A Contextual Bandits Framework for Personalized Learning Action Selection

Andrew S. Lan
Rice University
mr.lan@sparfa.com

Richard G. Baraniuk
Rice University
richb@sparfa.com

ABSTRACT

Recent developments in machine learning have the potential to revolutionize education by providing an optimized, personalized learning experience for each student. We study the problem of selecting the best personalized learning action that each student should take next given their learning history; possible actions could include reading a textbook section, watching a lecture video, interacting with a simulation or lab, solving a practice question, and so on. We first estimate each student's knowledge profile from their binary-valued graded responses to questions in their previous assessments using the SPARFA framework. We then employ these knowledge profiles as contexts in the contextual (multi-armed) bandits framework to learn a policy that selects the personalized learning actions that maximize each student's immediate success, i.e., their performance on their next assessment. We develop two algorithms for personalized learning action selection. While one is mainly of theoretical interest, we experimentally validate the other using a real-world educational dataset. Our experimental results demonstrate that our approach achieves superior or comparable performance as compared to existing algorithms in terms of maximizing the students' immediate success.

1. INTRODUCTION

In traditional classrooms, learning has largely remained a "one-size-fits-all" experience in which the instructor selects a single learning action for all students in their class, regardless of their diversity in backgrounds, learning goals, and abilities. The quest for a fully personalized learning experience began with the development of intelligent tutoring systems (ITSs) [6, 19, 38, 40]. However, to date, ITSs are primarily *rules-based*, meaning that building an ITS requires domain experts to consider every possible learning scenario that students can encounter and then manually specify the corresponding learning actions in each case. This approach is not scalable, since it is both labor-intensive and domain-specific.

Machine learning-based personalized learning systems [30] have shown great promise in reaching beyond ITS to scale to large numbers of subjects and students. These systems automatically create *personalized learning schedules*, a series of *personalized learning actions* (PLAs) for each individual student to take that maximizes their learning. Examples of PLAs include reading a textbook section, watching a lecture video, interacting with a simulation or lab, solving a practice question, etc. Instead of domain-specific rules, machine learning algorithms are used to select PLAs automatically by analyzing the data students generate as they interact with learning resources.

The general problem of creating a fully personalized learning schedule for each student can be formulated using the partially observed Markov decision process (POMDP) framework [31]. POMDPs utilize models on the students' latent knowledge states [23, 28] and their transitions [8, 11, 18, 22] to learn a PLA selection policy (a mapping from the knowledge state space to the set of learning actions) that maximizes a reward received in the possibly distant future (long-term learning outcome). Previous work applying POMDPs to personalized learning have achieved some degree of success [4, 9, 32, 33]. However, learning a personalized learning schedule using a POMDP is greatly complicated by the curse of dimensionality; the solution quickly becomes intractable as the dimensions of the state and action spaces grow [31]. Consequently, POMDPs have made only a limited impact in large-scale personalized learning applications involving large numbers of students and learning actions.

A more scalable approach to personalized learning is to learn a PLA selection policy using the *multi-armed bandits* (MAB) framework [10, 27], which is more suitable to optimizing students' success on immediate follow-up assessments (short-term learning outcome). The simplicity of the MAB framework makes it more practical than the POMDP framework in real-world educational applications, since it requires far less training data.

1.1 Contributions

In this paper, we study the problem of selecting PLAs for each student given their learning history using MABs. We first estimate each student's latent concept knowledge profile from their learning history (specifically, their binary-valued graded responses to questions in previous assessments) using the sparse factor analysis (SPARFA) framework [23]. Then, we use these concept knowledge profiles as contexts in the contextual (multi-armed) bandits framework to learn a policy to select PLAs for each student that maximize their performance on the follow-up assessment.

We develop two algorithms for PLA selection. The first algorithm, CLUB, has theoretical guarantees on its ability to identify the optimal PLA for each student. The second algorithm, A-CLUB, is more intuitive and practical; we experimentally validate its performance using a real-world educational dataset. Our experimental results demonstrate

that A-CLUB achieves superior or comparable performance to existing algorithms in terms of maximizing students' immediate success.

1.2 Related work

The work in [27] applies an MAB algorithm to educational games in order to trade off scientific discovery (learning about the effect of each learning resource) and student learning. Their approach is context-free and thus not ideally suited for applications with significant variation among the knowledge states of individual students. Indeed, it can be seen as a special case of our work in this paper when there is no context information available.

The work in [36] applies a contextual bandits algorithm to the problem of selecting the optimal PLA for each student given their previous exposure to learning resources. In their approach, each dimension of the context vector corresponds to the students' exposure to one learning resource. Thus, the context space quickly grows large as the number of learning resources increases. Our approach, in contrast, performs dimensionality reduction on student learning histories using the SPARFA framework and uses the resulting student concept knowledge profiles as contexts. This feature enables our approach to be applied to datasets where student learning histories contain a large number of learning resources.

The work in [29] collects high-dimensional student–computer interaction features as they play an educational game and uses them to search for a good teaching policy. We emphasize that our approach can be applied to almost all educational applications, not just computerized educational games, since it only requires graded response data of some kind.

The works in [10] and [20] both use some form of expert knowledge to learn a teaching policy. The approach of [10], in particular, uses expert knowledge to narrow down the set of possible PLAs a student can take. Our approach, in contrast, requires no expert knowledge and is therefore fully data-driven and domain-agnostic.

The work in [26] fuses MAB algorithms with Gaussian process regression in order to reduce the amount of training data required to search for a good teaching policy. Their work requires the policy to be parameterized by a few parameters, while our framework does not and can thus learn more complicated policies using only reward observations.

The work in [35] found that various student response models, including knowledge tracing (KT) [11], IRT models [28, 34, 5], additive factor models (AFM) [8], and performance factor models (PFM) [16] can have similar predictive performance yet lead to very different teaching policies. While these results are indeed interesting, we emphasize that the focus of the current work is to develop policy learning algorithms rather than comparing student models.

2. PROBLEM FORMULATION

We study the problem of creating a personalized learning schedule for each student by selecting the PLA they should take based on their prior learning experience. We assume that a student's learning schedule consists of a series of assessments with PLAs embedded in between, a setting that is typical in traditional classrooms, blended learning environments, and online courses like MOOCs [12, 13]. Each PLA can correspond to studying a learning resource, e.g., reading a textbook section, watching a lecture video, conducting an interactive simulation, solving a practice question, etc., or a combination of several learning resources.[1] Assessment could be a pop-quiz with a single question, a homework set with multiple questions, or a longer exam. Each student's personalized learning schedule can be visualized as in Figure 1, where a PLA is taken between consecutive assessments (starting after Assessment 1).

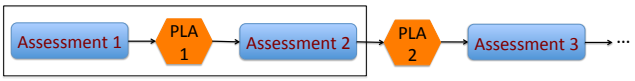

Figure 1: A personalized learning schedule.

The goal of this work is to select the optimal PLA for each student given their learning history (their graded responses to previous assessments) that maximizes their immediate success, i.e., the credit they receive on the following assessment. We aim to learn this learning action selection rule from data. For simplicity of exposition, we will place PLA 1 between Assessment 1 and Assessment 2 (as encased in the box in Figure 1) as a running example throughout the paper.

Let A denote the total number of PLAs available, let K denote the number of latent concepts covered up to Assessment 1, and let Q denote the number of questions in Assessment 2, with $s_i, i = 1, \ldots, Q$ the maximum credit of each question. Let $Y_{i,j}$ denote the binary-valued graded response of student j to question i, with $Y_{i,j} = 1$ denoting a correct response and $Y_{i,j} = 0$ an incorrect response. In order to pin down a feasible PLA selection algorithm, we make two simplifying assumptions: i) We assume that a reliable estimate of each student's latent concept knowledge vector (estimated from their graded responses to Assessment 1), denoted by $\mathbf{c}_j \in \mathbb{R}^K$, is available to the PLA selection algorithm. Such an estimate can be obtained using any IRT-like method, e.g., SPARFA [23]. ii) We assume that the PLA selected for each student will directly affect their performance on Assessment 2.

With this notation in place, we can restate our goal as selecting a PLA for student j, given their current concept knowledge[2] $\mathbf{c}_j$ in order to maximize their performance (i.e., their expected credit $\sum_{i=1}^{Q} s_i \mathbb{E}[Y_{i,j}]$) on Assessment 2.

2.1 Background on bandits

The multi-armed bandit (MAB) framework [3] studies the problem of a player trying to learn a policy that maximizes the total expected reward by playing (pulling the arms of) a collection of slot machines with a fixed number of trials and no prior information about each machine. Each machine has a fixed reward distribution that is unknown to the player. The key to maximizing the total expected reward is to find the right balance between exploration (playing

[1] Our notion of PLA is very general, and we do not restrict ourselves to studying a single learning resource.

[2] In practice, we augment $\mathbf{c}_j$ as $[\mathbf{c}_j^T \; 1]^T$ to add an "offset" parameter to each arm.

machines that might yield high rewards) and exploitation (repeatedly playing the machine with the highest observed reward). Analogously, a personalized learning system must strike a balance between testing the efficacy of every learning action (exploration) and maximizing the students' learning outcomes using observations on the actions (exploitation) [27].

Contextual (multi-armed) bandits [1, 2, 15, 24, 37] extend the MAB framework by accounting for the existence of additional information on the player and/or the machines, referred to as "contexts", in order to improve the policy. Our PLA selection problem fits squarely the contextual bandits framework, where the current estimates of students' concept knowledge correspond to the contexts and each PLA corresponds to an arm. Pulling an arm corresponds simply to selecting a PLA. In this paper, the context will include only information on the students. See Sec. 5 for a discussion on extending our framework to incorporate information on the learning resources into the contexts.

3. ALGORITHMS

The two algorithms we develop in this section are so-called upper confidence bound (UCB)-based algorithms [3]. These algorithms maintain estimates of the expected reward of each arm together with confidence intervals around these estimates, and iteratively update them as each new pull and its corresponding reward is observed. They then pull the arm with the highest UCB on the reward, which is equal to the expected reward plus the width of the confidence interval.

3.1 CLUB: An algorithm in theory

We first develop the *contextual logistic upper confidence bound* (CLUB) algorithm in order to provide theoretical guarantees for the PLA selection problem. We assume that the binary-valued student responses to the questions in Assessment 2 are Bernoulli random variables with success probabilities following a logistic model

$$p(Y_{i,j_{a_s}}=1)=\Phi_{\log}(\mathbf{c}_{j_{a_s}}^T\mathbf{w}_i^a)=\frac{1}{1+e^{-\mathbf{c}_{j_{a_s}}^T\mathbf{w}_i^a}},\ s=1,\ldots,n_a,$$

where $\mathbf{w}_i^a \in \mathbb{R}^K$ is the parameter vector that characterizes the students' responses to question i after taking PLA a. Also, j_{a_s} denotes the index of the s^{th} student taking PLA a, and n_a denotes the total number of students taking PLA a. $\Phi_{\log}(\cdot)$ denotes the inverse logit link function.

The maximum-likelihood estimate (MLE) of $\mathbf{w}_i^a$ is

$$\widehat{\mathbf{w}}_i^a = \arg\min_{\mathbf{w}} -\sum_{s=1}^{n_a}\log p(Y_{i,j_{a_s}}|\mathbf{c}_{j_{a_s}},\mathbf{w}), \qquad (1)$$

which can be computed using standard logistic regression algorithms [17] whenever the MLE exists (see [39, Sec. 5.1] for a detailed discussion on the conditions under which the MLE exists).

As detailed in Algorithm 1, CLUB maintains MLEs of the parameter vector $\mathbf{w}_i^a$ of each PLA together with a confidence interval around it. Then, after receiving a student's concept knowledge vector $\mathbf{c}_j$, CLUB selects the PLA with the highest UCB on the expected credit on the student's following assessment.

Algorithm 1: CLUB

Input: A set of student concept knowledge state estimates $\mathbf{c}_j$, $j=1,2,\ldots$, and parameters $\lambda_0,\delta,\eta,\epsilon$
Output: PLA a_j for each student, $j=1,2,\ldots$
$\text{MLE}_{\text{all exist}} \leftarrow$ False, $n_a \leftarrow 0,\ \forall\, a$
for $j \leftarrow 1$ **to** ∞ **do**
  if $\text{MLE}_{\text{all exist}}$ **then**
    Estimate $\widehat{\mathbf{w}}_i^a,\ \forall\, i,a$ according to (1)
    $\mathbf{\Sigma}_a \leftarrow \lambda_0\mathbf{I}_K + \sum_{s=1}^{n_a}\mathbf{c}_{j_{a_s}}\mathbf{c}_{j_{a_s}}^T,\ \forall\, a$
    $a_j \leftarrow$
    $\arg\max_a \sum_{i=1}^{Q} s_i(\Phi_{\log}(\mathbf{c}_j^T\widehat{\mathbf{w}}_i^a) + c_i(n_a)\sqrt{\mathbf{c}_j^T\mathbf{\Sigma}_a^{-1}\mathbf{c}_j})$
  else
    Randomly select a_j among PLAs where $\exists\, i$ s.t. $\widehat{\mathbf{w}}_i^a$ does not exist
  $n_{a_j} \leftarrow n_{a_j}+1$
  $\text{MLE}_{\text{all exist}} \leftarrow$ True
  for $a \leftarrow 1$ **to** A **do**
    for $i \leftarrow 1$ **to** Q **do**
      if $\widehat{\mathbf{w}}_i^a$ does not exist (verified via [39, Thm. 2]) **then**
        $\text{MLE}_{\text{all exist}} \leftarrow$ False

The constants in Algorithm 1 are given by $c_i(n_a) = \sqrt{2K(3+2\log(1+2a_m^2/\lambda_0))\log n_a K/\delta}/b_{i,a}$, where $a_m = \sqrt{K+2\sqrt{K\log(1/\eta)}+2\log(1/\eta)}$ and $b_{i,a} = 1/(2+e^{\|\mathbf{w}_i^a\|_2 a_m}+e^{-\|\mathbf{w}_i^a\|_2 a_m})$, and $0<\delta,\eta\ll 1$. Algorithm 1 exhibits theoretical optimality guarantees (omitted due to space constraints and available at `www.sparfa.com` [21]).

3.2 A-CLUB: An algorithm in practice

Since in practice we do not know the values of the constants $\Delta_{a,j}$ and also need to set the parameters ϵ, δ, and η, Algorithm 1 and its theoretical guarantees are not directly applicable. Furthermore, as the number of students grows, the confidence bounds around the estimates of each PLA's parameters might become overly pessimistic, causing the algorithm to over-explore [15]. Therefore, we now develop a second CLUB-like algorithm that leverages the asymptotic normality of the MLEs of the PLA parameters [14]. The asymptotic normality property states that, as the number of students grows large, the estimation error of the parameter $\mathbf{w}_i^a$ for each PLA converges to a normally distributed random vector with zero mean and a covariance matrix that is a scaled inverse of the Fisher information matrix

$$\mathbf{F}_a := \sum_{s=1}^{n_a}\frac{\mathbf{c}_{j_{a_s}}\mathbf{c}_{j_{a_s}}^T}{2+e^{\mathbf{c}_{j_{a_s}}^T\mathbf{w}_i^a}+e^{-\mathbf{c}_{j_{a_s}}^T\mathbf{w}_i^a}}.$$

Thus, we can build a confidence ellipsoid around the point estimate generated by (1), albeit asymptotically. In practice, since the true values of the parameters $\mathbf{w}_i^a\ \forall\, i,a$ are unknown, we will use their estimates $\widehat{\mathbf{w}}_i^a$ to approximate the Fisher information matrix.

Armed with the confidence ellipsoid, we can now compute the upper bound of the expected response of student j on each question in Assessment 2 after taking PLA a. This cor-

Algorithm 2: A-CLUB

Input: A set of student concept knowledge state estimates, $\mathbf{c}_j$, $j = 1, 2, \ldots$, parameter α
Output: PLA a_j for each student
$\text{MLE}_{\text{all exist}} \leftarrow$ False, $n_a \leftarrow 0$, $\forall a$
for $j \leftarrow 1$ **to** ∞ **do**
  if $\text{MLE}_{\text{all exist}}$ **then**
    Estimate $\widehat{\mathbf{w}}_i^a$, $\forall i, a$ according to (1)
    $\mathbf{F}_a \leftarrow \lambda_0 \mathbf{I}_K + \sum_{s=1}^{n_a} \frac{\mathbf{c}_j \mathbf{c}_j^T}{2+e^{\mathbf{c}_{j_s}^T \widehat{\mathbf{w}}_i^a} + e^{-\mathbf{c}_{j_s}^T \widehat{\mathbf{w}}_i^a}}$, $\forall a$
    $a_j \leftarrow$
    $\arg\max_a \sum_{i=1}^{Q} s_i \Phi_{\log}(\mathbf{c}_j^T \widehat{\mathbf{w}}_i^a + \sqrt{\alpha (\mathbf{c}_j^T \mathbf{F}_a^{-1} \mathbf{c}_j)/n_a})$
  else
    Randomly select a_j among PLAs where $\exists i$ s.t. MLE of $\mathbf{w}_i^a$ does not exist
  $n_{a_j} \leftarrow n_{a_j} + 1$
  $\text{MLE}_{\text{all exist}} \leftarrow$ True
  for $a \leftarrow 1$ **to** A **do**
    for $i \leftarrow 1$ **to** Q **do**
      if MLE does not exist for $\mathbf{w}_i^a$ (verified via [39, Thm. 2]) **then**
        $\text{MLE}_{\text{all exist}} \leftarrow$ False

responds to the following constrained optimization problem[3]

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & -\frac{1}{1+e^{-\mathbf{c}_j^T \mathbf{w}}} \\ \text{subject to} \quad & (\mathbf{w} - \widehat{\mathbf{w}}_i^a)^T \mathbf{F}_a (\mathbf{w} - \widehat{\mathbf{w}}_i^a) \leq \alpha / n_a, \end{aligned}$$

where α is a parameter controlling the size of the confidence ellipsoid and thus the amount of exploration. The solution to this problem is given by $\mathbf{w} = \widehat{\mathbf{w}}_i^a + \sqrt{\frac{\alpha}{n_a \mathbf{c}_j^T \mathbf{F}_a^{-1} \mathbf{c}_j}} \mathbf{F}_a^{-1} \mathbf{c}_j$. Therefore, we obtain an upper bound for the expected grade for student j on question i after taking PLA a as $\Phi_{\log}(\mathbf{c}_j^T \widehat{\mathbf{w}}_i^a + \sqrt{\alpha \mathbf{c}_j^T \mathbf{F}_a^{-1} \mathbf{c}_j / n_a})$. We thus arrive at Algorithm 2, which we dub asymptotic CLUB (A-CLUB).

4. EXPERIMENTS

In this section, we validate our algorithms experimentally on personalized cohort selection using a college physics course dataset. We will compare the performance of Algorithm 2 against other baseline (contextual) MAB algorithms. We do not compare Algorithm 1, since its theoretical bounds are usually too pessimistic in practice [15]. For comparisons using two additional datasets, see [21].

Dataset. The dataset consists of the binary-valued graded responses in a semester-long physics course administered on OpenStax Tutor [30] with $N = 39$ students answering 286 questions. Cognitive science experiments were conducted in this course to test the effect of spacing versus massed practice on the students' long-term retrieval performance of knowledge [7]. For this purpose, the students were randomly divided into two cohorts containing 19 and 20 students. There are

[3]We assume $\mathbf{c}_j$ is non-zero; otherwise we would simply select a PLA at random.

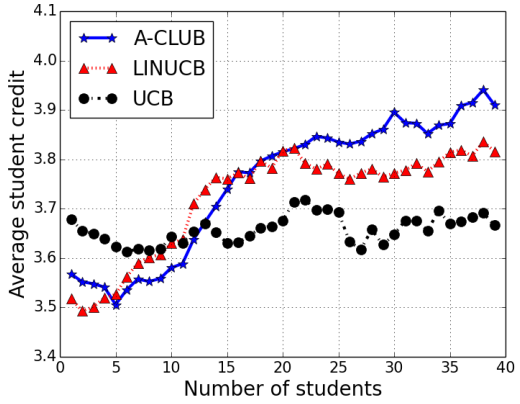

Figure 2: Average student credit on Assessment 5 vs. number of students used by three algorithms. Student performance on the follow-up assessment increases as the algorithms have access to more training data. Concretely, using data from 38 students, A-CLUB finds a PLA selection policy whereby students perform approximately 10% better than selecting randomly.

a total of 11 weekly assessments and 3 review assessments throughout the course. In the first three assessments, both cohorts received the same set of assessment questions. Starting from Assessment 4, apart from the same set of assessment questions both cohorts received on the concepts covered in the current week, each cohort also received additional, different questions. One cohort received spaced practice questions related to the concepts they learned several weeks earlier, while the other cohort received massed practice questions related to the concepts they learned in the current week. Each cohort received some spaced practices and some massed practices throughout the semester so that the sets of questions assigned to each cohort were identical in the end.

Experimental setup. Since the students in Cohorts 1 and 2 receive different sets of questions on Assessment 4, we investigate how this difference affects their learning on the concepts they learn next, i.e., their performance on Assessment 5. Treating each cohort as a PLA, our goal is to maximize the students' performance on Assessment 5 by assigning them to the cohort (selecting the PLA) that benefits them the most. Therefore, in our setting the number of PLAs is $A = 2$. We take the students' graded responses to questions in Assessments 1–3 and apply SPARFA to estimate each student's K-dimensional concept knowledge vector $\mathbf{c}_j$, which we use as the context. We set the number of concepts to $K = 3$.[4] Since Cohorts 1 and 2 also receive different questions for Assessment 5 as part of the spacing vs. mass retrieval practice experiment on new concepts covered in Week 5, we take the set of $Q = 5$ questions shared between the two cohorts to evaluate their performance. Since MAB algorithms analyze students sequentially, we randomly permute the order of the students and average our results over 2000 random permutations.

[4]In our experiments, we have found that the performance of SPARFA and A-CLUB is robust to the number of concepts K as long as $K \ll Q$.

	A-CLUB	LINUCB	UCB
Training set	**3.69**	3.68	3.65
Test set	**3.89**	3.77	3.70

Table 1: Performance comparison of A-CLUB against two baseline algorithms on personalized cohort selection on the physics course dataset. A-CLUB outperforms the other algorithms in terms of average student credit on the follow-up assessment (out of a full credit of 5) on both the training and test sets.

Evaluation method. We use the unbiased offline evaluation approach in [24, 25] to evaluate our algorithms. We use only the students that were actually assigned to the same cohort as chosen by our algorithms and ignore the other students. This approach evaluates the decision making algorithms under the scenario where the data is collected in a specific "off-line, off-policy" manner, i.e., the data is collected by selecting PLAs for each student uniformly at random across every PLA, as opposed to a more typical MAB setting where PLAs are chosen for students sequentially given the observed follow-up assessment performance of previous students. Such a scenario fits our experimental setup well and yields an unbiased estimate of the expected reward for each student [25]. We use the students' total credit on Assessment 5, i.e., $\sum_{i=1}^{Q} s_i Y_{i,j}$, as the metric to evaluate the performance of the algorithms.

Results and discussion. Figure 2 shows the students' average credit (out of a full credit of 5) on Assessment 5 vs. the number of students the algorithms use for the algorithms A-CLUB, LINUCB [24], and UCB [3]. The parameters in every algorithm were tuned for best performance. We see that the average student credit increases as the number of students the algorithms observe increases, i.e., the algorithms improve their PLA selection policy as they see more training data. As a concrete example, by comparing the average student credit at the first and last points on the curves, we see that A-CLUB has found a policy that yields students approximately 10% more credit than a policy that selects PLAs randomly.

Following the approach in [24], we also conduct an experiment by separating the dataset into a training set with 80% of the students and a test set with 20% of the students, to validate both the efficiency (performance on the training set) and efficacy (performance on the test set) of A-CLUB. We train the above three algorithms on the training set and apply the learned PLA selection policy to the test set, and report the average student credit obtained on both sets. Table 1 indicates that A-CLUB outperforms the other algorithms on both the training set and the test set. Better performance on the test set means that A-CLUB learns a better policy than the other algorithms, while better performance on the training set means that it learns this policy very quickly as the amount of training data increases.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a contextual (multi-armed) bandits framework for PLA selection that maximizes students' immediate success on a follow-up assessment, given the latent concept knowledge estimated from their binary-valued graded responses to questions in previous assessments. Our contextual logistic upper confidence bound (CLUB) algorithms learn such a policy and achieve better or comparable performance than baseline algorithms.

There are a number of avenues for future work. First, our context vectors are indexed by student features only, while in the general contextual bandits setting the contexts can be indexed by both student features and features of the learning resources. SPARFA-Trace [22], a recently developed framework for time-varying learning and content analytics, features a mechanism to analyze the content, quality, and difficulty of all kinds of learning resources (i.e., textbook sections, lecture videos, practice questions, etc.). We can apply this approach to extract features from the learning resources that we can integrate into the contexts in our algorithms. Second, we can incorporate an additional PLA that corresponds to "no action", due to the cost of taking actions, as considered in [36]. This extension would enable students with high knowledge on the concepts covered to avoid repeated practice and advance more quickly to new concepts. Third, we are interested in integrating our approach into more sophisticated contextual bandit algorithms, e.g., [37] to reap further performance improvements.

6. ACKNOWLEDGEMENTS

Thanks to Phillip Grimaldi, former pinball champion of Indiana, for collecting the physics course dataset and Mihaela van der Schaar for insightful suggestions. Visit our website `www.sparfa.com`, where you can learn more about the SPARFA project and purchase SPARFA t-shirts and other merchandise.

7. REFERENCES

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, Dec. 2011.

[2] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learning Research*, 3:397–422, Mar. 2003.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, May 2002.

[4] T. Barnes and J. Stamper. Toward automatic hint generation for logic proof tutoring using historical student data. In *Proc. 9th Intl. Conf. on Intelligent Tutoring Systems*, pages 373–382, June 2008.

[5] Y. Bergner, S. Droschler, G. Kortemeyer, S. Rayyan, D. Seaton, and D. Pritchard. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In *Proc. 5th Intl. Conf. on Educational Data Mining*, pages 95–102, June 2012.

[6] P. Brusilovsky and C. Peylo. Adaptive and intelligent web-based educational systems. *Intl. J. Artificial Intelligence in Education*, 13(2-4):159–172, Apr. 2003.

[7] A. C. Butler, E. J. Marsh, J. Slavinsky, and R. G. Baraniuk. Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review*, 26(2):331–340, June 2014.

[8] H. Cen, K. R. Koedinger, and B. Junker. Learning factors analysis – A general method for cognitive model evaluation and improvement. In *Proc. 8th Intl. Conf. on Intelligent Tutoring Systems*, pages 164–175, June 2006.

[9] M. Chi, K. VanLehn, D. Litman, and P. Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1-2):137–180, Jan. 2011.

[10] B. Clement, D. Roy, P. Oudeyer, and M. Lopes. Multi-armed bandits for intelligent tutoring systems. *J. Educational Data Mining*, 7(2):20–48, 2015.

[11] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253–278, Dec. 1994.

[12] Coursera. https://www.coursera.org/, 2016.

[13] edX. https://www.edx.org/, 2016.

[14] L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, Mar. 1985.

[15] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, Dec. 2010.

[16] Y. Gong, J. E. Beck, and N. T. Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Proc. 10th Intl. Conf. on Intelligent Tutoring Systems*, pages 35–44, June 2010.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2010.

[18] M. Khajah, R. Wing, R. Lindsey, and M. Mozer. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *Proc. 7th Intl. Conf. on Educational Data Mining*, pages 99–106, July 2014.

[19] K. R. Koedinger, J. R. Anderson, W. H. Hadley, and M. A. Mark. Intelligent tutoring goes to school in the big city. *Intl. J. Artificial Intelligence in Education*, 8(1):30–43, 1997.

[20] K. R. Koedinger, E. Brunskill, R. S. Baker, E. A. McLaughlin, and J. Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[21] A. S. Lan and R. G. Baraniuk. A contextual bandits framework for personalized learning action selection – Extended version. Technical report, Rice University, 2016.

[22] 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*, pages 452–461, Aug. 2014.

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

[24] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc.19th Intl. Conf. on World Wide Web*, pages 661–670, Apr. 2010.

[25] 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*, pages 297–306, Feb. 2011.

[26] R. Lindsey, M. Mozer, W. Huggins, and H. Pashler. Optimizing instructional policies. In *Advances in Neural Information Processing Systems*, pages 2778–2786, Dec. 2013.

[27] Y. Liu, T. Mandel, E. Brunskill, and Z. Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *Proc. 7th Intl. Conf. on Educational Data Mining*, pages 161–168, July 2014.

[28] F. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.

[29] T. Mandel, Y. Liu, S. Levine, E. Brunskill, and Z. Popovic. Offline policy evaluation across representations with applications to educational games. In *Proc. Intl. Conf. on Autonomous Agents and Multi-agent Systems*, pages 1077–1084, May 2014.

[30] OpenStaxTutor. https://openstaxtutor.org/, 2016.

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

[32] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by POMDP planning. In *Proc. 15th Intl. Conf. on Artificial Intelligence in Education*, pages 280–287, June 2011.

[33] A. N. Rafferty, M. M. LaMar, and T. L. Griffiths. Inferring learners' knowledge from their actions. *Cognitive Science*, 39(3):584–618, Apr. 2015.

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

[35] J. Rollinson and E. Brunskill. From predictive models to instructional policies. In *Proc. 8th Intl. Conf. on Educational Data Mining*, pages 179–186, June 2015.

[36] C. Tekin, J. Braun, and M. van der Schaar. eTutor: Online learning for personalized education. In *Proc. 40th IEEE Intl Conf. on Acoustics, Speech and Signal Processing*, pages 5545–5549, April 2015.

[37] C. Tekin and M. van der Schaar. RELEAF: An algorithm for learning and exploiting relevance. *IEEE J. Selected Topics in Signal Processing*, 9(4):716–727, June 2015.

[38] K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The Andes physics tutoring system: Lessons learned. *Intl. J. Artificial Intelligence in Education*, 15(3):147–204, Aug. 2005.

[39] D. Vats, C. Studer, A. S. Lan, L. Carin, and R. G. Baraniuk. Test size reduction via sparse factor analysis. *Preprint*, June 2014.

[40] B. P. Woolf. *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing E-learning*. Morgan Kaufman Publishers, 2008.