Dynamic Systems and Applications 30 (2021) No. 11, 1719 - 1732

REINFORCEMENT LEARNING BASED HANDOFF MECHANISM IN COOPERATIVE COGNITIVE RADIO NETWORKS

Vineetha Mathai[1*] And P. Indumathi[2]

[*1, 2]Department of Electronics Engineering, MIT Campus, Anna University, Chennai. Mail id:

vineethamathai@gmail.com

ABSTRACT. The spectrum handoff (SH) is a dynamic spectrum access technique which ensures effective channel utilization, fair resource allocation, as well as uninterrupted real-time connection. Facilitating SH across traffics of dissimilar characteristics in Cognitive Radio Networks (CRNs) is posing difficulty due to manifold interventions from Primary Users (PUs), disagreement among Secondary Users (SUs) and diversified Quality of Experience (QoE) demand. Here, we consider effective channel selection strategy (CSS) and put forward a learning-based handoff scheme to enhance QoE demand of users by the introduction of docition idea. A PU prioritized Markov method is introduced to represent the communications between PUs and SUs for even channel access. The reinforcement learning (RL) is applied to CSS to carry out proper channel selection. Numerical outcomes projects that proposed queuing model, suggested learning based handoff scheme and docitive learning enhances the quality of service by maintaining the average MOS of 3.6.

Keywords: Cognitive radio network, Spectrum handoff, Queuing Model, Reinforcement Learning, QoE.

1 INTRODUCTION

The progression of wireless communication towards 5G includes changes in network model and assessment of providing QoE for multimedia applications. The term CRN is coined to mitigate the effect of underutilization of spectrum resources [1],[2]. In CRN, unlicensed users (SUs) are having chance to access the spectrum only when it is not engaged by licensed users (PUs). If a PU returns on a channel, SU can either stay on it or shift (ie., handoff) to another one until the completion of PU's data transmission. If cognitive radio is shadowed by a high building over the sensing channel, then cooperative mechanism is included.

Proactive, reactive and hybrid handoff [10] are the various methods available in the literature. In the proactive method, to characterize PU's activities, to identify channels and to accomplish switching on revisit of PU, SUs uses the information of PU traffic model. So, handoff delay of this scheme is less but to get precise traffic model of PU is difficult. In the reactive mode, an SU does spectrum sensing initially when a PU interruption happens to identify vacant channels. So, channel status for handoff could be found without difficulty. However, it may bring delay. In hybrid method, a speedy method has combination effects of earlier methods by means of the proactive sensing and reactive handoff action [3-5].

Multimedia applications [12],[15] is difficult to introduce in CRN due to intervention of PUs and different requirements of QoE. In order to tackle previous problems we select a mixed preemptive and non-preemptive resume priority (PRP/NPRP) M/G/1 [31] queueing model to describe behavior of PUs and SUs on spectrum usage. Here, the former model is used to describe the queueing of the PUs and SUs and to ensure that PUs have control. To avoid an SU from intruding the current communication of other SUs, the queueing between them is modeled with latter model. When picking channels for SH, it is significant to study the transmission delay, channel quality and conditions.

The varying channel situations and traffic loads, the knowledge gained from prior SHs and earlier channel environments, a reinforcement learning-based [18]-[20],[22]SH scheme is proposed to adaptively achieve SH[7],[15-16]. The main parameter of QoE [30] is mean

Received JUL 12, 2021
www.dynamicpublishers.com;
www.dynamicpublishers.org;

ISSN1056-2176(Print); ISSN 2693-5295 (online)
$15.00 ©Dynamic Publishers, Inc.
https://doi.org/10.46719/dsa202130.11.03

opinion score (MOS), which is an end-user fulfillment measure. The offered system maximizes the total MOS through a RL [29] method where PUs co-occur with SUs by contacting the same frequency band, by considering a required interference limit to the PUs. The usage of MOS [24-28]lets coaching the nodes carrying different load as an outcome, the docitive [34] model is put in to examine influence of diverse docition situations where, a newcomer node being educated by practiced peers.

Contributions in this paper

1. A mixed queuing model is developed to provide differentiated service which considers conditions such as PU's interventions, prioritized traffic etc. of channels.

2. A new method of Q- learning based spectrum handoff for various traffic over CRN is suggested. For inclusion of heterogeneous multimedia applications this scheme considers QoE requirements of SU, packet loss rate of SU due to handoff delay etc.

3. A new concept called docition is applied for efficient resource allocation to investigate effect of different scenarios with acceptable MOS as performance metric. Here the new CR nodes are being taught by the already existing nodes so the learning time can be reduced. This proposed method effectively reduces the number of iteration required for convergence hence resource allocation can be done effectively.

Remaining paper is ordered as follows: Section 2 introduced system model with problem statement. Executed results and conclusion are explained in section 3 & 4 respectively.

2 REPRESENTATION OF CHOSEN SYSTEM AND IDENTIFICATION OF PROBLEM

CRN with M number of channels, M PUs and N number of SUs are considered here. To access a specific channel the PUs are given the utmost priority. The SUs get a chance to approach the channel after the PUs transmission is finished. If SU is sending information and a PU arrives at that time, the SU ought to stop sending and decide to change to the other channels or wait at the same till the PUs transmission get ended. Let for transmission and sensing, SU is set with two transceivers as in [10]. When PU enters in the current channel, the SU executes handoff in our model, as in the IEEE 802.22 standard [21]. Here a hybrid SH scheme by make use of proactive sensing of channel and reactive handoff activity[3] is utilized.

2.1 Queuing Model

PRP M/G/1 queuing network method is applied to describe the spectrum handling behavior of the PU and SU whereas actions between SUs are designed using NPRP M/G/1 queue. This mixed model considers higher priority queuing SUs for access, by not letting SUs with lower priorities to transmit the data until there is no high preference SUs in the queue. For delay sensitive SU applications this model suits well.

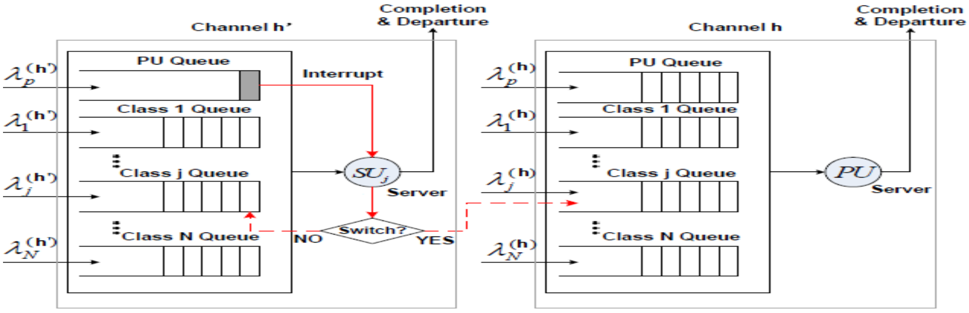


Figure 2.1 Considered queuing model

Figure 2.1 describes channels h and h' with one priority queue for a PU and for SUs. PUs and SUs are set high and low priority queues respectively. If on the arrival of PU (interrupt), then SU with priority j remains on the channel h' has to stay on same channel h' or move to another channel to finish its transmission and the remaining communication take place in front end of low priority queue while if it moves to another channel remaining communication take place in back end of the same queue.

The arrival pattern follows Poisson distribution with arrival rates $\lambda_{p\,1}, \lambda_{p\,2}, \ldots \lambda_{p\,N}$ and $\lambda_{s\,1}, \lambda_{s\,2}, \ldots \lambda_{s\,N}$. The service times obeys exponential distribution. On the arrival of PU, multiple interruptions may happen to SU. Each time SU has to perform sensing for finding unoccupied space to go on with uncompleted connections which causes handoff delay. On next interruption SU finds all the channels are busy and it waits on same space until PU completes its communication so handoff delay becomes addition of sensing period and the busy period obtained from multiple PUs. Citing [29] the average waiting time E$[W_p^{(h)}]$ of a PU connection and the average number E$[N_p^{(h)}]$ of PU connections in the M/G/1 queue can be obtained as

$$\mathrm{E}[W_p^{(h)}] = \frac{\lambda_p^{(h)} E[(X_p^{(h)})^2]}{2(1-\lambda_p^{(h)} E[X_p^{(h)}])} \tag{1}$$

$$\mathrm{E}[N_p^{(h)}] = \lambda_p^{(h)}\ \mathrm{E}[W_p^{(h)}] \tag{2}$$

where $\lambda_p^{(h)}$is the arrival rate on channel h of PU, $E[X_p^{(h)}] = \frac{1}{\mu_p^{(h)}}$ is wating time of PU at channel h with service rate $\mu_p^{(h)}$. The SH cases can be staying or changing phase. So the delay E $[D_{j\,i}^{(h)}]$ is found out to be

$$\mathrm{E}[D_{j\,i}^{(h)}] = \min\,(E[W'^{(h)}_j], E[W_j^{(h)}] + t_s) \tag{3}$$

where $E[W'^{(h)}_j]$, $E\ [W_j^{(h)}]$ are the expecting waiting times when SU stays at same channel and changes to other channel respectively. $\omega_i^{(h)}$is the SU's arrival rate on channel h. The fruitful service time of SU after dealing with ith interruption is $E[\phi_i^{(h)}]$, the maximum interruptions SU can tolerate is n_{max} and the switching time is t_s.

$$E[W'^{(h)}_j] = \frac{E[X_p^{(h)}]}{1-\lambda_p^{(h)} E[X_p^{(h)}]} \tag{4}$$

$$E[W_j^{(h)}] = \frac{\lambda_p^{(h)} E[(X_p^{(h)})^2] + \sum_{i=0}^{n_{max}} \omega_i^{(h)} E[(\phi_i^{(h)})^2] + \frac{(\lambda_p^{(h)})^2 E[(X_p^{(h)})^2] E[X_p^{(h)}]}{1-\lambda_p^{(h)} E[X_p^{(h)}]}}{2(1-\lambda_p^{(h)} E[X_p^{(h)}]) - \sum_{i=0}^{n_{max}} \omega_i^{(h)} E[(\phi_i^{(h)})^2]} \tag{5}$$

If the communication of SU is successful then it increases the service time and checks if the current communication is finished. Contrarily it increases the waiting time of current transmission and checks it reaches the threshold value. In this case SU drops one packet or else it continues communication.

2.2 Spectrum Handoff (SH)

For end user contentment, influence of several factors like data rate, length of the packet, packet loss, SINR, channel conditions end to end delivery time etc., on SH needs to be considered. A vital standard for QoE [30] is Mean Opinion Score (MOS) is selected to evaluate the user's viewpoint of quality. It can have values from 1 to 5. Since the channel is time varying, either loss of packet [8,9,11,17] due to delay created by SH or packet error rate (PER) due to low channel conditions can occur.

$$\text{MOS}=(r_1+r_2.FR + r_3.\ln(\text{SBR}))/\ (1+r_4.TPER+r_5.(TPER)^2) \quad (6)$$

where r_1,r_2,r_3,r_4,r_5 can be found by non linear regression. In this work sender bit rate (SBR) and frame rate (FR) are constant. The quality of channel is represented by total packet error rate (TPER), which showcases effect of failed transmission probability (FTP).

$$\text{TPER} = \text{PER} + \text{FTP} - \text{FTP. PER} \quad (7)$$

On arrival of PU, SUs connection breaks so for smooth SH, it is necessary to choose a channel with high MOS. To maximize MOS, handoff delay as well as PER should be minimal.

2.3 A Learning based Spectrum Handoff

The motive of bringing in reinforcement learning (RL) [32] in SH method is to reduce the failed transmission probability of SU so that it can have more fruitful data transmission link. The SH scheme based on RL chooses action in the present state in agreement with the values of the Q table and its selection policy (e.g. e-greedy policy)[14]. We modeled SH decisions as Markov decision process (MDP) [31] and reward of the action is MOS. A MDP can be represented as follows: S is states, A is set of actions, T is probabilities of state change and R is reward for particular action respectively. The iterative steps of MDP includes sensing of environment by intelligent agent followed by observing its state then find the suitable action to generate best reward based on transition to next state. Finally the policy gets updated and repeat it. To find optimal policies, we choose model free RL method which is Q learning. The Q learning for SH method is mentioned as follows.

a. States of SUs associations

The current state of SU, SU_j is indicated at channel h, when $(i+1)^{th}$ interruption happens as $s_{j,i} = \{\zeta^h_{j,i}, \omega^h_{j,i}, \phi^h_{j,i}\}$ where channel status is $\zeta^h_{j,i}$, arrival and service rate is $\omega^h_{j,i}, \phi^h_{j,i}$ respectively of (j,i) kind of SU. It reflects the interference caused by SUs.

b. Actions of SUs associations

Let $a_{j,i}=\beta^h_{j,i}$ represents actions of SUs on the state $s_{j,i}$ when intervention occurs and $\beta^h_{j,i}$ shows the probability of choosing channel h after intervention.

c. Rewards of SUs associations

Here Q learning tries to accelerate MOS, the reward with an attempt to stabilize the handoff delay based on priority assigned to SUs.

d. Learning of SU associations

The main aim is to identify the optimal action with current policy $\pi(s_{j,i}, a_{j,i})$ which increases MOS. Based on Q table entries, SH method selects the action in present state. The Q table updating process is denoted as

$$Q(S_t,A_t) = Q(S_t,A_t)+\alpha(R_{t+1}+ \gamma \max_a Q(S_{t+1},a) - Q(S_t,A_t)) \quad (8)$$

where α is the learning rate (0< α <1), γ is the rate of discount and for the time slot t, R_{t+1} is the reward function, S_t, A_t are the corresponding state and chosen action respectively.

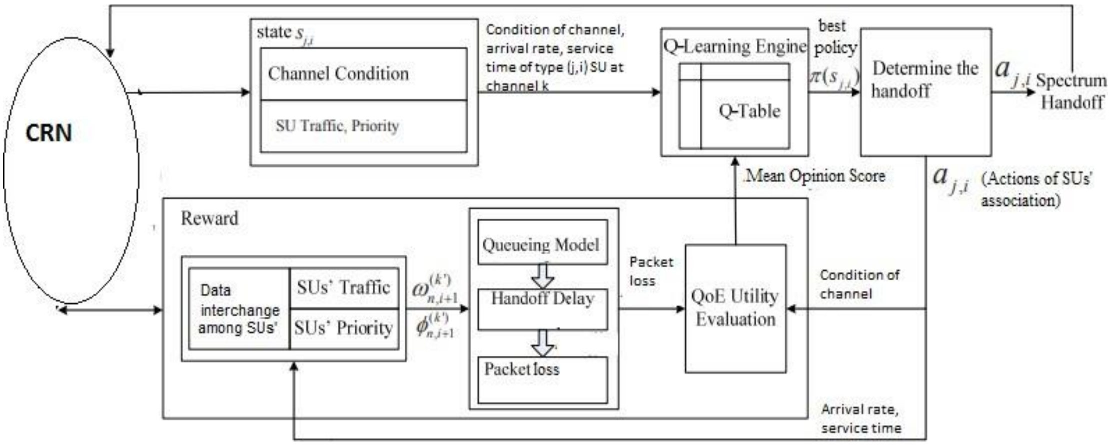


Figure 2.2 Proposed learning based SH scheme

The proposed methodology is illustrated in figure 2.2. Initially CRN observes the present state of SU at its ith intervention. Then after interruption CR node chooses a handoff activity for (i+1)th interruption caused by PU again based on present state from Q table[6,13]. So SU makes a move to next state $s_{j,i+1}$ as a result MOS is obtained as reward and further updates the Q table [32]. Learning based proposed SH is mentioned in pseudo- code 1.

Pseudo code 1: Q – learning based spectrum handoff method

Input: $\lambda^{(h)}_p$, $E[X^{(h)}_p]$, FTP$_h$ ∀ h

Output: $\pi(s,a)$ - the best policy

```
Initialize Q(s, a) randomly.
Create arrival rate by using $\lambda^{(h)}_p$
Redo for all events
  Initialize all state s.
  Redo
    if PU reaches at channel h
      if channel h is idle
        PU is present at channel h.
      else channel h is busy by other PU
        PU moves queue
      else channel h is occupied by SU
        // Performs spectrum handoff due to interruption
        Select an action for state resultant from Q-table.
        Performs spectrum handoff based on a
        if the latest channel = = the same channel
          Find wait time $E[W^{(h)}_j]$
        // Obtain total delay $d_{j,i}$ for i^th intervention $d_{j,i+1} = d_{j,i} + E[W^{(h)}_j]$
        if $d_{j,i+1}$ >=delay threshold of SU, drop packet and repeat.
        Update FTP.
Update MOS using (6), Q-table using (8), policy $\pi(s, a)$
else
    repeat the method.
else
    if service time >= T, the essential serve time
        SU successfully communicates.
    else
        Carry on the process.
```

2.4 Proposed Docition based learning for SU

By iteration, Q learning learns and for every action a reward is stored in Q table. Each cognitive SUs study about their environment and selects an action which yields high reward but its environment shows less variations and incompetent if we replay cognitive cycle. Q table portrays the properties of wireless environment. When a SU joins with already learnt scheme the awareness of the environment from Q table can be imparted to lessen learning time and progresses act of learning. The model is docitive radio. The motive of CR is to study while docitive radio gives attention on coaching. Under this new standard, for resolving a problem the nodes with more "experience" will teach fewer able nodes thus learning time [34] can be decreased and enhances the learning act. SUs present previously in scenario started their cognitive cycle with their Q table. When a less experienced node joins it resets the Q table by taking mean of the Q-table entries with different categories of load from SUs as expressed in (9).

$$Q_c = \frac{1}{N}\sum_{i=1}^{N} Q^{(i)} \tag{9}$$

The individual learning and docitive mechanism is explained in pseudo-code 2 and 3.

Pseudo- code 2: Individual learning

```
Initialize Q(s,a).
For t<tmax
    Choose action maximises Q(st,at)
    Update state, reward and Q value using (8).
end
```

Pseudo- code 3: Docitive (co-operative) learning

```
New SU joins the system as SUN+1
Initialize Qc using (9)
For $\forall$ $SU_i$ i= 1,…., N+1
    Start individual learning with available N+1 Q tables
End
```

We considerd six systems during simulations. Firstly considered called "individual learning" where all SUs carry out individual learning executed in pseudo-code 2, next, called "new comer with docition", "new comer with docition of similar traffic", "new comer with docition of dissimilar traffic", "new comer with docition of immediate neighbor" and "new comer with docition of random neighbor", respectively, letting one SU already learned the system. When first system executes individual learning for the joined SU, the others teach the new comer through the “docitive” approach implemented in pseudo code 3. By taking mean over Q-values various scenarios are evaluated and QoS enhancement is achieved.

3 RESULTS AND DISCUSSIONS

Here we perform investigations to assess the performance of suggested traffic adaptive SH technique. Let the service time of both connections obeys exponential distributions with values 0.6 and 0.5, time slot duration be 10msec [21]. Let number of channels and secondary connections be 4 and 6. The arrival rate of PU and SU take the values mentioned in table 4.1. The channel switching time is 1ms. Users with same priority follows first come first serve scheduling plan to avoid collisions. Primary network consists of one PU with bandwidth of 10MHz was considered. The signal to interference plus noise ratio (SINR) is fix to be 10dB. The noise power and transmit power are fix to be 1nW and 20mW respectively. SINR selected for simulation is from -5 to 15 dB for SUs. It transmits using BPSK or QAM modulation. The learning rate of $\alpha = 0.1$ and discounting factor of $\Upsilon=0.4$ is assumed.

Table 4.1 Parameters used in the simulation

PARAMETER	DESCRIPTION	VALUE
λ_p	Arrival Rate of PU	0.01 - 0.08 (arrivals/ms)
λ_s	Arrival Rate of Secondary Users	0. 02 (arrivals/ms)
μ_p	Service rate of Primary Users	0.6 (ms/arrival)
μ_s	Service rate of Secondary Users	0.5 (ms/arrival)
t_s	Switching Time	1 ms

3.1 Analysis of Handoff policy

The hybrid handoff decision efficiently allows to make the decision whether to stay on the same or shift to another channel based on the data delivery time of SUs. It is the total time taken by SU to finish its data transmission. Figure 3.1 compares this time of SH schemes as an activity of arrival rate of PU. When the arrival rate of PUs crosses the threshold value this method shifts to reactive mode it is giving almost 9.4% better than the other one. On converse, this method allows the SU to shift to proactive mode which yields 3.35% improved outcome than the other. Optimal value of arrival rate of primary user is found to be 0.05 (arrival/ms). Thus this handoff scheme allows a SU intelligently shifts between the two SH modes.

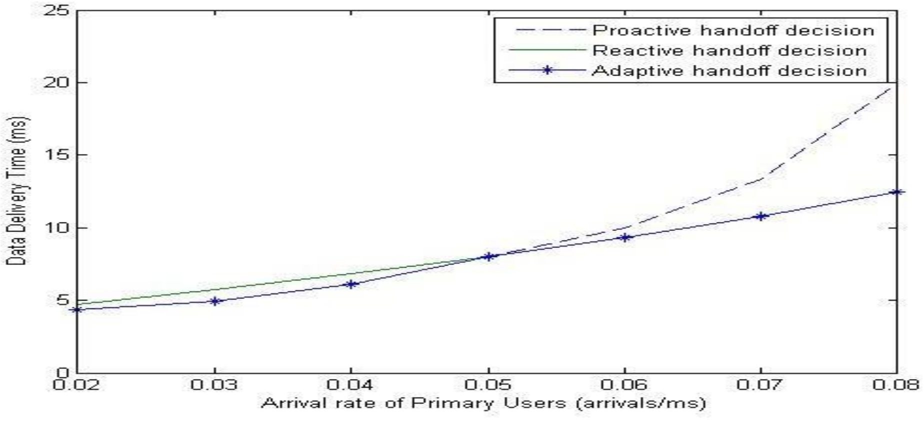
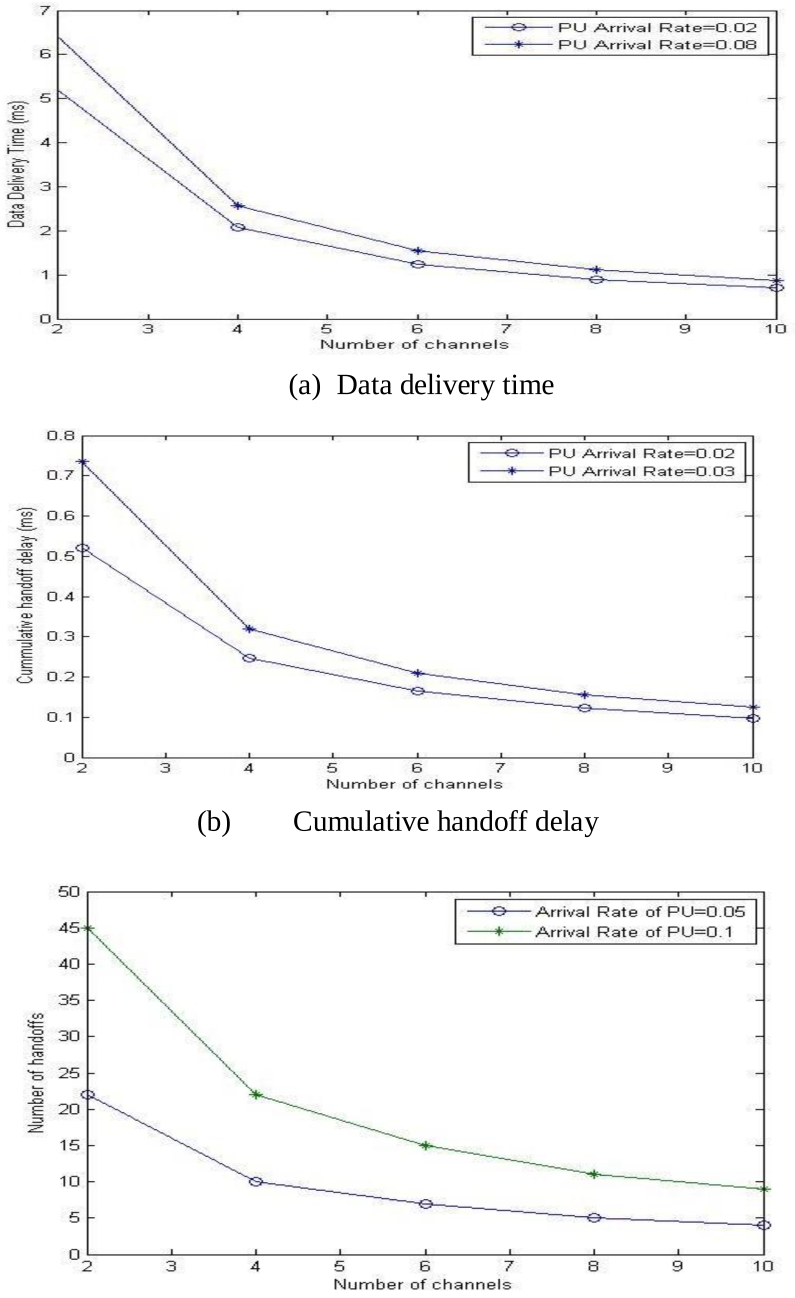


Figure 3.1 Adaptive spectrum handoff strategy

In figure 3.2 the execution of hybrid SH is investigated through varying the number of channels and SUs. Parameters from table 1 are used for simulation. The delivery time, cumulative handoff delay and the number of SHs decrease with the rising figure of channels is evident from figure 3.2 (a), (b) and (c) respectively. It is acceptable that with a growing figure of channels, the SUs encompass additional good time to fit the spectrum. The number of handoffs can be reduced by varying the arrival rate of PU.



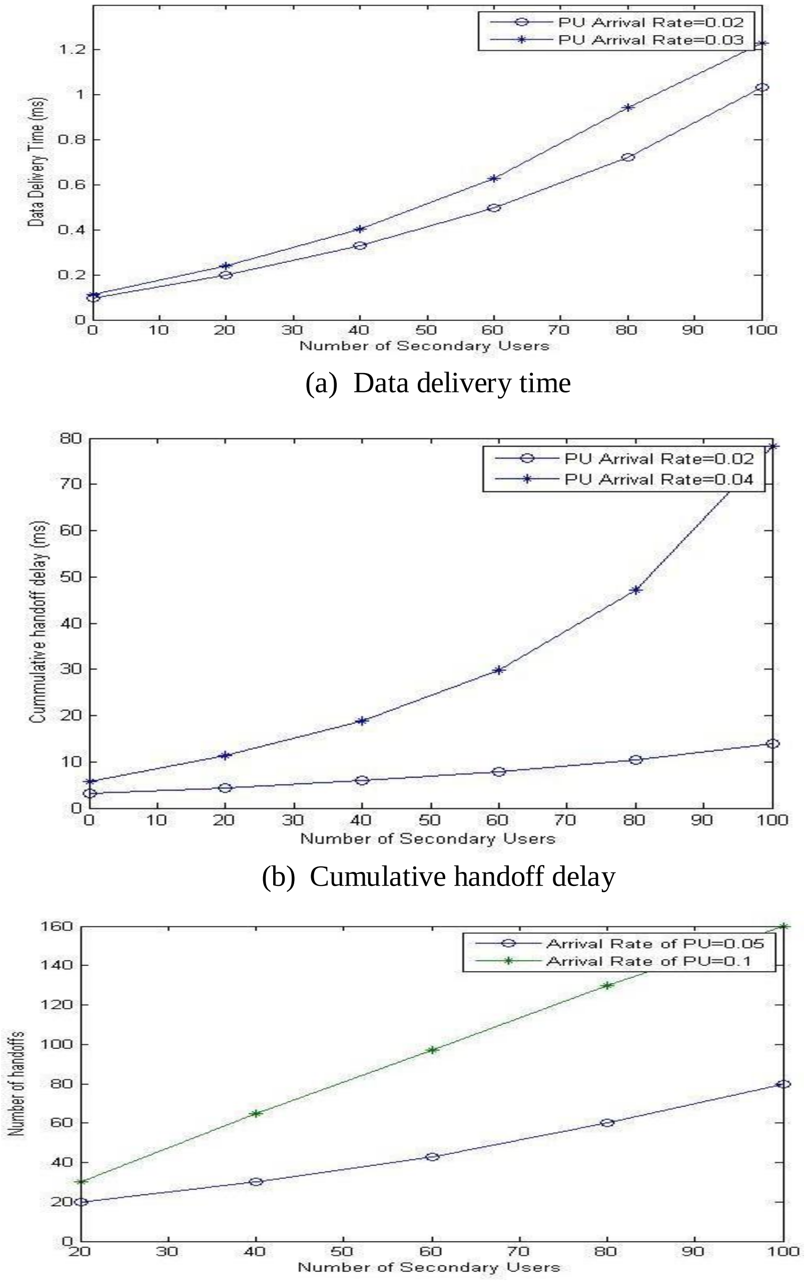
(a) Data delivery time

(b) Cumulative handoff delay

(c) Number of handoffs

Figure 3.2 Assessment of handoff in terms of the number of channels

With a rising figure of SUs the data delivery time, cumulative handoff delay along with the number of handoffs sufficiently increase which can be seen from figure 3.3 (a), (b) and (c) respectively. Due to heavy competition in channel access the number of handoffs requirement is more.



(a) Data delivery time

(b) Cumulative handoff delay

(c) Number of handoffs

Figure 3.3 Performance analyses with reference to the no. of SUs

3.2 MOS in secondary networks

From figure 3.4 MOS decreases as the amount of SUs increases is visible. When figure of users increases each SU tries to converge to low SINR value it further reduces average MOS. When we consider 26 users in the scenario, our algorithm obtains MOS>3.5. All the systems considered here have almost same MOS. By introducing docition among CR users with different traffic also maintains MOS of 3.6.

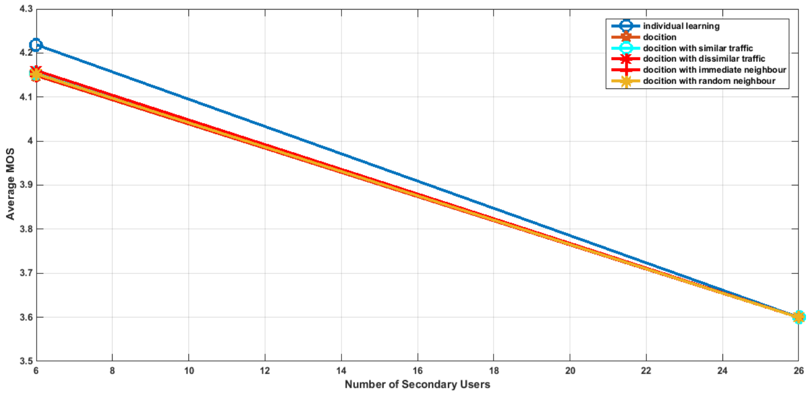


Figure 3.4 Average MOS Vs SUs

3.3 Performance review of resource allotment in terms of number of SUs

Figure 3.5 manifests the competence of employing docitive standard in altering knowledge of nearby environment to the fresh comer by skilled peers. It lowers the quantity of iterations wanted en route for reaching convergence. Compared to individual learning algorithm, cooperative learning reduces the iterations hence improving the performance. So the convergence can be reduced by using RL algorithm effectively even large numbers of SUs are considered.

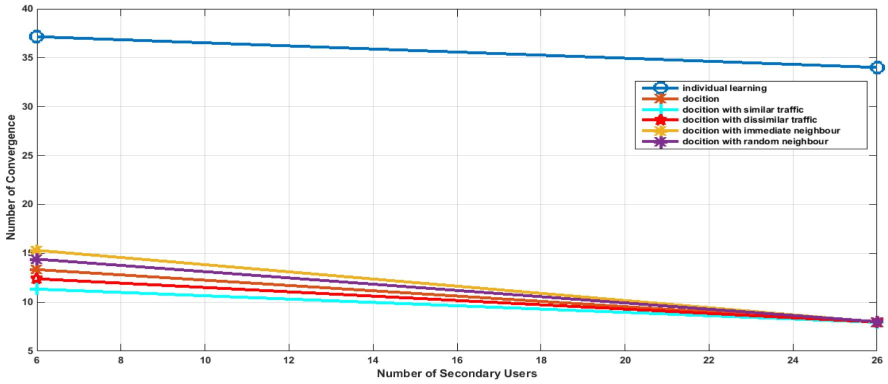


Figure 3.5 Iterations Vs SUs

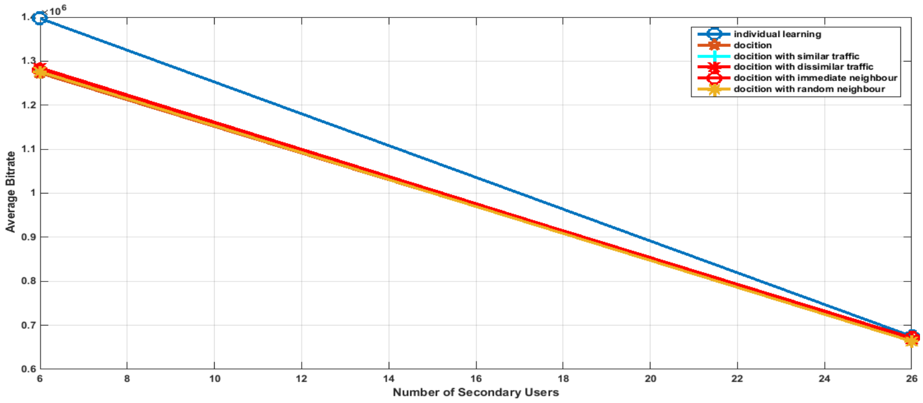


Figure 3.6 Average bitrate Vs SUs

From figure 3.6 it is evident that by satisfying interference conditions to PU communication, the average bit rate is maximized over all loads of unlike attributes (video and data) in the CR system considerably. Bit rate is reduced when cooperation is introduced due to congestion of SUs.MOS is the parameter used to evaluate the QoE of the end user requirement of 5G cognitive system. We look at diverse docition frameworks wherever a newly tied SU is trained by diverse classes of SUs with same and different loads and examine its influence on the overall QoE.

4 CONCLUSION

In this work we consider the secondary user admission method in the cognitive radio system by make use of feedback from PU's for efficient sensing mechanism. We studied the system with SU has a channel status indicator feedback. Real time PUs arrival rate obtained by SUs are not steady due to time varying channel conditions. This admission mechanism is designed as MDP and is decode using Q learning based approach. In comparison with existing literature it is obvious that this method requires less information about the PU actions to learn about the system. In existing works, QoE is not achieved properly due to the random behavior of PUs and the convergence performances reduce. Our proposed method does spectrum handoff effectively by learning channel status with the help of RL method. Besides we put in docition scheme it permits new SUs gain knowledge from their experienced peers to develop the learning procedure which reduces its convergence time. This work can be further extended by applying multi agent reinforcement learning on it.

REFERENCES

[1] M. Sherman et al. “IEEE Standards Supporting Cognitive Radio and Networks, Dynamic Spectrum Access, and Coexistence”. In: IEEE Communications Magazine 46.7 (2008), pp. 72–79.

[2] Ridhima and Avtar Buttar. “Fundamental Operations of Cognitive Radio: A Survey”. In: Feb. 2019, pp. 1–5. DOI: 10.1109/ICECCT. 2019.8869190.

[3] I. Christian, M. Sangman, I. Chung, and J. Lee, “Spectrum Mobility in Cognitive Radio Networks,” in IEEE Commun. Mag., vol. 50, Jun. 2012, pp. 114–121.

[4] Yang, W., Zu, Y. and Hou, B. (2016) ‘The Research of Reactive Spectrum Handoff Algorithm Based on Spectrum Prediction’, International Journal on Wireless Communications, Vol.4, pp.573-585.

[5] Xing, X., et al., (2013) ‘Spectrum prediction in cognitive radio networks’, IEEE Wireless Communications, Vol.20, pp.90-96.

[6] Sung Jang et al. “Reinforcement learning-based dynamic band and channel selection in cognitive radio ad-hoc networks”. In: EURASIP Journal on Wireless Communications and Networking 2019 (Dec. 2019).DOI: 10.1186/s13638-019-1433-1.

[7] Anita Garhwal and Partha Pratim Bhattacharya. “A survey on dynamic spectrum access techniques for cognitive radio”. In: arXiv preprint arXiv:1201.1964 (2012).

[8] S. A. Attalla et al. “Soft-Sensing CQI Feedback-Based Access Scheme in Cognitive Radio Networks”. In: IEEE Transactions on Cognitive Communications and Networking 4.3 (2018), pp. 486–499.

[9] A. M. Arafa et al. “A feedback-soft sensing-based access scheme for cognitive radio networks”. In: IEEE Transactions on Wireless Communications 12.7 (2013), pp. 3226–3237.

[10] Y. Song and J. Xie, “ProSpect: A Proactive Spectrum Handoff Framework for Cognitive Radio Ad Hoc Networks without Common Control Channel,” in IEEE Trans. Mobile Computing, vol. 11, Jul. 2012, pp. 1127–1139.

[11] Ibrahim Maina. “Channel Quality Indicator Feedback in Long Term Evolution (LTE) System”. In: IOSR Journal of Electronics and Communication Engineering 9 (Jan. 2014), pp. 14–19. DOI: 10.9790/2834- 09241419.

[12] Tam Nguyen, Frederic Villain, and Yann Guillou. “Cognitive Radio RF: Overview and Challenges”. In: VLSI Design 2012 (Feb. 2012). DOI:10.1155/2012/716476.

[13] R. H. Puspita et al. “Reinforcement Learning Based 5G Enabled Cognitive Radio Networks”. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC). 2019, pp. 555– 558.

[14] L. Gavrilovska et al. “Learning and Reasoning in Cognitive Radio Networks”. In: IEEE Communications Surveys Tutorials 15.4 (2013), pp. 1761– 1777.

[15] I.F. Akyildiz et al. “NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey”. In: Computer Networks 50.13 (2006), pp. 2127–2159.

[16] K. Eswaran, M. Gastpar, and K. Ramchandran. “Bits through ARQs: Spectrum Sharing with a Primary Packet System”. In: 2007 IEEE International Symposium on Information Theory. 2007, pp. 2171–2175. DOI: 10.1109/ISIT.2007.4557542.

[17] S. A. Attalla et al. “Soft-Sensing CQI Feedback-Based Access Scheme in Cognitive Radio Networks”. In: IEEE Transactions on Cognitive Communications and Networking 4.3 (2018), pp. 486–499. DOI: 10.1109/ TCCN.2018.2826539.

[18] Alex M. Andrew. “Reinforcement Learning: An Introduction by Richard S. Sutton and Andrew G. Barto, Adaptive Computation and Machine Learning series, MIT Press (Bradford Book), Cambridge, Mass., 1998, pp. 229–235.

[19] Yuxi Li. “Deep Reinforcement Learning: An Overview”. In: ArXiv abs/1701.07274 (2017).

[20] Junta Wu and Huiyun Li. “Deep Ensemble Reinforcement Learning with Multiple Deep Deterministic Policy Gradient Algorithm”. In: Mathematical Problems in Engineering 2020 (2020), pp. 1–12.

[21] IEEE Std 802.22-2011, Wireless Regional Area Network (WRAN) Specific Requirements Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Bands, in IEEE, Jul. 2011.

[22] Shaked Zychlinski. The Complete Reinforcement Learning Dictionary. https://towardsdatascience.com/the-complete-reinforcementlearning-dictionary-e16230b7d24e. Feb. 2019.

[23] Nadine Abbas and Karim Ahmad. "Recent advances on artificial intelligence and learning techniques in cognitive radio networks". In: Eurasip Journal on Wireless Communications and Networking 2015 (Dec.2015). DOI: 10.1186/s13638-015-0381-7.

[24] Faris B. Mismar and Brian L. Evans. "Deep Q-Learning for Self-Organizing Networks Fault Management and Radio Performance Improvement". In: CoRR abs/1707.02329 (2017). arXiv: 1707 . 02329. URL: http : //arxiv.org/abs/1707.02329.

[25] Neil Hosey et al. "Q-learning for cognitive radios". In: Proceedings of the China-Ireland Information and Communications Technology Conference (CIICT 2009). ISBN 9780901519672. National University of Ireland Maynooth. 2009.

[26] Avirup Das et al. "Q-Learning Based Co-Operative Spectrum Mobility in Cognitive Radio Networks". In: Oct. 2017, pp. 502–505. DOI: 10.1109/LCN.2017.80.

[27] Hao Jiang et al. "An Improved Sarsa() Reinforcement Learning Algorithm for Wireless Communication Systems". In: IEEE Access PP (Aug. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2019.2935255.

[28] J. Liu et al. "DeepNap: Data-Driven Base Station Sleeping Operations Through Deep Reinforcement Learning". In: IEEE Internet of Things Journal 5.6 (2018), pp. 4273–4282. ISSN: 2372-2541. DOI: 10.1109/ JIOT.2018.2846694.

[29] D. Bertsekas and R. Gallager, Data Networks. Upper Saddle River, NJ: Prentice Hall Inc., 1987.

[30] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of Experience- Driven Adaptation Scheme for Video Applications over Wireless Networks," in *IET Communications*, vol. 4, Jul. 2010, pp. 1337–1347.

[31] M. L. Puterman, *Markov Decision Process: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, 1994.

[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press, 1998.

[33] C. J. C. H. Watkins and P. Dayan, "Q-learning," in *Machine Learning*, vol. 8, May 1992, pp. 279–292.

[34] Q. Zhao, D. Grace, and T. Clarke, “Transfer learning and cooperation management: balancing the quality of service and information ex-change overhead in cognitive radio networks,” *Transactions on Emerging Telecommunications Technologies*, vol. 26, no. 2, pp. 290–301, 2015.