CONTEXTUAL MULTI-ARMED BANDIT ALGORITHMS FOR PERSONALIZED LEARNING ACTION SELECTION

Indu Manickam, Andrew S. Lan, and Richard G. Baraniuk

Rice University

ABSTRACT

Optimizing the selection of learning resources and practice questions to address each individual student's needs has the potential to improve students' learning efficiency. In this paper, we study the problem of selecting a personalized learning action for each student (e.g. watching a lecture video, working on a practice question, etc.), based on their prior performance, in order to maximize their learning outcome. We formulate this problem using the contextual multi-armed bandits framework, where students' prior concept knowledge states (estimated from their responses to questions in previous assessments) correspond to contexts, the personalized learning actions correspond to arms, and their performance on future assessments correspond to rewards. We propose three new Bayesian policies to select personalized learning actions for students that each exhibits advantages over prior work, and experimentally validate them using real-world datasets.

Index Terms— contextual bandits, personalized learning

1. INTRODUCTION

In today's classrooms, knowledge is typically passed from teachers to students using a "one-size-fits-all" approach where students listen to the same lectures and work on the same homework questions. This approach naturally leads to gaps in student knowledge, since one teaching style is not optimal for every student in class and students have diverse backgrounds, abilities, and goals. In such a setting, teachers often do not have the time or resources to (i) identify specific areas that each student needs remediation on or (ii) tailor classroom lectures to address each student's needs. Machine learning-driven personalized learning systems use machine learning algorithms to analyze student data (e.g. their graded responses) [1]. This analysis provides estimates of each student's knowledge in a scalable way, which can be used to automatically create learning schedules that are tailored to cater to the strengths and weaknesses of each individual student.

In this paper, we focus on the problem of generating personalized learning schedules for students given their learning history in order to maximize their future learning outcomes. Following the setting in [2], we define a personalized learning schedule as an alternating sequence of personalized learning actions (PLAs) and assessments. PLAs are enrichment or remediation activities intended to *improve* students' knowledge of a set of educational concepts. Examples of PLAs include reading a textbook, watching a lecture video, and practicing a homework question. Assessments consist of questions intended to *test* students' current concept knowledge level.

1.1. Contributions

We study the problem of selecting the optimal PLA for each student to maximize the student's performance on the follow-up assessment. We pose this problem as a *contextual multi-armed bandits* (MAB) problem; We estimate each student's prior knowledge by analyzing their responses to questions in previous assessments using the sparse factor analysis (SPARFA) framework [1], and use them as contexts. We use each available PLA as an arm, and use students' responses to questions in the follow-up assessment as rewards.

We propose three new policies to select PLAs that are specifically adapted for binary-valued rewards (the graded student responses). The first two, CPT and U-CPT, are based on Thompson sampling [3]. The CPT algorithm exhibits closed-form updates of the policy parameters, which leads to very low computational complexity; The U-CPT policy extends the basic Thompson sampling policy CPT to account for the uncertainty in the contexts (estimates of their prior knowledge states), which is of crucial importance when the number of questions the students have previously answered is limited. The third policy, KG, is an online knowledge gradient policy that tries to maximize the information gained on the PLAs with each PLA selection, and may therefore be preferable when the policy must learn from a very small number of students [4]. We experimentally show that the three proposed policies achieve comparable or better performance than existing PLA selection policies using two real-world educational datasets.

2. PROBLEM FORMULATION

In the MAB problem, a gambler is given a collection of A slot machines (each with a single arm), that each have with a fixed reward distribution that is unknown to the gambler. On each play, the gambler selects an arm to play and receives a reward independent of previous plays. The objective is to select arms

to play such that the expected cumulative reward over N plays is maximized. The key to the MAB problem is to strike a balance between exploration (finding the arm with the highest expected reward) and exploitation (capitalizing by selecting the arm with the highest observed reward so far).

The contextual MAB framework studies the MAB problem when additional context information is available to the player, and uses it to select an arm to play. The context contains information on each arm and/or the current play.

2.1. PLA Selection as a Contextual MAB Problem

We formulate the PLA selection problem using the contextual MAB framework as follows. Let there be N total students and A total PLAs. After completing the selected PLA, each student takes a follow-up assessment consisting of Q questions, each with full credit points s_i, $i = 1, \ldots, Q$, which we observe as rewards. PLA selections are completed sequentially, so each student corresponds to a "play" in the MAB framework. The graded responses of student j to question i in the follow-up assessment after taking PLA a is denoted as $Y_{i,j}^a$, where $Y_{i,j}^a = 1$ denotes a correct response and $Y_{i,j}^a = 0$ denotes an incorrect response.

We assume we are given $\mathbf{c}_j$, a K-dimensional vector that encodes student j's prior knowledge on every *concept*; $\mathbf{c}_j$ is obtained from student j's graded responses to previous assessment questions. Although in this paper we estimate c_j using SPARFA [1], c_j can also be estimated using any alternative approach, e.g., item response theory [5].

We model the graded student responses on the follow-up assessment, $Y_{i,j}^a$, as Bernoulli random variables where

$$p(Y_{i,j}^a = 1 | \mathbf{w}_i^a) = \Phi(\mathbf{c}_j^T \mathbf{w}_i^a). \tag{1}$$

Here, $\Phi(\cdot)$ denotes a link function; the two commonly used link functions are the inverse logit link function $\Phi_{\text{log}}(x) = \frac{1}{1+e^x}$ and the inverse probit link function $\Phi_{\text{pro}}(x) = \int_{-\infty}^{x} \mathcal{N}(t; 0, 1) \mathrm{d}t$ where $\mathcal{N}(t; 0, 1)$ denotes the standard normal distribution. The parameter we are interested in estimating through our measurements is $\mathbf{w}_i^a$, a K-dimensional parameter vector indexed by question i and PLA a that we assume governs students' responses to each question in the follow-up assessment after completing each PLA.

3. PLA SELECTION POLICIES

In this section, we detail the PLA selection policies, including a previously developed policy and three new policies. The three new policies we develop use the inverse probit link function to model student responses.

3.1. A-CLUB

We have previously proposed a contextual bandits policy, asymptotic contextual logistic upper confidence bound (A-CLUB), for the task of PLA selection. A-CLUB leverages the asymptotic normality of the maximum likelihood operator to build a confidence ellipsoid around the maximum likelihood estimates of $\mathbf{w}_i^a$, the parameter vector of each question and each PLA. Using this information, A-CLUB arrives at an upper confidence bound of the expected total credit of a student on the follow-up assessment after taking each PLA, and selects the PLA with the highest upper confidence bound. Details of this algorithm can be found in [2, Alg. 2].

Algorithm 1: CPT

Input: A set of student concept knowledge state estimates, $\mathbf{c}_j$, $j = 1, 2, \ldots, N$, parameter σ^2
Output: PLA a_j for each student
$\mathbf{m}_i^a \leftarrow \mathbf{0}$, $\mathbf{V}_i^a \leftarrow \sigma^2 \mathbf{I}$, $\forall\, i, a$
for $j \leftarrow 1$ **to** N **do**
  for $a \leftarrow 1$ **to** A **do**
    Sample $\widehat{\mathbf{w}}_i^a \sim \mathcal{N}(\mathbf{m}_i^a, \mathbf{V}_i^a)$, $\forall\, i$.
  $a_j \leftarrow \arg\max_a \sum_{i=1}^{Q} s_i \Phi_{\text{pro}}(\mathbf{c}_j^T \widehat{\mathbf{w}}_i^a)$
  Update $\mathbf{m}_i^{a_j}$ and $\mathbf{V}_i^{a_j}$, $\forall\, i$, according to (2) and (3)

3.2. CPT: Contextual Thompson Sampling

We now develop a new Bayesian PLA selection algorithm using Thompson Sampling [3, 6]. In each play, a sample of $\mathbf{w}_i^a$ is generated by sampling from its posterior distribution [7], and the PLA with the highest expected sample reward is selected. We dub this policy as *contextual probit bandits with Thompson sampling* (CPT). The reason that we use the inverse probit link function is that it enables a more computationally efficient rule to update the posterior distribution of $\mathbf{w}_i^a$ than the Laplace approximation technique used for the inverse logit link function [8]. CPT is also more computationally efficient than A-CLUB, which solves a set of optimization problems at each update [2].

We put a prior distribution on $\mathbf{w}_i^a$ as $\mathcal{N}(\mathbf{m}_0, \mathbf{V}_0)$. Consequently, the posterior distribution on $\mathbf{w}_i^a$ can be approximated by $\mathcal{N}(\mathbf{m}, \mathbf{V})$ after observing a graded response $Y_{i,j}^a$, where

$$\mathbf{m} = \mathbf{m}_0 + \left(2Y_{i,j}^a - 1\right) \frac{\mathbf{V}_0 \mathbf{c}_j}{\sqrt{1 + \mathbf{c}_j{}^T \mathbf{V}_0 \mathbf{c}_j}} \frac{\mathcal{N}(z)}{\Phi_{\text{pro}}(z)}, \tag{2}$$

$$\mathbf{V} = \mathbf{V}_0 - \frac{\mathbf{V}_0 \mathbf{c}_j \mathbf{c}_j{}^T \mathbf{V}_0}{1 + \mathbf{c}_j{}^T \mathbf{V}_0 \mathbf{c}_j} \left(z + \frac{\mathcal{N}(z)}{\Phi_{\text{pro}}(z)}\right) \frac{\mathcal{N}(z)}{\Phi_{\text{pro}}(z)}, \tag{3}$$

$$z = \left(2Y_{i,j}^a - 1\right) \frac{\mathbf{m}_0^T \mathbf{c}_j}{\sqrt{1 + \mathbf{c}_j{}^T \mathbf{V}_0 \mathbf{c}_j}}. \tag{4}$$

The prior parameter $\mathbf{m}_0$ is initialized as an all-zero vector $\mathbf{0}$, and $\mathbf{V}_0$ is initialized as $\sigma^2 \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. Details of this approximation can be found in [9]. Algorithm 1 summarizes the CPT policy.

Algorithm 2: U-CPT

Input: A set of maximum likelihood-type student concept knowledge state estimates, $\boldsymbol{\mu}_j$ and the uncertainty of these estimates $\boldsymbol{\Sigma}_j$, $j = 1, 2, \ldots, N$, parameter σ^2

Output: PLA a_j for each student

$\mathbf{m}_i^a \leftarrow \mathbf{0}$, $\mathbf{V}_i^a \leftarrow \sigma^2\mathbf{I}$, $\forall\, i, a$

for $j \leftarrow 1$ **to** N **do**

  for $a \leftarrow 1$ **to** A **do**

    Sample $\widehat{\mathbf{w}}_i^a \sim \mathcal{N}(\mathbf{m}_i^a, \mathbf{V}_i^a)$, $\forall\, i$.

  $a_j \leftarrow \arg\max_a \sum_{i=1}^{Q} s_i \Phi_{\text{pro}}(\frac{\mathbf{c}_j^T \mathbf{w}_i^a}{\sqrt{1+(\mathbf{w}_i^a)^T \boldsymbol{\Sigma}_j \mathbf{w}_i^a}})$

  Update $\mathbf{m}_i^{a_j}$ and $\mathbf{V}_i^{a_j}$, $\forall\, i$, using the Laplace approximation

3.3. U-CPT: CPT with Uncertain Contexts

Since the student contexts, i.e., the student knowledge states, are estimated from the students' previous assessments, the estimates can be inaccurate, especially when the number of questions a student answers in previous assessments is small. We propose a new policy, U-CPT, to tackle uncertainty in the contexts, a previously largely unexplored aspect in contextual bandits literature that arises in the current application.

We assume that estimates of the context vectors are normally distributed as $\mathbf{c}_j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The mean parameter $\boldsymbol{\mu}_j$ is simply chosen as the maximum likelihood-type estimates obtained from SPARFA, while the covariance parameter $\boldsymbol{\Sigma}_j$ can be approximated as the inverse of the Fisher Information matrix using the method described in [2, Sec. 3.2]. With this notation, the probability that student j answers question i in Assessment 2 correctly after taking PLA a can be written as

$$\begin{aligned} p(Y_{i,j}^a = 1|\mathbf{w}_i^a) &= \int_{\mathbf{c}_j} P(Y_{i,j}^a|\mathbf{c}_j, \mathbf{w}_i^a) P(\mathbf{c}_j|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\ &= \int \Phi(\mathbf{c}_j^T \mathbf{w}_i^a) \mathcal{N}(\mathbf{c}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \mathrm{d}\mathbf{c}_j \\ &= \Phi(\frac{\mathbf{c}_j^T \mathbf{w}_i^a}{\sqrt{1+(\mathbf{w}_i^a)^T \boldsymbol{\Sigma}_j \mathbf{w}_i^a}}), \end{aligned} \tag{5}$$

where the last identity comes from the derivation of (2) (see [9] for details). Intuitively, when uncertainty in the context estimates is low, U-CPT reduces to CPT. When uncertainty is high, the denominator term will be large and thus results in the predicted success probability $p(Y_{i,j}^a = 1)$ for each PLA are close to each other, decreasing the impact of the contexts.

U-CPT operates in the same manner as CPT, except that we use the Laplace approximation [8] to update $\mathbf{m}_i^{a_j}$ and $\mathbf{V}_i^{a_j}$, $\forall\, i$, since closed-form updates no longer exist due to the presence of the variable $\mathbf{w}_i^a$ in the denominator of (5). Algorithm 2 summarizes the U-CPT policy.

3.4. KG: Online Knowledge Gradient

We now develop a third PLA selection policy using the knowledge gradient (KG) method [10]. We use the same underlying probabilistic model and the same PLA parameter estimation strategy as CPT.

In the *offline* KG policy, in each play, we select the arm that provides the greatest expected increase in a quantity termed *value* (see [11] for further details). Specifically, for the PLA selection problem, we define the value of student j answering question i as the probability that they will respond to this question correctly after taking the optimal PLA, i.e., $V_{i,j} = \max_a \Phi_{\text{pro}}(\mathbf{c}_j^T \mathbf{w}_i^a)$. The KG of a particular PLA a is the expected increase in value for a future student $j+1$ with the same concept knowledge, given that we select PLA a for student j. It is formally defined as $v_i^{a,j} = \mathbb{E}_{Y_{i,j}^a}[V_{i,j+1} - V_{i,j}]$, where $V_{i,j+1}$ is a random variable representing the possible updated values for $\mathbf{w}_i^a$ based on the response observed for student j after taking PLA a. See Alg. 3 for the implementation details, and refer to [4] for the full derivation of the KG for binary observations. Note that we drop the constant term $V_{i,j}$ since it does not affect the location of the maximum value.

The offline KG policy only selects arms to maximize the knowledge gained from each single measurement, which comes at a price of sacrificing short-term rewards. To adapt the KG policy for the MAB problem, which aims to maximize the cumulative observed reward, we use an *online* version of the KG policy. The online KG policy balances between providing an optimal PLA for student j (exploitation) and improving PLA selection for future students (exploration). We formally state the policy for PLA selection for student j as

$$a_{KG,j} = \arg\max_a \sum_{i=1}^{Q} s_i (V_{i,j} + (N-j) v_i^{a,j}). \tag{6}$$

Note that the term $(N-j)v^{a,j}$ approaches 0 as j increases. Although this policy assumes a finite value for N, it is also possible to derive a similar policy that uses a tunable parameter rather than $(N-n)$ to balance exploration and exploitation, when we have no knowledge of the number of students [12]. Algorithm 3 summarizes the online KG policy.

4. EXPERIMENTS

In this section, we test the performance of the CPT, U-CPT, and KG policies on improving the students' learning outcomes using two real-world educational datasets, and compare their performance to A-CLUB.

We use two datasets, Dataset 1 and 2, that were recorded from two high school classes: AP Physics and Biology. Both datasets consist of the binary-valued graded responses of each student to questions in their homework assignments. Let N denote the number of students. After finishing Assessment 1, students were then individually assigned a question at random from a set of A questions, which we denote as the set of PLAs.

Algorithm 3: KG

Input: A set of student concept knowledge state estimates, $\mathbf{c}_j$, $j = 1, 2, \ldots, N$, parameter σ^2
Output: PLA a_j for each student
$\mathbf{m}_i^a \leftarrow \mathbf{0}$, $\mathbf{V}_i^a \leftarrow \sigma^2 \mathbf{I}$, $\forall i, a$
for $j \leftarrow 1$ **to** N **do**
  for $a \leftarrow 1$ **to** A **do**
    for $i \leftarrow 1$ **to** Q **do**
      Define $p_i^{a,\pm} = \Phi_{\text{pro}}(\mathbf{c}_j^T(\pm \mathbf{m}_i^a))$
      Let $\tilde{\mathbf{w}}^\kappa$ be the solution to (2) with $Y = \kappa$
      Let $\Psi^+ = \max_{\mathbf{w}'} \Phi_{\text{pro}}(\mathbf{c}_j^T \mathbf{w}')$
        where $\mathbf{w}' = \{\mathbf{m}_i^{a'}\}_{a' \in A \setminus a} \cup \tilde{\mathbf{w}}^1$
      Let $\Psi^- = \max_{\mathbf{w}'} \Phi_{\text{pro}}(\mathbf{c}_j^T \mathbf{w}')$
        where $\mathbf{w}' = \{\mathbf{m}_i^{a'}\}_{a' \in A \setminus a} \cup \tilde{\mathbf{w}}^0$
      Calculate $v_i^{a,j} = p_i^{a,+}\Psi^+ + p_i^{a,-}\Psi^-$
  $a_j \leftarrow \arg\max_a \sum_{i=1}^{Q} s_i(p_i^{a,+} + (N-j)v_i^{a,j})$
  Update $\mathbf{m}_i^{a_j}$ and $\mathbf{V}_i^{a_j}$, $\forall i$, according to (2) and (3)

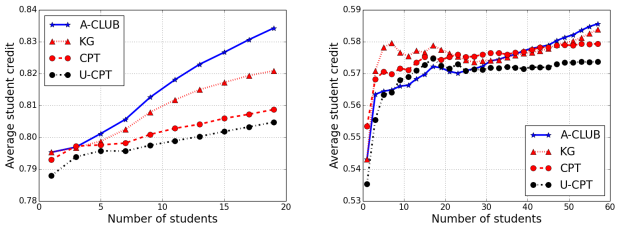
Students later completed the same homework set, Assessment 2, which consisted of Q questions. Note that all questions in this dataset have the same amount of credit ($s_i = 1$ for $i = 1 \ldots Q$). We have $N = 20$, $A = 2$, and $Q = 11$ for Dataset 1 and $N = 57$, $A = 4$, and $Q = 6$ for Dataset 2.

4.1. Experimental Setup and Evaluation Method

We formulate the PLA selection problem for this experiment as follows: Given students' estimated concept knowledge based on their responses to questions in Assessment 1 (contexts), our objective is to select a homework question (a PLA) to give each student out of the set of A available questions to maximize their performance on Assessment 2 (reward). We set K, the dimension of the concept knowledge vectors, to 3.[1] We tuned the algorithm parameters for each dataset to achieve the best performance.

We first generate estimates of the student context vector, $\mathbf{c}_j$, from SPARFA using the students' graded responses in Assessment 1. We then run each policy using these estimated context vectors. Since the dataset was collected in an "offline" way using a random PLA selection policy, we use the unbiased offline evaluation method [13] to evaluate the performance of the policies. We do so by using only students whose actual PLA selection matches those selected by the policies. Moreover, since MAB policies make PLA selections sequentially for each student, their performance depends on the order of the incoming students. Therefore, we randomly permute the ordering of students and average our results over 2000 random permutations.

[1] In our experiments, we observed that the choice of K has minimal impact on the performance of the policies as long as $K \ll L$, where L is the number of questions in Assessment 1.



(a) Dataset 1 (b) Dataset 2

Fig. 1. Average student credit on Assessment 2 vs. number of students the PLA selection policies train on for both datasets. Across all four policies, average student credit increases as the number of students increases.

4.2. Results and Discussion

Figure 1 plots the performance of each PLA selection policy over the number of students in the training set. We measure performance by the average student credit on Assessment 2 that is normalized to be in $[0, 1]$. The average credit for the policy that randomly selects PLAs is 0.791 for Dataset 1 and 0.573 for Dataset 2. Given more student data to train on, all four of our PLA selection policies achieve higher average test credits than the random policy.

While A-CLUB achieves better performance over the three Bayesian policies in Dataset 1, the results for Dataset 2 indicate that when the number of students in the training set is small (e.g., a typical high school class has no more than 20-30 students), the Bayesian policies (particularly KG) achieve superior performance. This suggests that the optimal policy is dataset-specific, and there may not be a single policy that works best for all classroom settings.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed three new policies for selecting personalized learning actions for students given estimates of their prior knowledge, in order to maximize their performance on a follow-up assignment. We validated our results using two real-world educational datasets and observe improvements in the average student credit as the policies train on data from more students. Our results demonstrate that our three new policies achieve comparable performance on improving students' future performance to existing contextual bandit policies for the task of selecting personalized practice questions.

Possible avenues of future work include i) extending the ideas of uncertainty in contexts to other algorithms and theoretically investigating its impact on regret, ii) exploring the correlation between the rewards of each PLA and each question, and iii) employing a Markov decision process [14] framework to extend the current work from selecting a single personalized learning action to crafting a fully personalized learning schedule for each student throughout the duration of a course.

6. REFERENCES

[1] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *J. Machine Learning Research*, vol. 15, pp. 1959–2008, June 2014.

[2] A. S. Lan and R. G. Baraniuk, "A contextual bandits framework for personalized learning action selection," in *Proc. 9th Intl. Conf. on Educational Data Mining*, June 2016, pp. 424–429.

[3] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, Dec. 1933.

[4] Y. Wang, C. Wang, and W. Powell, "The knowledge gradient for sequential decision making with stochastic binary feedbacks," in *Proc. 33rd Intl. Conf. on Machine Learning*, June 2016, pp. 1138–1147.

[5] M. D. Reckase, *Multidimensional Item Response Theory*, Springer, 2009.

[6] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Advances in Neural Information Processing Systems*, Dec. 2011, pp. 2249–2257.

[7] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. 25th Ann. Conf. on Learning Theory*, June 2012, vol. 23, pp. 39.1–39.26.

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning*, MIT Press, 2006.

[9] A. S. Lan, C. Studer, and R. G. Baraniuk, "Time-varying learning and content analytics via sparse factor analysis," in *Proc. 20th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Aug. 2014, pp. 452–461.

[10] P. Frazier, W. Powell, and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM Control and Optimization*, vol. 47, no. 5, pp. 2410–2439, 2008.

[11] W. B. Powell and I. O. Ryzhov, *Optimal Learning*, John Wiley & Sons, 2012.

[12] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, "The knowledge gradient algorithm for a general class of online learning problems," *Operations Research*, vol. 60, no. 1, pp. 180–195, Feb. 2012.

[13] L. Li, W. Chu, J. Langford, and X. Wang, "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms," in *Proc. 4th ACM Intl. Conf. on Web Search and Data Mining*, Feb. 2011, pp. 297–306.

[14] W. Powell, *Approximate Dynamic Programming: Solving The Curses of Dimensionality*, John Wiley & Sons, 2007.