Which to Choose? An Order-aware Cognitive Diagnosis Model for Predicting the Multiple-choice Answer of Students

Shuanghong Shen[1], Qi Liu[1], Enhong Chen[1], Shiwei Tong[1], Zhengya Huang[1], Wei Tong[1], Yu Su[1,2], and Shijin Wang[2]

[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology & School of Data Science,University of Science and Technology of China
{closer, tongsw, tongustc}@mail.ustc.edu.cn; {qiliuql, cheneh, huangzhy}@ustc.edu.cn
[2]iFLYTEK Research, iFLYTEK CO., LTD.
{yusu,sjwang3}@iflytek.com

Abstract

This paper presents our solution to the task 2 of the NeurIPS 2020 Education Challenge, which aims to predict the multiple-choice answer students choose for each question. We propose a novel Order-aware Cognitive Diagnosis model, shortly named OCD, to accomplish this task. Specifically, considering the attention span of students are limited during the whole quiz, we utilize the convolutional sliding windows in CNNs to model the attention span of students, in which the effects of question order are captured simultaneously. Moreover, the attention spans show certain variabilities between different students or at different learning phases. Therefore, we further present CNNs with different sizes of convolutional sliding windows to handle multi-scale attention spans. Finally, we take advantage of the attention mechanism to learn the similar weights of questions and give out the predictions. Our proposed OCD model has achieved the accuracy of 0.6800 on public evaluation and 0.6803 on private evaluation, which are both the best performances in this task.

1 Introduction

Recent decades have witnessed the fast evolution of educational data mining (EDM), which focuses on applying a wide variety of data mining tools and techniques to any educationally related data. Cognitive diagnosis, with the purpose of assessing students' knowledge state (knowledge proficiency of specific skills), is one of the fundamental and crucial tasks in EDM [1].

Existed works mainly focus on diagnosing the knowledge proficiency of students and predicting the binary values of whether they can answer correctly. In contrast, there is little attention paid to predict their specific choices. Generally, it is harder to realize the later target, because the multiple-choice answer can be determined by the characteristics of students, questions and tests. However, compared to just predicting right or wrong, we can relatively discover potential common misconceptions that students may have through predicting their choices. In task 2 of the NeurIPS 2020 Education Challenge, the organizers provide the training data, which is consisted of millions of records of answers given to questions by students, as well as some metadata about questions, students and answers (such as SubjectId, Gender and QuizId). The questions in the dataset are all multiple-choice, each with 4 potential choices and 1 correct choice, and we had to predict students' responses for a hidden, held-out subset of (StudentId, QuestionId) pairs.

Preprint. Under review.

To predict the multiple-choice answer of students, the effects of question order is critical which can not be ignored [2]. Taking a few minutes to think of a simple situation: after choosing *'B'* continuous for two or more times, would students choose *'B'* without hesitation for the next question? Actually, they tend to choose other options rather than *'B'*, especially when they are not very confident about the correct answer. Besides, the probability that several continuous questions have the same correct options is relatively small within the same quiz. The effects of question order on students' multiple-choice answers have complex micro and macro factors, the above situation is just an intuitive and easy-understanding case of the effects of question order. To this end, we propose a novel Order-aware Cognitive Diagnosis (shortly named OCD) model to predict the multiple-choice answer of students. Our proposed OCD model has achieved the accuracy of 0.6800 on public evaluation and 0.6803 on private evaluation, which are both the best performances in this task. In Section 2, we present the structures of our OCD model. In Section 3, we introduce how we process the data and detail the process of experiments to get the best results. In Section 4, we make a simple summarisation and discuss the potential of OCD.

2 Order-aware Cognitive Diagnosis Model

2.1 Modeling Order-aware Attention Span of Students

In an intelligent tutoring system, supposing there are the set of questions $\mathbf{Q} = \{q_1, q_2, ..., q_j, ..., q_J\}$. The practice records of a student are denoted as $\mathbf{X} = \{(q_1, c_1), (q_2, c_2), ..., (q_L, c_L)\}$, the tuple (q_l, c_l) represents a practice record (i.e., the *question-choice* pair) of the student, where q_l is the question, c_l represents the multiple-choice answer, L is the length of the records. We first represent the question q_l by random initialized vector $\mathbf{q}_l \in \mathbb{R}^{d_q}$, where d_q is the dimensions. The multiple-choice answer c_l is represented by one-hot encoding $\mathbf{c}_l \in \mathbb{R}^{d_c}$, where d_c is the number of choices. Then, we concatenate $\mathbf{q}_l$ and $\mathbf{c}_l$ to represent the record (q_l, c_l) as vector $\mathbf{r}_l = \mathbf{q}_l \oplus \mathbf{c}_l \oplus ... \oplus \mathbf{c}_l$, where $\mathbf{r}_l \in \mathbb{R}^{d_q + \alpha \times d_c}$. Here we repeat $\mathbf{c}_l$ for α times for balancing the length of $\mathbf{q}_l$ and $\mathbf{c}_l$. Finally, we concatenate all the practice records of a student in order of response time and get the practice embedding $\widetilde{\mathbf{s}} = \mathbf{r}_1 \oplus \mathbf{r}_2 \oplus ... \oplus \mathbf{r}_L$, where $\widetilde{\mathbf{s}} \in \mathbb{R}^{L \times (d_q + \alpha \times d_c)}$. Then a multi-layer perceptron (MLP) is applied to transform $\widetilde{\mathbf{s}}$ to the student embedding $\mathbf{s} \in \mathbb{R}^{L \times K_s}$, where K_s is the dimensions.

In our previous work [3], we have successfully applied convolutional sliding windows to capture the distinctions of learning rates in continuous learning interactions among students. A recent work has also indicated that CNNs can learn the absolute position information impliclitly [4]. For modeling the effects of question order within specific attention span of students, inspired by the idea of *learning windows* [3], we utilize convolutional sliding windows to simulate the attention span of students. The convolutional sliding window is represented as $\mathbf{W}_c \in \mathbb{R}^{K_c \times K_s}$, where K_c is the length of convolutional sliding window. After designing the convolutional sliding window, one-dimensional CNNs is utilized to take K_c continuous practice records to produce the new features of students' knowledge proficiency. For keeping the consistency of shapes, we do not set pooling operation and zero-padding is applied. Then we stack β same convolutional layers on top of each other to make up hierarchical convolutional layers and deeper fuse the practice records to output the knowledge proficiency of students. Finally, for each practice record s_l in $\mathbf{s}$, the whole length of its order-aware attention span is $(K_c - 1)\beta + 1$, where the first layer models the most recent K_c practice records and every upper layer covers $K_c - 1$ farther practice records.

2.2 Modeling Multi-scale Attention Spans

Generally speaking, the attention span is not static, but shows certain variabilities between different students or at different learning phases. In section 2.1, we have realised the order-aware cognitive diagnosis model with the attention span of $(K_c - 1)\beta + 1$. Now we are going to expand it to model multi-scale attention spans. Concretely, we take advantage of various sizes of convolutional sliding windows, where each window is corresponding to a specific attention span. Supposing the length of convolutional sliding window is K_{ci}, thus the window will be $\mathbf{W}_{ci} \in \mathbb{R}^{K_{ci} \times K_s}$ and the whole length of the order-aware attention span will be $(K_{ci} - 1)\beta + 1$, where the first layer models the most recent K_{ci} practice records and every upper layer covers $K_{ci} - 1$ farther practice records. We apply different convolutional sliding windows to model multi-scale attention spans and the average peformance on all attention spans will be the final knowledge proficiency of students.

2.3 Making Predictions

After multi-scale attention spans modeling, we transform the student embedding $\mathbf{s}$ to deep features about students' knowledge proficiency $\mathbf{h} \in \mathbb{R}^{L \times K_h}$ of each question in the test. Liu et al. [5] pointed that students may get similar scores on similar questions. Following this idea, we utilize a simple attention mechanism to calculate the similarity between questions for making predictions of the $(L+1)$-th question. Then the student state at $(L+1)$-th practice record is a weighted sum aggregation of all historical features of knowledge proficiency based on the correlations between question q_{L+1} and the historical ones $\{q_1, q_2, ..., q_L\}$. Formally, for predicting multiple-choice answer of question q_{L+1}, we define the attentive state vector h_{att} as:

$$\begin{aligned} h_{att} &= \sum_{l=1}^{L} \gamma_l h_l, \\ \gamma_l &= softmax(a_l), \quad l \in (1, L) \\ a_l &= q_{L+1} \cdot q_l. \qquad\quad l \in (1, L) \end{aligned} \tag{1}$$

Then we concatenate the attentive state vector h_{att} and question q_{L+1} and predict the student's choice on question q_{L+1} as following formulas:

$$\begin{aligned} \mathbf{I} &= f(\mathbf{W}_1 \cdot [q_{L+1} \oplus h_{att}] + \mathbf{b}_1), \\ \mathbf{p} &= f(\mathbf{W}_2 \cdot \mathbf{I} + \mathbf{b}_2), \end{aligned} \tag{2}$$

where $\oplus$ is the operation of concatenating, $\mathbf{W}_1 \in \mathbb{R}^{K_h \times K_i}$ and $\mathbf{W}_2 \in \mathbb{R}^{K_i \times d_c}$ are the weight matrices, $\mathbf{b}_1 \in \mathbb{R}^{K_i}$ and $\mathbf{b}_2 \in \mathbb{R}^{d_c}$ are the bias terms and f is the non-linear function. $\mathbf{p}$ indicates the probability that each option will be selected by the student and the option with max probability will be the result of prediction.

3 Experiments

3.1 Dataset

The training dataset of task 2 is consisted of 27613 questions, 118971 students and 15867850 practice records. Each question has four answer options, we have to predict students' choices for a hidden, held-out subset of practice records. We first cluster the practice records of each student based on the *quizid* of questions. Within each quiz, we rank the practice records in order of the response time. There are total 2004029 records pf quizzes for all students, we depict the length of these quizzes in Figure 1. Moreover, we have counted that over half of quizzes have a length of 10, and over 98% of quizzes have a length of less than 20. Therefore, we designed four ways to split the dataset based on the *quizid* as follows:

- *quiz-10*: in each quiz, we set the length of stundents' practices as 10.
- *quiz-20*: in each quiz, we set the length of stundents' practices as 20
- *quiz-sorted-10*: for every student, we first rank all the answered quizzes from long to short, and other settings are the same as *quiz-10*.
- *quiz-7days-10*: for all quizzes with more than 10 questions, if the interval between two adjacent practice records is more than seven days, we see them as a split point. Then we rank all the split quizzes from long to short, and other settings are the same as *quiz-10*.

3.2 Experimental results

In all experiments, the learning rate was set to be 0.001 and divided by 5 every 3 training epochs. d_q, α, K_s, K_h, K_i, β were 388, 10, 256, 256, 128, 3 respectively and there were three sizes of convolutional windows: 2, 3, 4. The 10-fold cross validation was used to train and evaluate our model. All the hyper-parameters are randomly initialized with uniform distribution, optimized with mini-batch Adam on the training set.

For each split of the dataset, we conducted 10-fold cross validation and the experimental results are shown in Table 2. As we can see from the table, *quiz-7days-10* gets the best performance, because it avoids the interaction between quizzes as much as possible. The performance on *quiz-20* shows a significant drop because the majority of quizzes have combined with other quizzes. For each fold, we

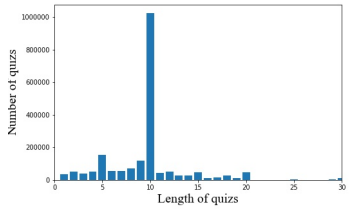

Figure 1: The length distribution of all quizzes.

Table 1: Ablation study

Models	accuracy
NeuralCD [6]	0.6580
OCD-shuffle	0.6656
OCD-1-attention span	0.6726
OCD	0.6736

Table 2: The 10-fold experimental results on different ways of dividing data.

Datasets	k-fold1	k-fold2	k-fold3	k-fold4	k-fold5	k-fold6	k-fold7	k-fold8	k-fold9	k-fold10
quiz-10	0.6698	0.6696	0.6713	0.6704	0.6713	0.6689	0.6701	0.6734	0.6727	0.6722
quiz-20	0.6599	0.6612	0.6623	0.6626	0.6617	0.6602	0.6601	0.6655	0.6641	0.6647
quiz-sorted-10	0.6710	0.6718	0.6725	0.6716	0.6723	0.6703	0.6710	0.6747	0.6735	0.6736
quiz-7days-10	0.6712	0.6717	0.6725	0.6729	0.6726	0.6707	0.6717	0.6756	0.6741	0.6745

can get the corresponding prediction results on the testing data, then we take the averaged predictions as the submission. For *quiz-7days-10*, we get the accuracy of 0.6777 on public evaluation. The best performance of 0.6800 on public evaluation and 0.6803 on public evaluation is achieved by averaging predictions on all four ways of dividing data.

Apart from this, we conduct some simple ablation study on *quiz-10*, the results are shown in Table 1. NeuralCD is a representive cognitive diagnosis model with state-of-art performance [6] and we choose it as the baseline model. OCD-shuffle is a variant of OCD without ranking the practices in order of response time and OCD-1-attention span is also a variant of OCD with only one kind of convolutional sliding window. The experimental results indicate the superior of our proposed order-aware cognitive diagnosis model. Besides, through model the effects of question order, OCD can well improve the performance of cognitive diagnosis.

4 Discussions

In this paper, we proposed the order-aware cognitive diagnosis to model the effects of question order in the cognitive process of students. Through utilizing convolutional sliding windows to simulate the attention span of students and designing different sizes of windows to handle multi-scale attention spans, our proposed OCD model achieved better diagnosis results and got the best performance on the task 2 of the NeurIPS 2020 Education Challenge (actually, OCD also performed well on task 1). These observations indicate that it is valuable to model the effects of question order in cognitive process and it is possible to capture certain student-specific variabilities under the idea of *learning windows*. Moreover, quite a few practices in the same quiz have same label of response time, which can not be ranked. Therefore, we believe the performance of OCD can have a considerable improvement if providing these missing information.

References

[1] Qi Liu, Runze Wu, Enhong Chen, Guandong Xu, Yu Su, Zhigang Chen, and Guoping Hu. Fuzzy cognitive diagnosis for modelling examinee performance. *TIST*, 9(4):1–26, 2018.

[2] Neal M. Kingston and Neil J. Dorans. Item location effects and their implications for irt equating and adaptive testing. *Applied Psychological Measurement*, 8(2):147–154, 1984.

[3] Shuanghong Shen, Qi Liu, Enhong Chen, Han Wu, Zhenya Huang, Weihao Zhao, Yu Su, Haiping Ma, and Shijin Wang. Convolutional knowledge tracing: Modeling individualization in student learning process. In *ACM SIGIR*, page 1857–1860, 2020. ISBN 9781450380164. doi: 10.1145/3397271.3401288. URL `https://doi.org/10.1145/3397271.3401288`.

[4] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[5] Qi Liu, Huang Zhenya, Yin Yu, Chen Enhong, Xiong Hui, Su Yu, and Hu Guoping. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE)*, pages 1–1, 2019.

[6] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *AAAI 2020*, 2020.