Exercise Hierarchical Feature Enhanced Knowledge Tracing

Hanshuang Tong[✉], Yun Zhou, and Zhen Wang

AIXUEXI Education Group Ltd., AI Lab, Beijing, China
`tonghanshuang2018@163.com`, `zhouyun.nudt@gmail.com`, `WangZhen__00@163.com`

Abstract. Knowledge tracing is a fundamental task in the computer-aid educational system. In this paper, we propose a hierarchical exercise feature enhanced knowledge tracing framework, which could enhance the ability of knowledge tracing by incorporating knowledge distribution, semantic features, and difficulty features from exercise text. Extensive experiments show the high performance of our framework.

Keywords: Knowledge tracing · Intelligent education · Deep learning

1 Introduction

Knowledge tracing is an essential and classical problem in intelligent education systems. By tracing the knowledge transition process, we could recommend specific educational items to a student based on one's weak knowledge. Existing methods try to solve knowledge tracing problems from both educational psychology and data mining perspectives, such as Item Response Theory (IRT) [7], Bayesian Knowledge Tracing (BKT) [1], Performance Factors Analysis (PFA) framework [9] and Deep knowledge tracing (DKT) [10]. Those models have been proved effective but still have limitations. They do not systematically consider the impact of different attributes of the exercises itself on the knowledge tracing problem. Exercise Enhanced Knowledge Tracing (EKT) [5] is the first method to take exercise text and attention mechanism into consideration. However, EKT extracts features of text by feeding the text of exercise directly into a neural network, which fails to extract hierarchical features from exercise (Fig. 1).
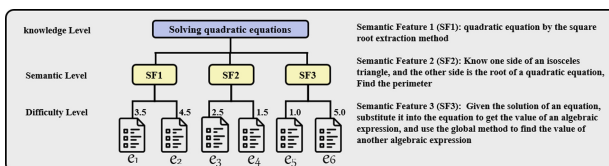
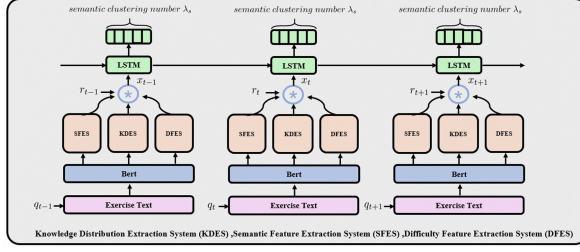**Fig. 1.** The illustration of hierarchical features of exercise

Fig. 2. Exercise hierarchical feature enhanced framework.

2 Exercise Hierarchical Feature Enhanced Framework

Framework Overview. Knowledge tracing task can be summarized as: In an online educational system, suppose we have M students and E exercises in total. Given any learners' exercise record $E = \{(q_1, r_1), (q_2, r_2) \ldots (q_m, r_m)\}$, predict one's performance on q_{t+1}. Here (q_t, r_t) represents that a learner practices question q_t and answers r_t at step t. The entire structure of the framework is shown in Fig. 2. In order to dig deeper into the information in the exercise text, first we utilize Bert [2] to generate embedding vector v_b. Then we feed them into three systems to generate knowledge distribution $v_t \in R^K$, semantic features s_t and question difficulty d_t separately. Let $\varphi(s_t)$ be the one-hot encoding of the semantic cluster where the question belongs at time t. Finally, we concatenate v_t, $\varphi(s_t)$, d_t, and r_t as x_t and feed x_t into a sequence model.

Subsystems Introduction. Two text classification systems, named KDES and DFES, are designed to predict the knowledge distribution and difficulty of the exercise respectively. The semantic feature extractor system (SFES) could be considered as an unsuperviesed clusering problems. The input of those systems is the Bert encoding of the exercise text. The knowledge labeled by teacher and the correct rate of a question [4] serve as ground truth and are predicted using TextCNN [8] in KDES and DFES systems. In KDES system, we use softmax results classified in the trained model to represent the knowledge distribution of an exercise. In DFES systems, we use neural networks to predict difficulty in order to solve the cold start problem. In SFES systems, we cluster the input using a Hierarchical Clustering method by calculating the cos distance between different semantic vectors [6].

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1}; \theta_t) \tag{1}$$

$$y_t = \sigma(W_{yh} \cdot h_t + b_y) \tag{2}$$

$$loss = -\sum_t (r_{t+1} * log(y_i^T \cdot \varphi(s_{t+1})) + (1 - r_{t+1}) * log(1 - y_i^T \cdot \varphi(s_{t+1}))) \tag{3}$$

Modeling Process. In the propagation stage, as shown in Eq. 1, we process x_t and the previous learner's hidden state h_{t-1} and then use RNN network to get current learner's hidden state h_t. Here we use LSTM as a variant of RNN since it can better preserve long-term dependency in the exercise sequence [3]. Finally, we use h_t to predict y_t which contains information about students' mastery of each semantic feature. Additionally, the dimension of y_t is same as the total number of different semantic clustering in DFES system. The θ_t, W_{yh}, b_y in the equation are the parameters of models. The goal of training is to minimize the negative log likelihood of the observed sequence of student response logs (shown in Eq. 3).

3   Experiment

3.1   Experimental Setting

Since there is no open dataset which could provide exercising records with text information. We derive an experimental dataset containing 132,179 students and 91,449,914 answer records from a large real-world online education system: aix-uexi.com.

The baselines of the experiments are as following: BKT, which is based on Bayesian inference; DKT, which uses recurrent neural networks to model student learning; EKTA, which incoporate exercise text features and attention mechanism into the recurrent neural networks; EHFKT_K/S/D, a simplified version of EHFKT, which only contains KDES/SFES/DFES system. The input of EHFKT series is the concatenation of problem encoding and the ouput of each system; EHFKT_T, which contains all subsystems. It diagnoses the transition of mastery of knowledge, while EHFKT diagnoses transition of the mastery of semantic features.

3.2   Experimental Results

Hierarchical Clustering Result. The SFES system uses Bert and Hierarchical Clustering to obtain semantic features of questions. Figure 3 shows the visualization of the clustering results of 11410 questions. The y-axis corresponds to the classification threshold and x-axis corresponds to each exercise. Table 1 implies the result of clustering when the number of clustering λ_s is 912.

Table 1. The result of clustering

Id	Question content	Knowledge	Cluster
Q35	Calculate factorization of $9a^2\left(2x - 3y\right) + 4b^2\left(3y - 2x\right)$	Factorization	SF3
Q37	Calculate $\left(2a^3 + a^2\right) \div a^2$, which of the following is true?	Factorization	SF3
Q38	If $\left(x^2 + y^2 + 2\right)\left(x^2 + y^2 - 2\right) = 0$, calculate $x^2 + y^2$	Factorization	SF4
Q36	Given two points on $y = -mx^2 + 2x$, calculate m	Factorization	SF5
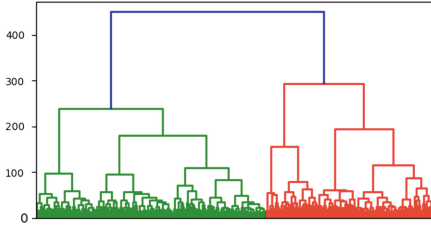
Fig. 3. Hierarchical clustering result
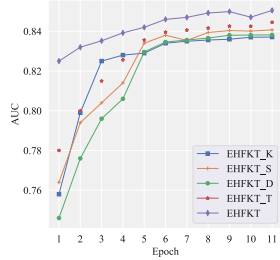
Fig. 4. AUC of EHFKT series

EHFKT Result. In this part, our experiment divides the dataset into a training set with 105,744 learners' logs and a test dataset with 26,435 learners' logs. Figure 4 shows the transition of AUC during the training process. Table 2 shows the overall comparing results in this task. The results indicate that EHFKT performs better than other baseline models. Thus, we could draw several conclusions from the result: In the knowledge tracing task, adding hierarchical features can better represent questions; Besides, tracing the mastery of semantic clusterings can predict students' performance more precisely. The reason is that the exercises contained in the same clusters have similar knowledge distribution, difficulty, and semantics; This result also demonstrates the instability of the tracing of knowledge mastery since the difficulty of an exercise is unpredictable.

Table 2. Evaluation metrics of different deep learning methods

Model	AUC	Model	AUC
BKT	0.6325 ± 0.0011	DKT	0.8324 ± 0.0031
EKTA	0.8384 ± 0.0036	EHFKT_S	0.8407 ± 0.0016
EHFKT_K	0.8371 ± 0.0022	EHFKT_ D	0.8382 ± 0.0035
EHFKT_T	0.8445 ± 0.0025	EHFKT	$\mathbf{0.8505 \pm 0.0021}$

4   Conclusions

In this article, we propose a novel knowledge tracing framework which could extract the knowledge distribution, semantic features and difficulty from the exercise. Besides, We introduce the diagnosis of semantic features of questions into knowledge tracing, which leads to more accurate performance prediction. Although the meaning of these semantic clusters is beyond people's understanding, in the future we will try extracting the meaning of the exercises in the same cluster by text sumarization technique to make the data-driven clusters result more understandable to human.

References

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: modeling the acquisition of procedural knowledge. User Model. User-Adap. Inter. **4**(4), 253–278 (1994)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
4. Hontangas, P., Ponsoda, V., Olea, J., Wise, S.L.: The choice of item difficulty in self-adapted testing. Eur. J. Psychol. Assess. **16**(1), 3 (2000)
5. Huang, Z., et al.: Ekt:Exercise-aware knowledge tracing for student performance prediction. IEEE Trans. Knowl. Data Eng. (2019)
6. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
7. Khajah, M.M., Huang, Y., González-Brenes, J.P., Mozer, M.C., Brusilovsky, P.: Integrating knowledge tracing and item response theory: a tale of two frameworks. In: CEUR Workshop Proceedings, vol. 1181, pp. 7–15. University of Pittsburgh (2014)
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Pavlik Jr., P.I., Cen, H., Koedinger, K.R.: Performance factors analysis-a new alternative to knowledge tracing. Online Submission (2009)
10. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)