Interpretable Cognitive Diagnosis with Neural Network

Fei Wang[1], **Qi Liu**[2], **Enhong Chen**[2], **Zhenya Huang**[1]
University of Science and Technology of China
[1]{wf314159,huangzhy}@mail.ustc.edu.cn
[2]{qiliuql,cheneh}@ustc.edu.cn

Abstract

In intelligent education systems, one key issue is to discover students' proficiency level on specific knowledge concepts, which called cognitive diagnosis. Existing approaches usually mine the student exercising process by manually designed function, which is usually linear and not sufficient to capture complex relations between students and exercises. In this paper, we propose a general Neural Cognitive Diagnosis (NeuralCD) framework, which incorporates neural networks to learn the complex interactions between student's and exercise's factor vectors. The interpretability of factor vectors is guaranteed with the monotonicity assumption borrowed from educational psychology. We provide NeuralCDM model as an implementation example of the framework. Further, we explore the text content for improving NeuralCDM to show the extendability of NeuralCD, and demonstrate the generality of NeuralCD by proving how it covers some traditional diagnostic models. Extensive experimental results on real-world datasets show the effectiveness of NeuralCD framework with both accuracy and interpretability.

1   Introduction

Cognitive diagnosis is a necessary and fundamental task in many real-world scenarios such as games [1], medical diagnosis [2], and education. Specifically, in intelligent education systems [3, 4], cognitive diagnosis aims to discover the states of students in the learning process, such as their proficiency on specific knowledge concepts [5]. Figure 1 shows a toy example of cognitive diagnosis. Generally, students usually first choose to practice a set of exercises (e.g., $e_1, \cdots, e_4$) and leave their responses (e.g., right or wrong). Then, our goal is to infer their actual knowledge states on the corresponding concepts (e.g., *Equation*). In practice, these diagnostic reports are necessary as they are the basis of further services, such as exercise recommendation and targeted training [6].
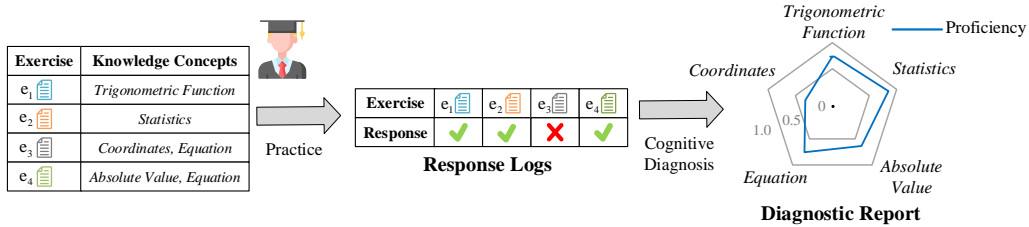

Figure 1: A toy example of cognitive diagnosis.

In the literature, massive efforts have been devoted for cognitive diagnosis, such as Deterministic Inputs, Noisy-And gate model (DINA) [7], Item Response Theory (IRT) [8], Multidimensional IRT (MIRT) [9] and Matrix Factorization (MF) [10]. Despite achieving some effectiveness, these works rely on handcrafted interaction functions that just combine the multiplication of student's and exercise's trait features linearly, such as logistic function [8] or inner product [10], which may not be sufficient for capturing the complex relationship between students and exercises [11]. Besides, the

design of specific interaction functions is also labor-intensive since it usually requires professional expertise. Therefore, it is urgent to find an automatic way to learn the complex interactions for cognitive diagnosis instead of manually designing them.

In this paper, we address this issue in a principled way of proposing a Neural Cognitive Diagnosis (NeuralCD) framework by incorporating neural networks to model complex non-linear interactions. Although the capability of neural networks to approximate continuous functions has been proved in many domains, such as natural language processing [12] and recommender systems [13], it is still highly nontrivial to adapt to cognitive diagnosis due to the following domain challenges. First, the black-box nature of neural networks makes them difficult to get explainable diagnosis results. That is to say, it is difficult to explicitly realize how much a student has mastered a certain knowledge concept (e.g., *Coordinates*). Second, due to functional restriction, it is hard for traditional non-neural models to leverage exercise text content. However, with neural network, it is worthy of finding ways to explore the rich information contained in exercise text content for cognitive diagnosis.

To address these challenges, we propose a NeuralCD framework to approximate interactions between students and exercises, yet preserving the explainability. We first project students and exercises to factor vectors and leverage multi-layers for modeling the complex interactions of student answering exercises. To ensure the interpretability of both factors, we apply the monotonicity assumption taking from educational property [9] on the multi-layers. Then, we propose two implementations on the basis of the general framework, i.e., NeuralCDM and NeuralCDM+. In NeuralCDM, we simply extract exercise factor vectors from traditional Q-matrix (an example is shown in figure 6) and achieve the monotonicity property with positive full connection layers, which shows feasibility of the framework. While in NeuralCDM+, we demonstrate how information from exercise text can be explored with neural network to extend the framework. We further show that our NeuralCD is a general framework that covers many traditional models such as MF, IRT and MIRT. Finally, we conduct extensive experiments on real-world datasets, and the results show the effectiveness of NeuralCD framework with both accuracy and interpretability guarantee.

2   Related Work

In this section, we briefly review the related works as follows.

Cognitive Diagnosis.   Existing works about student cognitive diagnosis mainly came from educational psychology area. DINA [7, 14] and IRT [8] were two of the most typical models among those works, in which each student and exercise was represented with trait features ($\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively). Specifically, in DINA, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ were binary, where $\boldsymbol{\beta}$ came directly from Q-matrix (a human labeled exercise-knowledge correlation matrix). The probability of student i correctly answering exercise j was modeled as $P(r_{ij} = 1|\boldsymbol{\theta}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}$, where $\eta_{ij} = \prod_k \theta_{ik}^{\beta_{jk}}$, g_j and s_j were guessing and slipping parameters of exercise j respectively. On the other hand, in IRT, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ were unidimensional and continuous latent traits, indicating student ability and exercise difficulty. The interaction between the trait features was modeled in a logistic way, e.g., a simple version is $sigmoid(a(\boldsymbol{\theta} - \boldsymbol{\beta}))$, where a is the exercise discrimination parameter. Although extra parameters were added in IRT [15, 16] and latent trait was extended to multidimensional(MIRT) [17, 9], most of their item response functions were still logistic-like. These traditional models depended on manually designed functions, which was labor-intensive and restricted their scope of applications.

Matrix Factorization.   Recently, some researches from data mining perspective have demonstrated the feasibility of MF for cognitive diagnosis. Student and exercise correspond to user and item in matrix factorization (MF). For instance, Toscher et al. [18] improved SVD (Singular Value Decomposition) methods to factor the score matrix and get students and exercises' latent trait vectors. Thai-Nghe et al. [19] applied some recommender system techniques including matrix factorization in the educational context, and compared it with traditional regression methods. Besides, Thai-Nghe et al. [20] proposed a multi-relational factorization approach for student modeling in the intelligent tutoring systems. Despite their effectiveness in student performance prediction task (i.e., predict students' scores on exercises with their diagnostic results), the latent trait vectors in MF is not interpretable for cognitive diagnosis.

Artificial Neural Network.   Techniques using artificial neural network have reached state-of-the-art in many areas, e.g., speech recognition [21], text classification [22] and image translation [23]. There are also some educational applications such as question difficulty prediction [24], code education [25] and formula transcribing from image [26]. To the best of our knowledge, deep knowledge tracing (DKT) [27] was the first attempt to model student learning process using recurrent neural network. However, DKT is unsuitable for cognitive diagnosis as its main goal is to predict
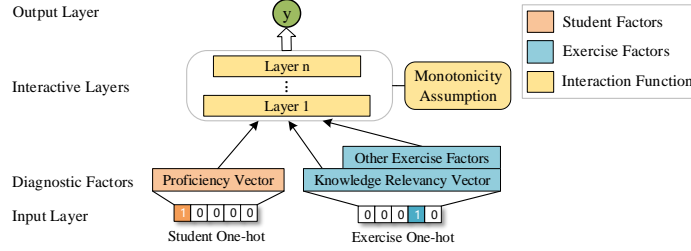
Figure 2: Structure of NeuralCD framework.

students' performance. Neural network performs poorly in parameter interpretation due to its inherent traits. Few works with neural network have high interpretability for student cognitive diagnosis. In this paper, we propose a neural cognitive diagnosis (NeuralCD) framework which borrows concepts from educational psychology and combine them with functions learned from data.

3 Neural Cognitive Diagnosis

We first formally introduce cognitive diagnosis task. Then we describe the details of NeuralCD framework. After that, we design a specific diagnostic network NeuralCDM with traditional Q-matrix to show the feasibility of the framework, and an improved NeuralCDM+ by incorporating exercise text content for better performance. Finally, we demonstrate the generality of NeuralCD framework by showing its close relationship with some traditional models.

3.1 Task Overview

Suppose there are N Students, M Exercises and K Knowledge concepts at a learning system, which can be represented as $S = \{s_1, s_2, \ldots, s_N\}, E = \{e_1, e_2, \ldots, e_M\}$ and $K_n = \{k_1, k_2, \ldots, k_K\}$ respectively. Each student will choose some exercises for practice, and the response logs R are denoted as set of triplet (s, e, r) where $s \in S, e \in E$ and r is the score (transferred to percentage) that student s got on exercise e. In addition, we have Q-matrix (usually labeled by experts) $\mathbf{Q} = \{Q_{ij}\}_{M \times K}$, where $Q_{ij} = 1$ if exercise e_i relates to knowledge concept k_j and $Q_{ij} = 0$ otherwise.

Problem Definition *Given students' response logs R and the Q-matrix $\mathbf{Q}$, the goal of our cognitive diagnosis task is to mine students' proficiency on knowledge concepts through the student performance prediction process.*

3.2 Neural Cognitive Diagnosis Framework

Generally, for a cognitive diagnostic system, there are three elements need to be considered: student factors, exercise factors and the interaction among them [11]. Figure 2 shows the structure of NeuralCD framework. For each response log, we use one-hot vectors of the corresponding student and exercise as input. After obtaining the student's and exercise's diagnostic factors, they are fed into neural interactive layers. The framework outputs the probability that the student correctly answers the exercise, and gets students' proficiency vectors simultaneously. Details are introduced as bellow.

Student Factors Student factors characterize the traits of students, which would affect the students' response to exercises. As our goal is to mine students' proficiency on knowledge concepts, we do not use the latent trait vectors as in IRT and MIRT, which is not explainable enough to guide students' self-assessment. Instead, we adopt the method used in DINA, but in a continuous way. Specifically, We use a vector F^s to characterize a student, and call it *proficiency vector*. Each entry of F^s is continuous, which indicates the student's proficiency on a knowledge concept. For example, $F^s = [0.9, 0.2]$ indicates a high mastery on the first knowledge concept but low mastery on the second. F^s is got through the parameter estimation process.

Exercise Factors Exercise factors denote the factors that characterize the traits of exercises. We divide exercise factors into two categories. The first indicates the relationship between exercises and knowledge concepts, which is fundamental as we need it to make each entry of F^s correspond to a specific knowledge concept for our diagnosis goal. We call it *knowledge relevancy vector* and denote it as F^{kn}. F^{kn} has the same dimension as F^s, with the ith entry indicating the relevancy between the exercise and the knowledge concept k_i. Each entry of F^{kn} is non-negative. F^{kn} is previously given (e.g., obtained from Q-matrix). The second type is optional factors. Factors from IRT and DINA such as knowledge difficulty, exercise difficulty and discrimination can be used if reasonable.

Interaction Function We use artificial neural network to obtain the interaction function for the following reasons. First, the neural network has been proven to be capable of approximating any continuous function [28]. The strong fitting ability of neural network makes it competent for capturing relationships among student and exercise factors. Second, with neural network, the interaction

3

function can be learned from data with few assumptions. This makes NeuralCD more general and can be applied in broad areas. Third, the framework can be highly extendable with neural network. For instance, extra information such as exercise texts can be integrated in with neural network. We formulate the output of NeuralCD framework as:

$$y = \varphi_n(\ldots \varphi_1(F^s, F^{kn}, F^{other}, \theta_f)), \tag{1}$$

where φ_i denotes the mapping function of the ith MLP layer; F^{other} denotes factors other than F^s and F^{kn} (e.g., difficulty); and θ_f denotes model parameters of interactive layers.

However, due to some intrinsic characteristics, neural networks usually have poor performance on interpretation [29]. In order to ensure the interpretation of student and exercise factors, we place a restriction on the diagnostic neural network based on the following monotonicity assumption [9]:

Monotonicity Assumption *The probability of correct response to the exercise is monotonically increasing at any dimension of the student's knowledge proficiency.*

This assumption should be converted as a property of the interaction function. For example, we assume exercise e contains knowledge k, and student s answered it correctly. During training, if the model predicts s to answer e incorrectly (i.e., outputs a value below 0.5), its optimization algorithm should increase the student's proficiency value of k (to raise the output). Monotonicity assumption is used in some IRT and MIRT models. It's general and reasonable in almost all circumstance. Thus it has less influence on the generality of NeuralCD framework.

The goal of NeuralCD framwork is to get students' knowledge proficiency, i.e., the values of F^s.

After introducing the structure of NeuralCD framework, we will next show some specific implementations. We first design a diagnostic model based on NeuralCD with extra exercise factors (i.e., knowledge difficulty and exercise discrimination), and further show its extendability by incorporating text information and generality by demonstrating how it covers traditional models.

3.3 Neural Cognitive Diagnosis Model

Here we introduce a specific neural cognitive diagnosis model (NeuralCDM) under NeuralCD framework. Figure 3 illustrates the structure of NeuralCDM.

Student Factors In NeuralCDM, each student is represented with a knowledge proficiency vector. The student factor F^s aforementioned is $\boldsymbol{h}^s$ here, and $\boldsymbol{h}^s$ is obtained by multiplying the student's one-hot representation vector $\boldsymbol{x}^s$ with a trainable matrix $\mathbf{A}$. That is,

$$\boldsymbol{h}^s = \text{sigmoid}(\boldsymbol{x}^s \times \mathbf{A}), \tag{2}$$

in which $\boldsymbol{h}^s \in (0, 1)^{1 \times K}, \boldsymbol{x}^s \in \{0, 1\}^{1 \times N}, \mathbf{A} \in \mathbb{R}^{N \times K}$.

Exercise Factors As for each exercise, the aforementioned exercise factor F^{kn} is $\boldsymbol{Q}_e$ here, which directly comes from the pre-given Q-matrix:

$$\boldsymbol{Q}_e = \boldsymbol{x}^e \times \mathbf{Q}, \tag{3}$$

where $\boldsymbol{Q}_e \in \{0, 1\}^{1 \times K}, \boldsymbol{x}^e \in \{0, 1\}^{1 \times M}$ is the one-hot representation of the exercise. In order to make a more precise diagnosis, we adopt other two exercise factors: knowledge difficulty $\boldsymbol{h}^{diff}$ and exercise discrimination h^{disc}. $\boldsymbol{h}^{diff} \in (0, 1)^{1 \times K}$, indicates the difficulty of each knowledge concept examined by the exercise, which is extended from exercise difficulty used in IRT. $h^{disc} \in (0, 1)$, used in some IRT and MIRT models, indicates the capability of the exercise to differentiate between those students whose knowledge mastery is high from those with low knowledge mastery. They can be obtained by:

$$\boldsymbol{h}^{diff} = \text{sigmoid}(\boldsymbol{x}^e \times \mathbf{B}) \quad \text{and} \quad h^{disc} = \text{sigmoid}(\boldsymbol{x}^e \times \mathbf{D}), \tag{4}$$

where $\mathbf{B}$ and $\mathbf{D}$ are trainable, and $\mathbf{B} \in \mathbb{R}^{M \times K}, \boldsymbol{D} \in \mathbb{R}^{M \times 1}$.

Interaction Function The first layer of the interaction layers is inspired by MIRT models. We formulate it as:

$$\boldsymbol{x} = \boldsymbol{Q}_e \circ (\boldsymbol{h}^s - \boldsymbol{h}^{diff}) \times h^{disc}, \tag{5}$$

where $\circ$ is element-wise product. Following are two full connection layers and an output layer:

$$\boldsymbol{f}_1 = \phi(\mathbf{W}_1 \times \boldsymbol{x}^T + \boldsymbol{b}_1), \tag{6}$$

$$\boldsymbol{f}_2 = \phi(\mathbf{W}_2 \times \boldsymbol{f}_1 + \boldsymbol{b}_2), \tag{7}$$

$$y = \phi(\mathbf{W}_3 \times \boldsymbol{f}_2 + b_3), \tag{8}$$

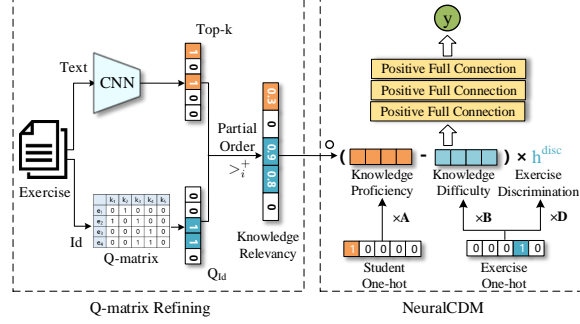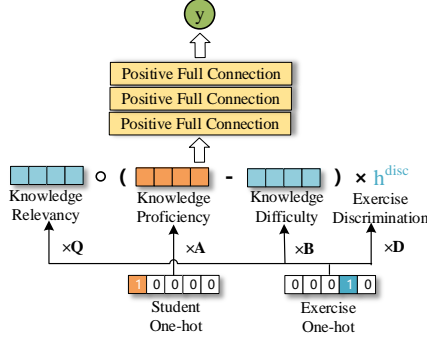where ϕ is the activation function. Here we use Sigmoid.

Figure 3: Neural cognitive diagnosis model.　Figure 4: Extended neural cognitive diagnosis model.

Different methods can be used to satisfy the monotonicity assumption. We adopt a simple strategy: restrict each element of $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$ to be positive. It can be easily proved that $\frac{\partial y}{\partial h_i^s}$ is positive for each entry h_i^s in $\boldsymbol{h}^s$. Thus monotonicity assumption is always satisfied during training.

The loss function of NeuralCDM is cross entropy between output y and true label r:

$$loss_{CDM} = -\sum_i (r_i \log y_i + (1 - r_i) \log(1 - y_i)). \tag{9}$$

After training, the value of $\boldsymbol{h}^s$ is what we get as diagnosis result, which denotes the student's knowledge proficiency.

3.3.1 NeuralCD Extension with Text Information

We now show the extendability of NerualCD through the use of exercise texts. In traditional methods, exercise texts are not used for modeling. However, these texts contain important information about the exercises which can be useful for diagnosis, such as exercise difficulty and related knowledge concepts. Here we use exercise texts to find possible relevant knowledge concepts, and use them to refine manually-labeled Q-matrix, which is deficient because of inevitable errors and subjective bias [30, 11]. For example, in Q-matrix, maybe only 'Equation' is labeled for an equation solving exercise. However, we may discover that 'Division' is also required due to the existence of '÷' in the text. We denote the extended model as NeuralCDM+, and present its structure in Figure 4.

Specifically, we first pre-train a CNN (convolutional neural network) to predict knowledge concepts related to the input exercise. CNN has advantage of extracting local information in text processing, thus it's able to capture important words from texts (e.g., words that are highly relative to certain knowledge concepts). The network takes concatenated word2vec embedding of words in texts as input, and output the relevancy of each knowledge concept to the exercise. Human-labeled Q-matrix is used as label for training. We define $V_i^k = \{V_{ij_1}, V_{ij_2}, \ldots, V_{ij_k}\}$ as the set of top-k knowledge concepts of exercise e_i outputted by the CNN.

Then we combine V_i^k with Q-matrix. Although there are defects in human-labeled Q-matrix, it still has high confidence. So we consider knowledge concepts labeled by Q-matrix are more relative than $\{k_j | k_j \in V_i^k \text{ and } Q_{ij} = 0\}$. For convenience, we define partial order $>_i^+$ as:

$$a >_i^+ b, \ if \ Q_{ia} = 1 \ and \ Q_{ib} = 0 \ and \ b \in V_i^k, \tag{10}$$

and define the partial order relationship set as $D_V = \{(i, a, b) | a >_i^+ \ b, i = 1, 2, \ldots, M\}$. To make Q-matrix continuous, we assume $\tilde{\mathbf{Q}}$ follows a zero mean Gaussian prior with standard deviation σ of each dimension, following the traditional Bayesian treatment. And define $p(a >_i^+ b | \tilde{\boldsymbol{Q}}_i)$ with a logistic-like function:

$$p(a >_i^+ b | \tilde{\boldsymbol{Q}}_i) = \frac{1}{1 + e^{-\lambda(\tilde{Q}_{ia} - \tilde{Q}_{ib})}}. \tag{11}$$

The parameter λ controls the discrimination of relevance values between labeled and unlabeled knowledge concepts. The log posterior distribution over D_V on $\tilde{\mathbf{Q}}$ is finally formulated as:

$$\begin{aligned}
\ln p(\tilde{\mathbf{Q}} | D_V) &= \ln \prod_{(i,a,b) \in D_V} p(a >_i^+ b | \tilde{\boldsymbol{Q}}_i) p(\tilde{\boldsymbol{Q}}_i) \\
&= \sum_{i=1}^{M} \sum_{a=1}^{K} \sum_{b=1}^{K} I(a >_i^+ b) \ln \frac{1}{1 + e^{-\lambda(\tilde{Q}_{ia} - \tilde{Q}_{ib})}} + C - \sum_{i=1}^{M} \sum_{j=1}^{K} \frac{\tilde{Q}_{ij}^2}{2\sigma^2},
\end{aligned} \tag{12}$$

5

where C is a constant that can be ignored during optimization. Sigmoid function is conducted on $\tilde{\mathbf{Q}}$ to restrict the range of each element to $(0,1)$. Let $\mathbf{M} \in \{0,1\}^{M \times K}$ be a mask matrix, where $M_{ij} = 1$ if $j \in V_i^k$ or $Q_{ij} = 1$; $M_{ij} = 0$ otherwise. Then $Sigmoid(\tilde{\mathbf{Q}}) \circ \mathbf{M}$ is used to replace $\mathbf{Q}$ in NeuralCDM. $\tilde{\mathbf{Q}}$ is trained together with the cognitive diagnostic model, thus the loss function is:

$$loss = -\ln p(\tilde{\mathbf{Q}}|D_V) + loss_{CDM}. \tag{13}$$

3.3.2 Generality of NeuralCD

NeuralCD is a general framework that can cover many traditional cognitive diagnostic models. Using Eq. (5) as the first layer, we now show the close relationship between NeuralCD and traditional models MF, IRT and MIRT.

MF $\boldsymbol{Q}_e$ and $\boldsymbol{h}^s$ can be seen as exercise and student latent trait vectors respectively in MF. By setting $\boldsymbol{h}^{diff} \equiv \mathbf{0}$ and $h^{disc} \equiv 1$, the output of the first layer is $\boldsymbol{x} = \boldsymbol{Q}_e \circ \boldsymbol{h}^s$. Then in order to work like MF (i.e., $y = \boldsymbol{Q}_e \cdot \boldsymbol{h}^s$), all the rest of layers need to do is to sum up the values of each entry in $\boldsymbol{x}$, which is easy to achieve. Monotonicity assumption is not applied in MF approaches.

IRT Take the typical formation of IRT $y = \text{Sigmoid}((h^s - h^{diff}) \times h^{disc})$ as example. Set $Q_e \equiv 1$, and let $\boldsymbol{h}^s$ and $\boldsymbol{h}^{diff}$ be unidimensional, the output of the first layer is $x = (h^s - h^{diff}) \times h^{disc}$, followed by a Sigmoid activation function. Monotonicity assumption is achieved by limiting h^{disc} to be positive. Other variations of IRT (e.g., $y' = C + (1 - C)y$ where C is guessing parameter) can be realized with a few changes.

MIRT One direct extension from IRT to MIRT is to use multidimensional latent trait vectors of exercises and student. Here we take the typical formation proposed in [17] as example:

$$y = \frac{e^{\mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e}}{1 + e^{\mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e}}. \tag{14}$$

Let $h^{disc} \equiv 1$, the output of the first layer given by Eq. (5) is $\boldsymbol{x} = \mathbf{Q}_e \circ (\boldsymbol{h}^s - \boldsymbol{h}^{diff})$. By Setting $\mathbf{W}_1 = [1 \quad 1 \quad \cdots \quad 1], \boldsymbol{b}_1 = \mathbf{0}$ and $\phi(x) = x$ in Eq. (6), we have $f_1 = \mathbf{Q}_e \cdot \boldsymbol{h}^s - d_e$ (where $d_e = \mathbf{Q}_e \cdot \boldsymbol{h}^{diff}$). All the rest of the layers need to do is to approximate the function $g(f_1) = 1 - \text{Sigmoid}(f_1)$, which can be easily achieved with two more layers. Monotonicity assumption can be realized if each entry of $\mathbf{Q}_e$ is restricted to be positive.

3.4 Discussion

We have introduced the details of NeuralCD framework and showed special cases of it. It's necessary to point out that the student's proficiency vector F^s and exercise's knowledge relevancy vector F^{kn} is the basic diagnostic factors needed in NeuralCD framework. Additional factors such as exercise difficulty and discrimination can be integrated in if reasonable. The formation of the first interactive layer is not limited, but it's better to contain the term $F^s \circ F^{kn}$ to ensure that each dimension of F^s corresponds to a specific knowledge concept. The positive full connection is only one of the strategies that implement monotonicity assumption. More sophisticated neural network structures can be designed as the interaction layers. For example, recurrent neural network may be used to capture the time characteristics of the student's learning process.

4 Experiments

We first compare our NeuralCD models with some baselines on the student performance prediction task. Then we make some interpretation assessments of the models.

Dataset We use two real-world datasets in the experiments, i.e., Math and ASSIST. Math is collected from a widely-used online learning system[1], which contains mathematical exercises and students data of high school examinations. ASSIST is an open dataset: Assistments 2009-2010 "skill builder"[2], which only provides student response logs and knowledge concepts. Table 1 summarizes basic statistics of the datasets.

Experimental Setup For dataset Math, we first choose response logs of objective exercises (response is binary, i.e., correct or incorrect) for diagnostic network. Then we filter all exercises with the same set of knowledge concepts, except those appear in logs, for the Q-matrix refining part of NeuralCDM+. Therefore we got 2,507 exercises with 497 knowledge concepts for diagnostic network. We perform a 80%/20% train/test split of each student's response log. As for ASSIST, we

---

[1]We omit system name due to the anonymity principle

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

Table 1: Dataset summary.

Dataset	Math	ASSIST
Students	10,268	4,163
Exercises	917,495	17,746
Knowledge concepts	1,488	123
Response logs	864,722	324,572
Average knowledge concepts per exercise	1.53	1.19

Table 2: Experimental results on student performance prediction.

Model	Math			ASSIST		
	Accuracy	RMSE	AUC	Accuracy	RMSE	AUC
DINA	0.593±.001	0.487±.001	0.686±.001	0.650±.001	0.467±.001	0.676±.002
IRT	0.782±.002	0.387±.001	0.795±.001	0.674±.002	0.464±.002	0.685±.001
MIRT	0.793±.001	0.378±.002	0.813±.002	0.693±.002	0.466±.001	0.713±.003
PMF	0.763±.001	0.407±.001	0.792±.002	0.657±.002	0.479±.001	0.732±.001
NeuralCDM	0.792±.002	0.378±.001	0.820±.001	**0.719±.008**	**0.439±.002**	**0.749±.001**
NeuralCDM+	**0.804±.001**	**0.371±.002**	**0.835±.002**	-	-	-

divide the response logs in the same way with Math, but NeuralCDM+ is not evaluated on this dataset as exercise text is not provided. All models are evaluated with 5-fold cross validation.

The dimensions of the full connection layers (Eq. (6)$\sim$ (8)) are 512, 256, 1 respectively, and Sigmoid is used as activation function for all of the layers. We set hyperparameters $\lambda = 0.1$ (Eq. (11)) and $\sigma = 1$ (Eq. (12)). For k in top-k knowledge concepts selecting, we use the value that make the predicting network reach 0.85 recall. That is, in our experiment, $k = 20$.

To evaluate the performance of our NeuralCD models[3], we compare it with previous approaches, i.e., DINA, IRT, MIRT and PMF. All models are implemented by PyTorch using Python, and all experiments are run on a Linux server with four 2.0GHz Intel Xeon E5-2620 CPUs and a Tesla K20m GPU. For fairness, all models are tuned to have the best performance.

Student Performance Prediction The performance of a cognitive diagnosis model is difficult to evaluate as we can't obtain the true knowledge proficiency of students. As diagnostic result is usually acquired through predicting students' performance in most works, performance on these prediction tasks can indirectly evaluate the model from one aspect. Considering that all the exercises in our data are objective exercises, we use evaluation metrics from both classification aspect and regression aspect, including accuracy, RMSE (root mean square error) and AUC (area under curve).

Table 2 shows the experimental results of all models on student performance prediction task. The error bars after '±' is the standard deviations of 5 evaluation runs for each model. From the table, we can observe that NeuralCD models outperform almost all the other baselines on both datasets, indicating the effectiveness of our framework. In addition, the better performance of NeuralCDM+ over NeuralCDM proves that the Q-matrix refining method is effective. Besides, it also demonstrates the importance of fine estimated knowledge relevancy vectors for cognitive diagnosis.

Model Interpretation To assess the interpretability of NeuralCD framework (i.e., whether the diagnostic result is reasonable), we further conduct several experiments.

Intuitively, if student a has a better mastery on knowledge concept k than student b, then a is more likely to answer exercises related to k correctly than b [31]. We adopt Degree of Agreement (DOA) [32] as the evaluation metric of this kind of ranking performance. Particularly, for knowledge concept k, $DOA(k)$ is formulated as:

$$DOA(k) = \frac{1}{Z} \sum_{a=1}^{N} \sum_{b=1}^{N} \delta(F_{ak}^s, F_{bk}^s) \sum_{j=1}^{M} I_{jk} \frac{J(j,a,b) \wedge \delta(r_{aj}, r_{bj})}{J(j,a,b)}, \tag{15}$$

where $Z = \sum_{a=1}^{N} \sum_{b=1}^{N} \delta(F_{ak}^s, F_{bk}^s)$. F_{ak}^s is the proficiency of student a on knowledge concept k. $\delta(x,y) = 1$ if $x > y$ and 0 otherwise. $I_{jk} = 1$ if exercise j contains knowledge concept k and 0

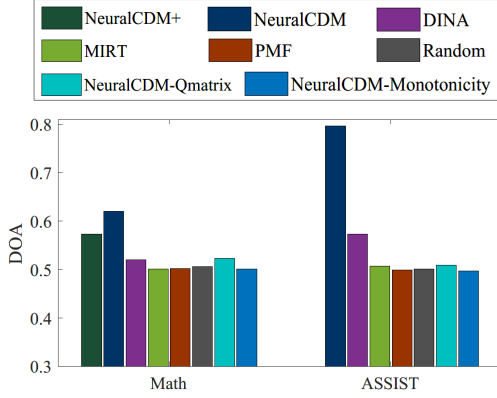[3]The code will be publicly available after the paper acceptance.

Figure 5: DOA results of models.

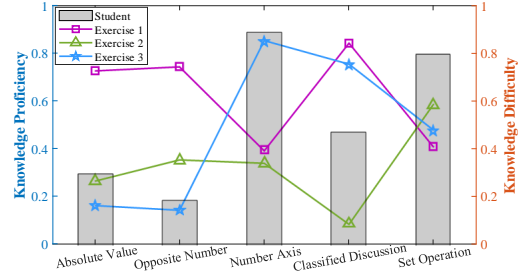|	Absolute Value	Opposite Number	Number Axis	Classified Discussion	Set Operation	Student Response
Exercise 1	1	1	0	0	0	✗
Exercise 2	0	0	1	1	0	✓
Exercise 3	0	0	0	0	1	✓

Figure 6: Diagnosis example of a student.

otherwise. $J(j, a, b) = 1$ if both student a and b did exercise j and 0 otherwise. We average $DOA(k)$ on all knowledge concepts to evaluate the quality of diagnostic result (i.e., knowledge proficiency acquired by models).

The dimension of students' latent trait vectors in PMF and MIRT are set to be equal to the number of knowledge concepts. IRT is not tested as it is unidimensional. Besides, we conduct experiments on two reduced NeuralCDM models. In the first reduced model (denoted as NeuralCDM-Qmatrix), knowledge relevancy vectors are estimated during unsupervised training instead of getting from Q-matrix. While in another reduced model (denoted as NeuralCDM-Monotonocity), monotonicity assumption is removed by eliminating the positive restriction on the full connection layers. These two reduced models are used to demonstrate the importance of fine-estimated knowledge relevancy vector and monotonicity assumption respectively. Furthermore, we conduct an extra experiment in which students' knowledge proficiencies are randomly estimated, and compute the DOA for comparison.

Figure 5 presents the experimental results. From the figure we can observe that DOAs of NeuralCDM and NeuralCDM+ are higher than all baselines, which proves that knowledge proficiencies diagnosed by them are reasonable. The low DOAs of two reduced NeuralCDM models indicate that the lack of information from Q-matrix or monotonicity assumption make the values of estimated knowledge proficiency vectors uninterpretable, making them incompetent for cognitive diagnosis. DOA of DINA is slightly higher than Random due to the use of Q-matrix, while MIRT and PMF perform nearly the same with Random. Besides, NeuralCDM performs much better on ASSIST than on Math. The reason may be that the number of knowledge concepts per exercise in ASSIST is smaller than that in Math, which makes the influence of knowledge concepts more focused. Fewer relevant knowledge concepts leads to sparser knowledge relevancy vectors in NeuralCDM, thus improves the model's performance on DOA, which only considers knowledge concepts contained in an exercise separately.

Case Study. Here we present an example of a student's diagnostic result of NeuralCDM on dataset Math. Figure 6 shows the Q-matrix of three exercises on five knowledge concepts and the response of a student to the exercises. The underneath subfigure presents his proficiency on the knowledge concepts and knowledge difficulties of the exercises. We can observe from the figure that the student is more likely to response correctly when his proficiency satisfies the requirement of the exercise. For example, exercise 3 requires the mastery of 'Set Operation' and corresponding difficulty is 0.47. The student's proficiency on 'Set Operation' is 0.79, which is higher than required, thus he answered it correctly. Both knowledge difficulty ($\boldsymbol{h}^{diff}$) and knowledge proficiency ($\boldsymbol{h}^s$) in NeuralCDM are explainable as expected.

5 Conclusion

In this paper, we proposed a neural cognitive diagnostic framework, NeuralCD framework, for students' cognitive diagnosis. Specifically, we first discussed necessary student and exercise factors in the framework, and placed a monotonicity assumption on the framework to ensure its interpretability. Then, we implemented a specific model NeuralCDM under the framework to show its feasibility, and further extended NeuralCDM by incorporating exercise text to refine Q-matrix. Extended experimental results on real-world datasets showed the effectiveness of NeuralCD models. We also showed that NeuralCD could be seen as the generalization of traditional cognitive diagnostic models (e.g., MIRT). The structure of the diagnostic network in our work is simple. However, with the high flexibility and potential of neural network, we hope this work could lead to further studies.

References

[1] Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 775–784. ACM, 2016.

[2] Xuan Guo, Rui Li, Qi Yu, and Anne R Haake. Modeling physicians' utterances to explore diagnostic decision-making. In *IJCAI*, pages 3700–3706, 2017.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*, pages 687–698. ACM, 2014.

[4] Hugh Burns, Carol A Luckhardt, James W Parlett, and Carol L Redfield. *Intelligent tutoring systems: Evolutions in design*. Psychology Press, 2014.

[5] Runze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. Cognitive modelling for predicting examinee performance. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[6] George D Kuh, Jillian Kinzie, Jennifer A Buckley, Brian K Bridges, and John C Hayek. *Piecing together the student success puzzle: research, propositions, and recommendations: ASHE Higher Education Report*, volume 116. John Wiley & Sons, 2011.

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

[8] Susan E Embretson and Steven P Reise. *Item response theory*. Psychology Press, 2013.

[9] Mark D Reckase. Multidimensional item response theory models. In *Multidimensional Item Response Theory*, pages 79–112. Springer, 2009.

[10] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[11] Louis V DiBello, Louis A Roussos, and William Stout. 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, 26:979–1030, 2006.

[12] Minjia Zhang, Wenhan Wang, Xiaodong Liu, Jianfeng Gao, and Yuxiong He. Navigating with graph representations for fast and scalable decoding of neural language models. In *Advances in Neural Information Processing Systems*, pages 6308–6319, 2018.

[13] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems*, pages 4957–4966, 2017.

[14] Matthias von Davier. The dina model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, 67(1):49–71, 2014.

[15] Gerhard H Fischer. Derivations of the rasch model. In *Rasch models*, pages 15–38. Springer, 1995.

[16] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[17] Raymond J Adams, Mark Wilson, and Wen-chung Wang. The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, 21(1):1–23, 1997.

[18] Andreas Toscher and Michael Jahrer. Collaborative filtering applied to educational data mining. *KDD cup*, 2010.

[19] Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, and Lars Schmidt-Thieme. Recommender system for predicting student performance. *Procedia Computer Science*, 1(2):2811–2819, 2010.

[20] Nguyen Thai-Nghe and Lars Schmidt-Thieme. Multi-relational factorization models for student modeling in intelligent tutoring systems. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*, pages 61–66. IEEE, 2015.

[21] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[22] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[24] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. Question difficulty prediction for reading problems in standard tests. In *AAAI*, pages 1352–1359, 2017.

[25] Mike Wu, Milan Mosse, Noah Goodman, and Chris Piech. Zero shot learning for code education: Rubric sampling with deep learning inference. 2019.

[26] Yu Yin, Zhenya Huang, Enhong Chen, Qi Liu, Fuzheng Zhang, Xing Xie, and Guoping Hu. Transcribing content from structural images with spotlight mechanism. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2643–2652. ACM, 2018.

[27] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[29] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[30] Jingchen Liu, Gongjun Xu, and Zhiliang Ying. Data-driven learning of q-matrix. *Applied psychological measurement*, 36(7):548–564, 2012.

[31] Yuying Chen, Qi Liu, Zhenya Huang, Le Wu, Enhong Chen, Runze Wu, Yu Su, and Guoping Hu. Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 989–998. ACM, 2017.

[32] Alain Pirotte, Jean-Michel Renders, Marco Saerens, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge & Data Engineering*, (3):355–369, 2007.