Bayesian Knowledge Tracing based on Transformer

1st Tingjiang Wei
Institute of Artificial Intelligence on Education
Shanghai Normal University

2nd Bingying Hu
College of Information,Mechanical and Electrical Engineering
Shanghai Normal University

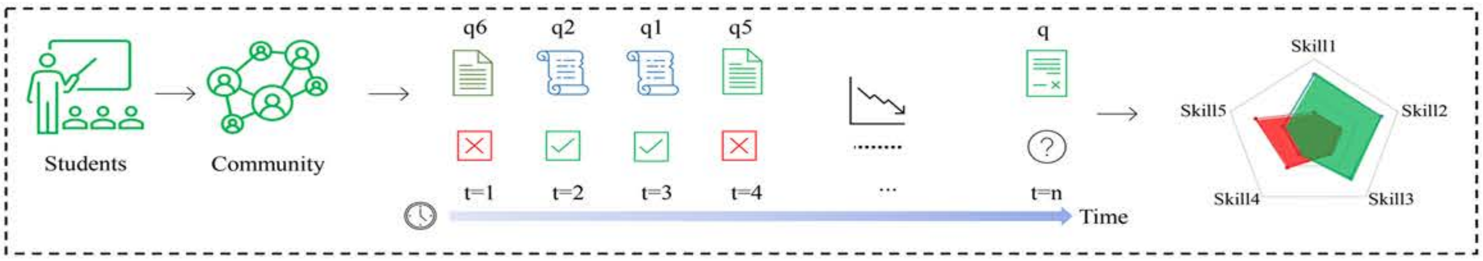
3rd Qin Ni
Institute of Artificial Intelligence on Education
Shanghai Normal University

INTRODUCTION

Personalized learning is an important research topic inadaptive learning systems. Educational data mining plays acrucial role in data analysis and intelligent tutoring systems. Adaptive learning systems require accurate assessmentof students' knowledge acquisition status. Knowledge tracing(KT) is a learning situation analysis task based on computer assisted large-scale data processing capabilities to track learning activities. KT assesses students' knowledge states (KS) by tracing their knowledge components (KCs).

Recurrent Neural Network (RNN) and Transformer-basedKT models have significantly developed, such as DKT andSAINT. Based on deep learning, the models do not need to encode the input knowledge explicitly and can be fitted to large scale data, showing great adaptability. However, teachers use simple and easy-to-understand teaching systems based on their tendency to use them in natural learning environments. Higher tracking accuracy does not substantially improve students'learning performance, as they have difficulty understanding the advice or guidance given by the system.

MATERIALS AND METHODS



Information about learners' input questions, skills, and responses are obtained from the dataset grouped by learners. The learning state of the learner s can be represented as a triad of questions, concepts and answers (q_i^s, c_i^s, r_i^s). So the learning sequence X can be represented as $\{(q_1, c_1, r_1), (q_2, c_2, r_2)...(q_i, c_i, r_i)\}$, where $q_i^s \in \mathbb{N}^+$ is the question number encountered in the learning process, $c_i^s \in \mathbb{N}^+$ represents the knowledge concept corresponding to the question, and $r_i^s \in \{0, 1\}$ indicates whether the learner has answered the question correctly.

As shown in [11], this model is trained by cross-entropy of the samples from the prior, the Prior-Data NegativeLog-Likelihood (Prior-Data NLL) is defined as (4).

$$\ell_\theta = \mathbb{E}_{D \cup \{x,y\} \sim p(\mathcal{D})} \left[-\log q_\theta(y \mid x, D) \right] \quad (4)$$

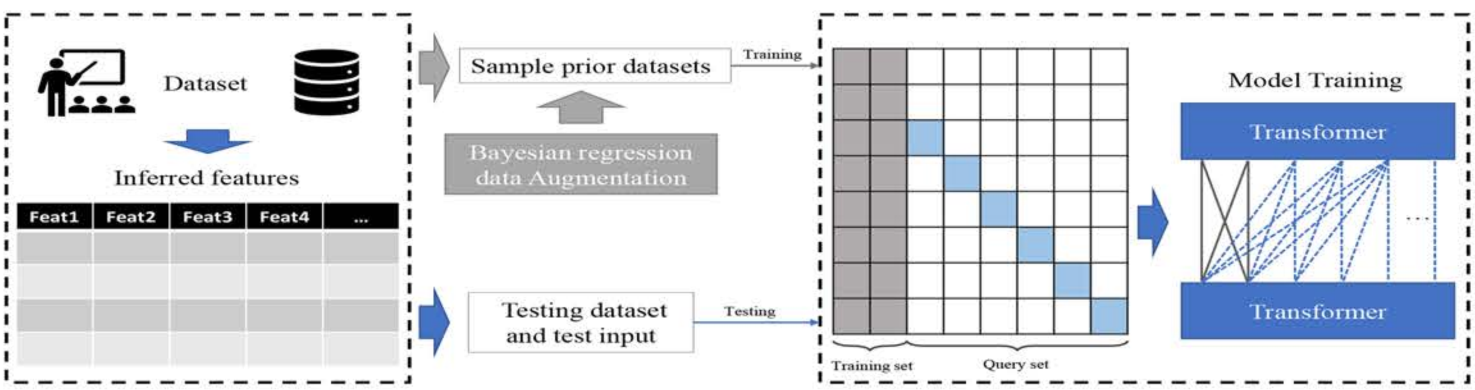
The prior $p(t)$ can be defined through the training set, then the posterior can be defined as $p(t|X)$, and thus the posterior prediction distribution can be as (2).

$$p(y \mid x, \mathcal{X}) = \int_t p(y \mid x, t) p(t \mid \mathcal{X}) \quad (2)$$

Define a transformer model with parameters q_θ and take the subset $Sub_X = \{(x_i, r_i)\}_{i=1}^n$ and the set Sub_{query} to be predicted as the input query, and the predicted output r can be obtained according to the query by the transformer.

$$Sub_X^{(i)} = X_{\text{train}}^{(i)} \cup \left\{ \left(x_{\text{test}}^{(i)}, r_{\text{test}}^{(i)} \right) \right\} \quad (3)$$

$$\mathcal{L} = -\sum_{i \in I} \left(r_i \log(p_i) + (1 - r_i) \log(1 - p_i) \right) \quad (5)$$



We sample datasets from a priori learning records and fit models on retained examples of these datasets. Given an actual dataset during testing, itand the test dataset are provided to the model and an approximation of Bayesian inference is obtained in a single forward propagation.

RESULTS

TABLE II
RESULTS.

Category	Methods	ASSIST2009 AUC	ASSIST2009 ACC
Regression-based model	IRT	0.664	0.653
	PFA	0.711	0.710
	DAS3H	0.742	0.723
Deep model	DKT	0.699	0.738
	SAKT	0.742	0.711
	SAINT+	**0.757**	**0.754**
	TBKT	**0.762**	**0.749**

Table II shows the prediction results of TBKT on theASSISTments2009 dataset. TBKT significantly outperformstraditional regression methods and the DKT model compared with other baseline models.

CONCLUSION

In the work of this paper, we explore the modeling of knowledge tracing tasks through a transformer-based Bayesian estimation approach. A Bayesian regression-based prior regression method is proposed to estimate the prior parameters, and a deep KT method with input feature distribution estimation is introduced. It is demonstrated that the remaining output can be inferred by observing only part of the data in the KT task by masking some of the values of the input training subset.

In the future, we plan to integrate more features in the a priori part and design more reasonable feature embedding methods for temporal type data. It is still worth discussing how to estimate the KT task using small sample data in a large-scale dataset.