Symbolic Cognitive Diagnosis via Hybrid Optimization for Intelligent Education Systems

Junhao Shen, Hong Qian*, Wei Zhang, Aimin Zhou

Shanghai Institute of AI for Education and School of Computer Science and Technology, East China Normal University,
Shanghai 200062, China
shenjh@stu.ecnu.edu.cn, {hqian, amzhou}@cs.ecnu.edu.cn, zhangwei.thu2011@gmail.com

Abstract

Cognitive diagnosis assessment is a fundamental and crucial task for student learning. It models the student-exercise interaction, and discovers the students' proficiency levels on each knowledge attribute. In real-world intelligent education systems, generalization and interpretability of cognitive diagnosis methods are of equal importance. However, most existing methods can hardly make the best of both worlds due to the complicated student-exercise interaction. To this end, this paper proposes a symbolic cognitive diagnosis (SCD) framework to simultaneously enhance generalization and interpretability. The SCD framework incorporates the symbolic tree to explicably represent the complicated student-exercise interaction function, and utilizes gradient-based optimization methods to effectively learn the student and exercise parameters. Meanwhile, the accompanying challenge is that we need to tunnel the discrete symbolic representation and continuous parameter optimization. To address this challenge, we propose to hybridly optimize the representation and parameters in an alternating manner. To fulfill SCD, it alternately learns the symbolic tree by derivative-free genetic programming and learns the student and exercise parameters via gradient-based Adam. The extensive experimental results on various real-world datasets show the superiority of SCD on both generalization and interpretability. The ablation study verifies the efficacy of each ingredient in SCD, and the case study explicitly showcases how the interpretable ability of SCD works.

Introduction

Cognitive diagnosis assessment (CDA) (Liu 2021; Liu et al. 2023b) is a fundamental task in intelligence education systems (Anderson et al. 2014; Burns et al. 2014) with plenty applications, such as educational recommendation systems (Xu and Zhou 2020) and exercise design (Jeckeln et al. 2021). An illustrative example of CDA is shown in Figure I of Appendix. There are three main factors in CDA: students, exercises and knowledge attributes which are also referred to as skills (e.g., calculation). The purpose of CDA is to model the student-exercise interaction via an interaction function based on response logs, and diagnose the students' cognitive states, i.e., inferring the proficiency levels on knowledge attributes.

In the past decades, extensive efforts have been dedicated to developing CDA methods. To name a few, item response theory (IRT) (Lord 1952), multidimensional IRT (MIRT) (Sympson 1978), deterministic inputs, noisy and gate model (DINA) (De La Torre 2008), neural cognitive diagnosis model (NCDM) (Wang et al. 2020a), knowledge-association neural cognitive diagnosis model (KaNCD) (Wang et al. 2022), and Q-augmented causal cognitive diagnosis model (QCCDM) (Liu et al. 2023a). Although these cognitive diagnosis models (CDMs) have achieved remarkable progress, applying them to educational scenarios still suffers from the dilemma of generalization and interpretability. In real-world education scenarios, generalization (e.g., prediction of students' responses) and interpretability (e.g., interaction function and proficiency levels) are of equal importance for evaluating a cognitive diagnosis method (Khosravi et al. 2022). The procedure and outcome of a CDM should be comprehensible and trustworthy for users such as students, teachers and parents. However, due to the complicated student-exercise interaction (DiBello, Roussos, and Stout 2006), most existing methods can hardly make the best of both worlds.

Specifically, on the basis of the $\mathrm{sigmoid}$ interaction function, models like IRT and MIRT exhibit the interpretability of interaction functions and relatively good generalization, but the diagnostic outcomes (e.g., proficiency levels and exercise difficulty parameters) lack interpretability due to the latent vectors vaguely corresponding to each knowledge attribute (Embretson and Reise 2013). Under the conjunctive assumption, models like DINA possess the interpretability of both interaction function and outcomes, but they may underperform in terms of generalization due to their simple forms of interaction function (De La Torre 2011). Through neural networks, models like NCDM show the strong generalization and interpretability of outcomes, but they could lack interpretability of the interaction function due to the black-box nature of neural networks (Murdoch et al. 2019; Du, Liu, and Hu 2020). Most existing methods struggle to well balance the generalization and interpretability mainly because of the dilemma of accurately modelling the complex interaction function in a non-linear way and its intelligibility. To alleviate this dilemma, a highly non-linear and explicable representation of interaction function, symbolic regression (SR), could provide a good recipe to make the best of both worlds. SR excels in finding a complicated non-linear function with high

[*] Corresponding Author: hqian@cs.ecnu.edu.cn.

interpretability thanks to the tree-structured expression.

Unfortunately, directly adapting SR to CDA is infeasible due to the following issues arising from education. Firstly, although SR excellently performs in regression tasks (Billard and Diday 2002), CDA is not only a regression problem but also a complex task that requires simultaneously obtaining the interaction function and diagnostic outcomes. Secondly, the interaction function achieved via SR could not satisfy the monotonicity assumption (Reckase 2009) which is vital common sense in education.

To this end, this paper proposes a symbolic cognitive diagnosis (SCD) framework to simultaneously boost interpretability of outcomes and interaction functions, while maintaining competitive generalization performance. The SCD framework employs the symbolic tree to explicably represent complicated student-exercise interaction functions, and utilizes gradient-based optimization techniques for effective learning of student and exercise parameters. At the same time, the accompanying challenge arises in reconciling discrete symbolic representation learning with continuous parameter optimization to enable model training. To address this challenge, we propose to hybridly optimize the representation and parameters in an alternating manner. To fulfill SCD, it alternately learns the symbolic tree by derivative-free genetic programming (Poli, Langdon, and McPhee 2008) and learns the student and exercise parameters via gradient-based Adam (Kingma and Ba 2015), resulting in the SCD model (SCDM). Specifically, preliminary student and exercise parameters are obtained via optimizing the manually designed initial interaction function. Then, these diagnostic outcomes are fixed to optimize the interaction function via symbolic regression. Afterwards, this complex interaction function is fixed for optimizing the parameters to obtain new diagnostic outcomes, and this process is repeated alternately. To satisfy the monotonicity assumption in education, the function set of symbolic regression only includes monotonic operators and the forms of trees are also subject to the constraints. The extensive experiment results on various real-world datasets show the excellence of SCD on both generalization and interpretability. The ablation study verifies the efficacy of each components of SCD. The case study explicitly showcases how the interpretable ability of SCD works in education scenarios, especially the interpretable form of the learned interaction function.

In the subsequent sections, we respectively recap the related work, introduce the preliminaries, present the proposed SCD, show the experimental results and analysis and finally conclude the paper.

Related Work

Cognitive Diagnosis Assessment. CDA is one of the most important research areas in educational psychology, with many representative CDMs. IRT (Lord 1952) models the student' response to a exercise as an outcome of the interaction between the latent traits of student and the traits of exercise. Specifically, the sigmoid function containing the traits gives the probability of student correctly answering the exercise. In the following decades, these artificially designed interaction functions, such as MIRT (Sympson 1978; Reckase 2009) and DINA (De La Torre 2008), have demonstrated increasing performance in CDA through parameter expansions (Fischer 1995) and increased dimensions (Sympson 1978; Chalmers 2012). However, the weakness also becomes obvious: the manually designed interaction functions hardly fully capture the complex student-exercise relationship (Reckase 2009).

Therefore, CDMs based on artificial neural networks have emerged. Given that neural networks have been theoretically demonstrated to possess the universal approximation property (Hornik, Stinchcombe, and White 1989), they can be employed to effectively capture the intricate student-exercise interactions in the CDA, replacing the traditional interaction function. Specifically, in NCDM (Wang et al. 2020a), KaNCD (Wang et al. 2022) and QCCDM (Liu et al. 2023a), the neural networks are designed as the multiple full connection layers to capture the complex student-exercise interactions. Nevertheless, interpretability is also of equal importance in education scenarios (Khosravi et al. 2022), leading to the insufficiency of those black-box neural networks.

Symbolic Regression. Symbolic regression (SR) is an effective approach to finding a suitable mathematical model to describe data (Billard and Diday 2002). Compared with traditional regression techniques like linear regression, there are no a priori assumptions on the specific form of the function. In other words, SR explores through the mathematical expression spanned by candidate mathematical operators to discover the most suitable solution (Zhang et al. 2022). SR plays a significant role in data processing and analysis across various fields due to its interpretability, e.g., dynamical systems prediction (Quade et al. 2016). There are also some educational applications like academic performance prediction (Ouyang et al. 2023). However, adapting SR to CDA remains challenging since CDA is not merely a prediction task. Currently, to the best of our knowledge, relevant work that adequately applies SR to CDA seldom exists. In this paper, we propose a SCD framework which balances the generalization and interpretability of both interaction function and diagnostic outcomes, especially the interpretable form of the learned interaction function.

Preliminaries

Task Overview. We at first introduce some notations and definitions in cognitive diagnosis. Let $S = \{s_1, \dots, s_N\}$, $E = \{e_1, \dots, e_M\}$ and $K = \{k_1, \dots, k_L\}$ respectively denote the sets of students, exercises and knowledge attributes, where $N = |S|$, $M = |E|$ and $L = |K|$ are the size of each set. Each student is required to finish some exercises for practice, and the corresponding response logs are denoted as a set of triplets $R = \{(s, e, r) | s \in S, e \in E, r \in \{0, 1\}\}$, where r represents the score (i.e., $r = 1$ means right while $r = 0$ means wrong) obtained by student s on exercise e. Besides, the Q-matrix (usually annotated by experts) is denoted as $\boldsymbol{Q} = \{Q_{i,j}\}_{M \times L}$, which captures the relationship between exercises and knowledge attributes. Herein, $Q_{i,j} = 1$ if exercise e_i is associated with knowledge attribute k_j, and $Q_{i,j} = 0$ otherwise. Furthermore, the monotonicity assumption (Reckase 2009), as Assumption 1, is introduced to enhance the interpretability of diagnostic outcomes.

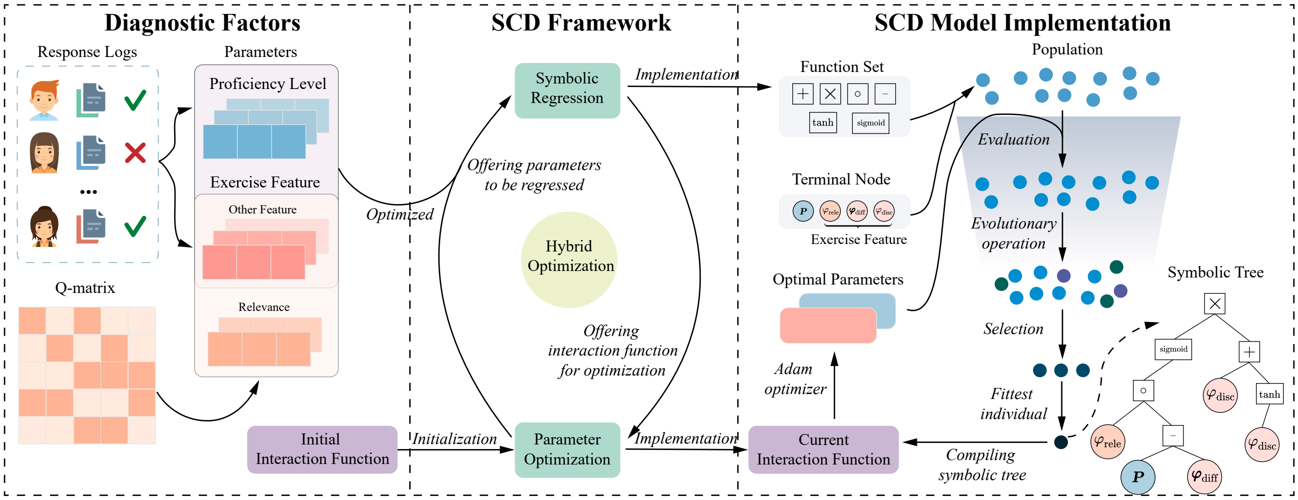


Figure 1: An overview of the proposed symbolic cognitive diagnosis (SCD)

Assumption 1 (Monotonicity Assumption) *The probability of correctly answering an exercise is monotonically increasing at any dimension of the student's proficiency level on relevant knowledge attributes.*

Problem Definition. Given the observed triplet logs of students R and the labelled Q-matrix $\boldsymbol{Q}$, the goal is to infer the students' proficiency on knowledge attributes with the process of simultaneously estimating the interaction function and proficiency level, where the interaction function models the student-exercise interaction.

Genetic Programming. Genetic programming (GP) (Koza 1994; Mei et al. 2023) stands out as one of the most prevalent methods used in SR. It aims to evolve solutions to a given problem following Darwin's theory of evolution, seeking the fittest solution over several generations. Instead of using binary code to represent chromosomes as in genetic algorithms (Forrest 1993), solutions in GP are represented as tree-structured chromosomes containing nodes and terminals, forming a symbolic tree. The tree comprises interior nodes denoting mathematical operators and terminal nodes denoting variables. Employing the depth-first search, the mathematical expression for each individual solution can be obtained through traversing the tree. The GP procedure will be clarified in the SCD Model (SCDM) Implementation Section.

Symbolic Cognitive Diagnosis (SCD)

This section presents the key ingredients of SCD. The overall framework of SCD is shown in Figure 1. Sequentially, we introduce the basic diagnostic factors of cognitive diagnosis in the SCD framework, elaborate the procedure of SCD framework, fulfill the SCD framework with GP and Adam optimizer to result in the proposed SCD model (SCDM), and finally discuss the flexibility of the SCD framework.

Diagnostic Factors

Generally, there are three factors in CDA need to be diagnosed and modeled, i.e., proficiency levels, exercise features and interaction function (DiBello, Roussos, and Stout 2006). Details are introduced as bellow.

Proficiency Levels. We aim to assess students' proficiency levels directly, whose each dimension corresponds to a specific knowledge attribute, without utilizing latent vectors like IRT or MIRT. Specifically, proficiency levels are denoted as $\boldsymbol{P} = \{P_{i,j}\} \in \mathbb{R}^{N \times L}$, and the i-th student's proficiency level is $\boldsymbol{P}_{i,\bullet}$, each entry of which is continuous among $[0, 1]$ and indicates the student's proficiency on a certain knowledge attribute. For instance, $\boldsymbol{P}_{i,\bullet} = [0.2, 0.5, 0.8]$ means the i-th student has a low mastery on the first knowledge attribute, middle on the second, and high on the last. Proficiency levels are deemed as the student parameter and learnt during the continuous parameter optimization module.

Exercise Features. Exercise features refer to the characteristics of each exercise. We divide exercise features into two categories. The first involves the relevance between the j-th exercise and knowledge attributes, and is denoted as $\boldsymbol{\varphi}_{\text{rele}_j} \in \{0, 1\}^L$, which is directly from the Q-matrix $\boldsymbol{Q}$ and not trainable during the parameter optimization. The second involves other optional trainable exercise parameters such as exercise difficulty and discrimination, and they can be included if necessary.

As shown in the left of Figure 1, proficiency levels and other exercise features are deemed as parameters, which are learnt during the parameter optimization.

Interaction Function. The interaction function models the process of students completing exercises and getting the response results. In the SCD, we utilize SR to obtain the interaction function for several reasons. Firstly, SR effectively captures non-linear and complex student-exercise relationship since it does not rely on a predefined functional form. Secondly, the result of model is illustrated by the symbolic tree as Figure 1, which exhibits high interpretability. Thirdly, the function set and terminal nodes can be expended to accommodate various real-world educational tasks. Formally, the output of SCDM can be formulated as

$$y_f = \sigma(f(\boldsymbol{P}_{i,\bullet}, \boldsymbol{\varphi}_{\text{rele}_j}, \boldsymbol{\varphi}_{\text{others}_j})) \,, \tag{1}$$

where y_f is the probability of the i-th student correctly answering the j-th exercise given by interaction function f compiled from symbolic tree, σ denotes the activate function, and $\boldsymbol{\varphi}_{\text{others}}$ denotes other factors except $\boldsymbol{P}$ and $\boldsymbol{\varphi}_{\text{rele}}$.

However, the symbolic tree constructed by a unlimited function set in Eq. (1) hardly meets Assumption 1. Thus, we need to design tailored strategies to address this problem.

Symbolic Cognitive Diagnosis Framework

This section introduces the proposed SCD framework shown as the middle of Figure 1. This framework can be divided into two modules: parameter optimization (PO) and SR. Due to the inherent differences between continuous and discrete optimization processes, they usually cannot be synchronously combined. Thus, the hybrid optimization is incorporated to asynchronously unify the continuous and discrete optimization processes and begins with PO.

Parameter Optimization. The reason for choosing to start with PO is that the quality of initialization significantly influences the model's training performance, and its initialization is easier compared with SR. Specifically, although both the student-exercise interaction function and parameters of students and exercises are unknown, manually designed simple interaction functions like sigmoid have been shown to effectively approximate for an unobserved interaction function in the field of education (Lord 1952; Sympson 1978; Reckase 2009). Conversely, the parameters lack suitable estimates, and random initialization is likely to lead symbolic regression astray. After initializing the interaction function in the PO module, the objective function is $\mathcal{L}_{r\in R}(y_f, r)$, where the $\mathcal{L}$ is loss function, r is the label from R, and y_f is defined in Eq. (1). After parameter optimization, the optimal parameters are proficiency levels and trainable exercise features (e.g., difficulty) and sent to SR.

Symbolic Regression. After gaining the optimal parameters, SR can discover the potential interaction function between students and exercises. This module does not have access to the interaction function in PO module. Some may ask: if SR directly finds the interaction function identical to the PO module, how do we discover more complex interaction relationships? Note that there are complexity requirements of SR. Typically, we aim to identify a suitable function as the regression result from complex candidates, which often outperforms simple ones in generalization (validated in ablation study). After regression, the optimal function is sent to the PO module as the current interaction function.

SCD Model (SCDM) Implementation

This section introduces the implementation of the SCD framework as the right of Figure 1. The PO is implemented by the Adam (Kingma and Ba 2015), SR by the GP (Billard and Diday 2002), and hybrid optimization by alternative optimization (AO). The SCDM implementation can be divided into two phases: continuous optimization and discrete optimization. Algorithm 1 shows the implementation of SCDM.

Algorithm 1: Symbolic Cognitive Diagnosis Model (SCDM)

Input: Response logs $R = \{s, e, r\}$, Q-matrix: $\boldsymbol{Q}$, and initial interaction function f_{init}.
Parameter: Maximum number of epochs T, generation T_{GP}, population size V, crossover rate p_{cr}, and mutation rate p_{mu}.
Output: Proficiency levels $\boldsymbol{P}$, exercise features $\boldsymbol{\varphi}_{\text{diff}}, \varphi_{\text{disc}}$, and fittest interaction function $\hat{f}$.

```
1: Initialize the $\boldsymbol{P}, \boldsymbol{\varphi}_{\text{diff}}, \varphi_{\text{disc}}$; $\hat{f} \leftarrow f_{\text{init}}, \boldsymbol{\varphi}_{\text{rele}_j} \leftarrow \boldsymbol{Q}_{j,\bullet}$
2: for $t = 1, 2, \ldots, T$ do
3:   $\boldsymbol{P}, \boldsymbol{\varphi}_{\text{diff}}, \varphi_{\text{disc}} \leftarrow \text{Adam}\, \mathcal{L}_{r\in R}(y_{\hat{f}}, r)$
4:   Randomly initialize a population of interaction functions $F = \{f_1, f_2, \ldots, f_V\}$
5:   for $t_1 = 1, 2, \ldots, T_{\text{GP}}$ do
6:     for $i = 1, 2, \ldots, V$ do
7:       if $i\%2 = 1$ and $\text{uniform}(0, 1) < p_{\text{cr}}$ then
8:         $\{f_i, f_{i+1}\} \leftarrow$ crossover$(f_i, f_{i+1})$
9:       end if
10:      if $\text{uniform}(0, 1) < p_{\text{mu}}$ then
11:        $f_i \leftarrow$ mutation$(f_i)$
12:      end if
13:    end for
14:    Evaluate all individuals in $F$, $F \leftarrow$ selection$(F)$
15:  end for
16:  $\hat{f} \leftarrow f$, where $f$ is the fittest in $F$
17: end for
18: return $\boldsymbol{P}, \boldsymbol{\varphi}_{\text{diff}}, \varphi_{\text{disc}}, \hat{f}$
```

Continuous Optimization. The continuous optimization for learning the student and exercise parameters $\boldsymbol{P}$, $\boldsymbol{\varphi}_{\text{diff}}$ and φ_{disc} is described in line 3 of Algorithm 1. The initial interaction function is inspired by MIRT and formulated as

$$f_{\text{init}} = \boldsymbol{\varphi}_{\text{rele}_j} \circ (\boldsymbol{P}_{i,\bullet} - \boldsymbol{\varphi}_{\text{diff}_j}) \times \varphi_{\text{disc}_j} \,, \tag{2}$$

where $\circ$ is inner product, $\boldsymbol{\varphi}_{\text{diff}_j} \in \mathbb{R}^L$ is the difficulty of the j-th exercise, and $\varphi_{\text{disc}_j} \in \mathbb{R}$ is the discrimination of the j-th exercise. Eq. (2) is simple and intuitive, but can correctly lead the optimization. The activate function σ is sigmoid which is widely used in CDMs. To use the Adam optimizer, we formulate the loss function of SCDM as the cross entropy between the i-th output of interaction function f denoted by $y_f^{(i)}$ and the corresponding label $r^{(i)}$, which is Eq. (3) below.

$$\mathcal{L}_{r\in R}(y_f, r) = -\sum_{i=1}^{|R|} (r^{(i)} \log y_f^{(i)} + (1 - r^{(i)}) \log(1 - y_f^{(i)})) \,. \tag{3}$$

After training Eq. (3), the parameters $\boldsymbol{P}$, $\boldsymbol{\varphi}_{\text{diff}}$ and φ_{disc} are the current diagnostic outcomes, which are sent to the discrete optimization as the parameters of symbolic regression.

Discrete Optimization. The discrete optimization for learning the symbolic representation tree is shown in line 5 to 15 of Algorithm 1. The procedure begins with an initial population of randomly generated individuals which must adhere to specific criteria (e.g., ensuring the output to be scalar). Through random operators like crossover and mutation, the current population is evolved. Then, each individual's fitness is evaluated by metric like accuracy. Finally, individuals are selected as parents in a certain way, and their offspring become the next generation. This process continues until it reaches the maximum generation. The fittest individual is the new interaction function and sent to continuous optimization. More detailed implementation is depicted in Appendix B.

Discussions. We discuss some points regarding SCD and SCDM. (1) Flexibility. Proficiency levels $\boldsymbol{P}$ and Q-matrix $\boldsymbol{Q}$ are essential in the SCD framework. SCDM includes difficulty and discrimination, and other exercise features can also be included if necessary. The initial interaction function is restricted to including $\boldsymbol{P}_{i,\bullet} \circ \varphi_{\text{rele}_j}$, ensuring each dimension of $\boldsymbol{P}_{i,\bullet}$ corresponding to a specific knowledge attribute. (2) Interpretability. SCD involves interpretability in two aspects: diagnostic outcomes via Assumption 1 and interaction functions guaranteed by GP. GP provides explicit operations and explainable object functions, and SR trees optimized by GP are transparent. (3) Implementation. We choose Adam (Kingma and Ba 2015) for optimization due to its wide use in deep learning, fast convergence, and to ensure fair comparison with existing methods that use it. The SCD framework is generic and can be implemented differently. For example, SR can be realized by other algorithms like transformer (Kamienny et al. 2022), and PO can be realized by evolutionary strategies (Beyer and Schwefel 2002).

Experiments

This section conducts extensive experiments on real-world datasets to answer the following crucial questions. The source code of SCDM is available at GitHub[1].

- **Q1:** How does SCDM perform when compared with existing CDMs in terms of generalization?
- **Q2:** How does SCDM perform when compared with existing CDMs in terms of interpretability?
- **Q3:** How do GP and gradient optimization contribute to the performance of SCDM respectively?
- **Q4:** How do hyperparameters influence SCDM?
- **Q5:** How does the interpretable ability of SCDM work in real-world educational scenarios?

Experimental Setup

Dataset Description. The experiments are conducted on four real-world datasets, i.e., Math1, Math2 (Liu et al. 2018), FracSub (Wu et al. 2015) and NeurIPS2020 (Wang et al. 2020b). Math1 and Math2 consist of response logs from high school students taking the final exams of their first and second senior years respectively. FracSub comprises of scores of middle school students on fraction subtraction objective problems. NeurIPS2020 containing two school years of students' answers to mathematics questions from Eedi. Details of these datasets are shown in Table 1.

Baselines and State-of-the-Art Methods. Over the past few decades, CDMs have been developing, and here we select some representative approaches for comparison.

- IRT (Lord 1952) is a classical CDM that employs a logistic-like interaction function.
- MIRT (Sympson 1978) is an extended model of IRT, which uses multidimensional $\boldsymbol{\theta}$ and $\boldsymbol{b}$ to model the latent traits of students and exercises.

Table 1: Statistics of real-world datasets for experiments.

Datasets	Math1	Math2	FracSub	NeurIPS2020
#Students	4209	3911	536	4129
#Exercises	15	16	20	44
#Knowledge Attributes	11	16	8	30
#Response Logs	63135	62576	10720	66638
Average Correct Rate	0.5515	0.4880	0.5339	0.5450

- DINA (De La Torre 2008) is a CDM based on the conjunctive assumption, where proficiency levels are represented by two discrete values, i.e., 0 and 1.
- NCDM (Wang et al. 2020a) is a CDM that replaces the traditional interactive function with neural networks, which outperforms the traditional CDA methods on most datasets with interpretability of the diagnostic outcomes.
- KaNCD (Wang et al. 2022) is a CDM based on the improvements of NCDM, which considers the implicit association between knowledge attributes, and reaches the state-of-the-art on most datasets.

Generalization Metrics. Assessing the performance of CDM can be challenging due to the difficulty in obtaining accurate proficiency levels of students. To overcome this, a widely accepted approach to evaluating them is through the prediction of students' test scores. Accordingly, similar to previous CDMs (Wang et al. 2020a), we assess how close the model's prediction (whether a student solves a question or not) is to the ground truth in the test set with classification metrics, i.e., accuracy, area under curve (AUC) and F1-score (F1). Besides, in the field of SR, the R^2 score is usually employed for evaluating symbolic trees (Zhang, Zhou, and Zhang 2022). As aforementioned, however, the true student-exercise interaction function is unknown, rendering this metric unavailable.

Interpretability Metric. Generalization metrics are only one aspect of evaluating the performance of CDMs. On the other hand, interpretability metrics hold equal significance in the education scenario. The interpretability of interaction functions is often assessed by the choice of function set and tree depth. But there are no other CDMs that utilize SR, we solely quantitatively assess the interpretability of the diagnostic outcomes via the degree of agreement (DOA) (Wang et al. 2020a, 2022). Intuitively, if student s_a has a greater accuracy in answering exercises related to k_i than student s_b, the proficiency level of s_a should be higher than that of student s_b on the knowledge attribute k_i, i.e., $\boldsymbol{P}_{a,i} > \boldsymbol{P}_{b,i}$. The DOA of k_i is defined as Eq. (4).

$$DOA(i) = \frac{1}{Z} \sum_{a=1}^{N} \sum_{b=1}^{N} \delta(\boldsymbol{P}_{a,i}, \boldsymbol{P}_{b,i}) \sum_{j=1}^{M} I_{j,i} \cdot \frac{J(j,a,b) \wedge \delta(r_{aj}, r_{bj})}{J(j,a,b)}, \quad (4)$$

where $Z = \sum_{a=1}^{N} \sum_{b=1}^{N} \delta(\boldsymbol{P}_{a,i}, \boldsymbol{P}_{b,i})$, and $\boldsymbol{P}_{a,i}$ is the proficiency level of student s_a on knowledge attributes k_i. $\delta(x, y) = 1$ if $x > y$ and otherwise $\delta(x, y) = 0$. $I_{ji} = 1$ if exercise e_j contains knowledge attribute k_i and otherwise $I_{ji} = 0$. $J(j, a, b) = 1$ if both student s_a and s_b completed exercise e_j and otherwise $J(j, a, b) = 0$. To evaluate the interpretability of diagnostic outcomes, we average the $DOA(i)$ on all knowledge attributes.

[1] https://github.com/shinkungoo/SymbolicCDM

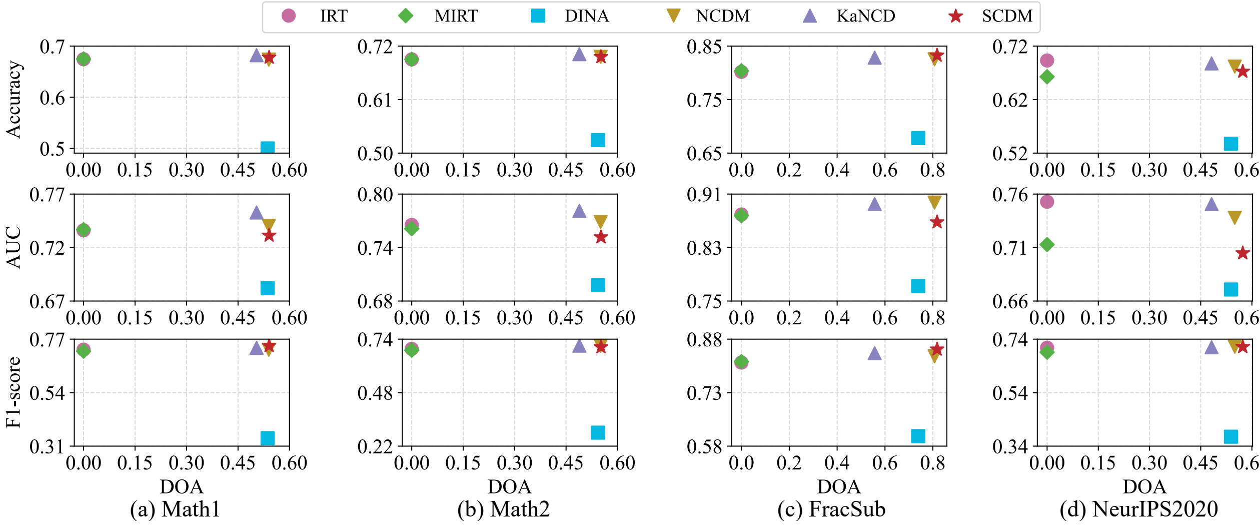


Figure 2: The Pareto performance of generalization and interpretability of SCDM and the compared methods

Detailed Settings. In the GP module implemented by DEAP (De Rainville et al. 2012), the population size V is 200, the number of generation T_{GP} is 10, the crossover and mutation rates are 0.5 and 0.1 respectively, the initial tree depth is 5, and the selection method is tournament selection. To meet the Assumption 1, the function set is $\{+, \times, \circ, -, \tanh, \mathrm{sigmoid}\}$. Most of the setting is referred to (Zhang, Zhou, and Zhang 2022), which shows to be effective in SR. In the continuous optimization module implemented by Pytorch (Paszke et al. 2019), we set the learning rate of Adam (Kingma and Ba 2015) to be 0.002, and initialize the feature parameters in the interaction function with Xavier normal initialization (Glorot and Bengio 2010). To evaluate performance, the size of test dataset is 0.2, and all experiments are repeated independently with 10 seeds.

Experimental Results

We conduct comprehensive experiments to answer the aforementioned questions. More details are in Appendix C.

Generalization Performance (Q1). As shown in Figure 2, SCDM performs competitively with the existing CDMs on each metric, without being dominated in any of them. It even outperforms other CDMs in terms of the F1-score. In fact, in the field of machine learning, enhancing model interpretability may weaken the generalization performance to a certain extent (Du, Liu, and Hu 2020). DINA is an extreme case, with a high DOA but inadequate generalization performance. Therefore, demanding an improvement in generalization while increasing interpretability is quite challenging. Hence, SCDM can achieve improved interpretability of interaction function and diagnostic outcomes while maintaining strong generalization, indicating a well-balanced trade-off between these two aspects.

Interpretability Performance (Q2). We assess the interpretability of diagnostic outcomes. As shown in Figure 2, the SCDM performs competitively or even outperforms all existing CDMs in terms of DOA, which shows SR also helps improve DOA. Notably, due to the utilization of latent vectors in IRT and MIRT, there does not exist an explicit correspondence between dimensions of latent vectors and knowledge attributes. Thus, we consider their DOA to be 0 in Figure 2.

Ablation Study (Q3). In order to comprehend the impact of the GP and Adam components on SCDM's performance (expressed in percentage), an ablation study was conducted. The results shown in Table 2, where the "SCDM w.o. GP" indicates SCDM solely employing the initial interaction function. "SCDM w.o. Adam" signifies the utilization of a derivative-free evolutionary strategy (see Appendix C.2 for details) instead of the gradient-based Adam. By t-test, SCDM is significantly better than them on all the performance metrics with significant level $\alpha = 5\%$, and the variance of each metric is less than 0.05, revealing that both components have a positive impact on the performance. Only using manually designed interaction functions may reduce SCDM's performance, but updating interaction functions with GP improves learning parameters, resulting in a performance enhancement. Besides, compared with the evolutionary strategy, the Adam optimizer makes learning parameters quicker and more accurate since it strictly adheres to the monotonicity assumption, leading to a significant enhancement in performance.

Hyperparameter Analysis (Q4). We conduct a hyperparameter experiment to study the effect of crossover rate, mutation rate, generations and population size on the FracSub dataset whose results are similar to other datasets. Figure 3 illustrates that our hyperparameter settings are good for most metrics. In practice, opting for a lower mutation rate and higher crossover rate promotes evolution, as the latter effectively blends various operators. A larger population size and more generations benefit evolution, but this entails a trade-off between performance and time. Besides, accuracy, F1-score and DOA are relatively insensitive to hyperparameters, whereas AUC shows higher sensitivity to them. Hence, when computational resources allow, opting for a larger population size and generations, coupled with suitable mutation

Table 2: Ablation study of SCDM. "SCDM w.o. GP" refers to the SCDM without using GP, and "SCDM w.o. Adam" refers to the SCDM without using Adam to optimize parameters. In each column, an entry is marked in bold if its mean value is the best. By t-test, SCDM is significantly better than them on all the performance metrics with significant level $\alpha = 5\%$.

	Math1				Math2				FracSub				NeurIPS2020			
	Accuracy	AUC	F1	DOA	Accuracy	AUC	F1	DOA	Accuracy	AUC	F1	DOA	Accuracy	AUC	F1	DOA
SCDM w.o. GP	67.22	72.46	73.95	53.13	69.05	75.07	69.94	54.92	82.67	86.33	84.33	81.16	63.43	65.98	67.50	56.03
SCDM w.o. Adam	67.30	72.86	73.81	50.97	68.43	75.08	68.67	51.58	76.42	83.12	78.62	68.68	63.27	68.70	68.62	51.37
SCDM	**67.78**	**73.13**	**74.14**	**53.75**	**69.79**	**75.17**	**70.15**	**55.08**	**83.26**	**86.80**	**85.15**	**81.78**	**67.25**	**70.48**	**71.11**	**57.32**

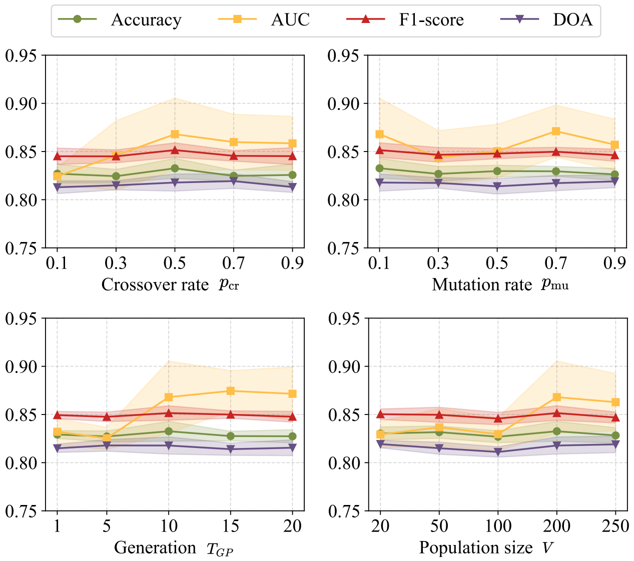
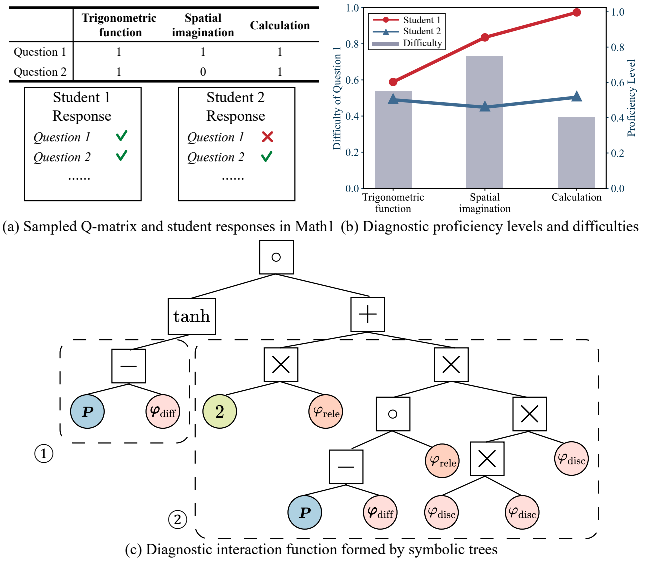


Figure 3: The performance of generalization and interpretability under different p_{cr}, p_{mu}, T_{GP} and V values on FracSub



	Trigonometric function	Spatial imagination	Calculation
Question 1	1	1	1
Question 2	1	0	1

(a) Sampled Q-matrix and student responses in Math1 (b) Diagnostic proficiency levels and difficulties

(c) Diagnostic interaction function formed by symbolic trees

Figure 4: Case study: diagnostic outcomes of two students, exercise features and interaction function

and crossover rates, enhances the SCDM's performance.

Case Study (Q5). Since each obtained symbolic tree is unique and there is no true function available as a standard, it is challenging to quantify the interpretability of the symbolic tree. In this part, we delve into a more detailed analysis of interpretability. Suppose an educator intends to diagnose the cognitive state of students participating in the test of dataset Math1, one can employ SCDM for cognitive diagnosis of the students, and the outcomes are presented in the Figure 4. The bars in (b) represent the difficulty on each knowledge attribute of Question 1, and the lines represent proficiency levels of Student 1 and Student 2 respectively. Figure 4 (b) depicts the necessity for students' proficiency levels to surpass the exercise difficulty levels to answer correctly, which is proved by the response logs in (a). Besides, the interpretability of the interaction function shown in (c) is noteworthy. Specifically, the tree depth is low and the choice of function set is highly interpretable since it satisfies the monotonicity assumption in education. For each part, Part ① in the symbolic tree involves proficiency level calculations, where deducing exercise difficulty from proficiency levels is corresponding to the findings in (a) and (b), and it is reasonable in education. Part ② contains complex computations concerning exercise discrimination and exercise-knowledge relevance. To be specific, the relevance φ_{rele} multiplied by discrimination suggests some underlying interconnections between knowledge attributes. These two parts result in the final interaction function.

Conclusion

This paper aims to simultaneously boost interpretability of outcomes and interaction functions, while maintaining competitive generalization performance. The proposed symbolic cognitive diagnosis incorporates the symbolic tree to explicably represent the complicated student-exercise interaction function and gradient-based optimization methods to effectively learn the student and exercise parameters. To effectively tunnel the discrete symbolic representation and continuous parameter optimization, and fulfill the SCD framework, we propose to hybridly optimize the representation and parameters in an alternating way. SCD possesses the merits of high intelligibility, generalization and flexibility. We sincerely hope this tentative work could pave the way for landing CDA in intelligent education systems. The future work of SCD includes theoretically disclosing the convergence behavior, further enhancing the generalization stability and exploring more intelligent education applications.

Acknowledgments

The authors would like to express our sincere thanks to the anonymous reviewers for their constructive and valuable comments. The authors also would like to thank Hengzhe Zhang for the helpful discussion, and Xinyu Shi for the reliable support. The algorithms and datasets in this paper do not involve any ethical issue. This work is supported by the National Natural Science Foundation of China (62106076), Defense Industrial Technology Development Program (JCKY2019204A007), National Natural Science Foundation of China (92270119), and Natural Science Foundation of Shanghai (21ZR1420300).

References

Anderson, A.; Huttenlocher, D. P.; Kleinberg, J. M.; and Leskovec, J. 2014. Engaging with Massive Online Courses. In *Proceedings of the 23rd International World Wide Web Conference*, 687–698. Seoul, Korea.

Beyer, H.-G.; and Schwefel, H.-P. 2002. Evolution Strategies–A Comprehensive Introduction. *Natural Computing*, 1: 3–52.

Billard, L.; and Diday, E. 2002. Symbolic Regression Analysis. In *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, 281–288. Springer.

Burns, H.; Luckhardt, C. A.; Parlett, J. W.; and Redfield, C. L. 2014. *Intelligent Tutoring Systems: Evolutions in Design*. Psychology Press.

Chalmers, R. P. 2012. MIRT: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48: 1–29.

De La Torre, J. 2008. An Empirically Based Method of Q-matrix Validation for the DINA Model: Development and applications. *Journal of Educational Measurement*, 45(4): 343–362.

De La Torre, J. 2011. The Generalized DINA Model Framework. *Psychometrika*, 76: 179–199.

De Rainville, F.-M.; Fortin, F.-A.; Gardner, M.-A.; Parizeau, M.; and Gagné, C. 2012. Deap: A Python Framework for Evolutionary Algorithms. In *Proceedings of the 14th Annual Conference Companion on Genetic and Evolutionary Computation*, 85–92. philadelphia, PA.

DiBello, L. V.; Roussos, L. A.; and Stout, W. 2006. A Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. *Handbook of Statistics*, 26: 979–1030.

Du, M.; Liu, N.; and Hu, X. 2020. Techniques for interpretable machine learning. *Communication of the ACM*, 63(1): 68–77.

Embretson, S. E.; and Reise, S. P. 2013. *Item Response Theory*. Psychology Press.

Fischer, G. H. 1995. Derivations of the Rasch model. In *Rasch Models: Foundations, Recent Developments, and Applications*, 15–38. Springer.

Forrest, S. 1993. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science*, 261(5123): 872–878.

Glorot, X.; and Bengio, Y. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the 13rd International Conference on Artificial Intelligence and Statistics*, 249–256. Sardinia, Italy.

Hornik, K.; Stinchcombe, M.; and White, H. 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5): 359–366.

Jeckeln, G.; Hu, Y.; Cavazos, J. G.; Yates, A. N.; Hahn, C. A.; Tang, L.; Phillips, P. J.; and O'Toole, A. J. 2021. Face Identification Proficiency Test Designed Using Item Response Theory. *CoRR*, abs/2106.15323.

Kamienny, P.-a.; d'Ascoli, S.; Lample, G.; and Charton, F. 2022. End-to-end Symbolic Regression with Transformers. In *Advances in Neural Information Processing Systems 35*, 10269–10281. New Orleans, LA.

Khosravi, H.; Shum, S. B.; Chen, G.; Conati, C.; Tsai, Y.; Kay, J.; Knight, S.; Martínez Maldonado, R.; Sadiq, S. W.; and Gasevic, D. 2022. Explainable Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 3: 100074.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, CA.

Koza, J. R. 1994. Genetic Programming as a Means for Programming Computers by Natural Selection. *Statistics and Computing*, 4: 87–112.

Liu, Q. 2021. Towards a New Generation of Cognitive Diagnosis. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 4961–4964. Montreal, Canada.

Liu, Q.; Wu, R.; Chen, E.; Xu, G.; Su, Y.; Chen, Z.; and Hu, G. 2018. Fuzzy Cognitive Diagnosis for Modelling Examinee Performance. *ACM Transactions on Intelligent Systems and Technology*, 9(4): 48:1–48:26.

Liu, S.; Qian, H.; Li, M.; and Zhou, A. 2023a. QCCDM: A Q-Augmented Causal Cognitive Diagnosis Model for Student Learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*, 1536–1543. Kraków, Poland.

Liu, Y.; Zhang, T.; Wang, X.; Yu, G.; and Li, T. 2023b. New development of cognitive diagnosis models. *Frontiers of Computer Science*, 17(1): 171604.

Lord, F. 1952. A Theory of Test Scores. *Psychometric Monographs*.

Mei, Y.; Chen, Q.; Lensen, A.; Xue, B.; and Zhang, M. 2023. Explainable Artificial Intelligence by Genetic Programming: A Survey. *IEEE Transactions on Evolutionary Computation*, 27(3): 621–641.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proceedings of the National Academy of Sciences*, 116(44): 22071–22080.

Ouyang, F.; Wu, M.; Zheng, L.; Zhang, L.; and Jiao, P. 2023. Integration of Artificial Intelligence Performance Prediction and Learning Analytics to Improve Student Learning in Online Engineering Course. *International Journal of Educational Technology in Higher Education*, 20: 1–23.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. British Columbia, Canada.

Poli, R.; Langdon, W. B.; and McPhee, N. F. 2008. *A Field Guide to Genetic Programming*. Springer.

Quade, M.; Abel, M.; Shafi, K.; Niven, R. K.; and Noack, B. R. 2016. Prediction of Dynamical Systems by Symbolic Regression. *Physical Review E*, 94(1): 012214.

Reckase, M. D. 2009. *Multidimensional Item Response Theory Models*. Springer.

Sympson, J. B. 1978. A Model for Testing with Multidimensional Items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*, 82–98. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometrics Methods Program.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020a. Neural Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 6153–6161. New York, NY.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Yin, Y.; Wang, S.; and Su, Y. 2022. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8312–8327.

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.; and Zhang, C. 2020b. Diagnostic Questions: The NeurIPS 2020 Education Challenge. *arXiv preprint arXiv:2007.12061*.

Wu, R.; Liu, Q.; Liu, Y.; Chen, E.; Su, Y.; Chen, Z.; and Hu, G. 2015. Cognitive Modelling for Predicting Examinee Performance. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 1017–1024. Buenos Aires, Argentina.

Xu, W.; and Zhou, Y. 2020. Course Video Recommendation with Multimodal Information in Online Learning Platforms: A Deep Learning Framework. *British Journal of Educational Technology*, 51(5): 1734–1747.

Zhang, H.; Zhou, A.; Qian, H.; and Zhang, H. 2022. PS-Tree: A piecewise symbolic regression tree. *Swarm and Evolutionary Computation*, 71: 101061.

Zhang, H.; Zhou, A.; and Zhang, H. 2022. An Evolutionary Forest for Regression. *IEEE Transactions on Evolutionary Computation*, 26(4): 735–749.