Sensors & Transducers, Vol. 164, Issue 2, February 2014, pp. 170-175

www.sensorsportal.com

Sensors & Transducers

http://www.sensorsportal.com

Multiagent Reinforcement Learning Dynamic Spectrum Access in Cognitive Radios

[1] Wu Chun, [2] Yin Mingyong, [2] Ma Shaoliang, [1] Jiang Hong

[1] School of National Defense Technology, Southwest University of Science and Technology, Mianyang 621000, Sichuan, China

[2] Institute of Computer Application, China Academy of Engineering Physics, Mianyang 621900, Sichuan, China

[1] Tel.: 86-816089890, fax: 86-816089890

[1] E-mail: soldier_wu@163.com

Received: 28 November 2013 /Accepted: 28 January 2014 /Published: 28 February 2014

Abstract: A multiuser independent Q-learning method which does not need information interaction is proposed for multiuser dynamic spectrum accessing in cognitive radios. The method adopts self-learning paradigm, in which each CR user performs reinforcement learning only through observing individual performance reward without spending communication resource on information interaction with others. The reward is defined suitably to present channel quality and channel conflict status. The learning strategy of sufficient exploration, preference for good channel, and punishment for channel conflict is designed to implement multiuser dynamic spectrum accessing. In two users two channels scenario, a fast learning algorithm is proposed and the convergence to maximal whole reward is proved. The simulation results show that, with the proposed method, the CR system can obtain convergence of Nash equilibrium with large probability and achieve great performance of whole reward. *Copyright © 2014 IFSA Publishing, S. L.*

Keywords: Cognitive radios, Multiagent reinforcement learning, Q-learning, Dynamic spectrum access.

1. Introduction

Under the trend of information innovation in current world economy and social development, the wireless communication technology has experienced rapid development. Cognitive radio (CR) [1] becomes a hot research topic in wireless communication domain owing to its advantages of dynamic spectrum access and intelligent adaptation to environment. The capability of high intelligence is one of significant key characteristics of CR, and the learning represents CR intelligence mostly. There are on-line learning and off-line learning methods applying in CR generally [2]. In on-line learning, the agent interacts with the environment, gets feedback reward, and learns from its own experience. The reinforcement learning is the representative on-line learning method.

The centralized solution is commonly used in traditional wireless communication for applying on-line learning to solve the issues of resource allocation. J. Nie presents a dynamic spectrum allocation method with centralized Q-learning in mobile communication systems [3]. S. Xergias makes use of the centralized E-FRTS (enhanced frame registry tree scheduler) to accomplish the schedule and allocation of multimedia traffic in IEEE 802.16 mesh networks [4].

On account of the autonomy and variety of CR users and the potential heterogeneity of cognitive

Article number P_1848

radio networks (CRN), the decentralized learning is more suitable for CRN than the centralized learning. To maximize the global reward of all users, multi CR users negotiate with each other ordinarily and that needs information exchange. J. E. Suris applies cooperative game model in the distributed spectrum sharing and proposes a distributed algorithm to achieve near optimal allocation, and the user exchange information of actions and rewards with each other in proposed algorithm [5]. P. Zhou studies applying Bush-Mosteller reinforcement learning to resolve power control issue in CRN [6]. Although CR users don't exchange strategy and reward information with each other, they still get the total jamming intensity (global reward) from the primary user. The information exchange between CR user and CR user (or primary user) need occupancy a certain amount of communication resource. Moreover, too frequent interactions may cause the overload of the communication. To overcome this shortcoming, an alternative method is self-learing (independent learning), with which the user only learns and acts based on itself reward, not exchanging information with each other. Currently, the researches on self-learning in CRN are scarce in literatures. This paper applies the repeated game to model the issue of multi users competing multi channels, and proposes a multiagent reinforcement learning method: multiuser independent Q-learning method with which the CR user implement self-learning to maximize global reward. The simulations validate the effectiveness of proposed method.

2. Stochastic Spectrum Access Model

The issue that *an M CR users (SUs, Second Users) access N channel not occupied by PUs (Primary Users) is* researched (only $M \le N$ is considered in this paper). Each CR user chooses channel independently according to itself tactics repeatedly and aims at achieving the maximal total reward of all users. The user does not exchange information with others in the whole learning and channel selecting process. The repeated game [7] is used to model the process of multi CR users competing channels. At each stage game, *M* users choose respective channel (in the whole *N* channels), and the reward of anyone is determined by the combined strategies of all users. This stage game is presented by a matrix game [8] $(M, A^{(1)}, \cdots, A^{(M)}, r^{(1)}, \cdots, r^{(M)})$, where M is the number of users in game, $A^{(m)}$ is actions assemble of user m and it includes N actions, i.e. choosing channel $n(n = 1, 2, \cdots N)$, A is combined action space $A = A^{(1)} \times \cdots \times A^{(M)}$, $r^{(m)}$ is reward function of user m . When user m chooses channel n and it conflicts with other user's choosing, the reward $r^{(m)}(n)$ equals to zero, i.e. $r^{(m)}(n) = 0$. The reward $r^{(m)}(n)$ is determined by channel gain with no channel conflict, $r^{(m)}(n) = b^{(m)}(n)$. The reward matrix of a game that two users compete two channels is shown as formula (1) and (2)

$$R^{(1)} = \begin{bmatrix} 0 & b^{(1)}(2) \\ b^{(1)}(1) & 0 \end{bmatrix}, \quad (1)$$

$$R^{(2)} = \begin{bmatrix} 0 & b^{(2)}(2) \\ b^{(2)}(1) & 0 \end{bmatrix}. \quad (2)$$

The item at line i , column j in matrix $R^{(m)}$ denotes the reward of user m when user *1* chooses channel i and user *2* chooses channel j .

3. Multiuser Independent Q-learning

The goal of the repeated game for accessing channels is to maximize the reward of the stage game $g_t = \sum_{m=1}^{M} r_t^{(m)}$ after executing the stage game for many times. Multiagent reinforcement learning (MARL) [9] is effective method to resolve the game. In most current literatures about multiuser game, the user need observe the rewards and the strategies of other users. In cognitive radio networks, such frequent information interactions of the rewards and the strategies between CR users will occupy a great quantity of communication resource. This paper proposes a multiuser independent Q-learning (MIQ) method with which the CR users don't require any information exchange between each other. The MIQ algorithm in the repeated game model expects achieving two goals: one is the convergence to Nash Equilibrium and the other is the total reward of all users reach the maximal value or the close maximal value.

Definition 1 The strategies assemble $(\sigma_m^*)_m$ is a Nash Equilibrium if for each user $m = 1, 2, \cdots M$ it has

$$r^{(m)}(\sigma_m^*, \sigma_{-m}^*) \ge r^{(m)}(\sigma_m, \sigma_{-m}^*), \quad \forall \sigma_m, \quad (3)$$

where σ_m^* is the strategy of user m , σ_{-m}^* is the combined strategies of the other users except user m , $r^{(m)}$ is the reward of user m . Among all possible combined strategies of every user choosing different channel, there must exist one combined strategies which has better or same reward than other combined strategies. That combined strategies is a Nash Equilibrium point. The Nash Equilibrium point exists apparently in above mentioned game, but the Nash Equilibrium point may be not unique. For instance, all orthogonal channel allocation strategies are Nash Equilibrium points when $M = N$.

The basic Q-learning method learns the optimal strategies in unknown environment by using learned knowledge and exploring new strategies with certain probability. In the situation of undetermined rewards

and actions, the updating formula of Q-value function is [10]:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a')) \quad (4)$$

Some improvements on basic Q-learnig are required for the proposed issue. Each user learns independently in game process, and its reward is affected by other users. The reward is uncertainty thus the slow updating of Q-value is a reasonable manner. In addition, owing to the status of users does not transform during repeated game, the new Q-value does not contain the contributions of delay reward. With the above two improvements, as well as proper exploring policy and control of learnig rate, a multiuser independent Q-learning method is proposed. The key to achieve the joint optimal solution by independent actions and learning is designing suitable autonomous learning policy and actions policy. Two principles are proposed and applied in the independent actions of each user: 1) The user prefers choosing the channel with high gain; 2) The user avoids channel conflict between others. Nevertheless, the two principles may conflict occasionally or frequently. The proposed MIQ algorithm executes iteration action tries under the two principles and gets the final channel allocation. The concrete implementation of MIQ algorithm is as below:

Step 1: Q-value table initializing. The Q-value table of user m is initialized as

$$Q^{(m)}(n) = \frac{\sum_{i=1}^{N} b^{(m)}(i)}{N}, n = 1,2,\cdots N. \quad (5)$$

The initialized Q-value of user m is the average value of the rewards on different channels. After initialization, each item in Q-value table has the same average value, therefore the user chooses any channel with same probability in first action. A more deep reason to initialize Q-value with average value is to conveniently realize the subsequent updating principle of Q-value: big reward makes Q-value increase slowly and small reward makes Q-value decrease slowly.

Step 2: Independent Q-learning process. Q-value update iteratively until reaching specified times of game.

a) Compute the probabilities of choosing each channel based on Q-value table, execute the actions with the probabilities in formula (6)

$$P^{(m)}(n) = \frac{(Q^{(m)}(n))^q}{\sum_{n=1}^{N} (Q^{(m)}(n))^q}, n = 1,2,\cdots N. \quad (6)$$

The bigger Q-value indicates the better channel and meanwhile results in the higher probability of channel choosing. In formula (6), q is the probability controlling factor. The selection of actions inclines to use learned knowledge with bigger q value and inclines to explore all possible choices with smaller q value. Due to there is no information interaction between users, adequate exploration is significant and necessary in the initial stage of learning. Generally, a small q is set at the very beginning of learning, along with the repeated learning process the q increase gradually and slowly to improve the convergence of learning.

b) After the actions, each user observes itself reward only.

$$r_t^{(m)}(n) = \begin{cases} b^{(m)}(n) & \text{user } m \text{ no conflict} \\ 0 & \text{user } m \text{ conflict} \end{cases}, \quad (7)$$

On the one hand, the definition of reward represents the quality of channel, i.e. the reward is the channel gain with no conflict. On the other hand, the reward reflects the conflict status of the channels. When action of user m conflicts with any other, the reward gets zero and that results in the decrease of corresponding Q-value. This can be deemed as punishment mechanism of channel conflict.

c) Update the Q-value table.

$$Q_{t+1}^{(m)}(n) = (1-\alpha_t^{(m)}(n))Q_t^{(m)}(n) + \alpha_t^{(m)}(n) b_t^{(m)}(n) \quad (8)$$

where $\alpha_t^{(m)}(n) = \frac{\beta}{1+\lambda_t^{(m)}(n)}$ is the updating rate of Q-value, $\lambda_t^{(m)}(n)$ denotes the times user m chooses channel n during the whole t repeated games, β denotes the controlling factor of Q-value updating rate. The updating rate of Q-value $\alpha_t^{(m)}(n)$ reduces gradually during learning process apparently and that contributes to the convergence of learning. When the user chooses a channel with high gain (higher than average gain) on the condition of no conflict, the Q-value increases, and higher gain lead to the more increase of Q-value. Meanwhile, if the user's selective channel conflict with others, the Q-value updating in formula (8) with current zero reward causes the decrease of Q-value, and the degree of decrease is much more than the degree of increase obtained in unconflict situation. Because of rather heavy punishment for channel conflict, the user could search high gain channel on the basis of no confliction.

The study on self-learning (independent learning) with which the agent learns and acts only by observing itself reward in MARL filed is very few. It is especially difficult to prove the convergence of self-learning on the condition that the multi users don't exchange information. Bowling proposed a multiagent learning method WOLF-PHC (win or learn fast policy hill-climbing) in which each user

learns using a variable learning rate and it achieves the maximal rewards of all users [11]. But the proof of the algorithm convergence is not provided and only an experiment on 2 users and 2 actions is done to evaluate the convergence property. This paper proposes a fast learning algorithm for 2 users and 2 channels scenario, and makes the proof of the convergence. The concrete implementation of the fast learning algorithm is as below:

Step 1: Phase of learning channel reward. Each user explores channels randomly thus get the reward value (channel gain) without channel conflict.

Step 2: Phase of fast greedy channel choosing. Each user chooses the channel of highest gain. The allocation of channels and learning process are finished if no channel conflict occurs. When channel conflict occurs, Step 3 executes subsequently.

Step 3: Phase of Q-learning. The learn process repeat for specified times.

(a) Initialize the Q-value table.

$$Q^{(1)}(1) = Q^{(1)}(2) = Q^{(2)}(1) = Q^{(2)}(2) = 0.5 . \quad (9)$$

(b) Choose action based on probability policy and observe the reward. The user chooses channel according to the probabilities in formula (6), observes the status of conflict, and calculates the reward.

When the user chooses the higher gain channel in the two candidates, the reward is

$$r_t^{(m)}(n) = \begin{cases} \Delta^{(m)} & \text{user } m \text{ no conflict} \\ -\Delta & \text{user } m \text{ conflict} \end{cases}, \quad (10)$$

where $\Delta^{(m)} = \left| b^{(m)}(1) - b^{(m)}(2) \right| / L$, Δ is an appropriate value between $\Delta^{(1)}$ and $\Delta^{(2)}$.

When the user chooses the lower gain channel in the two candidates, the reward is

$$r_t^{(m)}(n) = \begin{cases} \Delta & \text{user } m \text{ no conflict} \\ -\Delta & \text{user } m \text{ conflict} \end{cases}. \quad (11)$$

(c) Update the Q-value table.

$$Q_{t+1}^{(m)}(n) = Q_t^{(m)}(n) + r_t^{(m)}(n) . \quad (12)$$

Theorem 1 The proposed fast learning algorithm converges to the optimal solution.

Proof:

Suppose $b^{(1)}(1) > b^{(1)}(2), b^{(2)}(2) > b^{(2)}(1)$ or $b^{(1)}(1) < b^{(1)}(2), b^{(2)}(2) < b^{(2)}(1)$. It is clear that the algorithm converge to the optimal solution in the phase of fast greedy channel choosing.

Suppose $b^{(1)}(1) > b^{(1)}(2), b^{(2)}(1) > b^{(2)}(2)$. Assign the parameter L a sufficiently large value thus make $\Delta^{(1)}$ and $\Delta^{(2)}$ are small enough, and assign Δ an appropriate value between $\Delta^{(1)}$ and $\Delta^{(2)}$. In the first period of time of learning, user *1* performs $4n$ times of channel choosing. On the condition that Δ , $\Delta^{(1)}$ and $\Delta^{(2)}$ are extremely small, the Q-value table updates as

$$\begin{cases} Q^{(1)}(1) \doteq Q^{(1)}(1) + n\Delta^{(1)} - n\Delta \\ Q^{(1)}(2) \doteq Q^{(1)}(2) + n\Delta - n\Delta \\ Q^{(2)}(1) \doteq Q^{(2)}(1) + n\Delta^{(2)} - n\Delta \\ Q^{(2)}(2) \doteq Q^{(2)}(2) + n\Delta - n\Delta \end{cases}. \quad (13)$$

If $\Delta^{(1)} > \Delta^{(2)}$, there is $Q^{(1)}(1) > Q^{(1)}(2) \doteq Q^{(2)}(2) > Q^{(2)}(1)$. During the subsequent learn periods, $Q^{(1)}(1)$ increases continually meanwhile $Q^{(2)}(1)$ decreases continually. After a while, the learn process ends and final solution is that each user chooses the channel with bigger Q-value, i.e. the user *1* chooses channel *1* and the user *2* chooses channel *2*. This solution is the optimal solution owing to $b^{(1)}(1) + b^{(2)}(2) > b^{(1)}(2) + b^{(2)}(1)$. If $\Delta^{(1)} > \Delta^{(2)}$, the algorithm converges too.

Suppose $b^{(1)}(1) < b^{(1)}(2), b^{(2)}(1) < b^{(2)}(2)$, the convergence can be prove in the same way.

4. Simulation and Results

The MIQ algorithm is simulated and evaluated mainly in three aspects: the probability of convergence to Nash Equilibrium, the probability of convergence to the optimal solution and the normalization performance of proposed algorithm.

In the scene of M CR users selecting N channels, the reward of each user m choosing each channel n is initialized to uniformly distributed random numbers between 0.5 and 1,

$$b^{(m)}(n) = 0.5 + 0.5 * rand( ) . \quad (14)$$

Then each user carries out autonomic learning with MIQ algorithm independently. In policy updating procedure in Step 2(a), the probability controlling factor q adjusts dynamically. The q equals 0.5 when selecting channel for the first time, and the q increases very gradually until reaching specified learning times. In Step 2(c), the controlling factor of Q-value updating rate β is configured with 1. The number of times of repeated game is 10000, and the simulation process is executed 100 times with diverse random rewards to obtain the average performance of proposed algorithm.

Fig. 1 and Fig. 2 show performance of MIQ algorithm when the number of users M equal to the number of channels N. The strategies of multi users can converge to Nash Equilibrium at 100 % or near 100 % (as shown in Fig. 1). When $M = N = 2$, the total reward of all users converge to the maximal value with a probability of 98 %. The probability of convergence to maximal reward drops along with the increase of users/channels number, and the

probability drops to near 70 % when $M = N = 8$. Fig. 1 shows superficially that the MIQ algorithm behaves great performance under the scene of very few users while the MIQ algorithm is not a quite good method under the scene of many more users. More deep evaluation of the algorithm for average performance is described in Fig. 2. When $M = N = 2$ the normalization performance (the ratio of current total reward to the maximal total reward) $\eta = 0.9999$ and when $M = N = 3$ the normalization performance $\eta = 0.9995$. Along with increase of users/channels number, the normalization performance drops very slowly. When $M = N = 8$, the normalization performance maintains very high value, i.e. $\eta = 0.9978$. The reason that the normalization performance keeps high while the probability of convergence to maximal reward drops apparently is: the total reward under determined paradigm of channel allocation is very close to maximal reward. Therefore, even if the MIQ algorithm can not achieve absolute optimal performance, it achieves quite good performance which is very close to the absolute optimal performance.

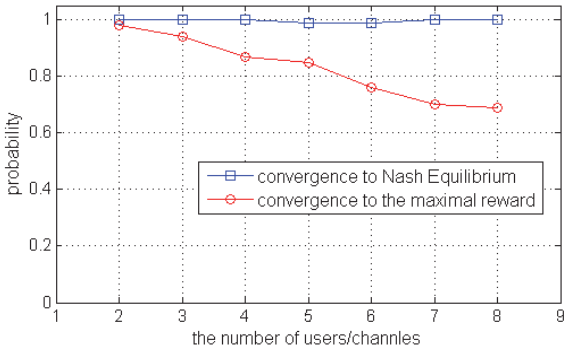

Fig. 1. The probabilities of convergence to Nash Equilibrium and the maximal reward $(M = N)$.

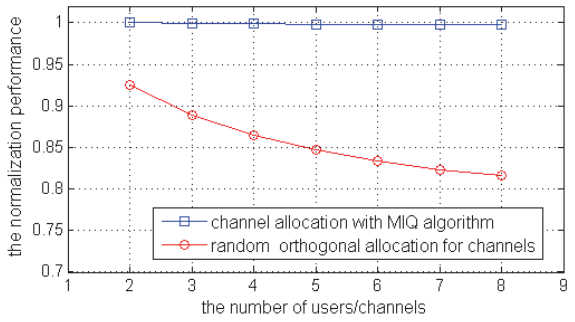

Fig. 2. The normalization performance of MIQ algorithm and random orthogonal allocation method $(M = N)$.

Fig. 3 shows the 100 samples of normalization performance obtained in simulations $(M = N = 8)$. It can be seen that in total 100 simulations, the MIQ algorithm converges to maximal reward for 69 times and the normalization performance η got in the remaining 31 simulations is greater than 0.98 mostly, even the worst performance is greater than 0.97.

On the condition that user number equals to channel number, the MIQ algorithm can reach unconflicted orthogonal channel allocation and the normalization performance obtained by MIQ algorithm is approximately 15 % higher than that of random orthogonal channel allocation method.

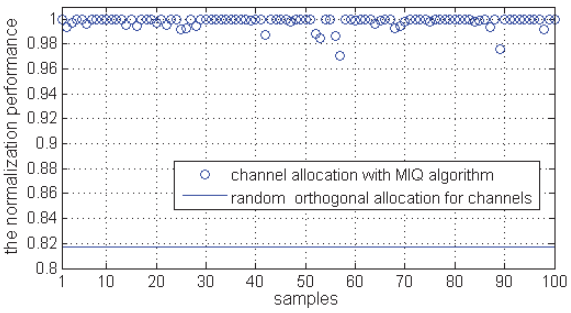

Fig. 3. The normalization performance distribution of 100 simulation samples for MIQ algorithm $(M = N = 8)$.

Fig. 4 and Fig. 5 show performance of MIQ algorithm when the number of users *M* is less than or equal to the number of channels *N*. Fig. 4 shows the probabilities of convergence to Nash Equilibrium and maximal total reward by proposed algorithm with different channel number $(M = 2, 3)$.

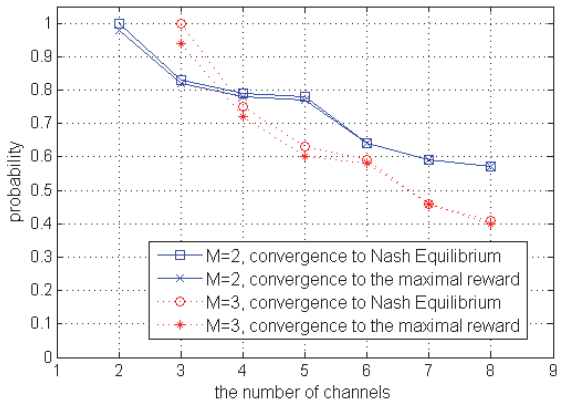

Fig. 4. The probabilities of convergence to Nash Equilibrium and the maximal reward $(M \le N)$.

Along with increase of channel number, not only the probability of convergence to maximal total reward drops obviously but also the probability of convergence to Nash Equilibrium drops similarly, and it is different from the case shown in Fig. 1. The channels conflict in MIQ learning process would lead to the decrease of Q-value and thus make final allocation of channels can avoid channels conflict

effectively. The strategies assemble realizing unconflicted allocation of channels is exactly Nash Equilibrium when $M = N$, so the probability of convergence to Nash Equilibrium is quite high (as shown in Fig. 1). When $M < N$, the unconflicted allocation of channels does not necessarily satisfy Nash Equilibrium, and that is why the probability convergence to Nash Equilibrium drops in Fig. 4. Fig. 5 shows the proposed MIQ algorithm has good performance too when $M \le N$.

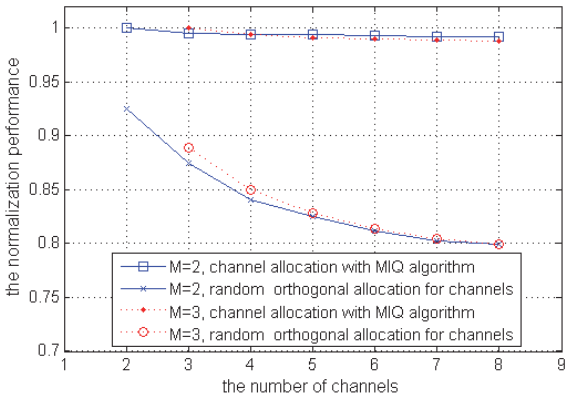

Fig. 5. The normalization performance of MIQ algorithm and random orthogonal allocation method ($M \le N$).

6. Conclusions

The independent learning without information exchange between each node is an alternative on-line learning method for resource allocation in cognitive radio networks. This paper uses the repeated game modeling multiuser dynamic spectrum accessing, and proposes a multiagent reinforcement learning method: multiuser independent Q-learning method with which the CR user coordinates in choosing best highest gain channel and avoiding conflict between each other. Moreover, a fast learning algorithm for 2 users and 2 channels case is presented and proved that it converge to Nash Equilibrium. The simulations show that user action can converge to Nash Equilibrium with high probability and achieved total reward is close to the maximal reward with proposed MIQ algorithm.

Acknowledgements

Project supported by the National Natural Science Foundation of China (Grant No. 61379005), and the National Basic Research Program of China (Grant No. 2009CB320403).

References

[1]. J. Mitola, Jr. G. Q. Maguire, Cognitive radio: making software radios more personal, *IEEE Personal Communications*, Vol. 6, Issue 4, 1999, pp. 13-18.

[2]. C. Wu, Y. Li, K. Yi, Research on GA-LSSVM offline learning in cognitive radios, *Journal of Beijing University of Posts and Telecommunications*, Vol. 35, Issue 2, 2012, pp. 90-93.

[3]. J. Nie, S. Haykin, A Q-learning-based dynamic channel assignment technique for mobile communication systems, *IEEE Transactions on Vehicular Technology*, Vol. 48, Issue 5, 1999, pp. 1676-1687.

[4]. S. Xergias, N. Passas, A. K. Salkintzis, Centralized resource allocation for multimedia traffic in IEEE 802.16 mesh networks, *Proceedings of the IEEE*, Vol. 96, Issue 1, 2008, pp. 54-63.

[5]. J. E. Suris, L. A. Dasilva, H. Zhu, et al., Cooperative game theory for distributed spectrum sharing, in *Proceedings of the IEEE International Conference on Communications*, Glasgow, Scotland, 2007, pp. 5282-5287.

[6]. P. Zhou, Y. Chang, J. A. Copeland, Reinforcement learning for repeated power control game in cognitive radio networks, *IEEE Journal on Selected Areas in Communications*, Vol. 30, Issue 1, 2012, pp. 54-69.

[7]. V. D. Schaar M, F. Fu, Spectrum access games and strategic learning in cognitive radio networks for delay-critical applications, *Proceedings of the IEEE*, Vol. 97, Issue 4, 2009, pp. 720-739.

[8]. C. Yang, J. Li, Mixed-strategy based discrete power control approach for cognitive radios: A matrix game-theoretic framework, in *Proceedings of the 2nd International Conference on Future Computer and Communication*, Wuhan, China, 2010, pp. 3806-3810.

[9]. L. Busoniu, R. Babuska, B. D. Schutter, A comprehensive survey of multiagent reinforcement learning, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, Vol. 38, Issue 2, 2008, pp. 156-172.

[10]. Tom M. Mitchell, Machine learning, *McGraw-Hill College*, 2005.

[11]. B. Michael, V. Manuela, Multiagent learning using a variable learning rate, *Artificial Intelligence*, Vol. 136, Issue 2, 2002, pp. 215-250.