Exploring Student Representation For Neural Cognitive Diagnosis

Hengyao Bao
Tencent Inc.
Chengdu, China
henniebao@tencent.com

Xihua Li
Tencent Inc.
Chengdu, China
lixihua9@126.com

Xuemin Zhao
Tencent Inc.
Chengdu, China
xueminzhao@tencent.com

Yunbo Cao
Tencent Inc.
Beijing, China
yunbocao@tencent.com

Abstract

Cognitive diagnosis, the goal of which is to obtain the proficiency level of students on specific knowledge concepts, is an fundamental task in smart educational systems. Previous works usually represent each student as a trainable knowledge proficiency vector, which cannot capture the relations of concepts and the basic profile(e.g. memory or comprehension) of students. In this paper, we propose a method of student representation with the exploration of the hierarchical relations of knowledge concepts and student embedding. Specifically, since the proficiency on parent knowledge concepts reflects the correlation between knowledge concepts, we get the first knowledge proficiency with a parent-child concepts projection layer. In addition, a low-dimension dense vector is adopted as the embedding of each student, and obtain the second knowledge proficiency with a full connection layer. Then, we combine the two proficiency vector above to get the final representation of students. Experiments show the effectiveness of proposed representation method.
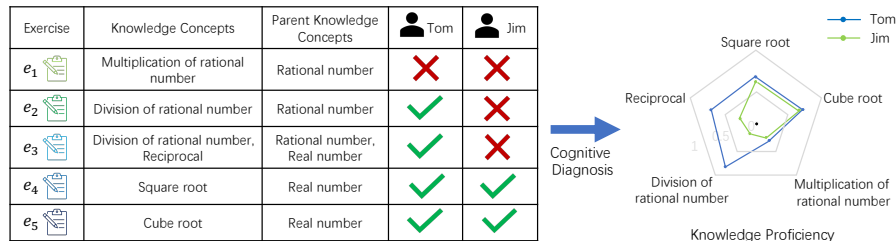
1 Introduction

Figure 1: An example of cognitive diagnosis

Cognitive diagnosis is an essential and fundamental technology in smart education systems, in which cognitive diagnosis can help to obtain the proper profiles of students and assist lots of education services, such as student learning report and adaptive exercise recommendation [1, 2]. Figure 1 shows an example of cognitive diagnosis process. Generally, with the exercise responses of students and

labels of exercises, cognitive diagnosis is to infer their relative abilities [3], such as proficiency on specific knowledge concepts (e.g. *multiplication of rational number*) [8].

Many classical methods have been developed to address this issue, such as Multidimensional Item Response Theory (MIRT) [4], Deterministic Inputs, Noisy And gate model (DINA) [5], Matrix Factorization (MF) [6], and Item response ranking framework [7]. Recently, deep neural network has also been applied in cognitive diagnosis. [8] proposed a Neural Cognitive Diagnosis framework (NCD) which utilizes a knowledge proficiency vector to represent student and formulates the students, exercises and responses with an MIRT-like multi-layer perceptron. In [9] an Educational context-aware Cognitive Diagnosis framework (ECD) was developed to model the context of student, e.g. highest education degree of parents and duration in early childhood education. However, since only a knowledge proficiency vector is used to represent the student, these methods are short of characterizing the complete profile of student, such as the comprehensive ability of student, or the average mastery on associated knowledge concepts. For Example, as shown in Figure 1, although Tom and Jim both answer e_1 incorrectly, it's still not suitable for the cognitive diagnosis system to give similar poor scores on concept *multiplication of rational number*, since Tom is obviously a good student and has a better performance in the domain of *rational number*.

In this paper, we develop a structure to enrich the representation of students by making use of the hierarchical relations of knowledge concepts and the embedding of students. First, the proficiency on parent knowledge concepts is used to represent the knowledge related profile of students, and the proficiency on child concepts is obtained with a parent-child concepts projection layer. Second, for a further exploration on the representation of students, we adopt a low-dimension dense vector as the embedding of each student, and obtain the second knowledge proficiency with a full connection layer. Then, we average the two proficiency above to get the final representation of students, and formulate the students, exercises and responses as a neural diagnosis network. Experiments show that it has a substantial improvement in terms of both response prediction and knowledge proficiency diagnosis.
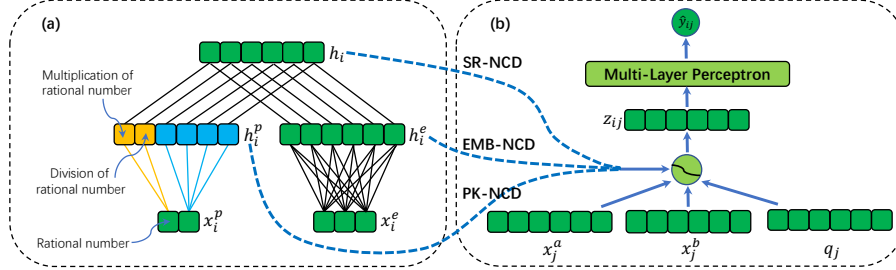
2 Model

Figure 2: The structure of proposed method: (a) student representation layer, (b) answer correctness prediction layer

2.1 Problem Definition

Suppose in a smart education system there are N students and M exercises, and define the responses of students as $R = \{\langle s_i, e_j, y_{ij}\rangle | 1 \leq i \leq N, 1 \leq j \leq M, y_{ij} \in \{0,1\}\}$, where s_i, e_j and y_{ij} denote the i-th student, j-th exercise and the relative response of student i on exercise j respectively. In addition, we define the Q-matrix (usually labelled by experts) as $\boldsymbol{Q} = \{q_{kj}\}_{K \times M}$, in which $q_{kj} \in \{0,1\}$ denote whether exercise e_j relates to the k-th knowledge concept, and K is the number of concepts. Then, given the responses of students R and the Q-matrix $\boldsymbol{Q}$, the goal of cognitive diagnosis is to estimate the knowledge proficiency of each student.

2.2 Student Representation

The proposed method of student representation is illustrated in Figure 2(a). We first notice that, the knowledge concepts to be diagnosed have related parent knowledge concepts that are labeled by experts in advance. As the case shown in Figure 1, *rational number* is the parent knowledge concept

of both *multiplication of rational number* and *division of rational number*. Generally, the proficiency in parent knowledge concepts can somehow indicate students' knowledge related profile and mastery in child knowledge concepts.[1] Therefore, the parent-child relations of knowledge concepts can be used to enrich the representation of students. Suppose the K knowledge concepts has L parent concepts, and we use a trainable vector $\boldsymbol{x}_i^p \in \mathbb{R}^{L \times 1}$ to represent the proficiency of i-th student in each parent concepts. Then the knowledge proficiency in child concepts $\boldsymbol{h}_i^p$ can be obtained by:

$$\boldsymbol{h}_i^p = \sigma(\boldsymbol{G}\boldsymbol{x}_e^p + \boldsymbol{b}^p) \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\boldsymbol{b}^p \in \mathbb{R}^{K \times 1}$ is a bias vector, and $\boldsymbol{G} = \{g_{kl}\}_{K \times L}$ is a parent-child map matrix in which g_{kl} is a trainable variable if the k-th child concept is descendent of the l-th parent concept, and $g_{kl} \equiv 0$ otherwise.

In the other hand, $\boldsymbol{h}_i^p$ only utilizes the relations of same knowledge concept family, and cannot fit the proficiency of different concept families, or the memory and comprehensive ability of students. Hence, we use a low-dimension dense vector $\boldsymbol{x}_i^e \in \mathbb{R}^{D \times 1}$ as the embedding of student s_i, and get the knowledge proficiency $\boldsymbol{h}_i^e$ by:

$$\boldsymbol{h}_i^e = \sigma(\boldsymbol{F}\boldsymbol{x}_i^e + \boldsymbol{b}^e) \tag{2}$$

where $\boldsymbol{b}^e \in \mathbb{R}^{K \times 1}$ is a bias vector, and $\boldsymbol{F} \in \mathbb{R}^{K \times D}$ is a projection matrix to knowledge concepts.

Then, we can get the final knowledge proficiency $\boldsymbol{h}_i$ by simply calculating the mean:

$$\boldsymbol{h}_i = (\boldsymbol{h}_i^p + \boldsymbol{h}_i^e) / 2 \tag{3}$$

Note that, like the work in [9], one also get $\boldsymbol{h}_i$ by a weighted sum of $\boldsymbol{h}_i^p$ and $\boldsymbol{h}_i^e$, in which the weight is also trainable. However, we do not see it has a substantial improve in our dataset.

2.3 Answer Correctness Prediction

With the knowledge proficiency obtained above, the task of cognitive diagnosis can be formulated as an answer correctness prediction problem [8]. The structure of prediction layer is shown in Figure 2(b). Specifically, for the exercise e_j, we define $\boldsymbol{q}_j$ as the j-th column of Q-matrix $\boldsymbol{Q}$, $\boldsymbol{x}_j^a \in \mathbb{R}^{K \times 1}$ and $\boldsymbol{x}_j^b \in \mathbb{R}^{K \times 1}$ as the discrimination and difficulty embedding respectively, and the prediction of answer correctness y_{ij} is obtained by:

$$\boldsymbol{\alpha}_j = \sigma(\boldsymbol{x}_j^a) \tag{4}$$

$$\boldsymbol{\beta}_j = \sigma(\boldsymbol{x}_j^b) \tag{5}$$

$$\boldsymbol{z}_{ij} = \boldsymbol{q}_j \cdot \boldsymbol{\alpha}_j \cdot (\boldsymbol{\beta}_j - \boldsymbol{h}_i) \tag{6}$$

$$\hat{y}_{ij} = MLP(\boldsymbol{z}_{ij}) \tag{7}$$

where $MLP(\cdot)$ denotes a multi-layer perceptron, and $\hat{y}_{ij}$ is the prediction result. Thus the cross entropy loss for student s_i on exercise e_j is defined as:

$$loss_{ij} = y_{ij}\log \hat{y}_{ij} + (1 - y_{ij})\log (1 - \hat{y}_{ij}) \tag{8}$$

In addition, to satisfy the *monotonicity assumption* [4] to ensure good performance and interpretability, we restrict $\boldsymbol{G}$ in (1) and weight matrix of the multi-layer perceptron (7) to be positive when training [8]. Thus, the higher each entry of $\boldsymbol{h}_i$ or $\boldsymbol{x}_i^p$ is, the more likely the student answers the exercise correctly. Also note that both the knowledge proficiency $\boldsymbol{h}_i^p$ and $\boldsymbol{h}_i^e$ can also be passed to (6) independently for prediction (for reason of same dimentions and same representation abilities). As shown in Figure 2(b), for simplification, we denotes the method using the parent knowledge $\boldsymbol{h}_i^p$ as PK-NCD(*P*arent *K*nowledge), method using the student embedding $\boldsymbol{h}_i^e$ as EMB-NCD(*EMB*edding), and method using the student representation $\boldsymbol{h}_i$ as SR-NCD(*S*tudent *R*epresentation) respectively.

3 Experiments

3.1 Datasets, Metrics and Setups

We test the cognitive diagnosis models with two datasets of real-world education scenarios, i.e. ASSIST [10] and XCLASS-MATH. See the datasets details in Appendix A. Besides, since there are

---

[1]Without specification, the term *knowledge concepts* or *child knowledge concepts* denotes the leaf nodes of concept tree, and *parent knowledge concepts* denotes the parent nodes of leaf nodes.

no ground-truth values for the knowledge proficiency of students, it is difficult to evaluate the models straightforwardly. Following the work in [7], we evaluate the performance of models from two perspectives. First, we use Accuracy (ACC) and Area Under the Curve (AUC) to test the classification abilities of models. Second, we adopt Degree Of Agreement (DOA) to assess the monotonicity of models. See the definition of DOA in Appendix B.

We evaluate the proposed EMB-NCD, PK-NCD and SR-NCD defined in Section 2.3 in the experiments. Since ASSIST does not have information about parent knowledge concepts, only EMB-NCD is tested in its experiments. Beside, we adopt two hidden layers in the MLP (7), and set the dimension as 512, 256 for ASSIST, and 128, 64 for XCLASS-MATH respectively, and the dimension of student embedding D is set to $K/4$. We also compare the performance of proposed methods with several previous works: DINA[5], MIRT[4], NCD[8].

3.2 Resutls

Table 1: Experimental results

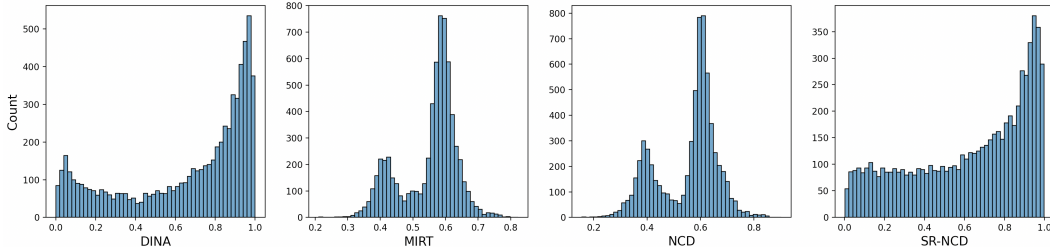| Model	ASSIST			XCLASS-MATH		
	ACC	AUC	DOA	ACC	AUC	DOA
DINA	0.682	0.727	0.603	0.670	0.712	0.629
MIRT	0.724	0.733	0.601	0.746	0.754	0.632
NCD	0.726	0.757	0.609	0.745	0.763	0.635
EMB-NCD	**0.735**	**0.771**	**0.681**	0.748	0.768	0.658
PK-NCD	-	-	-	0.753	0.768	0.656
SR-NCD	-	-	-	**0.757**	**0.780**	**0.664**

Figure 3: Distribution histogram of knowledge proficiency

The experimental results are shown in Table 1. One can observe that the proposed methods outperform all the other baselines on both datasets. Specifically, even though simply adding a student embedding layer, EMB-NCD can obtain a significant improvement compared with the original NCD. Furthermore, PK-NCD and EMB-NCD have similar performance, and by combining them we can acquire another obvious gain. Thus, the experimental results demonstrate the effectiveness of the proposed student representation methods.

Meanwhile, we also display the distribution histogram of knowledge proficiency obtained from XCLASS-MATH in Figure 3. It's interesting to notice that the knowledge proficiency of MIRT and NCD have almost the same distribution, since MIRT is a special case of NCD [8]. Besides, the distribution curve of DINA, MIRT and NCD are bimodal. On the contrary, SR-NCD has a convex and much more smooth curve, which indicates the proficiency acquired might be more discriminative.

4 Conclusion

In this paper, we considered the problem of student representation in cognitive diagnosis model. We developed a method of student representation with the exploration of the hierarchical relations of knowledge concepts and student embedding. Experiments demonstrate the effectiveness and interpretability of the proposed methods.

References

[1] Kuh G. D., Kinzie J., Buckley J. A., Bridges B. K., and Hayek J. C. (2011) *Piecing together the student success puzzle: research, propositions, and recommendations: ASHE Higher Education Report*, volume 116. John Wiley & Sons.

[2] Yuhao Zhou, Xihua Li, Yunbo Cao, Xuemin Zhao, Qing Ye, and Jiancheng Lv (2021) LANA: towards personalized deep knowledge tracing through distinguishable interactive sequences. In *Proceedings of the Educational Data Mining*.

[3] Embretson S. E., and Reise S. P. (2013) *Item response theory*. Psychology Press.

[4] Reckase, M. D. (2009) Multidimensional item response theory models. In *Multidimensional Item Response Theory*, pp. 79-112. Springer.

[5] De La Torre, J. (2009) Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics* 34(1):115–130.

[6] Koren Y., Bell R., and Volinsky C. (2009) Matrix factorization techniques for recommender systems. *Computer* 42(8), pp. 30–37.

[7] Shiwei Tong, Qi Liu, Runlong Yu, Wei Huang, Zhenya Huang, Zachary A. Pardos, and Weijie Jiang (2021) Item response ranking for cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1750-1756.

[8] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6153–6161.

[9] Yuqiang Zhou, Qi Liu, Jinze Wu, Fei Wang, Zhenya Huang (2021) Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2420–2428.

[10] Feng M., Heffernan N., and Koedinger K. (2009) Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction* 19(3): 243-266.

A   Datasets

Table 2: Dataset summary

Statistics	ASSIST	XCLASS-MATH
# students	4,163	165
# exercises	17,746	250
# knowledge concepts	123	74
# parent knowledge concepts	/	7
# response logs	324,572	13,574
# average logs per student	77.97	82.27
# average logs per exercise	18.29	54.30

The statistics of the datasets are summarized in Table 2. ASSIST (ASSISTments 2009-2010 "skill builder") is a widely used open dataset collected by the ASSISTments online tutoring systems[2]. XCLASS-MATH is a mathematical dataset collected by the smart education system XCLASS[3], in which teachers assign and correct students' homework online, and students do their homework with an e-ink pad. It mainly contains the mathematical homework logs within two months of the 7-th grade students of an middle school.

B   Degree of Agreement

The Degree of Agreement (DOA) is defined as:

$$DOA(k) = \frac{1}{Z} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(h_{ik}, h_{jk}) \frac{\sum_{l=1}^{M} I_{lk} \wedge J(l, i, j) \wedge \delta(y_{il}, y_{jl})}{\sum_{l=1}^{M} I_{lk} \wedge J(l, i, j) \wedge [y_{il} \neq y_{jl}]} \tag{9}$$

---

[2]https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010

[3]https://xclass.qq.com. XCLASS-MATH will be released later.

where $Z = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(h_{ik}, h_{jk})$, h_{ik} denotes the proficiency of student i on concept k, $\delta(x, y) = 1$ if $x > y$ and $\delta(x, y) = 0$ otherwise, $I_{lk} = 1$ if exercise l contains concept k and $I_{lk} = 0$ otherwise, and $J(l, i, j) = 1$ if both student i and j did exercise l and $J(l, i, j) = 0$ otherwise. The perspective of DOA is that, if student i has a higher proficiency on concept k than student j, then student i is more likely to answer exercise related to concept k correctly than student j. The average of $DOA(k)$ on all concepts is used in our experiments. Thus, the model with a higher DOA score might have a better monotonicity and interpretability.