



Online Opportunistic Routing in Cognitive Radio Ad-Hoc Network and Spectrum Management

Harshal Solao¹, Prof. R M Goudar², Prof. Sunita Barve³

^{1,2,3} Department of Computer Engineering,
MIT Academy of Engineering, Alandi,
University of Pune, Pune, Maharashtra, INDIA.

Abstract: Accessing under-utilized licensed spectrum band using cognitive radio ad-hoc network improves the spectrum usage effectiveness. Due to use of opportunistic spectrum access the performance of routing in cognitive radio ad-hoc network will improve where no need of prior knowledge about topology and channel statistics in Ad-Hoc Cognitive radio Network. To deal with requirements of Ad-Hoc Cognitive radio Network, this paper is proposing opportunistic routing algorithm with the help of Reinforcement Learning. This proposed routing scheme collectively deals with, prior time channel selection as well as a link and relay selection on the basis of overall success probabilities of data transmitting. This advance learning scheme successfully finds out the opportunities in the dynamic environment of Ad-Hoc Cognitive radio Network.

Keywords: Cognitive Radio Ad-hoc Network; Spectrum Management; Hidden Markov Model; Routing Scheme; Reinforcement learning

I. Introduction

Research on cognitive radio networks (CRNs) was effectively launched in 2002 following a report by the FCC [1]. The report challenged for the first time the common belief of spectrum scarcity by pointing out that only less than 10 percent of available spectrum is used. The FCC also spot lit that a large portion of the licensed spectrum is used periodically, and geographical fluctuations in the usage of licensed spectrum portions oscillate from 15 to 85 percent with a high variance in time. To exploit under use portions of the spectrum, i.e. White spaces /spectrum holes, that prompts the need of smart reprogrammable radios which are able to interference sensing, channel state learning, and spectrum access dynamically. Recently the research community has started working on Ad-hoc cognitive radio network to unleash the potential of it in exploring the wide range of pervasive applications [2], [3].

Cognitive Radio (CR) [4] is a novel solution to the scarce spectrum resource problem. In cognitive radio network, cognitive devices or nodes are reprogrammable devices that sense the wideband spectrum and as per the environment, it re-configurable to work in a dynamic environment. The CR nodes sense spectrum and get unoccupied frequency bands also known as Spectrum Opportunities (SOP), then use specific policy or also known as rule to select one candidate for them.

Ultimately the intension of Cognitive Radio (Cognitive device/node) is to access the unused licensed spectrum bands in a dynamic environment to improve the spectrum usage [6]. Spectrum scarcity and underutilization can be improved by allowing SU's to access the licensed bands opportunistically without disturbing the PU's activity. The Cognitive device has additional functionality for spectrum sensing and sharing licensed band so use of Cognitive devices/nodes in an ad - hoc network will improve the bandwidth utilization under the dynamic environment.

II. Related Work

Routing in ad-hoc Cognitive Radio Networks become challenging due to Dynamic environment. There are various approaches for routing depending on various schemes and Parameters. These approaches are studied and used in various ways worldwide. Conventional routing mechanisms where network attempts to find out fixed path to forward packets from source to destination [7], [8]. There is large number of routing protocols proposed for ad hoc cognitive radio network that finds a fixed path using route discovery processes [9]-[11]. Such fixed path strategies fail to address the uncertainties and dynamic changes in topology due to sporadic spectrum and route availability in network.

Allocation of the barely spectrum resource between cognitive nodes in stochastic environment increases the complexity due to uncertainty of the environment state [12]. Markov Decision Process (MDP) is promising mathematical framework for modeling environments like CRNs in which outcome is not only dependent on decision strategies but sometimes it can be random. MDP are useful for studying optimization problems solved via reinforcement learning [13-15]. Transition probability function and reward function of it can used to analyze the environment based on received feedback of any selected action.

There are many existing solutions for spectrum aware routing using conventional algorithms for ad-hoc cognitive radio networks were proposed. Channel selection as well as assignment and routing in semi static ad-

hoc networks is jointly addressed using a layered graph [5]. The layered graph plays layers equal to the number of available channels. Dynamic changes in the topology are captured using the link availability predictions based on the interference to the PU [16]. The resultant reliable topology is used for setting up multi-hop paths and to reduce rerouting. Cognitive Routing Protocol (CRP) gives explicit protection to the PU receiver [17]. It allows two classes of routes based on the service differentiation in cognitive radio network. Routing is achieved in two phases i.e. spectrum selection phase and next hop selection phase. Online opportunistic routing selects forwarding links based on the locally identified spectrum access opportunities [18].

This paper proposes the Online Opportunistic Routing Scheme for ad-hoc cognitive radio network. This Learning based online opportunistic routing scheme addresses the issue of routing packet in dynamic environment. Routing decisions are made with respect to the current state of the environment and actual transmission success probabilities. Proposed algorithm considers no knowledge of the topology and simultaneously learns channel statistics and good routes. The main contributions of this paper are as follows:

1. Channel availability prediction with the help of Hidden Markov Model
2. Markov Decision Process formulation using Temporal Difference (TD) implementation of Reinforcement Learning with incomplete and erroneous information of topology.
3. Softmax Action Selection Rule is used for balancing exploration and exploitation in relay selection process.

III. Hidden Markov Model

We see a cognitive radio network in which each node works as a cognitive node that can scan and sense multiple channels periodically to compute interference temperature at each channel. This process gives us an observation sequence O for a first T time slot for considering channel c . If a node can predict the observation sequence for next $2T$ time slot for same channel c then such prediction will be utilized by the node for selecting preferable communication channel prior periods. In order to resolve the above situation the Hidden Markov Model provides a suitable solution to the problem. The HMM has obtained noteworthy practical application in speech recognition [10] and bioinformatics [15]. The HMM is trained by using the observation sequence O also known as training observation sequence. In order to achieve the HMM, it is essential to define the followings:

1. The number of observed symbols, M
2. The number of states, N
3. The observation sequence $O = \{O_1, O_2, \dots, O_T\}$
4. The state transition probabilities, A_{ij} subject to condition $A_{ij} \geq 0$ and $\sum_{j=1}^N A_{ij} = 1$
5. The symbol observation probabilities, $B_j(m)$ subject to condition $B_j(m) \geq 0$ and $\sum_{m=1}^M B_j(m) = 1, 1 \leq j \leq N$
6. Initial State probabilities π

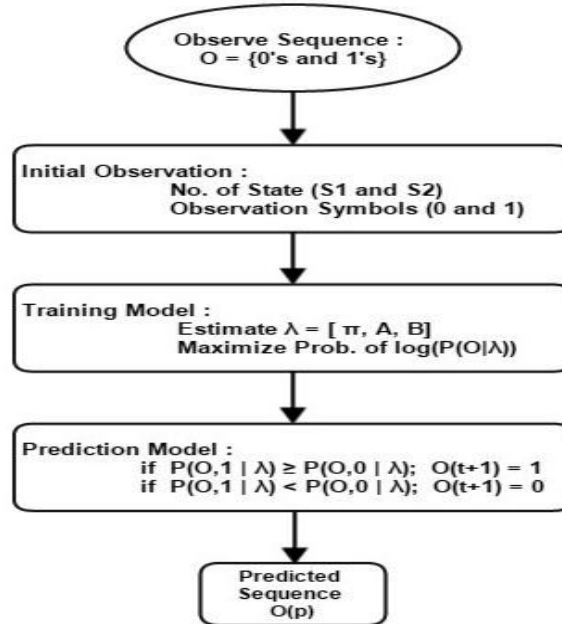


Figure 1: HMM training and Predictor

Here the observation sequence contains two symbols 0 and 1s also known as observation symbols $M = \{0, 1\}$. And set of states can be denoted as $S = \{s1, s2\}$.

The HMM can be denoted by the notation $\lambda = [\pi, A, B]$ that is estimated by iterative method of the Baum-Welch Algorithm. After training model the HMM predictor will predict the required sequence for next $2T$ time slot. The HMM prediction scheme is illustrated in Figure 1.

Where, the channel status is denoted by 1 and 0 for busy and ideal, respectively. The main goal of predictor is to predict the sequence $O(p)$ on the basis of past T slot observation. For that the HMM should be able to generate the observation sequence $O = \{O_1, O_2, \dots, O_T\}$ with maximum likelihood probability. Hence, the parameter $\lambda = [\pi, A, B]$ are used to maximize the likelihood probability of generating the observation sequence, i.e., maximize the probability $P(O|\lambda)$. After completing the training, the joint probability of observing the sequence O followed by a busy slot or an idle slot at instant $2T$ is predicted.

IV. Online Opportunistic Routing

The performance of online opportunistic routing in uncertain and dynamic wireless environment can be improved using Reinforcement Learning (RL) algorithm. RL can be used to learn optimal policy on-line from direct interaction with the environment for solving non-linear control problems [19].

A. Methodology

Methodology Using RL the agent learns how to achieve the given goal by trial and error interacting with environment. That improves performance of the agent from the received reward or punishment as a result of the performed action [20]–[22]. In online opportunistic routing, every cognitive node is considered as an intelligent agent and Ad-hoc network as environment. Temporal difference (TD) method learns directly from environment dynamics [23] without any knowledge of its model. It uses generalized policy iteration to update estimates i.e. value score, based in the parts on other learned estimates, without waiting for the final outcome. TD method wait only until the next time step, at $t+1$ it immediately form a target and make a useful update using observed reward r_{t+1} and the estimate $V(s_{t+1})$. So estimated value score $V(s_t)$ is updated by

$$V(s_t) \leftarrow V(s_t) + \alpha_{BRN}[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (1)$$

Where α_{BRN} , is a step-size parameter that influenced by BRN i.e. Best relay node among willing neighbor. Markov decision process the task that satisfies the markov Property, i.e. all decisions and values are functions of the current state only, is called Markov Decision Process (MDP) [24]. MDP is defined by its state and action sets and by the one-step dynamics of the environment. Given the current state and action, one-step dynamics enable us to predict the next state, expected next reward and iteratively all future states and rewards.

B. Proposed Work

One potential application of HMM that we investigate and propose in this paper is to use the sequences generated by the trained HMM in selecting preferable channel for communication in multi-channel cognitive wireless network. Consider a multi-channel cognitive wireless network with n channels. Let us assume that the training sequences for these channels is obtained by a designated cognitive nodes, which construct hidden Markov model for each channel using these training sequences, as explained in section III.

$$CAM_i = A_i^0 + \frac{1}{(OS^1/|OS|)} \quad (2)$$

Where, A_i^1 is average gap between any two 1's in given sequence, OS^1 is number of 1's in given sequence and $|OS|$ is length of sequence.

The node selects the channel having most eminent value of channel availability metric (CAM_i), as the most preferable channel for communication. Now each node has their own CAM value for each available channel in there vicinity. All the nodes in the network send beacon message to neighbor nodes with available channel CAM values because of this the each node got the neighbors channel availability set. Then each node calculates the common channel with neighbor using intersection.

Then sender node choose most available channel as a control channel for sending control packets and other as a data channel for sending data packets. After finding common channel between neighboring nodes the sender node sends the RREQ (relay request) to neighbor nodes on every available control channel. Now some of them those who are interested to be part of routing will send the RREP (relay reply) on control channel also known as willing neighbor WN_i to sender as shown in figure 2

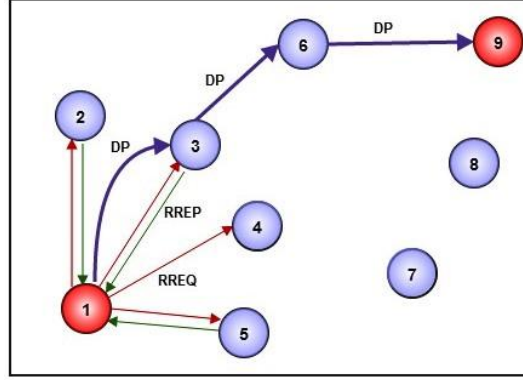


Figure 2: Relay Request (RREQ), Relay Reply (RREP) and Data Packet (DP) Transmission

Then sender node choose best channel as a control channel and other as a data channel. Then it will send data to the relay node then relay node send acknowledgment to sender on that basis sender update the estimated score for that relay node for future decision. Whenever in future the node wants choose the relay node for data transmission first it checks history record for available neighbor nodes score and as per statistics node will take decision quickly using Softmax action selection explained in section V.

The above process will be repeated by each relay node until the destination. After reaching the destination successfully it will get some rewards stored to source node for future decision. After reception of the packet at destination the cumulative link cost LC_c is subtracted from the fixed final reward received at the destination. Where,

$$LC_c = TT + LA + SC + CA + PD \quad (3)$$

Where, TT is time to transmit (estimated) , LA is link availability , SC is switching cost , CA is channel availability and PD is propagation distance. Selection of the lowest link cost at every hop finally increases reward at the destination.

So average per packet reward is maximized by selecting random hops h for the packets generated at the source s with index I for destination d is presented as,

$$RM = \frac{(\sum_{i=1}^M (R - \sum_{j=1}^h LC_c))}{M} \quad (4)$$

Where M is number of packet transmitted successfully, R is fixed reward received at destination.

V. Balancing Exploitation and Exploration

For increasing learning rate and to increase overall score exploration and exploitation must be balanced at time of taking action. Selection of relay with the help of routing score can increase complexity and load on that single node that also known as greedy selection in such case if node is choose with the help of softmax action selection rule i.e. one of the non-greedy actions. This alters to explore, as it allows improving the values of non-greedy actions. Softmax action selection rule is effective in exploring and exploiting the network opportunities by taking a sequence of decisions which will increase routing score of every willing neighbor WN. Suppose estimated value of action A at t^{th} play,

$$Q_t(A) = WN + 1 \quad (5)$$

Softmax action selection method is instinct in balancing exploration and exploitation in online dynamic routing. It can be represented as,

$$P = \frac{e^{Q_t(A)/T}}{\sum_{A=1}^n e^{Q_t(A)/T}} \quad (6)$$

Where, T is a positive parameter called temperature. Because of high temperature equi-probable action selection is done. This enables exploration by improving estimate of non-greedy actions. Because of low temperature greater difference in selection probability for actions that differs in their value estimate. When $T = 0$, softmax action selection method will act as greedy action select exploiting the network by selecting the best candidate node as relay.

VI. Algorithm and Flow Chart

1. Initialization
 - a. HMM predict sequence using given observation sequence.
 - b. Nodes share channel availability to neighbor node (N_s)
 - c. Select best channel as Control Channel (CC) and other as Data Channel (DC)
2. Relay Request

- a. Source (**S**) send a **RREQ** on CC for destination (**D**) using Control Packet (**CP**)
3. Relay Reply
 - a. Let **N(s)** is set that receive **CP**
 - b. **RREP** sent from **WN** $\in$ **N(s)**
 - c. Update Score for **WN**
4. Relay Selection and Data Transmission
 - a. Upon reception of **RREP** from **WN** node selects routing action **a** $\in$ **A(s)** using Softmax Action Selection Rule **P**.
 - b. Sends data packet (**DP**) on most available common data channel between source to next node at time **t**
5. Reception and Acknowledgement
 - a. Upon successful reception of **DP**, an acknowledgement is sent back to sending node
 - b. Reward is updated by subtracting LC_c
6. Step 1 to 5 are repeated until terminal node (Destination)
 - a. After successful reception of all **DP** average per packet reward is calculate
 - b. Acknowledgement is sent back to Source with reward

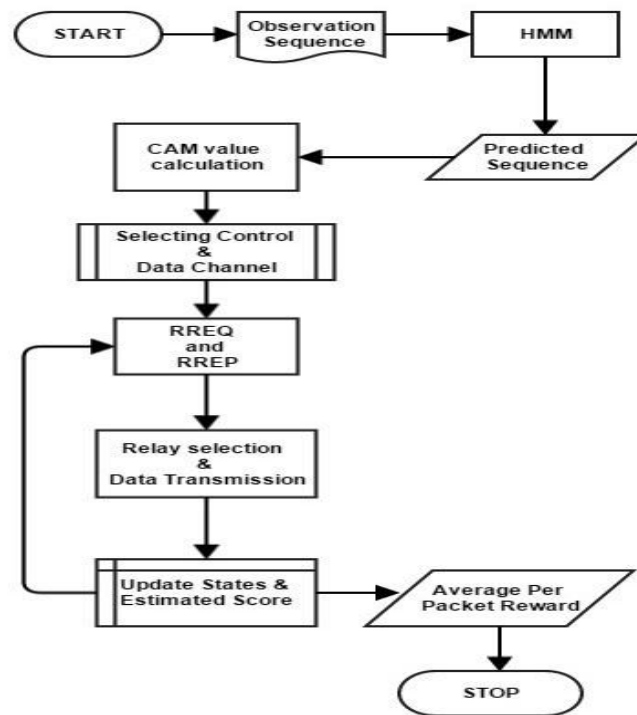


Figure 3: Flow Chart

VII. Future Work and Conclusion

1. Study and analyze the consequence of the reinforcement signal on convergence of the RL in Cognitive Radio network and effect of Q values for improving the learning efficiency of the agent.
2. Design of the routing protocol under circumstance of the multiple agents who are modeled as evolutionary game for using wireless spectrum.

In online opportunistic routing reinforcement learning based routing algorithm is designed successfully to explore the spectrum and routing opportunities in dynamic environment of Cognitive Radio Ad-hoc Network. In contrast to the conventional routing of finding out a static route from origin to destination, proposed algorithm routes the packets online without considering topology.

Temporal Difference (TD) implementation of RL based routing accomplishes an optimum performance evaluated using average per packet reward. It attains the good quality of service requirement of cognitive users by decreasing the route re-discoveries.

VIII. References

- [1] FCC, "Spectrum Policy Task Force Report," ET docket no. 02-155, Nov2002.
- [2] Sengupta S, Subbalakshmi KP (2013) Open research issues in multihop cognitive radio networks. IEEE Commun Mag 51(4):168–176.

- [3] Thomas Hou Yet al (2008) Spectrum sharing for multi-hop networking with cognitive radios. *IEEE J Sel Areas Commun* 26(1):146–155
- [4] J. Mitola III, “Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio,” Doctoral Dissertation, KTH, Stockholm, Sweden, May 2000.
- [5] C. Xin, B. Xie, and Shen, “A Novel Layered Graph Model for Topology Formation and Routing in Dynamic Spectrum Access Networks,” 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2005), pp. 308–317, November 2005, Baltimore, USA.
- [6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, March 1998.
- [7] L. P. Kaelbling, M. L. Littman, and A. P. Moore, “Reinforcement learning: A survey,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.
- [8] Royer EM, Toh CK (1999) A review of current routing protocols for ad hoc mobile wireless networks. *IEEE pers Commun* 6(2):46–55
- [9] How KC et al (2011) Routing and QoS provisioning in cognitive radio networks. *Elsevier Int J Comput Netw* 55(1):330–342
- [10] Cacciapuoti AS et al (2012) Reactive routing for mobile cognitive radio ad hoc networks. *Elsevier J Ad Hoc Netw* 10:803–815
- [11] Ian F. Akyildiz, Won-Yeon Lee, Mehmet C. Vuran and Shantidev Mohanty, Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey, *International Journal on Computer Networks* (Elsevier), Vol. 50, 2006, pp. 2127-2159.
- [12] Bertuccelli, L.F.; Wu, A.; How, J.P.; , “Robust Adaptive Markov Decision Processes: Planning with Model Uncertainty,” *Control Systems, IEEE* , vol.32, no.5, pp.96-109, Oct. 2012
- [13] Wipawee U., Javier B. , Reinforcement Learning for Resource Allocation in LEO Satellite Networks, *IEEE Transaction on Systems, Man and Cybernetics-Part B: Cybernetics*, Vol.37, No. 3, pp. 515-527, June 2007.
- [14] Dmitri A. D., Edmund H. D., Resource Allocation Among Agents with MDP-Induced Preferences, *Journal of Artificial Intelligence Research*, 27,505–549, 2006.
- [15] David V., A reinforcement learning approach to dynamic resource allocation, *Engineering Applications of Artificial Intelligence*, Volume 20, Issue 3, Pages 383-390, April 2007.
- [16] Guan Q et al (2010) Prediction-based topology control and routing in cognitive radio mobile ad hoc networks. *IEEE Trans Veh Technol* 59(9):4443–4452
- [17] Chowdhury KR, Akyildiz IF (2011) CRP: a routing protocol for cognitive radio ad hoc networks. *IEEE J Sel Areas Commun* 29(4): 794–804
- [18] Bhorkar AA et al (2012) Adaptive opportunistic routing for wireless ad hoc networks. *IEEE/ACM Trans Netw* 20(1):243–256
- [19] Wei-bing LIU, Xian-jia WANG (2009) Dynamic decision model in evolutionary games based on reinforcement learning. *Int J Syst Eng- Theory Pract* 2(3):28–33
- [20] Duana Yet al (2007) Application of reinforcement learning in robot soccer. *Eng Appl Artif Intell* 20:936–950
- [21] Usaha W, Barria JA (2007) Reinforcement learning for resource allocation in LEO satellite networks. *IEEE Trans Syst, Man Cybern-Part B: Cybern* 37(3):515–527
- [22] Nie J, Haykin S (1999) A dynamic channel assignment policy through Q-learning. *IEEE Trans Neural Netw* 10(6):1443–1455
- [23] Kaelbling LP et al (1996) Reinforcement learning: a survey. *J Artif Intell Res* 4:237–285
- [24] Sharma R, Gopal M (2010) Synergizing reinforcement learning and game theory-a new direction for control. *Elsevier Int J Application Soft Computer* 10:675–688