ISSN 2277-3061

Fast Iterative model for Sequential-Selection-Based Applications

Khosrow Amirizadeh, Rajeswari Mandava
Intelligent Systems Lab., School of Computer Sciences,
Universiti Sains Malaysia (USM), 11800 Penang, Malaysia
KhosrowAmirizadeh@yahoo.com,
mandava@cs.usm.my

ABSTRACT

Accelerated multi-armed bandit (MAB) model in Reinforcement-Learning for on-line sequential selection problems is presented. This iterative model utilizes an automatic step size calculation that improves the performance of MAB algorithm under different conditions such as, variable variance of reward and larger set of usable actions. As result of these modifications, number of optimal selections will be maximized and stability of the algorithm under mentioned conditions may be amplified. This adaptive model with automatic step size computation may attractive for on-line applications in which, variance of observations vary with time and re-tuning their step size are unavoidable where, this re-tuning is not a simple task. The proposed model governed by upper confidence bound (UCB) approach in iterative form with automatic step size computation. It called adaptive UCB (AUCB) that may use in industrial robotics, autonomous control and intelligent selection or prediction tasks in the economical engineering applications under lack of information.

Indexing terms/Keywords

Iterative MAB model; Fast action selection; Self-tuning of iterative algorithms; Step-size free adaptive algorithm.

Academic Discipline and Sub-Disciplines

Computer sciences – Machine learning – Artificial intelligence

SUBJECT CLASSIFICATION

Intelligent control / Robotic/ Machine Learning / Artificial intelligence

TYPE (METHOD/APPROACH)

Steepest Descent Optimization method

ISSN 2277-3061

INTRODUCTION

A growing number of models in autonomous and adaptive control applications operating based on intelligent learning approaches to make “sequential decisions” tasks. These prove a truly fundamental enhance from traditional control process to intelligent approaches. In most cases, these approaches should be able to perform sequential decision making with long control horizons that the exploration and exploitation trade-off is inherently considered. Recently, subjects such as “iterative learning control and reinforcement learning” in adaptive control and robotics, autonomous agents and intelligent decision making have widely developed. In this regard, multi-armed bandit (MAB) structure plays an important role in real world applications. With a simple search on the web, we may face with subjects such as, bandit algorithms for limited feedback learning, contextual bandits and optimal decision making, Bayesian reinforcement learning and optimal control for uncertain models. Ease to implementation is the main reason for this expansion however; the decision maker should never become unsafe in uncertain situations for this simplicity. For these reasons, necessity to move from traditional approaches in “sequential selections” problems to new approaches with rely more on-line, adaptive and autonomous is unavoidable [17, 19].

Because of popularity of multi-armed bandit (MAB) model in sequential decision making, some variant of this model have been used in different applications. Distributed networks operations [11, 15], industrial decision making [16], software engineering [13], games industries [14] and robotics [10, 18, 19, 20] are some samples of recent real-world applications of MAB. In these applications, the main objective is to compute accurate estimation of the “actual value” of each interest options that often referred as “usable plans", "solutions”, “operational actions”, "useable resources", etc and then, take the most valuable of them, at each step.

Multi-armed bandit (MAB) presents a general structure to perform sequential selections under insufficient information condition. At each step, the decision maker takes an option and receives a value as “reward”. This reward is taken from a distribution according to each action. The structure is similar to one state Markov decision process. This sequence should be repeated until agent/decision maker reaches the acceptable level of intelligence to make optimum decisions in the future. Since, at the beginning of the process, the agent is not enough intelligent, some selections should be taken randomly. Thus, the decision maker strikes to establish a good balance between two major tasks ” exploitation” and “exploration” to gain up the long term benefits.

There are several approaches to formulate MAB model but a general model that establishes optimal balance between the exploration and the exploitation tasks, in general, is scarce [3, 5]. The simplest and most popular model is known by ε-greedy family model. According to this approach, the agent takes, mostly, an action with higher “current value estimation” with probability 1- ε, greedy, and sometimes explores them with probability ε, randomly. The selection criterion in this category is the “action with higher value”. For maximizing the long-term benefit, one can optimally define the exploration rate ε. It may be fixed or computed dynamically according to the agent evolution; however, this computation is not a straightforward task and indicates a limitation in this category.

A wide range of theoretical MAB studies in statistics and decision making domains are presented based on defining an upper confidence bound (UCB). These models consider an extra term (i.e. a function of variance and exploration rate parameter) plus current value estimation to make the selection criterion. At each step, the agent selects the most valuable action with respect to this criterion. For example, authors in [6, 7] presented some models that have known as the main structures of UCB approach. However, UCB models are depended on the reward variance and consequently, they degrade with non-stationary observations. This is another limitation in MAB algorithms highlighted in UCB family approaches [4, 5].

Another approach has been presented in [3]. Authors introduced an online mirror descent (OMD) algorithm which operates based on Gradient descent technique. The goal is minimizing the total loss incurred. OMD uses a step size that plays a critical role in this iterative model. Similar to other Gradient-based approaches, OMD needs to step-size-tuning in each condition.

These mentioned limitations are usually noted in empirical evaluation under different settings. In this regard, some empirical comparisons have been conducted by Kuleshov and Precup [5]. The authors concluded models with good theoretical guarantees, sometimes, do not operate well in real-world applications. Furthermore, the presented models are usually addressed MAB problem with small set of actions under low variances.

This study aims to evaluate UCB approach of MAB model under different conditions and present an “iterative MAB algorithm” based on UCB approach to minimize the mentioned limitations. In this iterative model an automatic computation of step size is applied to eliminate the "parameter dependency" problem and increase “number of optimal selection” which is useful for on-line sequential decision making tasks.

The paper is organized as follows: a brief description of MAB problem and its mathematical model are presented in the next section. The proposed adaptive MAB model, AUCB, is expressed in section 3. Implementations and comparisons in section 4 exhibit the performance of new algorithm and finally, the paper ends with conclusion section.

Multi-armed bandit problem (MAB) and the solutions

Background and the basic mathematical model

MAB is a framework to study a learning task where an agent is expected to make successive selections without any knowledge about the reward of the selection made. In the general case, it contains a set of different options and a set of

rewards relating to each selection. The decision maker faces a row of these options, without any extra knowledge to indicate the prominent one, and decides which one must be selected such that, the total reward is maximized. Maximizing this cumulative reward is equivalent to minimizing the regret, the difference between true cumulative reward and sum of so far rewards relating to the best selection at each round. These may be formulated as: let $A = \{a_1, a_2 \dots a_N\}$ be a set of N usable action/option which have a set of probability distributions with expected values of $\{\mu^*_{a_1}, \mu^*_{a_2} \dots \mu^*_{a_N}\}$ and variances $\{\sigma_{a_1}, \sigma_{a_2} \dots \sigma_{a_N}\}$. The observations follow an independent and identical distribution (i.i.d). In non-stationary situation, the mean may vary with time. After choosing k_a times an action a, instant estimation of "actual value", $V^*(a)$ at step k is obtained through the sample-mean equation:

$$V_k(a) = \frac{1}{k_a} \sum_{i=1}^{k_a} r_i(a) \quad (1)$$

Where, $r_i(a)$ is the instant reward at each step i. Finding the best estimation of $V^*(a)$ is the main objective. Here, some samples of different categories are selected and empirically evaluated under different settings such as: increasing number of options and operating under different variances of observations. The comparisons are based on the percentage of optimal selections, the stability under different variances and the performance algorithms with larger set of options. The UCB approaches such as, UCB-1, UCB-V and UCB-Tuned are selected to be compared and evaluated with the proposed model AUCB.

UCB family approaches

Authors in [5, 6, 7, 9] have considered models that operate with an upper confidence bound (UCB) criterion to select the prominent option at each step, to minimize long-term regret value. These models have been derived by an objective functions that consider the total cumulative regret error which is formulated as:

$$CumReg(V) = \sum_{i=1}^{k} V_i^* - \sum_{i=1}^{k} V_i(a^*), \quad (2)$$

Where $V_i^* = max_a V(a)$ and $V_i(a^*)$ is the current value of winner action $a^* = argmax_a V(a)$ at step i. In this case, the goal is to minimize the expected regret $CumReg(V)$. UCB-1 uses a straightforward routine to take the winner action at every step. It considers the number of times that an arm has been selected after k rounds, namely k_a as well as the expected mean thus, the arm that maximizes the following criterion is selected.

$$a_k^* = argmax_a \left(V_k(a) + \sqrt{\frac{2ln\ k}{k_a}}\ \right) \quad (3)$$

Authors in [7] presented another criterion that considers empirical variance as well as a function of an exploration probability rate at each step. The simplified criterion is:

$$a_k^* = argmax_a \left(V_k(a) + \sqrt{\frac{2\ \sigma_k^2(a)\mathcal{E}(k)}{k_a}} + C1\left(\frac{3b\mathcal{E}(k)}{k_a}\right)\right) \quad (4)$$

Where $\mathcal{E}(k) = C_2 \log k$ is the exploration rate function for each action at each step and $C1, C2 \geq 0$, while the variance of observations must be in the domain [0 b]. This is called UCB-V. The variance is computed experimentally through $\sigma_k^2(a) = \frac{1}{k}\sum_i^k (r_i - V_i(a))^2$. Another approach, UCB-Tuned [5, 6], uses empirical variance with respect to a boundary 0.25 at each step. The criterion is:

$$a_k^* = argmax_a \left(V_k(a) + \sqrt{\left(\frac{ln\ k}{k_a}\right) \min(0.25\ , W_k(a))}\right) \quad (5)$$

H tshankar@vit.ac.inre, $W_k(a) = \sigma_k^2(a) + \sqrt{\frac{2ln\ k}{k_a}}$ is computed based on the current estimation of the variance [7].

These mentioned criteria select the optimum action/arm and make the relevant models UCB-1, UCB-V and UCB-Tuned. After each selection, the agent receives a reward and updates its value based on a uniform iterative structure that stated in next section.

The focus of this study is presenting iterative models with higher stability for on-line applications that operate under variable observations. Next section presents the new models and theirs specifications.

Proposed adaptive model AUCB

The following stochastic value function estimation is introduced to estimate "actual value" of an action a at time step k.

$$V_{k+1} = V_k + \eta_k\ [r_{k+1} - V_k] \quad (6)$$

Here, V_{k+1} is the "estimated value" associated to a selected action a. After each selection, the agent receives reward r_{k+1} and updates its "value" based on this equation. The term in the bracket is the temporal difference error. Sequence $\{\eta_k\}$ is a series of positive scalar gains or the step sizes taken from domain $0 < \eta_k \leq 1$. It plays important role in this iterative equation. Convergence of the Eq. (6) is guaranteed while the step sizes follow assumptions $\sum_{k=0}^{\infty} \eta_k = \infty$ and $\sum_{k=0}^{\infty} \eta_k^{r>1} < \infty$ [8]. With both stationary and non-stationary observations, step sizes should be precisely defined to compute the optimal performance. Authors in [2] showed that for non-stationary observations, the step sizes may be defined by a monotonically decreasing value at each time step. A general form that may be used is $\eta_k = {}^{n_1}/_{k^{n_2}}$, where $0 < (n_1, n_2) \leq 1$.

An iterative model with automatic step size computations technique may assist MAB algorithm to maintain its performance under variable variances or non-stationary observations. Assume that minimization the mean square error function $J(V_k) = \frac{1}{2} E[(V^* - V_k)^2]$ is the objective, where V^* is the "actual value" and V_k is the "estimated value" of an option a at step k. The Gradient descent approach introduces the path to the optimum point through the equation $V_{k+1} = V_k + \eta_k \nabla \mathrm{J_k}(V)$ where $\nabla \mathrm{J_k}(V) = (r_{k+1} - V_k)$. In TD learning, decision maker does not know the "actual value"; V^* instead, the approach operates with a temporal estimation, R_k that is, the expected mean of rewards at current step. It means that, $\lim_{\mathrm{k}\to\infty} \mathrm{E}[r_k] = \lim_{\mathrm{k}\to\infty} R_k = V^*$ and it is clear that, the estimation of value function, iteratively, is computed in V_k. In the steepest decent approach we may optimize the objective function with respect to the step size as:

$$\frac{\partial J\,(V_{k+1})}{\partial \eta_k} = \frac{\partial J}{\partial \eta_k} (E[\,(V^* - (V_k + \eta_k (r_{k+1} - V_k))^2]) = 0 \tag{7}$$

$$\frac{\partial J\,(V_{k+1})}{\partial \eta_k} = E[\big((V^* - V_k) - \eta_k (r_{k+1} - V_k)\big)\ (r_{k+1} - V_k)] = 0 \tag{8}$$

$$(E[\,V^*] - V_k)(r_{k+1} - V_k) - \eta_k (r_{k+1} - V_k)(r_{k+1} - V_k) = 0$$

$$(R_k - V_k)(r_{k+1} - V_k) = \eta_k (r_{k+1} - V_k)(r_{k+1} - V_k)$$

Finally, after reordering the elements, the step size at each step is computed by:

$$\eta_k = \frac{R_k r_{k+1} - R_k V_k - r_{k+1} V_k + V_k V_k}{r_{k+1} r_{k+1} - 2 r_{k+1} V_k + V_k V_k} \tag{9}$$

Automatic step size will be computed through a function of current reward, the value function and expected reward. The simplicity of the model for implementation and operating without any "parameter dependency" are two important advantages. In order to compute Eq. (9), the current R_k is iteratively estimated using the following stochastic equation:

$$R_k = R_{k-1} + \left(\frac{k_a}{k_a + 1}\right).(r_{k+1} - R_{k-1}) \tag{10}$$

The parameter k_a is the number of action a has been selected. The sequence $\{k_a/(k_a + 1)\}$ causes the error $(r_{k+1} - R_{k-1})$ damp to zero. This sequence in stationary situation may be changed to $\{1/k_a\}$ for best step size in stationary conditions. Based on computations in Eq.(9) and Eq.(10), at each step, optimal step size in iterative MAB model is computed. In addition, new MAB model is introduced by applying Eq. (9) and Eq. (10) in Eq. (6). Algorithm 1 shows pseudo code AUCB model.

Algorithm 1: Adaptive UCB model (AUCB).

1) For $k = 1$ to Plays
2) Select a_k^* ; // UCB-V approach Eq. (4)
3) Receive reward $r_{k+1}(a_k^*)$;
4) Compute $R_k(a_k^*)$; // Eq. (10)
5) Compute $\eta_k(a_k^*)$; // Eq. (9)
6) Update $V_k(a_k^*)$; // Eq. (6)
7) End

The proposed model AUCB, based on adaptive step size calculation in iterative MAB algorithm is presented. This is the major advantage of the new method and it may be useful for on-line applications. Experimental results will be presented in the next section, indicating the performance of proposed iterative model. The approach is not limited to reinforcement learning; it may be applied in adaptive models in control engineering, signal processing and pattern recognition to maintain their performance under non-stationary or variable stationary observations.

Experimental results and discussion

In order to evaluate the proposed model AUCB, some comparisons based on behavior of all mentioned models with different reward variances are considered. Each model run for 1000 plays and this task is repeated for 1000 times to get an appropriate average over these independent runs. The number of actions is N=5 that increases to N=10, 20, 40 and

60. All random rewards are taken from a normal distribution with mean $\mu^* \sim N(0,1)$ and the standard deviation $\sigma_{rew} \in \{0.01, 0.1, 1, 3\}$. Thus, the reward function is $rew(a) = \mu^* + \sigma_{rew} * Rand$. The $Rand$ function gives a random number from a normal distribution with mean zero and standard deviation one. Procedures use the general step size function as $\eta_k = 1/k^n$. Appropriate settings are separately defined as, UCB-1 (n=0.5), UCB-Tuned (n=0.6), UCB-V (n=0.4).

After each selection, the "optimal action" and the "selected action" are compared, and the number of correct selections is considered to plot the percentage of optimal selections. Fig.1 shows this quantity while low variances of observations are used. Usually, UCB family models have better performance than e-greedy algorithms in stationary variance conditions due to using complex criteria which are stated in the section "Background".

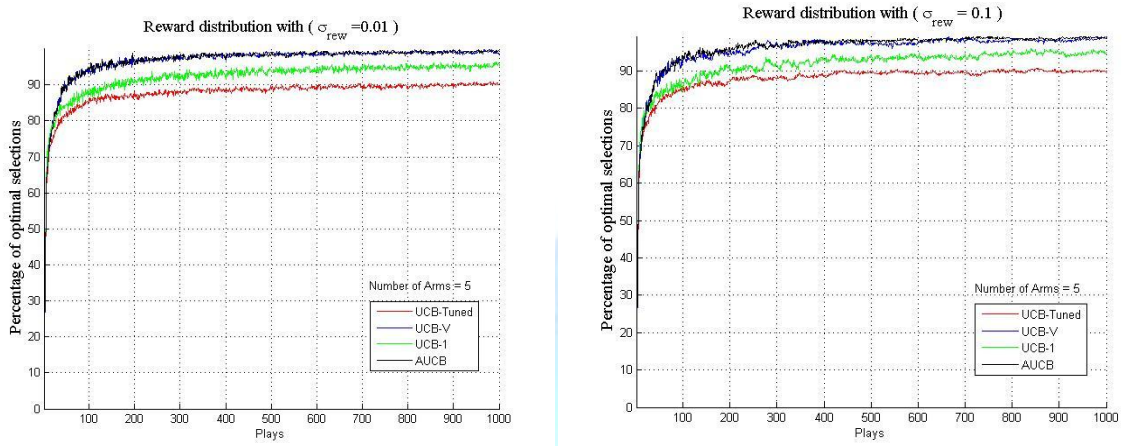

Figure 1: percentage of optimal selections with variance 0.01(left) and 0.1(right). UCB models operate well under lower variances as indicated. UCB-V and AUCB perform better and have higher curves (two above overlapped curves). All models are run with iterative structure stated in Eq.(6) to estimate value of each option, iteratively, which is useful for on-line adaptive sequential selection tasks. AUCB operates similar to the best MAB algorithm without any parameter setting.

Most bandit algorithms are depending on the variance of the reward [5]. It is important to know that which one is less sensitive. To gain an insight into this, we increase the variance to 1 and 3. The curves relating to percentage of optimal selections with high variances are plotted in Fig.2. In lower variances cases, Fig.1, UCB-V and AUCB select optimal objects more than 95%, quickly, while, in higher variances cases, as plotted in Fig. 2, only AUCB indicates the best performance, it means that the new structures may be useful to compensate the weakness under variable observations and is useful both in low and high variance conditions without any extra tuning.

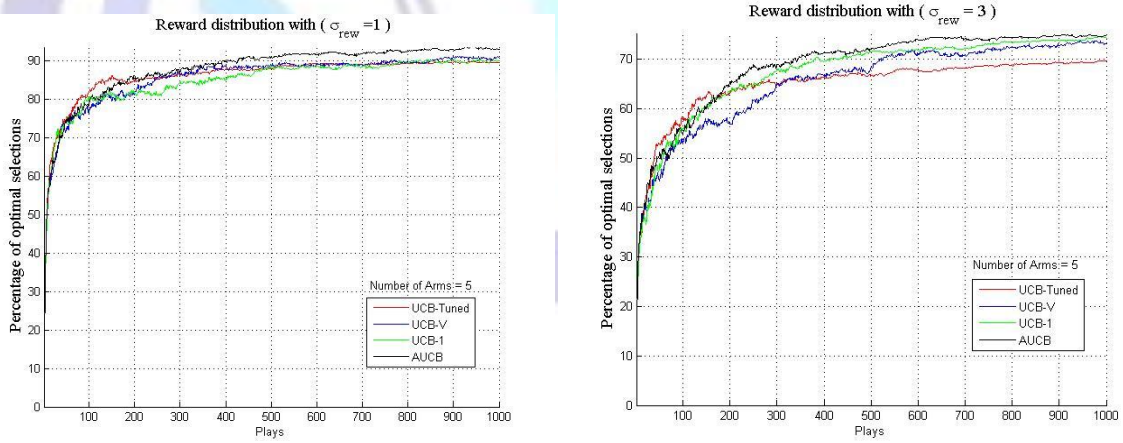

Figure 2: percentage of optimal selections under high reward variances 1(left) and 3(right) in stationary observation cases. AUCB is more tolerable under low and high variances. Among UCB models, UCB-1 has better performance in high variances, however, the new model, AUCB performed well in both cases.

The performance bandit algorithms degrade with larger set of actions [4, 5]. Figure 3 illustrates the performance all models with a larger set of options.

ISSN 2277-3061

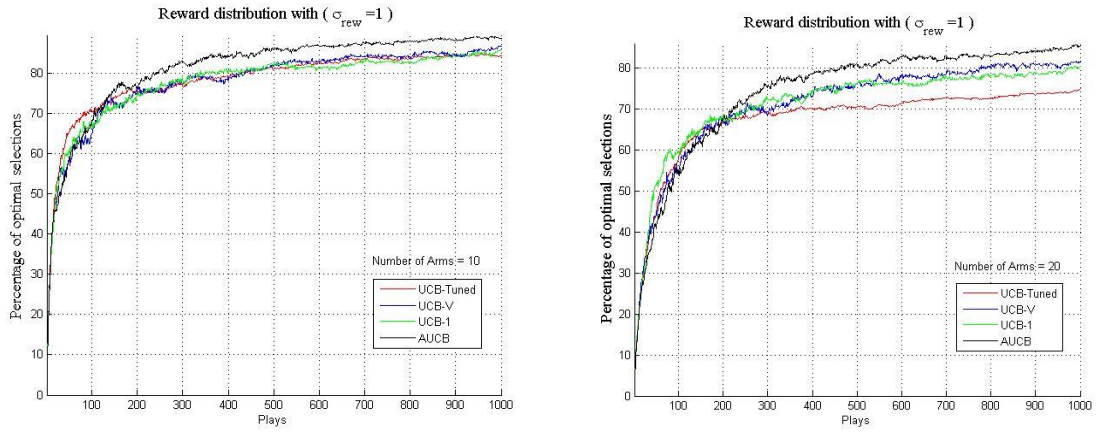

Figure 3: percentage of optimal selections with N=10 (left) and with N=20(right). Models, often, degrade with increasing the number of actions. Comparing these plots with Fig.2, it is noticed that, UCB models select about 90% percent optimal actions, while here, in the similar observations, all models lose the performance. AUCB is more stable with this condition.

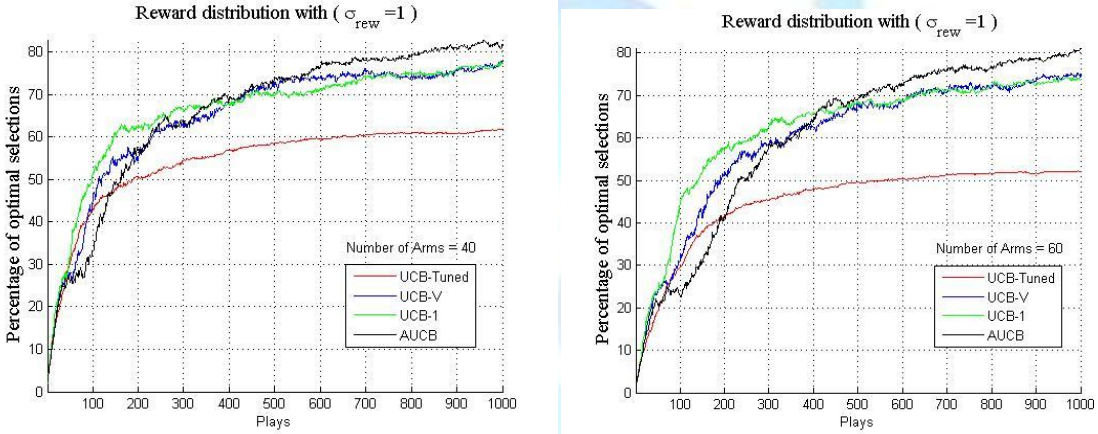

Figure 4: percentage of optimal selections with larger set of options. Left graph is resulted by N=40 and right is performed by N=60, both in stationary observation cases. Most UCB models degrade while UCB-Tuned failed. Still AUCB is more stable than other UCB models without any step-size tuning.

Conclusion

In this article, the iterative MAB model with automatic step size computation is presented. It called AUCB. The iterative structure with automatic computation of step size may present a more stable structure under different observations and conditions. Experimental results indicated these modifications may improve the performance MAB algorithms under different variances of observations and larger set of options. This iterative structure is important for on-line sequential decision making tasks where, automatic tuning of step sizes is the main concern. Some comparisons with different settings have been conducted to show the performance AUCB under variable observations, whereas similar models degrade under these conditions. It means the iterative model with automatic computation of step size may amplify the stability of the algorithms under these conditions. Variable variances and using larger set of actions are the set of concerns that may lose the efficiency of MAB algorithms. These results indicate that UCB approaches are depending on low stationary variances. AUCB may compensate this weakness, while it does not require any parameter tuning. In total, percentage of optimal selections is enhanced and stability under different variances is maximized. These are two major objectives due to applying AUCB model.

References

[1] Sutton, R. S., Barto, A. G. 1998. Reinforcement Learning: An introduction, Cambridge, MA: MIT Press.

[2] George, A.P., Powell, W.B. 2006. Adaptive step sizes for recursive estimation with applications in approximate dynamic programming. Journal of Machine Learning 65(1), 167-198.

[3] Bubeck, S., & Cesa-Bianchi, N. 2012. Regret analysis of stochastic and non stochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

[4] Vermorel, J., & Mohri, M. 2005. Multi-armed bandit algorithms and empirical evaluation. In *Machine Learning: ECML 2005* (pp. 437-448). Springer Berlin Heidelberg.

[5] Kuleshov, V., Precup, D. 2010. Algorithms for the multi-armed bandit problem, Journal of Machine learning research, 1, 1-48.

[6] Auer, P., Cesa-Bianchi, N., & Fischer, P. 2002. Finite-time analysis of the multi armed bandit problem. *Machine learning*, *47*(2-3), 235-256.

[7] Audibert, J. Y., Munos, R., & Szepesvári, C. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*(19), 1876-1902.

[8] Benveniste, A., Metivier, M., Priouret, P. 1990. Adaptive Algorithms and Stochastic Approximations, New York: Springer

[9] Audibert, J., Bubeck, S., Munos, R. 2010. Best arm identification in multi-armed bandits, in the proceeding of COLT 2010.

[10] Pini, G., Gagliolo, M., Brutschy, A., Dorigo, M., & Birattari, M. 2013. Task partitioning in a robot swarm: a study on the effect of communication. *Swarm Intelligence*, 1-27.

[11] Liu, K., & Zhao, Q. 2012. Adaptive shortest-path routing under unknown and stochastically varying link states. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on* (pp. 232-237). IEEE.

[12] Fialho, Á., Da Costa, L., Schoenauer, M., & Sebag, M. 2009. Dynamic multi-armed bandits and extreme value-based rewards for adaptive operator selection in evolutionary algorithms. In *Learning and Intelligent Optimization* (pp. 176-190). Springer Berlin Heidelberg.

[13] Scott, S. L. 2010. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, *26*(6), 639-658.

[14] Granmo, O. C., & Glimsdal, S. 2013. Accelerated Bayesian learning for decentralized two-armed bandit based decision making with applications to the goore game. *Applied Intelligence*, 1-10.

[15] Gai, Y., Krishnamachari, B., & Jain, R. 2010. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on* (pp. 1-9). IEEE.

[16] Dinesh Kumar, U., & Saranga, H. 2010. Optimal selection of obsolescence mitigation strategies using a restless bandit model. *European Journal of Operational Research*, *200*(1), 170-180.

[17] Kumar, P. R. 1985. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, *23*(3), 329-380.

[18] Schembri, M., Mirolli, M., & Baldassarre, G. 2007. Evolution and learning in an intrinsically motivated reinforcement learning robot. In *Advances in Artificial Life* (pp. 294-303). Springer Berlin Heidelberg.

[19] Sigaud, O., & Peters, J. 2010. *From motor learning to interaction learning in robots* . Springer Berlin Heidelberg.

[20] Abbeel, P., Coates, A., Quigley, M., & Ng, A. Y. 2007. An application of reinforcement learning to aerobatic helicopter flight. *Advances in neural information processing systems*, *19*, 1.

ISSN 2277-3061

Author' biography with Photo

Khosrow Amirizadeh received his B.S. degree in computer engineering from Shiraz University in 1990 and his M.S. degree in Artificial Intelligence from IAU- Researches and Sciences in 1998, Tehran, Iran. Currently he is a "PhD candidate" at "Universiti Sains Malaysia (USM)" and working on Reinforcement learning and application in Neuroimaging tasks, Intelligent control and decision making.

Rajeswari Mandava received the M.Tech in 1980 from Indian Institute of Technology, Kanpur and PhD degrees in 1995 from University of Wales Swansea. Join USM in 1982. Her main research interest is to process, analyze and to extract contents and information from the images; derive knowledge from the extracted information; to represent the knowledge and use the knowledge in various applications in addition to using it to guide the information extraction from the images. In the early stages of this research the focus was to extract information from the images and put into several applications that include automated visual inspection, and real time process control in industry; robot vision for intelligent assembly; image database retrieval and image segmentation. The major domain of research is in medical images and natural images.