Learning from Interpretable Analysis: Attention-Based Knowledge Tracing

Jia Zhu[1,2], Weihao Yu[1], Zetao Zheng[1](✉), Changqin Huang[1], Yong Tang[1], and Gabriel Pui Cheong Fung[3]

[1] Guangzhou Key Laboratory of Big Data and Intelligent Education, School of Computer Science, South China Normal University, Guangzhou, China
{jzhu,whyu,ztzheng,cqhuang,ytang}@m.scnu.edu.cn
[2] Deceneuron Intelligence Co., Ltd, Hong Kong, China
[3] Department of Systems Engineering and Engineering Management, The Chinese University of HongKong, Hong Kong, China
pcfung@se.cuhk.edu.hk

Abstract. Knowledge tracing is a well-established problem and non-trivial task in personalized education. In recent years, many existing works have been proposed to handle the knowledge tracing task, particularly recurrent neural networks based methods, e.g., Deep Knowledge Tracing (DKT). However, DKT has the problem of vibration in prediction outputs. In this paper, to better understand the problem of DKT, we utilize a mathematical computation model named Finite State Automaton(FSA), which can change from one state to another in response to the external input, to interpret the hidden state transition of the DKT when receiving inputs. And we discover the root cause of the two problems is that the DKT can not handle the long sequence input with the help of FSA. Accordingly, we propose an effective attention-based model, which can solve the above problem by directly capturing the relationships among each item of the input regardless of the length of the input sequence. The experimental results show that our proposed model can significantly outperform state-of-the-art approaches on several well-known corpora.

Keywords: Knowledge tracing · Interpretable analysis · Self attention

1 Introduction

With the development of modern technologies, online platforms for intelligent tutoring systems(ITS) and massive open online courses are becoming more and

Supported by the National Natural Science Foundation of China (No.61877020, No.U1811263 and No.61772211), the Key-Area Research and Development Program of Guangdong Province, China (No.2018B010109002) and the Science and Technology Project of Guangzhou Municipality, China (No.201904010393), as well as the Guangzhou Key Laboratory of Big Data and Intelligent Education (No.201905010009).

© Springer Nature Switzerland AG 2020

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 364–368, 2020.
https://doi.org/10.1007/978-3-030-52240-7_66

more prevalent. And knowledge tracing (KT) is considered to be critical for personalized learning in ITS. KT is the task of modeling students' knowledge state based on historical data, which represents the mastery level of knowledge.

One of the well-known methods to solve the KT problem is recurrent neural networks (RNNs) based model called deep knowledge tracing (DKT) [5]. Although DKT achieves impressive performance for the KT task, it still exists the vibration in prediction outputs [9]. This is unreasonable as students' knowledge state is expected to transit gradually over time, but not to alternate between mastering and not-yet-mastered.

To find out the root cause of the problem, we utilize FSA as an interpretable structure which can be learned from DKT because FSA has a more interpretable inner mechanism when processing sequential data [3]. We built an FSA for DKT referring [3] to interpret how elements on each input sequence affect the hidden state of DKT. When an input item was accepted by the FSA, it represents that this item has a positive effect on the final prediction outputs of the model, and vice versa. We display the acceptance rate of every input sequence in Fig. 1. We can draw the conclusion from Fig. 1 that the longer the input sequence, the higher the proportion of rejected items, and the lower prediction accuracy. This phenomenon is consistent with the description in [7], who points out that LSTM [2] has the weakness of capturing feature when the input sequence is too long. Accordingly, we proposed a model to solve the problem of long sequence input in KT and experiments show that our proposed model is effective in solving the problem we discovered above.

Our contributions are three-fold. Firstly, to the best of our knowledge, we are the first group to adopt FSA to provide deep analysis on KT task. By interpreting the learning state change using FSA, we can obtain a better understanding of the problem of existing RNN based methods. Secondly, according to the interpretable analysis, we propose a multi-head attention model to handle the problem of long sequence input in KT. Lastly, we evaluate our model on real-world datasets and the results show that our model improves the state-of-the-art baselines.

2 Proposed Models

In this section, we will describe the KTA in briefly. The overall structure of the model is shown in Fig. 2. **(1) Embedding Layer:** The tuples that contain the questions and the corresponding answers are first projected into real-value vectors, namely one-hot embeddings. **(2) Feature Extraction:** After that, The vectors are fed into a feature extractor, which aims at capturing the latent dependency relationships among the inputs. The main component of the feature extractor consists of N identical blocks. Each block has two sub-layers. The first is a multi-head self-attention mechanism [8], the critical element of the extractor, and the second is a fully connected feed-forward network [8]. Self-attention achieves the extraction of the global relationship by calculating the similarity of each item among the input sequence using the scaled dot-product attention [8]. Here, the attention is calculated h times, which allows the model to

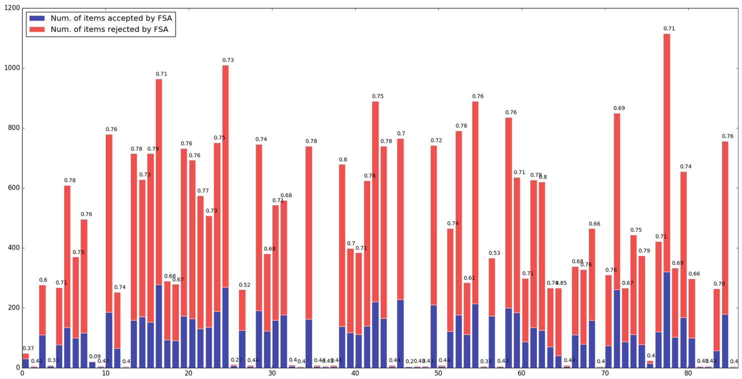

Fig. 1. Accept/Reject States of DKT. The values above each bar represent the proportion of the rejected items in an input sequence.

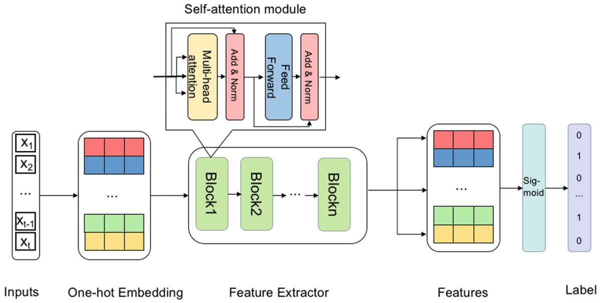

Fig. 2. An illustration of our KTA model.

learn relevant information in different representative sub-spaces, and making it so-called multi-head. **(3) Prediction and Loss:** On the prediction stage, only the topmost outputs of attention sub-layer are taken to a Sigmoid function to make the final decision. The prediction and optimization processes are the same as [9], we would not elaborate here.

3 Experiments

AUC Results. We evaluated our models on four popular datasets which are also used in [9]. We also select four popular methods for comparison, PFA [4], BKT [1], DKT [6], DKT+ [9]. Table 1 displays the AUC results of all the datasets.

According to Table 1, our proposed model achieves excellent results on four datasets on both evaluation metrics except for the Simulated-5. For example, KTA exceeds DKT+ 10% more on ASSIST2015 regards to AUC. Similar situations happened to the F1 score, and our model achieves notable improvement compared with other models. Moreover, we notice that on Simulated-5 dataset, the performance of our model is not very impressive. One reason is that there

Table 1. AUC result and F1 score for all datasets tested.

Model	BKT		PFA		DKT		DKT+		KTA	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
ASSIST2009	0.712	0.789	0.658	0.795	0.821	0.834	0.823	0.836	**0.833**	**0.841**
ASSIST2015	0.575	0.828	0.506	0.829	0.736	0.832	0.737	0.830	**0.811**	**0.840**
Statics2011	0.658	0.871	0.521	0.868	0.816	0.886	0.835	0.887	**0.841**	**0.909**
Simulated-5	0.599	0.753	0.522	0.752	0.825	0.794	**0.826**	**0.796**	0.654	0.732

is no long sequence in the dataset. Therefore, our model can not exploit the advantage of capturing the long sequence. Another reason is that all the data have the same length of questions, and every question appears only once. Thus the dependence between data is not as strong as other data.

Prediction Visualization. We also provide prediction visualization, as shown in Fig. 3, in order to give a better sense of the self-attention effect on the prediction results. The figure aims to display the change in the prediction of skill along with the number of questions, e.g., s33. Concretely, our model performs more smoothly compared with DKT.

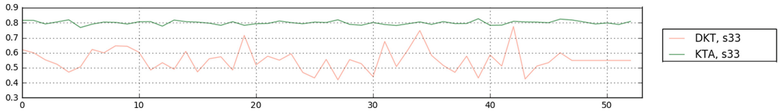

Fig. 3. Line plot for the skill 33 prediction of three models. The student interactions are extracted from ASSISTments 2009. Probability of correctly answering skill 33 is predicted by the trained models.

4 Conclusion

In this paper, we applied the FSA to interpret DKT and through the analysis of FSA, we discover that DKT can not handle the long sequence input. Therefore, we introduce a self-attention model, namely, KTA, which can directly capture the global dependency relationships by computing the similarity among each item of the input regardless of the length of the input sequence. The experimental results show that our proposed model can provide better predictions than existing models.

References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction pp. 253–278 (1995)
2. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

3. Hou, B.J., Zhou, Z.H.: Learning with interpretable structure from RNN. arXiv:1810.10708 (2018)
4. Pavlik, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis - a new alternative to knowledge tracing. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, pp. 531–538 (2009)
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Dickstein, J.S.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
6. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)
7. Tang, G., Müller, M., Rios, A., Sennrich, R.: Why self-attention? a targeted evaluation of neural machine translation architectures. arXiv preprint arXiv:1808.08946 (2018)
8. Vaswani, A., et al.: Attention is all you need pp. 5998–6008 (2017)
9. Yeung, C., Yeung, D.Y.: Addressing two problems in deep knowledge tracing via prediction-consistent regularization. arXiv:1806.02180 (2018)