Exploiting Non-Interactive Exercises in Cognitive Diagnosis

Fangzhou Yao1,2, Qi Liu1,2 *, Min Hou1,2, Shwei Tong1,2, Zhenya Huang1,2, Enhong Chen1,2, Jing Sha3, Shijin Wang2,3

1Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China
2State Key Laboratory of Cognitive Intelligence
3iFLYTEK AI Research (Central China), iFLYTEK Co., Ltd.

Abstract

Cognitive Diagnosis aims to quantify the proficiency level of students on specific knowledge concepts. Existing studies merely leverage observed historical students-exercise interaction logs to access proficiency levels. Despite effectiveness, observed interactions usually exhibit a power-law distribution, where the long tail consisting of students with few records lacks supervision signals. This phenomenon leads to inferior diagnosis among few records students. In this paper, we propose the Exercise-aware Informative Response Sampling (EIRS) framework to address the long-tail problem. EIRS is a general framework that explores the partial order between observed and unobserved responses as auxiliary ranking-based training signals to supplement cognitive diagnosis. Considering the abundance and complexity of unobserved responses, we first design an Exercise-aware Candidates Selection module, which helps our framework produce reliable potential responses for effective supplementary training. Then, we develop an Expected Ability Change-weighted Informative Sampling strategy to adaptively sample informative potential responses that contribute greatly to model training. Experiments on real-world datasets demonstrate the supremacy of our framework in long-tailed data.

1 Introduction

Cognitive diagnosis has been increasingly needed to assess and improve individual development in intelligent educational applications [Liu, 2021; Zhou et al., 2021]. Given the historical interaction logs of students, it aims to discover their latent cognitive states (proficiency levels) on knowledge concepts and reveal some exercise features such as difficulty and discrimination [Pandey and Karypis, 2019; Wu et al., 2020; Tong et al., 2021].

Existing cognitive diagnosis models (CDMs) mainly focus on assessing students’ proficiency level based on historical student-exercise interaction logs like Item Response Theory (IRT) [Lord, 1980] and NeuralCD [Wang et al., 2020]. However, in real scenarios, a large number of students interact with very few exercises [Lu et al., 2022], which leads to the fact that the distribution over students is quite imbalanced and even long-tailed. In Figure 1(a), we sort 10000 students randomly sampled from the well-known dataset Junyi (math practicing logs, description in Section 5.1) by the number of interaction times in a descending order. We notice a heavy long-tailed distribution that nearly 90% students’ interaction times are less than 50. The proposed works pay little attention to this problem. They validate model’s performance on datasets by filtering students with few interactions to ensure enough logs for executing diagnostic tasks, which runs counter with the task of diagnosing each student’s knowledge level. In Figure 1(b), we evaluate recent advanced CDMs’ performance with accuracy on different student groups by interaction times. It shows that all models exhibit lower accuracy in students group with limited interactions compared to students with plentiful interactions (more than 50 times). From this observation, it is reasonable to presume that limited interactions lead to inaccurate and uncertain diagnosis results. In other words, insufficient supervised signals result in poor robustness of CDMs in students with few interactions.

To tackle this problem, we naturally come up with incorporating auxiliary training signals from non-interactive exercises, and the relationship between student-exercise knowledge can be used to infer potential signals. Specifically, a student probably performs similarly on exercises that share
the same knowledge concept. An example is shown in Figure 2, student $s$ has a correct response to exercise $e_4$ and an incorrect response to exercise $e_3$. $s$ may have a higher proficiency level towards the knowledge concept $k_3$ behind $e_4$ than $k_1$ behind $e_3$. Thus, we may infer a partial order relationship that the probability of correctly answering $e_6$ with $k_3$ is much higher than $e_2$ with $k_1$. The partial order relationship has great potential in constructing additional training signals. Unfortunately, there are still many challenges in designing an effective solution to exploit this relationship and guide the training signal sample on unobserved data. On one hand, how to identify reliable signals from the massive potential partial order pairs is still an open issue. Due to the absence of ground-truth responses, partial order pairs bring information meanwhile noise. For example, a student’s response to an exercise with the same knowledge concept as a previously correctly responded one may be deemed as correct. However, the actual signal may be incorrect due to the new exercise being significantly more difficult than the previously responded one. On the other hand, how to effectively select informative signals is a non-trivial problem. When exploring the partial order pairs towards possible training signals, the scale of training samples increases dramatically. A common solution is randomly sampling instances from unobserved data [Rendle et al., 2012]. However, prior works [Rendle and Freudenthaler, 2014; Lian et al., 2020; Qin et al., 2020] demonstrate that it makes the model converge slowly, especially when the pool of instances is large. In order to speed up convergence, we need to design an efficient sampling strategy.

In this work, we propose a general framework called Exercise-aware Informative Response Sampling (EIRS) to overcome the long-tailed distribution problem in cognitive diagnosis. EIRS provides reliable and informative auxiliary training signals that can be seamlessly incorporated into existing cognitive diagnosis models. To achieve this goal, we first explore the partial-order relationship between student-exercise-knowledge in depth. We then design the Exercise-aware Candidates Selection (ECS) module to ensure the reliability of the training signals. Two indicators that measure the difficulty or discrimination of exercises are proposed to help identify reliable training pairs. For discriminating informative signals, we introduce Expected Ability Change weighted Informative Sampling strategy (EIS). EIS adaptively selects samples from two perspectives including contribution to model training and student ability change. Experiments on two real-world datasets demonstrate that the proposed EIRS improves CDMs’ robustness on long-tailed data and speeds up model convergence.

2 Related Work

2.1 Cognitive Diagnosis

The task of evaluating students’ knowledge level from historical response logs has been studied since 1950s. Item Response Theory (IRT) [Lord, 1980] is a standard statistical model cognitive diagnosis, which uses single-dimension variables to represent the trait features and logistic function. Later, Multidimensional Item Response Theory (MIRT) [Reckase, 2009] proposed to use multidimensional trait features instead of single dimension. Another classic model, Noisy “And” gate model (DINA) [De La Torre, 2009] diagnoses the mastery state by binary variables and considers students’ slip and guessing factors. Recently, the prevalence of deep learning motivates a rich line of work on cognitive diagnosis. Recently, some researchers introduce the deep learning into cognitive diagnosis [Wang et al., 2020; Ma et al., 2022; Gao et al., 2021]. Wang et al. proposed NeuralCD framework to learn the interaction function between students and responses with neural networks. Furthermore, Gao et al. designed a multi-layer student-exercise-concept relation map to model the interactive and structural relations. The methods described above learn trait parameters from entire historical response logs, which can suffer from a long-tailed problem and worsen across students’ diagnoses when there are insufficient supervision signals.

2.2 Sampling Strategy

One key component of our framework is to sample potential responses for the observed response anchor, which is most relevant to sampling strategy technology applied in some domains like natural language processing [Mikolov et al., 2013], recommendation [Rendle and Freudenthaler, 2014; Qin et al., 2019], etc. Static sampling strategies sample unobserved data based on a predefined distribution, such as uniform and popularity distribution corresponding to random sampling [Rendle et al., 2009] and popularity-based sampling [Caselles-Dupr´e et al., 2018; Mikolov et al., 2013] respectively. However, static methods cannot adjust to model training, suffering from low quality of samples. Adaptive sampling was proposed later, such as DNS [Zhang et al., 2013] which dynamically selects hard samples that are difficult for current model to discriminate. Inspired by generative adversarial learning [Goodfellow et al., 2014], some researchers have studied adversarial training between the sampling model (the generator) and the training model (the discriminator) [Wang et al., 2018; Park and Chang, 2019]. For example, Park [2019] proposed AdvIR to generate hard negatives by adding adversarial perturbations to them. However, since student-exercise responses depend on complex features such as knowledge concepts, difficulty and discrimination of exercises [Liu et al., 2021], sampling task in cognitive diagnosis is more challenging than sampling in other scenarios.
3 Problem Definition

Here we give a formal definition of cognitive diagnosis. Let $S = \{s_1, s_2, ..., s_N\}$ be the set of $N$ students, $E = \{e_1, e_2, ..., e_M\}$ represent the set of $M$ exercises, and $K = \{k_1, k_2, ..., k_L\}$ denote the set of $L$ knowledge concepts. We define the student-exercise interaction set of the entire space as $R = S \times E$. The observed interaction logs are a triplet set $R_O = \{(s, e, r_{se})| (s, e) \in R, r_{se} \in \{0, 1\}\}$ where $r_{se}$ represents a student’s response to an exercise (i.e., 0 indicates wrong answer while 1, otherwise). The number of interaction logs is much smaller than that of $|R|$, that is $|R_O| \ll |R|$. Besides, we have Q-matrix [Tatsuoka, 1995] labeled by experts, $Q = \{Q_{ij}\}_{M \times L}$, where $Q_{ij} = 1$ indicates the exercise $e_i$ relates to the knowledge concept $j$ and $Q_{ij} = 0$ otherwise.

Given: Students’ interactions logs $R_O$ and Q-matrix $Q$.

Goal: Quantify students’ knowledge level on specific knowledge concepts by modeling the student performance prediction process.

4 Methodology

In this section, we introduce Exercise-aware Informative Response Sampling (EIRS) framework which could be applied to all existing CDMs. In the following parts, we will first introduce the backbone cognitive module with the optimization task. Then we will explain the shortcomings of the current optimization task and show how to leverage the partial order to formulate a new ranking optimization task. After that, we will dive into the details of our proposed partial-order response sampling strategy and the learning algorithm.

4.1 Basic Cognitive Diagnosis Model

Cognitive diagnosis model (CDM) is for assessing students’ proficiency level according to their observed responses to exercises. Generally, CDM contains two steps: (1) the embedding layer to obtain the diagnostic factors of students and exercises, (2) the interactive layer to learn the interaction function among the factors and output the probability of correctly answering the exercises. After training, we get students’ proficiency vectors from the first step as diagnostic results.

Formally, given the student set $S$, exercise set $E$, and knowledge concept set $K$, through corresponding embedding-lookup layer, we represent them as $H_S \in \mathbb{R}^{N \times d}, H_E \in \mathbb{R}^{M \times d}, H_K \in \mathbb{R}^{L \times d}$. Each row of trainable metrics represents the representation of trait features (e.g., $h_s$ is the $s^{th}$-row of $H_S$ that represents the student $s$’s proficiency). For an exercise $e$, its difficulty $h_{diff}^{e}$ and discrimination $h_{dis}^{e}$ are two important characteristics, we further denote $h_e = [h_{diff}^{e} | h_{dis}^{e}]$. Here, we diagnose the cognitive state of student $s$ as $h_s$ and the characteristic of exercise $e$ as $h_e$. To verify the diagnosis, an interaction function $f_C$ is used to predict whether the student can answer the exercise correctly:

$$y_{se} = f_C(h_s, h_e),$$

where $y_{se}$ is the probability of the student $s$ correctly answering the exercise $e$. The architecture of the embedding layer and the interaction function $f_C$ can be arbitrary, all existing CDMs can be chosen, such as IRT [Lord, 1980], MIRT [Reckase, 2009], NeuralCD [Wang et al., 2020], etc.

When training the CDM, for each record in the observed response logs set $R_O = \{(s, e, r_{se})| s \in S, e \in E, r_{se} \in \{0, 1\}\}$, we calculate the loss function of basic cognitive diagnosis model as the cross-entropy loss between the prediction score $\hat{y}_{se}$ and the true label $r_{se}$.

$$L_P = - \sum_{(s, e, r_{se}) \in R_O} (r_{se} \log \hat{y}_{se} + (1 - r_{se}) \log(1 - \hat{y}_{se})).$$

The model is fit to predict the observed correct responses with value 1 and the incorrect responses with value 0. However, this can be problematic when there are not enough observed interactions, particularly for long-tailed students.

4.2 Partial-Order Ranking

In fact, optimizing the backbone cognitive diagnosis model only through traditional prediction tasks may not provide sufficient training signals, resulting in inferior diagnostic performance. Therefore, we aim to exploit the massive unobserved data. We propose creating item pairs from unobserved interactions as auxiliary training data and optimizing for correctly ranking item pairs instead of only scoring single observed item in cognitive diagnosis. In this section, we give the formulation of partial-order ranking learning.

For each student, we denote $E_I, E_U$ as the interactive and non-interactive exercises set. Thus, the interactive (non-interactive) exercises can be divided into positives $E_O^+(E_U^+)$ and negatives $E_O^-(E_U^-)$ based on the responses or potential responses to them, where +, - represent correct and incorrect responses respectively. According to the monotonicity theory [Rosenbaum, 1984] declaring that a learner’s proficiency is monotonic with the probability of correctly responding to a test item. It can be inferred that a student’s proficiency on a correct response is higher than an incorrect one.

Formally, for any exercises $e_o^+, e_o^-, e_u^+, e_u^-$ taken from $E_O^+, E_U^+, E_O^-, E_U^-$, respectively, we have the following partial order between interactive and non-interactive exercises:

$$y_o^+ > y_o^-, y_u^+ > y_u^-,$$

where $y_o^+ > y_o^-$ means that CDM should give a higher score to the observed correct response to exercise $e_o^+$ than the potential incorrect response to exercise $e_o^-$, $y_u^+ > y_u^-$ similarly. Inspired by the great success of the BPR loss [Rendle et al., 2012] in recommender systems which is defined as maximizing the difference between the predicted probability of a positive pair and a negative pair, we formulate the following constraint based on this theory:

$$L_R = - \sum_{s, e_o^+, e_u^-} \ln \sigma(y_o^+ - y_u^-) - \sum_{s, e_o^-, e_u^+} \ln \sigma(y_o^- - y_u^+).$$

4.3 Partial-Order Response Sampling

A key concern in optimizing cognitive models by ranking-based auxiliary training is how to construct reasonable partial-order pairs. The pivotal step is to screen out effective exercises from a large number of non-interactive exercises. If we randomly select non-interactive exercises, the corresponding partial order pairs will bring noise and limited information to the model, which can results in inaccurate
Given a large number of non-interactive logs set $R_U$, we will describe these two modules in detail.

**Exercise-Aware Informative Sampling Framework**

![Exercise-aware Informative Sampling Framework](image)

To validate this assumption, we conduct a hypothesis test on two datasets ASSISTments and Junyi (data description in Section 5.1). Specifically, we first give some important notions used in our testing without loss of generality. Student’s response results (right or wrong) of exercises $e_u$ are notated with $r_u$ and $r_b$, and $K_u, K_b$ are the knowledge concepts where $r_{se}$ is consistent with the student’s response in interactive logs $(s, e, r_{se})$ with overlapping knowledge concept.

Secondly, other vital properties of exercises like difficulty and discrimination have a huge impact on students’ responses. We discuss previously in Figure 2, both $e_5$ and $e_6$ have overlapping knowledge concept $k_3$ with $e_4$, while $e_5$ may be too difficult for the student to answer correctly based on his current knowledge states. We take into account the inherent properties (e.g., difficulty and discrimination) of exercises, then calculate the similarity between exercises as the sampling probability for each non-interactive exercise:

$$p(e_u|s, e) = \frac{h_{e_u}^T \cdot h_{e_i}}{\sum_{e_i \in N(e)} h_{e_i}^T \cdot h_{e_i}},$$

where $h_e = [h_e^{diff}, h_e^{dis}]$ represents exercise $e$’s characteristic factor from the basic CDM. A higher probability $p(e_u|s, e)$ score reflects a higher reliability level of the potential response based on observed responses. In this way, we sample some unobserved logs $R_u$ according to the probability and combine the unobserved logs set generated by each log to obtain a reliable candidate logs set $C_U = \bigcup_i R_u^i$. After acquiring the reliable candidate response logs set, we can select samples from the set $C_U$ to pair with observed responses for partial-order ranking learning.

It is worth noting that we used the key knowledge concept of the exercise as a criterion for candidate selection. Thus we suppose Consistency Assumption by student-exercise interaction and validate it on two real datasets.

**Consistency Assumption.** Students’ responses to similar exercises with overlapping knowledge concept are consistent.

To validate this assumption, we conduct a hypothesis test on two datasets ASSISTments and Junyi (data description in Section 5.1). Specifically, we first give some important notions used in our testing without loss of generality. Student’s response results (right or wrong) of exercises $e_u$ are notated with $r_u$ and $r_b$, and $K_u, K_b$ are the knowledge concepts...
Datasets  | $P_u$ (s) | $\alpha$ | p-value |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSISTments</td>
<td>0.638</td>
<td>0.05</td>
<td>1.89-74</td>
</tr>
<tr>
<td>Junyi</td>
<td>0.707</td>
<td>0.05</td>
<td>1.49-60</td>
</tr>
</tbody>
</table>

Table 1: One-sided test on real datasets.

that $e_a$, $e_b$ contain. Then, $P_{un}(r_a = r_b|K_a \cap K_b \neq \emptyset)$ represents the probability that student $u_m$ responses to $e_a$ and $e_b$ with overlapped knowledge are consistent. Let $P_{u1}, \ldots, P_{um}$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution, where $\mu$ is unknown and $\sigma^2$ is known.

Generally, if our assumption is valid, the probability $P_{un}$ should be over 0.5. Therefore, we perform a one-tailed test of significance with the null and alternative hypothesis:

$$H_0 : \mu \leq 0.5; H_1 : \mu > 0.5,$$

(7)

The mean of $P_{un}$ is noted as $\bar{P}_u$, and $\alpha$ is the level of significance. The $p$-value represents the probability of observing a given event under the null hypothesis. The testing results are reported in Table 1, and we observe that the $p$-values are much smaller than $\alpha$ on both datasets, so we reject the null hypothesis and accept the alternative hypothesis. Consequently, it can be concluded that students’ responses to exercises with overlapping knowledge are consistent.

**Expected Ability Change Weighted Informative Sampling Module**

Although we acquire reliable unobserved response candidates set $C_U$, the size of candidates is still large, which causes the inefficiency problem and it is impractical to traverse over the whole data to obtain the gradients. In that, we design the Expected Ability Change-weighted Informative Sampling strategy (EIS) to speed up the convergence with informative samples. This strategy evaluates the informativeness of potential responses from two perspectives: 1) contribution to model training, 2) student ability change.

1) Contribution to Model Training. In partial order ranking, we need to the sample potential exercises $e_o$ based on their responses from the candidates set $C_U$ to pair with observed responses to exercise $e_o$, while it is highly possible to sample low-quality instances. Since these responses already have a large gap with observed response, sampling them as potential responses hardly changes model parameters. Here, we want to select informative instances which will significantly change model parameters through partial order ranking. Following previous work [Rendle and Freudenthaler, 2014; Lian et al., 2020], we measure a sample’s contribution to ranking task by the gradient magnitude based on the objective function of pairwise ranking (Eq.4):

$$\Delta_{s, e_o, e_o} = 1 - \sigma(y_{so} - y_{so}), e_o \in E_O, e_o \in C_U,$$

(8)

which indicates that a response difficulty distinctly different from the opposite response (i.e. $\hat{y}_{so} - \hat{y}_{so} \to 0$) contributes much to gradient (i.e. $\Delta_{s, e_o, e_o} \to 1$). For example, if the current observed response is correct, incorrect samples with higher prediction scores make a greater contribution to the optimization. Then we define candidates’ difficulty level by the gap of their prediction scores in CDMS:

$$INF_u = |\hat{y}_{so} - \hat{y}_{su}|.$$

(9)

A smaller $INF_u$ for $e_o$ indicates a higher difficulty for CDMS to identify it from the known response to $e_o$. During training, we can reserve the top $k$ informative samples by their $INF$ values to faster training process.

2) Expected Ability Change. Beside significant contribution to partial ranking, we also aim to diagnose students’ knowledge states as soon as possible. Therefore, we hypothesize that if student feature $H_s$ undergoes a significant change by adding the unobserved response, the responded exercise can be considered informative.

However, it is challenging to calculate because the response to non-interactive exercise is unknown. Inspired by recent works [Bi et al., 2020], we calculate each sample’s importance by expected ability change (EAC). To formulate this, let $\Delta H_{seu}$ be the ability change of the target student $s$, $H_s(R_o)$ denote the student’s current ability with observed response $R_o$ and $H_s(R_o \cup r_{su})$ represent adding the unobserved response $r_{su}$. As such, the EAC weight is defined as follows:

$$\Delta H_{seu} = E_{r_{su} \sim f(c, h_{su})}[H_s(R_o \cup r_{su}) - H_s(R_o)],$$

$$w_u = \frac{\exp(\Delta H_{seu})}{\sum_u \exp(\Delta H_{seu})},$$

(10)

By considering the weight $w_u$ of response sample $e_o$, we reformulate the ranking loss in the following way:

$$L_R = - \sum_{s, e_o, e_o} w_u \cdot \ln\sigma(y_{so} - y_{so}) - \sum_{s, e_o, e_o} w_u \cdot \ln\sigma(y_{su} - y_{so}),$$

(11)

which enables our framework to pay more attention to instances that bring about a greater change to student ability.

### 4.4 Learning Algorithm

The observed responses can help optimize CDMS’ parameters by the objective of the basic cognitive model. In addition, we propose a response sampling strategy to sample reliable and informative unobserved responses. Thus, the partial order training signals can alleviate the long tail problem. Combining the prediction loss $L_P$ of the Basic Cognitive Diagnosis Model and the ranking loss $L_R$ of the Partial-Order Response Sampling Strategy, we obtain the complete loss function:

$$L = L_P + \lambda \cdot L_R,$$

(12)

**Algorithm 1 Exercise-aware Informative Response Sampling**

**Input:** Training Set $R_O = \{(s, e, r_{ui})\}$, Q-matrix

**Output:** CDMS Parameters $\Theta_C$

Initialize parameters randomly;

1: while not converge do
2: Sample a mini-batch $R_B \in R_O$ of size B.
3: for each observed response logs $(s, e, r_e)$ do
4: Get candidates set $C_U$ for student $s$ based on Eq.6
5: Select top $k$ instances from $C_U$ based on Eq.9
6: Evaluate instances’ quality $w_u$ based on Eq.10
7: Update parameters $\Theta_C$ w.r.t. Eq.12
8: end for
9: end while
### Baselines.
To evaluate the performance of our EIRS framework, we use four well-known CDMs as baseline methods: IRT [Lord, 1980], MF [Toscher and Jahrer, 2010], MIRT [Reckase, 2009] and NeuralCD [Wang et al., 2020]. In multidimensional models (i.e., MIRT and NeuralCD), we set the dimension of latent trait features of both student and item unitedly as the number of knowledge concepts, i.e., 112 in ASSISTments and 39 in Junyi. Furthermore, we compare popularity-based sampling (PopRS) method [Mikolov et al., 2013] with EIRS which calculates each exercise’s popularity based on the response rate in all students.

### Experimental Setup.
In our framework, we set the sample number from \{1,2,3,4,5\}. For the curriculum coefficient \(\lambda\), its initial value \(\lambda_0\) is chosen from the interval \((0, 1]\), and \(\lambda\) linearly increases from \(\lambda_0\) to 1 as the number of epochs increases. We employ the Adam algorithm [Kingma and Ba, 2015] for optimization, and all the hyper-parameters are tuned in the validation datasets. Our code is available at https://github.com/fannazya/EIRS.

### Evaluation Metrics.
Because the true knowledge proficiency is unknown, to directly evaluate the performance of a CDM is difficult. Following previous works, a reasonable solution is to measure performance by the prediction scores in diagnosis models as the diagnostic results can be acquired through learners performance prediction task. Here, we evaluate the model based on some classification and regression metrics such as AUC, Accuracy and RMSE.

### 5.2 Experimental Results

#### Performance Comparison (RQ1)
In order to validate the generality and effectiveness of the EIRS framework, we incorporate it into different existing CDMs and compare the performances both on the long-tailed data and whole data with baseline methods. The long-tail data extracted from test data consists of response logs of students whose interaction times are less than 50. Table 2 shows the results of EIRS with baselines, where ‘Origin’, ‘EIRS’ represent the baseline and baseline incorporating our framework respectively. The best results are shown in bold. There are the following findings. First, almost all the baselines’
results on the long-tailed data are inferior to those on the whole data, which indicates it is difficult to diagnose with few interaction logs. Second, both PopRS and EIRS perform better than the Origin baseline, showing that incorporating ranking-based training signals alleviates the long-tailed problem. Specifically, EIRS significantly outperforms both the original model and popularity sampling method on the long-tailed data, and has a fair contribution to improving performance on the whole data. Some results in the long-tailed data are even as good as those in the whole data. Hence, we are able to conclude from these observations that our framework is effective to alleviate the long-tail problem by adding auxiliary ranking-based training signals. Simultaneously, the non-interactive exercises selected by EIRS are reliable enough to provide proper training signals.

**Performance on Long-Tailed Data (RQ2)**

Our framework focuses on strengthening the robustness of CDMs, especially in long-tailed data where students interact with few exercises. We further divide them by interaction times into several groups to see the detailed improvements in different groups. The results are reported in Figure 4. We can see that the baseline MF and IRT applied with our framework was improved a lot, especially in students whose interaction times are very few (from 5 to 15). Therefore, we can conclude that EIRS’s contribution on diagnosing students knowledge level with limited response logs is prominent.

**Efficiency of Informative Sampling (RQ3)**

The above experiments show the effectiveness of samples generated by EIRS. Additionally, we compare the performance in terms of efficiency by the convergence speed of model training (the number at which an early stop occurs). We can observe that there is not much difference in convergence speed at the beginning. The reason for this is that we take a curriculum learning way (Section 4.4), where basic cognitive diagnosis is the dominant task at the beginning of training. As the epoch number increases, the test loss converges quickly when partial order ranking plays a vital role. The fast convergence speed indicates that the efficacy of EIS module to sample informative responses that bring much change to model learning.

**Effects of Sampling Number (RQ4)**

The key component in EIRS is the two-stage sampling consisting of ECS and EIS modules (Section 4.3) to generate reliable and informative exercises. To verify our sampling strategy’s superiority, we compare it with popularity-based sampling (PopRS). We vary the number of samples $K$ from set \{1, 2, 3, 4, 5\} to see the effects on the whole data. As shown in Figure 6, EIRS achieves high performance when the sample number is small and consistently outperforms PopRS at different sample numbers. Besides, our framework basically holds steady though the number changes. These give credit to the ECS module which generates reliable candidate logs through knowledge relations and exercises similarity, effectively avoiding data noise in unobserved data.

**6 Conclusion**

In this paper, we designed a general sample framework EIRS that exploits reliable and informative non-interactive data and can be seamlessly incorporated to existing cognitive diagnosis methods. Then we did experiments on two real datasets to validate the performance. The results showed the essence of non-interactive data and the superiority of our proposed framework on long-tailed data. We hope this work can stimulate more studies in the future leading to a prolonged period.
Acknowledgements
This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901003), the National Natural Science Foundation of China (Grants No. 62106244), the Students’ Innovation and Entrepreneurship Foundation of USTC (No. XY2022X12), the Open Research Fund of the State Key Laboratory of Cognitive Intelligence (iED2022-002), and the Hifyte joint research.

References


