Information Processing and Management 61 (2024) 103620

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Improving the performance and explainability of knowledge tracing via Markov blanket

Bo Jiang [a,b], Yuang Wei [b,*], Ting Zhang [c], Wei Zhang [b]

[a] Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China
[b] School of Computer Science and Technology (Shanghai Institute of AI for Education), East China Normal University, Shanghai, 200062, China
[c] Taizhou Branch of Zhenhai High School, Taizhou, 317503, China

ARTICLE INFO

Dataset link: https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198, https://sites.google.com/site/assistmentsdata/home/assistment2009-2010-data/skill-builder-data-2009-2010

Keywords:
Knowledge tracing
Markov blanket
Interpretable models
Educational AI
Causal discovery

ABSTRACT

Knowledge tracing predicts student knowledge acquisition states during learning. Traditional knowledge tracing methods suffer from poor prediction performance; however, recent studies have significantly improved prediction performance through the incorporation of deep neural networks. However, prediction results generated from deep knowledge tracing methods are typically difficult to explain. To solve this issue, a knowledge tracing model using Markov blankets was proposed to improve the interpretability of knowledge tracing. The proposed method uses the Markov blanket of the target variable as a subset of features and applies interpretable machine learning techniques to knowledge tracing. The results from the ablation experiments demonstrate that the feature subspace created by the Markov blanket is substantially effective for prediction. The proposed model also performs better than several other knowledge tracing models on two widely used datasets, i.e., Junyi and ASSISTments. Furthermore, the use of Markov blanket-based features provides high interpretability for predicting knowledge mastery states, elucidating the impact of these features on student knowledge acquisition. Moreover, this enables the use of previously considered low-correlation features, which may possess important latent causal relationships.

1. Introduction

A typical intelligent tutoring system comprises three models, i.e., domain, instruction, and learner (Koedinger et al., 2013). The learner model is used to reveal learner cognitive and noncognitive states, while driving the other two models to deliver personalized learning tasks. The learner model generally uses statistical models or machine learning algorithms to evaluate and trace student cognition, emotion, and attitude. Knowledge tracing (KT) has attracted considerable attention from the research community since the past two decades (Anderson et al., 1990; Cen, 2009; Piech et al., 2015). The KT generally uses data from learner historical responses to reveal the relationship between knowledge mastery status and performance in the learning process, extracting features of predictably mastered knowledge and adaptively modeling the progress of learner knowledge (Corbett & Anderson, 1995).

Most recently, the incorporation of deep neural networks has greatly improved the performance of KT, resulting in the construction of popular models such as Deep Knowledge Tracing (DKT) and Sequential Key-Value Memory Networks (SKVMN) (Yu et al., 2023). However, compared with the traditional KT models, such as BKT and AFM, DKT models, demonstrate weaker interpretability, making it challenging for users to understand the relationship among features, parameters, and results (Abdelrahman et al., 2023). Conversely, parameters such as error rates in BKT (Corbett & Anderson, 1995) and difficulty coefficients and learning

* Corresponding author.
E-mail address: philrain@foxmail.com (Y. Wei).

https://doi.org/10.1016/j.ipm.2023.103620
Received 25 July 2023; Received in revised form 15 November 2023; Accepted 23 December 2023
Available online 9 January 2024

List of abbreviations

AFM	Additive Factor Model
BKT	Bayesian Knowledge Tracing
DKT	Deep Knowledge Tracing
DT	Decision Tree
FGES	Fast Greedy Equivalence Search
FGES-MB	Fast Greedy Equivalence Search the Markov Blanket
GES	Greedy Equivalence Search
GS	Growing-Shrink
HMM	Hidden Markov Model
IAMB	Incremental Association Markov Blanket
inter-IAMB	interleaved Incremental Association Markov Blanket
IRT	Item Response Theory
KT	Knowledge Tracing
LR	Logistic Regression
MB	Markov Blanket
MB-DTKT	Markov Blanket-Decision Tree Knowledge Tracing
MB-KT	Markov Blanket-based Knowledge Tracing
MB-LRKT	Markov Blanket-Logistic Regression Knowledge Tracing
MB-RFKT	Markov Blanket-Random Forest Knowledge Tracing
MMMB	Min-Max Markov Blanket
PFA	Performance Factor Analysis
RF	Random Forest
S^2TMB	Score-based Simultaneous Markov Blanket
SLL	Score-based Local Learning
STMB	Simultaneous Markov Blanket

rates in AFM (Cen, 2009) can be used to explain prediction outcomes. Consequently, the tradeoff between interpretability and predictive performance presents a challenge for researchers. Nevertheless, in deep learning (DL) models, research has shown that feature selection impacts model performance and interpretability (Li et al., 2016, 2017), although the lack of interpretability in such methods hinders their use in the education domain.

Fortunately, causal models provide a valuable research tool that can lead to efficient model performance and interpretability in the physical and biological domains through the use of causal features (Chalupka et al., 2017). Moreover, the interpretability offered by causal feature theories is suitable for the education domain (Guyon et al., 2007). Unfortunately, there is currently no effective approach available to explore the significance and effectiveness of causal features in KT from a causal perspective.

In constructing a KT model with causal relationships, this study aims to accomplish two primary objectives.

1. To discover causal relationships and predict the knowledge state from online learning behavior data, we first developed a Markov Blanket (MB) discovery algorithm for the search model. This algorithm combines independence tests and scoring methods to identify causal feature subsets, with its effectiveness evaluated through ablation experiments.
2. To verify whether the causal feature subset learned from the MB improves the performance of the KT model, we combined the MB algorithm with various other machine learning algorithms to construct the Markov Blanket-based Knowledge Tracing (MB-KT) model. The effectiveness of the model was evaluated through comparison with other KT models. Furthermore, the interpretability of the MB itself, coupled with the use of interpretable machine learning techniques, guarantees that the KT model will be interpretable in application.

The rest of this paper is organized as follows. In Section 2, we review the literature related to KT and MB. The MB-KT method is described in Section 3. The dataset, experimental results, and analysis are presented in Section 4. Section 5 contains a discussion of this study. Finally, the conclusion is provided in Section 6.

2. Related work

2.1. Knowledge tracing

The KT refers to predicting the probability of mastering the knowledge components in a learning task for a learner based on their historical sequence of correct/incorrect answers. The KT, in essence, dynamically traces the knowledge state of the learner. From

machine learning perspective, KT presents a typical prediction problem in which latent variables (knowledge mastery) are estimated using noisy observation data (correct/incorrect sequence). Typically, KT models can be categorized into three types: Hidden Markov Model (HMM), Logistic Regression (LR), and Deep Learning (DL).

The HMM-based KT (Corbett & Anderson, 1995), of which Bayesian Knowledge Tracing (BKT) is the most used, treats the learner's performance on historical tasks as a variable and considers the knowledge mastery as state variables in its determination of the guessing and slipping probabilities of the learner. Using Bayesian probability to calculate the learner knowledge mastery probability under the given variables, the BKT models the learner knowledge state. However, the original BKT model did not factor in certain complex factors from the real learning environment, such as individual differences among learners, relationships between knowledge components, and knowledge forgetting. Ultimately, Pardos and Heffernan (2010) proposed the heuristic allocation of prior knowledge mastery probability to account for the initial individual differences of learners, whereas another study extended the BKT model parameters to accurately evaluate the learning and comprehensive abilities of students in an environment of various knowledge components (Mo et al., 2018).

The LR-based KT models are based on the Item Response Theory (IRT) (Harvey & Hammer, 1999) model, which assumes that various subjective and objective factors, such as learner ability, difficulty and discriminability of the item, guessing probability, and slipping probability, can affect learner response. The IRT model predicts the learner score based on these factors, using an LR (Kleinbaum & Klein, 2002) function. The Koedingerb research team at Carnegie Mellon University further expanded the parameters of the IRT model, adding learning rate and practice time as parameters and proposed the Additive Factor Model (AFM) (Cen, 2009). The AFM assumes that learning is a gradual process of change rather than a discrete transition, directly predicting the probability of the correct response of the learner to represent their knowledge state. The Koedingerb study (Pavlik et al., 2009) proposed that learning is not only influenced by the number of practice sessions but also by the performance in answering specific questions. Based on the AFM, the practice time parameter was further refined into the number of correct and incorrect answers with the assumption that the impact of the two answer scenarios on learning was different. On this basis, Performance Factor Analysis (PFA) model was proposed.

Owing to the remarkable performance of recurrent neural networks in handling time series tasks, researchers have attempted to apply them to KT tasks. The introduction of the DKT model has significantly improved the accuracy of KT compared with traditional BKT models (Piech et al., 2015). However, the DKT model also has limitations, including its inability to effectively trace the mastery of specific knowledge concepts and instability in prediction performance. To improve the DKT model, the regularization introduced for model loss function calculation, further improving model prediction accuracy and stability (Yeung & Yeung, 2018). Subsequently, another study introduced bidirectional long short-term memory networks into the original DKT model to obtain the textual features of the items (Yang & Cheung, 2018). An attention mechanism was also added to the prediction model, improving the model performance. Furthermore, forgetting has become an important feature to consider. For example, some researchers have simulated the KT process by considering the relationship between practice and forgetting behavior (Pandey & Srivastava, 2020). Some studies have modeled the KT process using learning curves and forgetting curves (Huang et al., 2020), while others have successfully enhanced predictive performance through the construction of DKT models that incorporate gate-controlled forgetting mechanisms and learning mechanisms (Zhao et al., 2023). In addition, researchers have tried to expand features such as answer time (Zhang et al., 2017), item information (Liu et al., 2019), and edge information (Wang et al., 2019) into recurrent neural networks, all of which have produced improvements in performance.

Subsequently, researchers proposed Relation-aware self-attention for Knowledge Tracing (RKT) model, considering the importance of exercise context information in EKT model. The RKT introduces the concept of relation coefficients to capture the relationships between exercises, which significantly improves upon the performance of KT models (Pandey & Srivastava, 2020). From a holistic data structure perspective, there exists a natural graph structure within knowledge components, and incorporating this graphical structure of knowledge components as additional information in the KT task is beneficial (Liu et al., 2021). Consequently, Graph-based Knowledge Tracing (GKT), which conceptualizes the underlying graph structure of knowledge components, has been adopted to enhance the performance of KT models (Nakagawa et al., 2019). Another study proposed Structure-based Knowledge Tracing (SKT), aiming to capture the multiple relationships within the knowledge structure and model the propagation of influence among concepts (Tong et al., 2020).

The existing research in this area demonstrates that performance improvements on KT models have been realized by adding features and incorporating graph structures. However, most of these studies have focused on exploring the relationships between features; however, uncovering the causal relationships among features remains a challenge.

2.2. Markov Blanket

Compared with DKT, KT based on HMM and LR offers substantial interpretability. Specifically, HMM-based KT involves the linkage of knowledge components in the model construction. However, owing to the significant time and computational requirements associated with learning a global Bayesian network in complex systems, the research community has begun to turn to local network structure learning algorithms as an alternative. The MB is a widely used local Bayesian network structure learning method, and researchers such as Koller have theoretically proven that the MB of a target variable is the optimal feature set of that target variable (Kohavi & John, 1997). The MB can shield the influence of other variables in the network structure on the target variable, making it an effective tool for mining relationships between target nodes. Since 1996, extensive studies have been performed building upon Koller's work, presenting various prominent algorithms such as Growing-Shrink (GS), Incremental Association Markov

Blanket (IAMB), and HITON-MB. Currently, MB learning algorithms can be described as two types: those based on the conditional independence test and those based on scoring.

The MB algorithms based on the conditional independence test use the conditional independence test method to determine the MB, such as the GS, IAMB, and HITON-MB. The GS (Margaritis & Thrun, 1999) algorithm is the first theoretically robust MB algorithm, proposed by Margaritis et al. The algorithm includes two stages, growing and shrinking, which are used to increase and decrease the candidate MB node set, respectively. In the growing stage, the GS algorithm traverses all nodes and adds nodes that are conditionally dependent on the target node of the candidate MB set. In the shrinking stage, the GS algorithm uses the conditional independence test method to check the conditional independence between each node in the candidate MB set and the target node and deletes erroneous nodes from the candidate MB set. The IAMB algorithm (Tsamardinos, Aliferis, Statnikov and Statnikov, 2003) is an improvement of the GS algorithm, in which the nodes with the highest correlation to the target variable are added at each iteration to the candidate MB set. Recently, researchers have improved the IAMB algorithm and proposed many variants such as Interleaved Incremental Association Markov Blanket (inter-IAMB) (Chang et al., 2018). However, the GS, IAMB, and their variants require an exponential relationship between the number of samples and size of the MB to obtain a reliable network structure. Therefore, to reduce the number of required samples, the Min-Max Markov Blanket (MMMB) (Tsamardinos, Aliferis and Statnikov, 2003) algorithm applies a divide-and-conquer approach to divide the learning of the MB of the target variable into learning of the parent–child nodes and spouse nodes. The HITON-MB algorithm (Aliferis et al., 2003) is an improved version of the MMMB algorithm, removing false parent–child nodes as early as possible from the candidate parent–child node set in the growing and shrinking stages. To reduce computational complexity, the Simultaneous Markov Blanket (STMB) algorithm was proposed (Gao & Ji, 2016), which also uses a divide-and-conquer strategy.

The MB algorithms based on scoring use a scoring function to evaluate the network structure and identify the Bayesian network structure with the highest score through the scoring function to determine the MB of the target variable. Typical algorithms include Score-based Local Learning (SLL) (Niinimäki & Parviainen, 2012) and Score-based Simultaneous Markov Blanket (S^2TMB) (Gao & Ji, 2017). The SLL algorithm finds the parent–child nodes of the specified target node based on the scoring function and then applies symmetry constraints to locate the spouse nodes on the parent–child nodes. The S^2TMB improves upon the SLL algorithm by eliminating the symmetry constraints in the spouse node search.

2.3. Causal and knowledge tracing

In the related work of knowledge tracking mentioned above, the benefits of graph structures have been explored for KT models, such as RKT (Pandey & Srivastava, 2020), GKT (Nakagawa et al., 2019), and AGKT (Long et al., 2022). In these studies, a type of associative relationship among knowledge components is examined to enhance model performance, which represents a pathway for improving KT. However, these associative relationships still maintain certain limitations. For example, they may not fully completely adhere to objective educational principles, leading to weaker model interpretability. In addition, most KT models consider data features as a flat structure, in which all features contribute to the final state of knowledge mastery (Kumar et al., 2023). Even studies that incorporate meaningful latent features (Minn et al., 2022) or skill hierarchical organization trees (Yang et al., 2021) face challenges in providing substantial pedagogical insights. In other words, the accurate prediction of future student performance becomes significantly more challenging in scenarios involving diverse instructional plans (variations in learning process features), rather than a fixed curriculum.

Exploration from these two perspectives has led to the consideration of studies that align more closely with the objective laws of the world, focusing on analyzing relationship structures and constructing causal relationships. Causal analysis tools are highly suitable for addressing these limitations in KT. The task of causal discovery, which involves using observed data to learn the causal relationships between different skills, is significant (Kumar et al., 2023). Causal networks or associations can assist educators in comprehending the prerequisite relationships among skills and guide them in sequencing topics in the curriculum, while assisting students in reviewing prerequisite information when facing difficulties (Desmarais, 2012). In addition, causal relationships between features facilitate causal reasoning, enabling the estimation of the effects of specific instructional treatments or interventions. Similarly, using causal inference can effectively enhance the interpretability of DKT (Huo et al., 2020), leading to the development of stable KT models (Zhu et al., 2023). For instance, in the context of high-performance DKT models, a universally interpretable approach, the Genetic Causal Explainer for Deep Knowledge Tracing (Li et al., 2023) is proposed. This approach explores the causal relationships between model inputs and outputs and introduces a causal attribution measurement method that considers the importance of the input units. This method helps avoid the influence of spurious correlations and allows the model to maintain excellent performance while providing interpretability in prediction and decision-making outcomes.

However, in the field of education, preinterpretable models are still considered superior to postinterpretable models owing to their transparency and comprehensibility. Nevertheless, the use of preinterpretable models relevant to KT application remains lacking in recent studies, particularly in studies that explore the causal relationships of input features within the models (Li et al., 2023).

3. Method

3.1. Markov Blanket learning algorithm considering prior knowledge

Presently used MB learning algorithms use either independence testing or scoring functions to select an appropriate search strategy for locating the MB of the target node. The MB learning algorithms using independence testing result in an undirected

Algorithm 1 FGES-MB

Input: Node dataset
Output: Markov blanket of target nodes
Step1: Determining node prioritization and maximum number of parent nodes
1. Calculate mutual information values among nodes, determine the maximum number of parent nodes, and prioritize nodes
Step2: Find neighboring node pairs of target node T
1. Given a scoring function $scoreI(node_1, node_2, Graph)$
2. Find nodes set $X = x_1, x_2, ...x_i..., x_n$ with $scoreI(X, T, G) > 0$, adding edges between x_i and T
3. Find nodes set $Y = y_1, y_2, ...y_j..., y_m$ with $scoreI(X, Y, G) > 0$, adding edges between x_i and y_j
Step3: Forward search
1.$(\forall x \in V, \forall y \in V | (x \neq y)) \cap y$ is not adjacent to x, $\Delta = scoreI(x \rightarrow y \subseteq edge) - scoreI(x \rightarrow y \not\subset edge)$
2. If $\Delta > \Delta_{max}, x \rightarrow y \subset \mathrm{MB}_T$
Step4: Backward search
1.$(\forall x \in V, \forall y \in V | (x \neq y)) \cap y$ is adjacent to x, $\Delta = scoreI(x \rightarrow y \subseteq edge) - scoreI(x \rightarrow y \not\subset edge)$
2. If $\Delta > \Delta_{max}, x \rightarrow y \subset \mathrm{MB}_T$
Step5: Return G,MB

graph, which cannot determine the directed edges between the target variable and its MB variables, thus failing to explain the causal relationship between variables. Most traditional MB learning algorithms using scoring functions require learning the global Bayesian network, which is a time-consuming process. To address the issues with existing MB learning algorithms, a combined approach using independence testing and scoring functions is proposed in this study to learn the MB of the target variable.

This study uses prior knowledge to restrict the scoring-based MB learning algorithm, reducing the search space and computation cost. Prior knowledge can be divided into node priority order and maximum parent node number, which are determined by combining node-sorting algorithms based on mutual information, expert knowledge, physical laws, and objective facts, and by calculating the mutual information among all nodes, respectively. We name this method Algorithm 1 Fast Greedy Equivalence Search-Markov Blanket (FGES-MB).

The FGES-MB algorithm is used to find the MB of the target node. The FGES-MB algorithm is based on the FGES (Ramsey et al., 2017) algorithm, with the addition of a constraint applied to determine the MB set containing the target variable. The search strategy and scoring function of the global Bayesian network structure learning algorithm are then used to learn the structure of the MB set. The FGES algorithm is an optimized and parallelized version of the GES (Meek, 1997) algorithm, further developed and researched by Chickering (2002). Unlike GES, FGES decomposes the global Bayesian network into local network structures and applies a scoring function to score the local network structures, with the process parallelized for operation on multiple network structures, which reduces the running time. In the FGES-MB algorithm, a target variable needs to be specified; in general, determining multiple MB sets containing the target variable is less time-consuming than characterizing the network structure of all variables.

We use the target node T as an example of the algorithm process. First, we calculate the set of associated nodes A that exhibit a high degree of correlation with the target node. Based on this set and the target node, we generate a randomly initialized graph. Next, we use the scoring function $scoreI(node_1, node_2, Graph)$ to compute the score of the current graph. To ensure the interpretability of the graph structure and reduce its complexity, we apply the Bayesian Information Criterion (BIC) scoring function in the learning model (Neath & Cavanaugh, 2012).

During the learning process of the MB structure, the associated nodes are divided into two categories: direct parent–child and sibling nodes. We represent this as (x, y). Starting from the given randomly initialized graph, the direct association structure is evaluated by checking if $scoreI(x, T, G) > 0$. Then, the presence of indirect association structures is determined by checking if $scoreI(x, y, G) > 0$. After each evaluation, a forward search is performed in the subset not containing indirect relationships. If a better network structure (i.e., a higher structural score) is found, the added directed edges are retained. Then, a backward search is conducted in the subset with indirect relationships to prune the edges and obtain the final Markov blanket.

3.2. Markov Blanket-based knowledge tracing algorithm

The MB is a set of nodes in a Bayesian network that D separates the target node T from all other nodes. It is denoted as *MB(T)*. In a Bayesian network, the *MB(T)* of a target variable T is the set of all parents *Pa(T)*, children *Ch(T)*, and spouses *Sp(T)* of the target node, i.e., $MB(T) = Pa(T) \cup Ch(T) \cup Sp(T)$. In Fig. 1, the blue node represents the *MB(T)* of the target node. Based on the properties of the Bayesian network structure, for a given *MB(T)*, the target node T is conditionally independent of all the other nodes in the Bayesian network. Under the faithfulness assumption, the *MB(T)* of the target variable is the optimal feature subset for classification tasks. Therefore, theoretically, constructing a KT model with the MB of the predictive variable as the feature set may lead to optimal predictive performance. Furthermore, some features with low relevance may have a direct causal relationship with the predictor variables. If this causal relationship could be understood and explained, models based on the MB would have higher interpretability, portability, and generalization ability than traditional correlation-based KT models.

To conduct concrete KT experiments, we first need to calculate the mutual information between the data nodes, determine the priority order of the nodes, and obtain the maximum number of parent nodes through Algorithm 1. Then, we construct the

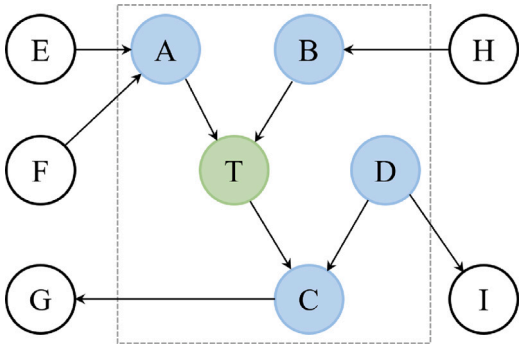


Fig. 1. Markov blanket: The circles in the diagram represent the nodes, lines with arrows represent the parent–child relationships, green nodes represent the target nodes, and blue nodes represent the Markov blankets of the target nodes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

MB-KT model based on the MB of the target variable, used to improve the KT model. From a machine learning perspective, KT can be abstracted as a sequence prediction problem, whose goal is to estimate the latent variable (e.g., knowledge mastery, learning motivation) based on noisy observation data (e.g., correct/incorrect sequences, learning behavior). Therefore, we use machine learning methods to improve the parameters and classifiers of the KT model. Machine learning models exhibiting good interpretability and model performance were selected, such as LR (Kleinbaum & Klein, 2002), Decision Tree (DT) (Quinlan, 1996), and Random Forest (RF) (Breiman, 2001) as classifiers for the model.

Specifically, if we use the MB feature set of the target variable to improve the existing LR-based KT model, we need to modify the model parameters and obtain the improved model, named Markov Blanket-Logistic Regression Knowledge Tracing (MB-LRKT). The formula for the model is as follows.

$$P(R_{ij} = 1|z) = \frac{1}{1 + e^{-z}} \tag{1}$$

$$Z = \beta_0 + \beta_1 mb_1 + \beta_2 mb_2 + \cdots + \beta_n mb_n \tag{2}$$

We integrate the significance of the parameters in the formula with the essence of the learning process, redefining the meaning of the formula using different features from the Markov blanket. We assume the KT problem to be a binary classification task (considering only the correctness of objective questions) (Mongkhonvanit et al., 2019). Here, $R_{ij} = 1$ represents a correct answer given by the learner. Thus, R_{ij} can be widely understood as the learning outcome of the learner, which in our algorithm represents as the target node in the Markov blanket. Therefore, independent variable Z represents a collection of factors that influence the learning outcome, and this collection is defined using the MB. Z encompasses all features that have causal effects on the target node. Through training, we obtain the regression coefficients β_n for different features to fit the underlying relationships in the data. In addition, the parameters automatically adapt and change as the learning progresses, enabling the dynamic capture of student learning outcomes. This allows for better KT performance along the time series.

Similar to the MB-LRKT model, the Markov Blanket-Decision Tree Knowledge Tracing (MB-DTKT) model was developed by combining the MB features of the target variable with DT. Using the target variable's MB features, the Markov Blanket-Random Forest Knowledge Tracing (MB-RFKT) model, a modified KT model, was also developed using RF for modeling. The procedure of the algorithm is shown in Algorithm 2.

4. Experimental results and analysis

4.1. Dataset

This study uses two real education datasets, the Junyi[1] dataset and the ASSISTments(2009–2010)[2] dataset, to evaluate the effectiveness of the model.

1. The Junyi dataset is a collection of educational data from the Junyi Academy, an online education platform developed by the Junyi Academy Foundation in Taiwan, China, to provide various educational resources for K-12 students. The Junyi dataset comprises learning log data collected from 247,606 students over a span of two years. The dataset primarily focuses on mathematics courses in which students have the opportunity to practice problems multiple times and can use hints during their learning process. The platform also provides personalized recommendations based on student knowledge states, allowing

[1] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198.

[2] https://sites.google.com/site/assistmentsdata/home/assistment2009-2010-data/skill-builder-data-2009-2010.

Algorithm 2 MB-KT

Input: Node dataset (*data*), FGES-MB graph (D_{MB})
Output: Knowledge tracing outcomes
Step1: Preprocessing method
1. Calculate the mutual information between nodes, and determine the priority order of target nodes
2. Set a threshold, the maximum number of parent nodes is determined by the number of features with a correlation greater than the threshold
3. Calculate MB using Algorithm 1: D_{MB}
4. Obtain the feature set that affects the target node

Step2: Parameter calculation
while $loss >= 0.015$:
Choose the machine learning model: LR, DT, RF
$\boldsymbol{\beta} = \beta_0, \beta_1, ..., \beta_n = f(data, D_{\mathrm{MB}})$
Get $\rightarrow P(True|\boldsymbol{\beta})$
Update$\rightarrow loss$

Step3: Dynamic tuning
while New round of response is TRUE:
Outputs $\rightarrow P(True|Step2_{\beta}, response)$

Return: Outputs

Table 1
Data field.

DataSet	Junyi(17 fields)		ASSISTments2009–2010(29 fields)		
Field name	user_id	hint_time_taken_list	order id	assignment id	user id
	suggested	problem_number	problem id	original	correct
	topic_mode	problem_type	attempt count	ms first response	tutor mode
	review_mode	time_done	answer type	sequence id	student class id
	time_taken	time_taken_attempts	position	problem set type	base sequence id
	correct	count_attempts	skill id	skill name	teacher id
	hint_used	count_hints	school id	hint count	hint total
	exercise	earned_proficiency	overlap time	template id	answer id
	points_earned		answer text	first action	bottom hint
			opportunity	opportunity original	

students to choose whether or not to attempt the recommended exercises. The Junyi dataset has been widely used in studies over the past few years. It consists of 17 fields and includes behavioral data and outcome data, making it suitable for studying KT models (see Table 1). Moreover, we have previously employed the Junyi dataset as the foundation for causal experiments in our research, achieving excellent results (Bo et al., 2023). Therefore, we have elected to continue using the Junyi dataset as our experimental data.

2. ASSISTments is an online education platform developed by Worcester Polytechnic Institute. Herein, the ASSISTments Math 2009–2010 dataset is used, which is a collection of math course learning records from the ASSISTments system during the 2009–2010 academic year. It includes various student interactions such as problem-solving attempts, hints used, and response correctness. The ASSISTments dataset is one of the most well-known publicly available large-scale datasets in the field of education, applied in various education studies. There are three versions: 2009–2010, 2015, and 2017. For our experiment, we have chosen the 2009–2010 version for several reasons. (1) First, ASSISTments2015 comprises only four fields, primarily intended for learning outcome prediction models. These fields are not particularly suitable for exploring causal relationships, which is a focus of our experiment. (2) ASSISTments2017 comprises more than 70 fields, making it excessively large for effectively studying the Markov blanket structure. In addition, these fields incorporate emotional factors derived from other models, introducing additional uncertainty. (3) ASSISTments2009–2010 comprises 29 fields, all of which capture objective interaction behavior and learning outcome data. These fields do not introduce any uncertain factors, demonstrating suitability for our experiment. Considering these factors, we have selected ASSISTments2009–2010 as the most appropriate dataset for our experiment.

4.2. Results and analysis of Markov Blanket learning

In this section, we will apply the proposed MB algorithm, which considers prior knowledge, to the Junyi dataset and ASSISTments (2009–2010) dataset. First, we calculate the mutual information values between the nodes on both datasets using the Pearson correlation coefficient method, which helps us determine the priority order of nodes.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3}$$

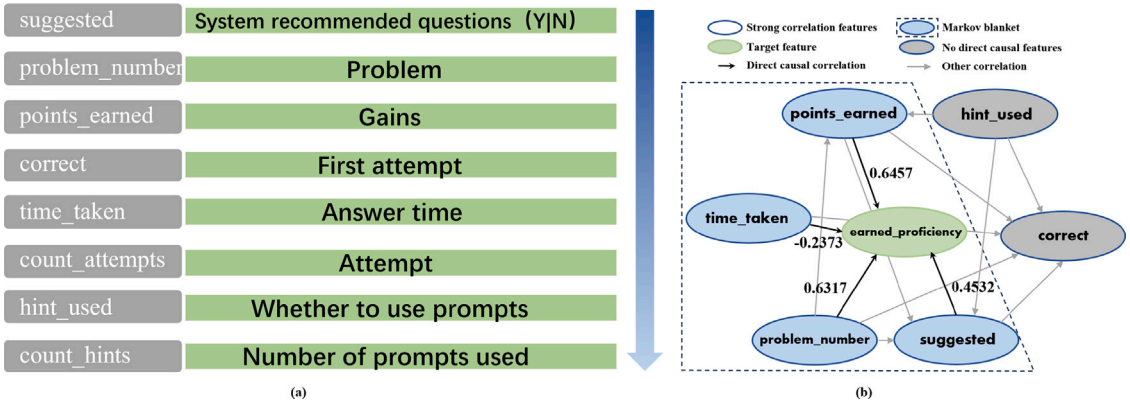


Fig. 2. Markov Blanket learning result graph based on Junyi dataset: (a) using the Pearson correlation coefficient, features with a correlation coefficient greater than 0.01 were identified and ranked accordingly. (b) By applying Algorithm 1, a network graph incorporating the variable *earned_proficiency* was generated and its associated MB was obtained.

cov(·) represents covariance, σ represents variance, μ represents mean, and E(·) represents expectation. Then, using the correlated features and their data, we perform network structure learning to obtain the MB of the specified node.

We select *earned_proficiency* from the Junyi dataset and *correct* from the ASSISTments (2009–2010) dataset as the target nodes for both datasets. Although the Junyi dataset also includes the *correct* feature, it was not chosen as the target feature for Junyi because of the following reasons: (1) for learning purposes, the correctness of a question is an indirect measure of student learning gains. When there are other types of learning gain features available, we prioritize the use of those types of features. (2) In our previous research, we already established a causal relationship between the *earned_proficiency* field in the Junyi dataset and other fields. Hence, we consider using *earned_proficiency* as the target variable. However, the ASSISTments (2009–2010) dataset only has the *correct* feature available; thus, we can only use *correct* as the target feature.

In the Junyi dataset, the *earned_proficiency* is selected as the target variable. With a threshold of 0.01, we establish the priority order of feature correlation coefficients, as illustrated in Fig. 2(a). Using the selected features and their associated data as prior knowledge, we apply the iterative loop of Algorithm 1 to construct a feature network graph that includes the *earned_proficiency* as the target variable. Ultimately, we obtain the MB associated with the *earned_proficiency*, as depicted in Fig. 2(b).

In the ASSISTments (2009–2010) dataset, the *correct* is selected as the target variable. With a threshold of 0.01, we establish the priority order of the feature correlation coefficients, as illustrated in Fig. 3(a). Using the selected features and their associated data as prior knowledge, we apply the iterative loop of Algorithm 1 to construct a feature network graph that includes the *correct* as the target variable. Ultimately, we obtain the MB associated with the *correct*, as depicted in Fig. 3(b).

Both datasets have relatively weak feature correlations, thus, a very small threshold was applied to select relatively correlated features for training. For the Junyi dataset, correlation analysis identified eight relatively correlated features. However, an examination of causal graph revealed that there is no direct causal relationship between *hint_used* and *correct* with *earned_proficiency*. This can be understood as *correct* represents the correctness of the initial response, which may not directly correlate with learning outcomes, as the first response could involve guessing or making mistakes. However, *hint_used* indicates whether the student used hints. If using hints is expected to lead to learning gains, it suggests that the hints provided are of high quality, which cannot be guaranteed. Therefore, the MB algorithm excludes these two features.

For the ASSISTments(2009–2010) dataset, the most significant difference between features with relatively strong correlations and the Markov blanket is the inclusion of the *hint_total* feature. By examining the causal graph, the varying degrees of causal effects that different features have on the target node can be observed based on the available data. In the feature analysis, it is noteworthy that the *hint_total* feature, which falls below the threshold, exhibits higher effect values in the graph than some other features. We hypothesize that the value of *hint_total* may determine the importance of a particular question and impact the student's ability to access multiple ways of solving that question. These factors are critical determinants of whether a student can successfully solve a particular problem. Consequently, through an extensive analysis of the data, the MB algorithm has uncovered this additional feature. To validate the effectiveness of the features obtained from the MB algorithm, the KT model can be constructed for verification.

4.3. Results and analysis of knowledge tracing

To demonstrate the impact of MB features on the performance of KT models, ablation experiments were conducted on the MB and other features in the datasets, with the results illustrated in Table 2. The measurement indicators are Accuracy (ACC), Precision, Recall, F1-score (F1), and area under the curve (AUC).

As shown in Table 2, the models constructed using the MB features generally outperformed those constructed using all other features on the Junyi and ASSISTments (2009–2010) datasets, further confirming that the MB features are effective in modeling learner knowledge states.

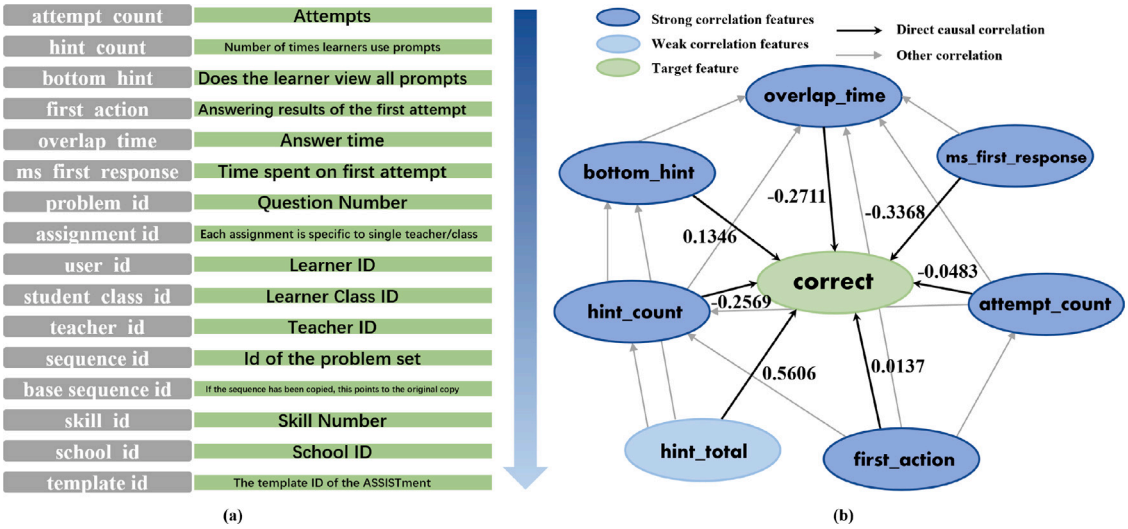
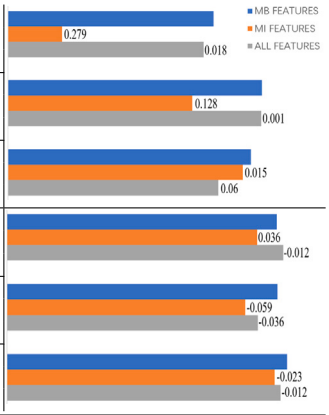


Fig. 3. Markov Blanket Learning result graph based on ASSISTments(2009–2010) Dataset: (a) using the Pearson correlation coefficient, features with a correlation coefficient greater than 0.01 were identified and ranked accordingly. (b) By applying Algorithm 1, a network graph incorporating the variable *correct* was generated and its associated MB was obtained.

Table 2
Comparison of ablation experiment results.

DataSet	Model	Accuracy	Precision	Recall	F1	AUC	Comparison of AUC
Junyi	MB-LRKT	0.666	**0.667**	**0.665**	**0.665**	**0.779**	
	MI FEATURES	0.674	0.478	0.500	0.488	0.500	
	ALL FEATURES	**0.956**	0.278	0.500	0.488	0.761	
	MB-DTKT	0.865	**0.891**	**0.865**	**0.863**	**0.868**	
	MI FEATURES	**0.968**	0.836	0.740	0.779	0.740	
	ALL FEATURES	0.956	0.278	0.500	0.488	0.867	
	MB-RFKT	**0.970**	**0.814**	**0.848**	**0.833**	**0.848**	
	MI FEATURES	0.969	0.813	0.848	0.831	0.833	
	ALL FEATURES	0.969	0.826	0.788	0.806	0.788	
ASSISTments (2009-2010)	MB-LRKT	0.897	0.914	0.897	0.896	0.897	
	MI FEATURES	0.861	0.890	0.861	0.859	0.861	
	ALL FEATURES	**0.908**	**0.919**	**0.909**	**0.908**	**0.909**	
	MB-DTKT	**0.897**	**0.914**	**0.898**	**0.897**	**0.898**	
	MI FEATURES	0.839	0.876	0.839	0.835	0.839	
	ALL FEATURES	0.861	0.889	0.862	0.858	0.862	
	MB-RFKT	**0.938**	**0.938**	**0.916**	**0.908**	**0.916**	
	MI FEATURES	0.893	0.894	0.893	0.893	0.893	
	ALL FEATURES	0.914	0.905	0.904	0.904	0.904	



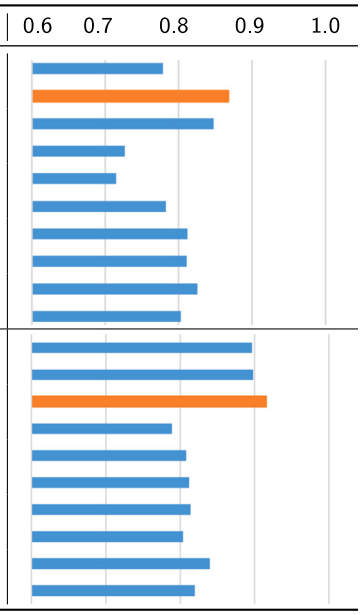
*MI FEATURES represent features with strong correlation, and ALL FEATURES represent the use of all graph nodes for prediction. The numeric values in the Comparison of AUC represent the performance improvement achieved by using the MB model.

Comparing the results from two datasets and three types of models, totaling six groups of experimental data, reveals that the performance of features with calculated mutual information (MI FEATURES) does not always surpass the baseline model that considers all features. First, we examine the MI features in the Junyi dataset, which comprises six strongly correlated features. Compared with the baseline model using all features, the performance is weak. This result may occur because there is more latent information within the additional features, and the strongly correlated features may also introduce confusion into the predictions. Similarly, in the ASSISTments dataset, the MI FEATURES, which includes six strongly correlated features, also exhibit weaker performance than the baseline model using all features. This further confirms that including more features introduces additional latent information, which leads to improved accuracy in prediction. However, when comparing the results obtained using causal features, the baseline model using all features performs poorly. This indirectly indicates that different features have varying effects on model prediction results, which can be positive or negative. This highlights the necessity of our study.

A more specific analysis revealed that models using MB features for KT outperformed models using ALL FEATURES in the Junyi dataset. However, in the case of the KT model built using LR on the ASSISTments dataset, the performance of the model using MB features did not significantly surpass the performance of the model using ALL FEATURES. This result can be attributed to the limitations of LR in adapting to different data and scenarios, compared with tree-based algorithms, which exhibit stronger adaptability. This adaptability may increase with the addition of more features, leading to a slight improvement in performance.

Table 3
Comparison of prediction results of knowledge tracing models.

DataSet	Model	AUC
Junyi	MB-LRKT	0.779
	MB-DTKT	**0.868**
	MB-RFKT	0.848
	AFM	0.727
	DKT	0.715
	DKT-CART	0.783
	DKT-RF	0.812
	DKT-GBDT	0.811
	SKVMN	0.826
	DKVMN	0.803
ASSISTments(2009-2010)	MB-LRKT	0.897
	MB-DTKT	0.898
	MB-RFKT	**0.916**
	AFM	0.789
	DKT	0.808
	DKT-CART	0.812
	DKT-RF	0.814
	DKT-GBDT	0.804
	SKVMN	0.840
	DKVMN	0.827



*DKT-CART, DKT-RF, DKT-GBDT come from Yang and Cheung (2018); SKVMN, DKVMN come from Abdelrahman and Wang (2019).

Although the model using ALL FEATURES demonstrated slightly higher performance compared with the model using MB features, it is important to consider practical application scenarios. In the case of the ASSISTment (2009–2010) dataset, if we exclude the target variable, the number of features is reduced to 28 (ALL FEATURES = 28), whereas the number of the MB features is 7. This means that the model using ALL FEATURES requires more than four times the computational effort of the model using the MB features. Considering the marginal performance improvement of less than 0.01, it is not cost effective to increase the computational burden to such an extent. Thus, from a practical standpoint, using the MB features is demonstrated to be a more efficient choice.

To verify that MB can improve the effectiveness and performance of KT, the MB-KT model was constructed using Algorithm 2. The MB of the predicted variables was selected as the feature set, and machine learning methods were used for modeling and model comparison. The comparison results are shown in Table 3.

An analysis of the results revealed that tree-based algorithms often exhibit good performance in KT, e.g., MB-DTKT, MB-RFKT, DKT-CART, DKT-RF, and DKT-GBDT. However, algorithms that use feature associations or concept relationships as prior knowledge for model construction also demonstrated improved performance in KT, e.g., MB-LRKT, MB-DTKT, MB-RFKT, and SKVMN. The use of tree-based algorithms leverages the ability to capture complex patterns and relationships within the data, which is crucial in KT tasks. In addition, the incorporation of feature associations or concept relationships as prior knowledge provides additional insights into the underlying structure of the KT process, leading to greater prediction accuracy. The combination of these two approaches allows the MB-KT model to achieve superior performance in KT, as demonstrated by the analysis results. Therefore, it can be observed that the ability to handle multidimensional features and capture complex patterns and relationships in the data is very important in KT tasks. In addition, incorporating feature associations or concept relationships as prior knowledge provides additional insights into the underlying structure of the KT process, securing accurate predictions. The combination of these two approaches allows the MB-KT model to achieve superior performance in KT, as demonstrated by the analysis results.

To illustrate a practical example of our model's performance in KT, we randomly selected a student from the ASSISTments dataset and made predictions of learning progress as a function of time. As shown in Fig. 4, the results indicate that our model effectively tracks the student's learning progress.

Furthermore, to demonstrate the impact of causal features on learner outcomes, we randomly selected 200 instances of positive and negative learning gains (true state and false state of *earned_proficiency*) from the Junyi dataset. We compared the performance relationship between the causal features and learning gains. Fig. 5 presents three of the four causal features obtained from Fig. 2. This choice was made as the influence of *problem_number* on learning gains is primarily reflected in its latent information, rather than its representational information. Therefore, we focused on showcasing the relationships between the remaining three features and learning gains. Considering the causal effects of the three features presented in Fig. 2, the *time_taken* feature shows a negative causal relationship with learning gains. However, in Fig. 5, the values of the blue triangles are lower under positive learning gains than under negative learning gains. However, the *points_earned* and *suggested* features demonstrate a positive causal relationship. Consequently, in Fig. 5, the values of the red circles and green squares are higher under positive learning gains than under negative learning gains.

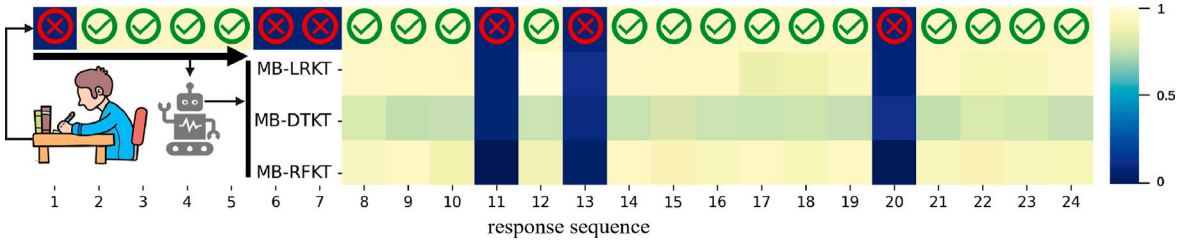


Fig. 4. Visualization depicting the predicted response results for a randomly selected student in the time series of the ASSISTments (2009–2010) datasets. The first row in the graph represents the original student response sequence, where light green indicates a correct response and dark blue represents an incorrect response. The following three rows represent the prediction results of three models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

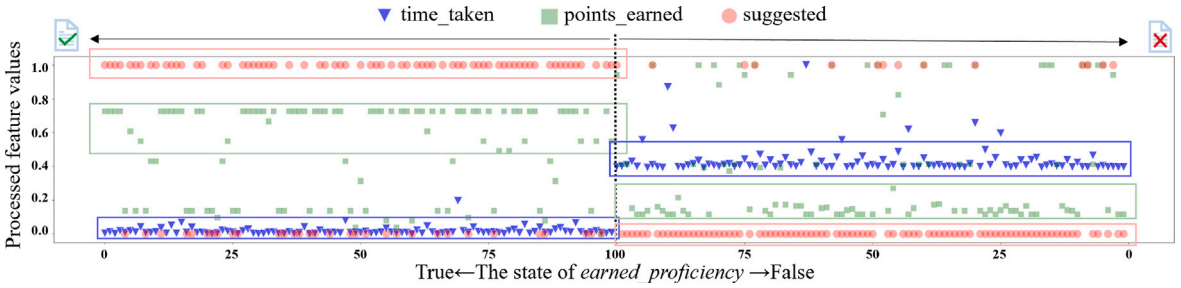


Fig. 5. Visualization of causal effects of features. The graph depicts the concrete effects of features on learning gains. The range of 0–100 represents the impact of features on positive learning gains, whereas the range of 100–0 represents the impact on negative gains. The vertical axis represents the feature data scaled to the interval (0, 1).

5. Discussion

The objective of this study is to propose an interpretable model for educational intelligence that enhances personalized teaching effectiveness, while simultaneously addressing the ethical concerns associated with using AI in sustainable education processes.

The proposed MB model can effectively identify features that exhibit causal relationships with the target variable. The integration of the MB model with KT models enhances the predictive performance, interpretability, and generalization ability of the overall models. We have presented comprehensive evidence, including algorithm structure and performance metrics, to substantiate the performance and interpretability of the model.

In some studies, certain models tend to overlook features with weak correlations to the target variable. However, these features may still demonstrate potential causal relationships with the target feature that can be interpreted. For instance, in experiments conducted with the ASSISTments dataset, the *hint_total* feature represents underlying information such as derived content, detailed guidance, and question quality related to a particular question. When students engage with this question, the latent information carried by the *hint_total* feature is often considered the source of their actual learning gains. Similarly, in the Junyi dataset experiment, the exclusion of the *hint_used* feature yields findings that support the previous explanation. The hints provided for questions indeed contain various latent content; however, the quality of this information is difficult to ensure when compared in equal quantities. Therefore, whether or not hints are used can introduce even greater errors in predicting student learning performance. As such, the integration of correlation information with causal relationships can significantly aid in the prediction and explanation of student learning performance and gains.

However, the learning method using MB represents a fundamental approach to causal network learning. We have only calculated the causal effects between features; however, the true strength of causality has not yet been defined. According to research conducted by Rajkumar et al. (2022), causal relationships between features may fall into two categories: weak and strong ties. Thus, it is advisable to perform intervention experiments and counterfactual experiments to reconstruct the network. Furthermore, there is scope for improvement in excluding confounding factors in this study. In the MB learning algorithm, the presence of unobserved latent variables can have a significant impact on causal inference. In addition, highly correlated features can affect the results of causal inference. Moreover, if the dataset contains noise or missing data, obtaining accurate causal relationships can be challenging. In the experiments conducted in this study, the handling of features was enhanced by incorporating prior knowledge; however, achieving complete automation would require further improvements in the learning algorithms.

This study only used transparent machine learning models, such as LR and DT. However, for better predictive performance, it is worth considering incorporating these learning models as prior knowledge into DL models. A similar study, conducted by Chen et al. (2018), suggests that prerequisite relationships between knowledge components should also be regarded as important features to be integrated into DKT models. Assuming that knowledge component 1 is a prerequisite for knowledge component 2 suggests

that mastering knowledge component 1 is critical for correctly answering exercises related to knowledge component 2. To address this scenario, a prerequisite relationship matrix was introduced as a constraint and added to the model input.

Subsequent research has further advanced the study of the interrelationships between knowledge components (Long et al., 2022; Nakagawa et al., 2019; Pandey & Srivastava, 2020). In another study (Chaudhry et al., 2018), the impact of hint information on student abilities was considered. Improvements were made by integrating the student usage of hints as a feature into the DKVMN model, transforming it from a single-task model to a multitask collaborative prediction model. The research direction of KT is gradually migrating toward the effectiveness of multiple features, including knowledge component labels (Yu et al., 2023), knowledge component prerequisites (Pandey & Srivastava, 2020), item prerequisites (Nakagawa et al., 2019), and learning process behaviors (Xu et al., 2023). Our model represents an exploration in this direction to investigate the true influence of multiple features on KT models from a causal perspective.

The goal of KT models is to serve education. From an educational perspective, our research offers a fresh perspective on personalized learning by examining the causal relationships that impact student learning outcomes and making predictions based on these associations. This approach not only assists educators in identifying and prioritizing what is essential during the teaching and learning process but also permits the early anticipation of potential changes in student performance. Importantly, our findings are both interpretable and comprehensible, which enhances their credibility among frontline educators and establishes them as reliable educational tools. The trustworthiness of such tools directly influences the advancement of sustainable education. For instance, if bias exists in the training data of AI systems, such as racial bias, it can lead to educational inequality, which is unacceptable. Similarly, algorithmic recommendations in personalized instruction may introduce bias and discrimination, creating an "information cocoon" that restricts learners to a narrow curriculum while disregarding flexibility and diversity in their learning and development. Therefore, when developing adaptive learning systems or personalized platforms driven by intelligent models, careful consideration should be given to the following: (1) identifying the tool user base; (2) evaluating the educational benefits it provides; and (3) ensuring the reasonability and interpretability of those benefits (Khosravi et al., 2022). Currently, trustworthy AI remains a research challenge; however, a growing number of researchers are dedicated to developing high-performance AI models with interpretability. Looking ahead, we anticipate the increased adoption of personalized learning approaches, interpretable educational intelligence, and affordable teaching tools. Considering the ongoing trajectory of AI advancement in education, achieving these objectives is within reach, bolstering the sustainable progress of education with the support of AI.

6. Conclusion

6.1. Implications

This study proposes a KT model based on MB. The improved MB-based KT model incorporates the MB structure learning algorithm to explore the causal mechanisms between features and predicted variable. It identifies the MB feature set that has a direct cause and direct effect on the predicted variable and uses it as a KT model parameter. Machine learning methods are employed for KT modeling while ensuring the interpretability of the model. Ablation experiments have revealed that models using a subset of causal features exhibit superior performance. In addition, when compared with other models, MB-KT demonstrates remarkable results on the Junyi dataset, achieving an AUC score of 0.868 and a score of 0.916 on the ASSISTments2009–2010 dataset, surpassing other models by a significant margin. Moreover, compared with DL models, MB-KT is inherently an interpretable model, signifying that it is explainable and computationally feasible. This characteristic reduces ethical risks when applying MB-KT in educational teaching contexts.

In previous research on KT models, it was widely believed that incorporating more learning features would lead to better model performance. However, our model has demonstrated that incorporating features with a causal relationship with the prediction variable yields better performance than simply increasing the number of features. This finding provides a novel perspective for future research on KT models.

6.2. Future work

Nevertheless, the model still has certain limitations, as discussed above, such as the challenge of assessing the strength of causal relationships between features and addressing the effects of confounding factors. In future studies, we plan to integrate domain expertise and prior assumptions into the construction of causal models. In addition, we will incorporate an understanding of confounding factors and latent variables into the process of structural learning. Furthermore, more feature data remain necessary, as we cannot guarantee that the currently used features are the most causally relevant to the target. Thus, further collection of additional features is required to obtain a more comprehensive causal network model. Fundamentally, our primary objective is to promote the development of sustainable education through the construction of more reliable AI models. We believe that these models can significantly improve personalized teaching effectiveness and contribute to the overall advancement of education. We plan to deploy our model on our self-developed adaptive learning system, which is already being used by several dozen elementary schools. Our aim is to track student progress in learning mathematical knowledge and proactively identify their mastery levels in specific areas. This would enable us to intervene and provide targeted teaching assistance accordingly. By integrating our model into the adaptive learning system, we anticipate an enhancement of the educational experience by tailoring instruction to meet individual student needs. Furthermore, there remains no doubt regarding the importance of ethical concerns related to the use of AI in education, and we aim to continue to develop models that are not only effective but also responsible and ethical in their application.

CRediT authorship contribution statement

Bo Jiang: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Yuang Wei:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Ting Zhang:** Conceptualization, Writing – original draft. **Wei Zhang:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available in the [Junyi] repository, [https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198]; [ASSISTments] repository, [https://sites.google.com/site/assistmentsdata/home/assistment2009-2010-data/skill-builder-data-2009-2010].

Funding

This research was funded by the National Natural Science Foundation of China grant number 61977058, 92270119, and the Natural Science Foundation of Shanghai, China grant number 23ZR1418500. All authors approved the final version of the manuscript.

Informed consent

Informed consent was obtained from all individual participants included in the study.

References

Abdelrahman, G., & Wang, Q. (2019). Knowledge tracing with sequential key-value memory networks. *ACM Computing Surveys*, 175—184.

Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, *55*(11), 1–37.

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: a novel Markov Blanket algorithm for optimal variable selection. In *AMIA annual symposium proceedings, Vol. 2003* (pp. 21–25). American Medical Informatics Association.

Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, *42*(1), 7–49.

Bo, J., Hengyuan, Z., & Wei, Y. (2023). How to evaluate the effectiveness of adaptive learning systems——Based on the causal structure analysis framework. *Modern Distance Education Research*, *35*(02), 95–101.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Cen, H. (2009). *Generalized learning factors analysis: Improving cognitive models with machine learning* (Ph.D. thesis), USA: Carnegie Mellon University, AAI3362263.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, *44*, 137–164.

Chang, K., Lee, J., Jun, C.-H., & Chung, H. (2018). Interleaved incremental association Markov blanket as a potential feature selection method for improving accuracy in near-infrared spectroscopic analysis. *Talanta*, *178*, 348–354.

Chaudhry, R., Singh, H., Dogga, P., & Saini, S. K. (2018). Modeling hint-taking behavior and knowledge state of students with multi-task learning. In *International educational data mining society* (pp. 21–31). ERIC.

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE international conference on data mining (ICDM)* (pp. 39–48). IEEE.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, *2*, 445–498.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM Sigkdd Explorations Newsletter*, *13*(2), 30–36.

Gao, T., & Ji, Q. (2016). Efficient Markov blanket discovery and its application. *IEEE Transactions on Cybernetics*, *47*(5), 1169–1179.

Gao, T., & Ji, Q. (2017). Efficient score-based Markov Blanket discovery. *International Journal of Approximate Reasoning*, *80*, 277–293.

Guyon, I., Aliferis, C., & Elisseeff, A. (2007). Causal feature selection. In *Computational methods of feature selection, Vol. 20071386* (pp. 63–85). CRC Press.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353–383.

Huang, Z., Liu, Q., Chen, Y., Wu, L., Xiao, K., Chen, E., Ma, H., & Hu, G. (2020). Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS)*, *38*(2), 1–33.

Huo, Y., Wong, D. F., Ni, L. M., Chao, L. S., & Zhang, J. (2020). Knowledge modeling via contextualized representations for LSTM-based personalized exercise recommendation. *Information Sciences*, *523*, 266–278.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, Article 100074.

Kleinbaum, D. G., & Klein, M. (2002). *Logistic regression (A self-learning text)*.

Koedinger, K. R., Brunskill, E., Baker, R. S. J., a Mclaughlin, E., & Stamper, J. (2013). New potentials for data-driven development and optimization. *AI Magazine*, *34*, 27–41.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324.

Kumar, N. A., Feng, W., Lee, J., McNichols, H., Ghosh, A., & Lan, A. (2023). A conceptual model for end-to-end causal discovery in knowledge tracing. arXiv preprint arXiv:2305.16165.

Li, Y., Chen, C.-Y., & Wasserman, W. W. (2016). Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, *23*(5), 322–336.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 1–45.

Li, Q., Yuan, X., Liu, S., Gao, L., Wei, T., Shen, X., & Sun, J. (2023). A genetic causal explainer for deep knowledge tracing. *IEEE Transactions on Evolutionary Computation*.

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019). Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 100–115.

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A survey of knowledge tracing. arXiv preprint arXiv:2105.15106.

Long, T., Liu, Y., Zhang, W., Xia, W., He, Z., Tang, R., & Yu, Y. (2022). Automatical graph-based knowledge tracing. In *Proceedings of the 15th international conference on educational data mining* (pp. 710–714). International Educational Data Mining Society.

Margaritis, D., & Thrun, S. (1999). Bayesian network induction via local neighborhoods. In *Advances in neural information processing systems, Vol. 12* (pp. 505–511).

Meek, C. (1997). *Graphical Models: Selecting causal and statistical models* (Ph.D. thesis), Ph.D. thesis, Carnegie Mellon University.

Minn, S., Vie, J.-J., Takeuchi, K., Kashima, H., & Zhu, F. (2022). Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 36* (pp. 12810–12818).

Mo, K., Wenjun, W., Xuan, Z., & Yanjun, P. (2018). Research on multi knowledge point knowledge tracking model and visualization. *e-Education Research*, *39*, 53–59.

Mongkhonvanit, K., Kanopka, K., & Lang, D. (2019). Deep knowledge tracing and engagement with moocs. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 340–342).

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *IEEE/WIC/ACM international conference on web intelligence* (pp. 156–163).

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

Niinimäki, T., & Parviainen, P. (2012). Local structure discovery in Bayesian networks. In *Proceedings of the twenty-eighth conference on uncertainty in artificial intelligence* UAI '12, (pp. 634—643). Arlington, Virginia, USA: AUAI Press.

Pandey, S., & Srivastava, J. (2020). RKT: relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1205–1214).

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *Proceedings of the 18th international conference on user modeling, adaptation, and personalization* UMAP '10, (pp. 255—266). Berlin, Heidelberg: Springer-Verlag.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis –a new alternative to knowledge tracing. In *Proceedings of the 2009 conference on artificial intelligence in education: Building learning systems that care: From knowledge representation to affective modelling* (pp. 531—538). NLD: IOS Press.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Proceedings of the 28th international conference on neural information processing systems - Volume 1* NIPS '15, (pp. 505—513). Cambridge, MA, USA: MIT Press.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys*, *28*(1), 71–72.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science*, *377*(6612), 1304–1310.

Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, *3*, 121–129.

Tong, S., Liu, Q., Huang, W., Hunag, Z., Chen, E., Liu, C., Ma, H., & Wang, S. (2020). Structure-based knowledge tracing: An influence propagation view. In *2020 IEEE international conference on data mining (ICDM)* (pp. 541–550). IEEE.

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 673–678).

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003). Algorithms for large scale Markov blanket discovery. In *FLAIRS conference, Vol. 2* (pp. 376–380). St. Augustine, FL.

Wang, Z., Feng, X., Tang, J., Huang, G. Y., & Liu, Z. (2019). Deep knowledge tracing with side information. In *Artificial intelligence in education: 20th international conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20* (pp. 303–308). Springer.

Xu, B., Huang, Z., Liu, J., Shen, S., Liu, Q., Chen, E., Wu, J., & Wang, S. (2023). Learning behavior-oriented knowledge tracing. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 2789–2800).

Yang, H., & Cheung, L. P. (2018). Implicit heterogeneous features embedding in deep knowledge tracing. *Cognitive Computation*, *10*, 3–14.

Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., Zhang, W., & Yu, Y. (2021). GIKT: a graph-based interaction model for knowledge tracing. In *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I* (pp. 299–315). Springer.

Yeung, C.-K., & Yeung, D.-Y. (2018). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the fifth annual ACM conference on learning at scale* (pp. 1–10).

Yu, W., Mengxia, Z., Shanghui, Y., Xuesong, L., & Aaoying, Z. (2023). Review and performance comparison of deep knowledge tracing models. *Journal of Software*, *34*(03), 1365–1395.

Zhang, L., Xiong, X., Zhao, S., Botelho, A., & Heffernan, N. T. (2017). Incorporating rich features into deep knowledge tracing. In *Proceedings of the fourth (2017) ACM conference on learning@ scale* (pp. 169–172).

Zhao, W., Xia, J., Jiang, X., & He, T. (2023). A novel framework for deep knowledge tracing via gating-controlled forgetting and learning mechanisms. *Information Processing & Management*, *60*(1), Article 103114.

Zhu, J., Ma, X., & Huang, C. (2023). Stable knowledge tracing using causal inference. *IEEE Transactions on Learning Technologies*, 1–11.