ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

Efficient Exploration for Reinforcement Learning Based Distributed Spectrum Sharing in Cognitive Radio System

U. Kiran[1], D. Praveen Kumar[2], K. Rajesh Reddy[3], M. Ranjith[4]

Assistant Professor, Dept. of ECE, Talla Padmavathi College of Engineering, Warangal, India[1,2,4]

Assistant Professor, Dept. of ECE, K.U. College of Engg. Warangal, India[3]

ABSTRACT: In this paper, we investigate how distributed reinforcement learning-based resource assignment algorithms can be used to improve the performance of a cognitive radio system. Today's decision making in most wireless systems include cognitive radio systems in development, depends purely on instantaneous measurement. Two system architectures have been investigated in this paper. A point-to-point architecture is examined first in an open spectrum scenario. Then, the distributed reinforcement learning-based algorithms are developed by modifying the traditional reinforcement learning model in order to be applied to a fully distributed cognitive radio system.

Keywords: cognitive radio, resource assignment, spectrum sensing, point-to-point architecture, distributed reinforcement learning.

I. INTRODUCTION

The assignment of spectrum to transmissions and to users is a fundamental issue of wireless communications. Numerous channel assignment methods have been proposed for sharing the limited physical resource. The traditional licensed spectrum allocation strategies employed by radio regulatory bodies is very restrictive and extremely inflexible, resulting in highly underutilized spectrum usage. A fully dynamic spectrum access technique called Cognitive Radio which was first introduced in [1, 2], has been considered as a potential way to improve the inefficient spectrum utilization. The inefficient usage of the existing spectrum can be improved through opportunistic access to the licensed bands without interfering with the existing users. The definition of cognitive radio suggested by ITU-R [3] is: 'a radio system employing a technology, which makes it possible to obtain knowledge of its operational environment, policies and internal state, to dynamically adjust its parameters and protocols according to the knowledge obtained and to learn from the results obtained'. The fundamental objective of cognitive radio is to enable an efficient utilization of the wireless spectrum through a highly reliable approach. Although a cognitive radio may be able to analyze the physical environment before it sets up a communication link, the best system performance is unlikely to be achieved by either a random spectrum sensing strategy or a fixed spectrum sensing policy.

Reinforcement learning (RL), a sub-area of machine learning, uses a mathematical way to evaluate the success level of actions [4, 5]. Its emphasis on individual learning from the direct interactions with the environment makes it perfectly suited to distributed cognitive radio scenarios. There are mainly two reasons to consider the reinforcement learning as the most suitable learning approach for cognitive radio systems. The first reason: Reinforcement learning is an individual learning approach where the learning agent learns only on local observations and the second is: Reinforcement learning learns on a trial-and-error basis that no environment model is required. This is also perfectly suited to cognitive radio systems which constantly interact with an 'unknown' radio environment on a trial-and-error basis.

This paper introduces the reinforcement learning-based distributed spectrum sharing (RL-DSS) scheme which enables efficient usage of spectrum by exploiting users past experience. In the proposed spectrum sharing scheme, a reward value is assigned to a used resource based on the reward function. Cognitive radio users select spectrum resources to use based on the weight values assigned to the spectral resources - resources with higher weights are considered higher priority. Furthermore we investigate and compare the system performance of different sets of reward values which effectively are the weighting factors in the reward function. In fact, we will show how different weighting

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

factor values have significant impact on the system performance, and that inappropriate weighting factor setting may cause some specific problems.

The reminder this paper is organized as follows. The cognitive radio based reinforcement learning model will be presented in section II. Reinforcement learning-based distributed spectrum sharing algorithm is described in section III. Section IV presents the key measurements for evaluating the system, Section V presents the simulation results to validate the analysis, and Section VI concludes the paper.

II. SYSTEM MODEL

The reinforcement learning model developed for the cognitive radio scenario is illustrated in Figure 1. The wireless spectrum is effectively the environment in which cognitive radio (CR) is the learning agent. The way we implement reinforcement learning in the CR scenario is slightly different from the original reinforcement learning model. This is caused by a few built-in features of cognitive radio. In the original reinforcement learning system, the value of the current state *s* under a policy π which is denoted by $V^{\pi}(s)$ is the basis to choose the action *A(s)*. An optimal policy is supposed to maximize $V^{\pi}(s)$ at each trial. $V^{\pi}(s)$ is formally defined as [4]:

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s') \quad (1)$$

Where *E* is the expectation operator, γ is a discount factor $(0 < \gamma < 1)$. $R(s, \pi(s))$ is the immediate reward if the agent chooses action $a = \pi(s)$ given a state *s*. $R(s, \pi(s)) = E\{r(s, \pi(s))\}$ is the mean value of $r(s, \pi(s))$. *s'* stands for the goal states which *s* will transit to by taking the action $\pi(s)$. Given that there may be multiple successor states *s'*, $P(s'|s, \pi(s))$ defines the probability of making a transition from state *s* to different successor states.

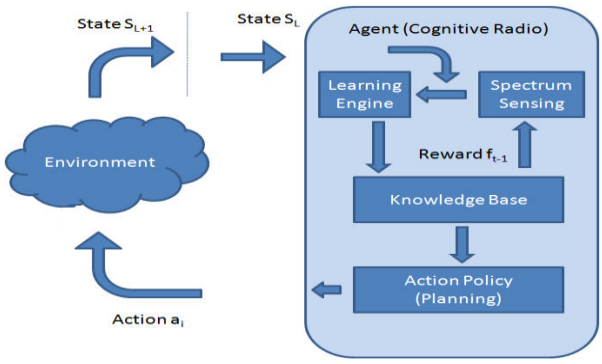


Figure 1. The Reinforcement Learning Model in Cognitive Radio Scenario

The optimal value function $V^{\pi*}(s)$ under the optimal policy π^* can be defined as:

$$V^{\pi*}(s) = \max_{a \in A} \left(R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi*}(s') \right) \quad (2)$$

Based on the optimal value function $V^{\pi*}(s')$, the optimal policy π^* is specified as:

$$\pi^*(s) = \arg \max_{a \in A} \left(R(s, \pi(s)) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi*}(s') \right) \quad (3)$$

$R(s, \pi(s))$ is effectively the cumulative reward in the state of *s*. The other part of the equation is the expected feedback of its successor states *s'*. It can be clearly seen from equation (1) to equation (3) that in order to obtain the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

optimal policy π^*, the information of *s'* is vital. Information like the number of potential successor states and the estimated value of each of the state's *s'* are essential.

Our strategy is to develop a policy π that maps memory (weight values) to action $\pi : W \rightarrow A$ instead of the original approach which maps the state of environment to action $\pi : S \rightarrow A$ [12]. On one hand, the agents are fully distributed in our strategy so that decisions are made only according to the local measurements. It is unlikely for a CR to obtain the information at the network level. Cognitive radio is able to sense the target spectrum before activation and it is not supposed to transmit data until unoccupied spectrum has been found. Choosing the most successful spectrum by reinforcement learning combined with spectrum sensing is the suggested method. A few amendments have been made to the learning model. The reinforcement learning model which we used consists of [4]: A set of memories, *W*. *W* is a set of weights of the performed actions which are stored in the knowledge base; a set of actions, *A*; A set of numerical rewards *R*.

A CR will access the communication resource according to the memory of reinforcement learning. The success level of a particular action, which is whether the target spectrum is suitable for the considered communication request, is assessed by the learning engine. Based on the assessment, a reward is assigned in order to reinforce the weight of the performed action in the knowledge base. Since the actions are all strongly connected to the target resources, the weight is practically a number which is attached to a used resource and this number reflects the successful level of the resource. Our goal is to develop an optimal policy mapping weight to action $\pi : W \rightarrow A$ that can maximize the value of the current memory $V^{\pi^*}(w)$. Given a set of available weights of used resources and a policy π, the selection of a specific action is denoted as $a = \pi(w)$. Then the optimal value function under the optimal policy π^* can be defined as:

$$V^{\pi^*}(w) = \max_{a \in A} \left(\sum_{w'} P(w' | w, \pi(w)) w' \right) \tag{4}$$

Where *w* is the weight of used resources of an agent at time *t*, *w'* is the expected values of weights after agent takes an action $\pi(w)$. $P(w' | w, \pi(w))$ is the probability of selecting an action after taking the action π^*. The optimal policy can be specified as:

$$\pi^*(w) = \arg \max_{a \in A} \left(\sum_{w'} P(w' | w, \pi(w)) w' \right) \tag{5}$$

At each communication request the agent chooses a resource which can maximize $V^*(w)$ according to its current memory. Based on the result, the learning engine updates the knowledge base by a reward *r*. The inner loop within cognitive radio in figure 1 will proceed constantly to update the knowledge base; the complexity of the communication system is reduced.

A key element of reinforcement learning is the value function [8]. A CR user updates its knowledge based on the feedback of the value function. In other words, the CR user adjusts its operation according to the function. The following linear function is used as the objective function to update the spectrum sharing strategy in this paper [6, 7]:

$$W_t = f_1 W_{t-1} + f_2 \tag{6}$$

Where W_{t-1} is the weight of a channel at time *t-1*, and W_t is the weight at time *t* according to previous weight which reflect the degree of responses of a learning agent towards the changes of environment, and the updated feedback from system. f_1 and f_2 are the weighting factors at time *t* that will take on different values depending on the localized judgment of current system states and the environment. The values of weighting factors are shown in table I. Based on the degree of success; either a reward or a punishment is assigned to the weight of the used spectrum.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

TABLE I
WEIGHTING FACTOR VALUES

Scheme	f_1		f_2	
	Reward	Punishment	Reward	Punishment
Mild Punishment	1	1	1	-1
Harsh punishment	1	0	1	0
Discounted Punishment	1	0.5	1	0

III. DISTRIBUTED REINFORCEMENT LEARNING - CR SPECTRUM SHARING SCHEME

The basic Reinforcement Learning-based distributed CR spectrum sharing algorithm is illustrated in Figure 2 [7]. We consider the CR users are a set of transmitting-receiving pairs of nodes, denoted as *U*, uniformly distributed in a square area and all the pairs $U_i \in U$ are spatially fixed. There are 3 main steps in the process:

Step 1: *Spectrum selection.* At the beginning of each activation, U_i chooses a channel according to the weights of the available resources. It starts with the channel with the highest weight, or picks up a channel randomly if all resources have same priority. The selected channel is denoted as C_k where $C_k \in C$ and C is the available channel set.

Step 2: *Spectrum sensing.* U_i senses the interference level on C_k. If the interference level *I* of C_k is below the interference threshold I_{thr}, U_i is activated. Otherwise if $I > I_{thr}$, the weight of C_k for U_i is decreased by a punishment weighting factor and U_i returns back to step 1.

Step 3: *SINR measuring*. After step 2, the existing users within the same channel can measure the Signal-to-Interference-plus-Noise Ratio (*SINR*) at their receivers. The purpose of measuring *SINR* is to maintain the communication quality of the channels. We set up a *SINR* threshold $SINR_{thr}$. If the *SINR* of the activated pair U_i is greater than the threshold $(SINR_i > SINR_{thr})$, U_i successfully uses the spectrum and the weight of the channel will be increased by a reward. If $SINR_i < SINR_{thr}$, U_i is blocked by the channel and the weight is updated with a punishment.

The CR users follow the above steps in every transmission process. One condition applies to the system that $N(U_i) < N_{\max}$, $N(U_i)$ denotes the number of sensed channels of U_i in each activation and $N_{\max}$ is the maximum number of channels which a CR user is allowed to scan in a single activation. If $N(U_i) < N_{\max}$, and U_i is still searching for an unoccupied resource, it is blocked and waits for the next activation. It is unrealistic to allow users to keep sensing and searching for a better resource without a time limit, because sensing is a power-intensive and time-consuming process.

IV. PERFORMANCE EVALUATION

In this paper we evaluated few performance parameters of the system capacity. Signal-to-Interference-plus-Noise-Ratio (SINR) is used to evaluate link quality, i.e. to determine whether the current user will lose its current service, or to determine the data rate depending on the adaptive modulation applied to the system. Blocking probability and dropping probability are normally used to evaluate link based wireless system, e.g. speech-oriented wireless service. The Cumulative Distribution Function (CDF) is used to process the initial data and to deliver the statistical behavior of the results.

1) Signal-to-Interference-plus-Noise-Ratio (SINR): Signal-to-Interference-and-Noise Ratio (SINR) [9], also known as Carrier-to Interference-and-Noise Ratio (CINR), is one of the fundamental parameters to measure the link quality of users in wireless communication. It is defined by the quotient of the average received signal power (*S* or *C*) and the average received co-channel interference power (*I*) plus the noise power from other sources (*N*). In point to point architecture the SINR has been derived:

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

$$\gamma_n = \frac{p_n g_{n,q}}{\sum_{i=1,i\neq n}^{M} p_i g_{i,q} + \sigma^2} \tag{7}$$

Where p is the transmit power of the n transmitter, g is the gain of the wireless link on channel q, is the noise power. A frequency separation of backhaul and access is assumed so that the backhaul network and the access network do not interfere with each other. Then for the backhaul network, SINR measured at ABS n (signal from HBS m in channel q and sub-channel r) can be derived as:

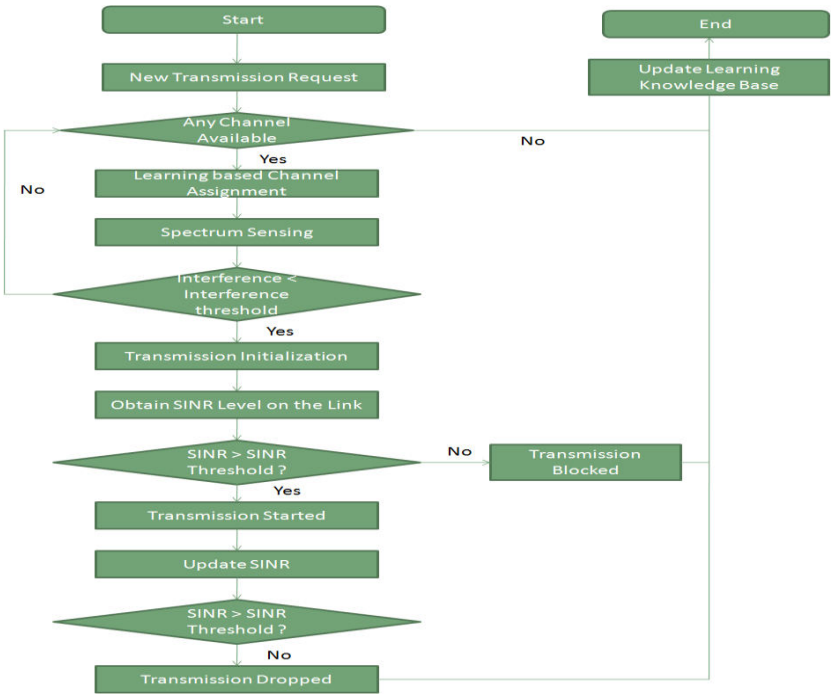


Figure 2 Reinforcement Learning-based Spectrum Sharing Algorithm

$$\gamma_{n,q,r}^{m,l} = \frac{p_m^{H,l} g_{q,r}^{B,m,l}}{\sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} p_i^{H,j} g_{q,r}^{B,i,j} + \sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{q,r}^{B,m,i} + \sigma^2} \tag{8}$$

Where $g_{q,r}^{B,m,l}$ is the gain of the wireless link from the l^{th} beam of HBS m to ABS n. $\sum_{i=1,i\neq m}^{M} \sum_{j=1}^{L} p_i^{H,j} g_{q,r}^{B,i,j}$ is the interference from other HBSs to ABS n. $\sum_{i=1,i\neq l}^{L} p_m^{H,i} g_{q,r}^{B,m,i}$ is the interference comes from other beams of HBS m, using the same channel q and sub channel r. σ^2 is the noise power. Similarly for the access network, the SINR received at MS k (signal from ABS n (associated with HBS m) in channel q and sub channel r) is:

$$\gamma_{k,q,r}^{n,m} = \frac{p_n^{A,m} g_{q,r}^{A,n,k}}{\sum_{i=1,i\neq m}^{M} \sum_{j=1}^{N} \sum_{u=1}^{K} p_j^{A,i} g_{q,r}^{A,j,u} + \sum_{i=1,i\neq n}^{N} \sum_{j=1}^{K} p_i^{A,m} g_{q,r}^{A,i,j} + \sigma^2} \tag{9}$$

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

Where $g_{q,r}^{A,n,k}$ is the link gain between ABS *n* and MS *k*. In the denominator first term is the interference from all the ABSs in other cells that are using the same frequency., and the second one is the interference from other ABSs in the same cell, and σ^2 is the noise power.

2) Cumulative Distribution Function (CDF): As we mentioned before, in order to obtain statistically accurate results we need to apply Monte Carlo simulation. However, a very large amount of unprocessed data can be expected by conducting Monte Carlo simulation. Appropriate mathematical analysis in this case is required to show the statistical behavior of the results. The cumulative distribution function is the main statistical method applied in this report. The CDF of *x* is defined as [10]:

$$CDF \equiv F(x) = \int_{-\infty}^{x} f(t)dt \tag{10}$$

where *f(x)* is the probability density function of *x*. The results of our simulation like blocking probability and dropping probability are mainly measured at regular points in the service area.

3) Blocking Probability and Dropping Probability: Blocking probability and dropping probability [11] are the measurements we use to evaluate the grade of service. The blocking probability at time *t* can be defined as:

$$P_B(t) = \frac{N_b(t)}{N_a(t)} \tag{11}$$

Where *P(t)* is the blocking probability at time *t*. $N_b(t)$ is the total number of blocked activations of the system by time *t* and $N_a(t)$ is the total number of activations of the system by time *t*. Similarly, the dropping probability is defined as follows:

$$P_D(t) = \frac{N_d(t)}{N_{sa}(t)} \tag{12}$$

Where $P_D(t)$ is the dropping probability by time *t*. $N_d(t)$ is the total number of dropped transmissions by time *t* and $N_{sa}(t)$ is the total number of accepted activations by time *t*.

V. SIMULATION RESULTS

In this paper we employed an event-based scenario and at each event a random subset of pairs are activated, system parameters used in this paper is shown in table II. The available spectrum will be partitioned autonomously by individual reinforcement learning and therefore CR users are able to avoid improper spectrum. Figure 3 (a)-(b) represent how the channel partitioning emerges during the simulation. A small number of 10 is used in this simulation to define the number of available channels and the number of users.

TABLE II
Simulation Parameters

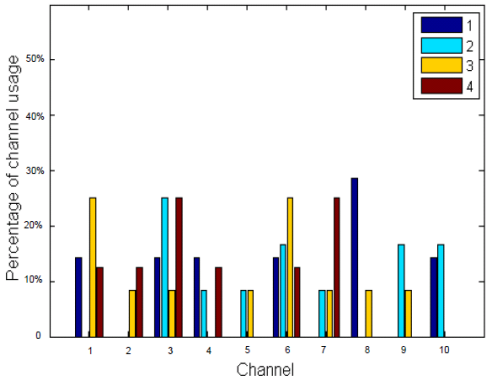
Parameter	Value
Service Ares	1000km^2
Number of pairs	1000
Maximum number of activated users	400
Link Length	200m – 1500m
Transmitter Antenna gain	0 dBi
Interference threshold	-40 dBm
SINR threshold	10dB
Noise floor	-137dBm

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

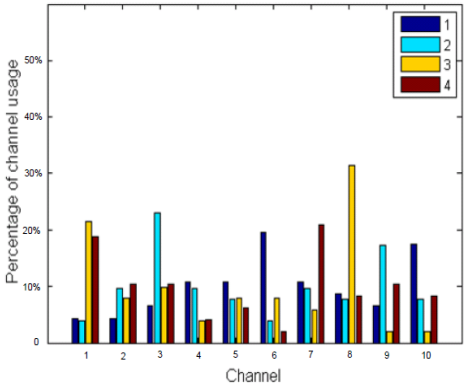
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013



(a) Event 50

At the beginning of the simulation (Figure 3 (a)), CR users use almost all resources equally. After a certain simulation time, at event 100 (Figure 3 (b)) a few channels already show their priority to certain users, like user 3 prefers channel 8 and user 2 prefers channel 3. However, the channel usage of user 1 is still fairly equal at this stage. It can be seen that a spectrum sharing equilibrium is established and therefore the channel usage converged to few preferred channels. The CR users are able to avoid collisions by utilizing their experience from learning consequently.



(b) Event 100

Figure 3 Channel Usage at (a) Event 50, (b) Event 100.

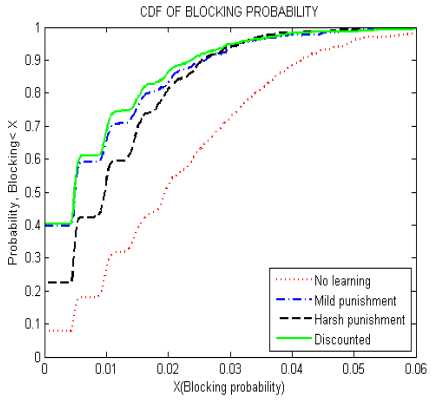


Figure 4 Cumulative Distribution Function of System Blocking Probability at Discrete Points over the Service Area

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

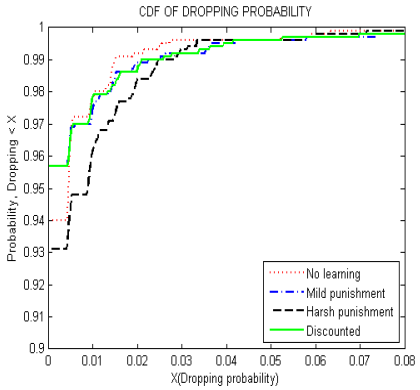


Figure 5 Cumulative Distribution Function of System Dropping Probability at Discrete Points over the Service Area

Figure 4 – Figure 5 illustrate the CDF of Blocking and Dropping probability respectively. Blocking probability is measured at regular points in the service area and a Cumulative Distribution Function (CDF) of system blocking probability at these points is derived. In order to analyze the level of system interruption, a CDF of dropping probability is calculated at the same time. All CR users' parameters are exactly the same for each scheme evaluation, with different system performance being caused only by different weighting factor values.

VI. CONCLUSION

In this paper, we introduced a reinforcement learning model for cognitive radio and a few basic reinforcement learning-based spectrum sharing schemes. By utilizing the ability of learning, cognitive agents can remember their preferred communication resources and enable an efficient approach to spectrum sensing and sharing accordingly. Simulation results show that reinforcement learning-based spectrum sharing algorithms achieve a better system performance compared to non-learning algorithms.

REFERENCES

[1] J. Mitola and G. Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communication,* vol. 6, pp. 13-18, Aug, 1999.
[2] J. Mitola, "Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio," Ph.D., Teleinformatics, Royal Institute of Technology (KTH), May, 2000.
[3] ITU-R. WRC-12 Agenda Item 1.19: Software-Defined Radio (SDR) and Cognitive Radio Systems (CRS). 2010.
[4] R. S. Sutton and A. G. Barto, *Reinforcement learning : An Introduction*: The MIT Press, 1998.
[5] L. P. Kaelbling, *et al.*, "Reinforcement Learning: A Survey," *Journal of artificial intelligence Research,* vol. 4, pp. 237-285, May. 1996.
[6] M. Bublin, *et al.*, "Distributed spectrum sharing by reinforcement and game theory," presented at the 5th Karlsruhe workshop on software radio, Karlsruhe, Germany, March. 2008.
[7] T. Jiang, *et al.*, "Performance of Cognitive Radio Reinforcement Spectrum Sharing Using Different Weighting Factors," presented at the International Workshop on Cognitive Networks and Communications (COGCOM) in conjunction with CHINACOM'08, , Hangzhou, China, August, 2008.
[8] S. Kapetanakis and D. Kudenko, "Reinforcement learning of coordination in cooperative multi-agent systems," presented at the Eighteenth national conference on Artificial intelligence, Edmonton, Alberta, Canada, 2002.
[9] S. Saunders, *Antennas and propagation for wireless communication systems*: Wiley, 1999.
[10] N. Drakos, "Introduction to Monte Carlo Methods," Computer Based Learning Unit, University of Leeds, Aug 1994.
[11] J. D. Gibson, *The Mobile Communications Handbook*, 1st ed.: IEEE Press, 1996.
[12] T. Jiang, *et al.*, "Two Stage Reinforcement Learning Based Cognitive Radio with Exploration Control," *accepted by IET Communications,* 2009.

BIOGRAPHY

U. Kiran, completed his B.Tech and M.Tech from Jawaharlal Nehru Technological University, Hyderabad, INDIA in 2008 and 2011. He published 4 research articles in various international Journals and Conferences. He is a member of professional bodies like ISTE and IETE. He is presently working as Assistant professor in department of Electronics and Communication Engineering. His area of research is in wireless Mobile Communication, LTE and Multiple input and Multiple output.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2013

D.Praveen Kumar, completed his B.Tech and M.Tech from Jawaharlal Nehru Technological University, Hyderabad, INDIA in 2006 and 2011. He published 10 research articles in various international Journals and Conferences. He is a member of international and national professional bodies like IEEE and IETE. He is presently working as an Assistant professor in department of Electronics and Communication Engineering. His area of research is in Wireless Mobile Communications, Cooperative Networking and Green Networks.

K.Rajesh Reddy, completed his B.Tech and M.Tech from Jawaharlal Nehru Technological University, Hyderabad, INDIA in 2008 and 2011. He published 8 wireless mobile communication related research articles in various international Journals and Conferences. He is a member of professional bodies like ISTE and IETE. He is presently working as Assistant professor in department of Electronics and Communication Engineering. His area of research is in Multiple input Multiple output, LTE and Orthogonal Frequency Division Multiplexing

M. Ranjith, completed his B.Tech and M.Tech from Jawaharlal Nehru Technological University, Hyderabad, INDIA in 2006 and 2010. He published 6 wireless communication related research articles in various international Journals and Conferences. He is a member of professional bodies like ISTE and IETE. He is presently working as Assistant professor in department of Electronics and Communication Engineering. His area of research is in Multiple input Multiple output, Heterogeneous Networking and Orthogonal Frequency Division multiplexing.