Multi-Armed Bandit, Dynamic Environments and Meta-Bandits

C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud and M. Sebag

LRI, Université Paris Sud, 91405 Orsay Cedex, France

Introduction

Multi-Armed Bandit problem

- K arms
- i-th arm : reward m_i (unknown)
  empirical reward $\hat{m}_i$
  numbers of trials n_i
- Best arm : reward m^*

goal: find a policy minimizing the regret $\mathcal{R}(T) = \sum_i n_i \times (m_i^* - \hat{m}_i)$

UCBT : play $i = Argmax\{\hat{\mu}_j + \sqrt{\frac{2 \log N}{n_j} \times min(\frac{1}{4}, V_j(n_j))}\}$, $n_i = n_i + 1$

From Multi-Armed Bandits to News recommendations

- K types of news (international, sports, politics, sciences, etc.)
- News submitted to customers → interestingness (in [0,1])

goal : provide every customer with the most interesting news

Challenge :

- The customer's interests and the news change quickly
- Bandits algorithms do not cope with fast dynamics

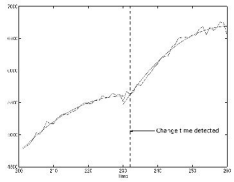
TOUCH CLARITY

The Adapt EvE algorithm = UCB-Tuned +

(1) changes detection

Change point detection based on Page-Hinkley Statistics :

$t \rightarrow \bar{x}_t = \frac{1}{t} \sum_{\ell=1}^{t} x_\ell$
$m_T = \sum_{t=1}^{T} (x_t - \bar{x}_t + \delta)$
$M_T = \max\{m_t,\ t = 1 \ldots T\}$
$PH_T = M_T - m_T$
change detected if $(PH_T > \lambda)$



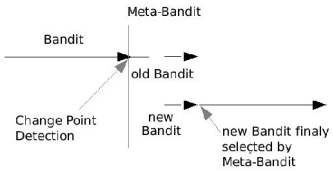
[Page, 1954; Hinkley, 1969-1971]

(2) Policy for transitory regime

(2.1) γ restart : at change point detection in UCB : $n_i = \gamma . n_i$

(2.2) Meta-Bandit : Play UCBT algorithm. At change point detection use UCBT to select between :
option 1 : old Bandit (do as before)
option 2 : new Bandit (restart)

(2.3) Discounted Meta-Bandit :
Meta-Bandits + in UCBT, replace $n_i = \gamma . n_i + 1$

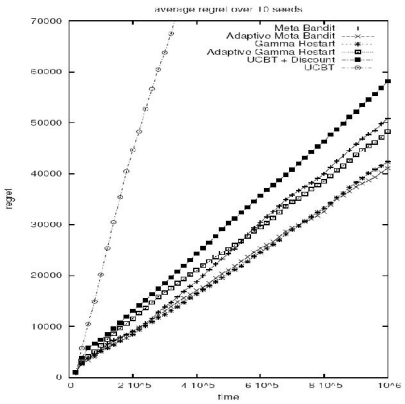


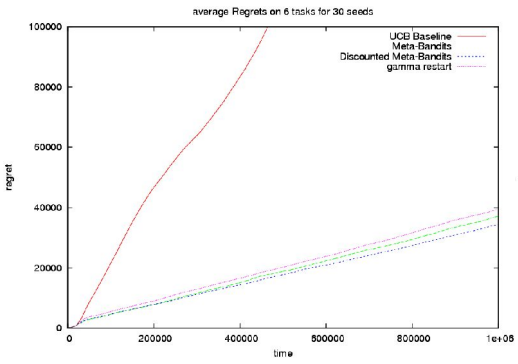
(3) adapting change

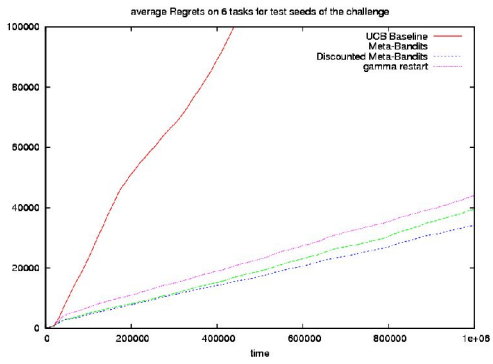
(3.1) change detection A posteriori : one knows whether the change was a true one (i.e. New bandit wins in Meta-Bandits) else the threshold λ is adjusted by e :

$$\lambda : \quad \lambda \times e$$
$$e = \begin{cases} (1 \quad \alpha \Delta\mu) & if\ true\ alarm \\ (1 \quad \beta \Delta\mu) & if\ false\ alarm \end{cases}$$

Results







Perspectives

- Theoretical analysis : relate γ with dynamics
- Extensions to many-armed Bandits
- Double cluster of options and customers

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ PARIS-SUD 11
INRIA FUTURS