Convolutional Knowledge Tracing: Modeling Individualization in Student Learning Process

Shuanghong Shen[1], Qi Liu[1], Enhong Chen[1,*], Han Wu[1], Zhenya Huang[1],
Weihao Zhao[1], Yu Su[2], Haiping Ma[2], Shijin Wang[2]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Data Science &
School of Computer Science and Technology, University of Science and Technology of China,
{closer, wuhanhan, huangzhy, zhaoweihao}@mail.ustc.edu.cn; {qiliuql, cheneh}@ustc.edu.cn;
[2]IFLYTEK Co.,Ltd., {yusu,hpma, sjwang3}@iflytek.com

ABSTRACT

With the development of online education systems, a growing number of research works are focusing on *Knowledge Tracing* (KT), which aims to assess students' changing knowledge state and help them learn knowledge concepts more efficiently. However, only given student learning interactions, most of existing KT methods neglect the individualization of students, i.e., the prior knowledge and learning rates differ from student to student. To this end, in this paper, we propose a novel Convolutional Knowledge Tracing (CKT) method to model individualization in KT. Specifically, for individualized prior knowledge, we measure it from students' historical learning interactions. For individualized learning rates, we design hierarchical convolutional layers to extract them based on continuous learning interactions of students. Extensive experiments demonstrate that CKT could obtain better knowledge tracing results through modeling individualization in learning process. Moreover, CKT can learn meaningful exercise embeddings automatically.

CCS CONCEPTS

• **Information systems** → *Data mining*; • **Social and professional topics** → **Student assessment**;

KEYWORDS

knowledge tracing; convolution neural networks; individualized learning; intelligent education

ACM Reference Format: Shuanghong Shen, Qi Liu, Enhong Chen, Han Wu, Zhenya Huang, Weihao Zhao, Yu Su, Haiping Ma, Shijin Wang. 2020. Convolutional Knowledge Tracing: Modeling Individualization in Student Learning Process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401288

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401288

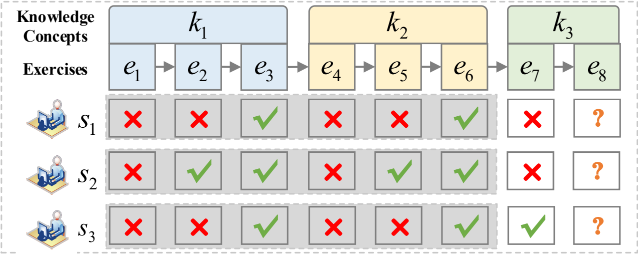

Figure 1: A toy example of the learning process.

1 INTRODUCTION

With the emergence of online education systems on the internet [10], increasing attention has been paid to knowledge tracing (KT). By assessing the knowledge state of students based on their learning interactions, KT can improve the learning efficiency of students and help them better understand their learning process [1]. Nevertheless, most previous KT methods assess the knowledge state without considering the individualization of prior knowledge and learning rates of various students. Actually, the prior knowledge and learning rates differ from student to student, which have proved to be of great significance in learning process [16].

In order to better illustrate the individualization of student, we give a toy example in Figure 1, where 3 students have answered 7 exercises related to 3 different knowledge concepts. As shown in Figure 1, student s_2 could master concepts k_1 and k_2 fastly after fewer mistakes, showing that s_2 had a faster learning rate than student s_1 and s_3. Meanwhile, student s_3 could correctly answer exercise e_7 for the first time, indicating that s_3 may have mastered concepts k_3 already. Unfortunately, the individualized prior knowledge and learning rates of different students are not given in advance, which makes it very challenging to measure them.

To address the challenges in modeling individualization of student, we propose a novel Convolutional Knowledge Tracing (CKT) model to measure individualized prior knowledge and learning rates of students from their learning interaction sequences. Specifically, for individualized prior knowledge, we assess it comprehensively according to students' historical learning interactions. For individualized student learning rates, we design hierarchical convolutional layers to extract the learning rate features by processing several continuous learning interactions simultaneously within a sliding window. Extensive experiments have been conducted on five public datasets to evaluate the performance of CKT, which shows that CKT could get better knowledge tracing results through modeling individualization in student learning process.

2 RELATED WORKS

Knowledge Tracing (KT). Most of existing methods for solving KT problem can be classified into the traditional Bayesian Knowledge Tracing (BKT) [1], and deep learning based methods such as Deep Knowledge Tracing (DKT) [12] and Dynamic Key-Value Memory Networks (DKVMN) [17]. BKT is a classic and widely-used model for modeling student learning [1], which defines two knowledge parameters and two performance parameters for all students. DKT introduces deep learning into KT for the first time [12] and it takes the learning sequence as the input of long short term memory networks (LSTMs) [6] and represents student knowledge states by hidden states. DKVMN pinpoints whether a student is good at specific concepts or not [17].

Convolutional Neural Networks (CNNs). CNNs are invented for computer vision originally [9]. CNNs operate over a fixed-size sliding window of the input sequence, which can extract the connections and changes between several continuous input elements. Moreover, the multi-layer convolutional architecture can extract deep features and creates hierarchical representations over the input sequence, where the nearby input elements interact at lower layers and the distant elements interact at higher layers [4, 11].

3 MODEL ARCHITECTURE

3.1 Problem Definition

Generally, the KT task can be formalized as follows: given the learning sequence $X_N = (x_1, x_2, ..., x_t, ..., x_N)$ with N learning interactions of a student, we aim to assess the student's knowledge state after each learning interaction. In the learning sequence, x_t is an ordering pair $\{e_t, a_t\}$, which stands for a learning interaction. Here e_t represents the exercise being answered at learning interaction t and $a_t \in \{0, 1\}$ indicates whether the exercise e_t has been answered correctly (1 stands for correct and 0 represents wrong).

3.2 Models of CKT

3.2.1 Embedding. Given the dataset with M different exercises in total, we randomly initialize $\mathbf{e}_t \in \mathbb{R}^K$ as the embedding of exercise e_t, which will be learned automatically in training process. Thus, exercises can be converted into an embedding matrix $\mathbf{E} \in \mathbb{R}^{N \times K}$, where K is the number of dimensions [15]. To distinguish influences of right and wrong responses on student knowledge state, inspired by [13], we extend the answer value a_t to a zero vector $\mathbf{a}_t = (0, 0, ..., 0)$ with the same K dimensions as $\mathbf{e}_t$ and express the embedding of learning interaction $\mathbf{x}_t \in \mathbb{R}^{2K}$ as:

$$\mathbf{x}_t = \begin{cases} [\mathbf{e}_t \oplus \mathbf{a}_t], & \text{if} \quad a_t = 1, \\ [\mathbf{a}_t \oplus \mathbf{e}_t], & \text{if} \quad a_t = 0, \end{cases} \tag{1}$$

where $\oplus$ concatenates two vectors. We represente the embedding of learning interaction sequence as $\mathbf{LIS} \in \mathbb{R}^{N \times 2K}$.

3.2.2 Individualized Prior Knowledge. Actually, the prior knowledge is hidden in students' historical learning interactions. First, several researches have proved that students may get similar scores on similar exercises. Second, the scoring rates of students could be seen as the reflection of their overall knowledge mastery [14]. Therefore, we measure individualized prior knowledge of students comprehensively based on their historical learning interactions from two perspectives: Historical Relevant Performance (HRP) and Concept-wised Percent Correct (CPC). Specifically, HRP reflects the concept-specific knowledge mastery of students in detail, while CPC concerns about overall knowledge mastery of students on all knowledge concepts roughly.

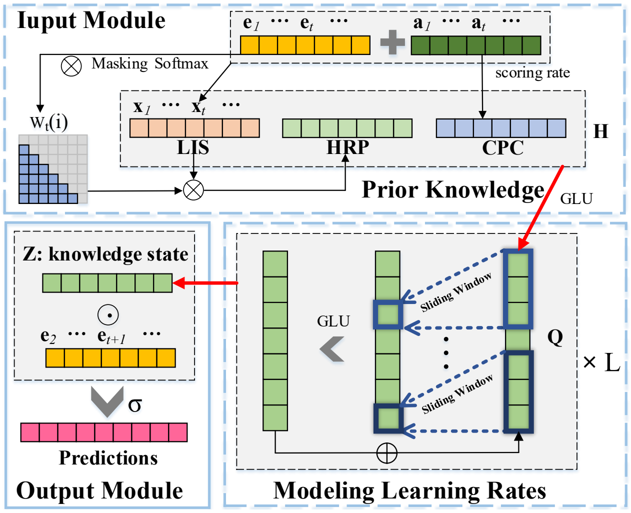

Figure 2: The architecture of CKT model.

(1) Historical Relevant Performance. HRP focuses on measuring student historical performance relevant to the exercise to be answered. To assess the relevance between present answering exercise $\mathbf{e}_t$ and previous answered exercise $\mathbf{e}_i$ ($i \in (1, t-1)$), we compute the relevant coefficients $w_t(i)$ by taking the softmax activation of the masked dot product between $\mathbf{e}_t$ and $\mathbf{e}_i$:

$$\begin{cases} r_t(i) = Masking(\mathbf{e}_i \cdot \mathbf{e}_t), & i \in (t, N), \\ w_t(i) = Softmax(r_t(i)), & i \in (1, N), \end{cases} \tag{2}$$

where *Masking* is the operation (i.e., setting to be $-\infty$) that excludes subsequent learning interactions and $Softmax(r_t(i)) = \frac{\exp(r_t(i))}{\sum_{i=1}^{N}\exp(r_t(i))}$. Then we take advantage of w_i to measure $\mathbf{HRP} \in \mathbb{R}^{N \times 2K}$ by the weighted sum of all historical learning interactions:

$$\mathbf{HRP_t(t)} = \sum_{i=1}^{t-1} w_t(i)\mathbf{x}_i. \tag{3}$$

(2) Concept-wised Percent Correct. CPC accounts for the overall knowledge mastery of student on all knowledge concepts. To measure global knowledge mastery of students, CPC is made up of student percent correct on each knowledge concept. We calculate $\mathbf{CPC} \in \mathbb{R}^{N \times M}$ by counting student scoring rate:

$$\mathbf{CPC_t(m)} = \frac{\sum_{i=0}^{t-1} a_i^m == 1}{count(e^m)}, \tag{4}$$

where $m \in (1, M)$ represents exercise related to the knowledge concept m, $count(e^m)$ is the number of times that exercise e^m has been answered, $\sum_{i=0}^{t-1} a_i^m == 1$ is the number of times that exercise e^m has been answered correctly.

Then we concatenate **LIS** with **HRP** and **CPC** as the matrix **H**. Inspired by the gate mechanism in LSTM [6], we pass **H** through the gated linear unit (GLU) [2], which plays the role of non-linearity and controls the information flowing in the learning process:

$$\begin{cases} \mathbf{H} = \mathbf{LIS} \oplus \mathbf{HRP} \oplus \mathbf{CPC}, \\ \mathbf{Q} = (\mathbf{H} * \mathbf{W_1} + \mathbf{b_1}) \otimes \sigma(\mathbf{H} * \mathbf{W_2} + \mathbf{b_2}), \end{cases} \tag{5}$$

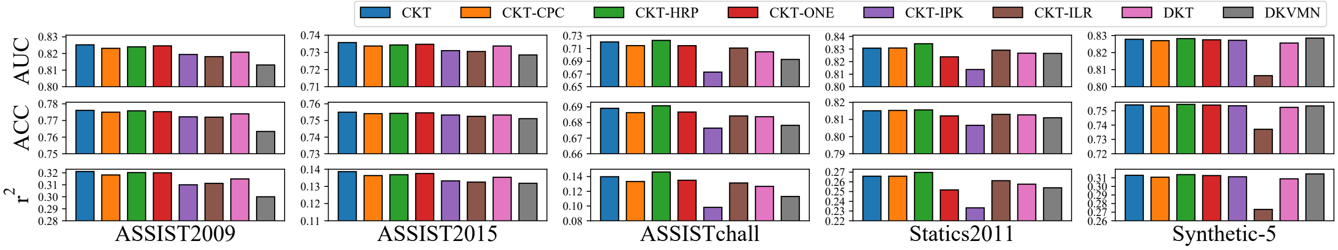

Figure 3: Results of comparison methods on student performance prediction.

where $\mathbf{W_1} \in \mathbb{R}^{(4K+M)\times K}$, $\mathbf{b_1} \in \mathbb{R}^{K}$, $\mathbf{W_2} \in \mathbb{R}^{(4K+M)\times K}$, $\mathbf{b_2} \in \mathbb{R}^{K}$ are the parameters to learn, σ is the sigmoid function, and $\otimes$ is the operation of point-wise multiplication.

3.2.3 *Individualized Learning rate.* Next, we are going to extract the learning rate features from the matrix $\mathbf{Q}$. Individualized learning rates represent the absorptive capacities for knowledge of different students. The sequence of student learning interactions could reflect learning rate in a way that students with high learning rates can master knowledge concepts fastly, while others have to spend more time trying and failing. Therefore, we design hierarchical convolutional layers to extract the learning rate features contained in continuous learning interactions of students.

In our design, one-dimensional convolution is applied. The sliding window is parameterized as $\mathbf{W} \in \mathbb{R}^{2d\times K}$, $\mathbf{b} \in \mathbb{R}^{K}$. We have masked out (i.e., set to be 0) the second half of the sliding window to prevent the convolutional operation involving subsequent learning interactions. Thus a sliding window takes d continuous learning interactions as input and maps them to a single output element. The number of feature maps is also set to be K. Then, GLU is utilized as non-linearity and realizes a simple gating mechanism over the output of the convolution layer, which controls whether the knowledge would be forgotten in learning. Besides, for speeding up the training process, we add residual connections [5] from the input to the output of the convolutional layer. Then we stack L same convolutional layers on top of each other to make up hierarchical convolutional layers, where lower layers capture the learning rates in recent period and higher layers could monitor farther range .

The output matrix of the hierarchical convolutional layers $\mathbf{Z} \in \mathbb{R}^{N\times K}$ represents student knowledge state. We utilize the dot product of present student knowledge state and the embedding of next coming exercise to predict student performance:

$$\begin{cases} y_{t+1} = z_t \cdot \mathbf{e}_{t+1}, \\ p_{t+1} = \sigma(y_{t+1}). \end{cases} \tag{6}$$

3.2.4 *Objective Function.* To learn all parameters in CKT and the exercise embedding matrix $\mathbf{E}$ in training process, we choose the cross entropy log loss between the prediction p_t and actual answer a_t as the objective function in CKT model, which was minimized using Adam optimizer [7] on mini-batches:

$$L = -\sum_{t=1}^{N}(a_t \log p_t + (1 - a_t)\log(1 - p_t)). \tag{7}$$

4 EXPERIMENTS

4.1 Datasets Description

Four real-world public datasets and one synthetic dataset have been used to evaluate the effectiveness of CKT. Table 1 shows the statistics of all datasets.

Table 1: Statistics of all datasets.

Datasets	Statistics			
	Students	Concepts	Records	Avg.length
ASSIST2009	4,151	110	325,637	78
ASSIST2015	19,840	100	683,801	36
ASSISTchall	1,709	102	942,816	552
Statics2011	333	1,223	189,297	568
Synthetic-5	4,000	50	200,000	50

- **ASSIST2009** is collected from the ASSISTments [3], an online tutoring system created in 2004. The data is gathered from skill builder problem sets where students need to work on similar exercises to achieve mastery.
- **ASSIST2015** also comes from ASSISTments, which covers response records in 2015.
- **ASSISTChall** is utilized in the 2017 ASSISTments data mining competition. Researchers collected it from a longitudinal study, which tracks students from their use of ASSISTments.
- **Statics 2011** is obtained from a college-level engineering statics course [8]. We have concatenated problem name and step name as the knowldge concept.
- **Synthetic-5** is published on the DKT paper [12], which simulates virtual students learning virtual concepts. It is worth noting that the simulated virtual learning process does not take the individualized student learning rates into account.

4.2 Comparison methods

We compare CKT with several variants of CKT and baselines. For a fair comparison, all these methods are tuned to have the best performances. To facilitate further research in CKT we have published our code [1] . The details of comparison methods are:

- **CKT-ONE** with only one concolutional layer.
- **CKT-HRP** measures prior knowledge only from HRP.
- **CKT-CPC** measures prior knowledge only from CPC.
- **CKT-ILR** only models individualized learning rate.
- **CKT-IPK** only models individualized prior knowledge.
- **DKT** leverages recurrent neural network to assess student knowledge state [12]. We utilized LSTM in our implemention.
- **DKVMN** takes advantage of memory network to get interpretable student knowledge state [17].

[1] https://github.com/bigdata-ustc/Convolutional-Knowledge-Tracing

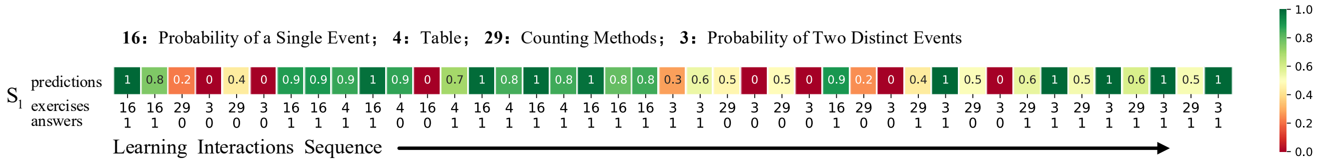

Figure 4: Visualization cases of individualized knowledge tracing result of student.

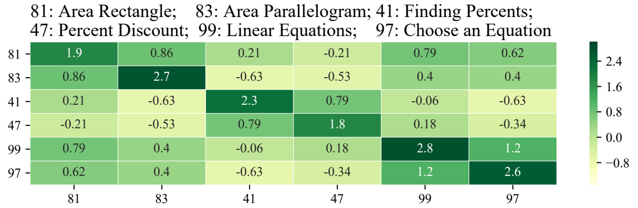

Figure 5: Visualization of exercise relevant coefficients

4.3 Experimental Results

In order to evaluate the performance of CKT, we conduct extensive experiments. For providing robust evaluation results, the performance was evaluated in terms of Area Under Curve (AUC), Accuracy (ACC) and the square of Pearson correlation (r^2).

4.3.1 Student performance prediction. In this experiment, we assess the effectiveness of CKT by predicting student performances on every learning interaction. The experiment results are depicted in Figure 3. From the figure, we can easily see that CKT and its variants get the best prediction results on all four real-world datasets. Moreover, CKT gets higher promotions on dataset ASSISTchall and Statics2011 with longer learning sequence length. This observation demonstrates that CKT can model individualization better with the longer learning sequences that provide more complete and precise prior information. On the other hand, for the synthetic dataset with no differences in student learning rates, CKT's variant CKT-ILR, which models only the individualized learning rate, declines significantly in predicting student performance.

4.3.2 Visualization of knowledge tracing results. As indicated in Figure 4, we present two cases of individualized knowledge tracing results on dataset ASSIST2009. The upper part of Figure 4 gives the knowledge concepts that different exercises are corresponding to. As can be seen from the figure, CKT can well model individualized learning rates and prior knowledge of various students as expected. As shown in the figure, student S_1 has a fast learning rate on the concepts *16: Probability of a Single Event* and *4: Table*, but gets stuck for a while in learning deeper concepts *3: Probability of Two Distinct Events* and *29: Counting Methods*.

4.3.3 Exercise embeddings learning. Exercise embeddings are used to be annotated by human experts, which is time-consuming and laborious. In CKT, we randomly initialize the exercise embeddings for training and CKT can automatically learn meaningful exercise embeddings in training process. As shown in Figure 5, we visualize the relevant coefficients of 6 various exercises corresponding to the concepts of either *Area*, *Equation* or *Percent* on dataset ASSIST2009. We can discover that the relevant coefficients tend to be higher between exercises related to the same concept and drop significantly among different concepts.

5 CONCLUSIONS

In this paper, we proposed a novel model called *C*onvolutional *K*nowledge *T*racing (CKT) to modeling individualization of student in KT task. Specifically, we measured individualized prior knowledge from students' historical learning interactions (i.e., HRP and CPC). Then, we designed hierarchical convolutional layers to extract individualized learning rates based on continuous learning interactions. Extensive experiment results indicated that CKT could get better knowledge tracing results through modeling individualization in student learning process.

ACKNOWLEDGMENT

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), the National Natural Science Foundation of China (Grants No. U1605251, 61922073), and the Iflytek joint research program.

REFERENCES

[1] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *UMUAI* 4, 4 (1994), 253–278.
[2] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *ICML*. JMLR. org, 933–941.
[3] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *USER-ADAP* 19, 3 (2009), 243–266.
[4] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*. JMLR. org.
[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[8] Kenneth R Koedinger, Ryan SJd Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. 2010. A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining* 43 (2010), 43–56.
[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[10] Antonija Mitrovic. 2012. Fifteen years of constraint-based tutors: what we have achieved and where we are going. *USER-ADAP* 22, 1-2 (2012), 39–72.
[11] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. 2020. Geom-GCN: Geometric Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
[12] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *NeurIPS*. 505–513.
[13] q. liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu. 2019. EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.
[14] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural Cognitive Diagnosis for Intelligent Education Systems. In *The 34th AAAI Conference on Artificial Intelligence (AAAI'2020)*.
[15] Hao Wang, Tong Xu, Qi Liu, Defu Lian, Enhong Chen, Dongfang Du, Han Wu, and Wen Su. 2019. MCNE: An End-to-End Framework for Learning Multiple Conditional Network Representations of Social Network. In *Proceedings of the 25th ACM SIGKDD*. 1064–1072.
[16] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *AIED*. Springer, 171–180.
[17] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *WWW*. 765–774.