Paper

Efficient wireless network selection by using multi-armed bandit algorithm for mobile terminals

Koji Oshima¹,²a), Takuma Onishi¹, Song-Ju Kim³, Jing Ma¹, and Mikio Hasegawa¹

¹ Department of Electrical Engineering, Tokyo University of Science
  6-3-1 Niijuku, Katsushika-ku, 125-8585, Japan
  ² Kozo Keikaku Engineering Inc.
  4-38-13, Hon-cho, Nakano-ku, 164-0012, Japan
  ³ Graduate School of Media and Governance, Keio University
  5322 Endo, Fujisawa, Kanagawa, 252-0882, Japan

a) koji.oshima.jp@ieee.org

Received May 7, 2019; Revised August 24, 2019; Published January 1, 2020

Abstract: Mobile wireless communication demand has increased rapidly in recent years, and advanced mobile devices such as smartphones have been widely deployed. Those devices are generally capable of connecting various wireless networks such as cellular networks and wireless LANs. Due to the software and hardware constraint of mobile devices, it is necessary to develop efficient and light-weight algorithms to select wireless networks. Previous studies have shown that multi-armed bandit algorithms can efficiently select wireless channel in cognitive radio. In this paper, we propose an efficient wireless network selection technique by using a multi-armed bandit algorithm called tug-of-war (TOW). We implement the proposed algorithm to a wireless device and show the effectiveness of the proposed method by experimental demonstrations.

Key Words: heterogeneous wireless network, cognitive radio, multi-armed bandit algorithm, tug-of-war model

1. Introduction

Wireless traffic and the number of wireless communication devices have increased rapidly in recent years. However, the frequency bands suitable for current technologies have already been exploited; thus, the resources are limited. Cognitive radio technologies [1, 2] have recently been developed to improve the radio resource usage of wireless networks under such situations. The basic concept of cognitive radio technology is the adaptation of the behavior of wireless systems through the recognition and learning of the radio environment. Cognitive radio systems observe and recognize the wireless network environment, make decisions, and take appropriate actions. Using this approach, various radio parameters can be optimized for the demand of wireless communications.
There are two types of cognitive radio systems; a spectrum sharing cognitive radio and a heterogeneous cognitive radio. As for spectrum-sharing cognitive radio, cognitive users utilize vacant frequency bands to improve radio resource usage. An effective channel sensing scheme is necessary for cognitive users, since the vacant radio resources change dynamically due to other users’ resource usage. Lai et al. modeled a cognitive radio as a multi-armed bandit (MAB) problem [3, 4]. The multi-armed bandit problem maximizes the total rewards provided from slot machines by optimizing the selection of slot machines that probabilistically provide rewards. In their model, channel selection of a cognitive radio is defined as a MAB problem under the assumption of probabilistic vacancy of each channel.

On the other hand, recent wireless devices are equipped with multiple wireless interfaces like smartphones. Various wireless networks such as 3G, LTE and Wi-Fi are available. Moreover, like access points in Wi-Fi, a mobile terminal can choose from multiple access networks. Ideally, in such a heterogeneous network, a user may choose the best wireless network through gathering various information on each network. Several literatures have studied heterogeneous wireless network selection. There are two types of approaches: network-centric and user-centric. In the network centric approach [6–8], the decision of the selection is done at a central controller. It assumes the detail information of each network status is available at the central controller. However, it is difficult to exchange information among various wireless networks which are operated independently. Therefore, in this paper, we focus on the selection of wireless networks at the mobile terminal.

Essentially, the traffic in the network is not static, nor stable. Moreover, the radio environment at the mobile terminal is unstable, especially in heterogeneous environment due to the interaction among various radio systems. The mobile terminal should seek the better solution as much as and as fast as possible under the probabilistically-changing environment. On the other hand, in many cases, the mobile terminal is forced to make “trial-and-error” decisions based on limited information of networks due to both software and hardware limitations. These constraint and requirement are closely similar to those in the MAB problem. Therefore, the MAB problem approach can be the aid of heterogeneous network selection.

Several studies investigated the wireless network selection at the mobile terminal [9–12]. Most of them needs various information of networks, or computational capacity. Authors have studied the optimizations of wireless systems based on cognitive cycle using machine learning [13]. In [13], mobile terminals gather various information of radio environments and communication performance, and build the performance model by machine learning. However, it is not always possible for mobile devices to gather information of all networks, nor to spare battery resources for complex calculations. It is important for the heterogeneous network selection to seek the better solution as much as and as fast as possible under the limited information about the networks. At the same time, it is also important for mobile devices to suppress the complexity of calculations for making decisions. From this point of view, in this paper, we use a novel algorithm called tug-of-war (TOW) to solve the MAB problem of heterogeneous network selection. The TOW algorithm has approximately highest performance as other MAB algorithms, whereas the implementation is very simple. In [5], the channel selection in wireless LANs using the TOW algorithm is proposed. In [21], the TOW algorithm is applied for massive IoT cognitive devices. These work show the effectiveness of the TOW algorithm in dynamic spectrum-sharing cognitive radio. However, the application for heterogeneous cognitive radio has not been studied yet.

As a summary, in this paper, we propose an efficient wireless network selection technique for heterogeneous cognitive radio using the tug-of-war MAB algorithm, and show the experimental evaluation. After the formulation of the algorithm is shown, an implementation and experimental results using a wireless device with Wi-Fi and LTE are shown.

2. Heterogeneous network selection as a multi-armed bandit problem
The major challenges in the selection of wireless network at the mobile terminal in heterogeneous environments are below:

• Making efficient decision under the situation where few information of each network is available.
• Practical algorithm which can be implemented on resource-constraint mobile devices.

To overcome these issues, several studies investigated algorithms and performances of them. In [9], a non-cooperative game formulation and analysis were given for Wi-Fi and cellular network selection on the mobile terminal. Results of computer simulations showed that the game can converge to Nash equilibria. However, the assumption that the mobile terminal can get the information of other mobile user is not always possible. In [12], a reinforcement learning solution and simulation analysis were given for heterogeneous cellular networks. Even though the simulation results showed the convergence speed and the suppression of overheads, it requires feedback information from the networks, which is only feasible in the cellular networks. It is important for the mobile terminal in heterogeneous network to select the better network without the coordination from the networks. At the same time, it is also important for mobile devices to suppress the complexity of calculations for making decisions. These constraint and requirement are closely similar to those in the MAB problem.

The multi-armed bandit (MAB) problem [14] is a simple machine learning problem, where a player attempts to obtain the maximum reward from multiple slot machines. The aim of the MAB is to decide which slot machine should be selected in order to obtain maximum reward through finite trials. The assumption is that the player does not have any prior information on each slot machine. The player starts to gather information on each slot machine through trying as many slot machines as possible. Then the player estimates which slot machine has the highest expected reward, and select that slot machine to play. Through this process, the player gets more rewards. There is a trade-off between exploitation and exploration. If the player takes long time for estimation, the player can estimate the reward more precisely, though the time for play the selected slot machine becomes short. If the player takes only short time for estimation, the player can take long time to play the selected slot machine, though the reward of that slot machine might be low.

![Fig. 1. Network selection in cognitive radio as a multi-armed bandit problem.](image)

In this paper, we apply the MAB problem to the wireless network selection in a cognitive mobile terminal. Figure 1 shows the model. The cognitive mobile terminal, as a player in the MAB problem, has \( n \) available wireless networks. Wireless networks correspond to the slot machines in the MAB problem. The terminal maintains an estimator \( Q_i \) of each wireless network \( i (i \in \{1, \ldots, n\}) \) according to the obtained rewards of the wireless network. The rewards can be anything such as throughput, delay, or other observables of the network performance, though the overheads to obtain them is preferable to be small. When the terminal communicates, it selects the network which can be expected highest rewards based on the estimator of each wireless network. To solve the MAB problem, we use a novel algorithm called tug-of-war (TOW) model described in the next section.

### 3. Multi-armed bandit algorithm

To solve MAB problems, several algorithms have been proposed [15–17], such as \( \epsilon \)-greedy algorithm, softmax algorithm, and UCB1-tuned algorithm. Although the UCB1-tuned algorithm is known as the best algorithm among parameter-free algorithms, the tug-of-war (TOW) model [18–20] has approximately the same performance as the UCB1-tuned algorithm [18]. Moreover, the TOW algorithm is very simple and the calculation cost is low: it has almost only addition and subtraction. Therefore,
the TOW algorithm is suitable for solving the problem of decision-making in cognitive mobile terminals which have the hardware and software limitations. The TOW algorithm is successfully used in cognitive dynamic channel selection [5] and cognitive massive IoT cognitive devices [21]. Therefore, the TOW algorithm is fitted for solving the problem of decision-making in heterogeneous cognitive radio.

3.1 Tug-of-war model

The tug-of-war (TOW) model is a multi-armed bandit algorithm inspired by the behavior of the amoeboid organism [18–20]. Unlike other algorithms for estimating the reward probability of each slot machine, the TOW dynamics uses a unique learning method which is equivalent to updating all machines’ estimates simultaneously based on the volume conservation law. In the TOW algorithm, the decision is made according to the displacements of the imaginary volume-conserving objects, which increase or decrease along with rewards, as shown in Fig. 2. The TOW algorithm of two machines are formulated as below. Imagine that the player plays a slot machine A or B at a time. When playing the machine A, if the player gets rewards, then 1 is added to an estimator $Q_A$, otherwise, $\omega$ is decreased (punishment). $\omega$ is a weighting parameter to be described below. After playing time step $t$, the displacement of machine A, $X_A (=-X_B)$, is expressed as followings:

$$X_A(t+1) = Q_A(t) - Q_B(t) + \delta(t),$$  \hspace{1cm} (1)

$$Q_A(t) = N_A (t) - (1 + \omega)L_A(t),$$  \hspace{1cm} (2)

where $\delta(t)$ is a fluctuation, $N_i(t) (i \in \{A, B\})$ counts the number of times that machine $i$ has been played until time $t$, $L_i(t)$ counts the number of punishments when playing machine $i$ until time $t$.

Let probabilities of providing rewards in the machines $i$ be $P_i$. Considering the ideal situation where the sum of the reward probabilities $\gamma = P_A + P_B$ is known to the player, the expected reward $Q'_i$ is given as

$$Q'_A = N_A \frac{N_A - L_A}{N_A} + N_B \left( \gamma - \frac{N_B - L_B}{N_B} \right) = N_A - L_A + (\gamma - 1)N_B + L_B,$$$$

Q'_B = N_A \left( \gamma - \frac{N_A - L_A}{N_A} \right) + N_B \frac{N_B - L_B}{N_B} = N_B - L_B + (\gamma - 1)N_A + L_A.$$$$

(3)

If we define $Q''_i = Q'_i / (2 - \gamma)$, we can obtain the difference of estimates in ideal situation as:

$$Q''_A - Q''_B = (N_A - N_B) - \frac{2}{2 - \gamma} (L_A - L_B).$$  \hspace{1cm} (5)

On the other hand, the difference of $Q_A$ and $Q_B$ from Eq. (2) is given by
\[ Q_A - Q_B = (N_A - N_B) - (1 + \omega) (L_A - L_B). \]  

From above two equations, we can obtain the nearly optimal weighting parameter \( \omega \) in terms of \( \gamma \) as:

\[ \omega = \gamma \left(1 - \gamma \right) / 2. \]  

If the number of machines is \( n (n > 2) \), \( \omega = \gamma / (2 - \gamma) \) is given by \( \gamma = P_1 + P_2 \), where \( P_1 \) and \( P_2 \) are the first- and second- highest reward probabilities, respectively [19]. Then Eq. (1) can be expressed as:

\[ X_i(t) = Q_i(t) - \frac{1}{n-1} \sum_{k=1; \neq i}^{n} Q_k(t) + \zeta_i(t), \]  

where \( \zeta \) is a fluctuation for each slot machine. The player selects the machine which has the highest \( X_i(t) \). We use the following \( \zeta_i(t) \) in this paper:

\[ \zeta_i(t) = A \cos \left( \frac{2\pi t}{n} + \frac{2(i-1)\pi}{n} \right), \]  

where \( A \) is the amplitude of the fluctuation.

4. Application of the MAB algorithm to wireless network selection

We propose the wireless network selection technique based on the MAB problem. As described in the previous section, we use the TOW algorithm to solve the MAB problem. Figure 3 shows the concept of proposed wireless network selection. The mobile terminal, capable of connecting various kinds of wireless networks \( S_i (i \in \{1, 2, \ldots, n\}) \), has the function of TOW algorithm. It observes the performance of the network \( Z_i \), where \( i \) is the selected network. \( Z_i \) can be any indexes of performance such as throughput, delay, or other metrics of wireless networks. The TOW algorithm then judges whether the reward or the punishment is to be added, by evaluating the obtained performance of network \( i \). If the reward is given, it updates the estimator as \( Q_i + 1 \), otherwise, updates the estimator as \( Q_i - \omega \). Then, \( X_1, \ldots, X_n \), the displacement of each network, are updated as described in Eq. (8). Note that all \( X_j (j \in \{1, 2, \ldots, i, \ldots, n\}) \), not only the selected network \( i \), are updated here. Also, \( P_i \) for \( \gamma \) in Eq. (7) is unknown, so they are estimated each time when the network \( i \) is selected as \( P_i = (N_i - L_i) / N_i \). The algorithm in the mobile terminal is as follows:

1. Start observing the performance of each wireless network \( i \). All networks are monitored at least once.

2. Update \( Q_i, P_i \), and all \( X_j \) based on the observation of network \( i \) by Eq. (2) and Eq. (8).

3. Select a wireless network \( i^* \) with highest \( X_i \).

![Fig. 3. Concept of the proposed wireless system which selects the wireless network based on the MAB algorithm.](image-url)
4. Observe the performance of the selected network and decide whether the reward or the punishment is to be given.

5. Back to 2 and continue.

5. Implementation and experiments

5.1 Implementation of the proposed scheme

In order to validate the proposed method in heterogeneous wireless network environment, we implement the proposed algorithm to a wireless device and perform experiments. The TOW algorithm is installed on Ubuntu Linux in Laptop PC as a cognitive mobile terminal, which is equipped both 802.11n/ac (2.4 GHz and 5 GHz) and LTE communication module. To judge the reward or punishment (1 or \(-\omega\)) in Eq. (2), we use average throughput of wireless networks as a threshold. If the observed throughput of wireless network \(i\) is larger than the average, then the reward is given for \(Q_i\), otherwise the punishment is given. Then the reward probability of selected network \(i\), \(P_i\), is updated.

We use the first- and second- highest reward probabilities among the networks as \(\gamma\) in Eq. (7). The experiments are conducted in and around the university building. The mobile terminal selects the wireless network from two Wi-Fi networks (2.4 GHz and 5 GHz) of the university infrastructure and LTE, to communicate to the server.

5.2 Experimental setup

We use iperf command to observe the throughput. The parameter \(A\) of the fluctuation in Eq. (9) is set to 5. Each iteration cycle has about 2 seconds. The locations of the experiments are (a) Laboratory room, (b) Inside the building, and (c) Outside. The average receive signal strength indicator (RSSI) of Wi-Fi is shown in Table I.

| Table I. Average RSSI of Wi-Fi. |
|-----------------------------|
| Location                   | Wi-Fi 2.4 GHz | Wi-Fi 5 GHz |
| (a) Laboratory room        | -36.0         | -37.0       |
| (b) Inside the building    | -39.0         | -76.0       |
| (c) Outside                | -73.0         | -81.0       |

5.3 Verification of the proposed scheme

Figure 4 shows an example of network selections of the proposed algorithm. The values \(X_i\) of the TOW algorithm in Eq. (8) are also shown. In each case, after the initial trial of all wireless networks, the selection of the wireless network is converged to the highest performance network. Note that the values \(X_i\) of unselected networks are also updated (decreased) through iteration. Even though the estimators \(Q\) of unselected networks in Eq. (2) are not updated, the displacements \(X_i\) of all networks are updated in Eq. (8). This is the unique characteristic of the TOW algorithm, as described in the previous section. In the case (a) and (b), the values \(X_i\) of Wi-Fi 5 GHz and 2.4 GHz, which have higher RSSI and throughput in laboratory room and inside the building, become larger according to the number of iteration, while \(X_i\) of unselected Wi-Fi and LTE become smaller. As a result, the selection is converged to Wi-Fi 5 GHz and 2.4 GHz. In the case (c), where the signal strength from Wi-Fi access point becomes much lower, the value \(X_i\) of LTE becomes larger according to the number of iteration, while \(X_i\) of unselected Wi-Fi 5 GHz and 2.4 GHz become smaller. As a result, the selection of the networks is converged to LTE.

5.4 Evaluation of the throughput performance of the proposed scheme

In order to verify the performance in heterogeneous environments, we examine the throughput in each place. Figure 5 shows the average throughputs of each wireless network and the proposed TOW algorithm. Each experiment is continued three times, and the throughputs shown here are the
The locations (c-1) and (c-2) are both outside the building but different in Wi-Fi traffic situation: (c-1) is more crowded. It is shown that the proposed TOW algorithm achieve as high average throughput among other wireless networks. In the laboratory room (a) and inside the building (b), the throughputs of the proposed system are as high as those of Wi-Fi in 5 GHz and 2.4 GHz, respectively. On the other hand, outside the building (c-1) and (c-2), where the signal strength from Wi-Fi access point becomes much lower, the throughput of the proposed system is as high as that of LTE. Moreover, as shown in the case (c-2), the proposed algorithm can achieve as high performance as LTE on average, where the differences in performance among all networks are rather small. An example of throughput variation by time is shown in Fig. 6. Throughputs when using only Wi-Fi 2.4 GHz, 5 GHz, or LTE are shown. The values are moving averages of 10 samples. This figure shows that sometimes the throughputs of Wi-Fi are higher than LTE, though the averages are lower than that of LTE. The result shows that the proposed algorithm can estimate the probabilities of rewards.
Fig. 5. Average throughputs of each wireless networks and the proposed TOW algorithm in different environments.

Fig. 6. An example of throughput variation by time in (c-2).

among the wireless networks properly.

6. Conclusion
Due to the advance in mobile wireless systems and the scarcity of the frequency spectrum, it is necessary for mobile devices to utilize wireless networks in a heterogeneous environment. Wireless network selection is one of the realistic problem in current mobile devices. Considering the limitation of hardware and software complexity in mobile devices, an efficient wireless network selection in a cognitive
way is important. The multi-armed bandit problem is a simple machine learning problem and applicable for cognitive radio problem. In this paper, we proposed a simple but powerful wireless network selection technique using the novel multi-armed bandit algorithm called tug-of-war. Through the implementation and experiments, the effectiveness of the proposed algorithm in a real heterogeneous wireless environment was shown. Since the proposed technique is based on a simple and light-weight algorithm, it can be applied to not only smartphones, but also IoT devices where the system resources are more limited. It is important to investigate the performance in real application services, and the candidates of reward indexes.

Acknowledgments
This work was supported by JSPS KAKENHI Grant Number JP17K00136.

References
[1] J. Mitola and G.Q. Maguire, “Cognitive radio: Making software radios more personal,” *IEEE Pers. Commun.*, vol. 6 no. 4, pp. 13–18, 1999.
[2] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” *IEEE J. Sel. Areas Commun.*, vol. 23 no. 2, pp. 201–220, 2005.
[3] L. Lai, H.E. Jiang, and H.V. Poor, “Medium access in cognitive radio networks: A competitive multi-armed bandit frame work,” *Proc. of IEEE 42 th Asilomar Conference on Signals, System and Computer*, pp. 98–102, 2008.
[4] L. Lai, H.E. Jiang, and H.V. Poor, “Cognitive medium access: Exploration, exploitation, and competition,” *IEEE Trans. on mobile computing*, vol. 10, no. 2, pp. 239–253, 2011.
[5] K. Kuroda, H. Kato, S.-J. Kim, M. Naruse, and M. Hasegawa, “Improving throughput using multi-armed bandit algorithm for wireless LANs,” *NOLTA*, vol. 9, no. 1, pp. 74–81, 2018.
[6] S. Singh and J.G. Andrews, “Joint resource partitioning and offloading in heterogeneous cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, 2014.
[7] C. Liu, M. Li, S.V. Hanly, and P. Whiting, “Joint downlink user association and interference management in two-tier HetNets with dynamic resource partitioning,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1365–1378, 2017.
[8] V. Sagar, R. Chandramouli, and K.P. Subbalakshmi, “Software defined access for HetNets,” *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 84–89, 2016.
[9] A. Keshavarz-Haddad, E. Aryafar, M. Wang, and M. Chiang, “HetNets selection by clients: Convergence efficiency and practicality,” *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 406–419, 2017.
[10] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, “RAT selection games in HetNets,” *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, pp. 998–1006, 2013.
[11] X. Wang, J. Li, L. Wang, C. Yang, and Z. Han, “Intelligent User-Centric Network Selection: A Model-Driven Reinforcement Learning Framework,” *IEEE Access*, vol. 7, pp. 21645–21661, 2019.
[12] D.D. Nguyen, H.X. Nguyen, and L.B. White, “Reinforcement learning with network-assisted feedback for heterogeneous RAT selection,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, 2017.
[13] K. Oshima, T. Kobayashi, Y. Taenaka, K. Kuroda, and M. Hasegawa, “Wireless network optimization method based on cognitive cycle using machine learning,” *IEICE ComEX*, vol. 7, no. 7, pp. 278–283, 2018.
[14] H. Robbins, “Some aspects of the sequential design of experiments,” *Bulletin of the American*, vol. 58, pp. 527–535, 1952.
[15] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, Massachusetts London, England, 1998.
[16] D. Daw, P. O’Doherty, P. Dayan, B. Seymour, and J. Dolan, “Cortical substrates for exploratory decisions in humans,” *Nature*, vol. 441, pp. 876–879, 2006.
[17] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” Machine Learning, vol. 47, pp. 235–256, 2002.

[18] S.-J. Kim and M. Aono, “Amoeba-inspired algorithm for cognitive medium access,” NOLTA, vol. 5, no. 2, pp. 198–209, 2014.

[19] S.-J. Kim, M. Aono, and E. Nameda, “Efficient decision-making by volume-conserving physical object,” New Journal of Physics, vol. 17, 083023, 2015.

[20] S.-J. Kim, M. Aono, and E. Nameda, “Decision maker based on atomic switches,” AIMS Material Science, vol. 3, no. 1, pp. 245–259, 2016.

[21] J. Ma, T. Nagatsuma, S.-J. Kim, and M. Hasegawa, “A machine-learning-based channel assignment algorithm for IoT,” Proc. ICAIIC, vol. 3, no. 1, pp. 245–259, 2016.