Enhancing Cognitive Diagnosis Using Un-interacted Exercises: A Collaboration-Aware Mixed Sampling Approach

Haiping Ma[1], Changqian Wang[1], Hengshu Zhu[2], Shangshang Yang[3*], Xiaoming Zhang[1], Xingyi Zhang[4*]

[1]Department of Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology, Anhui University, China
[2]Career Science Lab, BOSS Zhipin, China
[3] School of Artificial Intelligence, Anhui University, China
[4]School of Computer Science and Technology, Anhui University, China
hpma@ahu.edu.cn, xmzhang@ustc.edu,
{changqian.wang.dl, zhuhengshu, yangshang0308, xyzhanghust}@gmail.com

Abstract

Cognitive diagnosis is a crucial task in computer-aided education, aimed at evaluating students' proficiency levels across various knowledge concepts through exercises. Current models, however, primarily rely on students' answered exercises, neglecting the complex and rich information contained in un-interacted exercises. While recent research has attempted to leverage the data within un-interacted exercises linked to interacted knowledge concepts, aiming to address the long-tail issue, these studies fail to fully explore the informative, un-interacted exercises related to broader knowledge concepts. This oversight results in diminished performance when these models are applied to comprehensive datasets. In response to this gap, we present the **C**ollaborative-aware **M**ixed **E**xercise **S**ampling (**CMES**) framework, which can effectively exploit the information present in un-interacted exercises linked to un-interacted knowledge concepts. Specifically, we introduce a novel universal sampling module where the training samples comprise not merely raw data slices, but enhanced samples generated by combining weight-enhanced attention mixture techniques. Given the necessity of real response labels in cognitive diagnosis, we also propose a ranking-based pseudo feedback module to regulate students' responses on generated exercises. The versatility of the CMES framework bolsters existing models and improves their adaptability. Finally, we demonstrate the effectiveness and interpretability of our framework through comprehensive experiments on real-world datasets.

Introduction

Amid the rapid advancement of computer-aided education, cognitive diagnosis has garnered increasing attention (Lord 2012; Yang et al. 2023c; Qin et al. 2023). As a crucial task in intelligent education, cognitive diagnosis aims at evaluating students' proficiency levels across various knowledge concepts through exercises. As illustrated in Figure 1, existing cognitive diagnosis studies are based on the historical response logs between students and exercises, as well as the associations between exercises and knowledge concepts for modeling. They believe exercise interactions provide the greatest diagnostic value for students, while overlooking the information contained in un-interacted exercises. In practice, each student's interaction with exercises represents a mere fraction of the complete exercise bank, with the un-interacted exercises containing intricate and extensive information.

In this paper, we attempt to leverage these un-interacted exercise information. One challenge with leveraging such un-interacted exercise information is the absence of students' potential response labels. Recent work in EIRS (YAO et al. 2023) makes the assumption that students will perform comparably on exercises related to the same knowledge concepts. EIRS aims to mitigate the long-tail problem (where insufficient interaction data results in skewed distributions) through similarity-oriented sampling of exercises associated with previously interacted concepts. Since the attained sample exercises convey analogous information to the interacted ones, constraining the acquired knowledge, this circumstance culminates in the method being unable to realize optimal performance on full datasets.

Consequently, determining how to extract additional informative un-interacted exercises constitutes another challenge. Within the domain of recommender systems, informative negative samples are frequently utilized to train the system and enhance recommendation performance (Rendle and Freudenthaler 2014). Accordingly, substantial research has investigated techniques for sampling informative negative instances (Rendle et al. 2012; Wang et al. 2020b; Liu and Wang 2023). However, owing to the distinctive nature of cognitive diagnosis models, which encompass intricate interrelationships among students, exercises, and knowledge concepts, sampling approaches from recommender systems are not transferable to cognitive diagnosis models.

To address the aforementioned challenges, we propose a general framework, namely Collaborative-aware Mixed Exercise Sampling (CMES) for cognitive diagnosis models. CMES extracts more informative diagnoses from the pool of un-interacted exercises and obtains students' potential response labels. Specifically, to improve the quality and efficiency of sampling, we preclude sampling exercises affili-

*Corresponding Authors.

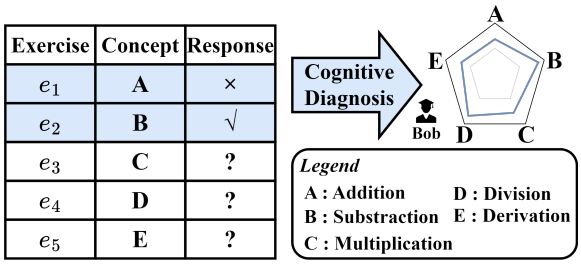

Figure 1: Illustration of cognitive diagnosis. Mainstream cognitive diagnosis models derive diagnosis results from students' response logs. e_1 and e_2 are the exercises that Bob has interacted with, while the remaining exercises are those that he has not interacted with.

ated with students' interacted concepts, as well as exercises with potentially similar information to interacted ones. We cluster students founded on their response capabilities and collaborations. Throughout the sampling progression, we sample from other students' interacted exercise sets in different clusters. As interacted exercises encompass robust diagnostic intimations while un-interacted ones retain prospective heterogeneous information, we use mixing techniques to mix the sampled exercise information by injecting interacted exercise information into sampled exercises, obtaining more informative mixed samples. Finally, we design a ranking based pseudo feedback module to predict potential response situations for the sampled exercises, which is combined with the cognitive diagnosis task for joint learning.

Our main contributions are summarized as follows:

- To fully leverage the latent information in un-interacted exercises for student diagnosis, we propose a generic sampling framework CMES for enhancing cognitive diagnosis tasks.
- We specially design a learning-based pseudo feedback module that defines a learning-to-rank task assisting in the training of the cognitive diagnosis task,
- We have conducted extensive experiments on real-world datasets to validate the effectiveness and scalability of our approach.

Related Work

In this section, we review the related work about cognitive diagnosis models and sampling strategies.

Cognitive Diagnosis

Cognitive diagnosis, a fundamental and critical task in education, aims to infer students' mastery of knowledge concepts. The early models IRT (Lord 1980) and DINA (De La Torre 2009) are two classic cognitive diagnosis models. Unlike IRT which hypothesizes unidimensional independence and adopts continuous latent variables to evaluate examinees' potential abilities, DINA is based on the attribute independence assumption and uses $0/1$ binary vectors to represent students' mastery of each attribute. MIRT (Reckase 2009), as an extension of IRT, discards the unidimensionality and proposes that student proficiency is multidimensional, thus utilizing multiple latent traits to characterize students more comprehensively. NCD (Wang et al. 2020a) firstly introduces neural networks into cognitive diagnosis, so as to capture the sophisticated student-exercise relationships. Afterwards, more neural network based approaches (Yang et al. 2023d,a; Ma et al. 2022; Yang et al. 2023b) are proposed, such as ECD (Zhou et al. 2021) incorporates contextual features to facilitate more precise diagnosis of students' cognitive status. RCD (Gao et al. 2021) attempts to explore student-exercise-concept associations via graphs and conducts more delicate modeling of the interactions. Recent work of ICD (Qi et al. 2023) further investigates the intrinsic correlations among knowledge concepts and quantitative relationships between exercises and concepts. Despite the remarkable progress, existing methods exclusively take advantage of interacted responses while overlooking the un-interacted yet more informative exercises.

Sampling Strategy

Sampling strategies are extensively utilized in recommender systems, where sampling informative non-interacted instances close to positive samples facilitates models to better learn the boundary between positive and negative samples. Conventional recommender systems often adopt random negative sampling (RNS) (Rendle et al. 2012) and static popularity-based negative sampling (PNS) (Caselles-Dupré, Lesaint, and Royo-Letelier 2018; Chen et al. 2017), through which the attained negative samples are typically of low quality and fail to train models effectively. Dynamic negative sampling (DNS) (Zhang et al. 2013) is an adaptive negative sampling approach, scoring each sample and using high-scored ones as negative samples for model training. Currently, GAN-based negative sampling (Wang et al. 2017; Ding et al. 2019; Guo et al. 2020) prevails in recommender systems. Despite explorations into GANs, existing GAN-based sampling strategies often suffer from poor interpretability and inferior performance due to training instability. A graph data augmentation based negative sampling (Huang et al. 2021) augments the positive samples with negative sample information to delude the recommender and enhance its ability to distinguish the boundary. Due to the sophisticated student-exercise interactions that not only reply on answering records but also associations between exercises and concepts, transplanting negative sampling strategies from recommender systems into cognitive diagnosis faces challenges. Although previous work EIRS (YAO et al. 2023) has introduced sampling strategies into cognitive diagnosis, it essentially performs similarity-based sampling, where the attained samples carry comparable information to interacted ones and fail to provide extra diagnostic values. Inspired by the high-quality negative samples achieved in recommender systems, we propose a novel sampling strategy to obtain informative samples.

Problem Statement

For cognitive diagnosis, we define three entity groups: the student set $S = \{s_1, s_2, ..., s_N\}$ of size N; the exercise set

$E = \{e_1, e_2, ..., e_M\}$ of size M; and the knowledge concept set $K = \{k_1, k_2, ..., k_C\}$ of size C. The exercise-concept relationship is defined by matrix $Q \in \mathbb{R}^{M \times C}$, where $Q_{i,j} = 1$ if exercise e_i involves concept k_j, else $Q_{i,j} = 0$. We also define interaction logs as triplets $(s_i, e_j, r_{ij}) \in R$, where e_j is called an interacted exercise of student s_i, $r_{ij} = 1$ if student s_i correctly answered exercise e_j, else $r_{ij} = 0$ and R is the interaction set. The un-interacted exercise set for student s_i is defined by $U_i = E \backslash E_i$, where E_i is the interacted exercise set of student s_i. The knowledge concepts associated with the interacted exercises by student s_i are called interacted knowledge concepts K_i, where $K_i \subset K$.

PROBLEM DEFINITION. *Given student entity, exercise entity, knowledge concept entity, students' exercising response logs, the un-interacted exercises set U_i for each student, and the exercise-knowledge relational matrix. Our goal is to leverage the un-interacted exercises to enhance the performance of cognitive diagnosis.*

The Proposed CMES Framework

In this section, we first briefly introduce the proposed framework, then elaborates on each module, and finally discusses how to train the cognitive diagnosis model with the proposed CMES framework.

Overview. The brief idea of this paper is to enhance cognitive diagnosis by sample augmentation using un-interacted exercises. For this aim, as shown in Figure 2, our CMES framework comprises three key components: the sample augmentation module, the pseudo feedback module, and the extensible diagnosis module. The initial two modules aim to sample and blend informative exercises from the pool of un-interacted exercises for individual students, simultaneously evaluating potential feedback labels for the mixed exercises. More precisely, within the sample augmentation module, we group students based on their response capabilities to mitigate interference from exercises with limited information. We then proceed to sample and mix exercise information from other clusters. Once we acquire informative samples, the pseudo feedback module leverages interacted information to deduce students' feedback, subsequently generating pseudo response labels for each mixed sample. The final module (i.e., the cognitive diagnosis module) employs the mixed exercises with pseudo labels and interaction records to deduce students' cognitive levels. Notably, our framework exhibits remarkable extensibility, seamlessly integrating supplementary data into existing methods, thereby enhancing their performance.

Sample Augmentation Module

To thoroughly augment the information encompassed within the samples for each student, we first sample more informative exercises for each student from un-interacted exercises by clustering students. Subsequently, the sampled exercises are combined with the interacted exercises to generate novel samples, facilitated by attention mechanisms.

Collaboration-aware Un-interacted Exercise Sampling This section focuses on the objective of selecting a specific number of exercises for each student s_i from the un-interacted exercise set U_i. We posit the existence of two types of exercises within U_i that offer limited supplementary information aimed at improving the accuracy of students' proficiency diagnosis. The initial type encompasses exercises related to the knowledge concepts within K_i that have been interacted with by student s_i. This is facilitated by the understanding that student s_i's proficiency with respect to the knowledge concepts in K_i can be discerned from the interaction records. The second type consists of exercises accomplished by students exhibiting similar proficiency levels, drawing inspiration from the notion of collaboration. Given that these details can be somewhat captured by prevailing cognitive diagnosis models.

Thus, we structure the sampling process in the following manner. Initially, we partition students into W groups based on their performance in exercises and the exercise-concept relational matrix Q. Subsequently, for the student s_i with an interaction set R_i of size t, we give preference to exercises that are commonly completed by peers within the remaining $W - 1$ clusters, as these exercises have garnered more feedback. In other words, for student s_i, we draw a sample of $2n$ exercises, forming the candidate set $U_i^{cand} = \{u_1, u_2, ..., u_{2n}\} \subseteq U_i$ which intentionally excludes exercises linked to the knowledge concepts in K_i, and the value of n serves as a hyperparameter.

Attention-based Sample Augmentation Using the sampled $2n$ exercises U_i^{cand} for student s_i, we combine these exercises with the interacted exercises E_i to create a newly generated sample set, thereby augmenting our samples. More precisely, for each interacted exercise $e_j \in E_i$, we randomly select n exercises (denoted $U_{i,j}^E$) from the set U_i^{cand}. Subsequently, we combine these $n + 1$ exercises, leveraging an attention mechanism to produce $n + 1$ new samples. Consequently, for student s_i who possesses an interacted exercise set E_i containing t exercises, we will generate a total of $t \times (n + 1)$ new samples. Interacted exercises consistently provide substantial information, whereas sampled un-interacted exercises offer a range of diverse and informative insights. This mixing operation serves to balance the informativeness and diversity of samples, thereby enhancing the robustness and precision of student s_i's diagnosis.

As we apply mixture to the vector representations of exercise instances, we initiate the embedding process of exercises by performing a matrix multiplication. Specifically, the one-hot vector x^{e_j} for each exercise e_j, along with x^{u_m} for exercise in $U_{i,j}^E$, is multiplied by a trainable matrix $E \in \mathbb{R}^{M \times d}$ to attain their initialized embedding representation $e_j^E, e_{u_m}^E \in \mathbb{R}^{1 \times d}$, where M is the number of exercises, d is the embedding size:

$$e_j^E = x^{e_j} \times E, \quad e_{u_m}^E = x^{u_m} \times E. \tag{1}$$

Cognitive diagnosis models commonly utilize e_j^E as the exercise e_j's feature vector. The dimensions of e_j^E correspond to the quantity of knowledge concepts, with each dimension representing an exercise attribute concerning the relevant concept. In this study, for a deeper understanding of exercises related to un-interacted knowledge concepts, it is essential to amplify the weights of correlated knowledge

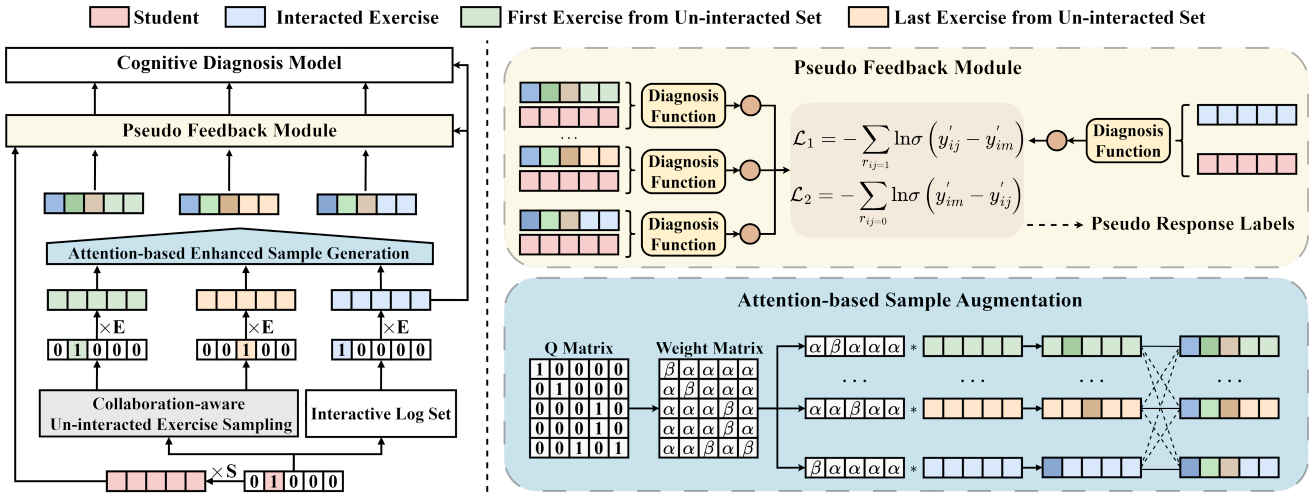

Figure 2: The overview architecture of CMES: (a) The Sample Augmentation Module, which consists of Collaboration-aware Un-interacted Exercise Sampling and Attention-based Sample Augmentation; (b) The Pseudo Feedback Module; (c) Model Training via the Cognitive Diagnosis Task.

concepts in the information mixing process. As such, we propose to construct a weight matrix based on the Q matrix denoted as $Q' \in \mathbb{R}^{\mathrm{M}\times\mathrm{C}}$, where M and C represent the number of exercises and knowledge concepts respectively:

$$Q'_{m,c} = \begin{cases} \alpha, & \text{if } Q_{m,c} = 0 \\ \beta, & \text{if } Q_{m,c} = 1 \end{cases}, \tag{2}$$

where $Q_{m,c}$ denotes exercise e_m is affiliated with knowledge concept k_c, α and β represent hyperparameters, subject to $\alpha < \beta$. Then, we multiply the knowledge concept weight vector Q'_j with the initialized embedding vector e_j^E of exercise e_j to obtain the exercise embedding vector $e_{j'}^E \in \mathbb{R}^{(n+1)\times d}$ with knowledge concept weight enhancement as follows:

$$e_{j'}^E = e_j^E \cdot Q'_j. \tag{3}$$

Based on the embedding representation vectors of exercises, for each interacted exercise $e_j \in E_i$ and the randomly selected n exercises $U_{i,j}^E$, we employ a self-attention network to mix them and obtain $n+1$ new embedding vectors. The embedding vectors of generated samples incorporate the learned target embedding as well as the information from one another. Here we adopt the *Scaled Dot-Product Attention* to capture the information among the sampled instances and interacted exercises:

$$\begin{aligned} Q, K, V =& e_{j'}^E \times W_Q,\ e_{j'}^E \times W_K,\ e_{j'}^E \times W_V \\ A_j =& softmax(\frac{Q \times K^T}{\sqrt{d}})V \end{aligned}, \tag{4}$$

where $W_Q \in \mathbb{R}^{d\times d}$, $W_K \in \mathbb{R}^{d\times d}$, and $W_V \in \mathbb{R}^{d\times d}$ are three trainable matrices. $A_j \in \mathbb{R}^{(n+1)\times d}$ is the result computed by the attention module, representing a weighted vector that captures information from other exercises. Then A_j is taken as the j-th item (i.e, $U_{i,j}^{E'}$) of the diagnosis-generated sample set $U_i^{E'} = \{U_{i,1}^{E'}, U_{i,2}^{E'}, \ldots, U_{i,j}^{E'}, \ldots, U_{i,t}^{E'}\}$, which encompasses $t \times (n+1)$ samples for student s_i.

Pseudo Feedback Module

The samples generated by the sample augmentation module lack genuine response labels, which is necessary for cognitive diagnosis models. Therefore, within this module, we introduce a learning-to-rank task (Cao et al. 2007) to deduce the corresponding pseudo response label for these generated samples, relying on the following assumption.

Assumption. We assume that the correct probability of student s_i answering the interacted sample e_j^E is greater than or equal to that of this student answering each generated sample, when $r_{ij} = 1$. Otherwise, when $r_{ij} = 0$, the ranking relationship is the opposite. This assumption is formally defined as follows:

$$\begin{aligned} P(r_{im}) \le P(r_{ij}) \le 1, r_{ij} = 1, \\ P(r_{im}) \ge P(r_{ij}) \ge 0, r_{ij} = 0, \end{aligned} \tag{5}$$

where r_{ij} is the real response label of student s_i on the exercise e_j, $r_{ij} = 1$ if student s_i correctly answered exercise e_j, otherwise $r_{ij} = 0$; $P(r_{im})$ and $P(r_{ij})$ denote the probabilities of student s_i correctly answering the exercises $e_{u'_m}^{E'} \in U_{i,j}^{E'}$ and e_j^E, respectively.

Based on the above assumption, we simply define a learning objective, which is to maximize the following function:

$$\prod_{e_{u'_m}^{E'} \in U_{i,j}^{E'}} r_{ij} \times P_i(e_j^E > e_{u'_m}^{E'}) + (1 - r_{ij}) P_i(e_{u'_m}^{E'} > e_j^E), \tag{6}$$

where $P_i(a > b)$ represents the probability of student s_i correctly answering exercise a is higher than that of correctly answering exercise b.

We employ the BPR (Bayesian Personalized Ranking) (Rendle et al. 2012) loss function to simplify the learning of the objective:

$$\begin{aligned} \mathcal{L}_{Feedback} = -\sum_{e_{u'_m}^{E'} \in U_i^{E'}} & (r_{ij} * \ln \sigma(y'_{ij} - y'_{im}) \\ & + (1 - r_{ij}) * \ln \sigma(y'_{im} - y'_{ij})) \end{aligned}, \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, y'_{ij} and y'_{im} are obtained by the diagnosis function (denoted by f_1) of the cognitive diagnosis model. Then, for each un-interacted exercise $e_{u'_m}$, we map y'_{im} into $\hat{y_{im}} \in \{0, 1\}$ as the pseudo feedback response label of student s_i on exercise $e_{u'_m}$, where 1 indicates student s_i may correctly answer the exercise $e_{u'_m}$, while 0 indicates possible wrong answers.

Cognitive Diagnosis Module

Learning Model with CMES Framework. Our framework is applicable to any prevailing cognitive diagnosis model. The Sample Augmentation Module and Pseudo Feedback Module cater to student s_i in the cognitive diagnosis model by providing personalized pairs of sampled exercises $U_i^{E'}$ and their corresponding pseudo feedback response labels.

Training. Through predicting students' proficiency levels, we derive the ultimate mastery status for each student. In addition to employing the generated sample set $U_i^{E'}$, we also leverage the interaction set R for diagnostic purposes. The loss function is defined as a composite of two components. Initially, we employ the frequently employed cross-entropy loss function (Yang et al. 2021, 2022) within conventional CD models on the data from the interaction set R:

$$\mathcal{L}^{inter} = -\sum_{(s_i, e_j, r_{ij}) \in \mathcal{R}} (r_{ij} \log y_{ij}) + (1 - r_{ij}) \log (1 - y_{ij}), \tag{8}$$

where y_{ij} represents the proficiency prediction for student s_i on exercise e_j attained through diagnosis function of the cognitive diagnosis model (denoted by f_2).

Then, we design a loss function for the mixed exercises:

$$\mathcal{L}^{un-inter} = -\sum_{s_i \in S} \frac{1}{|U_i^{E'}|} \sum_{e^{E'}_{u'_m} \in U_i^{E'}} \Big((\hat{y}_{im} \log y_{im}) + (1 - \hat{y}_{im}) \log (1 - y_{im}) \Big), \tag{9}$$

where y_{im} represents the proficiency prediction for student s_i on mixed sample $e^{E'}_{u'_m}$ attained through the cognitive diagnosis function f_2, $\hat{y}_{im}$ indicates the pseudo feedback label of student s_i on $e^{E'}_{u'_m}$. We optimize the cognitive diagnosis module using the following loss function:

$$\mathcal{L}_{CD} = \mathcal{L}^{inter} + \mathcal{L}^{un-inter}. \tag{10}$$

We optimize the entire framework using the following loss function:

$$\mathcal{L}_{CMES} = \mathcal{L}_{CD}(\Theta_1) + \alpha \cdot \mathcal{L}_{Feedback}(\Theta_2), \tag{11}$$

where α is a balancing hyper parameter that weighs the two loss functions, Θ_1 and Θ_2 represent the training parameters of the Pseudo Feedback Module and the Cognitive Diagnosis Module respectively. It is worth noting that the diagnosis functions f_1 and f_2 in the Pseudo Feedback Module and the Cognitive Diagnosis Module apply the same cognitive diagnosis model but different parameters.

Experiments

As the key contribution of this work is to extend existing cognitive diagnosis models (CDMs) to adaptively utilize un-interacted data, we compare the original CDMs and our optimized CDMs with the CMES[1] framework (denoted as Orginal-CDMs and CMES-CDMs respectively) on real-world datasets to address the following research questions:

- **RQ1:** Can CMES-CDMs outperform Original-CDMs in terms of performance?
- **RQ2:** How does our sample augmentation strategy outperform random sampling?
- **RQ3:** Whether the performance of CMES is sensitive to the setting of sampling number?
- **RQ4:** Whether the performance of CMES is sensitive to the setting of student cluster number?
- **RQ5:** How does CMES perform on different ratios of the training set?

Experimental Settings

Datasets Description. We conduct experiments on two real-world datasets ASSISTments (Feng, Heffernan, and Koedinger 2009) and Math, which both provide student-exercise interaction records and the exercise-knowledge concept relational matrix. ASSISTments is a publicly available dataset collected from the online tutoring system ASSISTments. Math is a proprietary dataset assembled by a renowned e-learning platform, comprising mathematics practice and examination records of elementary and secondary school students. For both datasets, we filter out students with less than 15 response logs to ensure sufficient data for model learning. After processing, the statistics of the two datasets are shown in Table **??**. We apply $70\% : 10\% : 20\%$ training/validation/test split for each student's response logs in the two datasets.

Statistics	ASSISTments	MATH
# Students	4,163	1,967
# Exercises	17,746	1,686
# Knowledge concepts	123	61
# Response logs	278,868	118,348
# Avg logs per student	67	60

Table 1: The statistics of the datasets.

Evaluation Metrics. Considering that there is no true knowledge mastery of students, in the literature, the mainstream approach is to indirectly evaluate the effectiveness of CDMs by using the knowledge mastery vector obtained to predict the student's exercising performance. Three famous metrics, i.e., the Root Mean Square Error (RMSE) (Tian et al. 2022), the Prediction Accuracy (ACC) (Tian et al. 2021) and Area Under an ROC Curve (AUC) (Bradley 1997) were chosen to evaluate predictive performance.

[1] https://github.com/WangCQ206/Intelligent-Education/tree/main/CMES

Metrics	ACC		RMSE		AUC	
	Orginal-CDMs	CMES-CDMs	Orginal-CDMs	CMES-CDMs	Orginal-CDMs	CMES-CDMs
IRT	68.89%	**70.59%**	0.4684	**0.4547**	70.45%	**74.40%**
MIRT	70.79%	**72.36%**	0.4634	**0.4368**	73.93%	**75.55%**
NCD	72.27%	**72.89%**	0.4335	**0.4283**	75.22%	**76.23%**
CDGK	72.08%	**73.01%**	0.4356	**0.4306**	74.83%	**75.51%**
ECD	72.47%	**72.80%**	0.4334	**0.4287**	74.97%	**76.25%**
RCD	72.99%	**73.06%**	0.4243	**0.4237**	76.40%	**76.51%**

(a) ASSISTments

Metrics	ACC		RMSE		AUC	
	Orginal-CDMs	CMES-CDMs	Orginal-CDMs	CMES-CDMs	Orginal-CDMs	CMES-CDMs
IRT	70.88%	**72.75%**	0.4505	**0.4460**	71.62%	**76.37%**
MIRT	72.99%	**74.60%**	0.4284	**0.4097**	75.31%	**78.04%**
NCD	74.13%	**74.94%**	0.4102	**0.4053**	77.14%	**78.81%**
CDGK	73.68%	**74.63%**	0.4121	**0.4068**	77.00%	**78.13%**
ECD	74.16%	**74.83%**	0.4101	**0.4077**	77.18%	**78.30%**
RCD	74.86%	**75.16%**	0.4063	**0.4055**	78.34%	**78.64%**

(b) MATH

Table 2: Experimental results on student performance prediction. The best results are highlighted in bold. Our CMES-CDMs significantly outperform the Orginal-CDMs with p <0.01.

Cognitive Diagnosis Models. To validate the effectiveness of CMES framework, we conducted the comparison experiments based on six representative CDMs, namely IRT (Lord 1980), MIRT (Reckase 2009), NCD (Wang et al. 2020a), CDGK (Wang et al. 2021), ECD (Zhou et al. 2021) and RCD (Gao et al. 2021).

Parameter Settings. We first initialized all the parameters in the networks with Xavier (Glorot and Bengio 2010) initialization and used the Adam (Kingma and Ba 2014) optimizer with a fixed batch size of 256 during the training process. For the multi-dimensional models (i.e., MIRT, NeuralCD, CDGK, ECD and RCD), we set the dimensions of latent features for both students and exercises to be equal to the number of knowledge concepts, i.e., 123 for ASSISTments and 61 for MATH datasets. Based on the parameter tuning, we set n to 20 for ASSISTments and 5 for Math respectively; we set W to 50 and 20 for ASSISTments and Math respectively. Finally, experimental results for all models are obtained by performing standard 5-fold cross-validation. The hyper-parameters of comparison approaches are tuned on the validation set according the original paper. All models are implemented in Pytorch, and all experiments are conducted on Linux servers with Tesla V100.

Performance Comparison (RQ1)

We compare six pairs of Orginal-CDMs and CMES-CDMs in terms of RMSE, ACC, and AUC. The experimental results are exhibited in Table 2. For each pair, better results are bolded. As shown in the table, for each pair, CMES-CDM outperforms Orginal-CDM in terms of all evaluation metrics on all datasets. Even for RCD that models the intrinsic correlations among knowledge concepts, our CMES can still improve its efficacy. These observations verify that our proposed CMES framework by excavating and leveraging the information within un-interacted exercises can match prevailing CDMs and boost the diagnosis performance of existing CDMs.

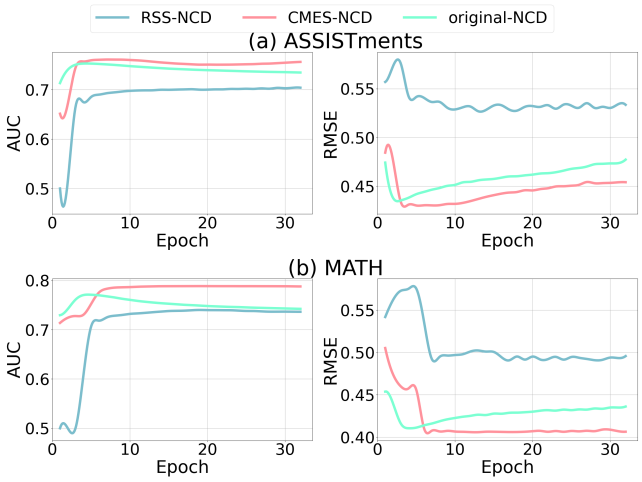

Figure 3: The comparison results between our sampling strategy CMES and the random sampling strategy (RSS).

Effectiveness of the Sampling Strategy (RQ2)

To validate the effectiveness of the sample augmentation strategy, we compare it with the random sampling strategy (RSS). RSS randomly samples exercises for each student s_i from U_i. These sampled exercises are directly used as extra training samples without the information mixture process. The randomly sampled exercises are then fed into the pseudo feedback module to assess the potential labels.

Figure 3 exhibits the comparison results among NCD

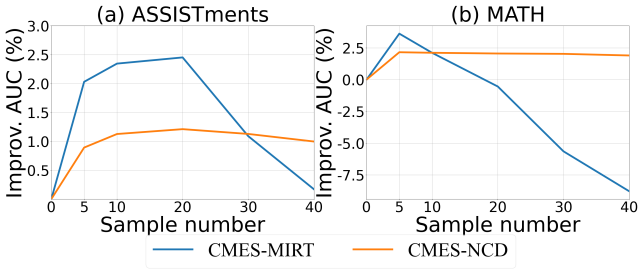

Figure 4: Impact of sampling number.

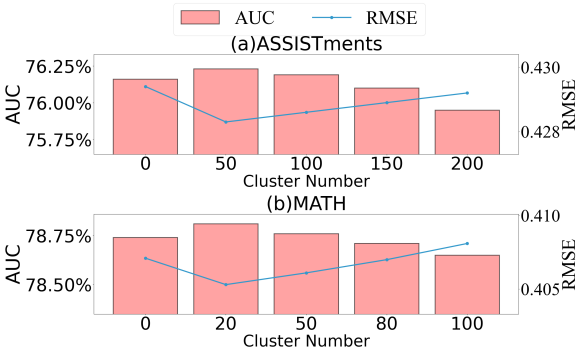

Figure 5: Impact of student cluster number.

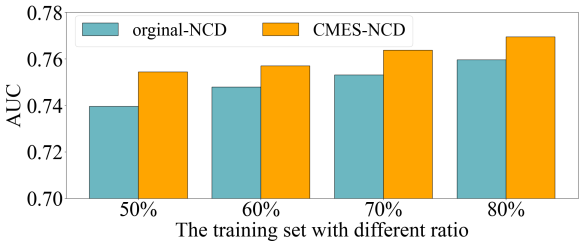

Figure 6: The training set with different ratio.

with our sample strategy (namely CMES-NCD), NCD with the random sampling strategy (namely RSS-NCD) and the original NCD (original-NCD). CMES-NCD markedly surpasses RSS-NCD and original-NCD across all metrics on both datasets, while the random sampling strategy deteriorates model performance. The information gathered from these randomly sampled exercises might lack diversity, resulting in ineffective diagnostic values. Even the redundant exercises delude the cognitive diagnosis model, hindering it from accurately diagnosing students' cognitive states. In contrast, the proposed CMES performs collaboration-aware sampling and mixes the information of exercises, enriching diagnostic information to enable the model to infer students' cognitive states more comprehensively.

Sensitivity Analysis of Sampling Number (RQ3)

We choose two representative CDMs (i.e., MIRT and NCD) combined with our CMES framework to investigate the performance change when varying the sampling number for each student (i.e., the parameter n) in the range of $\{5, 10, 20, 30, 40\}$. As shown in Table 4, on the ASSISTments, the optimal performance of CMES-MIRT and CMES-NCD is achieved at $n = 20$, while their peak performance is reached at $n = 5$ on the Math. It maybe attributed to the exercise pool size of the two datasets. As shown in Table **??**, the number of exercise in ASSISTments is more larger than that of Math. The performance of CMES-MIRT and CMES-NCD starts to decrease when $n > 20$ and $n > 5$ on ASSISTments and MATH respectively. The degradation is more significant for MIRT, because the simple student-exercise interaction function in MIRT model cannot capture fine-grained exercise information. Excessive exercises confuse the diagnosis model and lead to negative optimization.

Sensitivity Analysis of Student Cluster (RQ4)

We further used NCD combined with our CMES to probe the impact of the number of clusters W. Here we search W in the range of $\{0, 50, 100, 150, 200\}$ and $\{0, 20, 50, 80, 100\}$ for ASSISTments and MATH respectively. As depicted in Figure 5, the optimal values for W are set to 50 and 20 for ASSISTments and MATH. From this observation, on the one hand, the optimal setting for W seems to be related to the student size, as shown in Table **??**, the student number in ASSISTments is larger than that in Math. On the other hand, inappropriate setting for the student cluster number W will result in significant performance degradation, which manifests that the performance of CMES is sensitive to the setting of student cluster number and answers RQ4.

Case Study (RQ5)

We selected 20% of the dataset ASSISTments as the test set, and utilized data sets with sizes of 80%, 70%, 60% and 50% of the full dataset from the remaining data to train model respectively. As depicted in Figure 6, we observe that CMES-NCD trained by different size of training sets all demonstrate excellent performance. The performance of CMES-NCD trained on 60% of the data is on par with orginal-NCD trained on 80%. Additionally, the enhancement attained by CMES-NCD is most pronounced when trained on 50% of the data, which surpasses the performance of orginal-NCD trained on 70% of the data, and nears its performance trained on 80%. These observations validate that our CMES framework can mitigate data scarcity challenges by extracting more information from un-interacted exercises.

Conclusion

In this work, we attempted to explore informative, un-interacted exercises related to broader knowledge concepts with the aim of providing a more comprehensive diagnostic assessment of students. We proposed a generic framework **CMES** (**C**ollaborative-aware **M**ixed **E**xercise **S**ampling) that enables sampling of rich information from un-interacted exercises and facilitates the evaluation of potential true labels. Experimental results on real-world datasets demonstrate the effectiveness of the sampling strategy and the scalability of our framework. We intend to further investigate sampling strategies tailored to the characteristics of cognitive diagnostic models.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62107001, No. U21A20512, and No.62302010), in part by the Anhui Provincial Natural Science Foundation (NO.2108085QF272), in part by the University Synergy Innovation Program of Anhui Province (NO.GXXT-2021-004), in part by the Key Research and Development Project of Qinghai Province (NO.2023-GX-C13), and in part by the China Postdoctoral Science Foundation (No.2023M740015).

References

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7): 1145–1159.

Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, 129–136.

Caselles-Dupré, H.; Lesaint, F.; and Royo-Letelier, J. 2018. Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 352–356.

Chen, T.; Sun, Y.; Shi, Y.; and Hong, L. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 767–776.

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

Ding, J.; Quan, Y.; He, X.; Li, Y.; and Jin, D. 2019. Reinforced Negative Sampling for Recommendation with Exposure Data. In *IJCAI*, 2230–2236. Macao.

Feng, M.; Heffernan, N.; and Koedinger, K. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3): 243–266.

Gao, W.; Liu, Q.; Huang, Z.; Yin, Y.; Bi, H.; Wang, M.-C.; Ma, J.; Wang, S.; and Su, Y. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 501–510.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

Guo, G.; Zhou, H.; Chen, B.; Liu, Z.; Xu, X.; Chen, X.; Dong, Z.; and He, X. 2020. Ipgan: Generating informative item pairs by adversarial sampling. *IEEE transactions on neural networks and learning systems*, 33(2): 694–706.

Huang, T.; Dong, Y.; Ding, M.; Yang, Z.; Feng, W.; Wang, X.; and Tang, J. 2021. Mixgcf: An improved training method for graph neural network-based recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 665–674.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Liu, B.; and Wang, B. 2023. Bayesian Negative Sampling for Recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 749–761. IEEE.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. LAWRENCE ERLBAUM ASSCCIAATES.

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

Ma, H.; Zhu, J.; Yang, S.; Liu, Q.; Zhang, H.; Zhang, X.; Cao, Y.; and Zhao, X. 2022. A Prerequisite Attention Model for Knowledge Proficiency Diagnosis of Students. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4304–4308.

Qi, T.; Ren, M.; Guo, L.; Li, X.; Li, J.; and Zhang, L. 2023. ICD: A new interpretable cognitive diagnosis model for intelligent tutor systems. *Expert Systems with Applications*, 215: 119309.

Qin, C.; Zhang, L.; Zha, R.; Shen, D.; Zhang, Q.; Sun, Y.; Zhu, C.; Zhu, H.; and Xiong, H. 2023. A Comprehensive Survey of Artificial Intelligence Techniques for Talent Analytics. *arXiv preprint arXiv:2307.03195*.

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

Rendle, S.; and Freudenthaler, C. 2014. Improving pairwise learning for item recommendation from implicit feedback. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 273–282.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Tian, Y.; Pan, J.; Yang, S.; Zhang, X.; He, S.; and Jin, Y. 2022. Imperceptible and Sparse Adversarial Attacks via a Dual-Population-Based Constrained Evolutionary Algorithm. *IEEE transactions on artificial intelligence*, 4(2): 268–281.

Tian, Y.; Peng, S.; Yang, S.; Zhang, X.; Tan, K. C.; and Jin, Y. 2021. Action Command Encoding for Surrogate-Assisted Neural Architecture Search. *IEEE transactions on cognitive and developmental systems*, 14(3): 1129–1142.

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 AAAI conference on artificial intelligence*, volume 34, 6153–6161.

Wang, J.; Yu, L.; Zhang, W.; Gong, Y.; and Zhang, D. 2017. IRGAN: A minimax game for unifying generative and discriminative information retrieval models. *ACM SIGIR FORUM*.

Wang, X.; Huang, C.; Cai, J.; and Chen, L. 2021. Using knowledge concept aggregation towards accurate cognitive diagnosis. In *Proceedings of the 30th ACM Interna-*

tional Conference on Information & Knowledge Management, 2010–2019.

Wang, X.; Xu, Y.; He, X.; Cao, Y.; Wang, M.; and Chua, T.-S. 2020b. Reinforced negative sampling over knowledge graph for recommendation. In *Proceedings of the web conference 2020*, 99–109.

Yang, S.; Ma, H.; Zhen, C.; Tian, Y.; Zhang, L.; Jin, Y.; and Zhang, X. 2023a. Designing novel cognitive diagnosis models via evolutionary multi-objective neural architecture search. *arXiv preprint arXiv:2307.04429*.

Yang, S.; Tian, Y.; He, C.; Zhang, X.; Tan, K. C.; and Jin, Y. 2021. A gradient-guided evolutionary approach to training deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9): 4861–4875.

Yang, S.; Tian, Y.; Xiang, X.; Peng, S.; and Zhang, X. 2022. Accelerating Evolutionary Neural Architecture Search via Multifidelity Evaluation. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4): 1778–1792.

Yang, S.; Wei, H.; Ma, H.; Tian, Y.; Zhang, X.; Cao, Y.; and Jin, Y. 2023b. Cognitive diagnosis-based personalized exercise group assembly via a multi-objective evolutionary algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Yang, S.; Yu, X.; Tian, Y.; Yan, X.; Ma, H.; and Zhang, X. 2023c. Evolutionary Neural Architecture Search for Transformer in Knowledge Tracing. *arXiv preprint arXiv:2310.01180*.

Yang, S.; Zhen, C.; Tian, Y.; Ma, H.; Liu, Y.; Zhang, P.; and Zhang, X. 2023d. Evolutionary Multi-Objective Neural Architecture Search for Generalized Cognitive Diagnosis Models. In *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 1–10. IEEE.

YAO, F.; Huang, Z.; Hou, M.; Tong, S.; Liu, Q.; Chen, E.; Sha, J.; and WANG, S. 2023. Exploiting Non-Interactive Exercises in Cognitive Diagnosis. *IJCAI 2023*.

Zhang, W.; Chen, T.; Wang, J.; and Yu, Y. 2013. Optimizing top-n collaborative filtering via dynamic negative item sampling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 785–788.

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