Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070122
ISSN 2371-8412 Vol. 7 Num. 1

Research on Integrating Forgetting Behavior into Student Models for Online Learning Systems

Ze Song

University of Science and Technology of China, Hefei, Anhui, 230026, China
songze@mail.ustc.edu.cn

Keywords: Online Learning Systems, Student Modelling, Knowledge Tracing, Forgetting Behaviour, LSTM

Abstract: Modelling students accurately is an important task in online learning systems. In online learning systems, student models are usually built by dealing with students' historical data of answering questions as input, and eventually output to what extent the student has mastered a certain knowledge component. To evaluate a student model, a commonly used method, namely knowledge tracing, is to build multiple student models based on multiple continuous historical data as mentioned above, then predict whether or not a student can answer a question correctly, and finally compare predicted results with true results. However, the behavior of students is complicated and unpredictable, thus makes student modelling and knowledge tracing become very difficult tasks. Based on existing research of knowledge tracing, especially deep knowledge tracing which uses LSTM to model students, this paper proposes a novel method of student modeling. Compared with other state-of-the-art student modeling methods, the most significant feature of our modeling methods is our method can take students' forgetting behavior into consideration. Moreover, our modeling method can appropriately handle situations that one question corresponds to multiple knowledge components. To test the performance of our student model, this paper applies our student model to the knowledge tracing task. Based on our experiments in public datasets, when one question corresponds to multiple knowledge components and the length of students' historical data is greater, our model performs better in terms of all metrics compared to state-of-the-art knowledge tracing methods.

1. Introduction

In recent years, online learning systems have received increasing attention. Precisely modeling students who use online learning systems, particularly in terms of their mastery of different knowledge areas, is one of the important tasks of online learning systems. To verify the accuracy of the established student models, one can build models based on the sequence of students' historical answer data, and then predict whether the students' next answers will be correct or not. This prediction can be compared with the actual student answer records to measure accuracy and other metrics. This process is known as Knowledge Tracing [1]. Since this task was proposed, the field has seen continuous emergence of new research, producing various methods of student modelling.

In 2015, Deep Knowledge Tracing (DKT) [2] was introduced. This model uses a Recurrent Neural Network (RNN) as the main component of the model, which improved the ROC AUC by about 25% compared to traditional models, thereby quickly becoming the mainstream method for knowledge tracing.

Accurately modeling students is a challenging task due to the complexity of the human brains and learning process. The mastery level of different knowledge areas by students is influenced by multiple factors, including the learning curves for different concepts and the forgetting curves after learning these concepts, and so on; moreover, these factors are likely to vary from student to student. Therefore, intuitively, complex student models should be closer to representing real students than simpler models. This also explains the outperformance of DKT in knowledge tracing tasks.

However, most models only take into account the concepts covered in the student's answer records and whether the student's answers are correct as inputs to the model. Yet, current online learning systems can provide many other features beyond these two, such as the start and end time of the student's answer attempt, the type of questions, and which other functions of the online learning system the student has used, etc. Relying solely on the aforementioned two features as model inputs is clearly insufficient to represent the student's complete learning process, let alone to reflect the forgetting process after learning the knowledge. Therefore, this approach actually overlooks the impact of the forgetting process on learning.

To model students more accurately and overcome the aforementioned shortcomings, this paper improves upon existing DKT models and introduces an original student modeling method that considers the forgetting process. This method is applied in knowledge tracing tasks and has been tested on public datasets. Experiments demonstrate that, compared to previous deep knowledge tracing models, this model achieves improvements in accuracy, precision, and ROC AUC, especially in datasets with multi-labeled questions and long sequences of student history records.

2. Related Work

2.1. Knowledge Tracing

Before the popularization of deep learning methods, the most commonly used methods in the field of knowledge tracing included Bayesian Knowledge Tracing (BKT) [1] and Performance Factor Analysis (PFA) [3]. Many studies have attempted to improve the accuracy of student models by enhancing the BKT and PFA models. Some research has focused on improving the BKT model by adding parameters that can reflect differences between students [4, 5]. Other studies have considered incorporating the feature of the number of attempts a student makes on a question and combining it with the BKT or PFA models using boosting algorithms [6]. After the popularization of deep learning, the excellent performance in processing sequential data from Recurrent Neural Networks (RNN) and their variant, Long Short Term Memory (LSTM) networks, has drawn significant attention. Given that the inputs for knowledge tracing tasks are also temporal data, the DKT model [2] utilizing RNN or LSTM was proposed in 2015. Although the model was specifically designed for knowledge tracing tasks, they can predict the probability of a student correctly answering questions related to a concept. Thus, this probability can be used to reflect a student's mastery of the concept. Extensive experiments have shown that, despite a few exceptions, the DKT model significantly outperforms the earlier BKT and PFA models on most datasets in terms of performance [7].

After the proposal of the DKT model, deep learning methods began to be widely applied in the field of knowledge tracing. With the further development of deep learning theories, new models designed to replace RNN and LSTM, such as the Transformer model, have been introduced. These models have outperformed RNN and LSTM in areas like text classification and sentiment

recognition. Consequently, many studies [8-10] have attempted to apply these improved models to knowledge tracing tasks. However, existing research [11] has shown that these models do not necessarily perform better than the DKT model on public datasets. Currently, the DKT model remains state-of-the-art in the field of knowledge tracing.

2.2. Knowledge Tracing with Forgetting

One drawback of the BKT model is its complete inability to account for the forgetting process in students. Before the advent of DKT, some research attempted to incorporate the factor of student forgetting into the BKT model. For example, [12] added the transition from knowing to not knowing a concept into the BKT model and considered the number of days elapsed since learning the concept as a factor in forgetting. The limitation of this approach lays in the fact that, according to human forgetting patterns [13, 14], forgetting after learning occurs continuously and the rate of forgetting gradually decreases over time, rather than occurring in discrete day-by-day units.

After the introduction of DKT, although some research attempted to add additional features to the DKT model, few studies tried to incorporate features reflecting student forgetting into the DKT model. [15] attempted to represent student forgetting by segmenting the time interval between two quiz attempts into periods, but this approach is overly simplistic and fails to fully capture the forgetting process. [16] took a more comprehensive approach to consider various features that might influence student forgetting, including the time intervals between two attempts at answering questions, the intervals between attempts at questions involving the same concept, and the number of attempts at questions related to a specific concept. However, this model represents these features using vectors rather than matrices, which means it struggles to handle questions that relate to multiple concepts. Furthermore, when predicting the probability of a student correctly answering a question, it utilizes information such as the interval between the question being predicted and the previous question, which is not available in real-world student modeling scenarios. Although this approach can improve performance in knowledge tracing tasks, it also makes the model improper to be utilized as a student model.

3. Method

Inspired by the existing knowledge tracing models, especially the DKT model, this paper proposes the DKT-S model. This model is designed in the context of online learning systems as well as for knowledge tracing tasks. It incorporates the student's forgetting process into consideration. This chapter will provide a formal description of the inputs and outputs of the DKT-S model and present its architecture.

3.1. Task Formulation

We formulate the input and the output for knowledge tracing task as follows: Let $x_1, x_2, \cdots, x_m$ denote the student's history record with length m, and K denote the set of all concepts in online learning system. Then the input of model consists of the following:

$$X = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \boldsymbol{x}_m], \boldsymbol{x}_t = [q_t \ a_t]^T, q_t \in K \tag{1}$$

$$\boldsymbol{a} = [a_1 \ a_2 \cdots a_m], a_t \in A = \{0, 1\} \tag{2}$$

$$\boldsymbol{q}_t = [q_t^1 \ q_t^2 \ \cdots \ q_t^n] \in Q = \{0, 1\}^{|K|} \tag{3}$$

If only one concept is existent in problem $t, \boldsymbol{q}_t$ is one-hot; conversely, $\boldsymbol{q}_t$ is multi-hot when multiple concepts exist. The input to RNN is

$$X = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \cdots \boldsymbol{x}_m] \tag{4}$$

$$\boldsymbol{x}_t = [q_t^1 a_t \ \ q_t^1(1-a_t) \cdots \ \ q_t^z a_t \ \ q_t^z(1-a_t)]^T \tag{5}$$

that is, there are z dimensions in x_t representing a correct answer to a problem related to the concept, and other z dimensions representing an incorrect answer. As a result, in the model

$$\boldsymbol{x}_t \in \{0,1\}^{2|K|} \tag{6}$$

The output of the model is denoted as

$$\hat{a}_{t+1} \in [0,1] \tag{7}$$

Which is the predicted probability for the student to answer problem $t+1$ correctly.

3.2. Features for Forgetting

To model forgetting behavior in DKT framework, we consider the following three factors. Let k_p^t denote the indicator variable for whether or not student's question is related to concept p, and χ_t indicates the timestamp for student's question t. Then those factors are formulated as:

The duration of time δ_t for student to finish answering problem t;

$$\boldsymbol{s}_t = [s_1^t \ \ s_2^t \ \ \cdots \ \ s_z^t]^T \tag{8}$$

$$s_i^t = \begin{cases} \frac{B}{\chi_t - \chi_{m_i}} & k_i^t = 1 \ and \ \ \sum_{j=1}^{t-1} k_i^j > 0 \\ 0 & otherwise \end{cases} \tag{9}$$

Where B is a constant and

$$m_i = \underset{j}{\operatorname{argmax}} jk_i^j \tag{10}$$

$$\boldsymbol{c}_t = [c_1^t \ \ c_2^t \ \ \cdots \ \ c_z^t]^T; \ \ c_i^t = \sum_{j=1}^{t} k_i^j \tag{11}$$

In short, if concept i is included in the student's question t, we count for the appearance c_t of the concept in student's history, and the interval s_t between current time and the time when the student answered the previous problem that is related to concept i. As 0 denotes the lack of previous problem in s_t, We introduce a constant B and get its reciprocal. As a result, the input to DKT-S model, despite the aforementioned matrix X, consists of the following vectors or matrices:

$$S = [\boldsymbol{s}_1 \ \ \boldsymbol{s}_2 \cdots \boldsymbol{s}_m] \tag{12}$$

$$C = [\boldsymbol{c}_1 \ \ \boldsymbol{c}_2 \cdots \ \ \boldsymbol{c}_m] \tag{13}$$

$$\boldsymbol{\Delta} = [\delta_1 \ \ \delta_2 \cdots \ \ \delta_m] \tag{14}$$

Intuitively, both the S matrix and the C matrix encode the historical answer information of students. The process of a student answering questions on a certain concept is also the process of the student recalling that concept. The recall process weakens the impact of forgetting, so these features are related to forgetting. Similar features were also considered in [16], but that study used vectors to represent these features, while this paper uses matrices to represent them. The Δ vector represents the time each student spends answering each question; if a student takes a long time to answer, it is likely indicative that the student has forgotten the concept related to the question, making this feature also related to student forgetting.

3.3. Our Model

Our DKT-S model adopts an approach similar to [17], concatenating the features that represent student forgetting with the input of the original deep knowledge tracing model, which does not consider forgetting. Differently, to maximize the integrity of the input data and minimize the loss of information, the DKT-S model does not use an auto encoder structure to compress the input vector. The LSTM part of the model, as well as the sections following the LSTM, is consistent with the DKT model. Therefore, in this model, the parameters that need to be trained include the various parameters within the LSTM model, as well as the weights and biases of the linear layer.

For comparison, this paper also selected the original DKT model and the knowledge tracing model that considers forgetting, referred to as DKT-F, proposed in [16] for comparative experiments with the DKT-S model. The structure of all the abovementioned models is shown in Figure 1.

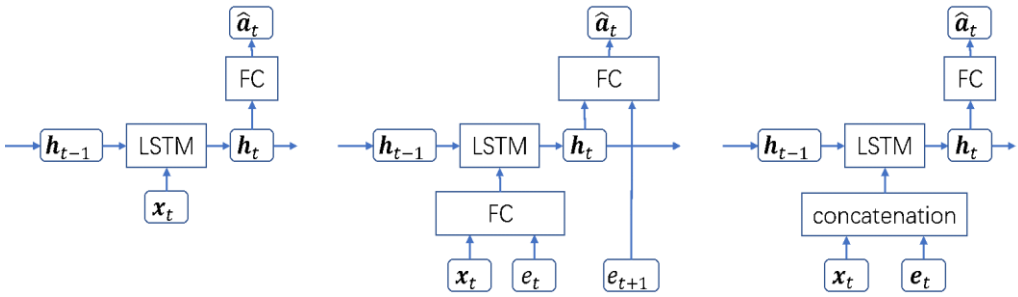

Figure 1: From left to right: The structures of (a) DKT (b) DKT-F (c) DKT-S model.

4. Experiments

4.1. Datasets

The ASSISTments dataset [18] is a public dataset provided by the ASSISTments online learning system, which was established in 2003 by Worcester Polytechnic Institute (WPI) with funding from the United States government. It is a free public service platform dedicated to helping students and teachers improve the quality of learning through intelligent tutoring technology. For a long time, this dataset was also one of the largest publicly available datasets for knowledge tracing [2].

The EdNet dataset, similar to the ASSISTments dataset, originates from historical data of the Santa online learning system in South Korea. The Santa online learning system is a multi-platform system dedicated to helping students prepare for the TOEIC (Test of English for International Communication). This dataset features a long time span (data collection spanned over two years) and a very large volume of data (comprising 784,309 students and 131,441,538 historical records). It is also highly suitable for knowledge tracing tasks and is currently the largest publicly available dataset for knowledge tracing [19].

Table 1: Details of selected datasets

#Students	16268	185498	61494	60593
#Records	419186	1148684	1925820	3074945
Avg. Records per student	25.77	6.19	30	50.75
Max. Records per student	70	7	30	70
Min. Records per student	4	5	30	40
Correctness rate	0.6855	0.4396	0.4021	0.5467
#Concepts	245	189	189	189

The experiments described in this paper will be conducted on the two datasets mentioned above. To further verify the performance of different student models on student history sequences of varying lengths, this paper splits the EdNet dataset according to the length of the student history sequences into three smaller datasets, namely EdNet-1 to EdNet-3. Additionally, to improve the training speed of the models, this paper preprocesses the original data from both the ASSISTments and EdNet datasets by removing student history sequences that are either too short or too long. The detailed information of the datasets used for the experiments is listed in Table 1.

4.2. Experimental Setup

As described earlier, this paper trains the original DKT model, the DKT-F model, and the DKT-S model on the above datasets. The data in the mentioned datasets were randomly divided into eight parts for 8-fold cross-validation [20]. During the training of the models, all three models were trained using the Adam optimizer and the following cross-entropy loss function for fair comparison:

$$L = \frac{-1}{m-1}\sum_{i=2}^{m} a_i \log \hat{a}_i + (1 - a_i)\log(1 - \hat{a}_i) \tag{15}$$

All three models uniformly use an LSTM with 200 hidden layer features and are trained under the aforementioned conditions until convergence. We evaluate the performance of the models based on three metrics: accuracy, precision, and ROC AUC.

4.3. Results and Analysis

The experimental results are shown in Table 2, and the values corresponding to the best-performing model are highlighted in bold in the table. Based on those results:

The DKT-S model and the DKT-F model both significantly outperformed the original DKT model on the ASSISTments dataset. Since the ASSISTments dataset was not divided according to the length of students' answer records, it can be inferred that taking students' forgetting into account as an additional factor indeed effectively improves the overall performance of the DKT model when considering a diverse student population.

Among the two models that consider student forgetting on the ASSISTments dataset, the DKT-F model performs slightly better. As mentioned before, the DKT-F model uses vectors to encode features that represent student forgetting, which can handle cases where a question corresponds to a single concept but struggles with questions that correspond to multiple concepts. However, in the ASSISTments dataset, each question contains only one concept, making the matrix representation of the DKT-S model more sparse compared to the vector representation of the DKT-F model, thereby affecting its performance to some extent.

Table 2: All experimental outcomes.

Model	DKT			DKT-F			DKT-S		
Dataset	Acc.	Prec.	AUC	Acc.	Prec.	AUC	Acc.	Prec.	AUC
ASSISTments	0.7157	0.7157	0.7434	**0.7532**	**0.7532**	**0.7984**	0.7369	0.7369	0.7703
Ednet-1	**0.7265**	**0.7283**	**0.7964**	0.6194	0.6487	0.6744	0.7017	0.7007	0.7692
Ednet-2	0.6872	0.6792	0.7277	0.6546	0.6408	0.6870	**0.6916**	**0.7031**	**0.7322**
Ednet-3	0.6522	0.6548	0.7008	0.6126	0.6135	0.6452	**0.6610**	**0.6558**	**0.7217**

In the three Ednet datasets, the DKT-F model shows a clear disadvantage compared to the DKT model and the DKT-S model. This is likely due to, in the Ednet dataset, the vast majority of questions cover more than one concept; thus, the DKT-F model can only combine these multiple concepts and treat them as a single concept. However, in the Ednet dataset, it is rare for two

questions to have exactly the same combination of concepts; therefore, the vectors used in the DKT-F model to represent the process of student forgetting degrade into vectors close to zero vector in this context. Since the DKT-F model also employs a fully connected layer to compute the original input matrix with the vectors representing the forgetting process, these features lead to a loss of information in the input matrix after computation, resulting in a significant decrease in performance.

In a horizontal comparison across the three Ednet datasets, the DKT model, which does not consider forgetting, performs best on the Ednet-1 dataset, where student history sequences are shorter. This is because the students in the Ednet-1 dataset typically answer 5 to 7 questions with similar concepts in a short period of time. This process can be seen as a continuous review of the same or similar concepts by the students, making the impact of forgetting very limited in this context.

In the Ednet-2 and Ednet-3 datasets, the DKT-S model proposed in this paper performs best. In these datasets, each student has 30 or more answer records, and unlike the Ednet-1 dataset, these answers are often not completed within a short period of time. Therefore, the process of forgetting has a more significant impact on students' mastery of various concepts, making the DKT-S model more advantageous in these datasets.

5. Conclusion

Accurately modeling students is a crucial task for the implementation of online learning systems. This paper proposes an original student modeling method and applies it to knowledge tracing tasks. Compared to other student modeling methods, this method has two distinct advantages. It incorporates student forgetting characteristics into the model and handles questions that correspond to multiple concepts. Through testing on the ASSISTments and Ednet datasets, the model proposed in this paper outperforms other leading-edge knowledge tracing models in scenarios where questions correspond to multiple concepts and where students have longer sequences of answer records. This indicates that the model indeed has enough advantages to meet the needs of the online learning systems described in this paper, making it practically applicable in online learning systems.

While this model was proposed within the context of online learning systems as a method for student modeling, experiments have already demonstrated that the model can also perform well in knowledge tracing tasks. This model has significant implications for the field of artificial intelligence combined with education, including but not limited to educational data mining, cognitive diagnostics, and beyond.

References

[1] T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, vol. 4, pp. 253–278, 1994.

[2] Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," Ad- vances in neural information processing systems, vol. 28, 2015.

[3] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger, "Performance factors analysis–a new alternative to knowledge tracing." Online Submission, 2009.

[4] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings 18. Springer, 2010, pp. 255–266.

[5] Y. Wang and N. T. Heffernan, "The student skill model," in Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11. Springer, 2012, pp. 399–404.

[6] Zhang Li "The "assistance" model: Leveraging how many hints and attempts a student needs," in Twenty-fourth international FLAIRS conference, 2011.

[7] M. Khajah, R. V. Lindsey, and M. C. Mozer, "How deep is knowledge tracing?" arXiv preprint arXiv:1604.02416, 2016.

[8] J. Zhang, X. Shi, I. King, and D.Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in Proceedings of the 26th international conference on World Wide Web, 2017, pp. 765–774.
[9] S. Pandey and J. Srivastava, "Rkt: relation-aware self-attention for knowledge tracing," in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1205–1214.
[10] Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. S. Choi, "Integrating temporal features for ednet correctness prediction," in Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA, 2021, pp. 12–16.
[11] S. Yang, M. Zhu, J. Hou, and X. Lu, "Deep knowledge tracing with convolutions," arXiv preprint arXiv: 2008. 01169, 2020.
[12] Y. Qiu, Y. Qi, H. Lu, Z. A. Pardos, and N. T. Heffernan, "Does time matter? modeling the effect of time with bayesian knowledge tracing." in EDM, 2011, pp. 139–148.
[13] L. Averell and A. Heathcote, "The form of the forgetting curve and the fate of memories," Journal of mathematical psychology, vol. 55, no. 1, pp. 25–35, 2011.
[14] H. Ebbinghaus, "Memory: A contribution to experimental psychology," Annals of neurosciences, vol. 20, no. 4, p. 155, 2013.
[15] Lalwani and S. Agrawal, "What does time tell? tracing the forgetting curve using deep knowledge tracing," in Artifi- cial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20. Springer, 2019, pp. 158–162.
[16] K. Nagatani, Q. Zhang, M. Sato, Y.-Y. Chen, F. Chen, and T. Ohkuma, "Augmenting knowledge tracing by considering forgetting behavior," in The world wide web conference, 2019, pp. 3101–3107.
[17] L. Zhang, X. Xiong, S. Zhao, A. Botelho, and N. T. Heffernan, "Incorporating rich features into deep knowledge tracing," in Proceedings of the fourth (2017) ACM conference on learning@ scale, 2017, pp. 169–172.
[18] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it as- sesses," User modeling and user-adapted interaction, vol. 19, pp. 243–266, 2009.
[19] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, Bae, B. Kim, and J. Heo, "Ednet: A large-scale hierarchical dataset in education," in Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21. Springer, 2020, pp. 69–73.
[20] Tang Y M. Impact of Self-Directed Learning and Educational Technology Readiness on Synchronous E-Learning. Journal of Organizational and End User Computing, 2021, pp. 1-20.