC) [1] told that CR has the capability to communicate and change its parameters by depending on intelligent data.

CR framework should address the following issues by:

1. Natural discernibleness and network models are restricted for configuring the transmission boundaries
2. How to co-ordinate the circulation of transmission hardware with restricted assets?
3. How to guarantee the assembly of the system under the state of contention between transmission devices?

Dynamic Spectrum-Allocation (DSA), which is characterized as a method for deciding the ideal planning between the accessible licensed channels and CR to accomplish ideal execution, is one of the fundamental instruments to organize interference in CRNs. The destinations of DSA in CRNs are:

1. To allocate the accessible channels to unlicensed users to accomplish professional spectrum usage.
2. To diminish the interference between unlicensed users.
3. To limit the interference among licensed and unlicensed users.

As of late, many scientists have examined and published many reviews on spectrum-allocation (SA) by concerning technical features of CR [2]. Strikingly, more and more specialists have as of late proposed various methodologies, for example, game-theory, fuzzy-logic, markov-randon-field and evolutionary algorithms, to take care of explicit issues and particularly to investigate the utilization of RL in SA (Figure-1).

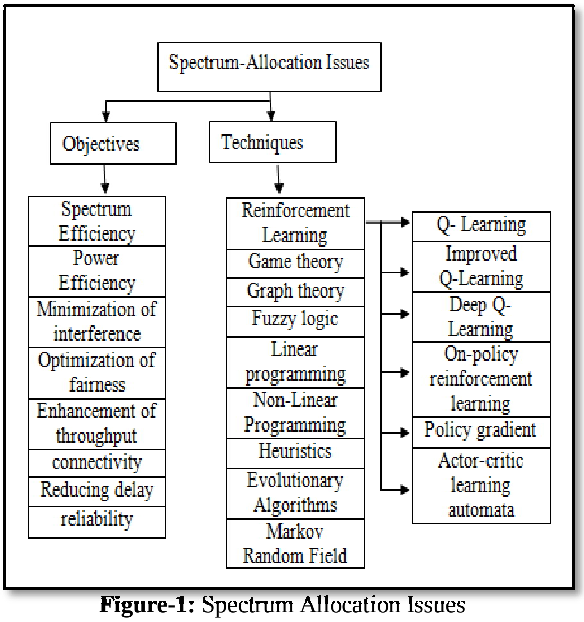

Figure-1: Spectrum Allocation Issues

The RL method is a significant class of AI that empowers users without any experience to constantly learn by experimentation and to boost the prize capacity or then again get the ideal technique. Thereave been a few pragmatic utilizations of RL, including Atari, AlphaGo and AlphaZero. RL can be utilized to understand huge scope (having huge state and activity spaces) issues with complex Markov Decision Process (MDP). Two principle highlights of RL are:

i. Trail & error process The early RL innovation was a normal experimentation learning framework, in which specialists execute activities to investigate nature with no earlier information. During the experimentation cycle, the key specialized issue is the harmony between investigations what's more, improvement. The delegate and momentous sans model RL calculations are worldly distinction (TD) learning and Q-learning. Current RL approach created from these calculations.

ii. Delay-rewards-Feedback signals are gotten by the operators from nature after taking activities. RL assignments can be ordered into two sorts as indicated by whether the decision-making undertakings are successive. Non-consecutive assignments underscore the thought of current prizes. As it were, the objective is to boost the compensations from state to activity right away [3]. On the other hand, consecutive undertakings are all the more testing and intend to amplify the long haul prizes from a progression of activities. There are a few focal points of RL for SA:

1. Effective observing of the ongoing powerful condition in a cordial way.
2. Continuous learning of new information to adjust to different extraordinary situations.
3. Easy recreation of an intricate domain that is hard to display precisely.
4. Access to significant information and data that are hard to track down and improved DSA performance.

5. Encourages breakthroughs in heterogeneous networks of data trade to enhance network performance.

Rather than the recently referenced overviews, our examination centers on various RL techniques in SA.

2. NETWORK MODEL

Obviously, the proposed algorithms are intended to take care of explicit issues in SA. Different strategies, including estimation-technique, game-theory, MDP, pricing theory and RL, have been applied to assemble a wide spectrum of numerical models of SA. Our review concentrates around the examination and use of RL in SA. Accordingly, RL isn't just the technique used to develop the models yet in addition the answer for different difficulties.

SA innovation can be partitioned into various classifications as indicated by different norms. Thinking about the strategy for admittance to spectrum assets, the innovation can for the most part be named spectrum underlay or spectrum overlay [5]. The system geography structure can be separated into two sorts: circulated and unified spectrum designation. Moreover, the collaboration mode can be named as cooperative SA and non-cooperative SA.

2.1 Spectrum underlay and overlay

Resources sharing of spectrum among licensed and unlicensed users can be partitioned into two sorts: spectrum underlay and overlay [6]. Unlicensed users have various limitations in the SA cycle; consequently, there are two unique sorts of spectrum asset sharing.

i. Spectrum-underlay

Ultra-wide-band (UWB) is commonly utilized as a spread spectrum innovation from which unlicensed users can access and utilize a similar frequency as that of authorized users. Accordingly, the frequency can be secured totally.

ii. Spectrum-overlay

The resources of spectrum might be gotten to astutely by unlicensed users without authorized users. Moreover, whenever authorized users involve their own spectrum data transmission, unlicensed users should compromise right away.

2.2 Distributive SA and centralized SA

i. Distributive SA

Each unlicensed user needs to identify whether there are authorized users in the ongoing condition [7]. At that point, they identify the data as indicated by the collaboration between unlicensed users and execute SA in blend with the technique of spectrum asset sharing.

ii. Centralized SA

The central-controller facilitates and oversees unlicensed users. Accordingly, the central-controller can distribute inactive spectrum to unlicensed users.

2.3 Cooperative SA & non-cooperative SA

i. Cooperative SA

Cooperative spectrum resource allocation can happen not just between unlicensed users and authorized users yet additionally between unlicensed users [8]. Productive sharing of spectrum assets is accomplished through data trade; accordingly, it can increment the use of spectrum assets and dodge obstruction all through the entire framework.

ii. Non-cooperative SA

The example of non-cooperative designation of spectrum assets can be portrayed as follows. A solitary secondary user watches the encompassing condition and afterward settles on spectrum choices, which don't perceive the shared data trade between users.

In this paper, the authors built up a numerical model for DSA issues dependent on the Q-learning, which is the most popular among RL strategies [9]. Here, we present a brief portrayal since a wide spectrum of RL strategies are created dependent on this model. To exploit the learning plan, it is important to build up the DSA as a unique programming issue or, identically, to decide the framework state x, the conduct a, the related cost r, and the following state y.

a. States

It is accepted that there are C cells and N available channels in the mobile communication system. M_t= (i, X(i))$_t$ is the meaning of the state M_t at time t, where i ∈ {1, 2, . . . , N} is known as the index. X(i) ∈ {1, 2, . . . , N} represents to the quantity of accessible directs in cell i at time t, which depends upon the direct use in cell i and its interfering cells I (i). To acquire X(i), the channel condition of cell q is characterized as a N-dimensional vector with every segment u_{qk} characterized as

u_{qk} = 1, I f divert k is being used in cell q

0, in any case (1)

where q = 1, 2, . . . , C, k = 1, 2, . . . , N.

Besides, an accessible vector s_q = $\{0, 1\}^N$ is framed as $s_q \in (s_q1, s_q2, . . . , s_qN)$, with every segment s_{qk} characterized as

s_{qk} = 0, I f channel k is accessible f or use in cell q

1,in any case (2)

where q = 1, 2, . . . , C, k = 1, 2, . . . , N.

We know that the channel condition of cell i and its interfering cells j ∈ I (i) and the accessibility vector s_i can be easily developed [13]. The comparing components are

$s_{ik} = \max\{u_{qk}|q \in I(i)\}$, k = 1, 2, . . . , N, I (i) = I(i) ∪ i. From s_{ik} , it is anything but difficult to get

$M(i) = \Sigma^N_{k=1} \bar{s}_{ik}$, where $\bar{s}_{ik}$ is the negation of s_{ik} .

b. Actions

Action ‘m’ is characterized by relegating a channel k from the accessible channel X(i) to the current call demand in cell i, that is, m = k, k ∈ {1, 2, . . . , N} and s_{ik} = 0.

Cost r (m, x) is utilized to speak to the prompt expense of making a action ‘x’ in state ‘m’, that is, r (m, k) = $c_1(k)r_1 + c_2(k)r_2 + c_3(k)r_3$, where $c_1(k)$ is the quantity of minimized cells in reference to cell i in which channel k is being utilized. $c_2(k)$ is the quantity of co-channel cells, furthermore, $c_3(k)$ is the quantity of other co-channel cells right now utilizing channel k.

Numerous strategies can be utilized to characterize the cost work contingent upon which boundaries is the principle thought [11].

c. Next-State

State advances from m_t to m_{t+1} are controlled by two irregular occasions.

3. RL-BASED SA ALGORITHMS

3.1 Q-learning algorithm based SA algorithms

Among RL procedures, the Q-learning is the most generally utilized. It continually refreshes the qualities in the Q-table that characterizes the states and afterward decides how to select an activity as per the present status. Q-learning is an off-procedure as Q-table updates esteems put together with respect to current experience as well as per past or indeed, evens others' encounters.

In the study [10] proposed a calculation dependent on Q learning out how to improve the presentation of DSA. In light of the attributes of the Q-learning search system, secondary users don't generally pick the channel with the biggest qualities however rather select another channel with an offered likelihood to investigate the earth for the ideal long haul rewards. The calculation not just accomplishes the self-rule of channel and force assignment yet additionally improves the information throughput and channel productivity.

The study [25] introduced an arbitrary picking dispersed SA conspires misusing RL in multicast earthbound correspondence frameworks that can fundamentally improve the force alteration by restricting the reassignment and dropping rates. In any case, this improvement is accomplished to the detriment of a higher blocking rate.

In [4] determined an assignment of resource scheme to licensed users (LUs) ideally assign the offered spectrum among various classes of unlicensed users (ULUs) while boosting the LUs' prizes. The Q-learning calculation is used to remove the ideal control strategy and to deal with the spectrum progressively. The secondary remote work organize can uphold the extra ULUs' traffic while ensuring the LUs QoS.

The problem of radio resource management the executives issue in open femto-cell networks has been completely concentrated [12]. Utilizing a smart asset allotment method, that is, the Q-learning algorithm, numerous femtocells can be spetrum ideally continuously.

In [20], a sale based Q-learning (QL-BA) calculation was proposed for an essential user (PU) and multiuser (OPMS) situation. The PUs in the model know all the data of the SUs, however the SUs know just their own data. The PUs and SUs can send data through the open channel. The bundle misfortune rate, portion effectiveness also, transmission rate execution are significantly improved.

Notwithstanding, in [20], the learning procedures of essential and auxiliary users are the equivalent, which isn't reasonable for SUs who know just their own data. Hence, in [21] a Q-learning calculation dependent on a twofold closeout (QL-DA) is proposed. This calculation mirrors the users' childishness in down to earth applications.

3.2 Improved Q-learning based SA algorithms

Despite the fact that Q-learning has been applied to SA, the normal issue with this calculation is that it includes indiscreet thought of the specific data and channel transmission states of user conduct [14]. Thinking about the existence of sensor nodes and other remote gadgets may share the spectrum, improved Q-learning algorithms for SA in distributive dynamic CRNs. The first process, called Q-learning+, utilizes precise channel inhabitance data to figure out how to improve the channel distribution choices. The second process, named Q-noise assesses channel transmission conditions by investigating the sign to-commotion proportion. The third algorithm, Q-Noise+, integrates Q-learning+ and Q-clamor into a solitary structure to consider both precise channel inhabitance data and channel transmission conditions. Examinations show that the exhibition of the three calculations is improved contrasted and that of the conventional Q-learning calculations [15]. Nonetheless, the improved calculations require precise channel-explicit data and along these lines increment the overhead of the channels.

A heterogeneous-network-self-optimization algorithm (HNSA) accomplishes self-advancement of a unique system dependent on an improved Q-learning strategy. This calculation respects the canny self-enhancing regulator (ISOC) of each radio access organize as an autonomous operator to change the variation issue in the reconfigurable frameworks into a multi-agent RL issue. This strategy decreases the framework blocking rate also, the unpredictability in DSA. Concentrating on 5G portable correspondences, an enemy of sticking MIMO NOMA power distribution furthermore, downlink transmission plot without knowing the channel and sticking condition, the NOMA framework is displayed as a lose-lose situation in which two systems are constructed imaginatively [16].

On one hand, the Dyna Architecture, which can be viewed as developing an encounter replay, quickens the speed of getting the ideal technique. Then again, the hot booting procedure, which utilizes a quick Q-learning calculation to get the instatement esteem, improves the combination contrasted with that of the zero introductions. Subsequently, the framework execution is extraordinarily improved. The depicted up to this point are all Q-learning calculations or their variations, that is, minor improvements made to Q-learning algorithms. Therefore, we present a basic summary in Table 1. In the accompanying areas, we will portray the different calculations proposed for the blend of RL and different philosophies.

3.3 Deep Q-networks-based SA algorithms

Lately, neural systems have gotten extensive consideration. The benefits of Q-learning with neural-networks are proposed in Q networks DQN. The greatest bit of leeway of DQN is the utilization of neural systems, as opposed to Q table, to gauge the state esteems, in which all states and activities must be put away. Confronting a enormous measure of information, DQN can conquer the impediments of PC equipment and incredibly improve information access and update speed [17]. A calculation like DQN was proposed and effectively applied in DSA. The calculation replaces the Q-table with a multilayer forward-engendering neural system by joining Q learning and back spread, which diminishes the outer sign obstruction and improves the system execution. The essential for application is that the verifiable action of the PU is excluded from channel determination, and every node must be gotten to trade data on the accessible channel [18].

Consequently, this technique expands the correspondence overhead. In the nonappearance of an impedance model and radio channel model, a SU can evade the hefty impedance area and select the recurrence jumping design by utilizing the DQN calculation. In particular, SUs use Q-figuring out how to acquire an ideal enemy of sticking correspondence procedure also, a profound convolution neural system

Method	Execution Method	Type of the Network	No.of Radios	Channel type	Advantages & Disadvantages
Q-learning based algorithm	Distributed	CRN	Multiple	N/A	It enhances the throughput, channel efficiency and controls the power
Q-learning based algorithm	N/A	Co-operative CRN	Single	N/A	Enhances the performance of anti-jamming and SINR
Q-learning based algorithm	Decentralized	Cognitive Wireless Mesh Networks	Multiple	Overlapping	Qurantees the LU's QoS and maintains ULU's traffic
Q-learning based algorithm	Distributed	public Femtocell Networks	Multiple	Orthogonal	Reduces the interference and enhances capacity of the network
Q-learning based auction algorithm	Centralized	MP MS-CRN	Multiple	Orthogonal	Improve the performance of transmission rate
Q-learning based double auction algorithm	Decentralized	MP MS-CRN	Multiple	Orthogonal	Shrinks the selfishness of the user
Q-learning+	Distributed	Industrial Networks	Multiple	N/A	Considers historic information
Q-noise+	Distributed	Industrial Networks	Multiple	N/A	Considers the channel QoS
Q-noise+	Distributed	Industrial Networks	Multiple	N/A	Identifies accurate historic information for channel quality
Fast Q-learning	Centralized	Wireless Networks	Multiple	Non-Orthogonal	Enhances the data rates

Table-1: Reinforcement Learning based SA algorithms

to quicken the learning speed [19]. The results show that the proposed technique accomplishes a quicker assembly rate, higher SINR, lower hostile to impedance cost and better execution of SUs against agreeable jammers contrasted and those of Q-learning alone.

3.4 On-policy RL-based SA algorithms

The Q learning algorithms referenced above are off-policy calculations. At the end of the day, the next state and activity are dubious when the calculation is refreshed. Paradoxically, the most delegate on-strategy calculation is the State Action Reward State Action (SARSA) calculation [22]. The SARSA calculation speaks to the entire cycle in a way. The next state and activity are the state and move it really makes and updates when the calculation has been distinguished. On the off chance that Q-learning is an avaricious and fearless calculation, SARSA is a moderately moderate and tentative calculation. Since the SARSA calculation is touchier to trial and error rate, in reasonable applications, if the harm is a significant thought, the SARSA calculation can accomplish better execution during preparing.

It is difficult to get exact MDP model for real complex circumstances. It utilizes the promptly accessible neighborhood condition data for learning users. Lessening the state space prompts decreased memory necessities and cost without huge execution misfortune [23]. Likewise, a disentangled SARSA calculation for spectrum redistribution utilizing just neighborhood data was recreated. The new call and exchanging blocking likelihood is nearly equivalent to the SARSA utilizing framework level impetus data, which demonstrates the practicality of the conveyed RL strategy.

3.5 Policy-gradient-based SA algorithms

This algorithm depends on strategy emphases. Its yield isn't the estimation of the activity yet the particular activity, which can be chosen on a ceaselessly appropriated activity [24]. To tackle the issue of DSA in CRNs, ULUs' channel access measure was worked as a confined mostly recognizable Markov choice cycle POMDP. In the confined POMDP, the prize capacity is utilized to gather ULUs immediate prizes; also, the cost work mirrors the prompt expense of the PU because of channel obstruction from ULUs. By utilizing the Lagrangian multiplier strategy to change over the first obliged POMDP into an unconstrained POMDP, the learning calculation dependent on the approach inclination gains the neighborhood ideal technique.

The authors relocated the genuine strategy inclination calculation, POMDP calculation and wordy support calculation to a CR mindfulness network. The exploratory outcomes show that the genuine approach slope calculation is better than the other two calculations regarding precision and information end to end delay execution.

Most RL algorithms are applied to discrete irregular issues, for example, the use of ISM band spectrum. Perception of the channel condition can decide the framework condition cutoff points and potential models. The following stage is to ascertain all potential answers dependent on MDP, which breaks down the conduct of the boundaries for future expectations before setting the last boundaries [25]. So, the learning cycle is the most proper methodology for the current condition. The framework at that point sets the underlying channel as the current channel that the sensor network utilizing to ascertain the following comparing channel dependent on the acquired strategy.

3.6 Actor-critic learning-automata-based SA algorithms

This algorithm, like Q-learning, is known as the pundit just calculation as a result of the force based cycles. Be that as it may, it is troublesome for a pundit just calculation to join. The inclination learning calculation dependent on learning automata (LA) permits the subject to straightforwardly become familiar with a fixed randomization system, so LA can be viewed as an entertainer just calculation that doesn't utilize any type of put away worth capacities [26].

This technique gives another calculation, the actor-critic LA (ACLA) calculation, which depends on the iterative calculation, to consider the state estimation of MDP and the strategy dependent on the LA calculation Figure-2. The upside of the algorithm is that it can perform refreshes in a solitary advance, making it quicker than the conventional policy gradient.

4 CHALLENGES AND OPEN ISSUES

In the past areas, we introduced a diagram and a concise portrayal of the proposed RL calculations. By the by, a few open issues remain. CRN issues

i. **Operating frequency** issues for SA is the one of the most significant parameter. The most remarkable component of secondary users is that they can change their working recurrence to another accessible working recurrence in view of their area or the sign quality [28]. In other words, ULUs can change to the most reasonable recurrence. This trademark should be thought of.

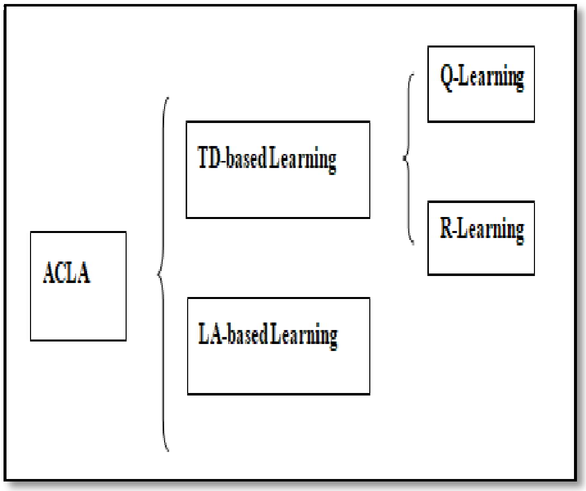

Figure-2: Relationship of ACLA

In a multiband secondary radio spectrum identifies calculation dependent on entertainer pundit LA [27]. The structure of the calculation in Figure- 3 was created in by incorporating a cross-layer advancement structure, and it is appropriate to situations with dynamic spectrum asset accessibility. We sum up the above strategies in Table 2.

ii. Transmission power is another significant pointer in SA. High transmission forces of SUs are unwanted sin light of the fact that they quickly exhaust the SUs' battery what's more, cause more impedance between the SUs and PU. To this end, the dynamic transmission power is organized inside the reasonable force go [29]. The force control ought to likewise consider the real channel conditions and the QoS prerequisites of the progressing SUs. Under the force control, the transmission power is decreased to a satisfactory least. In like manner, the relating impedance is diminished, which permits more users to share the accessible spectrum.

4.1 Channel-model

SA algorithms utilize a Rayleigh fading channel model, and few think about different models. The current calculations ought to be tried for various channel models, and new calculations ought to be proposed [31]. For various application scenarios, CR entity correspondence influences the qualities of remote channels, regardless of whether urban or provincial, in-door or out-of the calculation execution.

4.2 Distributed and mobile situations

The examination on circulated asset assignment for base level CRNs is restricted contrasted with that on incorporated calculations. For progressively complex correspondence frameworks, distributed should win to fulfill user versatility and high QoS principles. Thusly, research around there is valuable and empowering.

4.3 Experimental-testbed

It is essential to test these strategies in functional conditions to successfully assess the exhibition of

Method	Execution Method	Type of the Network	No.of Radios	Channel type	Advantages & Disadvantages
Q-Learning based network self-optimization algorithm	Distributed	Wireless LAN	Multiple	N/A	It lowers complexity and enhances performance of the network
Similar deep Q networks based algorithm	Distributed	Cognitive Wireless LAN	N/A	Non-Overlapping	Decreases interference and increases overhead
Deep Q networks based algorithm	Distributed	Anti-Jamming Communication CRNs	Multiple	N/A	lowers the interference and enhances the performance of Sus
Price search algorithm	Distributed	Cognitive Interference Networks	Single	N/A	Improves communication rate
True policy gradient algorithm	Distributed	Cognitive Wireless Sensor Networks	Multiple	N/A	High accuracy
POMDP algorithm	Distributed	Cognitive Wireless Sensor Networks	Multiple	N/A	Low accuracy
Episodic reinforcement algorithm	Distributed	Cognitive Wireless Sensor Networks	Multiple	N/A	Moderate accuracy
Actor-critic learning automata- based algorithm	Distributed	Multi-Band cognitive Radio Scenario	Multiple	N/A	Integrates cross-layer framework

Table-2: Reinforcement Learning based SA

CRN innovation. CR testbeds give an compelling route for analysts to precisely assess their outcomes in genuine circumstances. A great recreation condition improves the proficiency of the investigation as well as generously lessens costs. This issue has not been broadly considered in the current writing on fundamental CRNs, as existing test systems do not have an advanced CR module.

4.4 Reinforcement-learning issues

1. Seamless mix of RL and different methodologies: Most of the above works consider just RL calculations to tackle radio asset the board issues. To accomplish better asset assignment or more adaptable solutions, the points of interest of RL models and different other options can be joined. For instance, game-based examination applies just to the operator's learning elements, while natural elements are not expressly thought of. Subsequently, it is promising to take care of testing issues by joining numerous best in class calculations, for example, the mix of RL and game hypothesis.

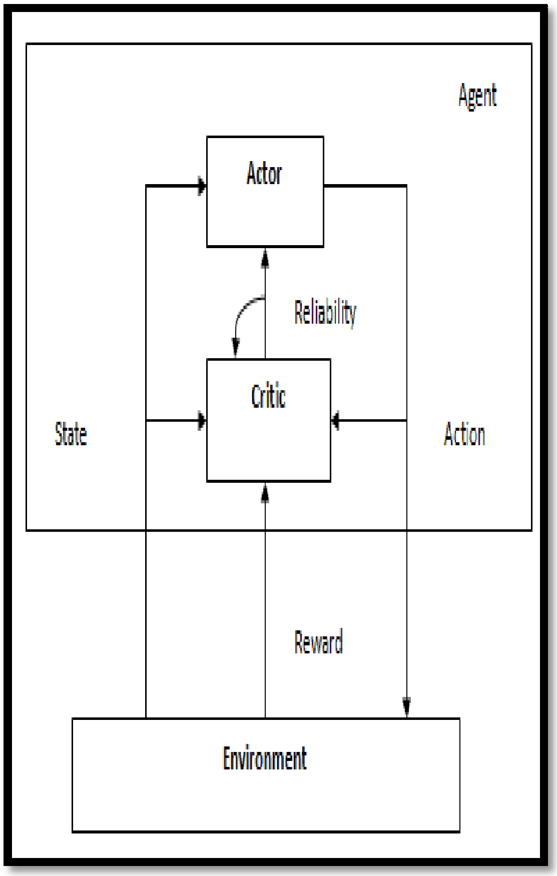

Figure-3: Actor-Critic Architecture

2. Improve the relevance DSA includes finding the focal recurrence as well as additionally choosing the recurrence and ideal transmission capacity the SU needs to access in an adaptable and constant way. Some developing RL calculations have not been applied in CRNs and remain an open territory for future examination. RL calculation for the investment of childish users in SA learning plans, the test of how they can profit by participation with different SUs is worth further examination.
3. Consider enormous scope and thick distribution At present, numerous RL calculations are as it were applied to some single DSA issues. Be that as it may, the state and activity space in actuality is large-scale also, consistent, so the current strategies are not appropriate [32]. Subsequently, to fortify the strategies to take care of different commonsense issues, further exploration is needed to enhance the present algorithms. For instance, in useful situations, psychological users are huge scope, and it is unreasonable to utilize conventional Q-learning with a Q table to store information. Therefore, we can consolidate neural systems with RL to improve the capacity to handle a lot of information.

5. CONCLUSION

DSA is a key structure issue for secondary radio innovation. In this paper, we group the current RL calculations in CRNs by introducing a topical scientific categorization and an overview of the cutting edge SA calculations. There are six parts, including Q-learning, improved Q-learning, profound Q-systems, on-strategy RL, strategy slope, and entertainer pundit learning automata. The basic parts of the current SA calculations are additionally investigated to decide their qualities and shortcomings. At last, we examine a few open issues and difficulties that have not been completely researched and that could be the reason for future work around there.

REFERENCES

1. Dhage, M. R., & Vemuru, S. (2018). A Effective Cross Layer Multi-Hop Routing Protocol for Heterogeneous Wireless Sensor Network. Indonesian Journal of Electrical Engineering and Computer Science, 10(2), 664-671.
2. Babu, K. S., & Vemuru, S. (2018). A low cost software defined radio based cognitive radio

testbed for LTE networks. International Journal of Engineering and Technology (UAE), 7(3.1).

3. Al-Rawi, H. A., Ng, M. A., & Yau, K. L. A. (2015). Application of reinforcement learning to routing in distributed wireless networks: a review. *Artificial Intelligence Review*, *43*(3), 381-416.
4. Alsarhan, A., & Agarwal, A. (2011). Profit optimization in multi-service cognitive mesh network using machine learning. *EURASIP Journal on Wireless Communications and Networking*, *2011*(1), 36.
5. Babu, K. S., & Vemuru, S. (2018). A low cost software defined radio based cognitive radio testbed for LTE networks. *International Journal of Engineering and Technology (UAE)*, *7*(3.1).
6. Faganello, L. R., Kunst, R., Both, C. B., Granville, L. Z., & Rochol, J. (2013, April). Improving reinforcement learning algorithms for dynamic spectrum allocation in cognitive sensor networks. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 35-40). IEEE.
7. Han, G., Xiao, L., & Poor, H. V. (2017, March). Two-dimensional anti-jamming communication based on deep reinforcement learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2087-2091). IEEE.
8. Anusha, M., & Vemuru, S. (2020). An Effective MAC Protocol for Multi-radio Multi-channel Environment of Cognitive Radio Wireless Mesh Network (CRWMN). In *First International Conference on Sustainable Technologies for Computational Intelligence* (pp. 21-35). Springer, Singapore.
9. Levorato, M., Firouzabadi, S., & Goldsmith, A. (2012). A learning framework for cognitive interference networks with partial and noisy observations. *IEEE transactions on wireless communications*, *11*(9), 3101-3111.
10. Li, Y., Feng, Z., Chen, S., Chen, Y., Xu, D., Zhang, P., & Zhang, Q. (2011). Radio resource management for public femtocell networks. *EURASIP Journal on Wireless Communications and Networking*, *2011*(1), 181.
11. Lv, C., Wang, J., Yu, F., & Dai, H. (2013, March). A Q-learning-based dynamic spectrum allocation algorithm. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering*. Atlantis Press.
12. Marinho, J., & Monteiro, E. (2012). Cognitive radio: survey on communication protocols, spectrum decision issues, and future research directions. *Wireless networks*, *18*(2), 147-164.
13. Mitola, J., & Maguire, G. Q. (1999). Cognitive radio: making software radios more personal. *IEEE personal communications*, *6*(4), 13-18.
14. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.
15. Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. N. (2020). Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms. *International Journal*, *8*(5).
16. Qadir, J. (2016). Artificial intelligence based cognitive routing for cognitive radio networks. *Artificial Intelligence Review*, *45*(1), 25-96.

17. Ru, M., Yin, S., & Qu, Z. (2017). Power and spectrum allocation in D2D networks based on coloring and chaos genetic algorithm. *Procedia Computer Science*, *107*, 183-189.
18. Anusha M, Dr.Srikanth V (2019) An effective resource utilization for multi-channel multi-radio environment in cognitive radio wireless mesh networks. International Journal of Recent Technology and Engineering (IJRTE).7(6):pp.393.
19. Salameh, H. A. B. (2011). Throughput-oriented channel assignment for opportunistic spectrum access networks. *Mathematical and Computer Modelling*, *53*(11-12), 2108-2118.
20. Babu, K. S., & Vemuru, S. Spectrum Signals Handoff in LTE Cognitive Radio Networks Using Reinforcement Learning (2018). Spectrum Signals Handoff in LTE Cognitive Radio Networks Using Reinforcement Learning, 119-125.
21. Teng, Y., Zhang, Y., Niu, F., Dai, C., & Song, M. (2010). Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks. In *2010 IEEE 72nd Vehicular Technology Conference-Fall* (pp. 1-5). IEEE.
22. Teng, Y., Yu, F. R., Han, K., Wei, Y., & Zhang, Y. (2013). Reinforcement-learning-based double auction design for dynamic spectrum access in cognitive radio networks. *Wireless Personal Communications*, *69*(2), 771-791.
23. Anusha, M., & Vemuru, S. (2018). Cognitive Radio Networks: State of Research Domain in Next-Generation Wireless Networks—An Analytical Analysis. In *Information and Communication Technology for Sustainable Development* (pp. 291-301). Springer, Singapore.
24. Wang, W., Kwasinski, A., Niyato, D., & Han, Z. (2016). A survey on applications of model-free strategy learning in cognitive wireless networks. *IEEE Communications Surveys & Tutorials*, *18*(3), 1717-1757.
25. Xiao, L., Li, Y., Dai, C., Dai, H., & Poor, H. V. (2017). Reinforcement learning-based NOMA power allocation in the presence of smart jamming. *IEEE Transactions on Vehicular Technology*, *67*(4), 3377-3389.
26. Mukherjee, A., Choudhury, S., Goswami, P., Bayessa, G. A., & Tyagi, S. K. S. (2019). A novel approach of power allocation for secondary users in cognitive radio networks. *Computers & Electrical Engineering*, *75*, 301-308.
27. Yang, R., & Ye, F. (2010, September). 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)* (pp. 1134-1137). IEEE.
28. Marouthu, A., & Vemuru, S. (2018). Cluster-Based Opportunistic Spectrum Allocation in CRWMN's Using Co-operative Mechanism. *Wireless Personal Communications*, *99*(2), 779-797.
29. Yau, K. L. A., Komisarczuk, P., & Teal, P. D. (2012). Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues. *Journal of Network and Computer Applications*, *35*(1), 253-267.
30. Pachipala A, Prachod E, Pavan Kumar T, Babu Rao D, Srinivas M (2020) A dynamic resource allocation approach to obtain load balancing in

fog computing environment. *Journal of Critical Reviews.* 7(6): pp. 15-19.

31. V.Manikandan, Dr. M. Sivaram.(2019). Energy Efficient Dynamic Routing in Wireless Sensor Networks. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(5), 2189- 2193.
32. Jojo F. Blanza, Lawrence Materum. (2019). Wireless Propagation Multipath Clustering: On Simultaneously Solving the Membership and the Number of Clusters. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(5), 1914-1919.