ICD: A New Interpretable Cognitive Diagnosis Model for Intelligent Tutor Systems

Tianlong Qi[a,c,d,*], Meirui Ren[a,b,c,d,e,*], Longjiang Guo[a,b,c,d,e,**], Xiaokun Li[d,f,***], Jin Li[a,c,d] and Lichen Zhang[a,b,c,d,e]

[a]Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an 710062, China
[b]Key Laboratory of Intelligent Computing and Service Technology for Folk Song, Ministry of Culture and Tourism, Xi'an 710119, China
[c]Engineering Laboratory of Teaching Information Technology of Shaanxi Province, Xi'an 710119, China
[d]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China
[e]Xi'an Key Laboratory of Culture Tourism Resources Development and Utilization, Xi'an 710062, China
[f]Postdoctoral Program of Heilongjiang Hengxun Technology Co. Ltd, Harbin 150090, China

ARTICLE INFO

Keywords:
Cognitive diagnosis
Learner modeling
Interpretability
Knowledge concepts interaction
Quantitative relation
Potential unknown ability

ABSTRACT

Cognitive diagnosis plays a crucial role in Intelligent Tutor Systems, which aims to diagnose learners' cognitive states according to learners' observable records, such as exercise records. However, this paper observes that the existing cognitive diagnosis models still have rooms for performance improvements: (1) they ignore the interaction between knowledge concepts; (2) they ignore the quantitative relation between exercises and concepts. In order to solve the aforementioned problems, this paper proposes a new interpretable cognitive diagnosis model named ICD based on a novel neural network designed by us, which can conduct answer records with different lengths. The neural network can fully utilize the interaction among concepts, the quantitative relation between exercises and concepts, and considers bias replacement by slip and guess, which makes the ICD more interpretable. The paper has conducted extensive experiments on four real publicity datasets, and the results show that both the performances and the interpretability of ICD are better than the latest state-of-the-art CDMs such as RCD, NCDM, and CDGK, and the classical CDMs such as DINA and MIRT.

1. Introduction

Intelligent Tutor System (ITS) can simulate the behavior of human tutors to help and guide learners to complete learning tasks (Castro-Schez et al., 2021), with no limitations regarding time and location (Nilashi et al., 2022a). Although ITS attracts many learners because of its convenience, there is a concern of high dropout rate due to "one size fits all" (Nabizadeh et al., 2020). Cognitive diagnosis provides the possibility to address the concern, which can infer learners' unobservable potential factors, for example, cognitive state, memory, and logical thinking ability, according to their observable records (Wu et al., 2015; DiBello et al., 2006), such as scores on exercises. As shown in Figure 1, both Yasser and Lisa got two exercises correctly and thus scored the same, but their cognitive states were very different. Based on the cognitive states, ITS can provide personalized learning services for them. In addition, cognitive states can not only help teachers and learners improve their educational goals (Cantabella et al., 2019) but also be used for learning early warning (Fischer et al., 2020), course recommendation (Nilashi et al., 2022b), adaptive learning and computerized adaptive testing (Liu, 2021). Therefore, cognitive diagnosis has received great attention from researchers.

Learners, exercises, and knowledge concepts are the most important components of cognitive diagnosis system (Gao et al., 2021). Generally, the proficiency of a learner in all knowledge concepts is called the learner's cognitive state. Existing Cognitive Diagnosis Models (CDMs) can be regarded as inferring learners' cognitive states by simulating the complex interaction between the above three components. CDMs can be roughly divided into two categories: CDMs based on statistical methods and CDMs based on neural networks. IRT (Embretson & Reise, 2013; Lord, 2012) and DINA (de la Torre, 2009) based on statistical methods are the most classic cognitive diagnosis models. IRT believes that the probability of a learner's correct answer to exercise depends on the learner's cognitive state and difficulty of the exercise, but IRT only uses a scalar to represent the learner's cognitive state, which is lack interpretability. DINA uses binary 0 or 1 to indicate whether a learner is proficient in a certain concept, and assumes that learners can answer an exercise correctly only if they are proficient on all concepts contained in the exercise, which has good interpretability.

However, CDMs based on statistical methods rely on artificially designed diagnostic functions, which is not sufficient for capturing the complex interaction between learners and exercises (Wang et al., 2020a). Thanks to the strong learning ability of neural network (Wei et al., 2022), CDMs based on neural network overcome the shortcomings of statistical methods to a great extent. For example, NCDM

*The first two authors contributed equally.
**Principal Corresponding authors. (Longjiang Guo: longjiangguo@snnu.edu.cn)
***Corresponding authors. (Longjiang Guo: longjiangguo@snnu.edu.cn; Xiaokun Li: li.xiaokun@163.com)
1873227344@qq.com (T. Qi); meiruiren@snnu.edu.cn (M. Ren); longjiangguo@snnu.edu.cn (L. Guo); li.xiaokun@163.com (X. Li); jin.li@snnu.edu.cn (J. Li); zhanglichen@snnu.edu.cn (L. Zhang)
ORCID(s): 0000-0001-8498-4569 (T. Qi); 0000-0001-5803-7365 (M. Ren); 0000-0003-0720-2505 (L. Guo); 0000-0002-6645-6890 (X. Li); 0000-0002-0260-3169 (J. Li); 0000-0002-6711-0533 (L. Zhang)

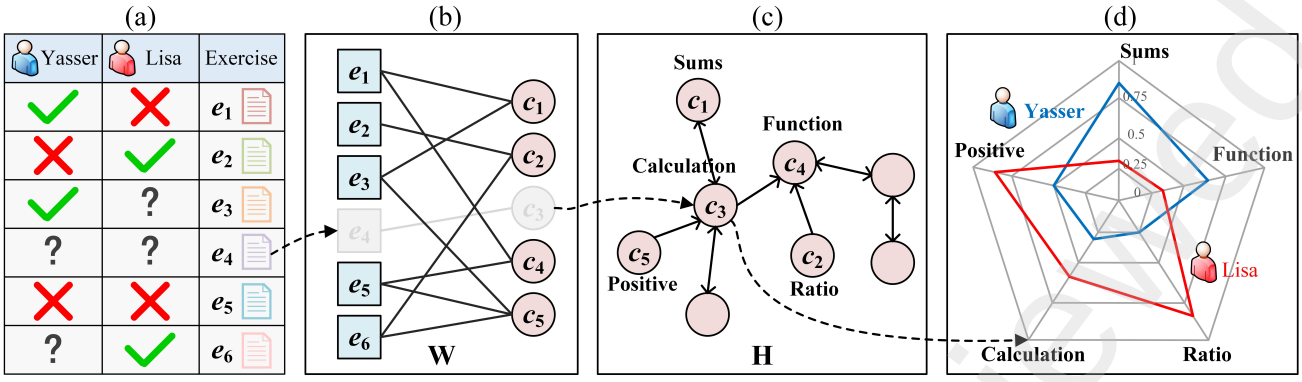

Figure 1: The illustrative examples of (a) answer records of Yasser and Lisa, "?" denotes unanswered; (b) relation between exercises and knowledge concepts, gray denotes inactive; (c) interaction among knowledge concepts; (d) cognitive states.

(Wang et al., 2020a) uses multiple neural layers to model learners and exercises, CDGK (Wang et al., 2021a) uses neural network to capture the interaction between exercise, learners' scores, and proficiency on concepts, and RCD (Gao et al., 2021) represents learners, exercises and concepts as nodes in three local relation graphs, a multi-layer attention network is constructed to aggregate the relation between nodes in graphs and the relation between graphs.

Although the above models have achieved excellent performance in cognitive diagnosis, this paper observes that they still have room for improvement. (1) Single learner usually answers only a small proportion of exercises in ITS, it is difficult to cover all concepts with these answered exercises. While most of the existing CDMs ignore the interaction among concepts (Gao et al., 2021), it is difficult to diagnose accurately the proficiency on concepts that have not been covered. (2) Existing CDMs only consider the *qualitative* relation between exercises and concepts, but ignore the *quantitative* relation between exercises and concepts.

This paper proposes a new **I**nterpretable **C**ognitive **D**iagnosis (ICD) model. ICD can mine the quantitative relation between exercises and concepts, and fully utilize the interaction among concepts. Let matrix **W** denote the quantitative relation between exercises and concepts, matrix **H** denote the interaction between concepts. The elements in **W** and **H** are the real number from 0 to 1. Firstly, learners' proficiency on the covered concepts is calculated based on learners' answer records and **W**, see Figure 1(b). Then, learners' cognitive states is diagnosed based on **H**, see Figure 1(c)(d). The details of **W** and **H** are given in section 4.2.

The contributions of this paper are as follows:

- This paper proposes a new interpretable cognitive diagnosis model named ICD based on a novel neural network designed by us, which can conduct answer records with different lengths. The neural network can fully utilize the interaction among concepts, the quantitative relation between exercises and concepts, and considers bias replacement by slip and guess, which makes the ICD more interpretable.
- To better provide learners with personalized learning services, the cognitive state diagnosed by CDMs should be able to effectively distinguish learners with different cognitive levels. The paper proposes the Degree of Distinguishing (DOD), and combines with widely accepted DOA (Wang et al., 2020a) to evaluate the interpretability of proposed ICD by two important experiments in section 5.5. Results show that ICD achieves the best interpretability.
- The paper has conducted extensive experiments on four real publicity datasets in section 5.4, and the results show that the performances of ICD on ACC, AUC, and RMSE, are better than the latest state-of-the-art CDMs such as RCD, NCDM, and CDGK, and the classical CDMs such as DINA and MIRT.

The rest of this paper is organized as follows. The related work of cognitive diagnosis is introduced in section 2. The relevant definitions are given in section 3. The details of ICD models are presented in section 4. The extensive experiment results of the baselines and ICD on four real datasets are shown in section 5. Finally, the section 6 concludes the whole paper.

2. Related Work

This section reviews the work related to cognitive diagnosis. Cognitive diagnosis models (CDMs) can be roughly divided into two categories: CDMs based on statistical method and CDMs based on neural network.

2.1. CDMs Based on Statistical Method

The early cognitive diagnosis models are based on statistical methods. IRT (Embretson & Reise, 2013; Lord, 2012) and DINA (de la Torre, 2009) are the most classical cognitive diagnosis models, which are derived from the theory of educational psychology. Many later CDMs are improved models based on these two classical models. IRT only uses a scalar to represent learners' proficiency in all concepts, and assumes that learners' scores on exercises increase with proficiency and decrease with the difficulty of exercises. MIRT

(Reckase, 2009; Reckase & McKinley, 1991) is a multidimensional improved model of IRT, which uses multidimensional vectors to represent learners' proficiency in all concepts. IRR (Tong et al., 2021) believes that if a learner is more proficient in a certain concept, the learner will score higher on exercises that contain the concept, and introduces this monotonicity into the process of model optimization. **However, there is no clear correspondence between the cognitive states output by the IRT-related model and concepts (Wang et al., 2020a), which is lack interpretability. This paper only uses MIRT as a representative model to compare the performance of such kind of models with ICD.**

Different from the model based on IRT, whether the learner is proficient in a certain concept is represented as 0 or 1 in DINA, which has good interpretability. DINA assumes that only learners who are proficient in all the concepts contained in an exercise can answer the exercise correctly, and introduces learners' guesses and slips on the exercise to fit noisy data. There are NIDA (Junker & Sijtsma, 2001), DINO (Templin & Henson, 2006) and other models similar to the principle of DINA, but DINA is still the most widely applicable. G-DINA (De La Torre, 2011) is a generalization model based on DINA, which has wider applicability. However, It is difficult to estimate the parameters of G-DINA. JRT-DINA (Zhan et al., 2018) introduces the time spent by learners on exercises into DINA and expands it. FuzzyCDF (Wu et al., 2015; Liu et al., 2018) applies fuzzy set theory to cognitive modeling, and re-expresses the proficiency of binary representation in DINA as real numbers from 0 to 1. It has better performance and better interpretability than DINA. In addition, FuzzyCDF also improves the guess and slip in DINA, so that it can be applied to both subjective and objective exercises. In NC-RUM (Hartz, 2002), abilities other than the learners' concept proficiency (such as logical ability, memory, etc.) are added to expand the cognitive diagnosis model. **Unfortunately, FuzzyCDF and NC-RUM can only handle smaller-scale datasets and are not suitable for ITS. In this paper, large-scale datasets are adopted, DINA, the most widely applicable model, is selected as the representative model.**

2.2. CDMs Based on Neural Network

CDMs based on statistical methods rely on artificially designed functions (Wang et al., 2020a), with limited performance and high complexity, it is difficult to perform tasks with large-scale data. **However, the number of learners and exercises in ITS is so large that CDMs based on statistical methods are no longer applicable.** In recent years, neural networks have achieve great progress and are widely used in many fields such as computer vision (de Almeida et al., 2022), natural language processing (Wang et al., 2021b), intelligent education (Grubišić et al., 2022), and so on. The rapid development has also led to a revolution in cognitive diagnostic models (Zhang et al., 2021). Not only improved models based on traditional models have emerged, such as Liu (Liu, 2021) who implemented the IRT, MIRT, and DINA models that use neural networks to estimate parameters, but also many new cognitive diagnosis models based on neural networks have been developed.

DIRT (Cheng et al., 2019) is based on IRT and uses deep learning to enhance the diagnostic process by the text of the exercise and the relation between the exercise and concepts. Deep-IRT (Yeung, 2019) uses a dynamic key-value memory network to estimate the difficulty of exercises and the cognitive state of learners. NCDM (Wang et al., 2020a) utilizes multiple neural layers to model learners and exercises, where monotonicity assumptions are applied to ensure model interpretability. MGCD (Huang et al., 2021) diagnoses the cognitive state of a group of learners, aiming to mine the group's proficiency on concepts and use the group's proficiency as the proficiency of each individual in the group. ECD (Zhou et al., 2021) uses hierarchical attention networks to mine the influence of contexts and cultures on learners, and then aggregate contexts and students' historical cognitive states to enhance cognitive diagnostic. **However, most of the publicity datasets lack the educational background information of learners, and ECD is not selected as the comparison model in this paper. The interaction among concepts is not considered in the above CDMs, so it is difficult to accurately diagnose the learner's proficiency in all concepts. This paper selects the most representative NCDM as a comparison model.**

In recent years, researchers have begun to combine graph structure and attention mechanism to deal with the interaction among concepts. DeepCDM (Gao et al., 2022) uses neural network and attention mechanism to learn the interaction among concepts, and the relation between exercises and concepts. **However, the interaction among concepts in DeepCDM depends on the keyword text of concepts, and the model is only suitable for small-scale datasets, so it cannot be applied to ITS.** CDGK (Wang et al., 2021a) applies a neural network to capture the interaction between exercise features, learners' scores, and learners' proficiency in concepts. **CDGK also aggregates concepts by transforming them into graph structures, but the aggregation operation decomposes the interaction among concepts into multiple subgraphs, which is not conducive to diagnosing learners' cognitive states on all concepts.** RCD (Gao et al., 2021) denotes learners, exercises, and concepts as nodes in three local relational graphs, respectively, and constructs a multi-layer attention network for node-level relation and graph-level relation aggregation. **However, there is no clear correspondence between RCD's cognitive states and concepts, which is lack interpretability. In this paper, CDGK and RCD are compared with ICD in section 5.**

3. Problem formulation

In this section, the relevant mathematical symbols, hypothesis, and theory are given. The formal problem description of cognitive diagnosis is also presented.

Let $U = \{u_i | 1 \leq i \leq \mathcal{N}\}$ denotes the set of $\mathcal{N}$ learners. $E = \{e_j | 1 \leq j \leq \mathcal{J}\}$ is the set of $\mathcal{J}$ exercises, $C = \{c_k | 1 \leq$

$k \leq \mathcal{K}\}$ is the set of $\mathcal{K}$ concepts. Cognitive states of all learners is represented as matrix $\mathbf{A} = (a_{ik})_{\mathcal{N}\times\mathcal{K}} \in \mathbb{R}^{\mathcal{N}\times\mathcal{K}}$. The i-th row $\mathbf{a}_i$ in $\mathbf{A}$ is the cognitive state of u_i, $a_{ik}(0 \leq a_{ik} \leq 1)$ represents the proficiency of u_i on the concept c_k.

Hypothesis 1: According to the monotonicity hypothesis (Rosenbaum, 1984), the more proficient the learner is in the concepts contained in one exercise, the higher the learner's score in the exercise.

Hypothesis 2: The learner's score on an exercise is mainly influenced by the concepts contained in the exercise, and different concepts have different influences on the exercise. The contained relation between exercises and concepts is represented as a matrix $\mathbf{Q} = (q_{jk})_{\mathcal{J}\times\mathcal{K}} \in \mathbb{Z}^{\mathcal{J}\times\mathcal{K}}$, $\mathbb{Z}$ is integers set, $q_{jk} \in \{0, 1\}$ denotes whether the exercise e_j contains the concept c_k.

Theory 1: According to pedagogical theory (Pinar et al., 1995; Kamii, 1986; Ellis, 1965), there is the interaction among concepts, that is, learners' proficiency in one concept will be influenced by other concepts. The interaction among concepts is qualitatively represented as a matrix $\mathbf{T} = (t_{lk})_{\mathcal{K}\times\mathcal{K}} \in \mathbb{Z}^{\mathcal{K}\times\mathcal{K}}$, $t_{lk} \in \{0, 1\}$ denotes whether the concept c_l has an influence on concept c_k. Obviously, each concept has an influence on itself, so set $t_{kk} = 1(1 \leq k \leq \mathcal{K})$.

Since there are almost no cognitive states available for reference in ITS and the existing datasets, existing works usually divide the exercises that each learner has answered into two parts, uses the scores on one part of the exercises to diagnose the learners' cognitive states $\mathbf{A}$, and then based on $\mathbf{A}$ to predict the learners' scores on the other part of the exercises, to evaluate the diagnostic effectiveness of the model indirectly. Usually, the learner only answers a part of the exercises in E. The exercises that learner u_i has already answered are represented as E_i, the answer record of u_i in E_i is represented as R_i=$\{r_{ij}|e_j \in E_i\}$, $r_{ij}(0 \leq r_{ij} \leq 1)$ denotes u_i's score on e_j. The part in E_i for diagnosing $\mathbf{a}_i$ is denoted as $E_i^{(X)}$, and the other part for prediction is denoted as $E_i^{(Y)}$, $E_i^{(X)} \cup E_i^{(Y)} = E_i$, $E_i^{(X)} \cap E_i^{(Y)} = \varnothing$ $(1 \leq i \leq \mathcal{N})$. The scores of u_i on $E_i^{(X)}$ are denoted as $X_i = \{x_{ij}|e_j \in E_i^{(X)}\}$, where the element $x_{ij}(0 \leq x_{ij} \leq 1)$ is the score of u_i on exercise e_j. The actual scores of u_i on $E_i^{(X)}$ are expressed as $Y_i = \{y_{ij}|e_j \in E_i^{(Y)}\}$, with $X_i \cup Y_i = R_i(1 \leq i \leq \mathcal{N})$ established. The predicted scores of u_i on $E_i^{(Y)}$ are denoted as $\hat{Y}_i = \{\hat{y}_{ij}|e_j \in E_i^{(Y)}\}$.

Problem Description: For $\forall u_i \in U$, given the part of exercises $E_i^{(X)}$ that learner u_i has answered and the score $X_i = \{x_{ij}|e_j \in E_i^{(X)}\}$ of u_i on $E_i^{(X)}$, there are two goals of cognitive diagnosis:

- To diagnose the cognitive state $\mathbf{a}_i$ of learner u_i, and finally get the cognitive states $\mathbf{A}$ of all learners;
- To predict the score $\hat{Y}_i = \{\hat{y}_{ij}|e_j \in E_i^{(Y)}\}$ of u_i in $E_i^{(Y)}$, make $\hat{Y}_i$ and Y_i as close as possible.

The important symbols used in this paper are shown in table 1. Generally, matrix is represented by bold capital letter, such as $\mathbf{T}$; set is represented by non-bold italic capital letter, such as E; the number is represented in cursive capital letter, such as $\mathcal{N}$; the vector is represented by bold lowercase letter, such as $\mathbf{a}_i$; the elements in set or matrix are represented by non-bold italic lowercase letters, such as c_k.

Table 1
Important symbols used in this paper

Symbol	Description
U	Learners set, $\|U\| = \mathcal{N}$, u_i is the i-th learner in the set.
E	Exercises set, $\|E\| = \mathcal{J}$, e_j is the j-th exercise in the set.
C	Knowledge concepts set, $\|C\| = \mathcal{K}$, c_k is the k-th concept in the set.
$\mathbf{Q}$	Contained relation matrix between exercises and concepts, $\mathbf{Q} \in \mathbb{Z}^{\mathcal{J}\times\mathcal{K}}$.
$\mathbf{W}$	Quantitative relation matrix between exercises and concepts, $\mathbf{W} \in \mathbb{R}^{\mathcal{J}\times\mathcal{K}}$.
$\mathbf{T}$	Interaction matrix among concepts, $\mathbf{T} \in \mathbb{Z}^{\mathcal{K}\times\mathcal{K}}$.
$\mathbf{H}$	Quantitative interaction matrix among concepts, $\mathbf{H} \in \mathbb{R}^{\mathcal{K}\times\mathcal{K}}$.
$\mathbf{A}$	Cognitive states of all learners, $\mathbf{A} \in \mathbb{R}^{\mathcal{N}\times\mathcal{K}}$.
$\mathbf{a}_i$	The i-th row of $\mathbf{A}$, denotes the cognitive state of u_i.
E_i	Exercises that the learner u_i has answered.
R_i	Answer record of u_i on E_i.
$E_i^{(X)}$	The part in E_i for diagnosing cognitive state of u_i.
X_i	Scores of u_i on $E_i^{(X)}$.
$E_i^{(Y)}$	The another part in E_i for prediction.
Y_i	Actual scores of u_i on $E_i^{(Y)}$, $X_i \cup Y_i = R_i$.
$\hat{Y}_i$	Predicted scores of u_i on $E_i^{(Y)}$.
$\mathbf{s}$	Slip rate of learners in each exercise, $\mathbf{s} \in \mathbb{R}^{1\times\mathcal{J}}$.
$\mathbf{g}$	Guess rate of learners on each exercise, $\mathbf{g} \in \mathbb{R}^{1\times\mathcal{J}}$.
P	potential unknown abilities set, $\|P\| = \mathcal{M}$, p_m is the m-th potential unknown ability in the set.
$\mathbf{D}$	Quantitative relation matrix between exercises and potential unknown abilities, $\mathbf{D} \in \mathbb{R}^{\mathcal{J}\times\mathcal{M}}$.
λ	Weight matrix of the prediction scores based on potential unknown abilities to final prediction scores, $\lambda \in \mathbb{R}^{1\times\mathcal{J}}$.
l, v	They are variables and are often used as subscripts.
$\|\cdot\|$	If the element in $\|\cdot\|$ is a scalar, it means to take the absolute value, and if the element is a set, it means to take the number of elements in the set.

4. Proposed interpretable cognitive diagnosis model (ICD)

This section designs a novel neural network, which can conduct answer records with different lengths and fully utilize the quantitative interaction among concepts $\mathbf{H}$ which is initialized by the qualitative interaction $\mathbf{T}$, and the quantitative relation between exercises and concepts $\mathbf{W}$, see section 4.2. The designed neural network also considers bias replacement by slip and guess, which makes the ICD more interpretable, see section 4.3 for details.

The core idea of ICD is given as follows. Firstly, ICD divides u_i's answer record R_i into X_i and Y_i, where R_i comes from the train set, see section 4.1 for details. Next, ICD diagnoses $\mathbf{a}_i$ based on X_i, $\mathbf{W}$ and $\mathbf{H}$, see section 4.2. Then, ICD predicts u_i's scores $\hat{Y}_i$ based on $\mathbf{a}_i$ and $\mathbf{W}$, see

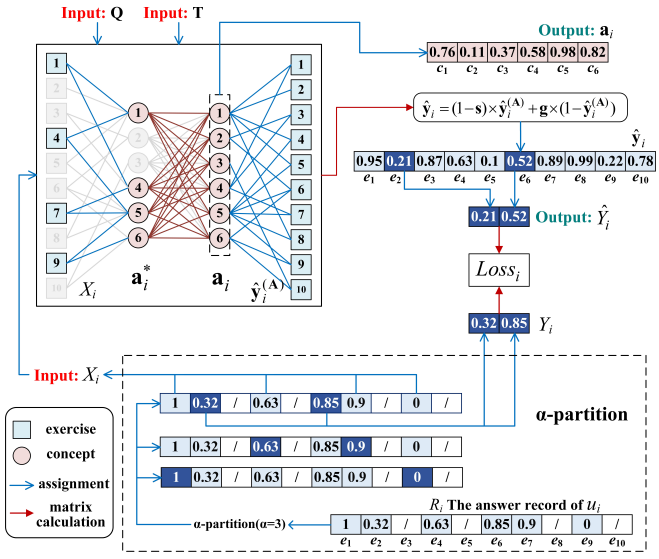

Figure 2: The schematic diagram of ICD-A. After α-partition, the answer record R_i of u_i is enhanced into α records. $\{X_i, Y_i\}$ is the first enhanced record. Gray denotes inactive.

section 4.3. Finally, $\hat{Y}_i$ compared with u_i's actual scores Y_i, the loss of ICD is calculated, and the parameters of ICD are updated by the back propagation algorithm Adam (Kingma & Ba, 2014), see section 4.4. Therefore, after several rounds of iteration, the model can predict learners' scores on one part of exercises by the scores on another part of exercises.

4.1. Divide answer records for data enhancement

Usually, a learner only answers a part of exercises set, as shown in Figure 2, there are 10 exercises, i.e. $\mathcal{J}$=10, and the learner u_i only answered 6 of them, i.e. $E_i = \{e_1, e_2, e_4, e_6, e_7, e_9\}$. To predict u_i's scores on one part of exercises by the scores on another part of exercises, it is necessary to divide u_i's answer record R_i into two parts: X_i and Y_i. In order to predict the scores on all exercises, this paper uses a data enhancement method similar to dividing data in cross validation, which is called α-partition. As shown in the Figure 2, the answer record of u_i, i.e. R_i={r_{i1}=1, r_{i2}=0.32, r_{i4}=0.63, r_{i6}=0.85, r_{i7}=0.9, r_{i9}=1}, is randomly divided into α (for example α=3) parts: {r_{i2}, r_{i6}}, {r_{i4}, r_{i7}}, and {r_{i1}, r_{i9}}. Each part is divided into Y_i in turn, and the union of the remaining parts is divided into X_i. After α-partition, R_i is enhanced as α records, i.e.

{ X_i={x_{i1}=r_{i1}, x_{i4}=r_{i4}, x_{i7}=r_{i7}, x_{i9}=r_{i9}}, Y_i={y_{i2}=r_{i2}, y_{i6}=r_{i6}} };
{ X_i={x_{i1}=r_{i1}, x_{i2}=r_{i2}, x_{i6}=r_{i6}, x_{i9}=r_{i9}}, Y_i={y_{i4}=r_{i4}, y_{i7}=r_{i7}} };
{ X_i={x_{i2}=r_{i2}, x_{i4}=r_{i4}, x_{i6}=r_{i6}, x_{i7}=r_{i7}}, Y_i={y_{i1}=r_{i1}, y_{i9}=r_{i9}} }.

After α-partition, R_i, the original answer record of u_i, is enhanced into α records. Each enhanced record is regarded as a regular answer record of certain learner and it is independent in the subsequent training process. The exercises answered by different learners are not exactly the same, the answer records of all learners are enough to cover the whole exercises set E. So, the prediction ability of ICD is enhanced by α-partition.

4.2. Diagnose cognitive states

This section introduces how does ICD diagnose cognitive states using the interaction among concepts, and the quantitative relation between exercises and concepts.

According to **Hypothesis 2**, different concepts have different influences on exercises. The influences of concepts on exercises are quantitatively denoted as matrix $\mathbf{W} = (w_{jk})_{\mathcal{J}\times\mathcal{K}} \in \mathbb{R}^{\mathcal{J}\times\mathcal{K}}$, where w_{jk} represents the influence of concept c_k on exercise e_j, here $0 < w_{jk} < 1$, if q_{jk}=1, otherwise w_{jk}=0, where q_{jk} is the element in the contained relation matrix $\mathbf{Q}$ between exercises and concepts. **This is, $\mathbf{W}$ represents the quantitative relation between exercises and concepts.**

According to **Theory 1**, there is interaction among concepts. This interaction is quantitatively denoted as matrix $\mathbf{H} = (h_{lk})_{\mathcal{K}\times\mathcal{K}} \in \mathbb{R}^{\mathcal{K}\times\mathcal{K}}$, $h_{lk}(0 < h_{lk} < 1)$ represents the influence of concept c_l on concept c_k, and for the k-th($1 \le k \le \mathcal{K}$) column of matrix $\mathbf{H}$, there is $\sum_{l=1}^{\mathcal{K}} h_{lk}$=1 established. **This is, $\mathbf{H}$ represents the quantitative interaction among concepts.**

The parameter matrices $\mathbf{W}'' \in \mathbb{R}^{\mathcal{J}\times\mathcal{K}}$ and $\mathbf{H}'' \in \mathbb{R}^{\mathcal{K}\times\mathcal{K}}$ are defined in ICD to calculate the quantitative relation matrices $\mathbf{W}$ and $\mathbf{H}$, respectively. Specifically,

$$w_{jk} = \frac{q_{jk}}{1+\exp(-w''_{jk})}, \tag{1}$$

where q_{jk} is the element in contained relation matrix $\mathbf{Q}$ between exercises and concepts, w''_{jk} is the element in $\mathbf{W}''$, and when initialized, w''_{jk} is a random number that follows the standard normal distribution, that is, $w''_{jk} \sim N(0,1)$.

$$h_{lk} = \frac{\exp(h''_{lk})}{\sum_{v=1}^{\mathcal{K}} \exp(h''_{vk})}, \tag{2}$$

where h''_{lk} is the element in $\mathbf{H}''$, and when initialized, $h_{lk} = \varepsilon \times t_{lk}$, where ε=5 is the empirical constant, t_{lk} is the element in qualitative interaction matrix $\mathbf{T}$ between concepts, here $\mathbf{T}$ can be obtained from dataset or can be calculated by the method provided by (Gao et al., 2021).

C_i denotes the concepts set covered by $E_i^{(X)}$, as shown in Figure 2, $E_i^{(X)}$={e_1, e_4, e_7, e_9}, for example, e_1 contains {c_1, c_4}, e_9 contains {c_5, c_6}, and so on, thus C_i={c_1, c_4, c_5, c_6}. Without considering the interaction among concepts, the proficiency of the learner u_i on C_i is denoted as $\mathbf{a}_i^*$, so, $|\mathbf{a}_i^*| = |C_i|$. $\mathbf{a}_i^*$ is calculated by eq. (3):

$$a_{ik}^* = \sum_{e_j \in E_i^{(X)}} w_{jk}^{(i)} x_{ij} \quad w_{jk}^{(i)} = \frac{w_{jk}}{\sum_{e_v \in E_i^{(X)}} w_{vk}}, \tag{3}$$

where a_{ik}^* is the element in $\mathbf{a}_i^*$, which denotes the proficiency of u_i on c_k without considering the interaction among concepts, and $x_{ij}(0 \le x_{ij} \le 1)$ is the score of u_i on exercise e_j, $w_{jk}^{(i)}$ is the normalized weight of x_{ij} to c_k, w_{jk} and w_{vk}

are the elements in the quantitative relation matrix $\mathbf{W}$ between exercises and concepts, which is calculated by eq. (1). As shown in Figure 2, $E_i^{(X)}=\{e_1, e_4, e_7, e_9\}$, and the exercises related to concept c_1 are $\{e_1, e_4\}$, so w_{71}=0, w_{91}=0. Suppose w_{11}=0.6, w_{41}=0.2, then $w_{11}^{(i)}$= 0.6/(0.6 + 0.2) = 0.75, $w_{41}^{(i)}$= 0.2/(0.6 + 0.2) = 0.25. If x_{i1}=0, x_{i4}=1, then a_{i1}^{*}= 0.75 × 0 + 0.25 × 1 = 0.25.

As shown in the example in Figure 2, there are no exercises related to concepts $\{c_2, c_3\}$ in $E_i^{(X)}$. If the interaction among concepts is not considered, the proficiency of u_i on concepts $\{c_2, c_3\}$ cannot be diagnosed. Therefore, it is necessary to consider the interaction among concepts during diagnosing the learners' proficiency on all concepts. After adding the quantitative interaction among concepts, the proficiency of learner u_i on all concepts is denoted as $\mathbf{a}_i$, which is calculated by eq. (4):

$$a_{ik} = \sum_{c_l \in C_i} h_{lk}^{(i)} a_{il}^{*} \quad h_{lk}^{(i)} = \frac{\exp(h''_{lk})}{\sum_{c_v \in C_i} \exp(h''_{vk})}, \tag{4}$$

where a_{ik} is the element in $\mathbf{a}_i$, which denotes the proficiency of u_i on c_k. At the same time, $\mathbf{a}_i$ is the i-th row of cognitive state matrix $\mathbf{A}$. $h_{lk}^{(i)}$ is the normalized weight of a_{il}^{*} to a_{ik}, h''_{lk} and h''_{vk} are the elements in the parameter matrix $\mathbf{H}''$.

4.3. Predict scores

This section introduces two methods of predicting scores. The first method depends on learners' cognitive states, see section 4.3.1. The second method not only depends on the cognitive states but also on learners' potential unknown abilities, see section 4.3.2 for details.

4.3.1. Predicting scores based on cognitive states

Based on the cognitive states $\mathbf{a}_i$ of u_i and the quantitative relation matrix $\mathbf{W}$ between exercises and concepts, the predicted score of u_i on all exercises is denoted as $\hat{\mathbf{y}}_i^{(\mathbf{A})}$, which is calculated as follows:

$$\hat{y}_{ij}^{(\mathbf{A})} = \sum_{k=1}^{\mathcal{K}} w_{jk}^{(j)} a_{ik} \quad w_{jk}^{(j)} = \frac{w_{jk}}{\sum_{v=1}^{\mathcal{K}} w_{jv}}, \tag{5}$$

$\hat{y}_{ij}^{(\mathbf{A})}$ is the element in $\hat{\mathbf{y}}_i^{(\mathbf{A})}$, which denotes u_i's predicted score on e_j based on $\mathbf{a}_i$ and $\mathbf{W}$, $w_{jk}^{(j)}$ is the normalized weight of a_{ik} to e_j. As shown in Figure 2, e_3 contains concepts $\{c_2, c_4\}$. Without loss of generality, let us suppose w_{32}=0.2, w_{34}=0.3, then $w_{32}^{(j)}$ = 0.2/(0.2+0.3) = 0.4, $w_{34}^{(j)}$= 0.3/(0.2 + 0.3) = 0.6. If a_{i2}=0.8, a_{i4}=0.5, then $\hat{y}_{i3}^{(\mathbf{A})}$ = 0.4 × 0.8 + 0.6 × 0.5 = 0.62.

However, in the real world, even if a learner is proficient in all concepts contained in an exercise, he/she may still answer the exercise incorrectly. In another case, even if the learner is not proficient in any concept contained in an exercise, he/she may still answer the exercise correctly. The former is called **slip**, and the latter is called **guess**. Guess and

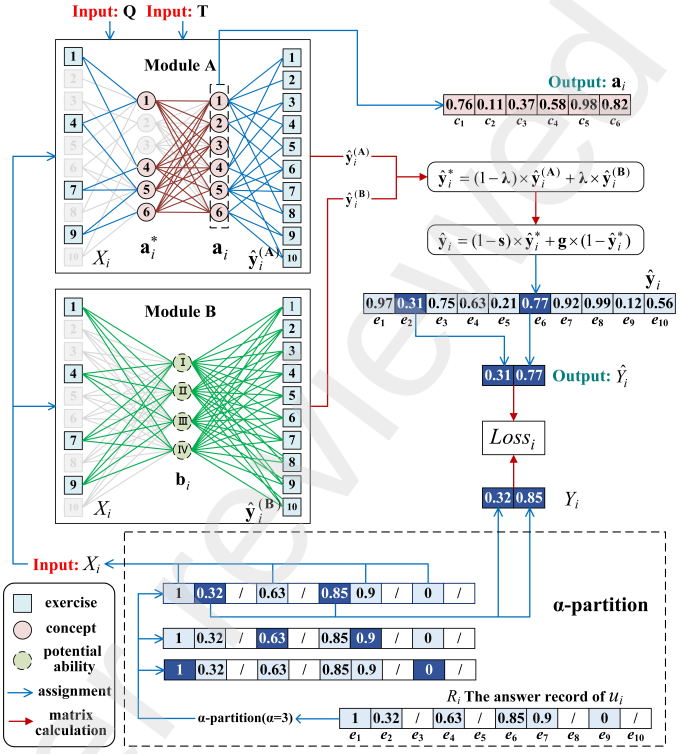

Figure 3: Schematic diagram of ICD+. Module A predicts scores based on cognitive state. Module B predicts scores based on potential unknown ability.

slip may happen at the same time (Wu et al., 2015). In this paper, the slip is denoted as $\mathbf{s} = (s_j)_{1\times\mathcal{J}} \in \mathbb{R}^{1\times\mathcal{J}}$, s_j denotes learners' slip rate on exercise e_j. The guess is denoted as $\mathbf{g} = (g_j)_{1\times\mathcal{J}} \in \mathbb{R}^{1\times\mathcal{J}}$, g_j denotes learners' guess rate on e_j. The parameter matrices $\mathbf{s}'' \in \mathbb{R}^{1\times\mathcal{J}}$ and $\mathbf{g}'' \in \mathbb{R}^{1\times\mathcal{J}}$ are defined in ICD to calculate $\mathbf{s}$ and $\mathbf{g}$, respectively. Specifically,

$$s_j = \frac{1}{1+\exp(-s''_j)} \quad g_j = \frac{1}{1+\exp(-g''_j)}, \tag{6}$$

where s''_j is the element in $\mathbf{s}''$, and when initialized, s''_j=φ, φ= − 2 is the empirical constant; g''_j is the element in $\mathbf{g}''$, and when initialized, g''_j=φ.

Usually, bias is used to fit noisy data so as to enhance the learning ability of neural networks. In this paper, slip and guess are used to replace the bias to fit what might happen for learners in the process of answering exercises. After adding slip and guess, the u_i's predicted score on all exercises is denoted as $\hat{\mathbf{y}}_i$, which is calculated by eq. (7):

$$\hat{y}_{ij} = (1-s_j)\hat{y}_{ij}^{(\mathbf{A})} + g_j(1-\hat{y}_{ij}^{(\mathbf{A})}), \tag{7}$$

$\hat{y}_{ij}$ is the element in $\hat{\mathbf{y}}_i$, which denotes the u_i's predicted score on e_j. Finally, the predicted scores on the exercises set $E_i^{(Y)}$ are extracted from $\hat{\mathbf{y}}_i$ and denoted as $\hat{Y}_i$, that is, $\hat{Y}_i = \{\hat{y}_{ij} | e_j \in E_i^{(Y)}\}$.

4.3.2. *Predicting scores based on cognitive states and potential unknown abilities*

In the real world, learners' scores on exercises are affected not only by cognitive states but also by the learners' other potential unknown abilities (such as learning ability, memory, and logical ability, etc.), and different potential abilities have different influences on exercises. $P = \{p_m | 1 \leq m \leq \mathcal{M}\}$ is the set of $\mathcal{M}$ potential unknown abilities. The influence of potential unknown abilities on exercises are quantitatively denoted as matrix $\mathbf{D} = (d_{jm})_{\mathcal{J}\times\mathcal{M}} \in \mathbb{R}^{\mathcal{J}\times\mathcal{M}}$, where d_{jm} denotes the influence of potential unknown ability p_m on exercises e_j. The parameter matrices $\mathbf{D}'' \in \mathbb{R}^{\mathcal{J}\times\mathcal{M}}$ is defined in ICD to calculate $\mathbf{D}$. Specifically,

$$d_{jm} = \frac{\exp(d''_{jm})}{\sum_{v=1}^{\mathcal{J}} \exp(d''_{vm})}, \tag{8}$$

where d''_{jm} is the element in $\mathbf{D}''$, and when initialized, d''_{jm} is a random number that follows the standard normal distribution, that is, $d''_{jm} \sim N(0, 1)$.

All learners' strength of potential unknown abilities is represented as the matrix $\mathbf{B} = (b_{im})_{\mathcal{N}\times\mathcal{M}} \in \mathbb{R}^{\mathcal{N}\times\mathcal{M}}$. The i-th row $\mathbf{b}_i$ in $\mathbf{B}$ is u_i's strength on all potential unknown abilities. $b_{im}(0 \leq b_{im} \leq 1)$ represents u_i's strength on p_m. b_{im} is calculated by eq. (9):

$$b_{im} = \sum_{e_j \in E_i} d^{(i)}_{jm} x_{ij} \quad d^{(i)}_{jm} = \frac{\exp(d''_{jm})}{\sum_{e_v \in E_i} \exp(d''_{vm})}, \tag{9}$$

where $d^{(i)}_{jm}$ is the normalized weight of score x_{ij} to b_{im}, both d''_{jm} and d''_{vm} are the element in $\mathbf{D}''$.

As shown module B in Figure 3, learners' scores on all exercises can also be predicted based on the strength of potential unknown abilities and the quantitative relation between potential unknown abilities and exercises. The predicted scores denoted as $\hat{\mathbf{y}}^{(\mathbf{B})}_i$,

$$\hat{y}^{(\mathbf{B})}_{ij} = \sum_{m=1}^{\mathcal{M}} d^{(j)}_{jm} b_{im} \quad d^{(j)}_{jm} = \frac{\exp(d''_{jm})}{\sum_{v=1}^{\mathcal{M}} \exp(d''_{jv})}, \tag{10}$$

where $\hat{y}^{(\mathbf{B})}_{ij}$ is the element in $\hat{\mathbf{y}}^{(\mathbf{B})}_i$, $d^{(j)}_{jm}$ is the normalized weight of b_{im} to exercise e_j.

The predicted scores $\hat{\mathbf{y}}^{(\mathbf{A})}_i$ based on cognitive state $\mathbf{a}_i$ are calculated by eq. (5), as shown Module A in Figure 3. The predicted scores $\hat{\mathbf{y}}^{(\mathbf{B})}_i$ based on strength of potential unknown abilities $\mathbf{b}_i$ is calculated by eq. (10). The prediction score considering both $\mathbf{a}_i$ and $\mathbf{b}_i$ is denoted as $\hat{\mathbf{y}}^*_i$, which is calculated by the following equation:

$$\hat{y}^*_{ij} = (1 - \lambda_j)\hat{y}^{(\mathbf{A})}_{ij} + \lambda_j \hat{y}^{(\mathbf{B})}_{ij}, \tag{11}$$

where $\hat{y}_{ij}$ is the element in $\hat{\mathbf{y}}^*_i$, $\lambda_j(0 < \lambda_j < 1)$ represents the weight of $\hat{y}^{(\mathbf{B})}_{ij}$ to $\hat{y}^*_{ij}$, which is the element in the weight matrix λ of $\hat{\mathbf{y}}^{(\mathbf{B})}_i$ to $\hat{\mathbf{y}}^*_i$. The parameter $\lambda'' \in \mathbb{R}^{1\times\mathcal{J}}$ is defined in the model to iteratively calculate λ. Specifically,

$$\lambda_j = \frac{1}{1 + \exp(-\lambda''_j)}, \tag{12}$$

where λ''_j is the element in λ'', and when initialized, $\lambda''_j = \phi$, $\phi = -2$ is the empirical constant.

Similar to eq. (7), after adding slip and guess, the u_i's predicted score on all exercises is denoted as $\hat{\mathbf{y}}_i$, which is calculated by eq. (13):

$$\hat{y}_{ij} = (1 - s_j)\hat{y}^*_{ij} + g_j(1 - \hat{y}^*_{ij}), \tag{13}$$

$\hat{y}_{ij}$ is the element in $\hat{\mathbf{y}}_i$, which denotes the u_i's predicted score on e_j. Finally, the predicted scores on the exercises set $E^{(Y)}_i$ are extracted from $\hat{\mathbf{y}}_i$ and denoted as $\hat{Y}_i$, that is, $\hat{Y}_i = \{\hat{y}_{ij} | e_j \in E^{(Y)}_i\}$.

4.4. Model Optimization

After α-partition, R_i, the original answer record of u_i, is enhanced into α records. For each enhanced record, ICD performs the following operations. Firstly, u_i's cognitive state $\mathbf{a}_i$ is diagnosed by combining the quantitative matrices $\mathbf{W}$ and $\mathbf{H}$, and u_i's strength $\mathbf{b}_i$ on all potential unknown abilities is calculated combining the quantitative matrix $\mathbf{D}$. Then, u_i's scores on all exercises are predicted based on $\mathbf{a}_i$ and $\mathbf{b}_i$ (or based on $\mathbf{a}_i$) combining the matrices $\mathbf{W}$, $\mathbf{D}$, $\mathbf{s}$, and $\mathbf{g}$. Finally, the predicted scores on the exercises set $E^{(Y)}_i$ are extracted from $\hat{\mathbf{y}}_i$ and denoted as $\hat{Y}_i$. The actual score of u_i on $E^{(Y)}_i$ is Y_i, that is, $Y_i = \{y_{ij} | e_j \in E^{(Y)}_i\}$. By comparing $\hat{Y}_i$ and Y_i, the loss of the model on the record of the learner can be calculated. The loss of the model on all enhanced records of all learners is:

$$Loss = \frac{1}{\alpha\mathcal{N}} \sum_{\substack{y_{ij}\in Y_i \\ \hat{y}_{ij}\in\hat{Y}_i}} -[y_{ij}\log(\hat{y}_{ij}) + (1 - y_{ij})\log(1 - \hat{y}_{ij})]. \tag{14}$$

The above training process is the forward propagation of ICD. During back propagation, the gradient of $Loss$ to model parameters $\{\mathbf{W}'', \mathbf{H}'', \mathbf{D}'', \mathbf{s}'', \mathbf{g}'', \lambda''\}$ are calculated in turn, and the model parameters are updated by Adam (Kingma & Ba, 2014) algorithm.

For the convenience of distinction, the prediction model based on $\mathbf{a}_i$ is denoted as ICD-A, and the prediction model based on $\mathbf{a}_i$ and $\mathbf{b}_i$ is denoted as ICD+. ICD-A is a simplified version of ICD+. In section 5, this paper compares the performance of ICD-A and ICD+, the experimental results show that the performance of ICD+ is better than ICD-A. Algorithm 1 shows the iterative training process of ICD+.

5. Experiment

To demonstrate the effectiveness of the proposed ICD-A and ICD+, firstly, this section introduces the datasets, baselines, and evaluation metrics, see sections 5.1 to 5.3.

Algorithm 1: The Iterative Process of ICD+

Input: $\{R_i|1 \le i \le \mathcal{N}\}$: Learners' answer records set; $\mathbf{Q}$: Contained relation matrix between exercises and concepts; $\mathbf{T}$: Qualitative interaction matrix; L: Max epoch; α: Parameter of α-partition.

Output: $\mathbf{A}$: Learner's cognitive states; $\hat{Y}$: Learner's predicted scores.

$w''_{jk} \leftarrow N(0,1)$; $h''_{lk} \leftarrow 5 \times t_{lk}$; $d''_{jk} \leftarrow N(0,1)$ $(1 \le j \le \mathcal{J})$ $(1 \le k \le \mathcal{K})$;

$s''_j \leftarrow -2$; $g''_j \leftarrow -2$; $\lambda''_j \leftarrow -2$ $(1 \le j \le \mathcal{J})$;

$\mathbf{A} \leftarrow \varnothing \in \mathbb{R}^{\mathcal{N}\times\mathcal{K}}$; $\hat{Y}_i \leftarrow \varnothing$; $Loss \leftarrow 0$; $\sigma \leftarrow$ sigmoid;

for $epoch \leftarrow 1$ *to* L **do**

  for $i \leftarrow 1$ *to* $\mathcal{N}$ **do**

    Divide R_i in the training set into X_i and Y_i by α-partition, there will produce α X_i and Y_i, denoted as $\{X_i^1, X_i^2, \cdots, X_i^\alpha\}$ and $\{Y_i^1, Y_i^2, \cdots, Y_i^\alpha\}$ respectively;

    $Loss_i \leftarrow 0$;

    for $z \leftarrow 1$ *to* α **do**

      $\hat{\mathbf{y}}_i \leftarrow \varnothing \in \mathbb{R}^{1\times\mathcal{J}}$;

      Denoted the element in X_i^z as x_{ij}, the element in $\mathbf{a}_i$ as a_{ik}, the element in $\mathbf{b}_i$ as b_{im}, the element in $\hat{\mathbf{y}}_i$ as $\hat{y}_{ij}$;

      $w_{jk} \leftarrow q_{jk} \times \sigma(w''_{jk})$; $w^{(i)}_{jk} \leftarrow \frac{w_{jk}}{\sum_{e_v \in E_i^{(X)}} w_{vk}}$;

      $h^{(i)}_{lk} \leftarrow \frac{\exp(h''_{lk})}{\sum_{c_v \in C_i} \exp(h''_{vk})}$; $d^{(i)}_{jm} \leftarrow \frac{\exp(d''_{jm})}{\sum_{e_v \in E_i} \exp(d''_{vm})}$;

      $a_{ik} \leftarrow \sum_{c_l \in C_i} \sum_{e_j \in E_i^{(X)}} h^{(i)}_{lk} w^{(i)}_{jl} x_{ij}$;

      $b_{im} \leftarrow \sum_{e_j \in E_i} d^{(i)}_{jm} x_{ij}$;

      $w^{(j)}_{jk} \leftarrow \frac{w_{jk}}{\sum_{v=1}^{\mathcal{K}} w_{jv}}$; $d^{(j)}_{jm} \leftarrow \frac{\exp(d''_{jm})}{\sum_{v=1}^{\mathcal{M}} \exp(d''_{jv})}$;

      $\hat{y}^{(\mathbf{A})}_{ij} \leftarrow \sum_{k=1}^{\mathcal{K}} w^{(j)}_{jk} a_{ik}$, $\hat{y}^{(\mathbf{B})}_{ij} \leftarrow \sum_{m=1}^{\mathcal{M}} d^{(j)}_{jm} b_{im}$;

      $\lambda_j \leftarrow \sigma(\lambda''_j)$; $s_j \leftarrow \sigma(s''_j)$; $g_j \leftarrow \sigma(g''_j)$;

      $\hat{y}^*_{ij} \leftarrow (1-\lambda_j)\hat{y}^{(\mathbf{A})}_{ij} + \lambda_j \hat{y}^{(\mathbf{B})}_{ij}$;

      $\hat{y}_{ij} \leftarrow (1-s_j)\hat{y}^*_{ij} + g_j(1-\hat{y}^*_{ij})$;

      Extract the predicted scores on exercises set $E_i^{(Y)}$ from $\hat{\mathbf{y}}_i$ and denoted as $\hat{Y}_i$;

      $Loss_i \leftarrow Loss_i - \sum_{y_{ij} \in Y_i, \hat{y}_{ij} \in \hat{Y}_i} [y_{ij} \log(\hat{y}_{ij}) + (1-y_{ij}) \log(1-\hat{y}_{ij})]$;

    $Loss \leftarrow Loss + Loss_i/\alpha$;

    if $epoch = L$ **then**

      Add $\mathbf{a}_i$ to set $\mathbf{A}$;

      Add $\hat{Y}_i$ to set $\hat{Y}$;

  $Loss \leftarrow Loss/\mathcal{N}$;

  Update $\mathbf{W}''$, $\mathbf{H}''$, $\mathbf{D}''$, $\mathbf{s}''$, $\mathbf{g}''$, and λ'' with $Loss$;

Return $\mathbf{A}$ and $\hat{Y}$.

Then, compares the performance of ICD-A, ICD+ and baselines on four real datasets, see section 5.4. Finally, analyze the interpretability of ICD-A and ICD+ based on DOA and DOD, see section 5.5.

5.1. Datasets and preprocessing

Datasets: This paper uses four real publicity datasets, namely ASSIST0910, ASSIST2017, JunYi and MathEC.

- **ASSIST0910**[1]

  ASSIST0910 is an publicity dataset collected by ASSISTments (an online tutor system), which contains the answer records of learners during the 2009-2010 school year and the contained relation between exercises and knowledge concepts, but ASSIST0910 does not provide the interaction among concepts. This paper uses the method provided by (Gao et al., 2021) to construct the interaction among concepts.

- **ASSIST2017**[2]

  ASSIST2017 comes from "The 2017 ASSISTments Datamining Competition", which provides learners' answer records from 2004 to 2007 and the contained relation between exercises and concepts. Similarly, there is no interaction information among concepts in ASSIST2017. This paper uses the method provided by (Gao et al., 2021) to construct the interaction among concepts.

- **JunYi**[3]

  JunYi comes from the online learning platform Junyi Academy, and the dataset collected answer records from October 2012 to January 2015. Each exercise contains only one concept, and one concept is contained by only one exercise. It provides the interaction among concepts marked by experts. Specifically, JunYi marks the dependency or similar relation between concepts as natural numbers from 1 to 9. The larger the marked value, the stronger the relation between the two concepts. Similar to the preprocessing operation of (Gao et al., 2021), this paper only retains the relations in which the marked value is not less than 5 as the interaction among concepts.

- **MathEC**[4]

  MathEC (Wang et al., 2020b) is from "NeurIPS 2020 Education Challenge", collected by the online education website Eedi, and contains the answer records from September 2018 to May 2020. It contains the relation between exercises, and concepts and the interaction among concepts. The interaction among concepts is represented as a tree structure. This paper

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

[2] https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017

[3] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198

[4] https://eedi.com/projects/neurips-education-challenge

Table 2
Datasets Summary

Dataset	ASSIST0910	ASSIST2017	JunYi	MathEC
#Learners	2380	1678	36591	118971
#Exercises	16804	2210	721	27613
#Concepts	110	101	721	388
#Concepts interaction	0	0	1918	387
#Concepts per exercise	1.2	1.27	1	1.1
#Answered exercises	257585	351530	1550016	15867850
#Exercises answered by per learner	108.23	209.49	42.36	133.38
#Concepts covered by per learner	14.87	52.56	42.36	34.05
Proportion of concepts covered by per learner	0.14	0.52	0.06	0.09
Average score on answered exercises	0.66	0.43	0.76	0.64

only retains the information between parent nodes and children nodes in the tree structure as the interaction among concepts.

Preprocessing: In all datasets, the same exercise may be answered several times by a learner, only the learner's first answer record is retained. To ensure that each learner has enough answer records for diagnosis, the same processing method as in the literature (Gao et al., 2021; Wang et al., 2020a) is adopted, that is, only learners who answer more than 15 exercises are retained. The statistical information of the four datasets is summarized in Table2. Where, "#" represents the number of statistics. "#Concepts covered by per learner" represents the average number of the concepts that can be covered by the answered exercises of each learner, "Proportion of concepts covered by per learner" means the proportion of the covered concepts to all concepts. "Average score on answered exercises" means the average score of all learners' answered exercises, see eq. (22) $\bar{r}$.

Experimental environment setting: The models proposed in this paper and baselines are implemented by pytorch 1.10.2 and python 3.8.3, and all experiments run on a Linux server equipped with a 3.60GHz Intel(R) Core(TM) i7-7820X CPU and RTX 2080 GPU, the running memory is 16GB. All experiments were repeated 10 times, and the average of the results of 10 times was taken as the final experimental result.

5.2. Baselines

To verify the effectiveness of ICD-A and ICD+, this paper compares them with two classical cognitive diagnosis models MIRT and DINA, as well as three state-of-the-art models NCDM, CDGK, and RCD.

- **MIRT** (Reckase, 2009; Reckase & McKinley, 1991):

  MIRT is a multidimensional improved model based on IRT, which uses multidimensional vectors to represent learners' cognitive states and the factors of exercises.

- **DINA** (de la Torre, 2009):

  DINA is one of the most classical cognitive diagnosis models, which uses 0 or 1 to indicate whether learners are proficient in one knowledge concept, and introduces learners' guesses and slips in exercises.

- **NCDM** (Wang et al., 2020a):

  NCDM is one of the earliest CDMs based on neural network. It uses multiple neural layers to model learners and exercises, and applies monotonicity hypothesis to ensure the interpretability of the model.

- **CDGK** (Wang et al., 2021a):

  CDGK uses a neural network to capture the interaction between exercises, learners' scores and cognitive states, and uses learners' guesses to adjust the predicted scores.

- **RCD** (Gao et al., 2021):

  RCD represents learners, exercises and concepts as nodes in three local relation graphs. A multi-layer attention network is constructed to aggregate the relation between nodes in graphs and the relation between graphs.

5.3. Evaluation metrics

Considering that the exercises in the datasets used in this paper are objective exercises, the scores of the exercises are 0 or 1, which denote incorrect or correct answer respectively. Both classification and regression metrics are adopted in this paper. In terms of classification, this paper uses Prediction Accuracy (ACC) and Area Under the Curve (AUC) to evaluate the prediction performance. In terms of regression, this paper uses Root Mean Square Error ($RMSE$) to measure the gap between predicted scores and the actual scores.

- **ACC**

$$ACC = \frac{1}{\sum_{i=1}^{\mathcal{N}} |E_i^{(Y)}|} \sum_{i=1}^{\mathcal{N}} \sum_{e_j \in E_i^{(Y)}} f(y_{ij}, \hat{y}_{ij}), \qquad (15)$$

where $E_i^{(Y)}$ is u_i's exercises set assigned in the test set, the same below. Here, $0 \leq y_{ij}, \hat{y}_{ij} \leq 1$, $f(y_{ij}, \hat{y}_{ij}) =$

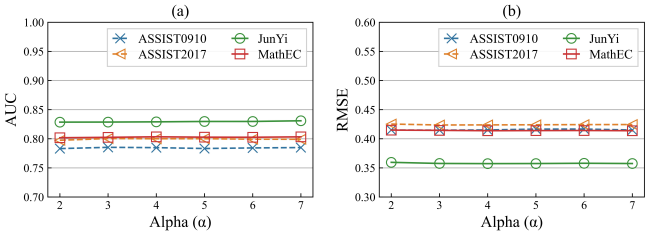

Figure 4: The impact of α.

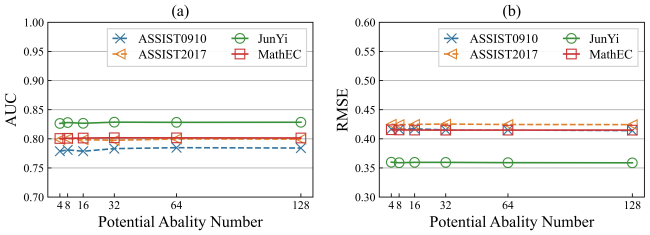

Figure 5: The impact of $\mathcal{M}$.

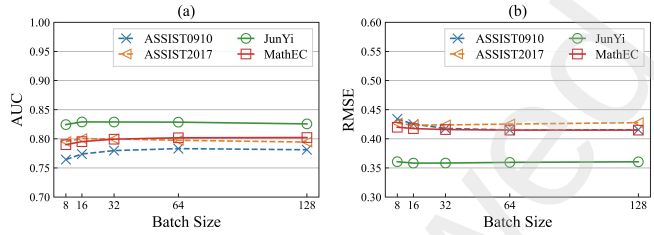

Figure 6: The impact of batch size.

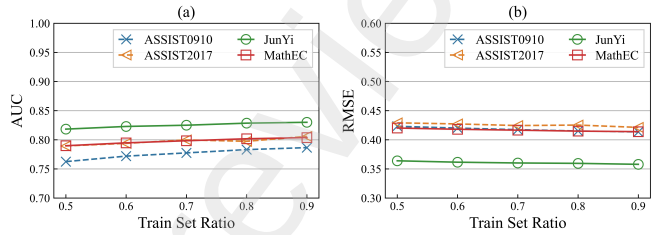

Figure 7: The impact of train set ratio.

1 if $y_{ij} = round(\hat{y}_{ij})$, $round(\cdot)$ denotes rounding operation; $f(y_{ij}, \hat{y}_{ij}) = 0$ if $y_{ij} \neq round(\hat{y}_{ij})$.

- **AUC**

  AUC represents the area under the ROC curve, and its value is between 0.5 and 1. The larger the value, the better the performance of the model. AUC can take into account both positive samples (such as correctly answered exercise) and negative samples (such as incorrectly answered exercise). AUC still has a good indication effect when the number of positive and negative samples is uneven. It is a commonly used evaluation index for classification tasks.

- **RMSE**

$$RMSE = \sqrt{\frac{1}{\sum_{i=1}^{\mathcal{N}} |E_i^{(Y)}|} \sum_{i=1}^{\mathcal{N}} \sum_{e_j \in E_i^{(Y)}} (y_{ij} - \hat{y}_{ij})^2}. \quad (16)$$

5.4. Experimental results and analysis

Diagnosed cognitive states cannot be used to evaluate the performance of CDMs directly, as there are almost no cognitive states available for reference in the ITS and existing datasets. In most existing work, learners' scores on exercises are predicted to evaluate the performance of models indirectly (Liu et al., 2018).

5.4.1. Hyperparameters sensitivity analyzes

The hyperparameters contained in ICD+ are α (parameters in α-partition), $\mathcal{M}$ (number of potential unknown abilities), Batch Size, and Train Set Ratio. This section will analyze the impact of these hyperparameters on the performance of ICD+, as well as, verify the robustness of ICD+. Since ICD-A is a simplified version of ICD+, the results of the sensitivity analysis apply to both models. Figures 4 to 7 shows the change trend on AUC and $RMSE$ of ICD+ when the hyperparameters are changed. AUC is chosen because the scores of learners in datasets are uneven, in this case, AUC can better reflect the actual performance of the model; $RMSE$ is chosen because it is widely used in multiple related works.

As shown in Figure 4 and 5, AUC and $RMSE$ hardly change with the increase of α and $\mathcal{M}$; As shown in Figure 6, when the batch size increases to 64, AUC and $RMSE$ are no longer improved. Finally, as shown in Figure 7, AUC increases with the increase of the Train Set Ratio, and $RMSE$ decreases with the increase of the Train Set Ratio, but the change is very small. To sum up, there is little impact of the above hyperparameters on the performance of ICD+, this is, both ICD-A and ICD+ are not sensitive to hyperparameters.

5.4.2. Model hyperparameters setting

Considering that the calculation amount of the models will increase with the increase of α and the number of potential unknown abilities ($\mathcal{M}$), and synthesizing the impact of the two hyperparameters on the performance of the models, α is set to 2 and $\mathcal{M}$ is set to 32 in all subsequent experiments. From section 5.4.1, when the Batch Size is increased to 64, the performance of the models is no longer improved. Therefore, set the Batch Size to 64. Since the ratio of the train set is usually set to 0.8 in other related works, that is, when dividing the dataset, 80% of the answer records of each learner are randomly divided into the train set, and the remaining 20% are divided into the test set. So, the ratio of the train set in this paper is also set to 0.8. The learning rates for both ICD-A and ICD+ are set to 0.03. Finally, because the data volumes of the four datasets are different, after

Table 3
Performance of the prediction scores of CDMs. Bold denotes the best results, and the second-best results are underlined.

CDMs	ASSIST0910			ASSIST2017			JunYi			MathCE		
	ACC ↑	AUC ↑	RMSE ↓	ACC ↑	AUC ↑	RMSE ↓	ACC ↑	AUC ↑	RMSE ↓	ACC ↑	AUC ↑	RMSE ↓
MIRT	0.5734	0.5570	0.6105	0.5546	0.5709	0.6172	0.7131	0.5252	0.5516	0.6192	0.6324	0.5880
DINA	0.6858	0.7117	0.4714	0.6199	0.7037	0.5054	0.5015	0.6238	0.5208	0.6576	0.7479	0.4751
NCDM	0.7299	0.7601	0.4353	0.6866	0.7447	0.4637	0.7370	0.6953	0.4550	0.7090	0.7444	0.4472
CDGK	0.7340	0.7660	0.4350	0.6871	0.7432	0.4532	0.8063	0.7982	0.3741	0.7390	0.7810	0.4230
RCD	0.7355	0.7721	0.4213	0.6560	0.7020	0.4649	0.7716	0.8262	0.3963	0.7013	0.7392	0.4398
ICD-A	0.7383	0.7727	0.4190	0.7194	0.7870	0.4311	0.8159	0.8180	0.3630	0.7359	0.7953	0.4182
ICD+	**0.7463**	**0.7843**	**0.4146**	**0.7299**	**0.8026**	**0.4229**	**0.8200**	**0.8284**	**0.3587**	**0.7413**	**0.8017**	**0.4149**

multiple optimization adjustments, the model sets iteration epochs to 8, 10, 1, and 2 respectively when training on the datasets ASSIST0910, ASSIST2017, JunYi, and MathEC.

5.4.3. Experimental results

Table 3 shows the performance of ICD-A, ICD+, and baselines on the four datasets. The results of CDGK on ASSIST0910 and MathEC are taken from their work (Wang et al., 2021a), the results of RCD on ASSIST0910 and JunYi are taken from their work (Gao et al., 2021).

From Table 3, ICD-A and ICD+ perform better than baselines on all metrics on the four datasets, which indicating that fully utilize of interaction among concepts and mining the quantitative relation between exercises and concepts can significantly improve the performance of CDMs. Among the baselines, the performance of NCDM, CDGK, and RCD is significantly better than that of MIRT and DINA, which verifies that the CDMs based on neural network can better simulate the complex interaction among learners, exercises, and concepts. CDGK and RCD outperform NCDM in most cases, which denotes that the interaction among concepts can improve the performance of CDMs.

The performance of ICD+ better than ICD-A also demonstrates that the model introducing potential unknown abilities can better predict learners' scores. The parameter λ defined in section 4.3.2 reflect the impact of learners' potential unknown abilities on the prediction scores. Relatively, 1-λ reflects the impact of learners' cognitive states on the prediction scores. After ICD+ convergence, the average values of λ on the datasets ASSIST0910, ASSIST2017, JunYi, and MathEC are 0.25, 0.48, 0.4, and 0.37, respectively. Relatively, the average values of 1-λ are 0.75, 0.52, 0.6, and 0.64, respectively. Therefore, when ICD+ predicting scores, the main influencing factors are learners' cognitive states, and the secondary influencing factors are learners' potential unknown abilities. This result verifies the **Hypothesis 2** proposed in this paper.

5.5. Model interpretability analysis

This paper adopts two metrics to evaluate and analyze the interpretability of cognitive diagnosis models, namely DOA and DOD. DOA is proposed by NCDM (Wang et al., 2020a), and DOD is proposed by this paper.

The experiments in this section are based on the cognitive states diagnosed by CDMs. Since MIRT represents the cognitive states of learners with multi-dimensional vectors, the elements in the vectors may be negative and have no clear correspondence with specific concepts. RCD does not depend on the cognitive states when predicting scores, and the network in RCD that outputs the cognitive states cannot be trained, there is no clear correspondence between the output cognitive states and specific concepts. Therefore, this section will not compare with MIRT and RCD.

5.5.1. Degree of agreement analysis

Monotonicity is one of the basic conditions of cognitive diagnosis theory (Tong et al., 2021). Whether the model is interpretable depends on whether the diagnosis results comply with the monotonicity hypothesis. According to the monotonicity hypothesis (Rosenbaum, 1984), the more proficient one learner is in a certain concept, the higher the learner's score should be on the exercises that contain this concept. Intuitively, if the learner u_i is more proficient in concept c_k than u_v, that is, $a_{ik} > a_{vk}$, then u_i's score on the exercise e_j containing c_k should also be higher than u_v's score on e_j, that is, $r_{ij} > r_{vj}$. To evaluate the interpretability of the cognitive diagnosis model, NCDM (Wang et al., 2020a) proposed the Degree of Agreement (DOA). NCDM believes the larger the DOA, the more the cognitive states diagnosed by the model conforms to the monotonicity hypothesis, that is, the better the interpretability of the model. DOA is calculated as follows:

$$DOA = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} DOA_k, \tag{17}$$

where,

$$DOA_k = \frac{1}{Z} \sum_{i=1}^{\mathcal{N}} \sum_{v=1}^{\mathcal{N}} \delta\left(a_{ik}, a_{vk}\right) \sum_{j=1}^{\mathcal{J}} I_{jk} \frac{J(j, u_i, u_v) \wedge \delta\left(r_{ij}, r_{vj}\right)}{J(j, u_i, u_v)}, \tag{18}$$

where,

$$Z = \sum_{i=1}^{\mathcal{N}} \sum_{v=1}^{\mathcal{N}} \delta\left(a_{ik}, a_{vk}\right), \tag{19}$$

a_{ik} is the proficiency of u_i on c_k, $\delta(x, y) = 1$ if $x > y$, otherwise $\delta(x, y) = 0$. $I_{jk} = 1$ if e_j contains c_k, otherwise

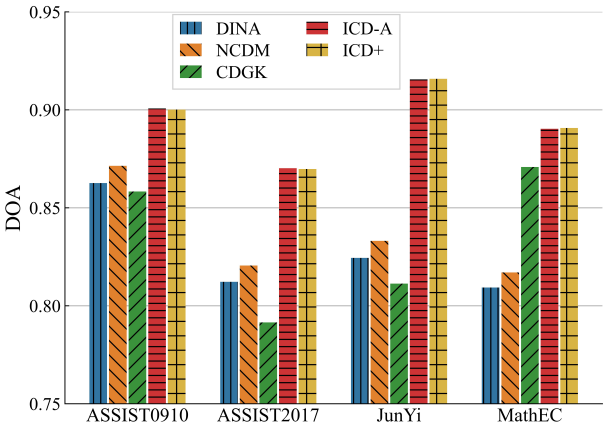

Figure 8: DOA of CDMs.

$I_{jk} = 0$. $J(j, u_i, u_v) = 1$ if both u_i and u_v answered e_j, otherwise $J(j, u_i, u_v) = 0$.

Figure 8 shows the comparison results of ICD-A, ICD+ and baselines on DOA. From the figure, the DOA of ICD-A and ICD+ are very close, and both are significantly higher than baselines, which proves that the ICD-A and ICD+ proposed in this paper are more interpretable. We also note that the DOA of NCDM is higher than DINA on four datasets, again demonstrating that CDMs based on neural network also outperforms CDMs based on statistical method in interpretability. Here's the surprise, the DOA of CDGK on ASSIST0910, ASSIST2017, and JunYi datasets is lower than that of DINA and NCDM, it indicates that CDGK still has room for improvement in interpretability.

5.5.2. Degree of distinguishing analysis

For ITS, in order to better provide learners with personalized learning services, the cognitive states diagnosed by CDMs should be able to effectively distinguish learners with different cognitive levels.

As shown in Figure 9, learners are clustered into four clusters using K-Means algorithm (Likas et al., 2003) according to their cognitive states. U_v is the v-th cluster. The average proficiency of learner u_i in U_v on all concepts is denoted as $\bar{a}_i$, and the average proficiency of U_v is denoted as $\bar{a}^{(v)}$, see eq. (20). Then, the clusters are sorted according to $\bar{a}^{(v)}$ from small to large, corresponding to the four categories: *poor*, *medium*, *good* and *excellent*.

$$\bar{a}_i = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} a_{ik} \quad \bar{a}^{(v)} = \frac{1}{|U_v|} \sum_{u_i \in U_v} \bar{a}_i, \tag{20}$$

From Figure 9, in most cases, the baselines' average proficiency of the four clusters overlap highly, and their average proficiency is mostly concentrated at about 0.5, that is, existing CDMs can not distinguish learners with different cognitive levels effectively. However, the average proficiency of ICD-A and ICD+ distinguish learners well.

To quantify the distinguishing ability of CDMs for learners with different cognitive levels, this paper proposes the

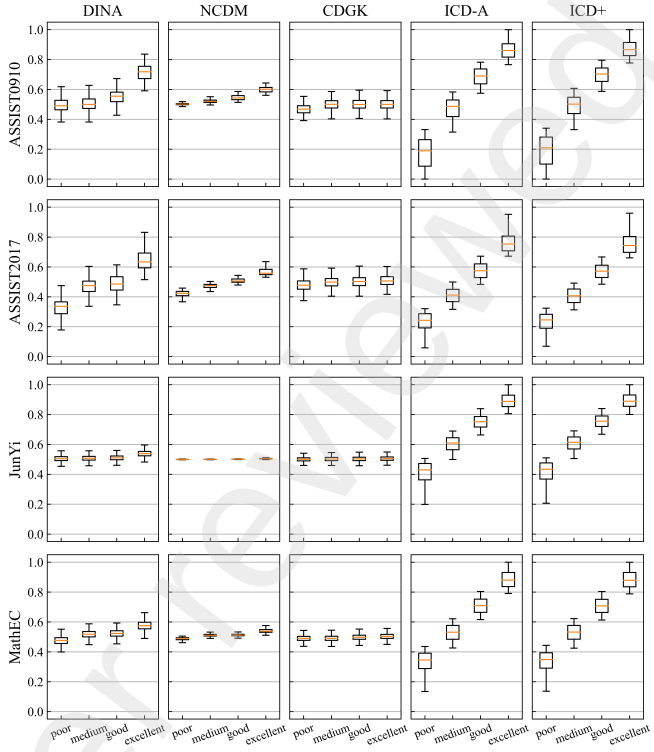

Figure 9: Clustering results of CDMs on four real datasets. The red line is the median, the black line is the boundary.

Degree of Distinguishing (DOD), see eq. (21). Intuitively, the larger the DOD, the stronger the model's ability to distinguish learners with different cognitive levels.

$$DOD = \frac{1}{\bar{r} C_n^2} \sum_{v=1}^{n-1} \sum_{l=v+1}^{n} \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \left| \bar{a}_k^{(v)} - \bar{a}_k^{(l)} \right|, \tag{21}$$

where, n represents the number of clusters,

$$\bar{r} = \frac{1}{\sum_{i=1}^{\mathcal{N}} |R_i|} \sum_{i=1}^{\mathcal{N}} \sum_{r_{ij} \in R_i} r_{ij} \quad \bar{a}_k^{(v)} = \frac{1}{|U_v|} \sum_{u_i \in U_v} a_{ik}, \tag{22}$$

scalar $\bar{a}_k^{(v)}$ represents the average proficiency of learners in U_v on the concept c_k, scalar $\bar{r}$ represents the average score on answered exercises of all learners. The introduction of $\bar{r}$ is to balance the uneven number of correct and incorrect answers in different datasets. Scalar C_n^2 represents the number of combinations of two elements arbitrarily extracted from n elements.

As shown in Table 4, ICD-A performs best on DOD, followed by ICD+. This indicates that the models proposed in this paper have the strongest ability to distinguish learners with different cognitive levels. The DOD of ICD-A is higher than ICD+ in most cases, but the gap is small, indicating that the introduction of potential unknown abilities will affect model's distinguishable ability, but the affect is very small. It is worth noting that the DOD of NCDM and CDGK are not

Table 4
DOD of CDMs. The larger the DOD, the stronger the CDM's ability to distinguish learners with different cognitive levels.

CDMs	ASSIST0910	ASSIST2017	JunYi	MathCE
DINA	0.2316	0.4288	0.0423	0.0935
NCDM	0.0906	0.1863	0.0046	0.0482
CDGK	0.2108	0.3763	0.2186	0.1803
ICD-A	**0.5754**	0.6845	**0.3538**	**0.4764**
ICD+	0.5714	**0.6851**	0.3512	0.4733

always higher than DINA, indicating that there is still room for improvement in distinguishing.

Comparing different datasets, this paper notices that when the proportion of concepts covered by per learner in the datasets is larger (see Table 2), the DOD is usually larger, which denotes that the factor of datasets itself will also affect the distinguishable ability of CDMs. Therefore, to better provide learners with personalized learning services, ITS should encourage learners to practice exercises set that can cover more concepts.

6. Conclusion

This paper proposes a new interpretable cognitive diagnosis model named ICD, which can not only fully utilize the interaction among knowledge concepts but also mine the quantitative relation between exercises and concepts. At the same time, ICD introduces learners' potential unknown abilities to derive an extended version of ICD named ICD+. The original version of ICD is denoted as ICD-A. The definition of degree of distinguishing (DOD) is proposed to quantify the distinguishing ability of CDMs for learners with different cognitive levels. Finally, the paper compares the performance of ICD-A, ICD+ with three latest state-of-the-art CDMs, and two classical CDMs on four real publicity datasets. The results show that both the performance and interpretability of ICD-A and ICD+ are the best.

CRediT authorship contribution statement

Tianlong Qi: Visualization, Conceptualization, Methodology, Software and Experiment, Data curation, Writing original draft, Writing review & editing. **Meirui Ren:** Supervision, Formal analysis, Validation, Project administration, Experimental report, Writing original draft, Writing review & editing. **Longjiang Guo:** Funding acquisition, Supervision, Formal analysis, Conceptualization, Experimental design, Project administration, Investigation, Writing original draft, Writing review & editing. **Xiaokun Li:** Funding acquisition, Validation, Writing – review & editing, Resources. **Jin Li:** Validation, Heuristic Discussion and Writing – review & editing. **Lichen Zhang:** Funding acquisition, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relations that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China under Grant No.61977044 and 62077035; The Second Batch of New Engineering Research and Practice Projects of the Ministry of Education of China under Grant No.E-RGZN20201045; the Natural Science Basis Research Plan in Shaanxi Province of China under Grant No. 2020JM-302 and 2020JM-303; the Key R&D Program of Shaanxi Province Grant No. 2020ZDLGY10-05.

References

de Almeida, P. R. L., Alves, J. H., Parpinelli, R. S., & Barddal, J. P. (2022). A systematic review on computer vision-based parking lot management applied on public datasets. *Expert Systems with Applications*, *198*, 116731. doi:https://doi.org/10.1016/j.eswa.2022.116731.

Cantabella, M., Martínez-España, R., Ayuso, B., Yáñez, J. A., & Muñoz, A. (2019). Analysis of student behavior in learning management systems through a big data framework. *Future Generation Computer Systems*, *90*, 262–272. doi:10.1016/j.future.2018.08.003.

Castro-Schez, J. J., Glez-Morcillo, C., Albusac, J., & Vallejo, D. (2021). An intelligent tutoring system for supporting active learning: A case study on predictive parsing learning. *Information Sciences*, *544*, 446–468. doi:10.1016/j.ins.2020.08.079.

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., Ma, H., & Hu, G. (2019). DIRT: deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 2397–2400). ACM. doi:10.1145/3357384.3358070.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, *26*, 979–1030.

Ellis, H. C. (1965). *The transfer of learning*. Macmillan.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*, 130–160. doi:10.3102/0091732X20903304.

Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. *Future Gener. Comput. Syst.*, *126*, 252–262. doi:10.1016/j.future.2021.08.019.

Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M., Ma, J., Wang, S., & Su, Y. (2021). RCD: relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 501–510). ACM. doi:10.1145/3404835.3462932.

Grubišić, A., Žitko, B., Gašpar, A., Vasić, D., & Dodaj, A. (2022). Evaluation of split-and-rephrase output of the knowledge extraction tool in the intelligent tutoring system. *Expert Systems with Applications*, *187*, 115900. doi:https://doi.org/10.1016/j.eswa.2021.115900.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. University of Illinois at Urbana-Champaign.

Huang, J., Liu, Q., Wang, F., Huang, Z., Fang, S., Wu, R., Chen, E., Su, Y., & Wang, S. (2021). Group-level cognitive diagnosis: A multi-task learning perspective. In *Proceedings of IEEE International Conference*

on Data Mining (ICDM) (pp. 210–219). IEEE. doi:10.1109/ICDM51629.2021.00031.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272. doi:10.1177/01466210122032064.

Kamii, C. (1986). The equilibration of cognitive structures: the central problem of intellectual development. *American Journal of Education*, *94*, 574–577.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*, 451–461. doi:10.1016/S0031-3203(02)00060-2.

Liu, Q. (2021). Towards a new generation of cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 4961–4964). ijcai.org. doi:10.24963/ijcai.2021/703.

Liu, Q., Wu, R.-z., Chen, E., Xu, G., Su, Y., Chen, Z., & Hu, G. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Trans. Intell. Syst. Technol.*, *9*, 48:1–48:26. doi:10.1145/3168361.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Nabizadeh, A. H., Leal, J. P., Rafsanjani, H. N., & Shah, R. R. (2020). Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications*, *159*, 113596. doi:https://doi.org/10.1016/j.eswa.2020.113596.

Nilashi, M., Minaei-Bidgoli, B., Alghamdi, A., Alrizq, M., Alghamdi, O., Khan Nayer, F., Aljehane, N. O., Khosravi, A., & Mohd, S. (2022a). Knowledge discovery for course choice decision in massive open online courses using machine learning approaches. *Expert Systems with Applications*, *199*, 117092. doi:https://doi.org/10.1016/j.eswa.2022.117092.

Nilashi, M., Minaei-Bidgoli, B., Alghamdi, A., Alrizq, M., Alghamdi, O., Khan Nayer, F., Aljehane, N. O., Khosravi, A., & Mohd, S. (2022b). Knowledge discovery for course choice decision in massive open online courses using machine learning approaches. *Expert Systems with Applications*, *199*, 117092. doi:10.1016/j.eswa.2022.117092.

Pinar, W. F., Reynolds, W. M., Taubman, P. M., & Slattery, P. (1995). *Understanding curriculum: An introduction to the study of historical and contemporary curriculum discourses* volume 17. Peter Lang.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional Item Response Theory* (pp. 79–112). Springer. doi:10.1007/978-0-387-89976-3_4.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373. doi:10.1177/014662169101500407.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425–435. doi:10.1007/BF02306030.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*, 287. doi:10.1037/1082-989X.11.3.287.

Tong, S., Liu, Q., Yu, R., Huang, W., Huang, Z., Pardos, Z. A., & Jiang, W. (2021). Item response ranking for cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1750–1756). ijcai.org. doi:10.24963/ijcai.2021/241.

de la Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, *34*, 115–130. doi:10.3102/1076998607309474.

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z., & Wang, S. (2020a). Neural cognitive diagnosis for intelligent education systems. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence* (pp. 6153–6161). AAAI Press. doi:10.1609/aaai.v34i04.6080.

Wang, X., Huang, C., Cai, J., & Chen, L. (2021a). Using knowledge concept aggregation towards accurate cognitive diagnosis. In *Proceedings of The 30th ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 2010–2019). ACM. doi:10.1145/3459637.3482311.

Wang, X., Kou, L., Sugumaran, V., Luo, X., & Zhang, H. (2021b). Emotion correlation mining through deep learning models on natural language text. *IEEE Transactions on Cybernetics*, *51*, 4400–4413. doi:10.1109/TCYB.2020.2987064.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2020b). Diagnostic Questions: The NeurIPS 2020 Education Challenge. *CoRR*, *abs/2007.12061*.

Wei, W., Gu, H., Deng, W., Xiao, Z., & Ren, X. (2022). Abl-tc: A lightweight design for network traffic classification empowered by deep learning. *Neurocomputing*, *489*, 333–344. doi:https://doi.org/10.1016/j.neucom.2022.03.007.

Wu, R., Liu, Q., Liu, Y., Chen, E., Su, Y., Chen, Z., & Hu, G. (2015). Cognitive Modelling for Predicting Examinee Performance. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1017–1024). AAAI Press.

Yeung, C. (2019). Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *CoRR*, *abs/1904.11738*. arXiv:1904.11738.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 262–286.

Zhang, J., Mo, Y., Chen, C., & He, X. (2021). GKT-CD: make cognitive diagnosis model enhanced by graph-based knowledge tracing. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. doi:10.1109/IJCNN52387.2021.9533298.

Zhou, Y., Liu, Q., Wu, J., Wang, F., Huang, Z., Tong, W., Xiong, H., Chen, E., & Ma, J. (2021). Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 2420–2428). ACM. doi:10.1145/3447548.3467264.