Learning Process-consistent Knowledge Tracing

Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang,
Wei Huang, Yu Yin, Yu Su, Shijin Wang

KDD, 2021

Reporter: Shuanghong Shen

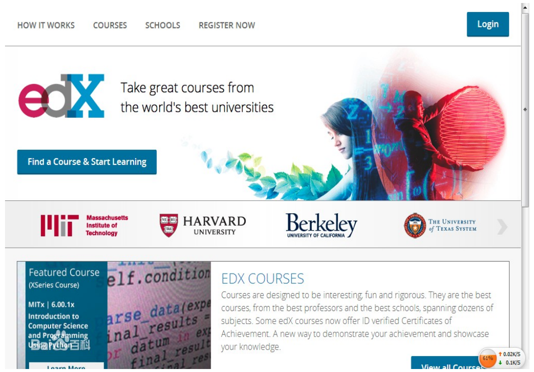
Outline

Background

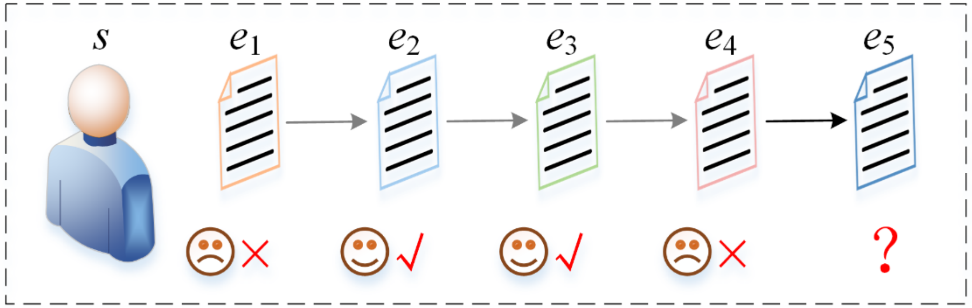
Knowledge tracing (KT)

A family of machine learning sequence models that are capable of using educationally related data to monitor the changing knowledge state of students.

Intelligent Tutoring System & MOOC



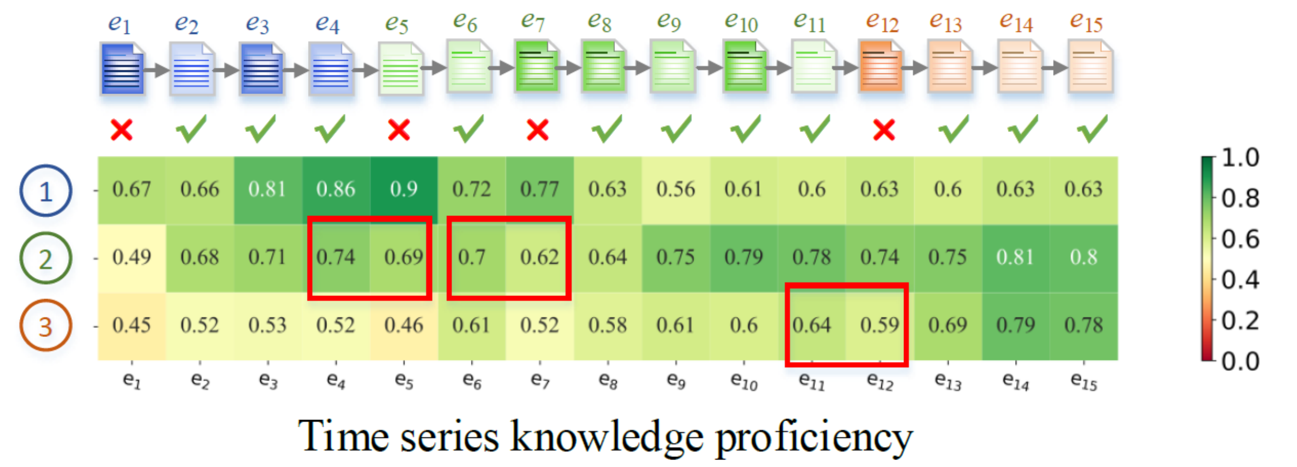
Knowledge State?



Background

A dilemma

Once the student gives wrong answers, existing KT models argue that his/her knowledge state on corresponding knowledge concepts will decline.



Exercises	Knowledge Concepts
e_1, e_2, e_3, e_4	①. Absolute Value
$e_5, e_6, e_7, e_8, e_9, e_{10}, e_{11}$	②. Addition and Subtraction Integers
$e_{12}, e_{13}, e_{14}, e_{15}$	③. Ordering Integers

Figure 1: A toy example of the evolution process of a student's knowledge states traced by DKT, where the student has answered 15 exercises on 3 knowledge concepts. In the left figure, the color of the heatmap or the number in the small box refers to the knowledge state of the student after answering the exercise. The red boxes indicate that DKT thinks the knowledge state will decline after wrong answers. The right table gives the relations between exercises and knowledge concepts.

Background

In fact

- mistakes are seen as natural elements of learning processes
- students can learn from errors and foster learning progress through a favorable error climate
- In this paper, we explore a new paradigm for the KT task by directly modeling students learning process

Background

Challenge

1. how to define the learning process and convert it into a proper form for modeling

2. the learning gain, which represents the knowledge that students acquire in learning, is implicit and changeable in learning process

3. students' knowledge will also decrease over time, which commonly manifests as forgetting, is also necessary to be considered in the KT task

Background

Related works

Learning Gain

- Learning gain is different from learning outcomes in that learning gain compares performance at two points in time, while learning outcome concentrates on the output level at a single point in time.
- Luckin et al. calculated learning gain as LG = post −pre
- Normalized Learning Gain (NLG) is a widely used adjusted measurement: NLG = $\frac{post - pre}{1 - pre}$,
- Quantized Learning Gain (QLG) : instantiate students' learning gains as High or Low

Forgetting

- *Ebbinghaus forgetting curve* theory indicates that students' knowledge proficiency may decline due to the forgetting factor
- Nedungadi and Remya incorporated forgetting based on the assumption that the learned knowledge decays exponentially over time
- Huang et al. model students' knowledge proficiency with both learning and forgetting theories

Outline

Preliminary

Definition of Learning Process:

- When an exercise is assigned to the student, he/she spends a certain time on answering it according to his/her learned knowledge. The learning process keeps repeating the above answering behavior on different exercises, where there is an interval time between adjacent answering interactions.

- Therefore, we denote the learning process of a student as $\boldsymbol{x} = \{ (e_1, at_1, a_1), it_1, (e_2, at_2, a_2), it_2, \ldots, (e_t, at_t, a_t), it_t \}$, where the tuple (e_1, at_1, a_1) represents a basic learning cell in learning process, e_t is the exercise, at_t is the answer time the student spent on answering e_t, and a_t represents the binary correctness label (1 represents correct and 0 for wrong), it_t stands for the interval time between the learning cells.

Preliminary

Definition of Knowledge Tracing:

Given students' learning sequence $\boldsymbol{x} = \{ (e_1, at_1, a_1), it_1, (e_2, at_2, a_2), it_2, \ldots, (e_t, at_t, a_t), it_t \}$, the KT task aims to monitor students' changing knowledge state during the learning process and predict their future performance at the next learning step t+1, which can be further applied to individualize students' learning scheme and maximize their learning efficiency.

Preliminary

Embeddings

Time Embedding

- We discretize the interval time by the minutes and the answer time by the seconds. Besides, we set all the interval time longer than one month as one month
- We represent the discretized answer time by two embedding matrices

Learning Embedding

- Learning embedding is the embedding of the basic learning cell
- We first represent the exercise set by an embedding matrix $\boldsymbol{E}$
- For the answer at , i.e., 0 or 1, we expand it to a all-zero or all-one vector
- For getting the learning embedding of the basic learning cell (e_t, at_t, a_t), we concatenate them together and apply a multi-layer perceptron (MLP) to deeply fuse the exercise embeddings, answer time embeddings, and answer embeddings

Embeddings

Knowledge Embedding

- Knowledge embedding is served to store and update the knowledge state of students during the learning process
- In LPKT, the knowledge embedding is initialized as an embedding matrix ***h***, each row of ***h*** represents the knowledge mastery of the corresponding knowledge concept.
- At each learning interaction, the learning gain on each knowledge concept modeled by LPKT are updated into the knowledge embedding, the forgetting effects are also included in it simultaneously.
- Manually-labeled Q-matrix may be deficient because of inevitable errors and subjective bias, we define an enhanced Q-matrix $\tilde{\boldsymbol{q}} \in \mathbb{R}^{J\times M}$ where q_{jm} will be set as a small positive value γ rather than 0 even if k_m is not in e_j
- We leave the exploration to learn the specific weights in the Q-matrix as a future work

Outline

The LPKT Model

Model Architecture

LPKT is consisted of three modules: (1) learning module, (2) forgetting module, and (3) predicting module after a student has answered an exercise:

- The learning module models learning gains compared with the previous learning interaction.
- The forgetting module is utilized to measure how much knowledge will be forgotten as time goes on.
- Then, the learning gains and forgotten knowledge will be taken advantage of to update the student's previous knowledge state for achieving their latest knowledge state.
- Finally, the predicting module is proposed to predict the student's performance on the next exercise according to his/her latest knowledge state.

The LPKT Model

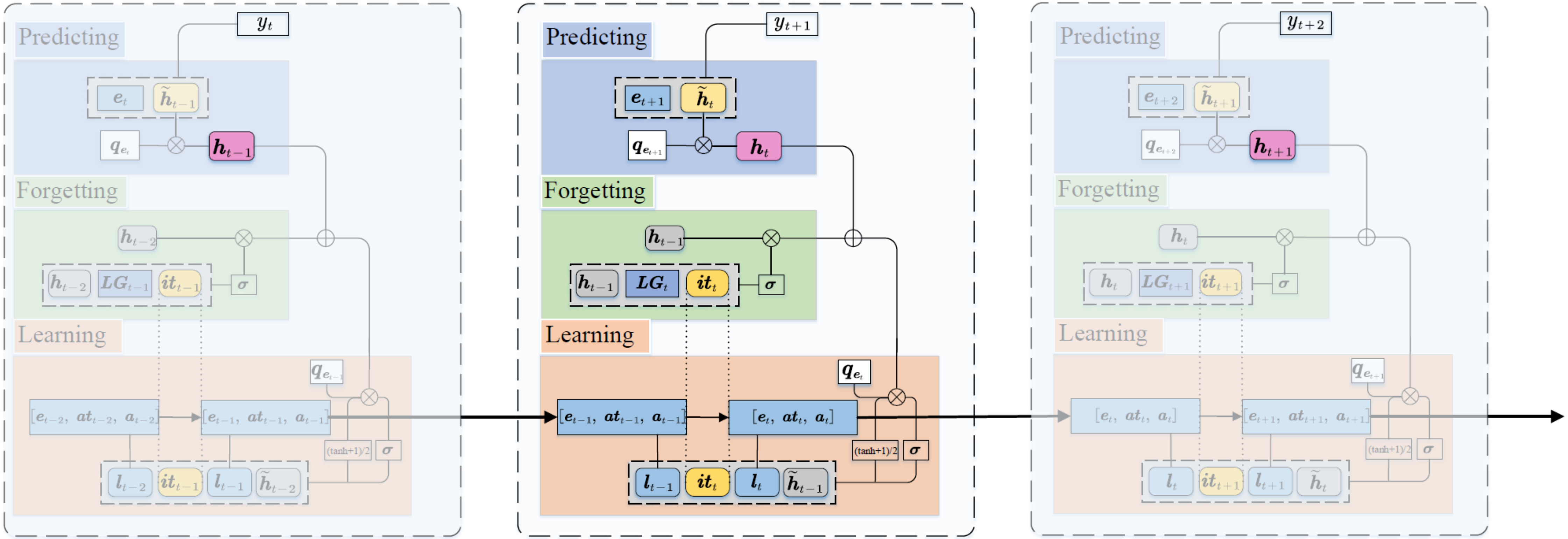


Figure 2: The architecture of the LPKT model.

The LPKT Model

Learning Module

- For modeling learning gains precisely, we should consider the differences in students‘ performance at two continuous learning interactions of students. In LPKT, we realize the modeling of learning gain through concatenating students' previous learning embedding $\boldsymbol{l}_{t-1}$ and present learning embedding $\boldsymbol{l}_t$ as the basic input element
- For the interval time, we concatenate it into the basic input element in the timeline between the two continuous learning embeddings. For previous knowledge state, to focus on the knowledge state on the related knowledge concepts of the present exercise, we first multiply $\boldsymbol{h}_{t-1}$ and the knowledge concept vector $\boldsymbol{q}_{e_t}$ of present exercise and get the related knowledge state $\widetilde{\boldsymbol{h}}_{t-1}$
- Besides, considering the diversity of learning gains, we incorporate two influencing factors of the learning gains into LPKT, which are the interval time and students' previous knowledge state respectively

$$\boldsymbol{lg}_t = tanh(\mathbf{W}_2^T[\boldsymbol{l}_{t-1} \oplus \boldsymbol{it}_t \oplus \boldsymbol{l}_t \oplus \widetilde{\boldsymbol{h}}_{t-1}] + \boldsymbol{b}_2),$$

The LPKT Model

Learning Module

- Considering that not all learning gains can be transformed into the growth of students' knowledge completely, we further design a learning gate to control the students' absorptive capacity of knowledge:

$$\Gamma_t^l = \sigma(\boldsymbol{W}_3^T[\boldsymbol{l}_{t-1} \oplus \boldsymbol{it}_t \oplus \boldsymbol{l}_t \oplus \widetilde{\boldsymbol{h}}_{t-1}] + \boldsymbol{b}_3),$$

-

- Due to the output range of *tanh* function is (-1, 1), we apply a linear transformation to project the range of $\boldsymbol{lg}_t$ from (-1, 1) to (0, 1). Therefore, the learning gains will be always positive, which is in line with our assumption that students' can consistently acquire knowledge at each learning interaction.

$$LG_t = \Gamma_t^l \cdot ((\boldsymbol{lg}_t + 1)/2),$$
$$\widetilde{LG}_t = \boldsymbol{q}_{e_t} \cdot LG_t,$$

The LPKT Model

Forgetting Module

- Forgetting phenomenon affects how much knowledge will be forgotten as time goes on.
- A simple manual-designed exponential decay function is not sufficient for capturing complex relations between knowledge state and the interval time.
- For modeling the complex forgetting effects, we design a forgetting gate in LPKT, which applies a MLP to learn the degree of loss information in knowledge matrix based on three factors: (1) students' previous knowledge state, (2) students' present learning gains, and (3) interval time

$$\Gamma_t^f = \sigma(\mathbf{W}_4^T[\boldsymbol{h}_{t-1} \oplus \boldsymbol{LG}_t \oplus \boldsymbol{it}_t]) + \boldsymbol{b}_4),$$

- Then, we can eliminate the influence of forgetting, and the knowledge state $\boldsymbol{h}_t$ after students accomplish the t-th learning interaction will be updated as follows:

$$\boldsymbol{h}_t = \widetilde{\boldsymbol{LG}}_t + \Gamma_t^f \cdot \boldsymbol{h}_{t-1}.$$

The LPKT Model

Predicting Module

- In a real learning environment, given an exercise to the student, he/she will try to solve it by applying his/her knowledge to the corresponding knowledge concepts. Therefore, we use the related knowledge state to infer the student's performance.

$$y_{t+1} = \sigma(\boldsymbol{W}_5^T [\boldsymbol{e}_{t+1} \oplus \widetilde{\boldsymbol{h}}_t] + \boldsymbol{b}_5),$$

Objective Function

- To learn all parameters in LPKT, we also choose the cross-entropy log loss between the prediction and actual answer as the objective function:

$$\mathbb{L}(\theta) = -\sum_{t=1}^{T} (a_t \log y_t + (1 - a_t) \log(1 - y_t)) + \lambda_\theta ||\theta||^2,$$

Outline

Experiments

Datasets

Statistics	Datasets		
	ASSIST2012	ASSISTchall	EdNet-KT1
Students	29,018	1,709	784,309
Exercises	53,091	3,162	12,372
Concepts	265	102	141
Answer Time	26,747	1,326	9,292
Interval Time	29,748	2,839	41,830
Avg.length	93.45	551.68	121.48

Table 1: Statistics of all datasets.

ASSIST2012: https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-withaffect

ASSISTchall: https://sites.google.com/view/assistmentsdatamining/dataset

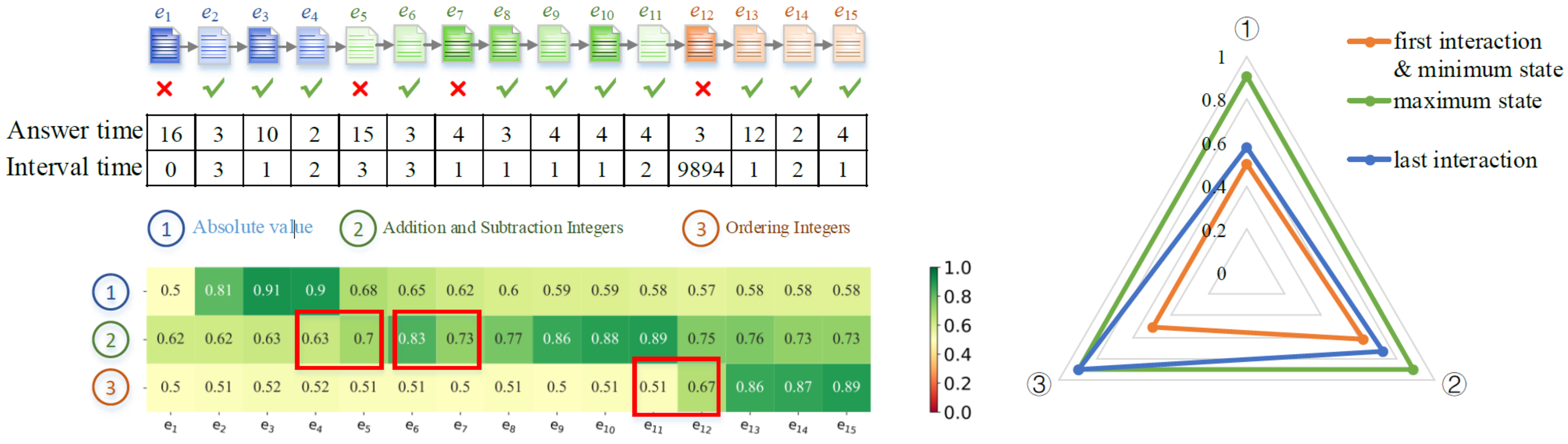
EdNet-KT1: http://ednet-leaderboard.s3-website-ap-northeast-1.amazonaws.com/

Experiments

Baseline Methods

- **DKT** leverages recurrent neural network to assess student knowledge state. We utilized LSTM in our implementation.
- **DKT+** is an extended variant of DKT , which attempts to solve two major problems in DKT. The first problem is that DKT fails to reconstruct the observed input and the second one is the predicted performance of DKT across time-steps is not consistent.
- **DKVMN** takes advantage of memory network to get interpretable student knowledge state. It defines a static matrix called *key* matrix to store latent knowledge concepts and a dynamic matrix called *value* matrix to store and update the corresponding knowledge state through read and write operations over time.
- **SAKT** applies the transformer structure to the KT task. It proposes a self-attentive model for knowledge tracing.
- **CKT** introduces convolutional windows in CNN to model the individualized learning rate of students in learning process
- **AKT** is the context-aware attentive knowledge tracing model. It uses the two self-attentive encoders to learn context-aware representations of the exercises and answers. The knowledge evolution model is referred to the knowledge retriever, which uses an attention mechanism to retrieve knowledge acquired in the past that is relevant to the current exercise.

Knowledge State Visualization



(a) Time series knowledge proficiency

(b) Radar diagram

Figure 3: The evolution process of a student's (the same student in Figure 1) knowledge states traced by LPKT. In sub-figure (a), the top part indicates his/her performance at each time step, the answer time and interval time. Sub-figure (b) is the radar diagram of the student's knowledge state at the first interaction and the last interaction, his/her maximum and minimum knowledge state in learning process are also depicted on it.

Knowledge State Visualization

LPKT can capture reasonable knowledge state of students, which is in consistent with their learning process.

- ✓ our proposed LPKT method can capture the student's learning gains from both wrong and right learning interactions
- ✓ if the student does not practice on some knowledge concepts, his/her knowledge state on these concepts will reduce gradually as time goes on
- ✓ the general changing process of the student's knowledge state is in line with his/her learning process. At the first learning interaction, his/her knowledge state is the minimum. During the learning process, the student keeps absorbing new knowledge and his/her knowledge state achieves the maximum, which can be reflected by the increased areas of the radar diagram that indicates the student's knowledge proficiency. At the last learning interaction, the student's knowledge state presents a certain degree of reduction in comparison with the maximum but is still better than the beginning.

Experiments

Student Performance Prediction

LPKT outperforms all other KT methods on all datasets and metrics.

Methods	ASSIST2012				ASSISTchall				EdNet-KT1			
	RMSE	AUC	ACC	r^2	RMSE	AUC	ACC	r^2	RMSE	AUC	ACC	r^2
DKT	0.4241	0.7289	0.7360	0.1468	0.4471	0.7213	0.6907	0.1425	0.4508	0.6836	0.6889	0.1008
DKT+	0.4239	0.7295	0.7254	0.1497	0.4502	0.7101	0.6842	0.1308	0.4601	0.6429	0.6733	0.0635
DKVMN	0.4261	0.7228	0.7329	0.1398	0.4503	0.7108	0.6842	0.1302	0.4538	0.6741	0.6843	0. 0913
SAKT	0.4258	0.7233	0.7339	0.1403	0.4626	0.6605	0.6694	0.0822	0.4524	0.6794	0.6862	0.0964
CKT	0.4234	0.7310	0.7365	0.1497	0.4455	0.7263	0.6924	0.1488	0.4519	0.6811	0.6871	0.0984
AKT	0.4100	0.7740	0.7554	0.2035	0.4317	0.7655	0.7141	0.2015	0. 4241	0. 7701	0.7287	0.2059
LPKT	**0.4069**	**0.7772**	**0.7583**	**0.2145**	**0.4153**	**0.8008**	**0.7424**	**0.2609**	**0.4234**	**0.7721**	**0.7300**	**0.2085**

Table 2: Results of comparison methods on student performance prediction. LPKT outperforms all baselines on all datasets.

Experiments

Ablation Experiments

We conduct some ablation experiments to further show how each module in LPKT affects final results

- ✓ the common phenomenon of forgetting plays a critical role in learning process, which can cause the biggest decline of the predictive results if we do not consider it.
- ✓ modeling the learning gain indeed performs better than modeling only the learning outcomes in knowledge tracing, because the learning gain can better reflect the dynamic changes of students' knowledge state.
- ✓ the answer time and interval time is essential and necessary information in the whole learning process, which is harmful to accurately model the learning process if omitted.

Methods	learning	forgetting	time	RMSE	AUC	ACC	r^2
LPKT-L	✓		✓	0.4112	0.7659	0.7531	0.1980
LPKT-F		✓	✓	0.4087	0.7734	0.7554	0.2075
LPKT(no time)	✓	✓		0.4077	0.7759	0.7571	0.2115
LPKT	✓	✓	✓	**0.4069**	**0.7772**	**0.7583**	**0.2145**

Table 3: Results of ablation experiments on ASSIST2012.

Experiments

Ablation Experiments

We also conduct experiments to evaluate that if LPKT can better model students' learning process

- ✓ we set four different lengths: 200, 100, 20, and 10, respectively. The shorter the learning sequence, the more incomplete the learning process
- ✓ the gap between LPKT and AKT becomes wider (i.e., the reduction of experimental results of LPKT is less than AKT) as the learning sequence is going shorter
- ✓ LPKT is less affected by incomplete learning sequences, so that LPKT indeed better models students' learning process.

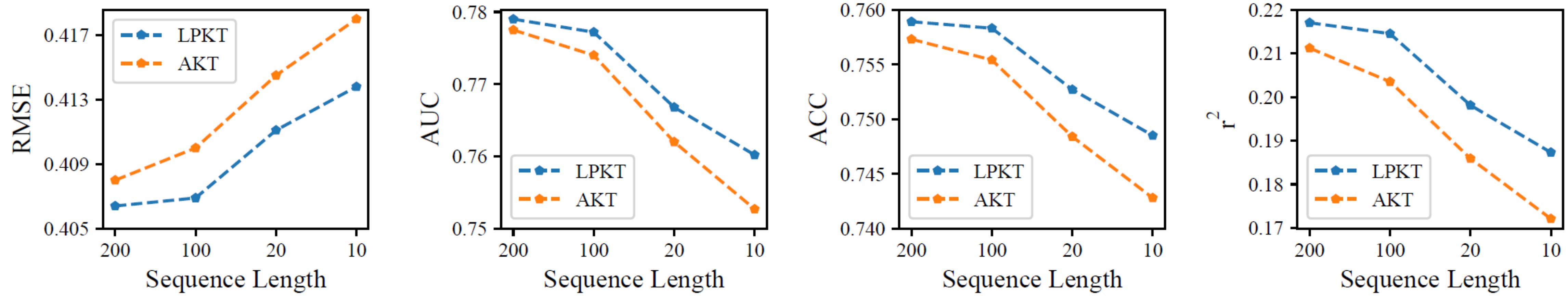


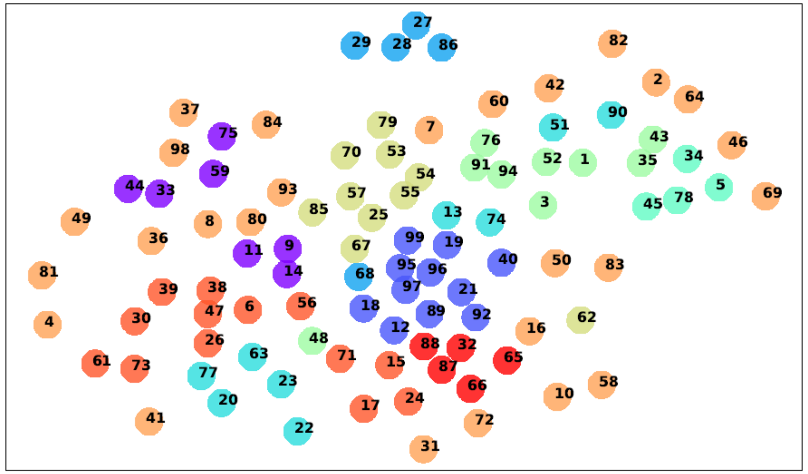
Figure 4: Comparison results of the influence of learning sequence length of LPKT and AKT on ASSISTchall.

Experiments

Exercises Clustering

LPKT can learn meaningful embeddings of exercises after training

- ✓ we randomly choose 100 exercises among the 3162 exercises in dataset ASSISTchall and visualize the embeddings of these exercises utilizing the T-SNE method . As shown in Figure \ref{f5}, we can see that the learned embeddings of exercises in LPKT can be clearly split into 10 concepts and the clustering results show well meanings.
- ✓ For example, exercises 89, 95, 96, 97 with same concept *subtraction* are split together and exercises 53, 54, 55, 57 with same concept *evaluating-functions* are also in the same cluster



(a) Exercises clustering results. Randomly selected 100 exercises in the 3162 exercises of ASSISTchall are clustered into ten concepts. Exercises under the same concept are labeled in the same color and the number stands for the index of exercises.

Knowledge Concepts	Index of Exercises	Knowledge Concepts	Index of Exercises
pattern-finding	13, 17, 18, 19, 20, 22, 28, 29, 30, 31, 32, 91, 92, 93, 94, 99, 100	equivalent-fractions-decimals-percents	47, 49, 51, 52
square-root	33, 34, 35, 36, 37, 38, 39, 40, 41	subtraction	62, 87, 88, 89, 90, 95, 96, 97, 98
symbolization-articulation	7, 8, 9, 10	transformations-rotations	23, 24
point-plotting	1, 2, 3, 5, 25, 26, 27	reading-graph	4, 14, 15, 16
supplementary-angles	58, 73, 74	inducing-functions	11, 12, 21, 42, 43, 44, 45, 46, 64, 69, 71
evaluating-functions	53, 54, 55, 56, 57, 63	addition	65, 66, 67, 68
transversals	59, 72	isosceles-triangle	60, 61
equation-solving	70, 82, 85, 86	equation-concept	6
interpreting-numberline	79, 80, 81	venn-diagram	75, 76
percent-of	77, 78	percents	83, 84
of-means-multiply	48	rate	50

(b) Manually-labeled knowledge concepts of the above 100 exercises.

Outline

Conclusions

- A new paradigm for knowledge tracing through modeling students' learning process
- A novel model named Learning Process-consistent Knowledge tracing (LPKT)
- Formalized the learning process as the basic learning cell and interval time
- Modeled the learning gain and its diversity learning process
- Designed a forgetting gate to determine the reduction of students' knowledge over time
- Extensive experiments on three public datasets demonstrated the interpretability and effectiveness of LPKT

Future Works

- keep exploring better ways to model students' learning process
- study how to automatically learn the specific weights in the Q-matrix
- mathematical theory proof

Q&A

Thank you for listening

Q & A