feedback models for multiple-plays bandits

ALICIA Meeting

Claire Vernade & Paul Lagrée, Olivier Cappé

February 12th, 2016

ALICIA – ANR Project

introduction

what's new from last alicia meeting ?

feedback :

- Bandit feedback : unknown function of the super-action;
- Semi-Bandit feedback : individual rewards of each arm in the super-action, may be partially observed.

Last time :

- one model for top to bottom scrolling with independent stopping;
- a lower bound on the regret for such model

Today :

- Two different models for different feedback situations;
- Lower bounds for each case;
- Algorithms based on the optimism-in-the-face-of-uncertainty principle.
- **But**, still no regret analysis of those algorithms.

modelling user's behaviour

already existing: cascade model

[Kveton 15'],[1]

1. at each round $t = 1, \cdots, T$, select a list A_t of L elements among K arms
2. feedback: each item until the first click is observed, nothing after.
3. reward is binary: 1 if one click, 0 otherwise

Placing the worst items in the beginning of the list is better since more information is obtained without increasing regret.

Not well suited for a recommendation context when no query.

[1] Branislav Kveton et al. "Cascading Bandits : Learning to Rank in the Cascade Model". In: Proceedings of the 32nd International Conference on Machine Learning. 2015.

the vanishing attention idea

One principle: the user's attention vanishes while looking at recommendations.

1. First model: **Scrolling-type**. The user scrolls from top to bottom and leaves when out of patience (departure independent from recommendations);
2. Second model: **Display-type**. The user is shown ads while browsing and pays variable but decaying attention to them.

scrolling-type model

1. at each round $t = 1, \cdots, T$, choose a list A_t of L items among K possible arms.
2. The user rates every item $k \in A_t$ until she decides to leave at position Λ_t (random variable)
3. The probability of scanning item in position l is modeled by $\kappa_l > 0$
4. reward: $\sum_{l=1}^{\Lambda_t} X_{t,l}$

display-type model

1. at each round $t = 1, \cdots, T$, choose a list A_t of L items among K possible arms.
2. The user does not pay attention to all of them: equivocal rewards "0" but full semi-bandit feedback.
3. The probability of effectively looking at and rating item in position l is modeled by $\kappa_l > 0$.
4. Reward: $\sum_{l=1}^{L} Y_l X_{t,l}$ where $Y_l \sim \mathcal{B}(\kappa_l)$.

lower bound on the regret

Lower bound for model 1:

$$\liminf_{T\to\infty} \frac{R(T)}{\log(T)} \geq \sum_{k=L+1}^{K} \frac{\cancel{\kappa_L}(\theta_L - \theta_k)}{\cancel{\kappa_L} KL(\theta_k, \theta_L)}$$

Comment: No kappas. The information loss at the denominator is exactly compensated by regret reduction.

Lower bound for model 2:

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{k=L+1}^{K} \min_{l\in\{1,\dots,L\}} \frac{\Delta_{v_l^k}(\theta)}{KL(\kappa_l\theta_k, \kappa_l\theta_L)}$$

Comment: Each suboptimal item has an optimal exploration position that depends on the parameters of the problem.

algorithms

Algorithms for Scrolling-type model: Observations are explicit, no modification needed, CUCB works with usual optimistic indices (UCB, KL-UCB or Thompson samples).

Algorithms for Display-type model: Observations are equivocal and all empirical estimators ar biased. New concentration results are needed ! Examples : for item k, if $\tilde{N}_k(t) = \sum_{l=1}^{L} \kappa_l N_{k,l}(t)$,

$$U_k^{UCB}(t,\delta) = \frac{S_k}{\tilde{N}_k} + \sqrt{\frac{N_k}{\tilde{N}_k}}\sqrt{\frac{\delta}{2\tilde{N}_k}}$$

or

$$U_k^{L}(t,\delta) = \sup_{q \in [\theta_k^{\min},1]} \left\{ q \,\middle|\, \sum_{l=1}^{L} N_{k,l} KL\left(S_{k,l}/N_{k,l}, \kappa_l q\right) \leq \delta \right\}$$

experimental results - algorithms

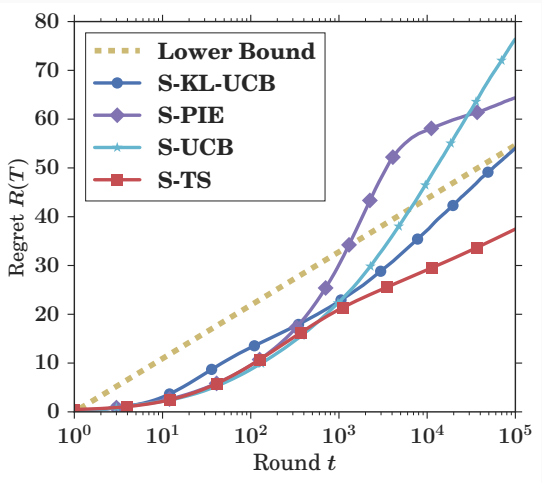

Figure 1: Scrolling model on synthetic data.

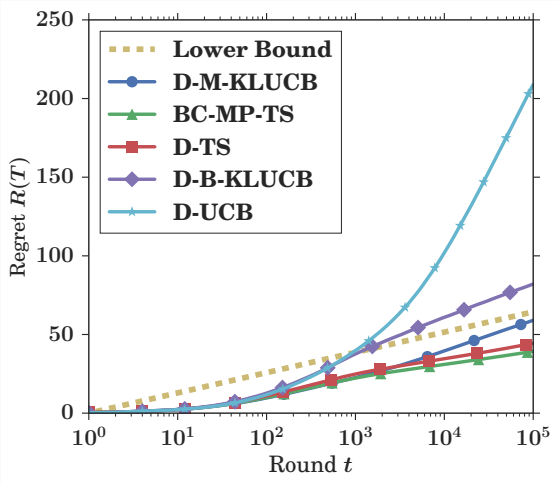

Figure 2: Display model on synthetic data.

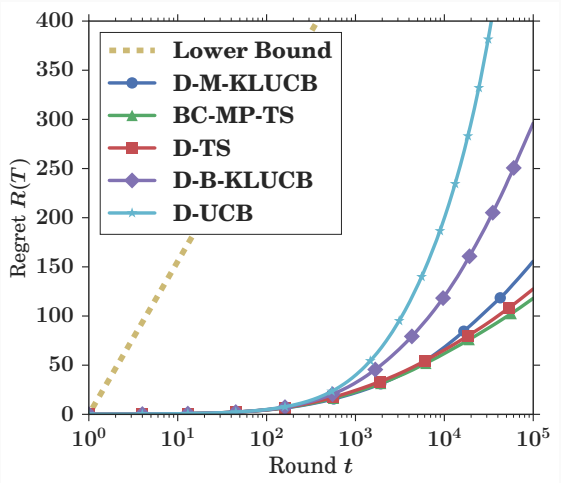

Figure 3: Display model on simulated real data.

future work

future work

Some perspectives can be considered:

- Analyze the regret of the proposed algorithms: seems technical and rather hard to achieve;
- Propose user models that include dependency between vanishing attention and item suggestions;
- Above all: real tests on real data.