中国科学技术大学
University of Science and Technology of China

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

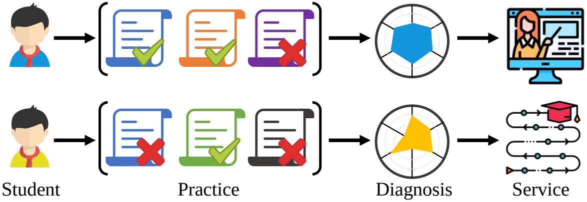
Haoyang Bi, Enhong Chen*, Weidong He, Han Wu,
Weihao Zhao, Shijin Wang, Jinze Wu

University of Science and Technology of China

2023.01.01

Content

- Personalized learning is a main component in *intelligent tutoring systems*
- Customize learning process and study experience for each individual student
- How to achieve personalized learning?
  - capture the student's personal state
  - provide downstream educational services
- Advantages of personalized learning:
  - lower practice burden
  - higher service quality



- Focus on the fundamental part of personalized learning
  - How to accurately capture personal states
- Cognitive Diagnosis (CD): a standard psychometric task for each student
  - Input: practice data (i.e., whether he/she answered questions correctly)
  - Output: cognitive state estimation (i.e., how well he/she grasps related knowledge)
- Cognitive Diagnosis Model (CDM): Math modeling to cognitive states
  - Example: Item Response Theory (IRT) Model
  - $p(r_{ij} = 1|\theta_i, \phi_j) = \text{sigmoid}(\theta_i - \phi_j)$
  - r_{ij}: response, θ_i: student's trait; ϕ_j: item's difficulty

Existing Challenges

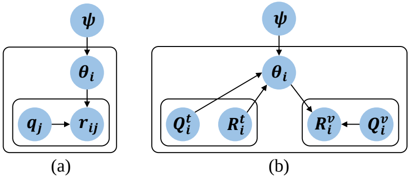
- Existing CDMs faces two challenges in personalized learning scenario
- Coping with data sparsity
  - Personalized learning requires minimum practice burden
  - However, fewer practice data lead to risk of overfitting
- Measuring reliability
  - Downstream applications heavily depend on the results of cognitive diagnosis
  - Being agnostic about an unreliable result may be dangerous
- Need a principled way to conduct CD tasks in personalized learning

- Propose a general Bayesian mETA-learned Cognitive Diagnosis framework (BETA-CD) to address the challenges
- Introduce Bayesian hierarchical modeling for CD task to unifiedly incorporate prior knowledge and model uncertainty
- Formulate a meta-learning objective to automatically exploit prior knowledge from historical data and solve it with gradient-based variational inference
- Conduct extensive experiments on various datasets and models to validate the effectiveness and generality of BETA-CD

- M students, $S = \{s_i\}_{i=1}^{M}$, N questions; $Q = \{q_j\}_{j=1}^{N}$; recorded triplets (s_i, q_j, r_{ij}); cognitive diagnosis model (CDM) $p(r_{ij} = 1 | q_j, \theta_i)$
- **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_*).

- Key idea
  - Prior knowledge exploitation: alleviate overfitting with meta-knowledge in massive data from other students
  - Model uncertainty quantification: enhance cognitive diagnosis results with probabilistic explanability
- Framework components
  - Bayesian Hierarchical Modeling: represent prior knowledge and model uncertainty
  - Meta-learned Prior Knowledge: meta-learning technique to exploit prior knowledge
  - Gradient-based Variational Inference: algorithm acceleration with approximation

- Traditional way of estimating point-wise cognitive states
  - $\theta_* = \arg\min_\theta - \log p(R_*|Q_*, \theta)$
  - Prone to overfit with few data
  - Little information about reliability
- View cognitive states in a probabilistic perspective
  - Prior distribution $p(\theta_i|\psi)$
  - Posterior distribution $p(\theta|Q_*, R_*, \psi)$
  - Bayesian inference $\theta_* \sim p(\theta|Q_*, R_*, \psi) = \frac{p(\theta|\psi)p(R_*|Q_*,\theta,\psi)}{\int p(\theta|\psi)p(R_*|Q_*,\theta,\psi)d\theta}$.



- The prior contains knowledge about the overall student population
  - e.g., average ability level
  - shared by all students
- The posterior represents the personal diagnosis result
  - can measure uncertainty such as entropy
  - specific to each student
- Next: how to specify a proper prior?

Meta-learned Prior Knowledge

- In the literature, the prior is manually determined
  - limited effect in preventing overfitting
- Discover prior knowledge automatically from practice data of historical students
  - optimizing the parametrized prior with a well-formulated meta-learning objective
- Key idea: exploit similar structures among CD tasks for each individual student
  - $\mathcal{T}_i = (Q_i^t, R_i^t, Q_i^v, R_i^v)$: 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)
- Meta-learning objective:

$$\min_{\psi} \sum_{i=1}^{M} \mathcal{L}_i^{(m)}(\psi) \equiv \sum_{i=1}^{M} -\log p(R_i^v | Q_i^v, Q_i^t, R_i^t, \psi)$$

$$= \sum_{i=1}^{M} -\log \mathbb{E}_{\theta_i \sim p(\theta_i | Q_i^t, R_i^t, \psi)} \left[p(R_i^v | Q_i^v, \theta_i) \right].$$

- Remaining problem: intractability of posterior
  - Especially in high-dimentional parameter space
- Solution: use a variational distribution $q(\theta_i; \lambda_i)$ to approximate the posterior
  - obtain $q(\theta_i; \lambda_i)$ by minimizing its KL divergence from the target distribution
- Posterior objective:

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

- Accordingly define a local loss

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

- And apply gradient-based optimization

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

- Advantages
  - Computationally efficient
  - Relate prior and posterior with gradient

Put it Together

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

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 $q(\theta_*; \lambda_*)$

Content

- Three real-world educational datasets
  - Different sizes and sources
- Four types of CDMs
  - Include classical and deep models
- Data split
  - For students: 60% as historical students, 20% as new students for evaluation/test
  - For records: 20% records of each student as validation item set

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

- Performance prediction
  - cognitive state is hard to observe
  - evaluate indirectly via student performance prediction task
  - essentially a binary classification task
  - use ACC and AUC as metrics
- Uncertainty quantification
  - whether the distributional result actually reflects uncertainty in prediction
  - reliability diagram: plotting the actual expected accuracy opposed to the output confidence of the model
  - expected calibration error (ECE): numerically measure the average distance in a reliability diagram

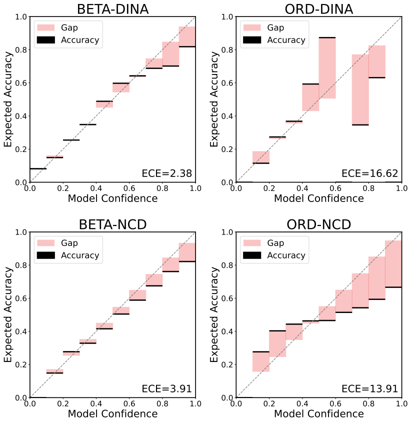
- Below show the overall performance comparison on the ACC metric
  - outperform on all datasets and base models
  - more advantage with fewer data

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**

- Evaluate with different components
  - No-ML: without meta-learning
  - No-BM: without Bayesian hierarchical modeling

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**

- Evaluate the effectiveness of uncertainty quantification
- The uncertainty information contained in BETA-CD is much more consistent with real predictive uncertainty



Content

Conclusion

- Proposed a general Bayesian mETA-learned Cognitive Diagnosis framework (BETA-CD)
- Introduced Bayesian hierarchical modeling, meta-learned prior knowledge and gradient-based variational inference
- Addressed prior knowledge exploitation and model uncertainty quantification for cognitive diagnosis in the context of personalized learning
- Validated the proposed method with extensive experiments

Thank you!