Contextual Combinatorial Cascading Bandits

Shuai Li1; Baoxiang Wang3; Shengyu Zhang1; Wei Chen2
1The Chinese University of Hong Kong, 2Microsoft Research

Motivation
- Cascading feedback
- Websites search results
- Recommended movies
- All are sequential lists
  - Users go through the list from top down
  - Stop at the first satisfactory item
  - Click it
- This online feedback helps improving future list quality
- Contexts
  - User profiles, search keywords
  - Important for search, recommendations, etc.
- Combinatorial
  - Action is selection of a sequence of items.
  - May have other combinatorial constraints (e.g. paths in networks)

Setting
- A finite set $E = \{1, \ldots, L\}$ of base arms.
- Let $S$ be the set of feasible actions, which are tuples of $E$ with length at most $K$.
- Position discounts $y_s(0.1)k \leq K$.
- $\alpha$-approximation oracle $O_k$
- At time $t$,
  - For each $a \in E$, a feature vector $x_{t,a} \in \mathbb{R}^{d+1}$ with $\|x_{t,a}\|_2 \leq 1$ is revealed to the learning agent.
  - Let $\mathcal{H}_a$ denote the history so far.
  - The learning agent recommends $A_t = (a_1, \ldots, a_k) \in S$ to the user.
  - The user checks from the first item of $A_t$ and stops at $O_k$-th item under some stopping criterion.
  - The learning agent observes the weights of first $O_k$ base arms in $A_t$, $w_t(a_1), \ldots, w_t(a_k) \leq O_k$.
  - Assume given $\mathcal{H}_t, w_t(a_1)'s$ are mutually independent $R$-sub-Gaussian random variables with $E[w_t(a_i)|\mathcal{H}_t] = \theta_t x_{t,a}$ for some unknown $\theta_t \in \mathbb{R}^{d+1}$ with $\|\theta_t\|_2 \leq 1$, $0 \leq \theta_t x_{t,a} \leq 1$.
  - Assume the expected reward of action $A$ is a function $f(A, w)$ of expected weight $w$ satisfying
    - monotonicity
    - $B$-Lipschitz continuity
  - The $\alpha$-regret of action $A$ on time $t$ is
    \[ R^\alpha(t) = \alpha_t - f(A, w_t). \]
  - Minimize $\alpha$-regret of $n$ rounds
    \[ R^\alpha(n) = \sum_{t=1}^n R^\alpha(t). \]

Algorithm: C$^3$-UCB
1. Parameters:
   \[ (y_t \in [0,1])_{k \in S}: \delta = \frac{1}{\sqrt{m}}, \lambda \geq C_p = \sum_{k=1}^K y_t^2 \]
2. Initialization:
   \[ \theta_t = 0, \theta_0(a) = 1, Y_0 = \emptyset, X_0 = \emptyset, Y_0 = \emptyset \]
3. For all $t = 1, 2, \ldots, n$ do
   1. Obtain context $x_{t,a}$ for all $a \in E$.
   2. For any $a \in E$, compute
      \[ U_t(a) = \min \{ \theta_t x_{t,a} + \beta_{t-1}(\delta), x_{t,a} \}
      \]
   3. Choose action $A_t$ using UCBs $U_t$
      \[ A_t = (a_1, \ldots, a_k) \sim \mathcal{O}_k(U_t) \]
   4. Play $A_t$ and observe $O_t$: $w_t(a_1) \geq O_t$.
   5. Update statistics
      \[ V_t = V_{t-1} + \sum_{i=1}^k y_t^2 x_{t,a_i}^2 x_{t,a_i} \]
      \[ X_t = X_{t-1} + \sum_{i=1}^k y_t^2 x_{t,a_i} - y_t x_{t,a_i} \frac{X_i}{X_{k+1}} \]
      \[ Y_t = \gamma_t(a_1, \ldots, a_k, X_{k+1}) = Y_{t-1} + \gamma_t(a_1, \ldots, a_k, X_{k+1}) \]
      \[ \theta_t = \theta_t - \frac{1}{\sqrt{\det(I_k + X_{k+1})}} X_{k+1} Y_t \]
      \[ \beta_t = \beta_{t-1} - \sqrt{\ln(\det(\mathcal{H}_t)(x_{t,a_i}^2 + \gamma^2))} + \gamma \]
      End for $t$

Results
Theorem 1. Suppose the expected reward function $f(A, w)$ is a function of expected weights and satisfies monotonicity and $B$-Lipschitz continuity. Then the $\alpha$-regret of our algorithm, C$^3$-UCB, satisfies
\[ R^\alpha(n) = O\left( \frac{d}{\sqrt{m}} \ln(C_p n) \right), \]
where $R$ is the sub-Gaussian constant and $C_p = \sum_{k=1}^K y_t^2 \leq K$.

Corollary 2. In the problem of cascading recommendation, the expected reward is disjunctive
\[ f(A, w) = \sum_{k=1}^K \left(1 - w(a_k)\right)w(a_k) \]
where $1 = y_1 \geq \cdots \geq y_K \geq 0$. Then the $\alpha$-regret of C$^3$-UCB satisfies
\[ R^\alpha(n) = O\left( \frac{d}{\sqrt{m}} \ln(C_p n) \right), \]
where $f^* = \max f_i$, the maximal expected reward in $n$ rounds.

Theorem 3. Suppose $1 = y_1 \geq \cdots \geq y_K \geq 1 - \frac{\alpha}{d}$, where $\alpha = \min f_i$. Then the $\alpha$-regret of C$^3$-UCB for the conjunctive objective
\[ f(A, w) = \sum_{k=1}^K \left(1 - y_k\right)w(a_k) \left(1 - w(a_k)\right) \]
satisfies
\[ R^\alpha(n) = O\left( \frac{d}{\sqrt{m}} \ln(C_p n) \right). \]

Table 1. Comparisons of our setting with previous ones.

<table>
<thead>
<tr>
<th>context</th>
<th>Cascading</th>
<th>Position discount</th>
<th>General reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combinatorial UCB</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Contextual Combinatorial UCB</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Comb-Cascade</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>C$^3$-UCB(ours)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Conclusions
- Formulate Contextual Combinatorial Cascading Bandits problem
- Propose C$^3$-UCB algorithm that can handle
  - contextual information
  - cascading feedback
  - position discount
  - general reward function
- Theoretical analysis and empirical evaluation

References
2. Shalev-Shwartz, Shai, and Amir Globerson. “Online Learning and Stochastic Approximations.” MIT. 2015.

Contact
Shuai Li
Email: shuai@acm.cuhk.edu.hk

Shengyu Zhang
Email: syzhang@cse.cuhk.edu.hk

Baoxiang Wang
Email: bwang@cse.cuhk.edu.hk

Wei Chen
Email: weic@Microsoft.com