Multi-Armed Bandits for Addressing the Exploration/Exploitation Trade-off in Self Improving Learning Environment

Louis Faucon[1], Pierre Dillenbourg[1]

Ecole Polytechnique Fédérale de Lausanne, Switzerland,
louis.faucon@epfl.ch

Abstract. This project proposes the use of machine learning techniques such as Multi-Armed Bandits to implement self-improving learning environments. The goal of a self-improving learning environment is to perform good pedagogical choices while measuring the efficiency of these choices. The modeling of students is done using the LFA model and fitted on a dataset of university courses to allow to simulate students. Three experiments with simulated students are carried out and show that the Multi-Armed Bandit approach improves learning outcomes.

1 Introduction

Many systems allow to fit models on students learning behavior and then predict their performance [1, 3, 4], but very rarely do these models directly give a strategy to select activities for delivering optimal teaching. In that case, a common practice is to choose the activity that leads to the maximum prediction, but more advanced methods for this have been studied such as Partially Observable Markov Decision Processes [5]. The selection of optimal pedagogical choices becomes more challenging when the system has no prior information about the teaching material. In that case, the system must evaluate the quality of its choices while still delivering the best teaching possible. This challenge is commonly called exploration-exploitation trade-off and usually solved using techniques such as Multi-Armed Bandit (MAB) algorithms. MAB approaches have already been used in the context of Intelligent Tutoring Systems [2]. This project uses MAB for optimizing in a context with no prior information about the teaching material. To the best of our knowledge this area has not received a lot of attention in educational research as most implemented systems rely on expert knowledge or previously generated data. Our belief is that data-driven approaches as presented in this paper would be more easily used on the large amount of pedagogical content already available online, for example in MOOCs.

2 Activity selection procedure

The model used for this work is similar to the LFA model [1]. It supposes that students gain a fixed amount knowledge when doing a learning activity independently of their current knowledge or performance (success or failure) on this

activity. Equation 1 shows the measure of knowledge. The parameters $\lambda_{i,j}$ represent the amount of knowledge useful to activity i acquired by doing activity j before, the parameter β_i is the prior knowledge of students and n_j is the number of times a student interacted with activity j. Equation 2 shows how the measure of knowledge is converted to a probability of success.

$$m(n,i) = \beta_i + \sum_{j \in \mathcal{A}} n_j * \lambda_{i,j} \tag{1}$$

$$p(n,i) = \frac{1}{1 + e^{-m(n,i)}} \tag{2}$$

Given the simplicity of this model, the task of selecting an optimal subset of learning activities is solved by a greedy approach. The challenge addressed in this work is the exploration-exploitation trade-off. It is the difficulty of making good pedagogical choices while testing enough different possibilities to estimate correctly the parameters of the model. Indeed, always selecting the optimal choice according to the first measurements made by the system directly leads to never testing other solutions that might end up being better than the believed optimal choice. The MAB Upper Confidence Bounds (UCB) algorithm proposes an elegant solution to this problem. The idea of the algorithm is to consider optimistic evaluations of the choices efficiency by adding a term which takes the measurement uncertainty into account. Equation 3 gives the score evaluated by the UCB algorithm to select an optimal activity to increase the chances of students on activity 0 according to the model in equations 1 and 2. In the equation, the variable t is the total number of choices made by the algorithm, t_i is the number of times the activity i has been selected.

$$s(i) = \frac{1}{1 + e^{-\beta_0 - \lambda_{0,i}}} + \sqrt{\frac{log(t)}{t_i}} \tag{3}$$

3 Methodology

The first step is to compute the parameters of the model for a given set of learning activities to then be able to simulate learning processes of the students using the probabilities given by the model. For this, we use a dataset containing the grades of about 6500 EPFL students on 36 courses and the dates at which the courses were taken. From our choice of dataset, it follows that the learning activities mentioned in the modeling are semester-long university courses. The simulation engine then uses the computed model to generate student learning processes. It also communicates with an optimization engine which selects the activities to perform and can observe students results. The goal of the optimization engine is to find the parameters of the real model from the observation of performance while providing optimal teaching to students. The experiment compares three types of optimization strategies for activity selection: the Multi-Armed Bandit strategy which selects activities according to the rule described on equation 3;

the Epsilon strategy, which first gives random activities to a small fraction of the students (exploration), then computes the model's parameters and optimizes the teaching (exploitation); the Random strategy which always selects activities at random.

	NA	NAS	NS	E1	E2	E3
EXP 1	2	1	1000	10	25	50
EXP 2	10	2	3000	100	500	1000
EXP 3	30	3	5000	200	1000	2000

Table 1: Parameters of the three experiments

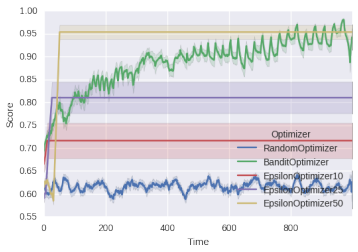
In the simulation students first learn through a sequence of activities, then do a test. Three experiments are repeated 100 times and run for different values of the following parameters (see table 1): the number of available activities (NA), the number of activities each student performs before the test (NAS), the number of students (NS) and the number of students assigned random activities for three Epsilon strategies (E1, E2, and E3).

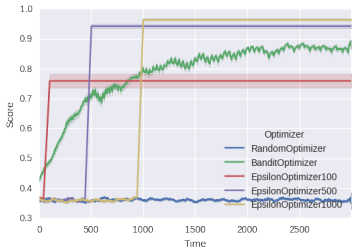
4 Results

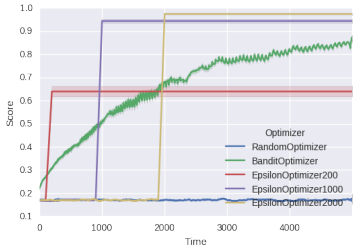
Figure 1 shows the evolution in time of the five strategies for our three experiments. The score in the three graphs is computed as the ratio of achieved learning gain divided by the optimal possible learning gain and the time axis represents the number of students. We can observe as expected that the three Epsilon strategies perform as well as a random selection at first, then reach higher performance after computing the model parameters. In the other hand, the Bandit strategy steadily increases its performance as it delivers activities to more and more students. The results show that the Bandit strategy did not perform better than the baseline Epsilon strategies when a good epsilon is chosen. It nevertheless reaches comparable results and scores higher than Epsilon strategies with inadequate choices of the parameter epsilon. As the choice of the epsilon parameter will be highly dependent on the teaching material while the Bandit strategy does not require any parameter, the result still appears as satisfactory for the Bandit strategy. It can also be noted that the Bandit strategy has the advantage of not intentionally delivering poor activity selection for the sake of exploration and thus is fairer than the Epsilon strategy.

5 Discussion and Future Work

The first limitation of this work is the simplicity of the chosen model. Indeed, it does not consider the order in which activities are selected neither does it consider individual differences between students. This results in the simulated students learning all in the same way, which is certainly not the case for real students. The simplicity of the modeling was useful to gather the focus of this







Strategy	EXP 1	EXP 2	EXP 3
Random	592	1714	2714
Bandit	657	2209	4037
Epsilon1	612	2177	3982
Epsilon2	635	**2263**	**4204**
Epsilon3	**669**	2168	3861

Fig. 1: Performance of optimization strategies of experiments 1 (top-left), 2 (top-right), and 3 (bottom-left) and results (bottom-right)

work on the MAB process, but similar approaches will benefit using individualized dynamic models. Another limitation of this work is the binary assessment. Even if the Bandit strategy does not surpass the Epsilon strategy, it still provides better teaching for the first students. It seems that a continuous scale could measure that failing students under the Bandit optimizer would have received better teaching than under the Epsilon optimizer. Given the success of the optimization procedures on simulated students, we made our implementation available [6]. We will then be trying to reproduce this results with real students, first in our own online learning environment, then at larger scale in a MOOC.

References

1. Cen, Hao, Kenneth Koedinger, and Brian Junker, 2006, Learning factors analysis-a general method for cognitive model evaluation and improvement, Intelligent tutoring systems. Vol. 4053.
2. Clement, Benjamin and Oudeyer, Pierre-Yves and Roy, Didier and Lopes, Manuel, 2014, Online optimization of teaching sequences with multi-armed bandits, Educational Data Mining 2014
3. Corbett, Albert T and Anderson, John R, 1994, Knowledge tracing: Modeling the acquisition of procedural knowledge User modeling and user-adapted interaction Springer
4. Pavlik Jr, Philip I and Cen, Hao and Koedinger, Kenneth R, 2009, Performance Factors Analysis–A New Alternative to Knowledge Tracing., Online Submission, ERIC
5. Rafferty, A.N., Brunskill, E., Griffiths, T.L. and Shafto, P., 2011, Faster Teaching by POMDP Planning, In AIED (pp. 280-287). Vancouver
6. https://github.com/chili-epfl/ASILE