Item Response Ranking for Cognitive Diagnosis

Shiwei Tong[1] , **Qi Liu**[1,*] , **Runlong Yu**[1] , **Wei Huang**[1] , **Zhenya Huang**[1] ,
Zachary A. Pardos[2] , **Weijie Jiang**[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
[2]University of California, Berkeley
{tongsw, yrunl, ustc0411}@mail.ustc.edu.cn, {qiliuql, huangzhy}@ustc.edu.cn,
{pardos, jiangwj}@berkeley.edu

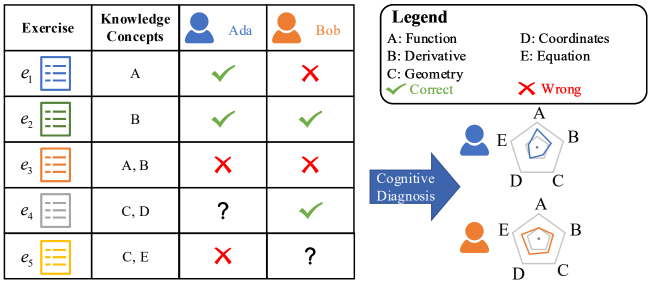
Abstract

Cognitive diagnosis, a fundamental task in education area, aims at providing an approach to reveal the proficiency level of students on knowledge concepts. Actually, monotonicity is one of the basic conditions in cognitive diagnosis theory, which assumes that student's proficiency is monotonic with the probability of giving the right response to a test item. However, few of previous methods consider the monotonicity during optimization. To this end, we propose Item Response Ranking framework (IRR), aiming at introducing pairwise learning into cognitive diagnosis to well model the monotonicity between item responses. Specifically, we first use an item specific sampling method to sample item responses and construct response pairs based on their partial order, where we propose the two-branch sampling methods to handle the unobserved responses. After that, we use a pairwise objective function to exploit the monotonicity in the pair formulation. In fact, IRR is a general framework which can be applied to most of contemporary cognitive diagnosis models. Extensive experiments demonstrate the effectiveness and interpretability of our method.

1 Introduction

Recently, online education systems has been widely used [Jiang *et al.*, 2019; Bi *et al.*, 2020; Liu *et al.*, 2019b]. These systems provide a variety of applications which can not only assist tutors to give proper instruction based on individual characteristics, e.g., strengths and weaknesses, of students, but also help students be aware of their learning progress [Pardos and Heffernan, 2010]. One of the key fundamental technologies supporting these systems is cognitive diagnosis, which tries to profile students by discovering their latent cognitive proficiency on knowledge concepts.

Massive efforts have been undertaken to improve the solutions of cognitive modelling. Generally, in cognitive diagnosis models (CDMs), students are characterized by their proficiency in specific knowledge concepts (e.g., *Circle Graph*, *Equivalent Fractions*). Classic CDMs (e.g., Item Response Theory (IRT) [Lord, 1952; Lord, 1980; Rasch, 1961] and Deterministic Inputs, Noisy "And" gate model (DINA) [De La Torre, 2009]) use single-dimension or discrete variables to represent the latent trait features of students and test items, and use simple manually designed interaction functions (e.g., logistic function in IRT). To improve the precision and interpretability, previous works mainly focus on the representation learning of trait features (e.g., extending trait features into multidimension [Reckase, 2009]) and interaction function design (e.g., using neural networks to automatically learn the complex interaction function [Wang *et al.*, 2020]). To preserve the monotonicity, previous works tend to apply a monotone function as the interaction function rather than considering it in the optimization criteria.



Exercise	Knowledge Concepts	Ada	Bob
e_1	A	✓	✗
e_2	B	✓	✓
e_3	A, B	✗	✗
e_4	C, D	?	✓
e_5	C, E	✗	?

Figure 1: Illustration of cognitive diagnosis. The left part presents the response logs of students on test items, and the right-bottom part shows corresponding diagnostic results.

In the literature, the monotonicity theory assumes that student's proficiency is monotonic with the probability of giving the right response to a test item [Rosenbaum, 1984; Wang *et al.*, 2020]. As shown in Figure 1, we can easily find that *Ada* with a right response to item e_1 is considered as having a higher proficiency level on the related concept A (i.e., *Function*) than *Bob* with a wrong response. This observation can be further interpreted from a monotonous point of view that students with correct responses should be more proficient than students with wrong responses. Therefore, based on the partial order between item responses, we can create a new optimization criteria to exploit the response pairs to enhance the monotonicity.

[*] Corresponding Author.

However, how to exploit the partial order between responses to improve the monotonicity within the optimization criteria is a challenging problem. First, it is hard for us to compare item responses across different items related to non-overlapped concepts. For example, as shown in Figure 1, it is hard to tell whether *Ada* is more skilled than *Bob* on concept B (i.e. *Derivate*) and E (*i.e., Equation*) when we only observe that *Ada* gives a wrong response to item e_5 and *Bob* gives a right response to item e_2. Second, there exist many unobserved responses such as Bob not giving a response to item e_5 in Figure 1. Last but not least, how to find an objective function so that the monotnicity can be directly optimized is another important challenge.

To this end, we propose the general Item Response Ranking framework (IRR) for cognitive diagnosis, which can be applied to most of contemporary CDMs. Specifically, we first design an item specific pair sampling method to resolve the potential non-overlapped problem, i.e., sampling responses from different students to the same item to keep related knowledge concepts the same. Then, to handle the unobserved responses along with the observed responses, we conduct a two-branch sampling method, i.e., positive sampling and negative sampling. After that, based on the sampled pairs, we introduce the pairwise learning to model the partial order among response pairs, where we use a pairwise objective function to better optimize the monotonicity. Extensive experiments on two real-world datasets show that CDMs with IRR not only significantly outperforms the baselines, but also effectively provides interpretable insights for understanding the cognitive diagnostic results of students.

2 Related Work

2.1 Cognitive Diagnosis

Cognitive diagnosis is a fundamental but important task in many real-world scenarios such as games[Chen and Joachims, 2016], medical diagnosis [Xu *et al.*, 2017], and especially, in education [Liu *et al.*, 2018; Liu *et al.*, 2019a; Huang *et al.*, 2020]. IRT [Lord, 1952] and DINA [De La Torre, 2009] are the two most fundamental but classic cognitive diagnosis models, which model the response result of a student answering an item as the interaction between the trait features of the student and the item. By extending the trait features into multidimensional, Reckase et al. [2009] proposed Multidimensional Item Response Theory (MIRT). Noticing the manually designed interaction functions of previous works may restrict the model scope of applications, NeuralCD [Wang *et al.*, 2020] exploited neural networks to automatically learn the interaction function. However, few of these works pay enough attention to the monotonicity, especially during optimization, which somehow limits their performance.

2.2 Pairwise Learning

Pairwise learning is an approach to explore the relations in a pair, which has been widely used in many areas (e.g., recommendation [Rendle *et al.*, 2012], natural language process [Huang *et al.*, 2017] and computer vision [Wang *et al.*, 2017]). For instance, Huang et al. [2013] used pairwise learning strategy to facilitate the training process for solving information retrieval. Rendle et al. [2012] constructed training pairs to solve the ranking problem in recommender system with implicit feedback. However, few works have been taken to integrate these methods into cognitive diagnosis.

3 Problem Definition

Let $S = \{s_1, s_2, ..., s_N\}$ be the set of all N students, $E = \{e_1, e_2, ..., e_M\}$ be the set of all M test items and $K = \{k_1, k_2, ..., k_L\}$ be the set of L knowledge concepts. Suppose the response logs $\mathcal{R}$ are denoted as set of triplet (s, e, r), where $s \in S$, $e \in E$ and r is the score (transferred to binary, i.e., 0 indicates wrong answer while 1, otherwise). For convenience, we also denote $\mathcal{R}$ as r_{se}. Furthermore, we have Q-matrix [Tatsuoka, 1995] labeled by experts, $Q = \{Q_{ij}\}_{M \times L}$, where Q_{ij} = 1 if the item e_i relates to the knowledge concept k_j and Q_{ij} = 0 otherwise.

Definition 1. ***Cognitive Diagnosis****: Given response logs $\mathcal{R}$ and Q-matrix Q, our goal is to mine students' proficiency on knowledge concepts.*

4 Preliminary

Before we step into our method, we would like to first briefly introduce Cognitive Diagnosis Models (CDMs). CDMs are developed to depict student's proficiency level on specific knowledge concepts based on her responses to several test items. To do this, a pointwise objective function is used to train CDMs on the Student Performance Prediction task. More concretely, CDMs are expected to minimize the difference of the predicted probability $P(y_{ie})$ of a student i giving the right response to the item e between the true response r_{ie}:

$$r_{ie} \leftarrow P(y_{ie}). \tag{1}$$

In the past decades, lots of CDMs have been proposed such as DINA and IRT. Generally, CDMs contain two parts: (1) the representations of trait features and (2) the interaction function. For example, IRT uses single-dimension variables to represent the trait features and logistic function as the interaction function as follows:

$$P(y_{ie}|\theta, a, b) = \frac{1}{1 + e^{-1.7a(\theta - b)}}, \tag{2}$$

where a and b represent the discrimination and difficulty of item e, and θ indicates the proficiency level of the student i. Using multidimensional vectors to represent latent traits of both test items and students, IRT is extended to MIRT:

$$P(y_{ie}|\boldsymbol{\theta}, \boldsymbol{a}, b) = \frac{1}{1 + e^{-\boldsymbol{a}\boldsymbol{\theta} + b}}. \tag{3}$$

However, CDMs trained by Eq. (1) cannot involve the optimization of the monotonicity while we argue that the monotonicity should be included during optimization.

5 Item Response Ranking

5.1 Overview

As we discussed before, the referred monotonicity declares that the student's proficiency is monotonic with the probability of giving the right response to a test item, which could be further expressed in a pairwise perspective: a more skilled student should have a higher probability to give the right response to a test item than an unskilled one. Formally, we have the following pairwise monotonicity:

Lemma 1. ***Pairwise Monotonicity*** *Given a specific test item, the students with right responses are more skilled than those with wrong responses.*

However, as we mentioned above, traditional methods use the pointwise optimization objective function (i.e., Eq. (1)) to train CDMs, which cannot optimize the monotonicity between responses. Therefore, instead of minimizing the difference of the predicted probability $P(y_{ie})$ between the true response r_{ie}, we would like to directly optimize the pairwise monotonicity in the objective function, i.e.,

$$(r_{ie} - r_{je}) \leftarrow P(y_{ie} - y_{je}), \tag{4}$$

where i and j represent different students.

In the following paragraphs, we will introduce: (1) how to construct the training pairs and (2) how to design the objective function so that the pair monotonicity can be promised.

5.2 Pair Construction

An important issue in IRR is to construct training pairs. For each response triplet $R = (u, e, r) \in \mathcal{R}$, we want to sample some triplets to form training pairs $T(R)$. In the following paragraphs, we will first conduct an item specific sampling, where we divide the students based on the item they give the response to. Then we will show how we handle the students with unobserved response, where we propose a two-branch sampling method. In the last part, we will present how to perform the training pair sampling.

Item Specific Sampling

For each item e, we first divide the students into two groups, one containing students who have completed the project and the other containing those who have not completed yet. The former one is called observed students $S^O(e)$ and the latter one is called unobserved students $S^U(e)$. Furthermore, we divide $S^O(e)$ into two subgroups based on their performance on e: (1) positive students $S^+(e)$ are those who correctly answer e and (2) negative students $S^-(e)$ are those who give the wrong answer:

$$\begin{aligned} S &= S^O(e) + S^U(e), \\ S^O(e) &= S^+(e) + S^-(e), \\ S^+(e) &= \{u | u \in S^O(e), r = 1\}, \\ S^-(e) &= \{u | u \in S^O(e), r = 0\}. \end{aligned} \tag{5}$$

Two-branch Sampling

Different from other scenarios like recommendation [Rendle *et al.*, 2012] where the unobserved data is usually treated as a negative sample, in cognitive diagnosis, the unobserved response log should not be simply treated as a negative one. Intuitively, we assume: the probability of any unobserved student u correctly answering e is $0 \le P(r_{s^-e}) \le P(r_{ue}) \le P(r_{s^+e}) \le 1$, which can be divided into two parts:

$$P(r_{ue}) \ge P(r_{s^-e}) \ge 0, \tag{6}$$

$$P(r_{ue}) \le P(r_{s^+e}) \le 1, \tag{7}$$

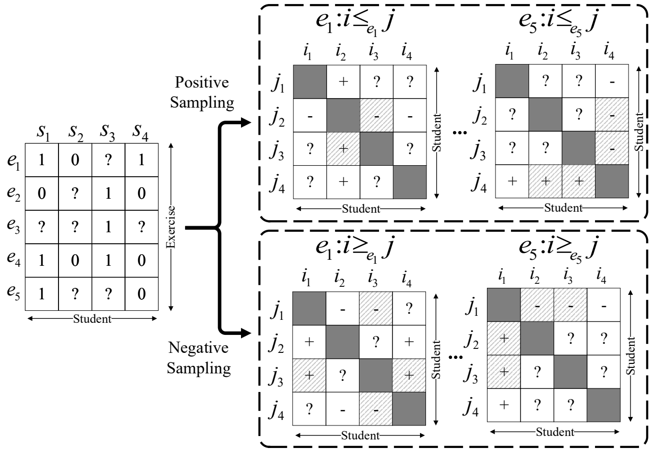


Figure 2: The observed responses are shown on the left part. Our method creates item specific pairwise partial responses $i \ge_e j$ and $i \le_e j$ between a pair of students. On the right part, (+) indicates the partial order from i to j and (–) vice versa (e.g., on the positive sampling branch, (+) means the student i have a higher probability to give the right response to the item e than the student j). The striped blocks highlight the pairs containing unobserved responses.

where $s^- \in S^-(e)$ and $s^+ \in S^+(e)$. As shown in Figure 2, based on Eq. (6) and Eq. (7), we have the two-branch sampling, i.e., positive sampling and negative sampling:

Positive sampling: Based on Eq. (6), for $s^- \in S^-(e)$, those unobserved students are considered as positive samples, which is shown in the right-top part of Figure 2.

Negative sampling: Based on Eq. (7), for $s^+ \in S^+(e)$, those unobserved students are considered as negative samples, which is illustrated in the right-bottom part of Figure 2.

Training Pair Sampling

The total number of training pairs is $|S^+(e) \times (S - S^+(e))| + |S^-(e) \times (S - S^-(e))|$, which is quite large. To accelerate the training speed, we apply the sampling method. Specifically, for each response triplet $R = (s, e, r)$, we randomly select at maximum of N^O observed samples and at maximum of N^U unobserved samples during each training step. If the student u in the response triplet $R = (s, e, r)$ is a positive student (i.e., $s \in S^+(e)$), the observed samples are selected from $S^-(e)$ and unobserved samples are selected from $S^U(e)$. We similarly select samples for the negative students and we select at maximum of N^O observed samples and at maximum of N^U unobserved samples to form training samples $T(R)$:

$$T(R) = \begin{cases} \{(s, s') | s' \in \underset{N^O}{\Lambda}(S^-(e)) \oplus \underset{N^U}{\Lambda}(S^U(e))\} & r_{se} = 1, \\ \{(s', s) | s' \in \underset{N^O}{\Lambda}(S^+(e)) \oplus \underset{N^U}{\Lambda}(S^U(e))\} & r_{se} = 0, \end{cases} \tag{8}$$

where $\underset{x}{\Lambda}(\mathcal{S})$ means randomly selecting at maximum of x elements from the set $\mathcal{S}$ to form a subset and $\oplus$ is the set addition. The discussion of how the sample number (i.e., N^O and N^U) affects the model performance is shown in Section 6.4.

5.3 Learning Model with IRR

After we construct the training pairs $T(R)$ for each response triplet, we are going to talk about the objective function.

Firstly, the log-likelihood of IRR is:

$$ln\, IRR = ln\, IRR^{+} + ln\, IRR^{-}, \tag{9}$$

$$\begin{aligned} IRR^{+} = \prod_{e\in E}\prod_{i\in S^{+}(e)}\prod_{j\in S-S^{+}(e)} & P(r_{ie} \geq r_{je}) \\ & \times (1 - P(r_{je} \geq r_{ie})), \end{aligned} \tag{10}$$

$$\begin{aligned} IRR^{-} = \prod_{e\in E}\prod_{i\in S^{-}(e)}\prod_{j\in S-S^{-}(e)} & P(r_{ie} \leq r_{je}) \\ & \times (1 - P(r_{je} \leq r_{ie})). \end{aligned} \tag{11}$$

Secondly, the objective function of IRR is:

$$min_{\Theta} - ln\, IRR + \lambda(\Theta), \tag{12}$$

where $\lambda(\Theta)$ is the regularization term and λ is a hyper-parameter. With the pair sampling mentioned in 5.2, we rewrite Eq. (12) as the following loss function:

$$\mathcal{L} = -\sum_{(i,j)\in T(R)} log \frac{exp(P(r_{ie}|\Theta))}{exp(P(r_{ie}|\Theta)) + exp(P(r_{je}|\Theta))} + \lambda(\Theta). \tag{13}$$

We can apply IRR to any fully differentiable CDMs (e.g., MIRT) and train them with Stochastic Gradient Descent.

6 Experiments

In this section, we first introduce the datasets and our experimental setups. Then, we conduct extensive experiments to compare the performances of CDMs optimized by the pointwise approach and our IRR (hereinafter referred to as pointwise-CDMs and IRR-CDMs respectively) to answer the following questions:

- RQ1: Can IRR-CDMs perform better in predicting student performance and preserve the monotonicity compared to the pointwise-CDMs?
- RQ2: Are the diagnostic results of IRR-CDMs monotonic on the knowledge level?
- RQ3: How does the number of samples (i.e., N^{O} and N^{U}) influence the performance of IRR-CDMs?
- RQ4: What are the differences of the diagnostic results between IRR-CDMs and pointwise-CDMs?

Our code is available at https://github.com/bigdata-ustc/EduCDM.

6.1 Dataset Description

We use two real-world datasets in the experiments, i.e., ASSISTments and MATH. ASSISTments (ASSISTments 2009-2010 “skill builder”) is an open dataset collected by the ASSISTments online tutoring systems [Feng *et al.*, 2009]. Collected from a widely-used online learning system, MATH contains mathematical test items and logs of high school examinations. Table 1 shows basic statistics of the datasets.

We filter out students with less than 15 and 30 response logs for ASSISTments and MATH respectively to guarantee that each student has enough data for diagnosis. For each dataset, we divide the students on each test item into training: test = 8:2. We use 90% of the training data to train model and apply grid search to adjust the hyper-parameters on the remaining 10% of the data (i.e., the validation dataset).

Statistics	ASSISTments	MATH
# users	4,163	10,268
# items	17,746	917,495
# knowledge concepts	123	1,488
# response logs	324,572	864,722

Table 1: The statistics of the dataset.

6.2 Experimental Setup

To evaluate the performance of our IRR, we apply our framework to four well-known CDMs, i.e., DINA, IRT, MIRT and NeuralCD. We make a statistics on the overall dataset to get the correct portion c_e of each item and the CDMs with IRR predict the top c_e percentage of students on each item giving the right responses. In multidimensional models (i.e., MIRT and NeuralCD), we set the dimension of latent trait features of both student and item unitedly as the number of knowledge concepts, i.e., 123 in ASSISTments and 1488 in MATH. All hyper-parameters are tuned in the validation datasets. λ is selected from $[0.1, 0.01, 0.001, 0.0001]$. N^{O} and N^{U} are selected from $[1, 5, 10, 30]$. Based on the performance on the validation datasets, we set $\lambda = 0.0001$ and $N^{O} = N^{U} = 10$.

We initialize parameters in all networks with *Xavier* initialization [Glorot and Bengio, 2010] and we use the Adam algorithm [Kingma and Ba, 2014] for optimization. All models are implemented by MXNet using Python and all experiments are run on a Linux server with two Intel(R) Xeon(R) E5-2699 v4 CPUs and a Tesla P100 PCIe GPU.

6.3 Evaluation Metrics

Classification and Ranking Metrics. Because we cannot obtain the true knowledge proficiency of students, it is hard to directly evaluate the performance of a cognitive diagnosis model. Following previous works [Wang *et al.*, 2020], as the diagnostic result is usually acquired through students performance prediction task, performance on the prediction task can indirectly evaluate the model based on some classification metrics such as *AUC*, *Precision*, *Recall* and *F1*. Besides, we need some metrics to investigate whether the monotonicity is maintained in models. Based on the monotonicity, we hope the models can correctly assign a higher predicted score to the the more skilled student. This is quite similar to a ranking problem. Therefore, we apply some commonly used ranking metrics [Yu *et al.*, 2018] to evaluate the monotonicity. The ranking metrics include *MAP* and *NDCG@k*. However, it is worth noticing that, we should not only focus on the top students (i.e., excellent students), but also need to pay attention to the underachievers (i.e., ranked at the bottom). Therefore, rather than evaluating the ranking accuracy result based on the descending order, we also evaluate it based on the ascending order. For example, given a descending ranking list $[r_1, ..., r_N]$ where the fronted students are predicted to have higher probability to correctly answer an item and r is true score, traditional NDCG@k is calculated by $NDCG@k = NDCG(r_1, ..., r_k)$. The inverse version *NDCG* is defined as $INDCG@k = NDCG(-r_N, ..., -r_{N-k+1})$. For convenience, we denote the top metrics as *MAP(E)*, *NDCG@k(E)* and *Precision@k(E)*, and represent the inverse

Metrics		DINA		IRT		MIRT		NeuralCD	
		Pointwise	IRR	Pointwise	IRR	Pointwise	IRR	Pointwise	IRR
Classification	AUC	0.786	**0.815**	0.776	**0.848**	0.777	**0.889**	0.817	**0.839**
	Precision	0.689	**0.713**	0.665	**0.752**	0.716	**0.785**	0.706	**0.734**
	Recall	0.485	**0.553**	0.585	**0.607**	0.737	**0.645**	**0.662**	0.579
	F1	0.518	**0.598**	0.534	**0.654**	0.666	**0.697**	0.625	**0.626**
Ranking	MAP(E)	0.840	**0.853**	0.817	**0.913**	0.824	**0.937**	0.867	**0.907**
	NDCG@5(E)	0.871	**0.881**	0.867	**0.894**	0.868	**0.912**	0.881	**0.889**
	MAP(U)	0.780	**0.799**	0.754	**0.880**	0.761	**0.908**	0.826	**0.875**
	NDCG@5(U)	0.464	**0.481**	0.462	**0.503**	0.462	**0.524**	0.487	**0.500**

(a) ASSISTments

Metrics		DINA		IRT		MIRT		NeuralCD	
		Pointwise	IRR	Pointwise	IRR	Pointwise	IRR	Pointwise	IRR
Classification	AUC	0.549	**0.567**	0.537	**0.666**	0.537	**0.686**	0.596	**0.608**
	Precision	0.717	**0.725**	0.637	**0.760**	0.583	**0.767**	0.686	**0.737**
	Recall	0.447	**0.707**	0.722	**0.743**	0.699	**0.750**	0.655	**0.719**
	F1	0.500	**0.711**	0.524	**0.747**	0.588	**0.754**	0.602	**0.724**
Ranking	MAP(E)	0.735	**0.740**	0.728	**0.808**	0.729	**0.814**	0.765	**0.772**
	NDCG@5(E)	0.852	**0.855**	0.845	**0.906**	0.844	**0.909**	0.853	**0.866**
	MAP(U)	0.356	**0.360**	0.345	**0.494**	0.347	**0.508**	0.431	**0.437**
	NDCG@5(U)	0.488	**0.498**	0.479	**0.654**	0.477	**0.677**	**0.587**	0.569

(b) MATH

Table 2: Experimental results on student performance prediction.

version of previous metrics as *NDCG@k(U)*, *NDCG@k(U)* and *NDCG@k(U)*, where E and U respectively represent *Excellent students* and *Underachievers*.

Degree of Agreement. Following Wang et al. [2020], we adopt Degree of Agreement (DOA) to further investigate the monotonicity based on concepts. Specifically, if student i has a better mastery on knowledge concept k than student j, then i is more likely to answer item l related to k correctly than j. For concept k, $DOA(k)$ is formulated as:

$$DOA(k) = \frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\delta(\theta_{ik},\theta_{jk})\frac{\sum\limits_{l=1}^{M} I_{lk}\wedge J(l,i,j)\wedge\delta(r_{ik},r_{jk})}{\sum\limits_{l=1}^{M} I_{lk}\wedge J(l,i,j)\wedge[r_{il}\neq r_{jl}]}}{Z}, \qquad (14)$$

where $Z = \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\delta(\theta_{ik},\theta_{jk})$. θ_{ik} is 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 item l contains concept k and $I_{lk} = 0$ otherwise. $J(l,i,j) = 1$ if both student i and j did item l and $J(l,i,j)$ = 0 otherwise. We average $DOA(k)$ on all concepts (i.e., $\overline{DOA}$) to evaluate the quality of diagnostic result.

6.4 Experimental Results

Student Performance Ranking (RQ1)

The experimental results are shown in Table 2. Each CDM result has two sub-columns, i.e., the left one shows the performance of the pointwise-CDM and the right one presents the IRR-CDM. Wilcoxon rank-sum statistical tests [Yu *et al.*, 2018] have been used to check whether the difference between original optimization strategy and our method are statistically significant (with a 0.05 significance level). From the table, we can see that for every CDM, the IRR-CDM significantly outperform the pointwise-CDM on classification metrics on all datasets. This indicates that our proposed framework can be applied to most of contemporary CDMs and can promote the diagnosis precision. Besides, we notice that, IRR-CDM also achieve higher scores in ranking metrics which means IRR can help CDMs to better maintain the monotonicity during training. Summarily, we have the conclusion that by exploiting the partial order between responses, IRR can not only help CDMs promote the diagnosis precision but also better maintain the monotonicity.

Knowledge Proficiency Monotonicity (RQ2)

Among CDMs, we only use DINA and NeuralCD as base models to verify whether the diagnostic results are monotonic on knowledge level, since for IRT, MIRT, there are no clear correspondence between their latent features and knowledge concepts. Figure 3 presents the experimental results. From the figure, we can observe that DOAs of IRR-CDMs are higher than pointwise-CDMs, which proves that the knowledge proficiency diagnosed by IRR is more monotonic.

Performance with Different Sample Number (RQ3)

In IRR, the number of samples (i.e., N^O and N^U) plays a crucial role. We use IRT and MIRT as base models to investigate the influence of different number of samples on the effectiveness of IRR. For convenience, we set $N^U = N^O$ and conduct the experiment by assigning different N^O from set $\{1, 2, 3, 4, 5, 10, 30\}$. As shown in Figure 4, as the sample number increases, the performance of IRR-CDMs increases at the beginning, but it converges afterwards both in two datasets. Meanwhile, we can find that compared with MATH,

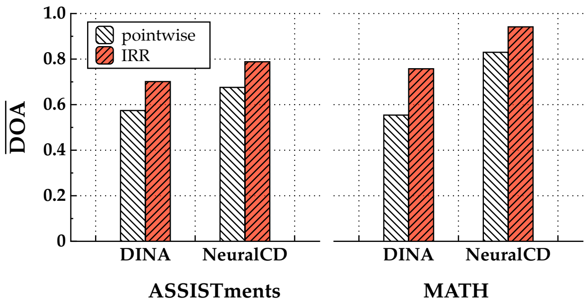


Figure 3: Results of knowledge proficiency estimation.

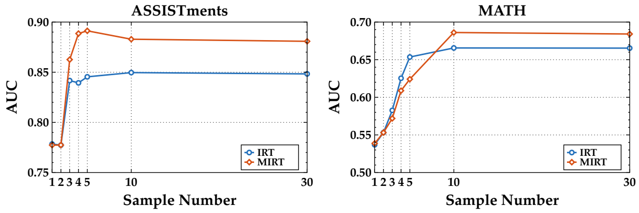


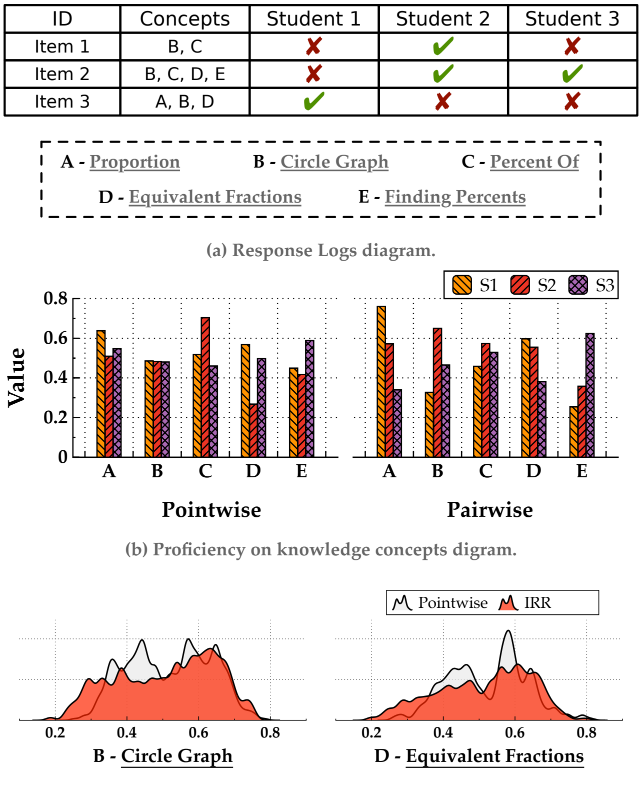
Figure 4: IRR performance with different sample number.

the performance of IRR-CDMs on ASSISTments converges earlier. The reason might be that the ASSISTments contains less data which means the models need fewer sampled data to acquire enough information. These results indicate that including more sampled pairs can help the IRR-CDMs promote the performance, but the promotion has a upper bound due to the converge of information increment.

Diagnostic Results Analysis (RQ4)

Here we present an example of a student's diagnostic results on dataset ASSISTments in Figure 5 to show the differences between pointwise-CDMs and IRR-CDMs, and we use NeuralCD as the base model. Figure 5 (a) shows the response logs of three students on three items and the related knowledge concepts of each item. Figure 5 (b) presents the cognitive diagnostic results from the pointwise-CDM (the left part) and IRR-CDM (the right part). We first look at the diagnosis report from the IRR-CDM. We can find that with correctly answering item 1 and item 2, which both contain the concept B (i.e., *Circle Graph*), student 2 has a higher proficiency on the concept B than student 1 and student 3. Similarly, we can observe that student 2 and student 3 are diagnosed to be more skilled on concept E (i.e., *Finding Percents*) than student 1, where student 2 and 3 correctly answer the related item 2 while student 1 not. Compared with the result in the left part, we can find that IRR-CDMs get a more precise and more discriminated ranking result.

Meanwhile, comparing the distribution of proficiency values in two diagnosis reports in Figure 5 (b), we notice that IRR-CDMs can give more discriminated diagnosis values. For better illustration, we visualize the all diagnosis proficiency values on concept B (*Circle Graph*) and concept D (*Equivalent Fractions*) in Figure 5 (c). We can find the distributions of the diagnosis proficiency values got by IRR-CDMs are more smooth, which means the proficiency values are more discriminated. Based on these observation, we have the conclusion that our method can help CDMs get a more discriminated diagnosis proficiency values, which accounts for why IRR-CDMs can well perform in cognitive diagnosis.



ID	Concepts	Student 1	Student 2	Student 3
Item 1	B, C	✗	✓	✗
Item 2	B, C, D, E	✗	✓	✓
Item 3	A, B, D	✓	✗	✗

A - Proportion B - Circle Graph C - Percent Of
D - Equivalent Fractions E - Finding Percents

(a) Response Logs diagram.

(b) Proficiency on knowledge concepts digram.

(c) Proficiency distribution digram.

Figure 5: An example of diagnostic results.

7 Conclusion and Future Work

In this paper, we present the monotonicity in a pair formulation and proposed the Item Response Ranking framework to incorporate the monotonicity into the optimization objective. With the proposed item specific two-branch sampling method, we manage to introduce the pairwise learning into cognitive diagnosis. As a general framework, IRR can be applied to most of CDMs. Extensive experiments demonstrate the effectiveness and interpretability of IRR framework.

In the future, we are going to optimize the response prediction with IRR by introducing difficulty prediction and fuzzy algorithm to better approximate the real correct portion. Meanwhile, we tend to improve the sampling strategy to find more distinguished samples during training so that the IRR can be more effective and efficient.

Acknowledgements

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61922073 and U20A20229) and the Fundamental Research Funds for the Central Universities (Grants No. WK2150110021).

References

[Bi *et al.*, 2020] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. Quality meets diversity: A model-agnostic framework for computerized adaptive testing. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 42–51. IEEE, 2020.

[Chen and Joachims, 2016] 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, 2016.

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

[Feng *et al.*, 2009] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3):243–266, 2009.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[Huang *et al.*, 2013] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

[Huang *et al.*, 2017] Jizhou Huang, Wei Zhang, Shiqi Zhao, Shiqiang Ding, and Haifeng Wang. Learning to explain entity relationships by pairwise ranking with convolutional neural networks. In *IJCAI*, pages 4018–4025, 2017.

[Huang *et al.*, 2020] Zhenya Huang, Qi Liu, Yuying Chen, Le Wu, Keli Xiao, Enhong Chen, Haiping Ma, and Guoping Hu. Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–33, 2020.

[Jiang *et al.*, 2019] Weijie Jiang, Zachary A Pardos, and Qiang Wei. Goal-based course recommendation. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 36–45, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[Liu *et al.*, 2019a] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2019.

[Liu *et al.*, 2019b] Qi Liu, Shiwei Tong, Chuanren Liu, Hongke Zhao, Enhong Chen, Haiping Ma, and Shijin Wang. Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 627–635, 2019.

[Lord, 1952] Frederic Lord. A theory of test scores. *Psychometric monographs*, 1952.

[Lord, 1980] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

[Pardos and Heffernan, 2010] Zachary A Pardos and Neil T Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 255–266. Springer, 2010.

[Rasch, 1961] Georg Rasch. On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 321–333, 1961.

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

[Rendle *et al.*, 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[Rosenbaum, 1984] Paul R Rosenbaum. Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49(3):425–435, 1984.

[Tatsuoka, 1995] Kikumi K Tatsuoka. Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. *Cognitively diagnostic assessment*, pages 327–359, 1995.

[Wang *et al.*, 2017] Tao Wang, Quan-Sen Sun, Qi Ge, Zexuan Ji, Qiang Chen, and Guiyu Xia. Interactive image segmentation via pairwise likelihood learning. In *IJCAI*, pages 2957–2963, 2017.

[Wang *et al.*, 2020] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6153–6161, 2020.

[Xu *et al.*, 2017] Jie Xu, Cheng Deng, Xinbo Gao, Dinggang Shen, and Heng Huang. Predicting alzheimer's disease cognitive assessment via robust low-rank structured sparse model. In *IJCAI: proceedings of the conference*, volume 2017, page 3880. NIH Public Access, 2017.

[Yu *et al.*, 2018] Runlong Yu, Yunzhou Zhang, Yuyang Ye, Le Wu, Chao Wang, Qi Liu, and Enhong Chen. Multiple pairwise ranking with implicit feedback. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1727–1730, 2018.