BETA-CD: A Bayesian Meta-learned Cognitive Diagnosis Framework for Personalized Learning

Haoyang Bi,[1, 2] **Enhong Chen**[*]**,**[1, 2] **Weidong He,** [1, 2] **Han Wu,** [1, 2]
Weihao Zhao, [1, 2] **Shijin Wang,** [2,3] **Jinze Wu** [3]

[1] Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China
[2] State Key Laboratory of Cognitive Intelligence
[3] iFLYTEK AI Research, iFLYTEK CO., LTD.
bhy0521@mail.ustc.edu.cn, cheneh@ustc.edu.cn, {hwd, wuhanhan, zhaoweihao}@mail.ustc.edu.cn,
{sjwang3, jzwu4}@iflytek.com

Abstract

Personalized learning is a promising educational approach that aims to provide high-quality personalized services for each student with minimum demands for practice data. The key to achieving that lies in the cognitive diagnosis task, which estimates the cognitive state of the student through his/her logged data of doing practice quizzes. Nevertheless, in the personalized learning scenario, existing cognitive diagnosis models suffer from the inability to (1) quickly adapt to new students using a small amount of data, and (2) measure the reliability of the diagnosis result to avoid improper services that mismatch the student's actual state. In this paper, we propose a general Bayesian mETA-learned Cognitive Diagnosis framework (BETA-CD), which addresses the two challenges by prior knowledge exploitation and model uncertainty quantification, respectively. Specifically, we firstly introduce Bayesian hierarchical modeling to associate each student's cognitive state with a shared prior distribution encoding prior knowledge and a personal posterior distribution indicating model uncertainty. Furthermore, we formulate a meta-learning objective to automatically exploit prior knowledge from historical students, and efficiently solve it with a gradient-based variational inference method. Extensive experiments upon various real-world datasets and models show the effectiveness and generality of the proposed BETA-CD.

Introduction

In intelligent tutoring systems, personalized learning is a promising educational approach which aims to deliver customized services based on the students' personal states to address their unique needs in study (Paramythis and Loidl-Reisinger 2003). As shown in Figure 1, each student has a few personal practice logs used to estimate his/her cognitive state (e.g., how well he/she has mastered the specific knowledge concepts), which is referred to as the procedure of *cognitive diagnosis* (Wang et al. 2020). Then the system is able to provide personalized services based on the diagnosis result, such as online course recommendation (Zhang et al. 2019) and learning path planning (Liu et al. 2019). Compared to conventional scenarios where students are provided with equally non-personalized practice questions and

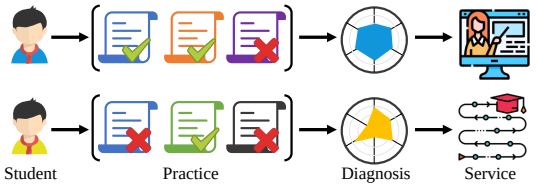

Figure 1: Illustrative procedure of personalized learning in intelligent tutoring systems.

educational services, personalized learning can essentially achieve *lower practice burden* and *higher service quality*. To this objective, the procedure of cognitive diagnosis is required to (1) rely on as few practice data as possible from students and (2) provide as an informative result as possible for downstream services. In the literature, massive efforts have been devoted to developing cognitive diagnosis models (CDMs) (Embretson and Reise 2013; Wang et al. 2020). However, existing works are faced with the following two challenges in the personalized learning scenario.

Firstly, it is challenging for existing CDMs to quickly adapt to new students within only a small amount of practice data. Typically, directly applying ordinary optimization methods to the current student's practice logs will bring severe overfitting problems. A major cause of such inability for fast adaptation is the lack of the exploitation of prior knowledge about the population of similar students. We take the human teacher as an explanatory metaphor. In practice, a proficient teacher can quickly judge a student's state through a few questions. A critical factor lies in the teacher's rich prior knowledge from her past teaching experience. For instance, assume that a student has only one correct response on a question at difficulty level 4 (out of 10). Taking no prior knowledge into consideration, since the student has a 100% correct rate, we might overestimate the student's ability to extremely high, say level 9. Contrastively, the teacher can refer to all the students in the same class who also correctly answered this question, and then give a more reasonable result, e.g., level 6, which is the average ability level of the similar students. Hence, it is necessary to exploit prior knowledge in cognitive diagnosis procedures for facilitating fast adap-

[*]Corresponding author.

tation with limited data.

Secondly, there lacks a principled way of measuring the reliability of the diagnosis results. Since downstream applications heavily depend on the results of cognitive diagnosis, being agnostic about an unreliable result may cause a serious mismatch between the provided services and the student's actual state (e.g., recommending an overwhelmingly hard online course), degrading the service quality. However, most of the existing CDMs solely provide the final value of the cognitive state estimation while conveying little information about how much we can trust it. To overcome this weakness, we propose to measure the reliability of the diagnosis results by quantifying the model uncertainty of the CDM in the procedure of cognitive diagnosis.

Summarily, in the context of personalized learning, cognitive diagnosis models should be ideally equipped with both prior knowledge exploitation and model uncertainty quantification. To this end, we propose a novel **B**ayesian m**ETA**-learned **C**ognitive **D**iagnosis (BETA-CD) framework that addresses both challenges in a unified manner. We first introduce Bayesian hierarchical modeling for the cognitive diagnosis task. Specifically, we incorporate prior knowledge as a globally parametrized prior distribution of the cognitive states for all students. Correspondingly, the locally inferred student-specific posterior distribution is naturally recognized as the cognitive diagnosis result which quantifies the model uncertainty. Furthermore, to effectively exploit prior knowledge, we formulate a meta-learning objective to automatically optimize an appropriate prior on historical data so that the resulting posterior has a good and quick adaptation to new students. To tackle the intractability of posteriors in optimization, we leverage a gradient-based variational inference method for scalable meta-optimization. Finally, a practical specification for both prior and approximate posterior is realized with a balanced consideration of reasonability, performance and computational efficiency. As a general and scalable framework, BETA-CD can be applied to a wide range of CDMs, especially the recent deep models (Wang et al. 2020; Gao et al. 2021).

The main contributions of our work are listed as follows. (1) We propose a general Bayesian meta-learned cognitive diagnosis framework (BETA-CD) which achieves both lower practice burden and higher service quality for personalized learning. (2) We introduce Bayesian hierarchical modeling for the cognitive diagnosis task to unifiedly incorporate prior knowledge and model uncertainty. (3) We formulate a meta-learning objective to automatically exploit prior knowledge from historical data and solve it with a scalable gradient-based variational inference method. (4) Extensive experiments on various datasets and models validate the effectiveness and generality of BETA-CD.

Related Works

Cognitive Diagnosis. As a fundamental task, cognitive diagnosis has been well studied for decades in the area of educational psychology. In general, a cognitive diagnosis model (CDM) links the parametrized trait features of the student (θ) and the question (ϕ) to the prediction of the student's response result (correct or wrong) (Wang et al. 2020). The most popular CDM is Item Response Theory (IRT) (Embretson and Reise 2013), in which $\theta, \phi \in \mathbb{R}$ are unidimensional continuous latent parameters indicating student ability level and question difficulty level, respectively, and the predictive response is modeled in a logistic way, $p(r_{ij} = 1|\theta_i, \phi_j) = \text{sigmoid}(\theta_i - \phi_j)$. Along this line, Multidimensional IRT (MIRT) (Liu et al. 2018c) incorporates higher dimensional parameters to extend the ability trait in IRT. Some other CDMs directly model the mastery level of the student on specific knowledge concepts (e.g., Trigonometric Function), such as Deterministic Inputs, Noisy-And gate (DINA) (De La Torre 2009). Recently, data-driven CDMs have been proposed, which leverage deep neural networks to facilitate automatic modeling for the latent traits as well as their complex relations (Wang et al. 2020; Gao et al. 2021).

Uncertainty Quantification. When developing mathematical models for various real-world tasks such as data analysis (Liu et al. 2018a; Zhao et al. 2022), image recognition (Hu et al. 2021) and user modeling (He et al. 2020), it is often needed to deal with the model uncertainty (Sullivan 2015; Abdar et al. 2021). In educational scenarios, being unaware of the uncertainty in cognitive diagnosis models will cause irrational behaviors of intelligent tutoring systems and thus leave terrible study experiences for the students. A natural solution of uncertainty quantification is Bayesian modeling (van de Schoot et al. 2021; Pei et al. 2020). Instead of training deterministic parameters, a Bayesian model relates its parameters with prior distributions and acquires their posterior by the Bayesian rule. There are some efforts to incorporating Bayesian modeling for specific classes of CDMs (Yamaguchi and Okada 2020; Ma and Jiang 2021; Zhan et al. 2019). However, there still lacks a principled way of dealing with uncertainty unifiedly for a wide range of CDMs, especially scaling to those deep models in which the computation of posteriors becomes intractable (Wang et al. 2020; Gao et al. 2021). To tackle this problem, we leverage a gradient-based variational inference method with a learnable prior to encode richer knowledge.

Meta-learning. Meta-learning aims at empowering the machine learning model with meta-knowledge extracted from a series of learning tasks so that the model can adapt fast to new tasks drawn from the same task distribution (Hospedales et al. 2021). With effective exploitation of prior knowledge, meta-learning has been widely applied to solve the data sparsity problem in various fields (Lee et al. 2019; Rakelly et al. 2019). The most relevant works to our work belong to so-called optimization-based meta-learning (Finn, Abbeel, and Levine 2017; Li et al. 2017; Rajeswaran et al. 2019). Furthermore, it has been shown that optimization-based meta-learning can be endowed with probabilistic interpretations by Bayesian hierarchical modeling (Grant et al. 2018), such that we can reason about uncertainty while adapting to new tasks (Finn, Xu, and Levine 2018; Ravi and Beatson 2018; Nguyen, Do, and Carneiro 2020). In the proposed BETA-CD, we formulate a meta-learning objective to facilitate automatic prior knowledge exploitation by treating the cognitive diagnosis procedure for each student as individual tasks.

Methodology

Problem Setup

Suppose there are M students, $S = \{s_i\}_{i=1}^{M}$, and N questions, $Q = \{q_j\}_{j=1}^{N}$, in an intelligent tutoring system. Each student practises on the system by responding to a few provided questions, and the results are recorded as triplets like (s_i, q_j, r_{ij}), where $r_{ij} \in \{1, 0\}$ denotes that student s_i responses question q_j correctly or wrongly. The set of questions that s_i has answered and the corresponding responses are denoted as Q_i and R_i, respectively. Through the practice data, we can discover the student's cognitive state with a cognitive diagnosis model (CDM). Principally, a CDM consists of two sets of parameters: the cognitive state parameters θ that are personalized for each student, and the question feature parameters $\Phi = \{\phi_j\}_{j=1}^{N}$ that are shared across students. In practice, the question features Φ are typically pretrained on historical data or directly calibrated by experts and then fixed for new students. For brevity of presentation, we omit Φ with more focus on the cognitive state parameters θ. Given the cognitive state θ_i of student s_i, the CDM predicts his/her response to any question $q_j \in Q$ as $p(r_{ij} = 1|q_j, \theta_i)$.

Problem Definition *Suppose an intelligent tutoring system with a cognitive diagnosis model parametrized by θ. Given the historical students $S = \{s_i\}_{i=1}^{M}$ with recorded practice data $\{(Q_i, R_i)\}_{i=1}^{M}$, for any new student $s_* \notin S$, our goal is to obtain a personalized cognitive state estimation θ_* via a small amount of new practice data (Q_*, R_*).*

Bayesian Meta-learned Cognitive Diagnosis

To achieve a lower practice burden and higher service quality for personalized learning, we introduce a Bayesian mETA-learned Cognitive Diagnosis framework, namely BETA-CD, including: (1) Bayesian hierarchical modeling that unifiedly incorporates prior knowledge and model uncertainty; (2) Meta-learning objective formulation to facilitate automatic prior knowledge exploitation; (3) Gradient-based variational inference for scalable meta-optimization.

Bayesian Hierarchical Modeling To solve the problem defined above, we should seek a proper objective for cognitive state parameter optimization. For a new student s_*, a conventional way of estimating his/her cognitive state is directly fitting the CDM to the practice data $\{Q_*, R_*\}$, i.e.,

$$\theta_* = \arg\min_{\theta} -\log p(R_*|Q_*, \theta), \tag{1}$$

where $p(R|Q, \theta)$ is a short form for $\prod_{q\in Q}^{r'\in R} p(r = r'|q, \theta)$ in this section. Although the objective is simple to optimize, it results in two challenges which are especially severe in the context of personalized learning. First, since personalized learning pursues a low practice burden for students, the data size, $|Q_*| = |R_*|$, is assumed to be small. As a result, the estimation obtained via Eq. (1) is prone to overfit. Second, because the point output θ_* is unable to quantify the model uncertainty, downstream applications have little information about the reliability of the cogitive diagnosis result, which may lead to seriously low-quality educational services for students.

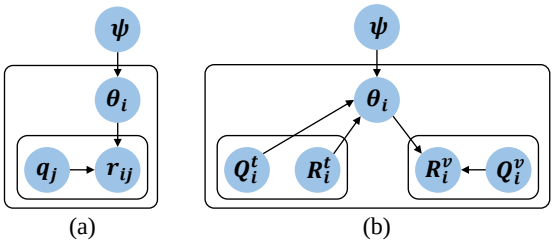

Figure 2: (a) Graphic model for Bayesian hierarchical modeling of cognitive diagnosis. (b) Graphic model after performing inference on θ_i.

To this end, we introduce Bayesian hierarchical modeling (Fei-Fei and Perona 2005) to view the above two challenges from a probabilistic perspective. As shown in Figure 2(a), we assume a global latent variable ψ and student-specific variables $\theta_i (i = 1, ..., M)$. ψ influences each θ_i by specifying a parametrized prior distribution $p(\theta_i|\psi)$, which can be treated as an encoding of prior knowledge about the overall students. For example, if a Gaussian prior is assumed, ψ consists of the mean and standard deviation of θ. The student-specific cognitive state θ_i determines the prediction of the student's response to any question q_j as $p(r_{ij} = 1|q_j, \theta_i)$, which has been supposed to be specified by the CDM. For a new student s_* with the observed practice data (Q_*, R_*), we can infer the posterior distribution of his/her cognitive state by the Bayesian rule:

$$\theta_* \sim p(\theta|Q_*, R_*, \psi). \tag{2}$$

Compared to Eq. (1), the above cognitive diagnosis result has two key differences. First, the overfitting problem caused by the limited observed data can be alleviated by leveraging external prior knowledge contained in the prior $p(\theta|\psi)$. Second, in contrast to a point result, the distributional one contains rich information about the model uncertainty.

Meta-learned Prior Knowledge In the literature, the prior $p(\theta|\psi)$ is determined either empirically or just for mathematical convenience (Lord 2012). However, such a crude prior typically acts as regularization and has limited effect in preventing overfitting. In contrast, we propose to automatically exploit prior knowledge from the practice data by optimizing the prior parametrization ψ with a well-formulated meta-learning objective (Hospedales et al. 2021). Specifically, for each historical student $s_i \in S$, we treat his/her cognitive diagnosis procedure as an individual task $\mathcal{T}_i = (Q_i^t, R_i^t, Q_i^v, R_i^v)$, where (Q_i^t, Q_i^v) denotes a train-validation split of Q_i, and so does (R_i^t, R_i^v). It is believed that these tasks have internally similar structures so that there exists shared prior knowledge that can be learned. To be specific, such prior knowledge is supposed to help the cognitive diagnosis model quickly adapt to each student, i.e., infer a student-specific posterior $p(\theta_i|Q_i^t, R_i^t, \psi)$ that fits well on the validation set (Q_i^v, R_i^v), as illustrated in Figure 2(b). Hence, we aim to obtain an appropriate prior parametrization ψ by learning to learn on each task with

minimizing the following meta-loss:

$$\begin{aligned}\min_{\psi} \sum_{i=1}^{M} \mathcal{L}_i^{(m)}(\psi) &\equiv \sum_{i=1}^{M} -\log p(R_i^v|Q_i^v, Q_i^t, R_i^t, \psi) \\ &= \sum_{i=1}^{M} -\log \mathbb{E}_{\theta_i \sim p(\theta_i|Q_i^t, R_i^t, \psi)} \left[p(R_i^v|Q_i^v, \theta_i)\right].\end{aligned} \tag{3}$$

Note that we assume ψ to be deterministic rather than treating it as a random variable. There are two reasons for this assumption. First, in typical intelligent tutoring systems, the number of historical students is much larger than the number of practice logs within each student. In this case, the uncertainty of the meta-learned prior knowledge in ψ should be low, i.e., $p(\psi)$ can be approximated with a Dirac delta distribution. Second, in personalized learning, it is the uncertainty in the student's cognitive state θ, rather than in the prior knowledge ψ, that is of our interest.

Gradient-based Variational Inference With the meta-learning objective (Eq. (3)) in mind, we now discuss how to train the meta-parameters ψ effectively and efficiently. The main difficulty lies in the intracbility of the posterior term $p(\theta_i|Q_i^t, R_i^t, \psi)$, especially when the cognitive state parameter θ is high-dimensional. To tackle this problem, we use a variational distribution $q(\theta_i; \lambda_i)$ to approximate the posterior, where $\lambda_i = \lambda_i(Q_i^t, R_i^t, \psi)$ denotes the parameters assumed in the same form as ψ for simplicity. In variational inference, we obtain $q(\theta_i; \lambda_i)$ by minimizing its Kullback-Leibler (KL) divergence from the target distribution:

$$\begin{aligned}\lambda_i &= \arg\min_{\lambda} \mathrm{KL}\left[q(\theta_i;\lambda) \| p(\theta_i|Q_i^t, R_i^t, \psi)\right] \\ &= \arg\min_{\lambda} \int q(\theta_i;\lambda) \log \frac{q(\theta_i;\lambda) p(R_i^t|Q_i^t, \psi)}{p(R_i^t|Q_i^t, \theta_i) p(\theta_i|\psi)} d\theta_i \\ &= \arg\min_{\lambda} \mathbb{E}_{\theta_i \sim q(\theta_i;\lambda)} \left[-\log p(R_i^t|Q_i^t, \theta_i)\right] \\ &\quad + \mathrm{KL}\left[q(\theta_i;\lambda) \| p(\theta_i|\psi)\right] + \underbrace{\log p(R_i^t|Q_i^t, \psi)}_{\text{constant w.r.t. } \lambda}.\end{aligned} \tag{4}$$

Intuitively, the first term maximizes the expected log-likelihood, while the second is a regularisation term that penalizes the difference between the approximate posterior and the prior. The last term is constant w.r.t. λ and thus ignored during optimization. Accordingly, we define a local loss used to obtain the approximate student-specific posterior $q(\theta_i; \lambda_i)$ for each student s_i:

$$\begin{aligned}\min_{\lambda_i} \mathcal{L}_i^{(l)}(\lambda_i) \equiv \mathbb{E}_{\theta_i \sim q(\theta_i;\lambda)} \left[-\log p(R_i^t|Q_i^t, \theta_i)\right] \\ + \eta\ \mathrm{KL}\left[q(\theta_i;\lambda_i) \| p(\theta_i;\psi)\right],\end{aligned} \tag{5}$$

where η is an empirical KL weighting parameter that we find useful for training stability. We optimize λ_i by minimizing the local loss as follows:

$$\lambda_i \leftarrow \psi - \mathrm{SGD}_{\lambda_i}^{K}(\mathcal{L}_i^{(l)}(\lambda_i); \alpha), \tag{6}$$

where $\mathrm{SGD}_{\lambda_i}^{K}(\mathcal{L}_i^{(l)}(\lambda_i); \alpha)$ denotes K-step stochastic gradient descent operations with per-step learning rate α in which the expectation terms are computed by Monte Carlo sampling. This gradient-based variational inference method has two advantages. First, it is much more computationally efficient than the original minimization problem (Eq. (4)). Second, by relating the prior and posterior with gradient-based operations, we can effectively use the approximate posteriors to compute the original meta-loss (Eq. (3)) and optimize the meta-parameters ψ via gradient descent as well. The whole implementation will be thoroughly presented next.

Algorithm 1: BETA-CD Meta-training

Input: Historical students $S = \{s_i\}_{i=1}^{M}$ with practice data $\{\mathcal{T}_i = (Q_i, R_i)\}_{i=1}^{M}$
Parameter: KL weighting parameter η; Mini-batch size T; Sampling sizes N_t, N_v; Number of local updates K; Local update rate α; Meta update rate γ
Output: Meta-parameters ψ

Initialize ψ randomly
while ψ not converged **do**
  Sample a mini-batch tasks $\mathcal{T}_i$, $i = 1 : T$
  for each task $\mathcal{T}_i$ **do**
    Train-validation split $\mathcal{T}_i = (Q_i^t, R_i^t, Q_i^v, R_i^v)$
    Initialize $\lambda_i \leftarrow \psi$
    for step $k = 1 : K$ **do**
      Sample $\hat{\theta}_i^{n_t} \sim q(\theta_i; \lambda_i), n_t = 1 : N_t$
      Compute local loss by sampling:
      $\mathcal{L}_i^{(l)}(\lambda) \approx \frac{1}{N_t} \sum_{n_t=1}^{N_t} -\log p(R_i^t|Q_i^t, \hat{\theta}_i^{n_t}) + \eta\ \mathrm{KL}[q(\theta_i;\lambda_i) \| p(\theta_i|\psi)]$
      Local Update: $\lambda_i \leftarrow \lambda_i - \alpha \nabla_{\lambda_i} \mathcal{L}_i^{(l)}(\lambda_i)$
    end for
    Sample $\hat{\theta}_i^{n_v} \sim q(\theta_i; \lambda_i), n_v = 1 : N_v$.
    Compute meta-loss by sampling:
    $\mathcal{L}_i^{(m)}(\psi) \approx -\log\left(\frac{1}{N_v} \sum_{n_v=1}^{N_v} p(R_i^v|Q_i^v, \hat{\theta}_i^{n_v})\right)$
  end for
  Meta Update: $\psi \leftarrow \psi - \gamma \cdot \frac{1}{T} \sum_{i=1}^{T} \nabla_{\psi} \mathcal{L}_i^{(m)}(\psi)$
end while
return ψ

Implementation Detail

The meta-training procedure of BETA-CD is summarized in Algorithm 1, following a bilevel paradigm that is common for optimization-based meta-learning (Finn, Abbeel, and Levine 2017). Specifically, in the inner loops, for each task $\mathcal{T}_i$ (i.e., the practice data of student s_i), we perform local updates w.r.t. λ_i by minimizing the local loss (Eq. (5)) on the training set to obtain the approximate posterior $q(\theta_i; \lambda_i)$. Then in the outer loops, we perform meta-updates w.r.t. ψ by minimizing the meta-loss (Eq. (3)) on the validation sets of a mini-batch of tasks.

By meta-training on massive historical students, the obtained ψ extracts rich prior knowledge about the overall cognitive states of the student population. Hence, when a new student arrives, we can infer his/her cognitive state via a small amount of practice data in aid of the prior knowledge. As illustrated in Algorithm 2, the meta-testing proce-

Algorithm 2: BETA-CD Meta-testing

Input: A new student s_* with practice data (Q_*, R_*)
Parameter: Trained meta-parameters ψ; Number of local updates K; Local update rate α; KL weighting parameter η; Sampling size N_t;
Output: Approximate posterior $q(\theta_*; \lambda_*)$

Initialize $\lambda_* \leftarrow \psi$
for step $k = 1 : K$ **do**
  Sample $\hat{\theta}_*^{n_t} \sim q(\theta_*; \lambda_*), n_t = 1 : N_t$
  Compute local loss by sampling:
  $\mathcal{L}_*^t(\lambda) \approx \frac{1}{N_t} \sum_{n_t=1}^{N_t} [-\log p(R_*|Q_*, \hat{\theta}_*^{n_t}) + \eta \, \mathrm{KL}[q(\theta_*; \lambda_*) \| p(\theta_*|\psi))]$
  Local Update: $\lambda_* \leftarrow \lambda_* - \alpha \nabla_{\lambda_*} \mathcal{L}_*^t(\lambda_*, \hat{\theta}_*^{n_t})$
end for
return $q(\theta_*; \lambda_*)$

dure on (Q_*, R_*) matches the local update on the training set (Q_i^t, R_i^t) of each task in meta-training, thus is supposed to adapt better on unseen validation data. Finally, the approximate posterior $q(\theta_*; \lambda_*)$ is output as the cognitive diagnosis result, in which we can easily observe the estimation uncertainty. In case a point result is needed (e.g., in some of our experiments), the cognitive state estimation or predictive response on some question q_j can also be obtained as $\mathbb{E}_q[\theta_*]$ and $\mathbb{E}_q\left[p(r_{*j}|q_j, \theta_*)\right]$, respectively.

In general, it is flexible to specify the parametrization forms for the prior $p(\theta_i|\psi)$ and the approximate posterior $q(\theta_i; \lambda_i)$, often aiming at a proper trade-off between performance and complexity. We consider both of them as fully factorized Gaussian distributions, i.e.,

$$p(\theta_i|\psi) \equiv \mathcal{N}(\theta_i|\boldsymbol{\mu_\theta}, \boldsymbol{\sigma_\theta} \boldsymbol{I}), \tag{7}$$

$$q(\theta_i; \lambda_i) \equiv \mathcal{N}(\theta_i|\boldsymbol{\mu_{\lambda_i}}, \boldsymbol{\sigma_{\lambda_i}} \boldsymbol{I}). \tag{8}$$

In other words, we define the parameters $\psi = \{\boldsymbol{\mu_\theta}, \boldsymbol{\sigma_\theta}\}$ and $\lambda_i = \{\boldsymbol{\mu_{\lambda_i}}, \boldsymbol{\sigma_{\lambda_i}}\}$ to be the mean and diagonal standard deviation of the Gaussian cognitive state. Besides convenience for implementation, there are two reasons to make this assumption. First, in educational psychology, the Gaussian distribution has long been recognized as a proper statistic model for the cognitive states of students (Liu et al. 2018b). Second, in most of the existing CDMs, different dimensions of the cognitive state parameters θ are modeled as independent factors, such as the mastery level on individual knowledge concepts, leading to a fully factorized prior as well.

We leverage some additional tricks in meta-optimization. For numerical stability, we implement the standard deviations in the logarithm form $\boldsymbol{\sigma} = \exp(\boldsymbol{\rho})$. The KL divergence $\mathrm{KL}\left[q(\theta_i; \lambda_i) \| p(\theta_i|\psi)\right]$ is computed efficiently in the closed form (Blundell et al. 2015) since both distributions are modeled as Gaussian. When sampling from $p(\theta_i|\psi)$ and $q(\theta_i; \lambda_i)$, we implement the re-parametrization trick (Kingma and Welling 2013). Finally, instead of using a constant learning rate α for local updates, we meta-learn an individual rate for each inner step along with the meta-parameters (Antoniou, Edwards, and Storkey 2018).

Dataset	#Students	#Questions	#Logs
ECPE	2,922	28	81,816
ASSIST	1,670	1,960	355,376
EXAM	3,750	1,179	158,178

Table 1: Statistics of the preprocessed datasets

Experiments

In this section, we first introduce the datasets and our experimental setups. Then, we conduct extensive experiments to compare the performances of CDMs optimized by the ordinary optimization approach and the proposed BETA-CD (hereinafter referred to as ORD-CDMs and BETA-CDMs, respectively) to answer the following questions:

- **RQ1:** Can BETA-CDMs gain greater accuracy in cognitive state estimation and thus perform better in predicting student performance?
- **RQ2:** How do the different designed parts in BETA-CD influence the performance of BETA-CDMs?
- **RQ3:** Are BETA-CDMs well calibrated as expected by incorporating model uncertainty quantification?
- **RQ4:** In what ways can model uncertainty quantification benefit personalized learning in practice?

Dataset Description

We evaluate our framework with three real-world datasets consisting of massive students' practice logs, i.e., ECPE, ASSIST and EXAM. ECPE (*Examination for the Certificate of Proficiency in English*), collected from a standard English test by the English Language Institute of the University of Michigan, is well adopted in educational psychology. ASSIST (*ASSISTments 2017 skill builder*) is a widely used educational dataset containing students' practice in mathematics with the ASSISTments system. Supplied by a famous online intelligent tutoring system from iFLYTEK Co., Ltd., EXAM contains mathematical test logs of high school examinations.

For each dataset, we filter out the questions answered by less than 10 students and the students that answered less than 20 questions. The statistics of the preprocessed datasets are presented in Table 1. We randomly divide the students in each dataset by 6:2:2, where the 60% partition contains the historical students in the intelligent tutoring system used to train the CDM, one 20% partition acts as new students to be diagnosed, and the other 20% partition is used for early stopping and hyperparameter tuning. For each new student, 20% of his/her logs are left out as a validation question set (i.e., Q^v, R^v). Among the rest 80%, we randomly select different numbers of logs as the training question sets (i.e., Q^t, R^t).

Experimental Setup

To evaluate the effectiveness and generality of BETA-CD, we apply our framework to four well-adopted CDMs, i.e., IRT (Embretson and Reise 2013), MIRT (Liu et al. 2018c), DINA (De La Torre 2009) and NCD (Wang et al. 2020). We call them BETA-CDMs with our framework applied

Dataset	Size	IRT		MIRT		DINA		NCD	
		ORD-	BETA-	ORD-	BETA-	ORD-	BETA-	ORD-	BETA-
ECPE	3	69.12	**72.86**	71.66	**72.84**	69.51	**71.77**	70.37	**72.54**
	5	70.48	**73.15**	72.17	**73.33**	70.58	**72.09**	70.93	**72.77**
	10	73.13	**73.80**	73.07	**73.91**	71.79	**72.90**	71.73	**73.42**
ASSIST	3	60.81	**67.46**	65.37	**67.24**	52.45	**63.67**	61.77	**64.93**
	5	63.51	**68.14**	65.71	**68.19**	53.13	**63.85**	61.86	**65.30**
	10	65.99	**69.28**	66.51	**68.97**	54.15	**64.14**	62.25	**65.84**
EXAM	3	70.06	**75.25**	75.01	**75.58**	63.07	**70.49**	69.42	**75.07**
	5	72.46	**75.62**	75.38	**75.65**	64.34	**71.15**	70.18	**75.16**
	10	74.63	**75.97**	76.13	**76.23**	65.81	**71.51**	71.03	**75.17**

(a) Results with ACC metric.

Dataset	Size	IRT		MIRT		DINA		NCD	
		ORD-	BETA-	ORD-	BETA-	ORD-	BETA-	ORD-	BETA-
ECPE	3	66.13	**71.40**	68.25	**71.55**	64.97	**67.99**	67.44	**70.92**
	5	68.11	**72.30**	69.19	**72.51**	66.33	**68.80**	68.60	**71.44**
	10	71.92	**74.10**	71.12	**74.17**	68.78	**70.67**	70.49	**73.29**
ASSIST	3	64.87	**72.52**	70.20	**72.36**	58.97	**68.43**	65.06	**69.44**
	5	67.12	**73.74**	70.75	**73.52**	59.77	**68.69**	65.24	**69.85**
	10	70.61	**74.93**	71.90	**74.70**	60.93	**69.00**	65.74	**70.50**
EXAM	3	69.40	**80.59**	79.78	**80.72**	65.82	**72.67**	74.02	**79.63**
	5	73.91	**80.87**	80.32	**80.87**	67.41	**72.94**	74.58	**79.66**
	10	78.41	**81.66**	81.24	**81.80**	68.71	**73.69**	75.06	**79.75**

(b) Results with AUC metric.

Table 2: Student performance prediction of BETA-CD compared with ordinary methods.

and ORD-CDMs otherwise, e.g., BETA-IRT and ORD-IRT. In ORD-CDMs, we implement ordinary gradient descent in cognitive state estimation for new students. In BETA-CDMs, the hyperparameters for meta-training and meta-testing are as follows. We set the mini-batch size $T = 8$, the sampling sizes $N_t = N_v = 4$ and the number of inner updates $K = 3$. The KL weighting parameter η in the local loss is set to 10^{-4}. The learning rate for local updates is initialized to $\alpha = 0.1$. We set the learning rate for meta-updates $\gamma = 10^{-4}$ and use the Adam algorithm (Kingma and Ba 2014) for meta-optimization. All the methods are implemented by PyTorch using Python and all the experiments are conducted on a Linux server with two 2.30GHz Intel(R) Xeon(R) Gold 5118 CPUs and one 11G GTX 1080ti GPU.

Evaluation Protocols

Performance Prediction Since cognitive states cannot be directly observed in practice, it is common to indirectly evaluate CDMs through the student performance prediction task on validation question sets (Wang et al. 2020), which is essentially a binary classification task. Specifically, we use two classification metrics, i.e., accuracy (ACC) and the area under the receiver operating characteristics curve (AUC).

Uncertainty Quantification To validate the effectiveness of model uncertainty qualification, we compute the reliability diagrams (Guo et al. 2017). In specific, a reliability diagram (e.g., Figure 3) visually measures how well calibrated the predictions of a model are by plotting the actual expected accuracy opposed to the output confidence of the model. A well-calibrated model with proper uncertainty quantification will have a small gap between its confidence and the actual accuracy, as it indicates that the predictive probability corresponds closely with how likely the prediction is actually correct, neither being overconfident or overcareful. More quantitively, we can compute the Expected Calibration Error (ECE) based on the diagram, which is a weighted average of accuracy-to-confidence differences (Guo et al. 2017).

Experimental Results

Student Performance Prediction (Q1) Table 2 shows the comparison results between BETA-CDMs and ORD-CDMs in student performance prediction using the ACC metric and the AUC metric, respectively. For reliability and comparability, data splits are conducted with 5 different random seeds, each time keeping the same split across all the CDMs. The table shows the averaged results, and Wilcoxon rank-sum statistical tests have been used to check whether the difference between the ORD-CDMs and our BETA-CDMs is statistically significant (with a 0.05 significance level). Specifically, on each dataset, we set the size of training data per new student as $3, 5, 10$. From the table, we can see that for every CDM, our BETA-CDM significantly outperforms the ORD-CDM on all the datasets. The results indicate that our framework is general to promote the cognitive state estimation of a wide range of CDMs by extracting useful prior knowledge from training data. Besides, there are two interesting obser-

Method	ECPE		ASSIST		EXAM	
	ACC	AUC	ACC	AUC	ACC	AUC
Ordinary	70.48	68.11	63.51	67.12	72.46	73.91
No-ML	72.33	69.36	65.91	70.69	74.61	78.79
No-BM	72.77	72.17	67.93	73.33	75.48	80.83
BETA-CD	**73.15**	**72.30**	**68.14**	**73.74**	**75.62**	**80.87**

Table 3: Ablation study results of BETA-CD with IRT.

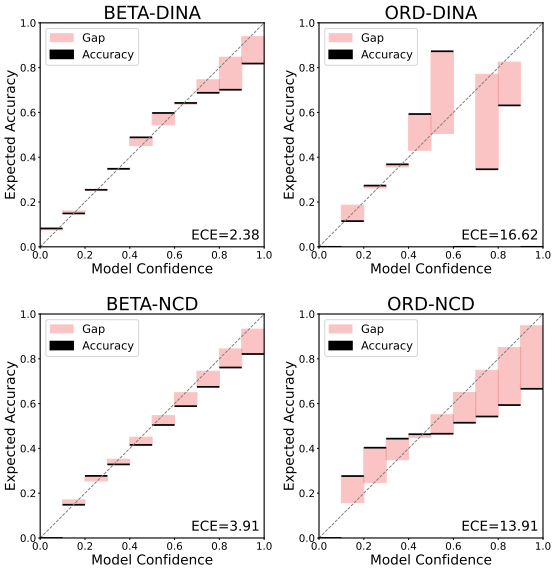

Figure 3: Comparison of reliability diagrams on ASSIST.

vations which confirms the role of the proposed framework for data sparsity problems. First, the BETA-CDMs has more obvious improvement compared to the ORD-CDMs when the amount of data is smaller. Second, BETA-CD has larger impact on the deep learning based model (i.e., NCD) than the other traditional models.

Ablation Study (Q2) To further examine different parts in the proposed framework, we compare BETA-CD to two variants. To validate the effectiveness of the Bayesian modeling, we remove uncertainty in cognitive state estimation by degenerating the prior $p(\theta_i|\psi)$ to $\delta(\theta_i - \psi)$ and the posterior $p(\theta_i|Q_i, R_i, \psi)$ to the maximum likelihood estimation (MLE), which is called *No-BM*. To validate the effectiveness of the meta-learning formulation, we manually specify a standard Gaussian prior $\psi = \{\mathbf{0}, \boldsymbol{I}\}$ rather than optimize ψ via meta-learning (Eq. (3)), which we call *No-ML*. Table 3 shows the results with IRT as the underlying CDM. The results on other CDMs are similar. As expected, meta-learning plays a key part in cognitive diagnosis. Besides, it is worth mentioning that Bayesian modeling also helps obtain better diagnosis results because a manually specified prior can incorporate prior knowledge as well.

Model Uncertainty Quantification (Q3) To validate the uncertainty quantification of BETA-CDMs, we use MIRT and NCD as base CDMs to compute their reliability diagrams and associated Expected Calibration Error (ECE) across different students on ASSIST dataset. We have similar results with other CDMs on other datasets. As shown in Figure 3, the predictive probability of BETA-MIRT and BETA-NCD is much closer to the actual accuracy than that of ORD-MIRT and ORD-NCD, respectively, indicating better calibration under our BETA-CD framework. Besides, the ECE scores quantitively support the same results. Hence, we conclude that BETA-CD can effectively quantify the model uncertainty such that we can really believe in the CDM when it gives confident predictions, and vice versa.

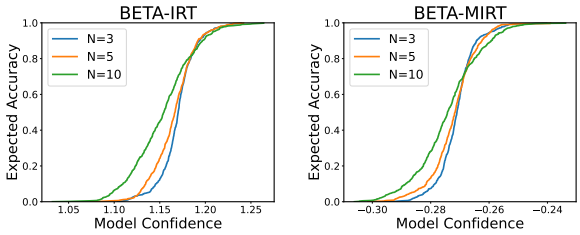

Figure 4: Empirical CDF curves on ECPE.

Benefits of Uncertainty Awareness (Q4) In principle, it is beneficial for intelligent tutoring systems to be informed of the reliability of the cognitive diagnosis through the model uncertainty, so that the systems can make a wise decision on whether to provide further services or collect more practice data. To validate this benefit, we construct a circumstance in which the model is supposed to be more uncertain, and check if BETA-CD can consistently detect larger model uncertainty. Specifically, we consider the cases where the numbers of practice logs from each student are $3, 5, 10$, respectively. In each case, we calculate the entropy of each student's cognitive state posterior, and plot the empirical Cumulative Distribution Function (eCDF) of the entropy values across all the students. Figure 4 shows the results of BETA-IRT and BETA-MIRT on ECPE dataset, and the results on other CDMs and datasets are similar. We can see that when there are fewer practice data, the eCDF is lower on small entropies and higher on large entropies, indicating an overall tendency of reporting larger uncertainty. Summarily, we conclude that BETA-CD is able to appropriately measure the reliability of the results via model uncertainty.

Conclusion

In this paper, we proposed a general Bayesian mETA-learned Cognitive Diagnosis framework (BETA-CD), which unifiedly addresses prior knowledge exploitation and model uncertainty quantification for cognitive diagnosis in the context of personalized learning. We firstly introduced Bayesian hierarchical modeling consisting of a shared prior encoding prior knowledge and a student-specific posterior conveying uncertainty. Furthermore, we formulated a meta-learning objective to automatically exploit prior knowledge and efficiently solved it with a gradient-based variational inference method. Extensive experiments have shown the effectiveness and generality of our framework.

Acknowlegements

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2021YFF0901003), and the Iflytek joint research program.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297.

Antoniou, A.; Edwards, H.; and Storkey, A. 2018. How to train your MAML. In *International Conference on Learning Representations*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 1613–1622. PMLR.

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

Embretson, S. E.; and Reise, S. P. 2013. *Item response theory*. Psychology Press.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 524–531. IEEE.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 1126–1135. PMLR.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic model-agnostic meta-learning. *NIPS*, 31.

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.

Grant, E.; Finn, C.; Levine, S.; Darrell, T.; and Griffiths, T. 2018. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.

He, W.; Li, Z.; Lu, D.; Chen, E.; Xu, T.; Huai, B.; and Yuan, J. 2020. Multimodal dialogue systems via capturing context-aware dependencies of semantic elements. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2755–2764.

Hospedales, T. M.; Antoniou, A.; Micaelli, P.; and Storkey, A. J. 2021. Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hu, B.; Zha, Z.-J.; Liu, J.; Zhu, X.; and Xie, H. 2021. Cluster and scatter: A multi-grained active semi-supervised learning framework for scalable person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2605–2614.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1073–1082.

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*.

Liu, Q.; Tong, S.; Liu, C.; Zhao, H.; Chen, E.; Ma, H.; and Wang, S. 2019. Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 627–635.

Liu, Q.; Wu, H.; Ye, Y.; Zhao, H.; Liu, C.; and Du, D. 2018a. Patent Litigation Prediction: A Convolutional Tensor Factorization Approach. In *International Joint Conferences on Artificial Intelligence*, 5052–5059.

Liu, Q.; Wu, R.; Chen, E.; Xu, G.; Su, Y.; Chen, Z.; and Hu, G. 2018b. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(4): 1–26.

Liu, Y.; Magnus, B.; O'Connor, H.; and Thissen, D. 2018c. Multidimensional item response theory. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 445–493.

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

Ma, W.; and Jiang, Z. 2021. Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied psychological measurement*, 45(2): 95–111.

Nguyen, C.; Do, T.-T.; and Carneiro, G. 2020. Uncertainty in model-agnostic meta-learning using variational inference. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3090–3100.

Paramythis, A.; and Loidl-Reisinger, S. 2003. Adaptive learning environments and e-learning standards. In *Second european conference on e-learning*, volume 1, 369–379.

Pei, H.; Yang, B.; Liu, J.; and Chang, K. 2020. Active surveillance via group sparse Bayesian learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rajeswaran, A.; Finn, C.; Kakade, S. M.; and Levine, S. 2019. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32.

Rakelly, K.; Zhou, A.; Finn, C.; Levine, S.; and Quillen, D. 2019. Efficient off-policy meta-reinforcement learning via

probabilistic context variables. In *International Conference on Machine Learning*, 5331–5340. PMLR.

Ravi, S.; and Beatson, A. 2018. Amortized bayesian meta-learning. In *International Conference on Learning Representations*.

Sullivan, T. J. 2015. *Introduction to uncertainty quantification*, volume 63. Springer.

van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M. G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1): 1–26.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6153–6161.

Yamaguchi, K.; and Okada, K. 2020. Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, 45(5): 569–597.

Zhan, P.; Jiao, H.; Man, K.; and Wang, L. 2019. Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, 44(4): 473–503.

Zhang, J.; Hao, B.; Chen, B.; Li, C.; Chen, H.; and Sun, J. 2019. Hierarchical reinforcement learning for course recommendation in MOOCs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 435–442.

Zhao, H.; Cheng, Y.; Zhang, X.; Zhu, H.; Liu, Q.; Xiong, H.; and Zhang, W. 2022. What is Market Talking about Market-oriented Prospect Analysis for Entrepreneur Fundraising. *IEEE Transactions on Knowledge and Data Engineering*.