A Multi Armed Bandit Formulation of Cognitive Spectrum Access

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider a cognitive network where a cognitive user attempts to access the channel if not occupied by primary users. The problem is formulated as a multi-armed bandit (MAB) problem. After reviewing several existing MAB algorithms, we propose a new MAB algorithm. The simulation results demonstrate the advantage of the proposed scheme compared to other listing algorithms when applied to a cognitive spectrum access problem.

1 Introduction

Recently, the overwhelming increase of wireless services and devices results in overcrowded wireless networks and the lack of spectrum resources. The problem stimulated the generation of a new paradigm of wireless communication, referred as cognitive communications [1]. The basic idea of this communication technique is to take advantage of unused portions of licensed spectrum resources. In a cognitive network, users are classified into primary users and secondary users. Primary users always gains the permission to transmit, while secondary users, also known as cognitive users, first senses the channel and transmits its information if the channel is not occupied. Extensive attention has been paid to develop efficient schemes for the cognitive users to access the spectrum. In this paper, we propose to cast the media access problem of cognitive users into the frame of a multi-armed bandit (MAB) problem. Each channel is considered as a slot machine with certain expected reward while the cognitive user is considered as a gambler playing on several slot machines.

The MAB has been well investigated in the context of machine learning. The UCB algorithm proposed in [2] is proven to be optimal if the reward distribution is stationary. On the other hand, with non-stationary reward distributions, Whittle's index [3] is proven to be asymptotically optimal. However, these algorithms assume infinite time, therefore cause problem when applied into the spectrum access problem of cognitive users. Moreover, the very nature of a wireless channel is that it is normally time varying, which also should be treated carefully when applying exiting MAB algorithms into cognitive communication. In this paper, we introduce and evaluate several existing MAB algorithms, and also proposed a new algorithms which is a combination of existing schemes. However, the new algorithms take account of both the finite-time and time varying nature of a wireless channel.

The remainder of the paper is organized as follows: In section 2, we describe the network model and formulate the spectrum access problem of cognitive communication as a MAB problem. Section 3 introduces several existing MAB algorithms as well as the proposed algorithm. Simulation results are provided in section 4, followed by the concluding remarks in section 5.

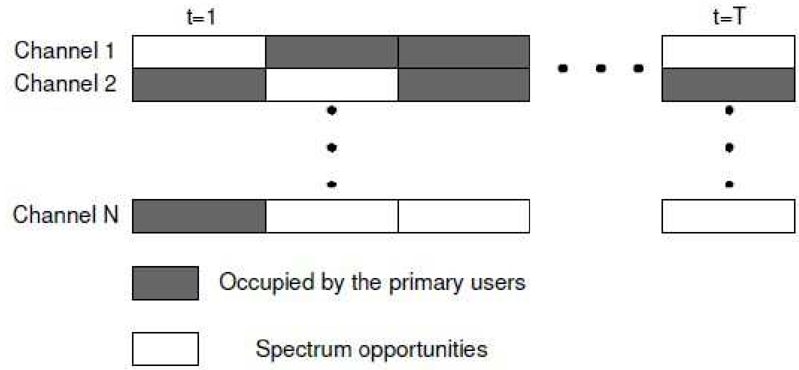

Figure 1: Channel model.

2 Network Model

Fig. 1 shows the network model of interest in this paper[1]. Consider a network consisting of total N channels, $\mathcal{N} = \{1, \ldots, N\}$. The primary users have the priority to access all the channels, while a cognitive user tries to use these channels when they are not occupied by the primary users. The channels are accessed in a time-slotted fashion. Let i refer to the channel index, j refer to the time slot index and k denote the cognitive user index. Assume that at each time slot, channel i is free with probability p_i and let $\mathbf{p} = (p_1, \ldots, p_N)$. Let $b_i(j)$ be a random variable that equals 1 if channel i is available at time slot j and equals 0 otherwise. For the wireless channel, we assume a block varying model, i.e., the value of $\mathbf{p}$ is static for a block of T time slots. Normally, the cognitive user assumed to be unaware of $\mathbf{p}$ a priori.

In the network model, the cognitive user seeks to exploit the free channels by sensing a channel at the beginning of each time slot. In particular, at time slot j, the cognitive user selects channel $s(j) \in \mathcal{N}$ to access. If the sensing result shows that channel $s(j)$ is free, i.e., $b_{s(j)}(j) = 0$ then the cognitive user can send one unit of information over this channel; otherwise the cognitive user have to wait until the next time slot and choose again a channel to access. The problem is that which channel the cognitive user should choose to sense at each time slot. Therefore, we can compute the total number of units of information that the cognitive user is able to send over one block as

$$W = \sum_{j=1}^{T} b_{s(j)}(j)\,. \tag{1}$$

and the problem can be generalized as characterizing strategies that maximize

$$\mathbb{E}\{W\} = \mathbb{E}\left\{\sum_{j=1}^{T} b_{s(j)}(j)\right\}\,. \tag{2}$$

Intuitively, we can observe that the essence of the problem is a trade-off between exploitation and exploration. By exploitation, it refers to that the cognitive user performs myopic action by selecting the channel with th highest probability of being free according to all the observations. On the other hand, by exploration, it means in order to learn the true value of $\mathbf{p}$[2], the cognitive user will try to choose to different channel to access at different time slots. The above observation allows us to interpret the problem in a bayesian approach and to further reformulate the problem as a MAB problem.

[1]We use a network model and notations similar to [4].

[2]It is assumed there is a true value of $\mathbf{p}$ in the real world.

2.1 Problem Formulation

We can use the following typical MAB example to illustrate our problem properly: A gambler is sequentially choose one of N machines to play. If he wins, there will be one unit of reward. The ith machine has winning probability p_i, which is unknown to the gambler. But he has observations of the outcomes of past plays. The goal is to maximize the overall reward after a total of T plays.

Denote a medium access strategy of the cognitive user, i.e., a strategy of how to choose channels, by Γ. Therefore, Γ is a function of the previous $j-1$ observations:

$$\Phi(j) = \{s(1), b_{s(1)}(1), \dots, s(j-1), b_{s(j-1)}(j-1)\}, j \geq 2\,. \tag{3}$$

Note that $s(j)$ is the channel chosen by adopting strategy Γ at time j, i.e., $s(j) = \Gamma(\Phi(j))$.

The payoff function is the expected units of informations the cognitive user is able to transmit through a block

$$W_\Gamma = \mathbb{E}\left\{\sum_{j=1}^{T} b_{s(j)}(j)\right\} = \sum_{j=1}^{T}\sum_{i=1}^{N} p_i \Pr\{\Gamma(\Phi(j)) = i\}\,. \tag{4}$$

and the regret function is

$$R_\Gamma = \sum_{j=1}^{T} p^* - \sum_{j=1}^{T}\sum_{i=1}^{N} p_i \Pr\{\Gamma(\Phi(j)) = i\}\,, \tag{5}$$

where $p^* = \max\{p_1, \dots, p_N\}$.

With the MAB problem well formulated, we now are ready to proceed to learning algorithms.

3 Learning Algorithms

3.1 Upper Confidence Bound

In [5], Agrawal defines a family of policies based on the man value of the reward. These policies are referred as the Upper Confidence Bound (UCB) algorithms. The main idea of UCB is to add a bias factor to the mean value of the reward. The algorithm first selects each channel once. Then, at time slot j, UCB chooses channel $s(j)$ such that

$$s(j) = \arg\max_{i\in\mathcal{N}}\left(\frac{x_i(j)}{y_i(j)} + \sqrt{\frac{\sigma \log j}{y_i(j)}}\right)\,, \tag{6}$$

where $y_i(j)$ is the number of times channel i has been chosen to access till time $j-1$, $x_i(j) = \sum_{t=1}^{j} v_i(t)$, $v_i(t)$ is the number of time slots for which the cognitive user has sensed channel i to be free till time $t-1$, and σ is a design parameter chosen to be 2 in [5].

3.2 Upper Confidence Bound Tuned (UCBT)

The UCBT algorithm was first proposed by Auer *et al.* in [6]. The main characteristic of the UCBT is the use of empirical variance in the bias sequence. Thus, the exploration is reduced for the channels with small reward variance. The UCBT algorithm chooses channel $s_i(j)$ such that

$$s_i(j) = \arg\max_{i\in\mathcal{N}}\left(z_i(j) + \sqrt{\frac{(z_i(j) - (z_i(j))^2)\sigma \log j}{y_i(j)} + \frac{c \log j}{y_i(j)}}\right)\,, \tag{7}$$

where $z_i(j) = \frac{x_i(j)}{y_i(j)}$ and c is also a design parameter free to adjust.

3.3 Discounted UCB (DUCB)

The discounted UCB [7] adds a discount factor to the original UCBT algorithm. The average reward are weighted as

$$\hat{z}_i(j) = \frac{\sum_{t=1}^{T} \gamma_i^{T-t} x_i(t)}{\hat{n}_i(j)}\ , \ \hat{n}_i(j) = \sum_{t=1}^{T} \gamma_i^{T-t} \mathbf{1}_{s_(t)=i}\}\ , \tag{8}$$

where $0 < \gamma_i < 1$ is the discount factor for channel i. The factor γ_i represents how fast channel i changes. The discounted UCB is especially suitable for wireless channels because of the time varying nature of wireless environment. The algorithm assigns less weight for old data and more weight for fresh data.

3.4 Sliding Window UCB (SWUCB)

Another practical algorithm the sliding window UCB [8]. The difference between SWUCB and DUCB is that SWUCB only uses a window of length l and only consider the average reward within this window. The window length decreases as the dynamic environment changes faster.

3.5 Combined UCBT and DUCB

In this section, we proposed a novel UCB which combines the UCBT and the DUCB algorithms. The combined algorithm adopts the Equation (8) as average reward function and uses the selection criteria of DUCB. Therefore, the selection criteria of the new algorithm is expressed as

$$s_i(j) = \arg\max_{i \in \mathcal{N}} \left(\hat{z}_i(j) + \sqrt{\frac{(\hat{z}_i(j) - (\hat{z}_i(j))^2)\sigma \log j}{y_i(j)}} + \frac{c \log j}{y_i(j)} \right)\ , \tag{9}$$

where $\hat{z}_i(j)$ is given in Equation (8).

The combined algorithm enjoys the benefits of both UCBT and DUCB, therefore it considers the effect of the empirical variance, as well as the time varying nature of wireless channels.

4 Simulation Results

In this section, we provide the simulation results for all the MAB algorithms introduced in this paper as well as the proposed new algorithm. The test scenario includes 20 channels with time block length $T = 100$ and 2000 blocks in total. The wireless channels are generated according to the IEEE standard 802.11. The simulation results including average regret, variance of regret and the percentage of time choosing the optimal channel are plotted in Figure 2, 3, and 4. It can be observed that, although UCB exhibits the highest average regret and regret variance, it performs best in terms of the percentage of time choosing the optimal channel. UCBT performs best in terms of regret variance and SWUCB exhibits the best average regret. The performance of the proposed algorithm lies in between that of UCBT and SWUCB. However, it has better optimal channel chosen percentage than those two algorithms.

5 Concluding Remarks

In this paper, we propose to make use of the MAB problem model to formulate the spectrum access problem in cognitive radio in the context of wireless communication. Several existing algorithms for solving the MAB problem are introduced. We also proposed a novel algorithm, the combined UCBT and SWUCB algorithm to address the problem. Performance of these algorithms are evaluated under wireless channels generated by the IEEE 802.11 standard model.

Several aspects worth further investigation as potential future work. First, although the simulation results demonstrates its advantage of the proposed scheme, it is necessary to derive the theoretical bounds on regrets in order to evaluate exactly how good the scheme is. Moreover, multiple cognitive users can be included in the network model. Finally, the work can be extended by adding the actual behavior model of the primary users to generate the probability distribution of channels being free.

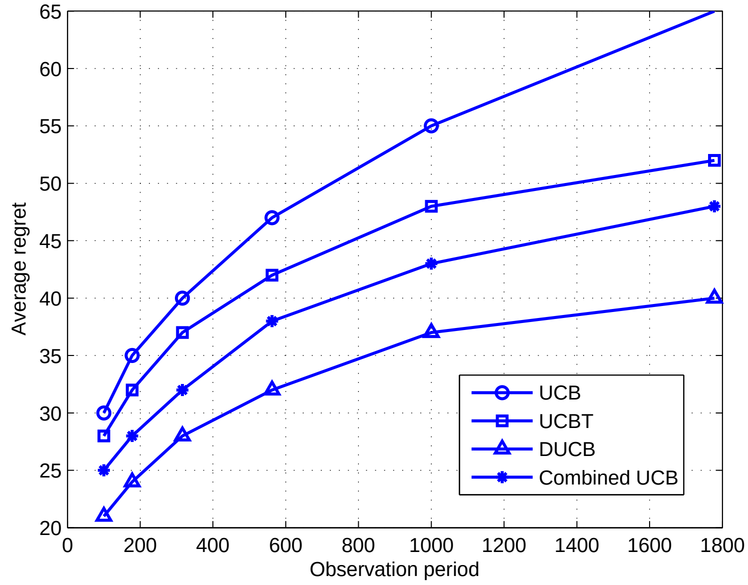

Figure 2: Average regret.

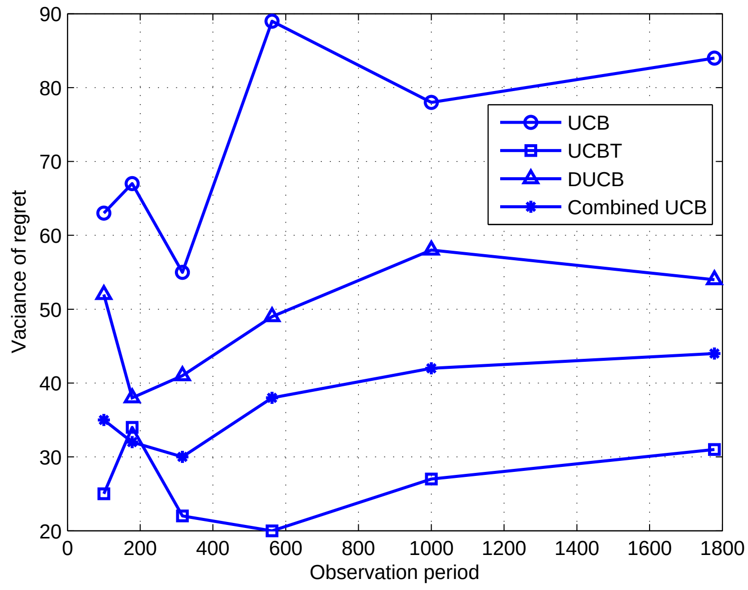

Figure 3: Variance of regret.

References

[1] Mitola, J. (2000) Cognitive radio: an integrated agent architecture for software defined radio. Royal Institute of Technology (KTH), Stockholm, Sweden.

[2] Gittins, J. & Jones, D. (1974) A dynamic allocation indices for the sequential design of experiments. *Progress in Statistics*, *European Meeting of Statisticians*, vol. 1, pp. 241-266.

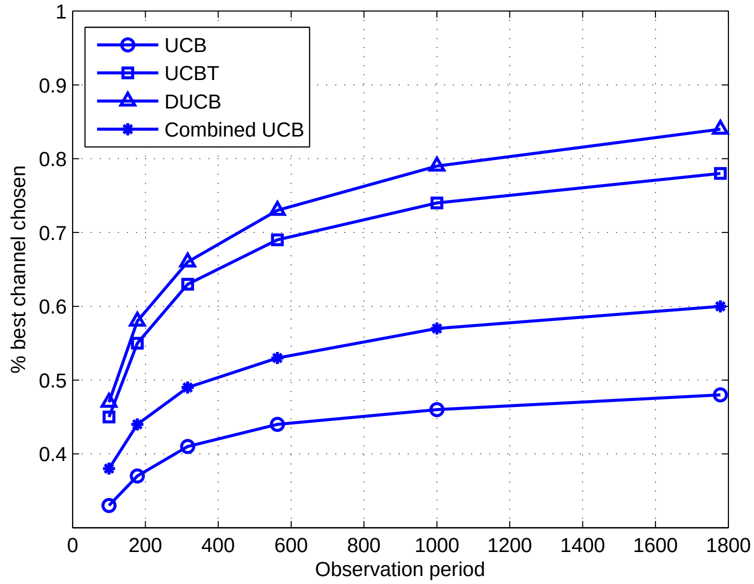

Figure 4: Percentage of best channel chosen.

[3] Whittle, P. (1988) Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, vol. 25.

[4] Lai, L. & Gamal, H. El & Jiang, H. & Poor, H. V. (2007) Cognitive medium access: exploration, exploitation and competition. *IEEE/ACM Trans. on Networking*, vol.10, no. 2, pp. 239-253.

[5] Agrawal, R. (1995) Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, vol. 27, pp. 1054-1078.

[6] Auer, P. & Cesa-Bianchi, N. & Fisher, P. (2002) Finite time analysis of the multiarmed bandit problem. *Machine learning*, vol. 47, pp. 235-256.

[7] Kocsis, L. & Szepesvari, C. (2006) Discounted UCB. *2nd Pascal Challenge Workshop*.

[8] Garivier, A. & moulines, E. (2008) On upper-confidence bound policies for non-stationary bandit problems. Available from http://arxiv.org/PS_case/arxiv/pdf/0805/0805.3415v1.pdf