17th ICEE 2009
Iranian Conference on Electrical Engineering
www.icee.ir/2009
IUST

Sequential optimistic ad-hoc methods for nonstationary multi_armed bandit problem

Majid Mazouchi*, Farzaneh Tatari*, Mohammad B. Naghibi S.**

*Cognitive Computing Lab, Department of Electrical Engineering, Ferdowsi University of Mashhad, majid.mazouchi@gmail.com, fa_tatari@yahoo.com
**Department of Electrical Engineering, Ferdowsi University of Mashhad, mb-naghibi@ferdowsi.um.ac.ir

Abstarct: *One of the common ways for showing the trade_off between exploration_exploitation in reinforcement learning problems is the multi_armed bandit problem. In this paper we consider the MABP in a nonstationary environment which features change during the period of learning. The represented learning algorithms are intuition based solutions to the exploration_explotation tarde_off that are called ad hoc method. These methods include action_value methods with e-greedy and softmax action selection rules, the probability matching method and finally the adaptive pursuit method. For producing near optimal results we change the ad hoc methods to sequential optimistic ad hoc methods which provide us completely better results.*

Key words: Sequential optimistic ad hoc methods, Exploration_exploitation, Multi_armed bandit, Reinforcement learning, Action selection.

1. Introduction

In the original form of multi_armed bandit problem the agent is faced repeatedly with a choice among n different actions. After each choice agent receives a numerical reward based on the selected action. The objective of the agent is to maximize the total reward during a certain number of plays. One of the most interesting variant of MABP is the non_stationary bandit problem. In this kind of MABP the agent is located in an environment that generates rewards changing during the learning period, therefore the optimal action may change up to the last episode. Four well_known ad hoc algorithms exist that are well suited for non_stationary bandit problem. These learning algorithms do not balance exploration and exploitation in a sophisticated way but have the advantage of not relying on strong theoretical assumptions while in the same time can be fine-tuned to produce near-optimal results [4].

The paper is organized as follows. In the next section the problem definition is represented. In section 3 ad hoc techniques are introduced. In section 4 the results of ad hoc techniques applied to non_stationary MABP are discussed. In section 5 we represent sequential optimistic ad hoc methods and we consider their results in the next section. Finally a conclusion can be found in section 7.

2. Problem definition

The non_stationary MABP can be illustrated as follows. An agent is presented, n different actions, a_i, $i = 1,2,...,n$. The agent is repeatedly asked to select only one of the available actions for a finite number of episodes t_j, $j = 1,2,..., p$. For each action selection the agent is given a reward, $r_{t_j}(a_i)$, dependant on the selected action. The reward is chosen from the actual value of the actions which are initially unknown to the agent. At each episode the estimated value of the actions takes new value. The agent objective is to maximize the sum of the collected rewards J:

$$Max\ J = \max \sum_{j=1}^{p} r_{t_j} \qquad (1)$$

In this paper we consider an environment with 10 available actions. The time horizon is set to 1000 episodes and we show the results which are the averages over 100 bandits. Each selected action receives a uniformly distributed reward r_a between the respective boundaries

$r_{10} = u[9\ 10]$, $r_9 = u[8\ 9]$, $r_8 = u[7\ 8]$,
$r_7 = u[6\ 7]$, $r_6 = u[5\ 6]$, $r_5 = u[4\ 5]$,
$r_4 = u[3\ 4]$, $r_3 = u[2\ 3]$, $r_2 = u[1\ 2]$,
$r_1 = u[0\ 1]$ (2)

After a fixed time interval, $\Delta T = 100$, rewards of the actions would change. Therefore actions rewards change 10 times over 1000 episodes with the following pattern:

$$0123456789 \longrightarrow 8901234567 \longrightarrow 6789012345 \longrightarrow$$
$$4567890123 \longrightarrow 2345678901 \longrightarrow 0123456789 \longrightarrow$$
$$8901234567 \longrightarrow 6789012345 \qquad (3)$$

3. Action value estimates update rule and ad hoc techniques

We apply four ad hoc methods for the non_stationary bandit problem. These methods lack strong theoretical foundations but can be fine-tuned in order to produce near-optimal results, a quality that makes them very interesting, especially from a practitioner's point of view. These methods are (i) the probability matching algorithm, (ii) the e-greedy action selection, (iii) the softmax action selection and finally, (iv) the adaptive pursuit method. As all RL methods are based on action value estimates $Q_{t_j}(a_i)$_ i. e. the estimated reward for taking action a_i in the t_j th episode, all the mentioned algorithms maintain estimates of the actions actual expected reward that are updated every time where action a_i is tried, by the following rule called Action value estimates update rule:

$$Q_q(a_i) = Q_{q-1}(a_i) + \alpha[r_{t_j} - Q_{q-1}(a_i)] \qquad (4)$$

where action a_i is selected for the qth time in episode t_j and α is a positive constant , $0 < \alpha < 1$, which is called learning rate.

3.1 probability matching technique

One of the ways to set the selection probability $P(a_i)$, of action a_i can be proportional to its current estimated value:

$$P_{t_j}(a_i) = \frac{Q_{t_j}(a_i)}{\sum_n Q_{t_j}(a)} \qquad (5)$$

However, this approach does not exclude the possibility that at some point the probability with which a certain action is being selected reaches near 0, rendering this action practically unavailable to the agent for the rest of the task. That is exactly what needs to be avoided in a non-stationary bandit task. This can be solved by setting a minimum value $P_{\min}$ for all the selection probabilities and consequently setting $P_{\max} = 1-(n-1)P_{\min}$, where n is the number of actions. Therefore the selection probabilities are updated in every episode t_j as below:

$$P_{t_j}(a_i) = P_{\min} + (1 - nP_{\min})\frac{Q_{t_j}(a_i)}{\sum_n Q_{t_j}(a)} \qquad (6)$$ *3.2.*

3.2 e_greedy action selection

By the e_greedy action selection the agent selects the greedy action with probability 1-e and selects other actions with probability e which cause exploratory choices. E is a small positive parameter, 0<e<1, which is usually chosen in the space of [0.01 0.1].

3.3 softmax action selection

softmax action selection method gives the highest selection probability to the greedy action and all other actions are ranked and weighted according to their value estimates. The most common softmax method uses a Gibbs or Boltzman distribution. It chooses action a_k in the play t_j with probability:

$$P(a = a_k) = \frac{e^{Q_{T_j}(a_k)/T}}{\sum_{i=1}^{n} e^{Q_{T_j}(a_i)/T}} \qquad (7)$$ T is a positive parameter called temperature. High temperatures cause the actions to be nearly equiprobable but low temperatures cause a greater difference in selection probability that for actions that differ in their value estimates. In the limit as $T \to 0$ softmax action selection becomes the same as greedy action selection. With appropriate adjustment of T this method can be applied to non_stationary bandit problems.

3.4 adaptive pursuit method

Pursuit methods were initially proposed by Thathachar and Sastry [9]. Theirens [8] proposed a variant suitable of non_stationary environments called the adaptive pursuit algorithm.

Similarly to probability matching technique, all selection probabilities can not fall under the threshold $P_{\min}$ or exceed the maximum value $P_{\max} = 1-(n-1)P_{\min}$. Let a_g be the greedy action in episode t_j. The probability of selecting a_g is incremented toward $P_{\max}$ while all the other remaining probabilities are decremented toward $P_{\min}$.

$$P_{t_j}(a_g) = P_{t_{j-1}}(a_g) + \beta(P_{\max} - P_{t_{j-1}}(a_g)) \qquad (8)$$

$$P_{t_j}(a_i) = P_{t_{j-1}}(a_g) + \beta(P_{\min} - P_{t_{j-1}}(a_g)), \forall i \neq g \qquad (9)$$

Where β is a small positive parameter, $0 < \beta < 1$.

4. Optimistic ad hoc results

The obtained results for probability matching method, e_greedy action selection, softmax action selection and adaptive pursuit methods are shown in fig. (1)~(4). While in all methods $\alpha = 0.1$ and $P_{\min} = \frac{1}{1000n}$ which ensures $P_{\min} < \frac{1}{n}$. The initial estimated values have been optimistically chosen for all the actions.

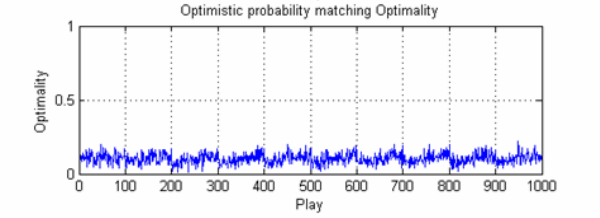

Fig 1: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; alpha = .1; result are average over 100 runs .

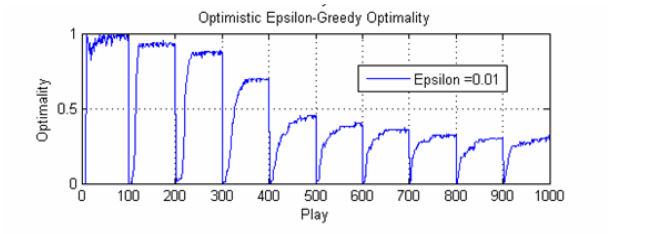

Fig 2: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10;epsilon=.01; alpha = .1; result are average over 100 runs.

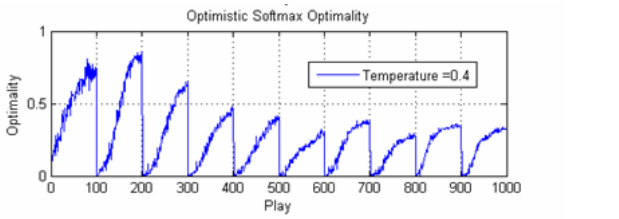

Fig 3: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; alpha = .1; temperature=.4; result are average over 100 runs.

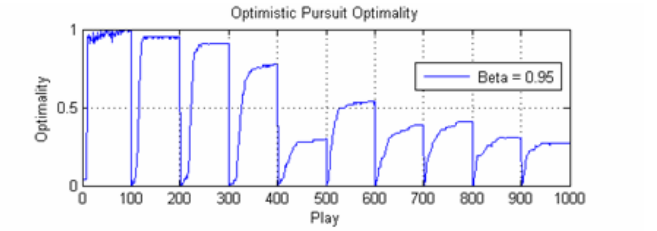

Fig 4: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; beta=.95; alpha = .1; result are average over 100 runs.

According to the results, e_greedy and adaptive pursuit methods are performing the same and probability matching exhibit the poorest performance in comparison with the results of other methods. What is obvious in all results is that by progressing through the episodes, selecting the optimal actions percentage are decreasing. The main reason is that this kind of optimistic initializing of the estimated values is not well suited for non_stationary problems and its drive for exploration is inherently temporary [7]. In the following section we represent a new approach to optimistic initializing of the estimated values which helps the non_stationary bandit problems solutions gaining better results.

5. Sequential optimistic ad hoc methods

In the new approach we assumed that the agent has some incomplete knowledge about the non_stationary environment. It just assumes that whenever the environment switches to a new set of actions rewards, it takes at least k episodes to face another change in the actions rewards ,which k is the number of the actions. Therefore after the first k episodes the agent checks the numerical gap which is between the rewards of any two sequential episodes. If the difference between these two rewards is considerable, the agent can conclude that the environment has switched to a new set of actions rewards. Consequently it replaces the calculated estimated values with the optimistic estimated values and starts to find the new optimal action in first k following episodes ,then exploite the new optimal action . This process will continue up to the last episode, while after each replacement of estimated values, agent stops the evaluation of optimizing condition of estimated values just during the next k episodes.

Finally, we can now specify the sequential optimistic initialize algorithm:

1-Specify the gap

2-While temp < 0
  If reward(play-1)> (reward(play)+gap)
    Value estimated= initialize optimistic
  End
  temp= number of arms
 End
temp= temp – 1

however, you should notice that if the number of actions are k and the environment period of changing is ΔT, then the probability of selecting the optimal action is $\frac{\Delta T-(k-1)}{\Delta T}$. Therefore if $\Delta T \to \infty$ and $\Delta T \gg k$, then $\frac{\Delta T-(k-1)}{\Delta T} \to 1$. In the other words, this method is suitable for the environment that ΔT is greater than $2k$. Consequently the probability of selecting the optimal action is always greater than $1-\frac{k-1}{2k} \geq 0.5$.

6. Sequential optimistic ad hoc results

The results of sequential optimistic ad hoc methods are as below:

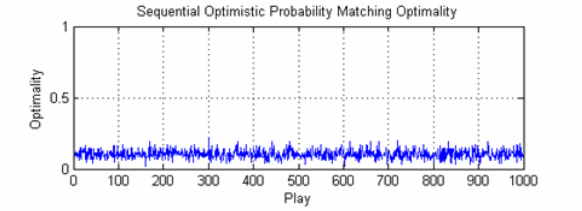

Fig 5: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; alpha = .1; gap=.85 ; result are average over 100 runs .

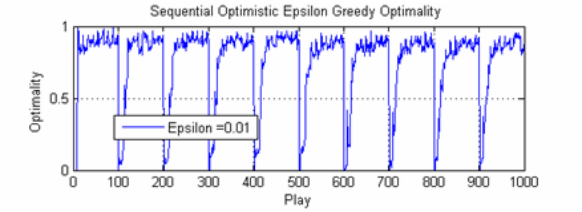

Fig 6: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10;epsilon=.01; alpha = .1; gap=.85 ; result are average over 100 runs.

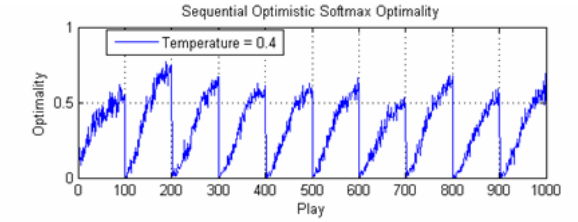

Fig 7: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; alpha = .1; temperature=.4; gap=.85 ; result are average over 100 runs.

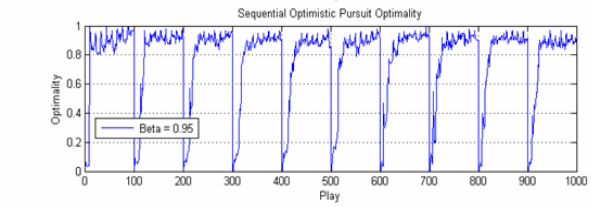

Fig 8: the probability of selecting the optimal action at each time step in the non stationary environment with switching interval $\Delta T = 100$ Time steps ; $P_{min} = .0001$; number of actions k=10; beta=.95; gap=.85 ;alpha = .1; result are average over 100 runs .

By applying the sequential optimistic ad hoc methods to non_stationary bandit problems, the obtained results improved considerably in comparison to optimistic ad hoc methods. The best results are achieved by sequential optimistic adaptive pursuit and sequential optimistic e_greedy action selection that have performed nearly the same. We can observe the improved performance of softmax action selection by applying sequential optimistic softmax action selection in fig. (7).

The sequential optimistic probability matching technique that results the weakest performance among others, does not show considerable results improvement in comparison with optimistic probability matching technique.

7. Conclusion

MABP is one of the common ways of balancing exploration and exploitation. In the non_stationary MABP the pattern of the actions rewards changes during the learning time. We obtained poor results for the non_stationary MABP by applying optimistic ad hoc methods, while we improved the results by proposing the sequential optimistic ad hoc methods. According to the results it can be easily concluded that sequential optimistic ad hoc methods are one of the most powerful methodologies for non_stationary MABP with fast changes. However we should notice this method is suitable for the problems that the number of actions are at least less than the half of the environment period of changing.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. "The nonstochastic multiarmed bandit problem". *SIAM j. Computing* Vol.32, No.1, pp.48–77, 2002.

[2] J.C. Gittins, "Multi-armed Bandit Allocation Indices", Wiley, New York, 1989.

[3] D.E. Goldberg. "Probability matching, the magnitude of reinforcement, and classifier system bidding". *Machine Learning*. Vol.5, pp. 407–425, 1990.

[4] D.E. Koulouriotis and A. Xanthopoulos, " Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems", Applied Mathematics and Computation 196 pp 913–922, (2008).

[5] P.P. Varaiya, J.C. Walrand and C. Buyukkoc, "Extensions of the multi armed bandit problem: the discounted case", IEEE Transactions on Automated Control 30 (5) (1985) 426–439.

[6] H. Kaspi and A. Mandelbaum, "Multi-armed bandits in discrete and continuous time", Annals of Applied Probability 8 (4) (1998) pp 1270–1290.

[7] R.S. Sutton and A.G. Barto, "Reinforcement Learning: An Introduction", MIT Press, Cambridge, MA, 1998.

[8] D. Thierens, "An adaptive pursuit strategy for allocating operator probabilities", in: Proceedings of the Genetic and Evolutionary

[9] Computing Conference (GECCO 2005), 2005, pp. 1539–1546.

[10] M.A.L. Thathachar and P.S. Sastry, "A class of rapidly converging algorithms for learning automata", IEEE Transactions on Systems, Man and Cybernetics SMC-15 (1985) 168–175.