DIRT: Deep Learning Enhanced Item Response Theory for Cognitive Diagnosis

Song Cheng¹, Qi Liu¹,* , Enhong Chen¹, Zai Huang¹, Zhenya Huang¹, Yuying Chen¹,², Haiping Ma³, Guoping Hu³
¹Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, {chsong, huangzai, huangzhy, cyy3322}@mail.ustc.edu.cn; {qiliuql, cheneh}@ustc.edu.cn; ²Ant Financial Services Group, yuying.cyy@antfin.com ³IFLYTEK Co., Ltd., {hpma, gphu}@iflytek.com

ABSTRACT

Cognitive diagnosis is the cornerstone of modern educational techniques. One of the most classic cognitive diagnosis methods is Item Response Theory (IRT), which provides interpretable parameters for analyzing student performance. However, traditional IRT only exploits student response results and has difficulties in fully utilizing the semantics of question texts, which significantly restricts its application. To this end, in this paper, we propose a simple yet surprisingly effective framework to enhance the semantic exploiting process, which we termed Deep Item Response Theory (DIRT). In DIRT, we first use a proficiency vector to represent student proficiency on knowledge concepts and represent question texts and knowledge concepts by dense embedding. Then, we use deep learning to enhance the process of diagnosing parameters of student and question by exploiting question texts and the relationship between question texts and knowledge concepts. Finally, with the diagnosed parameters, we adopt the item response function to predict student performance. Extensive experimental results on real-world data clearly demonstrate the effectiveness and the interpretability of DIRT framework.

CCS CONCEPTS

• Information systems → Data mining; • Social and professional topics → K-12 education;

KEYWORDS

Cognitive diagnosis; Item response theory; Deep learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:


A large number of educational systems (e.g., massive open online courses) provide a series of computer-aided applications for better tutoring, such as computer adaptive test [7] and knowledge tracing [9]. Among these applications, the cognitive diagnosis that discovering the latent traits of students is becoming increasingly important. To execute cognitive diagnosis more effectively, the classic framework of Item Response Theory (IRT) [10] has been proposed, which introduces interpretable parameters with item response function to analyse students’ performance.

Though IRT has achieved great successes in cognitive diagnosis area, there is still an important issue limits its usefulness. Specifically, it only considers student responses, right (e.g., 1) or wrong (e.g., 0)—that is, it ignores the rich semantics in the other question materials. As shown in Figure 1, the question texts and the knowledge concepts on the underline with the same color are closely related, which is helpful for modelling questions [5]. It motivates us to integrate semantics to improve and enhance traditional IRT.

To this end, we propose a novel and general deep item response theory (DIRT) framework to enhance item response theory. Specifically, we first create a proficiency vector to represent the student proficiency on each knowledge concept and embed questions. Then, to diagnose the latent trait 𝜃 of students, the discrimination 𝑎 and the difficulty 𝑏 of questions [10], we introduce the deep learning methods (e.g., DNN, LSTM) for parsing semantics from question tests and the relationship between question texts and knowledge concepts. Finally, with the parameters diagnosed by deep learning methods, it can predict whether the student can answer the question correctly by item response function. Extensive experimental results present that DIRT surpasses traditional IRT by a large margin.

Figure 1: A toy example of student question records

1 INTRODUCTION

1 INTRODUCTION

A large number of educational systems (e.g., massive open online courses) provide a series of computer-aided applications for better tutoring, such as computer adaptive test [7] and knowledge tracing [9]. Among these applications, the cognitive diagnosis that discovering the latent traits of students is becoming increasingly important. To execute cognitive diagnosis more effectively, the classic framework of Item Response Theory (IRT) [10] has been proposed, which introduces interpretable parameters with item response function to analyse students’ performance.

Though IRT has achieved great successes in cognitive diagnosis area, there is still an important issue limits its usefulness. Specifically, it only considers student responses, right (e.g., 1) or wrong (e.g., 0)—that is, it ignores the rich semantics in the other question materials. As shown in Figure 1, the question texts and the knowledge concepts on the underline with the same color are closely related, which is helpful for modelling questions [5]. It motivates us to integrate semantics to improve and enhance traditional IRT.

To this end, we propose a novel and general deep item response theory (DIRT) framework to enhance item response theory. Specifically, we first create a proficiency vector to represent the student proficiency on each knowledge concept and embed questions. Then, to diagnose the latent trait 𝜃 of students, the discrimination 𝑎 and the difficulty 𝑏 of questions [10], we introduce the deep learning methods (e.g., DNN, LSTM) for parsing semantics from question tests and the relationship between question texts and knowledge concepts. Finally, with the parameters diagnosed by deep learning methods, it can predict whether the student can answer the question correctly by item response function. Extensive experimental results present that DIRT surpasses traditional IRT by a large margin.
2 PRELIMINARIES

2.1 Cognitive diagnosis Task
Suppose there are $L$ students, $M$ questions and total $P$ knowledge concepts. The history records that $L$ students do $M$ questions are represented by $R = \{R_{ij}\}_{1 \leq i \leq L, 1 \leq j \leq M}$, where $R_{ij} = (S_j, Q_j, r_{ij})$ denotes the student $S_j$ obtains score $r_{ij}$ on question $Q_j$, $Q_j = (QT_j, QK_j)$ is composed of question texts $QT_j$ and knowledge concepts $QK_j$. Given students’ responses $r_{ij}$, question texts $QT_j$ and knowledge concepts $QK_j$, our goal is to build a model $M$ to diagnose students’ proficiency on each knowledge concept. Since there is no ground truth for diagnosis results, following previous works [14], we adopt performance prediction task to validate the effectiveness of cognitive diagnosis results.

2.2 Related Models

2.2.1 Item Response Theory. IRT is one of the most important psychological and educational theories which roots in psychological measurement [10]. With the student latent trait $\theta$, question discrimination $a$ and difficulty $b$ as parameters, IRT can predict the probability that the student answers a specific question correctly with item response function. The item response function is defined as follow:
\[
P(\theta) = \frac{1}{1 + e^{-Da(\theta - b)}},
\]
where $P(\theta)$ is the correct probability, $D$ is a constant which often set as 1.7.

2.2.2 Multidimensional Item Response Theory. MIRT is extended from IRT to meet the demands of multidimensional data [13]. With student latent traits $\theta = (\theta_1, ..., \theta_m)^T$, knowledge concept discriminations $a = (a_1, ..., a_m)^T$ and intercept term $d$ of the question as parameters, MIRT can also predict the probability of the student answers a specific question correctly with multidimensional item response function. The multidimensional item response function is defined as follow:
\[
P(\theta) = e^{a^T \theta + d},
\]
where $P(\theta)$ is the probability same as IRT.

3 DIRT FRAMEWORK
To enhance item response theory for cognitive diagnosis, DIRT contains three modules, i.e., input, deep diagnosis and prediction module. Input module initializes a proficiency vector in each knowledge concept for the student, and embeds question texts and knowledge concepts to vectors. Deep diagnosis module diagnoses latent trait, discrimination and difficulty with deep learning to enhance the model. Prediction module predicts the probability that the student answers the question correctly with item response function. In the section bellow, we give a specific implementation of DIRT which is shown in Figure 2.

3.1 A Specific Implementation of DIRT

3.1.1 Input Module. Given a student $S$, we initialize a proficiency vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_P)$ with randomly, it is not belong to the training process, where $\alpha_l \in [0, 1]$ represents the degree a student masters the knowledge concept $l$.

For a question $Q$, question texts are composed of a sequence of words $Q_T = \{w_1, ..., w_u\}$, where $u$ is the length of $Q_T$, $w_1 \in \mathbb{R}^{d_0}$ is a $d_0$-dimensional Word2Vec [8] vector, as for mathematical formulas, we regard each symbols as a word. Knowledge concepts are represented by one-hot vectors $Q_K = \{K_1, ..., K_v\}$, $K_l \in \{0, 1\}^P$, where $v$ is the number of knowledge concepts. Then, we utilize a $d_1$-dimension dense layer to acquire the dense embedding for each knowledge concept $K_l$ for better training, the dense embedding of $K_l$ as $k_l$, and $k_l \in \mathbb{R}^{d_1}$:
\[
k_l = K_lW_k,
\]
where $W_k \in \mathbb{R}^{P \times d_1}$ are the parameters of the dense layer.

3.1.2 Deep Diagnosis Module. Deep diagnosis module is mainly achieved by deep learning techniques (e.g., DNN, LSTM) to diagnose latent trait, discrimination and difficulty. The details are as follows.

Latent Trait. Latent trait $\theta$ has strong interpretability for students’ performance on questions, it is closely related to the proficiency of knowledge concepts [13]. In order to learn high-order features for latent trait diagnosing, we may use some nonlinear models (e.g., DNN), here we adopt deep neural network [15]. Specifically, given the proficiency vector $\alpha = (\alpha_1, ..., \alpha_P)$ of the student $S$ and a question $q$, we multiply the corresponding proficiency in $\alpha$ with the concepts dense embedding of the questions and get a $d_1$-dimension vector $\Theta \in \mathbb{R}^{d_1}$. Then we input $\Theta$ into DNN to learn the latent trait, which is defined as follow:
\[
\Theta = \text{DNN}_d(\Theta), \quad \Theta = \alpha \circ k = \sum_{k_l \in K_q} \alpha_l k_l,
\]
where $K_q$ is the set of the knowledge concepts of question $q$.

Discrimination. Discrimination $a$ can be applied to analyse student performance distribution on the question. Inspired by the relationship between Multidimensional Item Discrimination (MDISC) and knowledge concepts [13], we learn question discrimination $a$ from knowledge concepts corresponded to the question. Also, since deep neural network can learn high-order nonlinear features automatically [15], we use another DNN to diagnose question discrimination $a$. Specifically, we sum the dense embedding of knowledge concepts in $K_q$ to get a $d_1$-dimensional vector $A \in \mathbb{R}^{d_1}$. Then, we input $A$ into the DNN to diagnosis question discrimination. We normalize the discrimination to meet the requirements that the range of $a$ should be $[-4, 4]$ [1] and the definition of $a$ is as follow:
\[
a = 8 \times \text{sigmoid} (\text{DNN}_d(A) - 0.5), \quad A = k \circ k = \sum_{k_l \in K_q} k_l.
\]
where the structure of DNN$_a$ is same as DNN$_b$, but the parameters are not shared between them.

**Difficulty.** Difficulty $b$ determines how difficult the question is. The first perspective is to diagnose difficulty by exploiting semantics of question texts [5]. Following previous works [11], LSTM can handle and represent long time sequence texts from semantic perspective which have strong robustness, we adopt LSTM to model difficulty $b$ from question text perspective. As for the second perspective, the depth and width of knowledge concepts examined by the question also have a great impact on difficulty. The deeper and wider the knowledge concepts are examined, the more difficult the question is. Obviously, the depth and width of the examined concepts can be reflected by the relevance between question texts and knowledge concepts. We adopt an attention mechanism to capture the relationship between question texts and knowledge concepts. Totally, we design an attention-based LSTM to integrate question texts and knowledge concepts for diagnosing question difficulty $b$. Specifically, the sequence input to this LSTM is $x = (x_1, x_2, ..., x_T)$, where $N$ is the max step of the attention-based LSTM. The $t$-th input step of attention-based LSTM is defined as follow:

$$x_t = \sum_{k_i \in \mathcal{K}_q} \text{softmax} \left( \frac{\xi_j}{\sqrt{d_0}} \right) k_i + w_t, \quad \xi_j = w^T_i k_i, \quad (6)$$

where $\sqrt{d_0}$ is the scaling factor. $\xi_j$ is the relevance between word $w_t$ and the knowledge concepts in $\mathcal{K}_q$. After that, an average-pooling operation is utilized to obtain parameter $b$. Also, we normalize the difficulty to meet the requirements that the range of $b$ should be $[-4, 4]$ [1] and the definition of $b$ is as follow:

$$b = 8 \times \text{sigmoid}(\text{averagePooling}(h_N)) - 0.5, \quad (7)$$

where $\text{averagePooling}$ is an operation that calculates the mean of all elements in the last step vector $h_N$ of LSTM.

3.1.3 **Prediction Module.** The prediction module is used to preserve the ability of performance prediction and the interpretation power of student latent trait, question discrimination and difficulty in traditional item response theory. We input parameters diagnosed by deep diagnosis module into the item response function Eq.(1) [10] to predict the student performance on the specific question.

3.1.4 **DIRT Learning.** The whole parameters to be updated in DIRT mainly exist in two parts: input module and deep diagnosis module. In input module, the parameters need to be updated contain proficiency vector $\alpha$, question embedding weights and knowledge concept dense embedding weights $\{W_Q, W_K\}$. In the deep diagnosis module, the parameters need to be updated contain the weights of three neural networks $\{W_{DNN_a}, W_{DNN_b}, W_{LSTM}\}$ which are used to learn the latent trait, discrimination and difficulty respectively. The objective function of DIRT is the negative log likelihood function. Formally, for student $i$ and question $j$, let $r_{ij}$ be the actual score, $\hat{r}_{ij}$ be the score predicted by DIRT. Thus the loss for student $i$ on question $j$ is defined as:

$$L = r_{ij} \log \hat{r}_{ij} + (1 - r_{ij}) \log (1 - \hat{r}_{ij}), \quad (8)$$

in this way, we can learn DIRT by directly minimizing the objective function using Adam optimization [6].

**EXPERIMENTS.**

4.1 **Dataset Description**

Since DIRT needs to exploit question texts, only one private dataset can be used, which is composed of the mathematical data supplied by iFLYTEK Co., Ltd collected from Zhixue. We filter out the students with less than 15 records and the questions that have not been answered by students. After pruning, the distribution of knowledge concepts number and question texts length are shown in Figure 3. Also, some statistics of the dataset are shown in Table 1. We can observe that each student has done about 62.09 questions, and each question requires about 1.49 knowledge concepts.

4.2 **Baselines and Evaluation Metrics.**

We compare the performance of DIRT with several methods: IRT [10] and DINA [3] are continuous and discrete cognitive diagnosis methods respectively, MIRT [13] is a multidimensional cognitive diagnosis method extend from IRT, PMF [4] and (NMF) [12] are matrix factorization methods, DIRTNA is a variant of DIRT without attention mechanism.

We evaluate the performance of DIRT from two perspectives, regression perspective [2]: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and classification perspective [14]: Area Under Curve (AUC) and Prediction Accuracy (ACC).

4.3 **Experimental Results**

4.3.1 **Performance Prediction Task.** Here, we conduct extensive experiments on performance prediction task at different data sparsity by splitting dataset into training and testing dataset with different ratio: 60%, 70%, 80%, 90%. The results on all metrics are shown in Figure 4. We can observe that compares with all the baselines, especially IRT, MIRT. DIRT performs the best, it illustrates that DIRT can make full use of question texts, benefiting the prediction. Comparing with DIRTNA, DIRT performs better, it proves that attention mechanism is effective for exploiting the relationship between question texts and knowledge concepts and helpful for prediction. We can also observe that DIRT and IRT perform better than MIRT, which is mainly because MIRT is sensitive to the concept on which student has high proficiency. Therefore, DIRT

![Figure 3: Distribution of words and knowledge concepts.](https://www.zhihu.com)

**Table 1: The statistics of the dataset.**

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Original</th>
<th>Pruned</th>
</tr>
</thead>
<tbody>
<tr>
<td># of history records</td>
<td>65,368,739</td>
<td>5,068,039</td>
</tr>
<tr>
<td># of students</td>
<td>1,016,235</td>
<td>81,624</td>
</tr>
<tr>
<td># of questions</td>
<td>1,735,635</td>
<td>13,635</td>
</tr>
<tr>
<td># of knowledge concepts</td>
<td>1,412</td>
<td>621</td>
</tr>
<tr>
<td>Avg. questions per student</td>
<td>62.09</td>
<td>1.49</td>
</tr>
<tr>
<td>Avg. concepts per question</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

1http://www.zhixue.com
framework is more reliable than MIRT to the concept on which students have a high proficiency.

4.3.2 Case Study. Here, we give an example of cognitive diagnosis of student knowledge proficiency. As shown in Figure 5, the radar chart shows a student’s concepts proficiency diagnosed by IRT, MIRT and DIRT. Since IRT only diagnoses student latent trait which has the same value on all questions, so the diagnosis result of IRT is a regular polygon in Figure 5. Thus, DIRT can provide more accurate diagnosis results on knowledge concepts than IRT. We can also observe that DIRT predicts all three concepts correctly, but IRT gets a wrong result on the second question, because IRT obtains a wrong value -0.171 of difficulty b compares with DIRT and MIRT. Also, MIRT gets a wrong result on the third question, which is because MIRT is sensitive to concepts on which student has high proficiency [13] such as K7. Totally, DIRT can enhance traditional IRT with deep learning for cognitive diagnosis by exploiting question texts.

5 CONCLUSIONS

In this paper, we proposed a general DIRT framework to enhance traditional IRT to exploit the rich semantics in the question texts, as well as the relationship between question texts and knowledge concepts for cognitive diagnosis. Extensive experiments on a large scale real-world dataset clearly validated the effectiveness and the interpretation power of DIRT.

ACKNOWLEDGMENT

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), the National Natural Science Foundation of China (Grants No. 61672483, U1605251), and the Science Foundation of Ministry of Education of China & China Mobile (No. MCM20170507). Qi Liu gratefully acknowledges the support of the Young Elite Scientist Sponsorship Program of CAST and the Youth Innovation Promotion Association of CAS (No. 2014299).

REFERENCES

[14] Run ze Wu, Guandong Xu, Enhong Chen, Qi Feng Liu, and Wan Ng. 2017. Knowledge or Gaming?: Cognitive Modelling Based on Multiple-Attempt Response. In WWW.