Contextual Bandits with Delayed Feedback and Semi-supervised Learning (Student Abstract)

Luting Yang, Jianyi Yang, Shaolei Ren
University of California, Riverside
Riverside, California 92521
{lyang029, jyang239, shaolei}@ucr.edu

Abstract

Contextual multi-armed bandit (MAB) is a classic online learning problem, where a learner/agent selects actions (i.e., arms) given contextual information and discovers optimal actions based on reward feedback. Applications of contextual bandit have been increasingly expanding, including advertisement, personalization, resource allocation in wireless networks, among others. Nonetheless, the reward feedback is delayed in many applications (e.g., a user may only provide service ratings after a period of time), creating challenges for contextual bandits. In this paper, we address delayed feedback in contextual bandits by using semi-supervised learning — incorporate estimates of delayed rewards to improve the estimation of future rewards. Concretely, the reward feedback for an arm selected at the beginning of a round is only observed by the agent/learner with some observation noise and provided to the agent after some a priori unknown but bounded delays. Motivated by semi-supervised learning that produces pseudo labels for unlabeled data to further improve the model performance, we generate fictitious estimates of rewards that are delayed and have yet to arrive based on already-learnt reward functions. Thus, by combining semi-supervised learning with online contextual bandit learning, we propose a novel extension and design two algorithms, which estimate the values for currently unavailable reward feedbacks to minimize the maximum estimation error and average estimation error, respectively.

Problem Formulation

Given context information at each round $t = 1, 2, \cdots, T$, the agent/learner needs to select an arm. We denote $x_{a,t} \in \mathcal{R}^M$ as the context, which is a representation of the environment information or feature regrading arm a at the t-th round, for $a \in \mathcal{A} = \{1, 2, \cdots, K\}$ and $t = 1, 2, \cdots, T$. For a selected arm a at round t, we denote the resulting reward as $y_{a,t} \in \mathcal{R}$. Nonetheless, due to feedback delays, the agent can only receive the reward feedback at the beginning of the $(t + d_t)$-th round, where $d_t \geq 1$ is the delay for the arm selected at round t. Note that it is possible that the learner simultaneously receives multiple feedback signals for arms selected in prior rounds. Assume that at round t, the agent receives a set of reward feedbacks for arms selected at rounds belonging to the set $\mathcal{S}_t$, i.e. $\forall \tau \in \mathcal{S}_t, \tau + d_\tau = t$.

We use the kernel method to model non-linear reward functions in terms of the context and arm. Given a kernel function $k(x, x') = \phi(x)^\top \phi(x'), \forall x, x' \in \mathbb{R}^M$, we can express the expected reward function as $g(x_{a,t}) = \phi(x_{a,t})^\top \theta$ in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ corresponding to the kernel function $k(x, x')$. Specifically, the actual reward feedback $y_{a,t}$ received by the agent (after a delay of d_t rounds) for its arm a selected at round t is written as

$$y_{a,t} = g(x_{a,t}) + \epsilon_t = \phi(x_{a,t})^\top \theta + \epsilon_t,$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon)$ is the reward observation noise. The goal of the agent is to maximize its total expected reward, or equivalently minimize its cumulative regret, over T rounds. We define the best arm given context $x_{a,t}$ at round t as the arm that leads to the highest expected reward, i.e.,

$$a_t^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[y_{a,t}] = \arg\max_{a \in \mathcal{A}} g(x_{a,t}).$$

Thus, the agent needs to find an arm selection policy based on the received feedback signals and context-arm history to minimize the cumulative regret $R_T = \sum_{t=1}^{T} \mathbb{E}\left[y_{a_t^*,t} - y_{a_t,t}\right]$, where the expectation is taken over the observation noises.

Bandit with Semi-supervised Learning

In bandits with immediate reward feedback, the reward parameter θ can be estimated based on all the reward-context pairs in the history rounds. However, in the considered bandits with delayed feedback, since the rewards of some history rounds are not fed back, only unsupervised data (context) of these rounds is available, which can increase the reward estimation error and cause a large regret. Despite this, semi-supervised learning which exploits the unsupervised data can be used to improve the reward estimation (Zhu and Goldberg 2009). In this section, we will consider two ways of semi-supervised learning to obtain better reward estimations, which are described in Algorithm 1.

Minimizing the Maximum Estimation Error

Based on the reward function learnt so far, the agent can estimate the upper and lower bounds of average rewards for those delayed feedbacks while waiting for them to arrive. Thus, we can view the upper and lower bounds as the perturbation range of the delayed rewards, and provide a robust learning algorithm.

Algorithm 1 Contextual UCB with Semi-supervised Learning

Inputs : kernel function k and parameter α and λ.
for $t = 1, \cdots, T$ **do**
  if $length(\hat{\mathbf{y}}_t) = 0$ **then**
    Randomly choose arm a_t
  else
    if $|\mathcal{S}_t| \neq 0$ **then**
      $\forall \tau \in \mathcal{S}_t$, augment $y_{a_\tau,\tau}$ into $\hat{\mathbf{y}}_t$, and append $\phi(x_{a,\tau})$ into $\hat{\mathbf{\Phi}}_t$ and remove it from $\tilde{\mathbf{\Phi}}_t$.
    Receive context $x_{a,t}$, $a = 1, \cdots, K$
    for $a \in \mathcal{A}$ **do**
      MinMax: Calculate $\tilde{g}^c_{a,t}$ and $\tilde{w}_{a,t}$
      Or MinAvg: Calculate $\bar{g}_{a,t}$ and $\bar{w}_{a,t}$
    MinMax: $a_t = \arg\max_{a\in\mathcal{A}} \tilde{g}^c_{a,t} + (\alpha+\lambda)\tilde{w}_{a,t}$
    Or MinAvg: $a_t = \arg\max_{a\in\mathcal{A}} \bar{g}_{a,t} + (\alpha+\lambda)\bar{w}_{a,t}$
    Augment $\phi(x_{a,t})$ into $\mathbf{\Phi}_t$ and $\tilde{\mathbf{\Phi}}_t$

First, we use $\tilde{\mathbf{\Phi}}_t$ to store contexts without feedback yet and $\hat{\mathbf{\Phi}}_t$ to store contexts whose reward feedbacks have arrived and the corresponding rewards are $\hat{\mathbf{y}}_t$. Thus, we can use $\mathbf{\Phi}_t$ to represent all the experienced contexts up to the beginning of round t, such that $\mathbf{\Phi}_t^\top = [\tilde{\mathbf{\Phi}}_t^\top, \hat{\mathbf{\Phi}}_t^\top]$. Once a delayed reward feedback is provided, it will be appended to $\hat{\mathbf{y}}_t$ and its corresponding context information will be transferred from $\tilde{\mathbf{\Phi}}_t$ to $\hat{\mathbf{\Phi}}_t$. Based on the feedback rewards, we can get a primary estimation of reward parameter by kernel ridge regression (Deshmukh, Dogan, and Scott 2017) as

$$\hat{\theta}_t = \hat{\mathbf{C}}_t^{-1}\hat{\mathbf{\Phi}}_t\hat{\mathbf{y}}_t, \tag{1}$$

where $\hat{\mathbf{C}}_t^{-1} = \hat{\mathbf{\Phi}}_t {\Sigma_t}^{-1}\hat{\mathbf{\Phi}}_t^\top + \lambda\mathbf{I}$ and λ is a hyper-parameter.

Then based on the primary parameter estimation (1), we can get an ambiguous range of each delayed reward feedback that have not arrived. Specifically, denote the reward feedbacks that have not arrived before round t are contained in $\tilde{\mathbf{y}}_t$. By using the estimation error bound of kernel ridge regression (1), with probability at least $1-\delta, \delta \in (0,1)$, the k-th element $\tilde{y}_t^k$ in the vector $\tilde{\mathbf{y}}_t$ corresponding to context x_a^k can be bounded by the confidence width as follows:

$$\left|\tilde{y}_t^k - \phi(x_a^k)^\top\hat{\theta}_t\right| \leq (\alpha+\lambda)\sqrt{\phi(x_a^k)^\top\hat{\mathbf{C}}_t^{-1}\phi(x_a^k)} \tag{2}$$

where $\alpha = \sqrt{\frac{1}{2}\ln\frac{2KT}{\delta}}$.

Next we can estimate the reward function in a robust manner given the ambiguous range shown in Eqn. (2). Specifically, the robust estimation of $\tilde{\theta}_t$ is obtained by solving the MinMax optimization problem:

$$\min_{\tilde{\theta}}\left\{\max_{\tilde{\mathbf{y}}_t} \|\tilde{\mathbf{y}}_t - (\tilde{\mathbf{\Phi}}_t)^\top\tilde{\theta}\|^2 + \|\hat{\mathbf{y}}_t - (\hat{\mathbf{\Phi}}_t)^\top\tilde{\theta}\|^2 + \lambda\|\tilde{\theta}\|^2\right\}$$

Finding a closed-form solution $\tilde{\theta}_t$ can be intractable due to the maximization operator. we can use a sampling approach to solve the problem approximately. First, we generate random samples within the ambiguous range in Eqn. (2). We use $\mathbf{Y}_t$ to store all the generated samples for those delayed rewards that have not yet been received. Then, for each sample $\mathbf{y}_t^i \in \mathbf{Y}_t$, we can solve the kernel-based ridge regression as

$$\tilde{\theta}_t^i = \arg\min_{\tilde{\theta}}\left\{\|\mathbf{y}_t^i - (\tilde{\mathbf{\Phi}}_t)^\top\tilde{\theta}\|^2 + \|\hat{\mathbf{y}}_t - (\hat{\mathbf{\Phi}}_t)^\top\tilde{\theta}\|^2 + \lambda\|\tilde{\theta}\|^2\right\}.$$

Based on (Deshmukh, Dogan, and Scott 2017), we obtain $\tilde{\theta}_t^i$ and the estimated reward

$$g_{a,t}^i = \phi(x_{a,t})^\top\tilde{\theta}_t^i = \phi(x_{a,t})^\top\tilde{\mathbf{C}}_t^{-1}\mathbf{\Phi}_t\tilde{\mathbf{y}}_t^i \tag{3}$$

, where $\tilde{\mathbf{y}}_t^i = [\mathbf{y}_t^i, \hat{\mathbf{y}}_t]$ and $\tilde{\mathbf{C}}_t = \mathbf{\Phi}_t{\mathbf{\Phi}_t}^\top + \lambda\mathbf{I}$. In total, we can have $|\mathbf{Y}_t|$ candidate reward estimates. Here, we choose the worst-case candidate $\tilde{g}^c_{a,t}$ which has the largest estimation error as the estimated reward, i.e.

$$c = \arg\max_{i=1,\cdots,|\mathbf{Y}_t|}(\|\mathbf{y}_t^i - (\tilde{\mathbf{\Phi}}_t)^\top\tilde{\theta}_t^i\|^2 + \|\hat{\mathbf{y}}_t - (\hat{\mathbf{\Phi}}_t)^\top\tilde{\theta}_t^i\|^2 + \lambda\|\tilde{\theta}_t^i\|^2). \tag{4}$$

Like other UCB algorithms, we also add an exploration term of confidence width $\tilde{w}_{a,t} = \sqrt{\phi(x_{a,t})^\top\tilde{\mathbf{C}}_t^{-1}\phi(x_{a,t})}$ into the estimated reward.

Minimizing the Average Estimation Error

The second robust method is to minimize the average reward estimation error to estimate reward functions. Also, we first generate random samples $\mathbf{y}_t^i$ in $\mathbf{Y}_t$, based on the ambiguous range in Eqn. (2), as pseudo-feedback for contexts in $\hat{\mathbf{\Phi}}_t$. Instead of solving the MinMax optimization, we calculate $\bar{\theta}_t$ by minimizing the average reward estimation error of all samples in $\mathbf{Y}_t$ by kernel-based ridge regression:

$$\bar{\theta}_t = \arg\min_{\bar{\theta}}\left\{\frac{1}{|\mathbf{Y}_t|}\sum_i \|\mathbf{y}_t^i - (\tilde{\mathbf{\Phi}}_t)^\top\bar{\theta}\|^2 + \|\hat{\mathbf{y}}_t - (\hat{\mathbf{\Phi}}_t)^\top\bar{\theta}\|^2 + \lambda\|\bar{\theta}\|^2\right\}$$

Therefore, we can get the expected reward candidate as

$$\bar{g}_{a,t} = \phi(x_{a,t})^\top\bar{\theta}_t = \phi(x_{a,t})^\top\bar{\mathbf{C}}_t^{-1}\bar{\mathbf{\Phi}}_t\bar{\mathbf{y}}_t, \tag{5}$$

where $\bar{\mathbf{y}}_t = [\frac{1}{\sqrt{|\mathbf{Y}_t|}}\mathbf{y}_t^{1\top}, \ldots, \frac{1}{\sqrt{|\mathbf{Y}_t|}}\mathbf{y}_t^{|\mathbf{Y}_t|\top}, \hat{\mathbf{y}}_t^\top]^\top$, $\bar{\mathbf{\Phi}}_t$ is created by concatenation of $\frac{1}{\sqrt{|\mathbf{Y}_t|}}\tilde{\mathbf{\Phi}}_t$ itself $|\mathbf{Y}_t|$ times with one $\hat{\mathbf{\Phi}}_t$, and $\bar{\mathbf{C}}_t = \bar{\mathbf{\Phi}}_t\bar{\mathbf{\Phi}}_t^\top + \lambda\mathbf{I}$.

We still use UCB-based arm selection where the exploration term is expressed as $\bar{w}_{a,t} = \sqrt{\phi(x_{a,t})^\top\bar{\mathbf{C}}_t^{-1}\phi(x_{a,t})}$.

We will perform regret analysis and evaluations in our future work.

References

Deshmukh, A. A.; Dogan, U.; and Scott, C. 2017. Multi-task learning for contextual bandits. In *Advances in neural information processing systems*, 4848–4856.

Zhu, X.; and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3(1): 1–130.