Will You Get It Right Next Week: Predict Delayed Performance in Enhanced ITS Mastery Cycle

Xiaolu Xiong, Shoujing Li, Joseph E. Beck
Department of Computer Science
Worcester Polytechnic Institute
100 Institute Road, Worcester, MA, USA
xxiong@wpi.edu, sli@wpi.edu, josephbeck@wpi.edu

Abstract

Researchers of Intelligent Tutoring Systems (ITS) and Educational Data Mining (EDM) have focused increasing attention on predicting students' long-term retention performance as well as attempting to find effective methods to help improve student knowledge retention. Wang and Beck proposed a system which allows ITS to strive for student long-term mastery learning. This paper describes our implemented work of such a system for improving student retention along with a model to predict student performance for delayed retention tests; this incorporates features of student behavior and performance levels in the system. Using this model, we analyzed the data of 27,451 mathematical problems that 662 students in the 2012 fall semester attempted to solve or were successful in solving. We found that after students successfully master the skill, the number of those who attempted solving problems during the process of achieving mastery is predictive of delayed retention test performance. Specifically, on the 7-day retention test, 82% of students who try to master a skill in 3 or 4 attempts did so correctly, while students who required 5 to 8 attempts to master a skill achieved a rate of 70%. Furthermore, we propose that using the prediction model to guide the improvement of our tutorial decision-making on when we should test students also help them to better retain skills.

Keywords: Educational data mining, intelligent tutoring system, performance factors analysis, student modeling, knowledge retention

1. Introduction

Currently, most ITS (Beck 2003) present a sequence of problems and evaluate student performance directly after the student finishes solving or attempts to solve these problems to see if the student mastered the given skill. The exact definition of mastery varies, but it typically involves recent students' performance level. This process of detecting mastery has neither the mechanism for the system to review students' knowledge after a time period; nor does it know about students' long-term performance. However, studies of psychology (Anderson 1993, Cepeda 2006, George and John 1994) and EDM suggested that students do not always retain what they have learned. The local measure of student performance is insufficient and dangerous for ITS to promote a student just on the basis of short-term performance. This applies specifically to a cumulative subject such as mathematics: we are more concerned with students' capability to remember the knowledge that they acquired over a long period of time. Some researchers have carried out work on long-term performance prediction. Qiu et al. (Qiu, et al 2011) extended the Knowledge Tracing (KT) model, to take into account that students exhibit the forgetting feature when a day elapses between problems in the tutor system. Pavlik and Anderson (Pavlik and Anderson 2010) studied alternative models of practice and forgetting what had been learnt; this confirmed most importantly the standard spacing effect in various conditions and showed that wide-spacing of practice provides increasing benefits as practice accumulates. This leads to students forgetting less afterwards as well. Furthermore in Wang and Beck's work (Wang and Beck 2012), the notion of mastery learning was expanded to take into account the long-term effect of learning and this identified several features; which are relevant to students' long-term knowledge. In addition, they proposed an enhanced system of an ITS mastery cycle that can be used to discover new problems in the EDM field which can then lead to a higher mastery learning rate. Figure 1 shows the structure of this system.

2. ASSISTments and ARRS

Inspired by the design of the enhanced ITS mastery cycle, we developed and deployed an extension called the Automatic Reassessment and Relearning System (ARRS) in the ASSISTments platform (www.assistments.org). The ASSISTments is a non-profit, web-based tutoring project for 4th through 10th grade mathematics tutoring (approximately 9 through 16 years of age). In the school year of 2011 to 2012, it served approximately 20,000 active students nationwide. One of the important compounds of ASSISTments is the Mastery learning problem set, which simplifies the notion of skill mastery to three consecutive correct responses with the number of attempted problems before students achieve mastery (this is called the *mastery speed*). Note that three problems for a skill represent the lower boundary for the amount of practice students require. However, if students make mistakes, they are required to obtain three correct answers in a row to additional problems. In fact, some students require over 20 practice attempts to reach mastery. ASSISTments limits the daily practice number for a skill at 10 attempts, so these students need multiple days to master a skill. In the summer of 2012, we adapted the idea of enhanced ITS mastery cycle by spacing practice effects as well as utilizing Mastery learning problem sets to create ARRS: this was consequently utilized by ASSISTments in September of 2012.

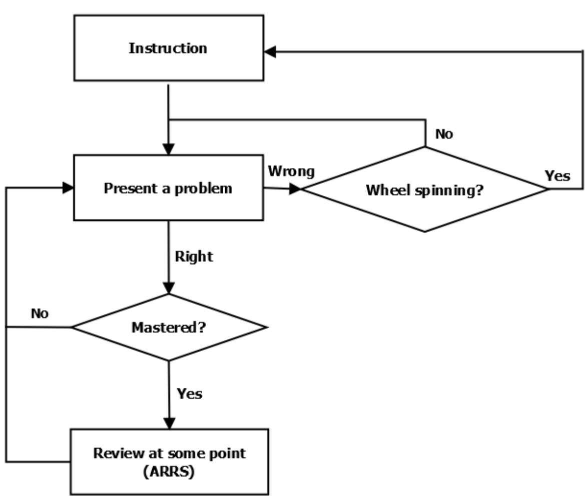

Figure 1. The enhanced ITS mastery cycle

The current workflow of ARRS is relatively simple: after classroom teaching of a certain skill, teachers using ASSISTments to assign Mastery learning problem set of that skill to students and students should first master the skill by completing the Mastery learning problem set; ARRS will then automatically reassess students on the same skill 7 days later with a reassessment test built from the same sets of problems the student already mastered. If students answer the problem correctly, we treat them as if they are still retaining this skill, and ARRS will test them two weeks later, a month later, and then finally two months after that. If students fail on one of the reassessment tests, they will be given an opportunity to relearn the topic with relearning problems and be re-tested again after the same amount of days in between tests. Note: in order to ensure every student completes retention tests, the system will assign tests to students even if they have not yet started acquiring a skill or have not started to achieve skill mastery; in other words, if they have not yet started the practice on the Mastery learning problem sets.

Two months after the deployment of ARRS in ASSISTments, 182 classes from 50 schools were using this system. As a result, we have 3422 students who finished 83,159 reassessment test problems and several hundred relearning problems. Each problem record was recorded straight after a student answered a problem and this contains relevant information including the identity of the student, the identity of the problem, the correctness of the answer, the date when the student answered this problem and the time that the student spent on solving this problem, as well as the required skill to answer the problem and the school grade of the problem. One of the important characteristics of this data is that it represents the students' long-term performance on different delayed time periods. Therefore we believe that we can use this data to build models to predict if students will remember a skill after a certain period of time and this helps in solving ITS decision-making problems. For example, if our models can tell for certain that a student appears likely to retain a skill, it is probably not necessary to keep presenting the item of that particular skill; and if it seems likely that a student will not master a skill, then it may be a good time to allow the student to relearn what he or she has already forgotten.

3. ARRS Data Analyses and Modeling

In this paper, we focused on the part of data that recorded reassessment tests; since most of the data was gathered during the first 7-day retention tests and 14-day retention tests, we conducted our analysis and study only on these pieces of data. We decided to build an extended version of Pavlik's Performance Factors Analysis (PFA; Pavlik et al. 2009) model that predicts students' performance on the delayed retention tests for these two different delay periods. Although we are not explicitly modeling students' retention and forgetting process, our data driven approach captures aspects of performance that relate to students' long-term retention of the material. PFA models track the number of correct and incorrect responses the student has made on this skill. In the scenario of ASSISTments, we argue that the number of correct and incorrect answers can be replaced by the *mastery speed* we mentioned in the previous section. This means that we needed to first look into the relation between mastery speed and test performance.

3.1 Relation between Mastery Speed and Delayed Test Performance

In order to not over-fit the data we collected, we only considered students who answered 10 or more retention tests and only considered skills with at least 100 retention tests. Skills with very few items suffer not only from over-fitting, but selection bias as well, since they were probably only assigned by one or two teachers. After this filtering, we have the data that contains 662 students and 27,451 rows of problem records. In this data, we found that students performed very differently on the *mastery speed*; so we separated the possible mastery speed into interpretable bins:

- 3-4 attempts: students answered 3 problems correctly in a row or answered the first problem incorrectly but three consequent problems correctly after that;
- 5-8 attempts: students had approximate equal numbers of correct and incorrect attempts;
- more than 8 attempts: students endured very long sequences of problems, but eventually achieved three correct answers in a row;
- failed at mastering: students started the practice of Mastery learning, but did not complete it before doing the retention tests;
- did not attempt mastery: students didn't start to master the skills (they didn't start the practice on the Mastery learning problem sets)

Table 1 shows the relationship between student *mastery speed* and performance on delayed retention tests.

mastery speed	% correctness on retention tests	
	7-day delay	14-day delay
3-4 attempts	82%	76%
5-8 attempts	70%	62%
> 8 attempts	59%	49%
failed at mastering	44%	29%
did not attempt to mastery	63%	40%

Table 1. Relationship between mastery speed and retention test performance

From Table 1 we can observe that in general; the more practice opportunities a student required in mastering a skill, the lower the probability the student can answer the problems in the retention test correctly. More interestingly, there is approximately a 10% decrease in percentage correctness between each level of mastery. This result is somewhat surprising, as most ITS have a simple threshold for mastery, yet these results suggest that even a relatively simple disaggregation of how students mastered the skill reveals substantial differences in how well it was learned

For students who did not master the skill, it is likely that they had great difficulties in understanding the skills at the outset or were gaming the system. On the other hand, we suspect that some students who skipped the initial assignment for mastery learning did not understand the material, although many of them felt that they understood the material but did not wish to spend time on the assignment. Table 1 also confirms our intuition here, as students who tried to master the material and failed performed less well than students who did not attempt to master the skill. Note, we are not asserting that this relationship is *causal* (Pearl 2009, Rai and Beck 2011), that is, failing at the mastery exam did not make the students less able to answer the retention test. Rather, this relationship is diagnostic, i.e., knowing that students cannot master the material is very predictive, relative to being less sure with students who did not attempt the exercises at all.

3.2 Predicting Delayed Test Performance

We defined a student as retaining a skill if he or she was able to respond correctly after a delay. In our model, the dependent variable is whether a student responded correctly on the delayed test problem, treating incorrect responses as a 0 and correct responses as a 1. Note that in the mastery cycle of ARRS, students who failed on the retention tests received repeat delayed tests, but for this study, we were only predicting the performance of the first retention tests of each delayed period. As well as, considering the *mastery speed*, the *problem_set_id* and *class_id* as factors, we also used the following factor features:

- *grade_diff_binned*: the binned value of grade difference. We computed the grade difference by students' current grade minus their skill grade and then further grouped these difference values into five different bins, which are above grade, on grade, one year ago, longer ago, and others;
- *on_grade*: whether this is a skill that belongs to the same grade which students are in.

We had the following features as covariates:

- *item_difficulty_binned*: the binned values of problem difficulty. The problem difficulty is represented by using the percentage of correctness for this problem across all answers and all students. The higher this value is, the more likely the problem can be answered correctly;

- *num_first_tests*: the number of repeat 7-day delayed tests. Students who failed on the retention tests received repeat relearning assignments and delayed tests, some students took many repeat tests. This feature was used only on 14-day delayed test prediction. It was designed to capture the information of students' 7-day test performance and a number of relearning opportunities.

After training the model with our ARRS data, we got a R^2 of 0.208 for 7-day delayed tests and a R^2 of 0.187 for 14-day delayed tests. Since these are results that fit the training data, they are optimistic and strong enough to predict the students' delayed retention test performance.

We first looked at the Beta coefficient values and p-values for the prediction of 7-day delayed tests. We noticed that only *mastery speed* is a reliable predictor factor. This confirms our observation in Section 3.1 that *mastery speed* has a strong connection with long-term retention. Table 2 shows the Beta coefficient values and p-values of *mastery speed*. The positive Beta values indicate that the larger the covariate is, the more likely the student responded to this problem correctly. We took the group of students who did not attempt to master the skill as the base line in this model. We can see that the other three groups of students who achieved mastery then had a better chance of answering the retention test correctly.

mastery speed	Beta	p-value
3-4 attempts	0.718	0.000
5-8 attempts	0.403	0.000
>8 attempts	0.130	0.056
failed at mastering	-0.483	0.000
did not attempt to mastery	0.000	0.000

Table 2. Parameter table of master speed in prediction of 7-day delayed tests

In terms of the only covariate of the 7-day delayed test prediction, we found that *item_difficulty_binned* is a reliable feature, the Beta coefficient value of it is 0.487.

In the experiment of predicting 7-day delayed test prediction, we can take *mastery speed* and *item_difficulty_binned* as reliable predictors for predicting retention test performance. When looking at the prediction of 14-day delayed tests, *item_difficulty_binned and num_first_tests* are both reliable covariate features as well as *mastery_speed* which works as an important factor feature for the prediction. Table 3 and Table 4 show Beta coefficient values and p-values for reliable features.

We also built a test data to validate these two models. For information which did appear in training data, we used the mean values of coefficients to replace them with. The R^2 of the 7-day delayed model is 0.176, and 0.168 for 14-day delayed model; results that indicate a reasonable fit in-line with other PFA models.

mastery speed	Beta	p-value
3-4 attempts	0.793	0.000
5-8 attempts	0.576	0.000
>8 attempts	0.232	0.058
failed at mastering	-0.221	0.157
did not attempt to mastery	0.000	0.000

Table 3. Parameter table of master speed in the prediction of 14-day delayed tests

Covariate	Beta	p-value
item_difficulty_binned	0.579	0.000
num_first_tests	0.131	0.000

Table 4. Parameter table of covariates in the prediction of 14-day delayed tests

The coefficients of the two experiments confirmed our intuition about *master speed* as a predictor of students' delayed retention tests; this also indicated that student knowledge retention varies by their *mastery speed* across different periods of delay. In the prediction of the 14-day delayed tests, we appended a covariate feature *num_first_tests* to keep track of the number of 7-day delayed tests that a student had on the same skill. The larger this number, the more chances the student had failed to retain the skill and had to relearn it. Given that we had a positive coefficient in the prediction model, this roused our curiosity as to how the relearning and re-mastery of problems could affect students' retention performance. We extended our training data to include repeat retention tests, and added a new factor feature *relearn_speed.* The *relearn_speed* factor is similar to *mastery speed*; it captures the number of attempted problems in the process of re-mastery between two retention periods. Our hypothesis here is that the relearning performance could reliably influence the next retention test performance. Consequently, we conducted another two experiments to predict the 7-day retention tests and the 14-day retention tests using this extended data. Unfortunately, *relearned_speed* is not a reliable predictor in the 14-day retention tests performance; this could suggest that relearning practices can only help the 7-day retention, and skill retention is going to decrease with a longer-delay. We are still exploring methods to help us understand how relearning practices work with the mastery cycle.

4. Contribution

This paper makes three contributions. First, the work behind this paper deployed the Enhance ITS mastery cycle model (Wang and Beck 2012) within the field. Through the participation of thousands of students, we carried out a randomized controlled trial to test the idea of reviewing students' long-term performance. As the first study on such experiment, the paper explores a new path for improving ITS to help students achieve long-term mastery learning.

The second contribution of this paper is the extension of the PFA model with new features that are likely to be relevant for mastery learning and retention. The majority of preceding works (Pavlik, Cen, Koedinger 2009, Gong, Beck, and Heffernan 2010) have only focused on features such as student performance and item difficulties. Our study adopted features which have characteristics of high pertinence to student retention and relearning. In comparison with some studies that took in to account the time gap from the student last seeing a skill as an important factor, we fixed the time factor in our study and we conformed that the notion of mastery speed is relevant to student delayed performance. This model can be easily applied to the prediction of longer delayed tests; it could also become an important mechanism in helping ITS in the decision-making process.

The third contribution of this paper is the discovery of a method in which it is possible to estimate one of the factors of students' robust relearning. Previous work (Beck, Gowda and Corbett 2012) presented models which distinguish between shallow learning from robust learning with features focused on skill transferring. These, however, are very limited in investigation on the importance of retention in robust learning. This work provided a new concept of features relevant to student retention to help in the detection of robust learning.

5. Future Work and Conclusions

This enhanced ITS mastery cycle and its implementation in ASSISTments have been presented to the field for just a few months, so we are still at the initial phase of study and there are many further problems that we are interested in: can *mastery speed* as a feature affect all predictions on delayed test performance? Besides *master speed*, are there other cognitive indicators that the student is not learning the skill in long-term mastery? Can we craft our model to make good quality predictions on longer-term retention tests?

Most importantly, there are some very challenging problems that we believe can be answered in our study. Firstly, when should we reassess a student? And if a student fails to retain a skill, what is the best strategy to help him or her relearn the skill?

This paper presents the latest development and study of the enhanced ITS mastery cycle. With the data we collected from this system, we aimed to predict the delayed test performance and introduced some useful features to extend the PFA model in retention performance prediction.

Acknowledgments

We want to acknowledge the funding on NSF grant DRL-1109483 as well as funding of ASSISTments. See here (http://www.webcitation.org/67MTL3EIs) for the funding sources for ASSISTments.

References

Anderson, J.R. *Rules of the Mind*. 1993. Psychology Press.

Beck, J.E., et al. 2003. Predicting student help-request behavior in an intelligent tutor for reading. The 9th International Conference on User Modeling.

Cepeda, N.J., et al. Distributed practice in verbal recall tasks: A review and quantitative synthesis. 2006. Psychological Bulletin, Vol 132(3). 354-380

Baker, R.S.J.d.; Gowda, S.; Corbett, A.; Ocumpaugh, J. 2012 Towards Automatically Detecting Whether Student Learning is Shallow. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, 444-453.

George B.S., John A.E. Knowledge Taught in School: What Is Remembered? 1994. Review of Educational Research Summer, Vol. 64, No. 2, 253-286

Gong, Y., Beck, J.E., Heffernan, N. 2010. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In Intelligent Tutoring Systems, 35-44. Springer Berlin/Heidelberg.

Pavlik, P. I., Anderson, J. R. 2010. Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. Cognitive Science, 29(4), 559-586.

Pavlik, P. I., Cen, H.; Koedinger, K. 2009. Performance Factors Analysis - A New Alternative to Knowledge. In *Proceedings the 14th International Conference on Artificial Intelligence in Education*, 531-538

Pearl, J. 2009. Causal inference in statistics: An overview. Statistics Surveys, 3, 96-146.

Qiu, Y., et al. Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing. 2011. In *Proceedings of Fourth International Conference on Educational Data Mining*.

Rai, D., Beck, J.E. 2011. Causal Modeling of user data from a math learning environment with game-like elements, In Proceedings of the 15th International Conference on Artificial Intelligence in Education

Wang, Y., Beck, J.E. 2012. Incorporating Factors Influencing Knowledge Retention into a Student Model, In *Proceedings of the 5th International Conference on Educational Data Mining*, 201-203.