Incorporating Forgetting in the Personalized, Clustered, Bayesian Knowledge Tracing (PC-BKT) Model

Prema Nedungadi
Amrita Create
Email: prema@amrita.edu
Amrita Vishwa Vidyapeetham

Remya M S
Amrita Create
Email: ms.remya7@gmail.com
Amrita VishwaVidyapeetham

***Abstract*— Personalization and adaptation are at the core of Intelligent Tutoring Systems. The Bayesian Knowledge Tracing (BKT) Student Model is a time-tested method that maintains information about students' knowledge levels for the different skills in the topic domain. In our previous work, we had proposed the Personalized, Clustered, Bayesian Knowledge Tracing (PC-BKT) model that individualizes the learning of skills for each student and additionally improves the prediction for the cold start problem. A clustering of both students and skills based on a student and skill capability matrix was used to learn the prior skills to deal with the cold start problem, which is the prediction for either new skills or new students. Both the BKT and the PC-BKT models assume that a skill once learnt is never forgotten. But forgetting is pervasive. If a previously learnt skill is not used for a while, there is a higher chance of forgetting it. One of the factors that influence the forgetting is the time duration before the current attempt at using a skill and the previous attempt. We incorporate forgetting as a time decay function in the BKT and PC-BKT models and show significant increase in the accuracy of the student prediction.**

Keywords*— *Bayesian Knowledge Tracing, BKT, PC-BKT, Student Models, User Models, Intelligent Tutoring Systems, ITS

I. INTRODUCTION

Computer systems have been used for educational purposes since the early 1960s [6]. An Intelligent Tutoring System (ITS) is a type of e-learning system that determines the sequence and presentation of content based on the performance of students.

Intelligent tutoring systems aim to help students learn effectively in an efficient manner by presenting the content that most contributes to learning. Shute et al. (1989) compared the learning of a traditional economics course with an ITS and found that the traditional course took twice the amount of time, while learning was comparable. A major benefit of ITS is in the correct diagnosis and efficient delivery of learning materials that best suit the learners.

Personalizing to student needs are what differentiate an ITS from an e-learning systems without intelligence. ITS may personalize the content to present, the method of presentation, the sequence of presentation and the presentation modality.

Hence, the ITS needs to maintain student models, content models and context models to adapt appropriately to the students' learning needs. An accurate student model forms the heart of the ITS and helps tailor instruction and assessment that is most effective for the student.

While many student models have been the study of intense research, the Bayesian Knowledge Tracing model has both theoretical foundation and successful real world applications. In our previous work, we had proposed the Personalized, Clustered, Bayesian Knowledge Tracing (PC-BKT) that personalizes learning of prior skills for each student and deals with the cold start problem. A clustering of both students and skills based on a student and skill capability matrix was used to learn the prior skills to deal with the cold start problem, which is the prediction for either new skills or new students.

The BKT models assume that a skill once learnt is never forgotten. But forgetting is pervasive. If a previously learnt skill is not used for a while, there is a higher chance of forgetting it. One of the factors that influence the forgetting is the time duration before the current attempt at using a skill and the previous attempt. In this paper, we incorporate forgetting as a time decay function in the student prediction model and show significant reduction in the false positive case and an increase in the accuracy of the student prediction for both the base BKT and PC-BKT models.

II. COGNITIVE LEARNING & FORGETTING

The durability of learning shows a strong correlation to the temporal distribution of multiple sessions [15, 3].

Research in the benefits of distribute practice suggest that long-term memory and retention generally improve when the sessions are spaced in time. The ideal temporal distribution of study can lead to almost double the retention [4]. Similar benefits of spaced studied are shown in acquiring conceptual skills and cognitive skill [2], in medicine [11] and in the classroom [16] and overall educational outcomes [8, 7].

The beneficial effects of temporal distribution of practice have been shown in domains that require motor skills such as music [17] and sports [10].

978-1-4799-7171-8/15/$31.00 ©2015 IEEE

Though there are computational models to explain this phenomenon known as the distributed practice effect, there is not yet an application in the student models of Intelligent Tutoring Systems. This important cognitive principle has not yet been applied to real life applications [2, 7].

As a first step towards this, we model forgetting a skill in the Bayesian Knowledge Tracing algorithm. Qiu and Qi et al. explain that forgetting is a possible cognitive explanation for the over prediction of KT when considering the time students take to finish their tasks. However, they do not attempt to model forgetting within the BKT algorithm or update the knowledge components of a student.

III. STUDENT MODELING

A key feature of ITS systems is to guide adaptive behavior in e-learning: customize the learning content, select the items and the right sequence of items, customize content to devices and provide various types personalized feedback to the student. Two important requirements for this personalization are the estimation of student knowledge for skills, called student modeling and the ability to predict student performance.

Student performance prediction aims to predict the success or failure of a student for a given step of a problem based on the estimated skills required to solve the step.

We first estimate the latent student skills by learning from sequential data about student performance by considering factors such as answers to exercises and time taken to answer. A student model that maintains the skill levels for a student and predicts future performance of a student can be used for determining tutoring actions such as the choice of the next topic to teach, whether to provide help or thinking clues. Reports of student levels, performance and predictions assist teachers in tailoring appropriate interventions.

A. Bayesian Knowledge Tracing Model

The Bayesian Knowledge Trading model [5] is a four parameter model; with two knowledge parameters, the initial knowledge and the rate of learning, and two performance parameters the guess rate and slip rate.

- P (L0) is the initial probability that the student knows a particular skill.
- P (G) is probability of guessing correctly, if the student doesn't know the skill.
- P (S) is probability of making a slip, if student does know the skill.
- P (T) is probability of learning the skill, if the student does not know the skill. Note that P (T) is assumed to be constant over time.

B. Knowledge Tracing Algorithm for BKT

The Knowledge Trading Algorithm is used to predict P (Ln) and to find out whether the student has mastered a particular skill. The probability that the student has learned/mastered the skill at opportunity (time) t is computed as follows

We find P (Ln|Action_n), the probability that the student has learned a skill just after completing step j given student performance Oj on previous steps, where Oj = {o1, o2, o3,,…,oj} is the student performance on the first j opportunities and Oi can be either correct or incorrect. This conditional probability obeys the recursion [1]:

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * P(G)}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * (P(S)) + (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_{n-1}|Action_n) = P(L_{n-1}|Action_n) + (1 - P(L_{n-1}|Action_n) * P(T)$$

C. PC-BKT Model

In our previous work, we proposed an enhanced BKT model called the PC- BKT model that uses individual priors for each student and skill and combined it with dynamic clustering of students based on changing learning ability to deal with the cold start problem and show that PC-BKT significantly increases the accuracy of student prediction in both the general and the cold start problem [12].

Though we show a significant improvement in the overall prediction accuracy, there is room to improve the false positive. We posit that the lower accuracy in this case is because both BKT and PC-BKT assume that a skill once learnt is never forgotten and hence predicts success when a student may fail due to forgetting of skills.

D. Incorporating Forgetting into the BKT Model

Learning and forgetting are affected by the temporal distribution of study. It is well established that temporally spaced learning results in better learning outcomes when compared to the entire practice in one session [3].

The BKT model predicts student performance based on past student performance. The learning component of every attempt is tracked as a transition probability in the BKT mode. However, the BKT model does not take into account the temporal distribution of items attempted. We learn the transition probability forget from the data and find a small improvement in the prediction accuracy over the base model. We propose an alternative method of incorporating forgetting by updating the knowledge level for a skill based on the duration of the previous attempt at using the skill. An exponential decay function based on the time lag between the previous use of the skill is used update the knowledge level and show a significant increase in prediction accuracy over the base models.

IV. PROPOSED MODEL WITH DECAY

Our model is based on the assumption that the amount of learnt material that is remembered decays exponentially over time [19]. We use an exponential decay function to update the knowledge level so as to increase the prediction accuracy. Any quantity that decays by a fixed percent at regular intervals is said to possess exponential decay. Exponential rate of change,

$$N(t) = \boldsymbol{N_{(t-30)} e^{-\lambda(t/30)}}$$

Where N is the knowledge level, lamda is an exponential decay constant set to 0.1 [13] and t is the time interval. We assume that the chance of forgetting will increases if a student doesn't attempt the skill within 30 days.

Algorithm 1. Algorithm for Forgetting in PC-BKT

N = #students; K = #skills; J = #items, D = #days

Procedure PC-BKTwithForget (N, K, J, D)

Compute Knowledge Sequence (K) Using Empirical Probabilities.

Compute Guess, Slip and transition rate Using K.

Calculate Initial Learning rate of student to each skill as per first performance.

Calculate Capability Matrix, B.

for i=0 to N **do** /*Training Phase*/

for k=0 to K **do**

Initialize D (k) to 0.

for j=0 to J **do**

if (k is not in j **&&** (timestamp (j) - TS (k))! =0) **then**

D (k) = timestamp (j) – TS (k);

else

if (D (k) >= 30) **then**

$\boldsymbol{P_{ik}(L_{t-1}) = P_{ik}(L_{d-30}) e^{-0.1\,(d/30)}}$

End if

Apply the update equation of PC-BKT

D (k) = 0

TS (k) = timestamp of j.

End if

end for

end for

end for

Apply clustering on B

Calculate Mean P (L) for each cluster.

Predict the performance of Student on the task at time t+1.

end Procedure

V. EXPERIMENTS AND RESULTS

We enhance both BKT and PC_BKT models with the above to incorporate forgetting as a time decay function. Our proposed methods are compared with both the base models and the model with a fixed value for "forgetting".

A. Dataset Used

For testing the proposed approached, we use the Cognitive Tutor dataset, from the Knowledge Discovery and Data Mining Challenge 2010 KDD Cup 2010 [9] from two different tutoring systems from multiple schools over multiple school years. The dataset contains 19 attributes. There are mainly 5 data sets: 3 development data sets and 2 challenge data sets from 2 different tutoring systems [9]. We used the Challenge dataset for this work which contains 3,310 students with 9,426,966 steps [9] and their original information is described as in Table 1 & 2. The technical challenges of this dataset include a sparse data matrix, the temporal dimension of the data and that the problem a given student sees is determined in part by student choices or past success history.

We used 240,000 log data of 200 students to test our models.

Datasets	Size	#Attributes	# Instances
Algebra 2008-2009 train	3.1 GB	23	8,918,054
Algebra 2008-2009 test	124 MB	23	508,912
Bridge to Algebra 2008-2009 train	5.5 GB	21	20,012,498
Bridge to Algebra 2008-2009 test	135 MB	21	756,386

Table 1. KDD CUP 2010 Challenge Datasets

Challenge Datasets	Students	Steps
Algebra I 2008-2009	3,310	9,426,966
Bridge to Algebra 2008-2009	6,043	20,768,884

Table 2. Information of Original Datasets

B. Results

We compare the misclassification percentage of basic BKT model, the BKT with a fixed forget value that is learnt from data and BKT with decay function that depends on the time delay from the previous use of the skill. Fig. 1 shows that BKT with the learnt forget value has slightly lower misclassification as compared to the base BKT. But the percentage of misclassification in BKT with decay function is significantly lower as compared to other models.

Similarly, Fig. 2 shows the comparison of misclassification of the PC-BKT model with and without decay function. The comparison shows that model with decay function performs well as compared to other. Our experiments included two assumptions on the time interval, a minimum of 30 days or 45 days for forgetting. The figure shows that assumption with 30 days time interval performs better than the 45 days.

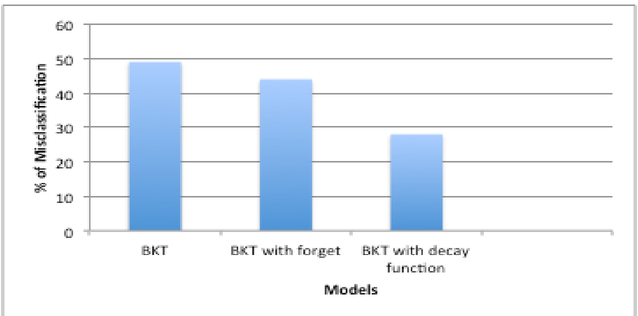

Fig 1. Comparison of BKT model with and without decay function.

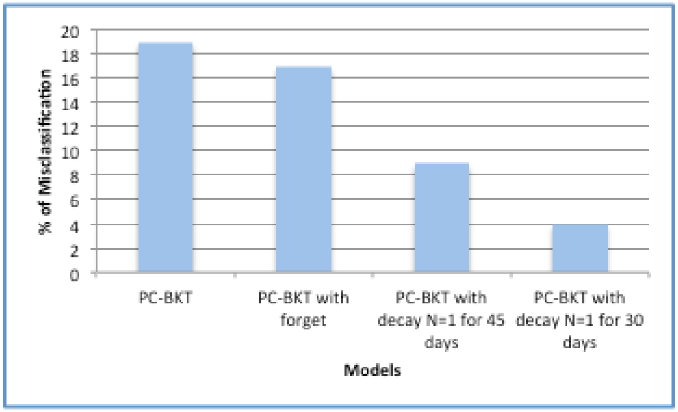

Fig 2. Comparison of PC-BKT model with and without decay function.

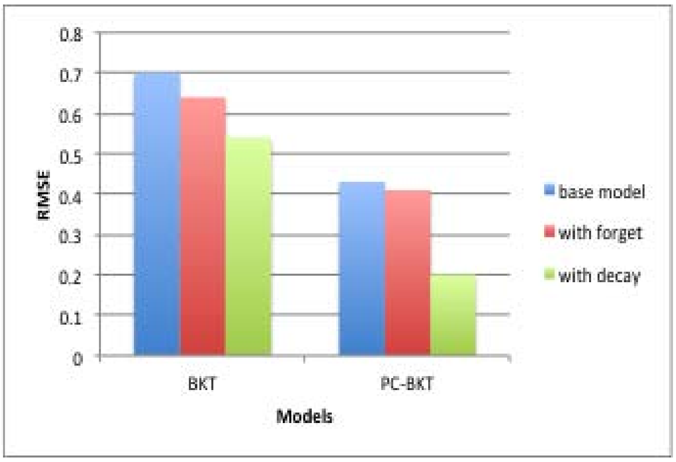

Fig 3. Comparison of RMSE of models

Fig 3 shows the comparison of Root Mean Square Error of Models. This comparison also proves that the decay function increases the performance accuracy.

Table 3 shows the comparison of accuracy of different models. This comparison also shows that using decay function in models improve their performances. The accuracy in the negative class increases by including forgetting as a time decay function.

Models	Accuracy	+ve class	-ve class	RMSE
BKT	50.8	57.6	24.1	0.7
BKT with forget	58.7	67.1	26.4	0.64
BKT with decay	70.9	78.8	40.4	0.54
PC-BKT	80.3	90.6	40.8	0.43
PC-BKT with forget	83.1	92.2	48.2	0.41
PC-BKT with decay	95.9	97.6	86.1	0.2

Table 3. Comparison of Prediction Accuracy in %

VI. CONCLUSION

The BKT model typically sets the forget parameter to 0. We propose an enhancement in the original BKT model and our enhanced PC-BKT model by using a decay function to model forgetting of skills.

We use 240,000 log records of 200 students of the Knowledge Discovery and Data Mining Challenge 2010 dataset, and show that, in both the BKT and the PC-BKT models, the percentage of misclassification significantly reduces as our algorithm adjusts the learnt rate of a skill based on the time duration between the last use. We find that the duration of one month of not using a skill to be a reasonable amount of time for a student to start forgetting the skill. Setting the time duration to 45 days shows a smaller reduction in misclassification suggesting that forgetting starts earlier than 45 days, in this case by 30 days. Future work involves additional evaluations to estimate the time that forgetting begins to set. Future work also involves the study of other mathematical functions to model forgetting within the BKT models.

Acknowledgments. The paper derives ideas, inspiration and guidance from Shri Mata Amritanandamayi Devi, the Chancellor of Amrita University.

REFERENCES

1. Baker, R., Corbett, A., & Aleven,V.: More accurate student modeling through contextual estimation of slip & guess probabilities in BKT.In ITS , B.Woolf, E.Ameur Eds, Lecture Notes in Computer Science.Springer Berlin/Heidelberg 406-415 (2008)

2. Carpenter, S., Cepeda, N., Rohrer, D., Kang, S. K., & Pashler, H.: Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. Educational Psychology Review, 24, 369-378. doi: 10.1007/s10648-012-9205-z (2012)
3. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D.: Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psych. Bull., 132, 354–380 (2006)
4. Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H.: Spacing effects in learning: A temporal ridgeline of optimal retention. Psych. Sci., 19, 1095–1102 (2008)
5. Corbett, A., & Anderson, J.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling & User-Adapted Interaction, 4, 253-278 (1995)
6. Corbett, A.T., Koedinger, K.R. & Anderson, J.R.: Intelligent tutoring systems. In M.G. Helander, T.K. Landauer & P.V. Prabhu, eds., Handbook of human- computer interaction, 849–874, Elsevier Science B. V., Amsterdam. 1, 2 (1997)
7. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T.: Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14, 4–58 (2013)
8. Kerfoot, B., Fu, Y., Baker, H., Connelly, D., Ritchey, M., & Genega, E.: Online spaced education generates transfer and improves long-term retention of diagnostic skills: A randomized controlled trial. Journal of the American College of Surgeons, 211, 331–337 (2010)
9. K.Koedinger, R.Baker, K.Cunningham, A.Skogsholm, B.Leber, and J.Stamper.: A data repository for the edm community: The pslc datashop. In C. Romero, S. Ventura, M. Pechenizkiy and R. Baker, eds., Handbook of Educational Data Mining, Lecture Notes in Computer Science, CRC Press. 5, 13, 16, 18, 20, 26,54, 153 (2010)
10. Lawton, T. W., Cronin, J. B., & Lindsell, R. P.: Effect of interrepetition rest intervals on weight training repetition power output. The Journal of Strength & Conditioning Research, 20, 172-176 (2006)
11. Moulton, C.-A. E., Dubrowski, A., Macrae, H., Graham, B., Grober, E., & Reznick, R. (2006).
12. Nedungadi, P., Remya, M. S.: Predicting Students' Performance on Intelligent Tutoring System - Personalized Clustered BKT (PC-BKT) Model, presented at The 43rd ASEE/IEEE Frontiers in Education Conference (FIE 2014), Madrid, Spain (in press).
13. R. Pelánek.: Application of Time Decay Functions and the Elo System in Student Modeling. Educational Data Mining (EDM) (2014)
14. Shute, V. J., Woltz, D. J., & Regian, J. W.: An investigation of learner differences in an ITS environment: There's no such thing as a free lunch. Paper presented at the 4th international conference on artificial intelligence and education—AI-ED '89, Amsterdam, Holland (1989)
15. Shea, C. H., Lai, Q., Black, C., & Park, J. H.: Spacing practice sessions across days benefits the learning of motor skills. Human Movement Science, 19, 737–760. doi:10.1016/ S0167-9457(00)00021-X (2000)
16. Sobel, H., Cepeda, N., & Kapler, I.: Spacing e□ects in real-world classroom vocabulary learning. Applied Cognitive Psychology, 25, 763-767 (2011)
17. Simmons, A. L: Distributed practice and procedural memory consolidation in musicians' skill learning. Journal of Research in Music Education, 0022429411424798 (2011)
18. Qiu, Y., Qi, Y., Lu, H., Pardos, Z. & Heffernan, N.; Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing. In Pechenizkiy, M., Calders, T., Conati, C., Ventura, S., Romero , C., and Stamper, J. (Eds.) Proceedings of the 4th International Conference on Educational Data Mining. pp. 139-148. (2011)
19. Loftus, Geoffrey R. "Evaluating forgetting curves." Journal of Experimental Psychology: Learning, Memory, and Cognition 11.2 (1985): 397.