BDAA USTC
Anhui Province Key Laboratory
Of Big Data Analysis and Application

IJCAI 2021 MONTREAL

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, {tongsw,yrunl,ustc0411}@mail.ustc.edu.cn, {qiliuql, huangzhy}@ustc.edu.cn,
[2]University of California, Berkeley, {pardos, jiangwj}@berkeley.edu

Introduction

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.

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

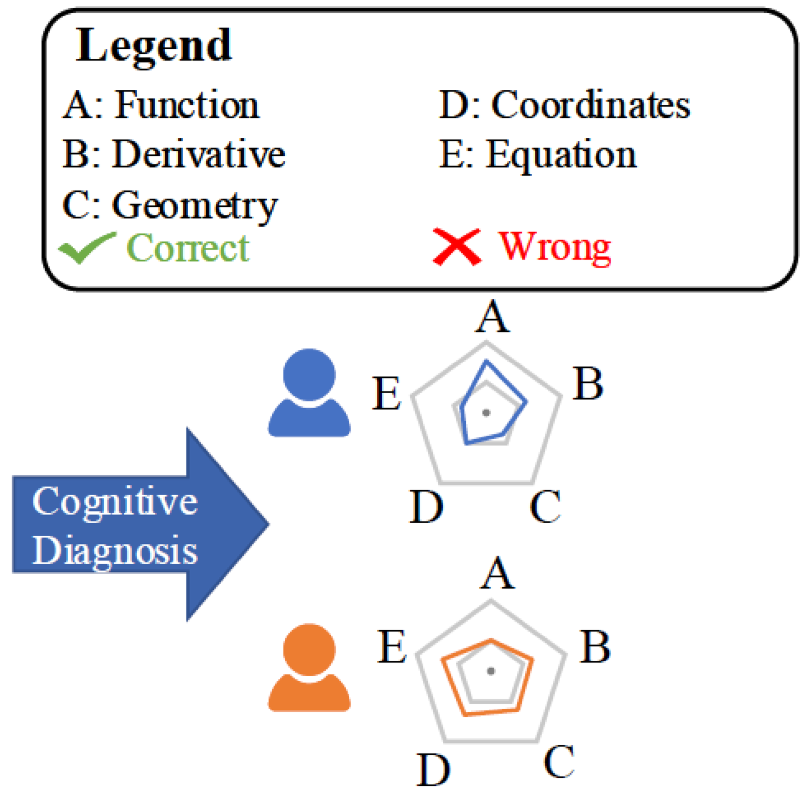


Figure 1: Illustration of cognitive diagnosis. The left part presents the response logs of students on test items, and the right-bottom part shows corresponding diagnostic results.

Observation

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.

Induction

Students with correct responses should be more proficient than students with wrong responses.

Motivation

Based on the partial order between item responses, we can create a new optimization criteria to exploit the response pairs to enhance the monotonicity.

Key Problem

How to exploit the partial order between responses to improve the monotonicity within the optimization criteria.

Challenges

1. It is hard for us to compare item responses across different items related to non-overlapped concepts.
2. There exist many unobserved responses.
3. How to find an objective function so that the monotonicity can be directly optimized

Problem Statement

Given

- N students: $S = \{s_1, s_2, \dots, s_N\}$
- M test items: $E = \{e_1, e_2, \dots e_M\}$
- L knowledge concepts: $K = \{k_1, k_2, \dots, k_L\}$
- Response logs $\mathcal{R} = \{(s, e, r)\}$
- Q-matrix $Q = \{Q_{ij}\}_{M \times L}$

Goal

Mine students' proficiency on knowledge concepts

Item Response Ranking

Monotonicity

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.

$$r_{ie} \leftarrow P(y_{ie})$$

Pairwise Monotonicity

Given a specific test item, the students with right responses are more skilled than those with wrong responses.

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

Pair Construction

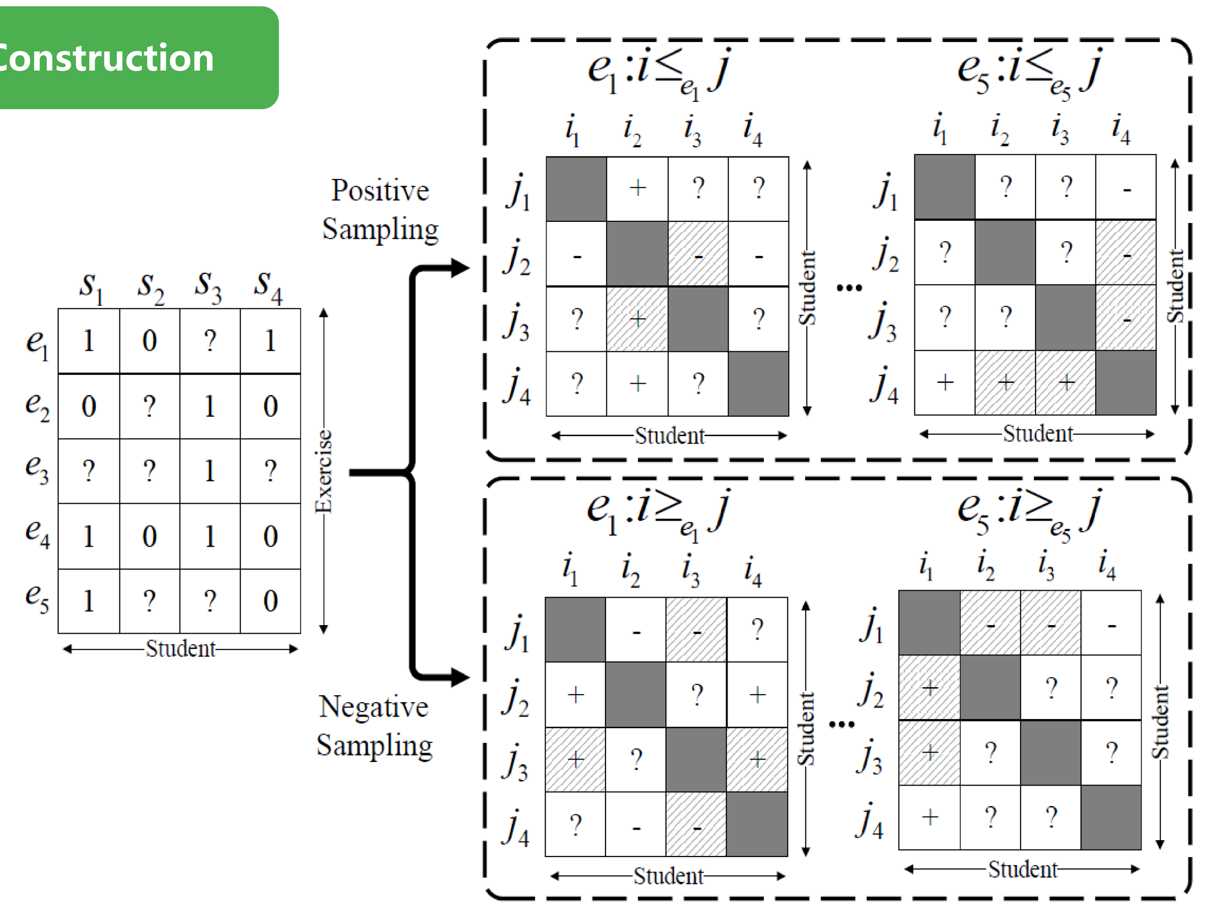


Figure 2: The observed responses are shown on the left part. Our method creates item specific pairwise partial responses $i \geq_{e_j}$ and $i \leq_{e_j}$ between a pair of students. On the right part, (+) indicates the partial order from i to j and (–) vice versa. The striped blocks highlight the pairs containing unobserved responses.

Objective Function

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

Experiments

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.

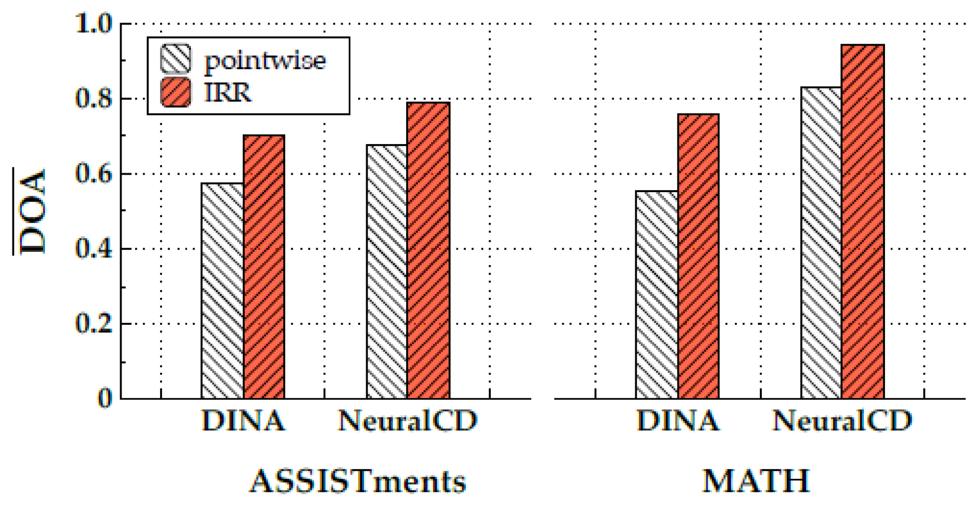


Figure 3: Results of knowledge proficiency estimation.

Metrics		DINA Pointwise	DINA IRR	IRT Pointwise	IRT IRR	MIRT Pointwise	MIRT IRR	NeuralCD Pointwise	NeuralCD 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 Pointwise	DINA IRR	IRT Pointwise	IRT IRR	MIRT Pointwise	MIRT IRR	NeuralCD Pointwise	NeuralCD 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

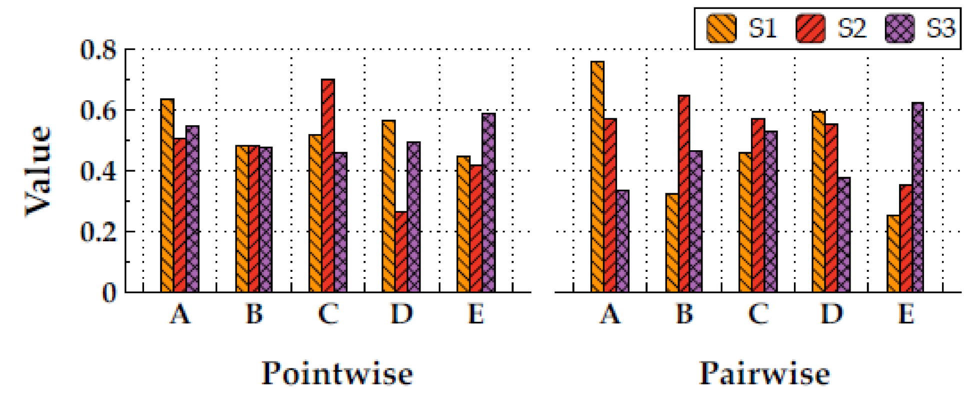
(b) MATH

Table 2: Experimental results on student performance prediction.

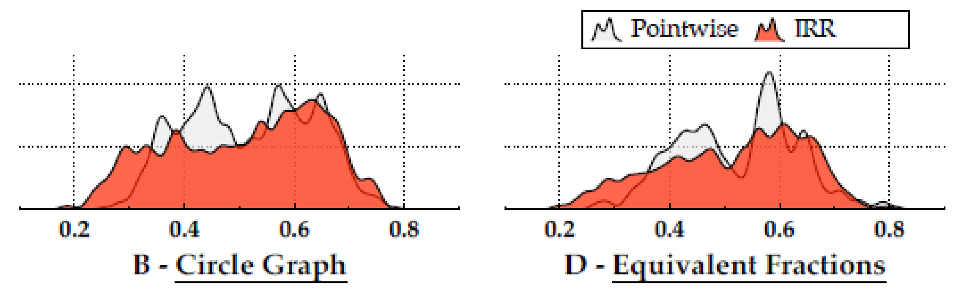
ID	Concepts	Student 1	Student 2	Student 3
Item 1	B, C	✘	✔	✘
Item 2	B, C, D, E	✘	✔	✔
Item 3	A, B, D	✔	✘	✘

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 4: An example of diagnostic results.

Observation

- The IRR-CDM significantly outperform the pointwise-CDM on classification metrics on all datasets.
- IRR-CDM achieve higher scores in ranking metrics.
- DOAs of IRR-CDMs are higher than pointwise-CDMs.
- The distributions of the diagnosis proficiency values got by IRR-CDMs are more smooth

Conclusion

- Our proposed framework can be applied to most of contemporary CDMs and can promote the diagnosis precision.
- IRR can help CDMs to better maintain the monotonicity during training.
- IRR-CDMs can give more discriminated diagnosis values.