HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework

Jiatong Li
School of Data Science, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
satosasara@mail.ustc.edu.cn

Fei Wang
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
wf314159@mail.ustc.edu.cn

Qi Liu
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
qiliuql@ustc.edu.cn

Mengxiao Zhu
Department of Communication of Science and Technology, University of Science and Technology of China
Hefei, China
mxzhu@ustc.edu.cn

Wei Huang
School of Data Science, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
ustc0411@mail.ustc.edu.cn

Zhenya Huang*
School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
huangzhy@ustc.edu.cn

Enhong Chen
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
cheneh@ustc.edu.cn

Yu Su
Hefei Normal University & Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
Hefei, China
yusu@hfnu.edu.cn

Shijin Wang
iFLYTEK AI Research (Central China) & State Key Laboratory of Cognitive Intelligence
Hefei, China
sjwang3@iflytek.com

ABSTRACT

Cognitive diagnostic assessment is a fundamental task in intelligent education, which aims at quantifying students' cognitive level on knowledge attributes. Since there exists learning dependency among knowledge attributes, it is crucial for cognitive diagnosis models (CDMs) to incorporate attribute hierarchy when assessing students. The attribute hierarchy is only explored by a few CDMs such as Attribute Hierarchy Method, and there are still two significant limitations in these methods. First, the time complexity would be unbearable when the number of attributes is large. Second, the assumption used to model the attribute hierarchy is too strong so that it may lose some information of the hierarchy and is not flexible enough to fit all situations. To address these limitations, we propose a novel Bayesian network-based Hierarchical Cognitive Diagnosis Framework (HierCDF), which enables many traditional diagnostic models to flexibly integrate the attribute hierarchy for better diagnosis. Specifically, we first use an efficient Bayesian network to model the influence of attribute hierarchy on students' cognitive states. Then we design a CDM adaptor to bridge the gap between students' cognitive states and the input features of existing diagnostic models. Finally, we analyze the generality and complexity of HierCDF to show its effectiveness in modeling hierarchy information. The performance of HierCDF is experimentally proved on real-world large-scale datasets.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **E-learning**.

KEYWORDS

Cognitive Diagnosis, Attribute Hierarchy, Bayesian Network

ACM Reference Format:
Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539486

*Zhenya Huang is the corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539486

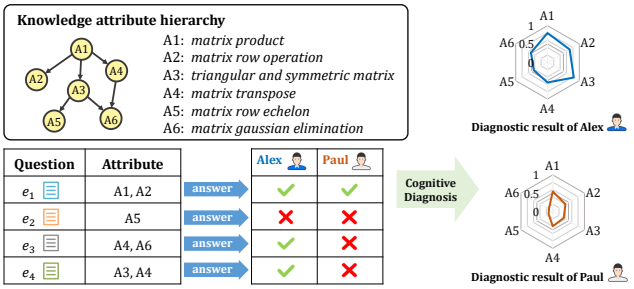


Figure 1: A toy example of cognitive diagnosis. The left part includes the attribute hierarchy, test information and response logs. The right part is the diagnostic report.

1 INTRODUCTION

Cognitive diagnostic assessment (CDA) [23] is a fundamental task in the application of intelligent education systems [13], such as course recommendation [33] and question design [16]. As Figure 1 shows, there are three main components in CDA, including students (i.e., Alex and Paul), questions (i.e., $e_1 \sim e_4$), and attributes (i.e., A1 ~ A6, also known as skills or knowledge concepts). The goal of CDA is to diagnose students' cognitive states on the attributes (presented in radar graphs in Figure 1) through their response to questions.

To successfully answer test questions, students need to master one or more attributes corresponding to these questions. At the same time, the learning of an attribute might be dependent on that of another, as shown in Figure 1. For instance, students' mastery of *matrix product* (A1) supports their mastery of *matrix row operation* (A2) because cognitively the former is the basis of the latter. As a result, in an attribute hierarchy, students' cognitive levels on parent attributes would significantly impact that on child attributes.

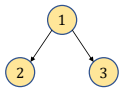
In the literature, there exist two main lines of the research of diagnosing students' cognitive states. The first line estimates students' latent abilities without considering their cognitive levels on specific attributes, such as Item Response Theory (IRT) models [2]. The second line enables diagnosing students on specific attributes using Cognitive Diagnostic Models (CDMs) [23], by either classifying students into binary attribute mastery pattern (e.g., Deterministic Input, Noisy "And" gate model (DINA) [5]) or representing students with continuous vectors indicating their proficiencies on different knowledge attributes (e.g., NeuralCD [32], Relation map driven Cognitive Diagnosis (RCD) [6]). However, most CDMs, including DINA, NeuralCD, etc., treat knowledge attributes as independent and ignore the aforementioned attribute hierarchy which contains important information for cognitive diagnosis. Some CDMs, such as the Attribute Hierarchy Method model [22] consider the attribute hierarchy but have strong assumptions that might not be suitable for different situations (we use AHM as the abbreviation of this type of CDMs). For example, the hierarchical cognitive assumption (defined in Sec. 3.2) in AHM indicates that an attribute could be mastered only if its parent attribute(s) is/are mastered. Taking our toy model in Figure 1 as an example, if a student has not mastered *matrix product*, then it is considered impossible to master *matrix transpose*. This may not be true when the influence of the parent attribute is not strong on its child attribute(s), or when the manually labeled attribute hierarchy is imprecise. Moreover, the AHM models have to enumerate all attribute mastery patterns (2^K mastery patterns given K attributes), which is time-consuming and thus impractical when facing a large number of attributes.

To this end, we aim to model the attribute hierarchy efficiently and integrate it for better cognitive diagnosis. Inspired by hierarchical multi-label classification [8, 12] and hierarchical topic mining [18, 25, 31] where attribute hierarchy is usually characterized in a top-down way, we model the effect of attribute hierarchy on students' cognitive level by propagating the cognitive level on parent attributes to that on child attributes. Along this line, we propose a novel Hierarchical Cognitive Diagnosis Framework (HierCDF) that combines the modeling of attribute hierarchy with traditional diagnostic models. Specifically, we first utilize a Bayesian network [7] to reasonably and efficiently model students' cognitive states in the attribute hierarchy. Then, by designing a CDM adaptor that transforms students' cognitive states and questions' features, we integrate the hierarchical cognitive diagnosis to several traditional diagnostic models (e.g., IRT, MIRT, MF, and NeuralCD). We further compare HierCDF with AHM and show its superiority from two aspects. The first is a generality that HierCDF makes weaker assumptions on the dependencies and mines the dependence information in the attribute hierarchy from response data. The second is efficiency that HierCDF has only linear time complexity with the number of attributes. Extensive experiments on real-world datasets illustrate the performance of HierCDF on modeling students' cognitive level and its ability of modeling attribute hierarchy.

2 RELATED WORK

Cognitive Diagnostic Assessment (CDA). Existing diagnostic models characterize student profile by latent factors (e.g., Item Response Theory (IRT) [2], Multidimensional Item Response Theory (MIRT) [27]), or by attribute mastery patterns (e.g., Deterministic Input, Noisy "And" gate model (DINA) [5], Neural Cognitive Diagnosis (NeuralCD) [32], and Relation map driven Cognitive Diagnosis (RCD) [6]). The former is also known as Latent Factor Models (LFM), while the latter is named Cognitive Diagnosis Models (CDM). As an example of LFM, IRT represents the student i's ability as a single variable θ_i, and uses a logistic function to model the interaction. For instance, the 2PL-IRT defines $P(r_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1+e^{-a_j(\theta_i - b_j)}}$, where r_{ij}, a_j, b_j are student i's response score of question j, question j's discrimination parameter and question j's difficulty parameter respectively. As an representative CDM, DINA uses independent binary variables to model the mastery state (0 for "unmastered" and 1 for "mastered"). Another example of CDM, i.e., NeuralCD, leverages independent continuous variables in $(0, 1)$ to model students' mastery degree on attributes, and exploits a neural network to capture the complex interaction between students and questions. Besides, RCD models the interactive and structural relations from the multi-layer student-question-concept relation map. However, few of LFM and CDM utilize the attribute hierarchy to help model students' cognitive states.

Attribute Hierarchy Method (AHM). AHM [22] is a class of rule-based cognitive diagnosis models that characterizes students' cognitive states under the attribute hierarchy (AH). AHM model characterizes AH by the hierarchical cognitive assumption (HCA) that the mastery of parent attributes is the prerequisite of the mastery of child attributes. As a representative AHM model, Hierarchical Diagnostic Classification Model (HDCM) [28] combines HCA



Valid	✓	✗	✗	✗	✓	✓	✓	✓
A1	0	0	0	0	1	1	1	1
A2	0	1	0	1	0	0	1	1
A3	0	0	1	1	0	1	0	1

Figure 2: An example of latent space. The left part is an attribute hierarchy. The right part is the set of corresponding attribute patterns.

with Log-Linear Cognitive Diagnosis Model (LCDM) [11]. However, the implementation of HCA is time-consuming, because all attribute mastery patterns (i.e., the composite of attribute mastery state) have to be enumerated. For example, for the AH shown in Figure 2, AHM firstly enumerates all $2^3 = 8$ attribute mastery patterns, then removes 3 attribute mastery patterns that are inconsistent with HCA. The set of unremoved attribute mastery patterns is called the latent space. The AHM then classifies each student to an attribute pattern in the latent space that has the largest probability to generate her response log. Although efforts have been made to boost the efficiency of AHM (e.g., using a lattice-theoretical approach to construct the latent space [20]), the time complexity is still large. Moreover, there are limitations for HCA in real-world scenarios, because the effect of a parent attribute on its child attribute(s) may vary a lot, and manually labeled attribute hierarchy might be imprecise. As a result, AHM models are impractical when facing a large number of attributes and a complex or imprecise AH.

Bayesian Network (BN). Bayesian network [7] is a probabilistic graphical model (PGM) [21] that enables computers to reason and infer relationships between entities organized in a directed acyclic graph (DAG), where each directed edge represents a logical dependence relationship. The Bayesian network uses conditional probabilities to model the dependence relationship between entities and uses posterior probabilities to estimate the distribution of an entity given its priors. For example, given the attribute hierarchy in Figure 2, the marginal distribution of A1 to A4 is factorized to the product of prior and conditional probabilities below:

$$P(\mathrm{A1}, \mathrm{A2}, \mathrm{A3}) = P(\mathrm{A1})P(\mathrm{A2}|\mathrm{A1})P(\mathrm{A3}|\mathrm{A1})\,. \tag{1}$$

In intelligent education, Bayesian networks have been applied to student modeling [4] and knowledge tracing [17, 26]. For example, Conati et al. [4] applied BN to the Andes [30], an intelligent education system for Newtonian physics, to model the uncertainty in students' reasoning and learning process. In knowledge tracing [24], Pelánek [26] systematically introduced Bayesian Knowledge Tracing (BKT) that uses the Bayesian network to deduce latent student variables in a knowledge tracing model. Furthermore, Käser et al. [17] utilized Dynamic Bayesian Networks (DBN) to model skill topologies in knowledge tracing. However, in cognitive diagnosis, the application of the Bayesian network is still under-explored.

3 PRELIMINARIES

3.1 Task Overview

Basic concepts. $S = \{s_1, s_2, \ldots, s_N\}$ and $E = \{e_1, e_2, \ldots, e_M\}$ are utilized to represent the student set and question item set respectively, where $N = |S|$ and $M = |E|$. Let $\mathcal{R} = \{(s, e, y)|s \in S, e \in E, y \in \{0, 1\}\}$ be the set of response logs, where y is the response score. Furthermore, we define $C = \{1, 2, \ldots, K\}$ as the attribute set where $K = |C|$. Next, we define $Q = (q_{ij})_{M\times K}$ as the Q-matrix where $q_{ij} = I(\text{item } e_i \text{ requires attribute } j)$. $I(\cdot)$ is the indicator function that $I(\mathcal{E}) = 1$ if $\mathcal{E}$ is true, and $I(\mathcal{E}) = 0$ otherwise. All these data are given in advance of cognitive diagnosis assessment.

Attribute hierarchy. The Attribute Hierarchy (AH) is defined as *the cognitive dependency structure of attributes in the cognitive states.* That is to say, attribute a is the ancestor of attribute b in AH only if the learning of a is the basis of the learning of b. Formally, the attribute hierarchy is a directed acyclic graph (DAG), i.e., $G = (C, \mathbf{E})$ with C being the node set (i.e., the attribute set) and $\mathbf{E}$ being the edge set. In an AH, each node represents an attribute, and each edge represents a dependency between two attributes.

Given these basic concepts and the attribute hierarchy, the cognitive diagnosis task is defined as below:

Definition 3.1. ***Cognitive Diagnosis Task.*** *Given the attribute hierarchy G, the response log set $\mathcal{R}$ and the Q-matrix Q, our goal is to mine students' attribute mastery pattern (i.e., the composite of students' proficiency on knowledge attributes).*

3.2 Hierarchical Cognitive Assumption

The hierarchical cognitive assumption (HCA) is proposed by [22], where the authors describe the assumption as a constraint of the rule space of attribute patterns. Here we give the formal definition of the hierarchical cognitive assumption:

Definition 3.2. ***Hierarchical Cognitive Assumption.*** *Given the attribute hierarchy G, the mastery of a parent attribute is the prerequisite of the mastery of a child attribute.*

For example, *matrix product* is the parent of *matrix row operation* in G (see Figure 1). If a student has not mastered *matrix product*, it is impossible for him to master *matrix row operation* under HCA.

HCA is a strong assumption whose validity is influenced by the preciseness of manually labeled attribute hierarchy, and may not always be correct in different situations. Thus, we do not directly use the HCA as a hard constraint as AHM does. Instead, as will be introduced in Sec. 4.2, we use a Bayesian network to learn the variant dependencies among attributes based on the hierarchy and response data. We will compare the modeling of HierCDF and AHM in Sec. 4.5, and prove that HCA can be learned from data if it is indeed statistically contained in data. Furthermore, in Sec. 5.6.4 we will demonstrate that HierCDF could discover more reasonable dependencies among attributes.

3.3 Monotonicity Assumption

The monotonicity assumption [27] qualitatively describes the relationship between attribute mastery pattern and the probability of correct answer. The definition is given as below:

Definition 3.3. ***Monotonicity Assumption.*** *The probability of correctly answering a question is monotonically increasing with any dimension of the student's attribute mastery pattern.*

For instance, both student a and student b plan to do a question that requires the attribute *Euclidean geometry*. If a masters better on *Euclidean geometry* than b, then the probability of a correctly answering the question is higher than that of b. The assumption is the key point to keep model explainability and is the basis of traditional models such as IRT [2], MIRT [27] and NeuralCD [32].

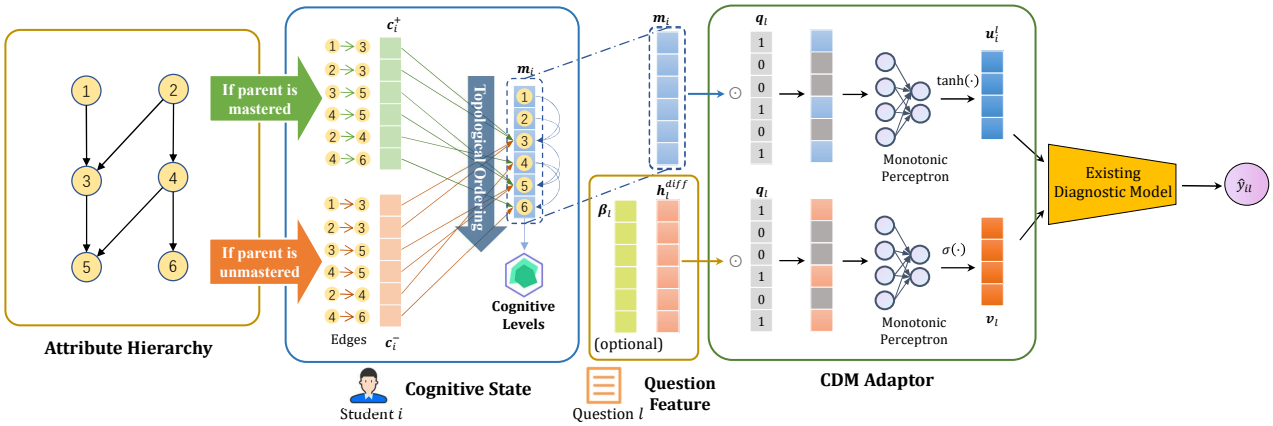


Figure 3: An overview of the hierarchical cognitive diagnosis framework.

4 HIERARCHICAL COGNITIVE DIAGNOSIS FRAMEWORK

4.1 Model Overview

The structure of the Hierarchical Cognitive Diagnosis Framework (HierCDF) is presented in Figure 3. HierCDF consists of three main components, i.e., the cognitive state module, the question feature module, and the CDM adaptor. The cognitive state module uses a Bayesian network isomorphic to the attribute hierarchy to infer students' cognitive states. Simultaneously, to provide cognitive diagnosis models (CDMs) with question factors, the question feature module characterizes questions by extendable question parameters. Next, the CDM adaptor transforms students' cognitive states and questions' feature to latent factors as the input feature of the existing diagnostic model. In this way, HierCDF can be combined with many diagnostic models like IRT, MIRT, etc. Finally, transformed latent factors are input to the existing diagnostic model to predict students' scores on questions.

4.2 Cognitive State Modeling

In this part, we introduce the modeling of students' cognitive states in the attribute hierarchy. As Sec. 3.1 did, we use the term "node" to represent an attribute in the attribute hierarchy. We firstly represent student i's cognitive level on attribute k as $m_{ik} = P(\Theta_{ik} = 1)$, where $\Theta_{ik} = I(\text{student } i \text{ masters attribute } k)$. Furthermore, we use $c^+_{i,k|j} = P(\Theta_{ik} = 1|\Theta_{ij} = 1)$ (the element of $\boldsymbol{c}^+_i$) and $c^-_{i,k|j} = P(\Theta_{ik} = 1|\Theta_{ij} = 0)$ (the element of $\boldsymbol{c}^-_i$) to characterize student i's cognitive level on attribute k when she has mastered or not mastered the parent j. Then, to propagate the influence of parent attributes on the student's cognitive level on child attributes, for each attribute k, m_{ik} is inferred in a topological ordering. Mathematically, let $p(k)$ be the parent node set of k. Given each m_{ij} for $j \in p(k)$ and the conditional mastery probabilities $c^+_{i,k|j}$ and $c^-_{i,k|j}$, m_{ik} can be represented as their function, as shown in Eq.(2):

$$m_{ik} = \mathcal{G}(m_{ij_1}, \ldots, m_{ij_q}, c^+_{i,k|j_1}, \ldots, c^+_{i,k|j_q}, c^-_{i,k|j_1}, \ldots, c^-_{i,k|j_q}), \quad (2)$$

where $\mathcal{G}(\cdot)$ models the relationship between m_{ik} and parent nodes. For node k in different positions, students' learning method on the attribute differs a lot. We classify nodes into three types by the number of parents nodes, and analyze the impact of parent nodes on the cognitive level of child attributes.

Root nodes without parent. If attribute k is a root node without parent (e.g., $k = 1$ in Figure 3), students can learn the attribute from scratch. Thus, student i's mastery of the attribute is independent of any of the other attributes. Formally, m_{ik} is an independent trainable parameter:

$$m_{ik} = t^*_{ik}, \quad (3)$$

where "$*$" means that m_{ik} is optimal in predicting response scores.

Nodes with a single parent. If attribute k is a node with a single parent (e.g., $k = 6$ in Figure 3), student i's cognitive level on the attribute depends on that on the parent node j (e.g., $j = 4$ in Figure 3). To this end, we use positive and negative conditional mastery probabilities $c^+_{i,k|j}$ and $c^-_{i,k|j}$ to characterize the cognitive level of attribute k given that of parent attributes. In this way, HierCDF models attribute hierarchy more flexibly than the hierarchical cognitive assumption. For example, if $c^-_{i,k|j}$ is high, the attribute k tends to be mastered even if j is unmastered. Thus, in student i's learning process, the dependence degree of the mastery of k to the mastery of j is low. Besides, if $c^+_{i,k|j}$ is low, the attribute k tends to be unmastered even if parent j is mastered. So for student i, the difficulty of attribute k is high.

Next, we introduce the constraint on conditional mastery probabilities. Intuitively, the better student i masters a parent attribute, the better she masters the child attribute. So we limit that $c^+_{i,k|j}$ is always larger than $c^-_{i,k|j}$ for any parent-child attribute pair (j, k):

$$c^+_{i,k|j} > c^-_{i,k|j}. \quad (4)$$

To satisfy the constraint, we define a penalty term in Eq.(5) for the loss function presented Sec. 4.4:

$$J(\Omega) = \sum_{i,j,k} ReLU\left(\log \frac{\left(1 - c^+_{i,k|j}\right) c^-_{i,k|j}}{\left(1 - c^-_{i,k|j}\right) c^+_{i,k|j}}\right), \quad (5)$$

where Ω composes of all optimizable parameters in the HierCDF. Then, for attribute k with a single parent attribute, m_{ik} is the expectation of conditional mastery probabilities, as shown in Eq.(6):

$$m_{ik} = c^+_{i,k|j} m_{ij} + c^-_{i,k|j}(1 - m_{ij}). \quad (6)$$

The inference of m_{ik} is reasonable because both cases of the mastery state of the parent attribute are considered. If the cognitive level of the parent attribute is low, then the negative conditional mastery probability plays a leading role in the inference of m_{ik}.

Nodes with multiple parents. If attribute k is a node with multiple parents (e.g., $k = 5$ in Figure 3), the effect of parent attributes (e.g., $j_1 = 3, j_2 = 4$ in Figure 3) on student i's cognitive level on the attribute differs a lot. Let $p(k) = \{j_1, \ldots, j_q\}$ be the parent set of k, then the effect of each parent $j \in p(k)$ on m_{ik} is characterized by positive and negative conditional mastery probabilities, i.e., $c^+_{i,k|j}$ and $c^-_{i,k|j}$, with $c^+_{i,k|j} > c^-_{i,k|j}$ defined in Eq.(4). However, Bayesian networks only model the joint effect of all parent nodes on the student's cognitive level on k by the joint conditional mastery probability, i.e., $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$. To this end, in HierCDF, $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$ is estimated as the function of $\boldsymbol{c}^+_i$ and $\boldsymbol{c}^-_i$, as shown in Eq.(7):

$$P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q}) = \mathcal{H}(c_{i,k|j_1}(\Theta_{ij_1}), \ldots, c_{i,k|j_q}(\Theta_{ij_q})), \tag{7}$$

where $c_{i,k|j_z}(\Theta_{ij_z}) = \Theta_{ij_z} * c^+_{i,k|j_z} + (1 - \Theta_{ij_z}) * c^-_{i,k|j_z}$ for $1 \le z \le q$, and $\mathcal{H}(\cdot)$ can be specified by many functions. We next analyze the constraints on $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$, and present our implementation of $\mathcal{H}(\cdot)$.

There are two constraints on $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$ theoretically. The first constraint is that the higher the student's cognitive level of a parent attribute (i.e., m_{ij} for $j \in p(k)$), the higher the $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$. For instance, attribute A3 has two parents A1 and A2. For student i, the higher the cognitive level of A1/A2 when fixing the other, the higher the cognitive level of A3. The second constraint is that if at least one of the parent attributes is unmastered, the attribute k tends to be unmastered (i.e., if one of m_{ij} for $j \in p(k)$ is small, $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$ is also small), because students' learning of an attribute is *conjunctive*[1] [11].

Considering these constraints, we find that the geometric mean is a fine implementation of $\mathcal{H}(\cdot)$, as shown in Eq.(8).

$$P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q}) \approx \prod_{z=1}^{q} \sqrt[q]{c_{i,k|j_z}(\Theta_{ij_z})}. \tag{8}$$

The geometric mean is monotonically increasing with any of conditional mastery probabilities, and it is more sensitive to small values than other mean estimators such as arithmetic mean. For instance, $X = (0.1, 0.9)$, then $\overline{X}_{arithmetic} = 0.5 > 0.3 = \overline{X}_{geometric}$. Successive product (i.e., $\prod_{z=1}^{q} c_{i,k|j_z}(\Theta_{ij_z})$) which is normally used in the naive Bayesian model is not considered because the estimated value drops dramatically as the number of parents increases, which is unfair for nodes with many parents. For example, for three parents with conditional mastery probabilities equal to 0.6, 0.6, 0.65 respectively, the product is $0.6 \times 0.6 \times 0.65 = 0.234$.

Next, student i's cognitive level on attribute k with multiple parent attributes is characterized by the expectation of $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$ in HierCDF, with the corresponding weight as $P(\Theta_{ij_1}, \ldots, \Theta_{ij_q})$. For parent nodes $j_1, \ldots, j_q \in p(k)$, their correlation has been contained in the inference of cognitive levels on them (e.g., if $j_1, j_2 \in p(k)$ and $j_1 \in p(j_2)$, then m_{ij_2} is inferred through m_{ij_1}). So we assume that $P(\Theta_{ij_1}, \ldots, \Theta_{ij_q}) = \prod_{l=1}^{q} P(\Theta_{ij_z})$. Because $P(\Theta_{ij_z} = 1) = 1 - P(\Theta_{ij_z} = 0) = m_{ij_z}$ for $z = 1, \ldots, q$, the estimation is equivalent to:

$$P(\Theta_{ij_1}, \ldots, \Theta_{ij_q}) = \prod_{z=1}^{q} \left((1 - \Theta_{ij_z})(1 - m_{ij_z}) + \Theta_{ij_z} m_{ij_z}\right). \tag{9}$$

Finally, m_{ik} is inferred as the expectation of $P(\Theta_{ik} = 1|\Theta_{ij_1}, \ldots, \Theta_{ij_q})$, which is shown in Eq.(10):

$$m_{ik} = \sum_{\Theta_{ij.} \in \{0,1\}} P(\Theta_{ij_1}, \ldots, \Theta_{ij_q}) \prod_{z=1}^{q} \sqrt[q]{c_{i,k|j_z}(\Theta_{ij_z})}. \tag{10}$$

4.3 Question Modeling and CDM Adaptor

4.3.1 Question Modeling. Diagnostic models assess students' cognitive states through their response of questions. Thus, it is crucial not only to model students' feature but also to model questions' feature. In question modeling, each question l is characterized as a extendable variable $\psi_l = (\boldsymbol{h}^{diff}_l, \boldsymbol{\beta}_l)$, where $\boldsymbol{h}^{diff}_l \in (0, 1)^{K\times 1}$ indicates l's attribute dificulty and $\boldsymbol{\beta}_l$ is the set of optional parameters like discrimination $h^{disc}_l \in (0, 1)$. The necessity of optional parameters depends on the selection of cognitive diagnosis models.

4.3.2 CDM Adaptor. Different diagnostic model uses different form to characterize student and question features. To build a bridge between students' cognitive states, question features, and the input feature of existing diagnostic models, we introduce the CDM adaptor. Generally, given student i and question l, diagnostic models predict student performance score $\hat{y}_{il}$, as shown in Eq.(11):

$$\hat{y}_{il} = \mathcal{F}\left(\boldsymbol{u}^l_i, \boldsymbol{v}_l\right), \tag{11}$$

where $\boldsymbol{u}^l_i$ is the student latent feature, and $\boldsymbol{v}_l$ is the question latent feature. $\mathcal{F}(\cdot)$ represents the existing diagnostic model, and can be specified with many models like IRT, MIRT, etc. To this end, CDM adaptor transforms $\boldsymbol{m_i} = (m_{i1}, \ldots, m_{iK})^\top$ and ψ_l to $\boldsymbol{u}^l_i$ and $\boldsymbol{v}_l$ respectively, as shown in Eq.(12) and Eq.(13):

$$\boldsymbol{u}^l_i = \tanh\left(W_S(\boldsymbol{m}_i \odot \boldsymbol{q}_l) + \boldsymbol{b}_S\right), \tag{12}$$

$$\boldsymbol{v}_l = \text{sigmoid}\left(W_E\left(\boldsymbol{h}^{diff}_l \odot \boldsymbol{q}_l\right) + \boldsymbol{b}_E\right), \tag{13}$$

where $\boldsymbol{q}_l = (q_{l1}, \ldots, q_{lK})^\top$ is the transpose of the l-th row of Q-matrix, and W_S and W_E are weight matrices of perceptrons. The input vector of each perceptron is the element-wise product of the attribute pattern and $\boldsymbol{q}_l$, because we only focus on attributes that are required by the question. Besides, all elements of W_S are positive because $\boldsymbol{u}^l_i$ must be kept monotonically increasing with $\boldsymbol{m}_i$ to ensure the monotonicity assumption. As for the activation function, our choice is based on two reasons. First, most diagnostic models limit student and question features to a fixed range. Second, for those models with an inner-product operation such as MIRT and MF, $\boldsymbol{v}_l$ must be kept positive to ensure the monotonicity assumption.

4.4 Loss Function

The main term of the loss function of our HierCDF is the cross entropy between the output $\hat{y}_{il}$ and the true response score y_{il}. The $J(\Omega)$ is the penalty term defined in Eq.(5). Then the overall loss function is defined as:

$$\mathcal{L}(\Omega) = -\sum_{i,l} \left(y_{il} \log \hat{y}_{il} + (1 - y_{il}) \log(1 - \hat{y}_{il})\right) + \lambda \cdot J(\Omega), \tag{14}$$

where Ω is the optimizable parameter of our HierCDF, and λ is a hyperparameter given in advance. Generally, λ is a relatively small value such as 0.001 so that the penalty term will not be much larger than the cross-entropy term.

[1] In the learning of a child attribute, the mastery of a parent attribute cannot make up for the non-mastery of another.

4.5 A Comparison of HierCDF and Attribute Hierarchy Method (AHM)

4.5.1 Modeling Attribute Hierarchy. We compare the ability of modeling attribute hierarchy of AHM and HierCDF. The analysis result shows that without direct limitation to students' cognitive level, HierCDF can still learn the hierarchical cognitive assumption (HCA) when it is consistent with data distribution. This property of HierCDF is also illustrated in experiments in Sec. 5.6.4.

AHM. In AHM, for student i, if the parent attribute j is unmastered by her (i.e., $\Theta_{ij} = 0$), then it is sure that the child attribute k cannot be mastered by her (i.e., $\Theta_{ik} = 0$) either. The hierarchical cognitive assumption is rigid even if the data distribution is inconsistent with the assumption.

HierCDF. In HierCDF, if student i has not mastered the parent attribute j, m_{ij} would be small. Then in the inference of m_{ik}, the weight of $c^-_{i,k|j}$ would be larger than that of $c^+_{i,k|j}$, thus the former plays a dominant role. Since we have limited that $c^-_{i,k|j} < c^+_{i,k|j}$, m_{ik} would also be a small value. However, the property is not rigid in HierCDF. If response logs of student i show that she always answers correctly on questions requiring k even if she has not mastered parent node j, she still has chances to master k without mastering j. So $c^-_{i,k|j}$ would be large, and the dependence degree (i.e., $d(j,k)$) of her cognitive level on attribute k on parent node j is very low. As a result, $d(j,k)$ is monotonically decreasing with $c^-_{i,k|j}$.

4.5.2 Model Complexity. We assume that the number of students is fixed and a set of K attributes and the corresponding attribute hierarchy with W edges are given.

AHM. In the construction phase, the AHM enumerates all binary attribute mastery patterns and reserves only that is consistent with the hierarchical cognitive assumption. The construction process of AHM has been introduced in Sec. 2. The time and space complexities are both $O(2^K)$. In the diagnosis phase, let $Q(K)$ be the number of reserved attribute mastery patterns. Since AHM compares each pattern to the targeted student, the time complexity is $O(Q(K))$. $Q(K)$ is usually much larger than K.

HierCDF. In the construction phase, HierCDF only saves the attribute hierarchy (AH) and its parameters (i.e., mastery probabilities of root attributes, and positive and negative conditional mastery probabilities of each parent-child attribute pair). Thus the time complexity is $O(1)$, and the space complexity is $O(K + W)$. In the diagnosis phase, given a student, HierCDF needs to infer her cognitive level on each attribute. Because the in-degree of attributes is usually small, the average time complexity is approximately $O(K)$. As a result, in both phases, HierCDF is more efficient than AHM.

5 EXPERIMENT

5.1 Experiment Overview

In this section, we introduce datasets and the experimental setup. Next, we conduct experiments on the original and the HierCDF[2] version of diagnostic models to answer the following questions:

- RQ1. Can HierCDF improve the performance of diagnostic models on predicting students' response scores?
- RQ2. What is the explainability of the diagnostic output of HierCDF?

[2]https://github.com/CSLiJT/HCD-code

Table 1: The statistics of datasets

Statistics	Junyi	Junyi-s	MATH-2021
#Attributes	734	8	662
#Edges of G	929	8	673
Diameter of G	37	3	21
#Students	10,000	2,401	14,826
#Question items	734	8	13,114
#Response logs	408,057	6,176	555,625
#Responded attributes	707	8	588
#Attributes per question	1	1	1
#Response per student	40.8	2.6	37.5

Table 2: The hypothesis test results of datasets ($\alpha = 0.05$)

Dataset	$\overline{Y}$	Test statistic	P-value	H_a
Junyi	0.0641	2.37×10^9	$< 10^{-10}$	Accepted
MATH-2021	0.0429	1.10×10^9	$< 10^{-10}$	Accepted

- RQ3. Can HierCDF effectively model AH?
- RQ4. What are the features of the output of HierCDF?

5.2 Dataset Description

We conduct experiments on two real-world datasets, i.e., the Junyi Academy Math Practicing Log (*Junyi*) dataset [3] [3] from Junyi Academy and the *MATH-2021* dataset supplied by iFLYTEK Co., Ltd., which is collected from the iFLYTEK Learning Machine[4]. Both datasets contain K-12 mathematical attribute hierarchies, questions, and student response logs. A subset of the Junyi dataset, namely *Junyi-s*, is sampled to compare the performance of HierCDF and AHM. That's because the AHM is unsuitable for the whole Junyi dataset due to the huge time complexity to build the model. The statistics of datasets are described in Table 1.

We reserve only the first attempt and the first day's response log for Junyi Academy Math Practicing Log and MATH-2021 respectively to ensure that the attribute state of students is static. We filter out students with less than 15 response logs to guarantee that every student has adequate response logs for diagnosis. We randomly select 10,000 students from these students for Junyi Academy Math Practicing Log. We divide 80% response logs randomly for each student to compose the train set, and the rest 20% to compose the test set. To ensure fairness, we divide 90% as train data and 10% as validation data respectively from the train set to tune hyperparameter for all models using grid search. All models are trained and tested from scratch 10 times repeatedly and assessed by the average performance.

5.3 Validation of Hierarchical Cognitive Assumption on Datasets

The hierarchical cognitive assumption (HCA) is a theoretical assumption about students' cognitive states. Since our model does not depend on the HCA and learns the dependencies from the data, we first validate the HCA on our datasets.

For any student, the ratio of correctly answering questions that require an attribute (i.e., the attribute response ratio) is increasing with her attribute mastery probability. As a result, if the HCA is valid, then for each student, her response ratio on the parent

[3]https://pslcdatashop.web.cmu.edu/Files?datasetId=1198
[4]https://xxj.xunfei.cn/

Table 3: Experimental results on student performance prediction

Dataset	Metrics	IRT		MIRT		MF		NeuralCD		Random
		Original	HierIRT	Original	HierMIRT	Original	HierMF	Original	HierNCD	
Junyi	AUC	0.7541	**0.7848**	0.7514	**0.7842**	0.7530	**0.7843**	0.7768	**0.7848**	0.4999
	ACC	0.7288	**0.7491**	0.7355	**0.7514**	0.7288	**0.7503**	0.7460	**0.7516**	0.5000
	F1	0.8125	**0.8247**	0.8171	**0.8290**	0.8143	**0.8284**	0.8214	**0.8292**	0.5773
	RMSE	0.4234	**0.4153**	0.4267	**0.4093**	0.4306	**0.4093**	0.4135	**0.4098**	0.7074
MATH	AUC	0.7050	**0.7362**	0.6520	**0.7375**	0.6634	**0.7313**	0.7048	**0.7272**	0.4991
	ACC	0.7008	**0.7008**	0.6680	**0.7170**	0.6615	**0.7133**	0.6933	**0.7114**	0.4999
	F1	0.7978	**0.8095**	0.7636	**0.8109**	0.7504	**0.8072**	0.8115	**0.8057**	0.5766
	RMSE	0.4448	**0.4340**	0.4861	**0.4304**	0.5058	**0.4322**	0.4413	**0.4348**	0.5777

Table 4: Experimental results on Junyi-s

Model	AUC	ACC	F1	RMSE
Random	0.4999	0.5001	0.5283	0.7070
AHM	0.5567	0.5466	0.5373	0.6745
HierIRT	**0.6367**	**0.5589**	**0.7071**	**0.4976**

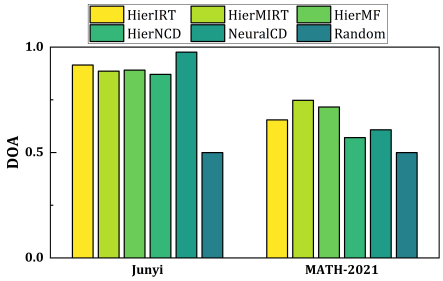


Figure 4: DOA@10 for datasets.

attributes should be larger than that on the child attributes. To test this hypothesis, we use the paired *Wilcoxon Signed-Rank test*, which is a non-parameter method for data with unknown distribution. Formally, let $U = r_{iu}$ and $V = r_{iv}$ be student i's attribute response ratio on the parent attribute u and the child attribute v respectively, then the null and alternative hypothesis are

$$H_0 : E[U] \leq E[V] \Leftrightarrow H_a : E[U] > E[V].$$

We further let $Y = U - V$ and transform the above test to the test of whether or not $E[Y] \leq 0$. The test result is shown in Table 2. $\overline{Y}$ is the mean of Y, and α is the level of significance. For both datasets, the p-values are less than 10^{-10}, thus the H_0 is rejected, and the H_a is accepted, indicating that $E[Y]$, i.e., the difference between U and V, is significantly larger than 0. As a result, the hierarchical cognitive assumption is valid in both datasets statistically.

5.4 Experimental Setup

To assess the performance of HierCDF, we apply the model framework to four diagnostic models, i.e., IRT, MIRT, MF, NeuralCD, to get the corresponding HierCDF implementations[5], i.e., HierIRT, HierMIRT, HierMF, HierNCD. Then we compare the student score prediction performance of HierCDF to these baseline approaches. We also test the performance of HierIRT and AHM [22] on the same task on the Junyi-s dataset to compare HierCDF with rule-based methods. The threshold of score prediction is set as 0.5. For multidimensional latent factor models (i.e., MIRT and MF) and all HierCDF models except HierIRT, the hidden dimension is set as 16. The dimensions of the full connection layers of NeuralCD are 512, 256, and 1 respectively as in [32]. We set the hyperparameter $\lambda = 0.001$. All parameters are initialized with *Xavier normal* method [9], and we use the *Adam* algorithm [19] for model optimization. All models are implemented with PyTorch using Python, and all experiments are run on a Linux server with four 2.30GHz Intel Xeon E5-2620 v3 CPUs and a Tesla P40 GPU.

5.5 Evaluation Metrics

In this section, we introduce evaluation metrics that measure the performance of diagnostic models from various aspects.

[5]The implementations of HierCDF are presented in Appendix A.1

Student Score Prediction Metrics. Since the students' true attribute mastery pattern is unobservable, it is difficult to directly evaluate the performance of the diagnostic models. A common practice is to assess the diagnostic models through predicting students' test score [5, 27, 32]. We follow this approach, and measure the performance of the diagnostic models by their predictions of students' test scores. We evaluate diagnostic models similarly as the evaluation of classification models and regression models, and choose *Accuracy* (ACC), *F1-score* [29], *Area Under Curve* (AUC) [1] and *Rooted Mean Squared Error* (RMSE) as evaluation metrics.

Explainability Metrics. The explainability of attribute mastery pattern is crucial for any diagnostic model. Intuitively, if student a's response accuracy on attribute k is larger than student b, then a's mastery probability of k should also be larger than b's, i.e., $m_{ak} > m_{bk}$ [15]. Therefore, we adopt *Degree Of Agreement* (DOA) as our explainability metrics, which is defined in Eq.(15):

$$DOA_k = \frac{\sum_{a,b \in S} \delta(m_{ak}, m_{bk}) \frac{\sum_{j=1}^{M} q_{jk} \wedge J(j,a,b) \wedge \delta(r_{aj}, r_{bj})}{\sum_{j=1}^{M} q_{jk} \wedge J(j,a,b) \wedge I(r_{aj} \neq r_{bj})}}{Z}, \quad (15)$$

where $Z = \sum_{a=1}^{N} \sum_{b=1}^{N} \delta(m_{ak}, m_{bk})$. $\delta(x, y) = I(x > y)$. q_{jk} is the (j, k) element of Q-matrix, indicating whether question j requires attribute k. $J(j, a, b) = 1$ if both student a and b answered question j and $J(j, a, b) = 0$ otherwise. We take an average of the DOAs of the top 10 attributes with the largest number of response logs, and use as the DOA of the model (DOA@10).

Attribute Hierarchy Modeling Metrics. In Sec. 5.3, we statistically validated the hierarchical cognitive assumption (HCA). However, the assumption may not hold for all parent-child attribute pairs. To thoroughly evaluate how well the attribute hierarchy is modeled by diagnostic models, we propose two evaluation metrics. For the conditions when the HCA is ignored, a student's cognitive

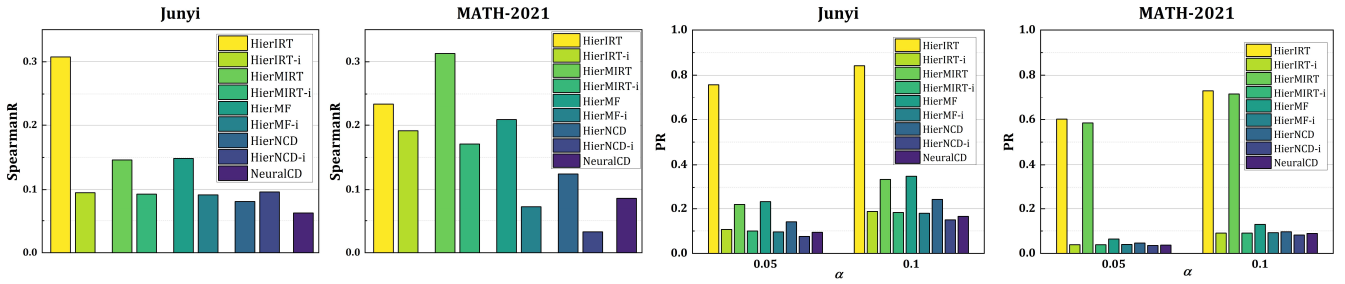


Figure 5: Spearman rank correlation and passing ratio results of models.

level on child attribute(s) may still be correlated to that on parent attribute(s). As a result, use the *Spearman's rank correlation coefficient* [10] as the evaluation metrics, which is shown in Eq.(16):

$$r_s(X, Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{16}$$

where $d_i = rank(x_i) - rank(y_i)$, $i = 1, 2, \ldots, n$, and r_s is mapped for each student. Here X and Y are the child and parent attribute mastery degrees separately. We calculate the average value $\overline{r_s}$ about students as the evaluation metrics.

Furthermore, we evaluate how well the model satisfies the HCA. If the assumption is well modeled, then for each student i, her cognitive level on a parent attribute should be no less than that on a child attribute. So we adopt a Wilcoxon-signed-rank-test *Passing Ratio* (*PR*) [14] on students, as defined in Eq.(17). Specifically, for every student and a fixed confidence level α, we conduct a test with alternative hypothesis (H_a) that $E_{k \in C}[m_{ij} - m_{ik}] > 0$ where (j, k) is a parent-child attribute pair. We then calculate the ratio of students that accept H_a among all students as the *PR*. The larger the *PR*, the better the model mines the HCA.

$$PR_\alpha = \frac{1}{N}\sum_{i=1}^{N} I(\text{p-value}_i < \alpha). \tag{17}$$

5.6 Experimental Results

5.6.1 RQ1. Student Score Prediction. We conduct a student score prediction experiment on baseline models and HierCDF. As Table 3 and Table 4 shows, HierCDF outperforms almost all baseline models in three datasets, which indicates the effectiveness of HierCDF on predicting student score. Besides, the hyperparameter λ also affects the performance of HierCDF (see Appendix A.2).

5.6.2 RQ2. Explainability of Diagnostic Results. The explainability experimental result is shown in Figure 4. For traditional CDMs, we only choose NeuralCD as the baseline model because for IRT, MIRT, and MF, there are no clear connections between attribute mastery degrees and latent factors [32]. It is observed that the HierCDF has a competitive DOA compared to the DOA of NeuralCD. As a result, the diagnostic reports of HierCDF are reasonable.

5.6.3 RQ3. Modeling Attribute Hierarchy. Experiment results are presented in Figure 5. An ablation study is included to prove the effectiveness of the Bayesian network in the modeling of the attribute hierarchy (AH). For each Hier$X$ where $X \in$ {IRT, MIRT, MF, NeuralCD}, $\mathbf{E} = \emptyset$ ($\mathbf{E}$ is the edge set of the AH) to build Hier$X$-i. Every Hier$X$ outperforms the Hier$X$-i and NeuralCD in most cases, indicating that the Bayesian network is effective in modeling AH. Another observation is that HierNCD performs worse than other Hier$X$ in both datasets. The reason is that HierNCD suffers from the problem of vanishing gradient. For a student and attributes

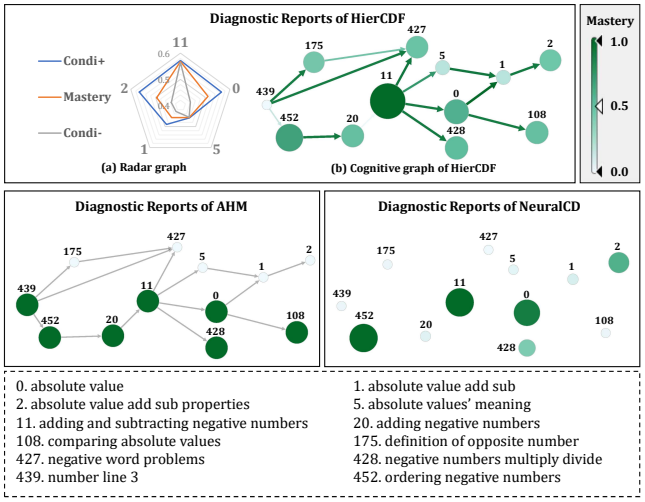


Figure 6: Diagnostic reports of CDMs in Junyi. The deeper the color of a node, the larger the cognitive level. The deeper the color of an edge, the larger the dependence degree.

with parents, the variation of her conditional mastery probabilities is so low that it is hard to distinguish the cognitive levels on these attributes. However, in comparison, the HierCDF is effective on any of the presented diagnostic models in modeling attribute hierarchy.

5.6.4 RQ4. Diagnostic Output Analysis. We randomly select one student ($i = 2493$) in the Junyi dataset to generate her diagnostic outputs. Diagnostic reports are presented in Figure 6, where HierCDF is implemented as HierIRT. Characteristics of the output of HierCDF are demonstrated in the following.

First, HierCDF provides richer information than other CDMs. HierCDF provides students with cognitive levels on attributes (as shown in the radar graph and the nodes of the cognitive graph) and attribute dependencies (as shown in the edges of the cognitive path), while other CDMs provide students with only the former. In the radar graph of HierCDF, *Mastery* is m_{ik} given student i and attribute k, and *Condi+/-* is the student's cognitive level on the attribute given her mastery/non-mastery on parent nodes (e.g., for student i and attribute 0, *Condi+* is $c^+_{i,0|11}$). In the cognitive graph, the dependencies are also visualized. We use $d(j,k) = -\text{sigmoid}^{-1}(c^-_{i,k|j})$ to indicate the dependency of child k on parent j for student i . In Figure 6, diagnostic outputs are normalized to $(0, 1)$. Colors and sizes represent the cognitive level and dependency. For nodes, the deeper the color and the larger the size, the higher the cognitive level. For edges, the deeper the color and the larger the size, the stronger the dependency.

Second, HierCDF can learn the dependencies between different attributes from data instead of making strong assumptions. In the diagnostic reports of HierCDF, students' cognitive level is both affected by their response logs and the attribute hierarchy (AH). In HierCDF, except for the attribute nodes 1 and 439, the student's cognitive levels on all other nodes are consistent with the hierarchical cognitive assumption (HCA). The inconsistency with HCA on node 1 and 439 can be explained by the diagnostic report of NeuralCD, which is the result of response logs where the student's correctness on questions requiring 1 and 439 is actually low. However, compared to NeuralCD, the variation of the student's cognitive level on 1, 439 and child attributes in HierCDF is smaller, and the dependence degrees are low. The result shows that HierCDF strikes a balance between practical response log distribution and the theoretical HCA. In summary, HierCDF can learn the feature of AH based on data distribution rather than a strong assumption.

6 CONCLUSION

In this paper, we presented a novel Hierarchical Cognitive Diagnosis Framework, which utilizes the Bayesian network to efficiently model the attribute hierarchy and integrates with traditional models for better diagnosis. Specifically, we first used a Bayesian network isomorphic to the attribute hierarchy to learn the dependencies among students' cognitive levels on attributes. There are two advantages of our method. First, the dependencies are learned from data instead of strong assumptions. Second, the time and space complexities are both linear, which is much more efficient than traditional models such as AHM. Then, we designed a CDM adaptor to transform the attribute mastery patterns to input features of traditional diagnostic models. This enables traditional models to integrate the attribute hierarchy and therefore expand their diagnostic capabilities. Extensive experiments on real-world datasets showed that HierCDF is both effective and efficient for cognitive diagnosis with consideration of attribute hierarchy and at the same time provides abundant and interpretable diagnostic results. We hope this work provides a new perspective in the area of cognitive diagnosis and inspires further improvements in the future.

ACKNOWLEDGEMENTS

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2021YFF0901005), the National Natural Science Foundation of China (Grants No. 61922073, U20A20229, 62106244, and 62177044), and the iFLYTEK joint research program.

REFERENCES

[1] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 7 (1997), 1145–1159.
[2] Justyna Brzezinska. 2020. Item response theory models in the measurement theory. *Commun. Stat. Simul. Comput.* 49, 12 (2020), 3299–3313.
[3] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach. In *EDM*. International Educational Data Mining Society (IEDMS), 532–535.
[4] Cristina Conati, Abigail S. Gertner, and Kurt VanLehn. 2002. Using Bayesian Networks to Manage Uncertainty in Student Modeling. *User Model. User Adapt. Interact.* 12, 4 (2002), 371–417.
[5] Jimmy de la Torre. 2009. DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (2009), 115–130.
[6] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. RCD: Relation Map Driven Cognitive Diagnosis for Intelligent Education Systems. In *SIGIR*. ACM, 501–510.
[7] Dan Geiger and Judea Pearl. 1990. Logical and algorithmic properties of independence and their application to Bayesian networks. *Ann. Math. Artif. Intell.* 2 (1990), 165–178.
[8] Eleonora Giunchiglia and Thomas Lukasiewicz. 2020. Coherent Hierarchical Multi-Label Classification Networks. In *NeurIPS*.
[9] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS (JMLR Proceedings, Vol. 9)*. JMLR.org, 249–256.
[10] Andréas Heinen and Alfonso Valdesogo. 2020. Spearman rank correlation of the bivariate Student t and scale mixtures of normal distributions. *J. Multivar. Anal.* 179 (2020), 104650.
[11] Robert Henson, Jonathan Templin, and John Willse. 2009. Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika* 74 (06 2009), 191–210.
[12] Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach. In *CIKM*. ACM, 1051–1060.
[13] Zhenya Huang, Xin Lin, Hao Wang, Qi Liu, Enhong Chen, Jianhui Ma, Yu Su, and Wei Tong. 2021. DisenQNet: Disentangled Representation Learning for Educational Questions. In *KDD*. ACM, 696–704.
[14] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *AAAI*. AAAI Press, 1352–1359.
[15] Zhenya Huang, Qi Liu, Yuying Chen, Le Wu, Keli Xiao, Enhong Chen, Haiping Ma, and Guoping Hu. 2020. Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students. *ACM Trans. Inf. Syst.* 38, 2 (2020), 19:1–19:33.
[16] Géraldine Jeckeln, Ying Hu, Jacqueline G. Cavazos, Amy N. Yates, Carina A. Hahn, Larry Tang, P. Jonathon Phillips, and Alice J. O'Toole. 2021. Face Identification Proficiency Test Designed Using Item Response Theory. *CoRR* abs/2106.15323 (2021).
[17] Tanja Käser, Severin Klingler, Alexander G. Schwing, and Markus H. Gross. 2014. Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks. In *Intelligent Tutoring Systems (Lecture Notes in Computer Science, Vol. 8474)*. Springer, 188–198.
[18] Saurabh Kataria, Krishnan S. Kumar, Rajeev Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. 2011. Entity disambiguation with hierarchical topic models. In *KDD*. ACM, 1037–1045.
[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.
[20] Hans-Friedrich Köhn and Chia-Yi Chiu. 2019. Attribute Hierarchy Models in Cognitive Diagnosis: Identifiability of the Latent Attribute Space and Conditions for Completeness of the Q-Matrix. *J. Classif.* 36, 3 (2019), 541–565.
[21] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.
[22] Jacqueline P. Leighton, Mark J. Gierl, and Stephen M. Hunka. 2004. The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement* 41, 3 (2004), 205–237.
[23] Qi Liu. 2021. Towards a New Generation of Cognitive Diagnosis. In *IJCAI*. ijcai.org, 4961–4964.
[24] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. *IEEE Trans. Knowl. Data Eng.* 33, 1 (2021), 100–115.
[25] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. In *KDD*. ACM, 1908–1917.
[26] Radek Pelánek. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Model. User Adapt. Interact.* 27, 3-5 (2017), 313–350.
[27] Mark D. Reckase. 2009. *Multidimensional Item Response Theory Models*. Springer New York, New York, NY, 79–112.
[28] Jonathan Templin and Laine Bradshaw. 2014. Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika* 79 (01 2014).
[29] C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth.
[30] Kurt Vanlehn, Collin Lynch, K. Schulze, Joel Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. 2005. The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education* 15 (01 2005), 147–204.
[31] Chi Wang, Xueqing Liu, Yanglei Song, and Jiawei Han. 2015. Towards Interactive Construction of Topical Hierarchy: A Recursive Tensor Decomposition Approach. In *KDD*. ACM, 1225–1234.
[32] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural Cognitive Diagnosis for Intelligent Education Systems. In *AAAI*. AAAI Press, 6153–6161.
[33] Wei Xu and Yuhan Zhou. 2020. Course video recommendation with multimodal information in online learning platforms: A deep learning framework. *Br. J. Educ. Technol.* 51, 5 (2020), 1734–1747.

A APPENDIX

A.1 Implementations of HierCDF

HierCDF is a general framework that can be implemented with many diagnostic models. Here, we take model $X \in$ {IRT, MIRT, MF, NeuralCD} as an example. We use latent factors generated by the CDM adaptor of HierCDF as input features of the model. Then we specify $\mathcal{F}(\cdot)$ defined in Sec. 4.3 with each model X to implement the corresponding Hier$X$.

IRT. In HierIRT, latent factors are scalars. That is to say, $\boldsymbol{u}_i^l, \boldsymbol{v}_l \in (0, 1)$. The $\mathcal{F}(\cdot)$ is a logistic-like function:

$$\hat{y}_{il} = \text{sigmoid}(h_l^{disc} \cdot (\boldsymbol{u}_i^l - \boldsymbol{v}_l)). \tag{18}$$

MIRT. In HierMIRT, latent factors are multidimensional, i.e., $\boldsymbol{u}_i^l, \boldsymbol{v}_l \in (0, 1)^{D\times 1}$ where $D > 1$. The $\mathcal{F}(\cdot)$ is shown as below:

$$\hat{y}_{il} = \text{sigmoid}(\boldsymbol{v}_l^\top \boldsymbol{u}_i^l + h_l^{disc}). \tag{19}$$

MF. The HierMF uses the inner product as the interaction function. We further use a sigmoid function to compress the predicted value to $(0, 1)$, i.e.,

$$\hat{y}_{il} = \text{sigmoid}(\boldsymbol{v}_l^\top \boldsymbol{u}_i^l). \tag{20}$$

NeuralCD. The HierNCD inputs multidimensional latent factors, and uses a multilayer perceptron (MLP) to capture the complex interaction between students and questions:

$$\hat{y}_{il} = \phi(h_l^{disc} \cdot (\boldsymbol{u}_i^l - \boldsymbol{v}_l)), \tag{21}$$

where $\phi(\cdot)$ is a three layer full connected neural network with non-negative weights to keep explainability.

A.2 Hyperparameter Experimental Results

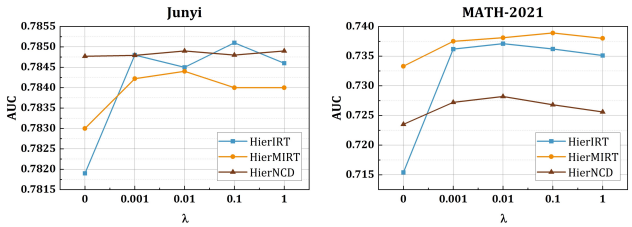


Figure 7: The result shows that HierCDF with $\lambda > 0$ outperforms those with $\lambda = 0$, proving the effectiveness of the parameter constraint in HierCDF. However, the performance of HierCDF drops when λ becomes too large (e.g., $\lambda = 1.0$) because the penalty term exceeds the cross-entropy term in the loss function.