JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 10, pp. 2499-2507, Oct. 2023

Check for updates

다차원 특징 데이터에 강인한 XGBoost 기반 지식 추적 연구

조 현 진[1] · 박 동 욱[1] · 김 태 희[1] · 한 정 규[2*] · 김 현 석[2*]
[1]동아대학교 컴퓨터공학과 학사과정
[2*]동아대학교 컴퓨터공학과 교수

Study on XGBoost-based Knowledge Tracing Robust to Multidimensional Features

Hyun-Jin Cho[1] · Dong-Uk Park[1] · Tae-Hee Kim[1] · Jungkyu Han[2*] · Hyunseok Kim[2*]

[1]Bachelor Course, Department of Computer Engineering, Dong-A University, Busan 49315, Korea
[2*]Assistant Professor, Department of Computer Engineering, Dong-A University, Busan 49315, Korea

[요 약]

온라인 교육 플랫폼으로의 전환은 대량의 학습 활동 데이터를 제공하며 학생의 지식 추적(Knowledge Tracing) 모델을 생성할 수 있다. 기존 순환 신경망과 결합한 지식 추적 기법들로 학생의 교육 상태를 수치화하고 맞춤형 학습을 추천할 수 있다. 하지만 최근 게임 형태로 발전하고 있는 온라인 교육 플랫폼은 다차원 특징 데이터를 제공하고 있어, 기존 순환 신경망 기반 지식 추적 방식만으로는 성능 향상을 기대하기 어렵게 되었다. 따라서, 본 논문에서는 다차원 특징 데이터임에도 빠른 학습이 가능한 앙상블 학습 알고리즘 XGBoost(eXtreme Gradient Boosting) 기반 지식 추적 방법을 제안한다. 또한, 공개된 온라인 학습 데이터와 더불어 최근 Kaggle에서 완료된 게임 형태의 온라인 교육 플랫폼 학습 대회 데이터를 이용하여, 기존 LSTM 기반 딥러닝 지식 추적 방법보다 제안된 방법이 AUC 결과가 높음을 제시한다. 다차원 특징 데이터에도 강인하게 이용할 수 있는 XGBoost기반 지식 추적 방법은 온라인 교육 플랫폼의 성능 향상에 기여할 것이다.

[Abstract]

The change to online education can teach a student's knowledge tracing model through a large amount of collected learning activity data. With the advent of knowledge tracing techniques combined with recurrent neural networks, it is possible to quantify the educational status of students and recommend customized learning through them. However, in the case of multi-dimensional feature data collected from an online education platform that has recently developed into a game form, the recurrent neural network-based knowledge tracking method has a limitation in that its performance deteriorates. Therefore, in this paper, we propose knowledge tracing based on the ensemble learning algorithm XGBoost (eXtreme Gradient Boosting), using online learning data and recently completed learning data of an online education platform in the Kaggle competition. The results indicate that the XGBoost-based knowledge tracing method can be used robustly even for multidimensional feature data, and will contribute to changes in various online education platforms.

색인어 : 지식 추적, 깊은 지식 추적, 익스트림 그래디언트 부스팅, 순환 신경망, 장단기 메모리

Keyword : Knowledge Tracing, Deep Knowledge Tracing, XGB, Recurrent Neural Network, Long-Short-Term Memory

http://dx.doi.org/10.9728/dcs.2023.24.10.2499

Received 24 August 2023; **Revised** 19 September 2023
Accepted 22 September 2023

***Corresponding Author; Hyunseok Kim, Jungkyu Han**

Tel: +82-51-200-7928
E-mail: hertzkim@dau.ac.kr

Ⅰ. 서 론

COVID-19 팬데믹으로 학생들의 교육 플랫폼은 오프라인에서 온라인으로 급격하게 확장하고, 온라인 교육에서 수집된 대량의 학습 활동 데이터를 통해 학생의 지식 추적(Knowledge Tracing; KT) 모델을 생성할 수 있다. 특히, 생성된 지식 추적 모델을 통해 학생의 교육 상태를 수치화하여 아직 풀지 않은 문제에 대한 정답률을 예측하고, 문제 유형별 학습 성취도를 추정하며 부족한 유형에 대한 개인별 맞춤형 교육을 제공할 수 있다. 이러한 이유로 최근 온라인 교육 플랫폼의 핵심 기술로서 지식 추적이 연구되고 있다.

온라인 교육 플랫폼에서 수집된 학습 활동 데이터는 학생의 문제 유형별 교육 성취도를 순차적인 형태로 기록하고 있어, 순환 신경망(Recurrent Neural Network) 구조를 사용하는 Deep KT[1] 방법이 많이 사용되어 왔다. Deep KT는 학생의 순차적 문제 풀이 기록이 포함된 온라인 교육 데이터를 분석하여, 학생이 이전에 푼 문제와 그에 대한 정확도, 시간 등의 단편적인 정보만을 기반으로 학습자의 지식 추적을 진행한다. 하지만, 깊은 신경망의 구조적 한계로 인해, 수치형 특징과 범주형 특징이 혼합된 데이터인 경우 성능이 저하되는 문제점이 있다. 또한 실시간 피드백 제공이 가능한 온라인 교육 플랫폼의 요구사항에 부합하기에는 신경망 기반 지식 추적 모델이 학습이 오래 걸려 적합하지 않다.

최근 온라인 교육 플랫폼은 단순한 단답형 문제 풀이를 벗어나 학습자의 집중도를 향상하거나 재미있는 요소가 포함된 게임 형태로 발전하고 있다. 2023년 6월에 완료된 캐글(Kaggle.com) ‘Predict Student Performance from Game Play(PSPG)’[2] 대회는 게임 형태의 온라인 교육 플랫폼에서 수집된 다차원 데이터를 제공하고 있다. 본 대회에서는 총 23가지 유형의 문제가 주어지며, 실시간으로 다음 단계의 문제를 맞힐 확률을 제시하는 지식 추적 모델의 성능으로 대회 순위를 결정한다. 제공된 데이터는 학생 정보, 문제 유형, 문제 번호, 정답 유무와 같은 단편적 특징 정보뿐 아니라, 힌트를 사용한 횟수, 문제를 풀기 위해 시도된 횟수, 소요된 시간과 같은 부가적인 특징 정보도 포함하고 있다. 또한, 이전 단계에서 질문의 정답 여부를 예측하는 것에 따라 다음 단계에서 질문을 맞출 것인지를 실시간으로 추정하는 것이 지식 추적 모델의 중요한 평가 요소로 간주되고 있다. 따라서, 다차원 특징 데이터에서도 빠르게 지식 추적 모델을 학습하는 강인한 방법이 요구된다.

본 논문에서는 수치형 특징과 범주형 특징이 혼합된 데이터로부터 학습이 가능한 결정 트리 계열의 알고리즘을 사용한다. 특히, 적은 메모리 사용으로 빠르게 학습이 가능한 앙상블 결정 트리 계열의 최신 알고리즘인 XGBoost(eXtreme Gradient Boosting)를 지식 추적 모델 학습에 이용한다. 이를 통해 복잡한 문제 유형과 다양한 관점으로 취득된 수치형 데이터, 게임 요소와 같은 부가적인 범주형 데이터가 혼합된 다차원 특징 데이터에도 강인하게 지식 추적 모델을 학습할 수 있는 XGBoost 기반 지식 추적(XGBoost-based Knowledge Tracing; XKT)를 제안한다. 비교 검증을 위해 순환 신경망의 단점을 보완한 장기 단기 메모리(Long-Short Term Memory)기반 지식 추적 모델 LSTM-KT (DKT)[3]를 이용한다. 또한, 공개된 온라인 학습 데이터 3종과 Kaggle PSPG 데이터를 사용하여 학습자의 문제 정답률을 Area Under the Curve(AUC) 측정을 통해 다차원 특징에 대한 XKT의 강건성을 비교 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 지식 추적, 딥러닝을 이용한 지식 추적 연구 현황, 그리고 온라인 학습 지식 추정에 필요한 XGBoost의 특징을 소개한다. 3장에서는 다차원 데이터에 강인한 XGBoost기반 지식 추적 기술을 제안하고, 4장에서는 DKT와 XKT의 비교 분석을 한다. 끝으로 5장에서는 결론 및 향후 연구에 관해서 기술한다.

Ⅱ. 관련 연구

2-1 Knowledge Tracing (KT)

Knowledge Tracing은 교육 분야에서 학습자의 학습 과정을 추적하고 모니터링하는 기술이다. 이 기술은 주어진 시험 또는 문제를 학습자가 얼마나 정확하게 푸는지를 평가하고, 학습자의 지식수준을 추정하는 데 사용된다. 이를 통해 개인화된 학습 경로를 제시하고 학습자의 학습 행동을 정량적으로 분석하는 데에 활용되는 등, 최근까지 활발하게 연구되고 있다[4]-[6]. 학습자의 지식수준을 효과적으로 추적하기 위해서는 학습자가 이전에 푼 문제와 그에 대한 정확도, 시간 등의 정보를 기록하고 분석해야 한다. 하지만 기존의 KT는 학습자의 이전 상태 정보를 반영하지 않고, 오직 현재 문제 결과만을 기반으로 학습자의 이해도를 추적하고 있어 순차적 특성을 갖는 학습 데이터를 처리하는 데 어려움이 있다[4]. 이러한 한계를 극복하기 위해 KT 기술에 딥러닝 기술을 도입한 Deep KT가 등장하여 활발하게 응용되고 있다[1].

2-2 Deep Knowledge Tracing (Deep KT)

Deep KT는 교육 분야에서 학습자의 학습 과정을 딥러닝을 활용하여 추적하고 모니터링하는 기술이다[1]. Knowledge Tracing을 발전시킨 Deep KT는 순환 신경망(RNN)이나 장기 단기 메모리(LSTM)와 같은 딥러닝 모델을 활용하여 학습자의 학습 행동을 시계열 데이터로 처리하고 이해도를 추정한다. Deep KT는 학습자가 이전에 푼 문제와 그에 대한 정확도, 시간 등의 시계열 정보를 기반으로 학습자의 지식수준을 예측하고 모델링한다. 이를 통해 학습자의 개별 학습 패턴과 지식 변화를 더 정확하게 추적하고 개인화된 학습 경로를 제시하는 여러 연구가 진행되었다[7]-[9]. 하지만 Deep KT도 딥러닝

모델을 사용하는 만큼 몇 가지 한계점이 존재한다. 우선, 대량의 데이터와 긴 학습 시간이 필요하다. 특히 시계열 데이터를 처리하는 과정에서 학습의 복잡성이 더 높아질 수 있다. 또한, 딥러닝 모델의 특성상 많은 파라미터를 가지고 있는데, 특히 데이터가 부족한 경우에는 과적합(overfitting)이 발생할 가능성이 있다[3].

2-3 Recurrent Neural Network (RNN)

Deep KT에서 사용되었던 RNN (Recurrent Neural Network)은 딥러닝의 한 종류로, 시퀀스 데이터를 처리하는 데 특화된 네트워크 구조이다. RNN은 일련의 입력 데이터를 순서대로 처리하면서 각 시점(time step)에서의 hidden state를 출력으로 계산하며 이전 시점의 정보를 현재 시점에 전달한다. 이러한 순환구조를 통해 Deep KT에서는 학습자의 이력을 시계열로 기록하고 특징을 포착하여 학습자의 이전 학습 행동을 추적하는 데 사용한다. 그러나 RNN의 구조는 간단하고 유연하지만, 긴 시퀀스 데이터에서 발생하는 기울기 소실(Vanishing Gradient) 또는 폭발적인 기울기 문제로 인해 학습이 어려워질 수 있다. 특히, RNN은 시퀀스가 길어질수록 이전 시점의 정보를 현재 시점으로 전달하기 어려워진다. 이러한 한계로 긴 시퀀스 데이터를 처리하는 문제점을 보완하기 위해 LSTM이 등장하였다[10].

2-4 Long Short Term Memory (LSTM)

LSTM(Long Short-Term Memory)은 RNN의 한계를 보완하여 시퀀스 데이터의 장기적인 의존성을 더욱 잘 모델링할 수 있도록 설계된 특별한 종류의 RNN이다. LSTM은 내부적으로 기억셀(memory cell)이라는 구조를 사용하여 이전 시점에서의 정보를 기억하고 이를 현재 시점으로 전달함으로서 장기적인 의존성을 관리한다. 기억셀은 세 개의 중요한 게이트(forget gate, input gate, output gate)로 구성되며, 게이트 메커니즘을 활용한 LSTM은 기울기 소실 문제를 해결하고 긴 시퀀스 데이터에서도 효과적으로 장기적인 의존성을 학습할 수 있는 특징 때문에 Deep KT에서 학습자의 학습 패턴 및 지식 추적에 주로 사용되고 있다[10].

2-5 EXtreme Gradient Boosting (XGBoost)

XGBoost(eXtreme Gradient Boosting)는 Tianqi Chen [11]에 의해서 2016년에 8월 소개되었으며 2015년에는 Kaggle에 게시된 29개의 챌린지 우승 솔루션 중 17개의 솔루션이 XGBoost를 사용한 바가 있다. XGBoost는 Gradient boosting 알고리즘으로 의사결정 나무(Decision Tree) 기반의 앙상블 모델이다. Gradient boosting 알고리즘은 약한 분류기를 통해 학습한 후, 나타나는 오차를 다음 약한 분류기에서 학습하며 오차를 줄여나가 강한 예측 모델을 만드는 Boosting 기법이다. 손실 함수 미분값의 크기가 점차 줄어드는 방향으로 가중치를 업데이트하며 손실 함수의 최솟값을 찾아가는 경사하강법을 접목하여 실제값-예측값을 의미하는 잔차(Residual)를 줄여가는 방식이다[12]. XGBoost는 Gradient Boosting 알고리즘을 기반으로 하고 있어 높은 예측 성능을 보여주며, 병렬 실행을 통해 학습 속도 및 분류 속도에서 뛰어난 성능을 보여주고 있다. 또한, L1 정규화 및 L2 정규화를 제공하여 데이터 양이 증가할 때 발생할 수 있는 과적합을 방지하고 일반화 성능을 향상할 수 있다[13],[14].

Ⅲ. XGBoost 기반 지식 추적 (XKT)

3-1 온라인 학습 환경에서 지식 추적 문제 정의

기존의 오프라인 학습에서는 데이터의 크기가 크지 않고 데이터의 종류도 정답 여부와 학습 시간에 대한 단순한 정보만 주어졌다. 온라인 학습으로 넘어오면서 LSTM-based Deep Knowledge Tracing (DKT) 모델은 빅데이터에 대한 느린 연산속도와 다양한 특징을 고정된 벡터에 압축하는 과정에서 정보 손실이 증가하는 문제가 발생한다[15]. 또한, 온라인 학습에서는 실시간으로 계속 주어지는 다양한 학습 데이터를 빠르게 처리할 필요가 있다. 본 연구에서는 이러한 문제를 해결하기 위해 그림 1과 같은 XGBoost-based Knowledge Tracing (XKT) 지식 추적 방법을 제안한다. DKT 모델은 인코딩과 임베딩을 통해 데이터의 전처리를 해야하지만, 결정 트리(Decision Tree) 기반 XGBoost 모델은 다양한 특징들의 전처리를 Feature Engineering만을 통해 정렬과 군집화를 수행할 수 있다. 가공된 데이터로 XGBoost 모델을 학습시키는 중, 실시간으로 데이터가 추가되어도 앙상블 방식의 XGBoost 모델은 이전의 결과값에 새로운 데이터를 계속 학습시켜 주는 장점이 있다.

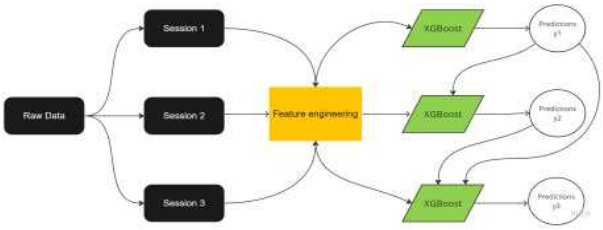


그림 1. XGBoost 기반 지식 추적 모델 구조
Fig. 1. XGBoost based knowledge tracing model

표 1과 같이, 기존 DKT 학습에서 사용된 데이터(학습자 ID, 문제 번호 및 유형, 정답 유무)와는 다르게 게임 기반 온라인 학습 시스템은 많은 양의 데이터를 수집할 수 있다. 게임 형태의 온라인 학습 시스템의 경우 학습자들이 게임 플레이 방식의 다양한 문제들을 푸는 과정에서 그에 대한 정확도

와 시간 등의 부가 정보를 수집할 수 있다. 이때, 학습자의 이해도를 추적하고 개인화된 학습 경로를 제공하기 위해 DKT를 사용할 수 있지만, 다차원 특징에 대한 학습률이 떨어진다는 단점이 있다. 따라서 본 연구에서는 DKT의 단점을 보완하고 모델링의 성능을 향상하기 위해 XGB(eXtreme Gradient Boosting)를 적용한다. XGB는 2-5장에서 설명한 바와 같이 Gradient Boosting 알고리즘으로 높은 성능과 빠른 학습 속도를 제공하며, Decision Tree를 기반으로 하기 때문에 다차원 특징들을 효과적으로 처리하여 학습자의 지식 추적에 활용될 수 있다. XGB를 적용한 지식 추적 모델(XKT)을 통해 학습자 별 개개인에 특화된 학습 방향성을 정확하게 제시할 수 있게 된다.

표 1. 캐글 데이터셋
Table 1. Kaggle dataset

Columns	Description
session_id	the ID of the session the event took place in
index	the index of the event for the session
elapsed_time	how much time has passed (in milliseconds) between the start of the session and when the event was recorded
event_name	the name of the event type
name	the event name (e.g. identifies whether a notebook_click is is opening or closing the notebook)
level	what level of the game the event occurred in (0 to 22)
page	the page number of the event (only for notebook-related events)
room_coor_x	the coordinates of the click in reference to the in-game room (only for click events)
room_coor_y	the coordinates of the click in reference to the in-game room (only for click events)
screen_coor_x	the coordinates of the click in reference to the player's screen (only for click events)
screen_coor_y	the coordinates of the click in reference to the player's screen (only for click events)
hover_duration	how long (in milliseconds) the hover happened for (only for hover events)
text	the text the player sees during this event
fqid	the fully qualified ID of the event
room_fqid	the fully qualified ID of the room the event took place in
text_fqid	the fully qualified ID of the
fullscreen	whether the player is in fullscreen mode
hq	whether the game is in high-quality
music	whether the game music is on or off
level_group	which group of levels - and group of questions - this row belongs to (0-4, 5-12, 13-22)

3-2 온라인 학습 환경에서 지식 추적 데이터

DKT와 XKT의 데이터 종류 수에 따른 학습 지표로는 AUC를 사용하였고 학습 데이터로는 표 2와 같은 세 가지 데이터를 사용하였다. 각각의 데이터는 서로 다른 이름의 열들을 가지고 있으며, 각 열의 특성에 따라 열의 이름을 통일시켜 주고 학습을 진행하였다.

표 2. 학습 데이터셋 통계
Table 2. Training dataset statistics

Dataset	Student	Interactions
ASSISTment 2009	4,217	525,534
Algebra 2005	575	813,661
Bridge to Algebra 2006	1,146	3,656,871

표 3. 학습 데이터셋 항목별 특성
Table 3. Training dataset features

Data	Description
user_id	Learner's ID
problem_id	Unique ID of the issue
skill_name	Type of problem
correct	Correct answer
hints	Hint usage count

1) 데이터 설명

본 연구에서는 DKT와 XKT 학습에 사용되었던 데이터 3종에 특징을 추가하여 실험에 사용하였다. 표 2는 사용된 데이터와 그에 해당하는 설명을 나타낸다.

(1) user_id: 학습자의 고유 식별자로, 각 학습자마다 고유한 ID가 부여되며, 학습자를 식별하는 데 사용된다.
(2) problem_id: 학습할 문제의 고유 식별자로, 시스템에서 제공되는 각 문제마다 고유한 ID가 부여되며, 학습자들은 이 문제들을 풀게 된다.
(3) skill_name: 학습할 문제(problem_id)의 유형을 나타낸다. 문제의 난이도, 주제, 혹은 카테고리 등이 이 정보에 해당될 수 있다.
(4) correct: 학습자가 해당 문제를 정답, 오답 여부를 나타낸다. 이진(binary) 형태로 표현되며, 1은 정답, 0은 오답을 의미한다.
(5) hints: 학습자가 해당 문제를 풀기 위해 사용한 힌트의 횟수를 나타낸다. 학습자들은 문제 해결을 위해 여러 번의 힌트 사용을 시도할 수 있다.

이와 같이 다양한 특징들을 포함한 데이터를 사용하여 DKT와 XKT 모델을 학습하고 평가하여 게임 기반 온라인 학습 시스템에서의 학습자 모델링을 비교하고 개선하는 것이 본 연구의 목표이다. 추가된 특징들을 통해 학습자의 이해도 예측과 개인화된 학습 경로 제공에 더 나은 성능을 기대할 수 있다.

2) 데이터 전처리

각 각 dataset을 섞어서 train set과 test set으로 나누고 문제를 단계별로 학습시키기 위해 세션 개수만큼 일정한 비율로 분할한다. 이후 각 모델에 맞게 데이터를 전처리 해준다. 이때 DKT는 시퀀스 데이터를 처리하는 모델이기 때문에 위의 데이터를 전처리를 통해 시퀀스로 만들어준다. XKT는 전처리를 통해 데이터를 특성으로 변환하고, 다음 문제의 정답 여부를 예측 할 수 있도록 데이터를 구성한다.

3-3 모델 구현

1) DKT (LSTM-based Deep Knowledge Tracing)

DKT 모델은 크게 Embedding Layer와 LSTM(Long Short-Term Memory) Layer 두 부분으로 구성된다. Embedding Layer를 통해 문제, 힌트, 정답 여부 등의 입력 데이터를 하나의 임베딩 벡터로 변환시켜 과거 정보를 현재 상태에 반영할 수 있다. 이후 이 결과를 LSTM Layer를 통과시키는데, 이는 앞서 2-4장에서 설명했던 것처럼 RNN(Recurrent Neural Network)의 한 종류로, 과거 정보를 기억하고 현재의 입력과 결합하여 다음 상태를 예측하는데 사용된다. LSTM Layer에 Embedding 레이어의 출력을 입력으로 받아 LSTM의 hidden state를 계산한다. 여기서 말하는 hidden state란 학습 데이터 시퀀스에서 현재까지의 정보를 인코딩한 벡터를 말한다. 이후 LSTM 레이어의 출력인 hidden state를 Linear Layer인 Output Layer에 통과시켜 학생의 지식 수준을 예측하는데 사용한다. 이 출력값들은 학생의 각 문제별 응답을 예측한 값이며 학습 과정에서 과적합을 방지하기 위해 일부 뉴런을 무작위로 끼워준다.

2) XKT (XGBoost-based Knowledge Tracing)

XKT모델은 기본적으로 앙상블 기법을 사용하기 때문에 별도의 LSTM과 같은 시퀀스 처리를 하지 않는다. 먼저 전처리한 데이터로 효율적인 모델 학습을 위해 D-matrix 형태로 변환시킨다. 이후 XGBClassifier를 사용해 분류 모델을 생성한다. 이때 objective, max_depth, learning_rate, eval_metric 등의 하이퍼 파라미터를 같이 설정한다. 이렇게 모델을 생성시킨 후 학습과 평가를 진행한다. 학습이 진행됨에 따라 새로운 데이터가 들어오면 과거의 정보와 새로 들어온 정보를 학습하던 XGBoost 모델에 넣어 학습과 평가를 반복하며 진행한다.

Ⅳ. 실험

4-1 실험 설계

본 연구에서는 Deep Learning 기반인 LSTM을 사용한 DKT[3]와 Decision Tree 기반인 XGBoost[11]를 기반으로 본 논문에서 제안된 XKT 두 가지 모델에 대해 특징 개수에 따른 성능 비교 실험을 수행하였다. 이러한 선택은 Deep Learning과 Decision Tree 기반 모델 간의 성능 비교를 통해 각 모델의 장점과 한계를 보기 위함이다. 해당 실험에는 "ASSISTment 2009,"[16] "Algebra 2005,"[17] 그리고 "Bridge to Algebra 2006"[18] 세 가지 데이터 셋을 활용하였다. "ASSISTment 2009" 데이터 셋은 학습자들이 온라인 수학 학습 플랫폼에서 진행한 활동을 기록한 것으로, 다양한 문제 유형과 난이도를 포함하고 있다 "Algebra 2005"와 "Bridge to Algebra 2006" 데이터셋은 대부분 대중교육 학습과정에서 활용되는 Algebra 과목과 관련된 데이터로, 학습자들의 수학적 이해도를 다양한 상황에서 반영하고 있다. 각 데이터셋은 3-2장에서 설명한 방식대로 Train과 Test로 분할한 후, 각 모델에 맞게 데이터를 전처리하여 학습시켜 모델 간의 공정한 비교와 성능 평가를 할 수 있도록 하였다. DKT와 XKT의 비교 실험은 표 4, 표 5와 같이 파라미터들을 설정해주고 진행하였다.

본 연구에서는 초기 데이터 사이즈를 256으로 설정하였으며, 이 결정은 학습 프로세스의 속도와 성능 간의 균형을 찾기 위한 실험적인 접근에 기반한다. 작은 데이터 사이즈로 시작할 경우 모델 학습 속도가 향상되지만, 성능은 저하될 우려가 있었다. 반면, 큰 데이터 사이즈를 사용하면 모델의 성능이 개선될 것으로 예상되었지만 학습 속도가 저하될 우려가 있었다. 따라서 다양한 데이터 사이즈에서 실험을 진행하고 그 결과를 통해 최적의 데이터 사이즈를 256으로 선정하였다.

표 4. DKT 모델의 파라미터
Table 4. DKT model parameters

Name	Value	Description
Leaning_rate	0.001	Learning rate
Batch_size	256	Number of samples in each batches
embedding_size	256	Number of embeddings
Hidden_size	256	Number of hidden layers
Epochs	30	Training times

표 5. XKT 모델의 파라미터
Table 5. XKT model parameters

Name	Value	Description
Leaning_rate	0.02	Learning rate
max_depth	7	Depth of decision tree
num_boost_round	100	Number of boosting iteration
Epochs	30	Training times

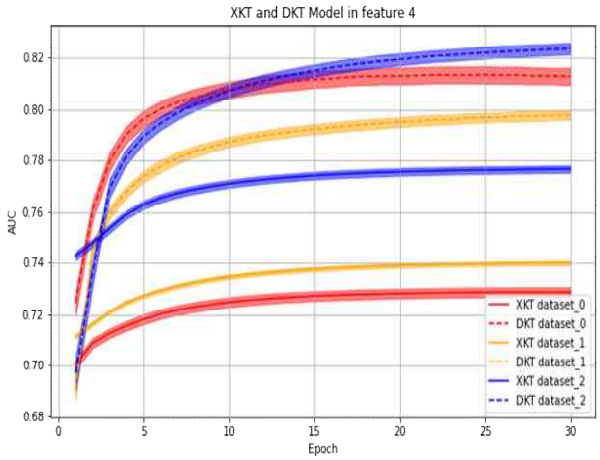


그림 2. 특징 4개인 경우 DKT와 XKT의 AUC 점수 비교
Fig. 2. DKT vs. XKT of AUC in 4 features

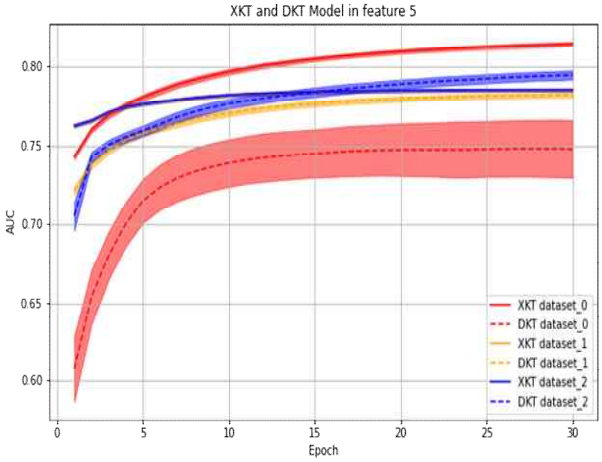


그림 3. 특징 5개인 경우 DKT와 XKT의 AUC 점수 비교
Fig. 3. DKT vs. XKT of AUC in 5 features

4-2 실험결과

실험 결과를 효과적으로 시각화하기 위해, 각 특징 개수에 대한 두 모델의 AUC 변화를 Epoch 별로 list에 저장하는 과정을 50번 반복하여 얻은 실험 결과로 평균과 분산을 계산하여 그래프로 작성하였다. 그림 2에서는 "user_id", "problem_id", "skill_name", "correct" 4개의 특징에 따른 DKT와 XKT의 성능을 비교하였다. 이 실험에서는 모든 데이터셋에서 DKT가 XKT에 비해 평균 약 8.5% 우수한 성능을 나타내는것을 볼 수 있다. 이는 DKT의 Deep Learning 기반 구조로 인해 학생 개별의 학습 패턴을 더 정확하게 학습하고 예측할 수 있는 능력 때문이다[3]. 그림 3과 같이, 데이터 셋의 크기가 증가함에 따라 DKT와 XKT의 성적 차이가 줄어드는 경향을 관찰할 수 있다. 이는 XKT의 앙상블 모델 특성으로 설명할 수 있다[19]. 큰 데이터셋에서는 XKT가 특징간의 상호작용을 더 잘 파악하며 예측하기 때문에, XKT의 성능이 향상되고 DKT와의 성적 차이가 축소된 것으로 해석할 수 있다. 그림 3의 결과는 이미 앞서 DKT와 XKT의 특징에 대해 설명한 내용과 일치한다. "hints"라는 새로운 특징를 추가하여 두 모델의 성능을 비교한 결과, XKT가 DKT에 비해 평균 약 3.5% 우수한 성적을 냈다는 것을 확인할 수 있다. 이를 통해 우리가 이전에 언급한 DKT와 XKT의 특징이 실제 결과에 어떤 영향을 미쳤는지 명확하게 확인할 수 있다. 또한, 실험 결과를 종합적으로 분석한 결과, 다차원 특징를을 사용하는 경우에 XKT가 DKT보다 우수한 성능을 보여준다. 3개의 데이터셋 중 2개의 경우에 XKT가 더 우수한 성적을 보여주었으며, 세번째 데이터셋에서도 15 epoch까지는 XKT가 더 좋은 성능을 보여준다. 특히 DKT의 경우 특징 5개일 때 성능이 떨어진 모습을 보였으며 Epoch의 크기가 커질수록 DKT의 성능이 저하되는 반면에 XKT의 성능이 향상되는 모습을 실험을 통해 확인하였다. 이는 이전 연구에서 언급된 seq2seq 모델의 한계에 직면했다는 점을 의미한다[20]. 반면에 XKT는 특징 5개에서와 같이 다차원 특징 데이터일수록 성능이 향상된 모습을 보인다. 이는 Decision Tree 기반인 XGB 모델의 장점으로 다양한 정보를 포착하며 비선형 관계를 학습하고 특징 간의 상호작용을 높일 수 있다는 점을 의미한다. 본 실험 결과와 같이 XKT가 다차원 특징 데이터에서 더 강인함을 확인할 수 있다.

4-3 검증

본 연구에서는 "특징 개수가 많아질수록 XKT의 성능이 향상된다"는 점을 보다 확실하게 입증하고 실험 결과의 일관성과 신뢰성을 강화하기 위해 Kaggle의 "Predict Student Performance from Game Play"이라는 Competition에 참가하였다. 이 대회는 “Education Resource and Information Center (ERIC) Archive”에서 수집한 실제 게임인 “Jowilder”의 플레이 데이터를 활용하여 학습자의 성취도를 예측하는 것을 목표로 하고 있다. 이 게임 데이터를 분석하여 학습자의 게임 플레이 패턴과 학습 성과를 이해하고, 이를 기반으로 성취도를 예측하는 모델을 개발하는 것이 주된 목표이다. 본 대회의 데이터는 실험에 사용한 데이터에 존재하는 학습자의 정답 여부와 더불어 다른 여러 특징을 가진 다차원의 데이터라는 점에서 검증에 적합하다고 판단하였고, 실험 결과의 신뢰성을 검증하기 위해 XGBoost 모델을 Baseline으로 설정하였다. 또한, CatBoost, LightGBM [21],[22]과 RNN과 같은 모델이 본 대회에서 널리 사용되었음을 확인하였다. 표 1에서 확인할 수 있듯이 게임 기반의 온라인 학습 시스템은 많은 데이터가 존재하기 때문에, Decision Tree 계열의 Boosting 알고리즘이 많이 채택되었다고 파악할 수 있다. 대회에서 사용한 XKT 모델의 파라미터 값은 표 6과 같다.

표 6. Kaggle PSPG 대회에 사용된 XKT 모델의 파라미터
Table 6. XKT model parameters in Kaggle PSPG Competition

Name	Value
Leaning_rate	0.02
tree_metohd	hist
n_estimators	249
alpha	8
max_depth	4
colsample_bytree	0.5

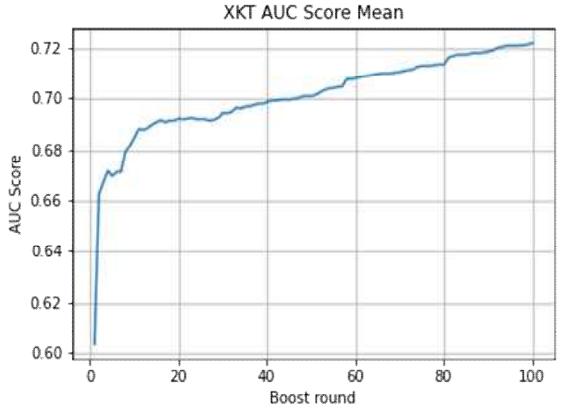


그림 4. 캐글 대회에서의 XKT 모델 AUC 점수 평균
Fig. 4. Mean AUC of XKT model in Kaggle PSPG

대회에서의 데이터를 가지고 본 연구에서 제안하는 알고리즘을 평가한 결과는 다음과 같다. 평가는 100번의 Boost Round를 학습시키는 과정을 50번 반복하여 얻은 결과로 평균을 계산하여 그래프로 작성하였다. 그림 4와 같이 AUC 점수는 꾸준히 상승하였고 그림 5와 같이 Error는 줄어드는 모습을 보였다. 이를 통해 다차원 특징 데이터에서도 XKT를 통해 강인한 지식 모델 학습이 가능함을 확인할 수 있다.

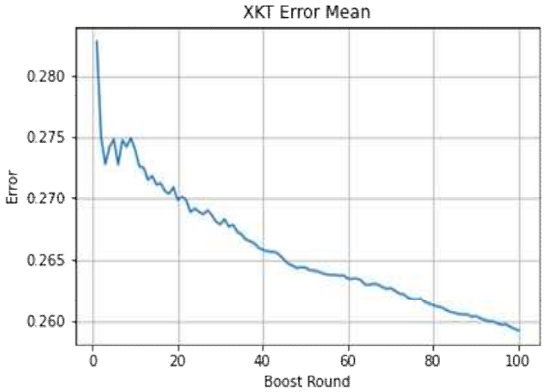


그림 5. 캐글 대회에서의 XKT 모델 Error 평균
Fig. 5. Mean error of XKT model in Kaggle PSPG

4-4 한계점

본 연구는 DKT와 XKT를 활용하여 학습자의 이해도 예측 성능을 비교하고 평가하는 것에 초점을 맞추었지만 몇 가지 한계점이 존재한다. 먼저, 사용한 데이터셋은 제한된 범위 내에서 수집되었고, 데이터의 양과 다양성이 충분하지 않을 수 있다. 또한, 데이터 전처리 과정에서 어떤 정보를 무시하거나 변형하므로서 정확한 분석을 제한할 수 있다. 두 번째로, 실험에서 사용한 특징을 선택하고 조합하는 과정에서 주관적 판단이나 경험이 작용할 수 있다. 따라서 더 나은 Feature Engineering 방법을 고려하거나 다양한 특징 조합을 시도한다면 더 정확한 결과를 얻을 수 있을 것이다. 세 번째로, 학습자의 개인적인 특성이나 환경 요인 등은 고려하지 않았다. 이러한 추가적인 변수들이 학습자의 성능에 영향을 줄 수 있으며, 이를 고려하는 모델링이 더 정확한 예측을 가능하게 할 수 있을 것이다. 향상된 예측 모델을 개발하고자 한다면 이러한 한계점들을 보완하여 다양한 요인을 종합적으로 고려하는 연구를 진행하는 것이 필요하다.

V. 결 론

본 연구에서는 온라인 교육 플랫폼에서 생성된 다차원 학습 데이터를 활용하여, 새로운 지식 추적 기술을 제안하고 성능을 비교 평가하였다. 실험에서는 기존의 순환 신경망 기반 지식 추적(DKT)과 새로운 접근인 XGBoost 기반 지식 추적(XKT)을 비교하였다. 실험 결과 DKT와 비교하여 XKT가 다차원 특징 데이터셋에서 성능이 우수하였다. 특히, XKT는 다양한 특징 데이터를 활용하면서도 빠른 학습 속도를 유지하는 장점을 보였다.

본 연구 결과를 통해 온라인 교육플랫폼을 위한 혁신적인 지식 추적 기술로서 XKT가 학습자의 학습 상태를 정확히 예측하고 맞춤형 교육을 제공할 수 있음을 확인하였다.다차원 특징 데이터에서도 강인한 지식 추적이 가능한 XKT를 제안하며 온라인 교육 플랫폼의 성능 향상과 학습 경험 개선을 위한 중요한 도구로 활용될 수 있을 것으로 기대한다.

감사의 글

이 논문은 동아대학교 교내연구비 지원에 의하여 연구되었음.

참고문헌

[1] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, “Deep Knowledge Tracing,” arXiv:1506.05908, June 2015. https://doi.org/10.48550/arXi

v.1506.05908 .

[2] Kaggle. Predict Student Performance from Game Play [Internet]. Available: https://www.kaggle.com/competitions/predict-student-performance-from-game-play.

[3] S. Minn, Y. Yu, M. C. Desmarais, F. Zhu, and J.-J. Vie, "Deep Knowledge Tracing and Dynamic Student Classification for Knowledge Tracing," *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 1182-1187, November 2018. https://doi.org/10.1109/ICDM.2018.00156

[4] A. T. Corbett and J. R. Anderson, "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, Vol. 4, pp. 253-278, December 1994. https://doi.org/10.1007/BF01099821

[5] S. Pandey and G. Karypis, "A Self-Attentive Model for Knowledge Tracing," arXiv:1907.06837, July 2019. https://doi.org/10.48550/arXiv.1907.06837

[6] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge Tracing: A Survey," *ACM Computing Surveys*, Vol. 55, No. 11, pp. 1-37, February 2023, https://doi.org/10.1145/3569576

[7] L. Lyu, Z. Wang, H. Yun, Z. Yang, and Y. Li, "Deep Knowledge Tracing Based on Spatial and Temporal Representation Learning for Learning Performance Prediction," *Applied Sciences*, Vol. 12, No. 14, 7188, 2022, https://doi.org/10.3390/app12147188

[8] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent Advances in Recurrent Neural Networks," arXiv preprint arXiv:1801.01078, February 2018, https://doi.org/10.48550/arXiv.1801.01078

[9] Y. Lim, J. Moon, E. Choi, and J. Lee, "Quantized Correctness Rate Embedding Method for Performance Enhancement of Knowledge Tracing Models," *Journal of the Korean Institute of Information Scientists and Engineers*, Vol. 50, No. 4, pp. 329-336, April 2023, https://doi.org/10.5626/JOK.2023.50.4.329

[10] S. Yang, "Deep Knowledge Tracing with Convolutions," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*, pp. 1561-1569, October 2021. https://doi.org/10.48550/arXiv.2008.01169

[11] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp. 785-794, August 2016. https://doi.org/10.1145/2939672.2939785.

[12] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, Vol. 29, No. 5, pp. 1189-1232, October 2001. https://doi.org/10.1214/aos/1013203451

[13] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267-288, January 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[14] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 42, No. 1, Special 40th Anniversary Issue, pp. 80-86, February 2000. https://doi.org/10.2307/1271436

[15] J. Zhang, X. Shi, I. King, and D. Y. Yeung, "Dynamic Key-Value Memory Networks for Knowledge Tracing," in *Proceedings of* WWW '17: Proceedings of the 26th International Conference on World Wide Web, pp. 765-774, April 2017. https://doi.org/10.1145/3038912.3052580.

[16] M. Feng, N. T. Heffernan, and K. R. Koedinger, Addressing the Assessment Challenge in an Intelligent Tutoring System that Tutors as It Assesses, *The Journal of User Modeling and User-Adapted Interaction*, Vol. 19, pp. 243-266, 2009.

[17] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger, *Algebra I 2005-2006*, Development Data Set from KDD Cup 2010 Educational Data Mining Challenge, 2010. Available: http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

[18] J. Stamper, A. Niculescu-Mizil, S. Ritter, G. J. Gordon, and K. R. Koedinger, *Bridge to Algebra 2006-2007*, Development Data Set from KDD Cup 2010 Educational Data Mining Challenge, 2010. Available: http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

[19] A. Althnian, A. Alhanoof, D. AlSaeed, H. Al-Baity, A. Samha, A. Bin Dris, ... and H. Kurdi, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, Vol. 11, No. 2, 796, January 2021. https://doi.org/10.3390/app11020796.

[20] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A Critical Review of Recurrent Neural Networks for Sequence Learning," arXiv preprint arXiv:1506.00019, May 2015. https://doi.org/10.48550/arXiv.1506.00019.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS '18)*, Vol. 32, No. 2, pp.

6639-6649, December 2018. https://doi.org/10.48550/arXiv.1706.09516.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, ... and T. Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4874-4884, 2017.

조현진(Hyun-Jin Cho)

2019년～현　재: 동아대학교 컴퓨터공학과 학사과정
※관심분야 : 강화학습, 인공지능, 머신러닝

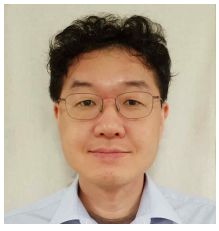
박동욱(Dong-Uk Park)

2019년～현　재: 동아대학교 컴퓨터공학과 학사과정
※관심분야 : Data Science, Data Analysis, Recommendation System

김태희(Tae-Hee Kim)

2019년～현　재: 동아대학교 컴퓨터공학과 학사과정
※관심분야 : Software

한정규(Jungkyu Han)



2005년 : 서울대학교 컴퓨터공학부 (공학학사)
2007년 : 서울대학교 전기컴퓨터공학부 (공학석사)
2018년 : Waseda University, Computer Science and Communications Engineering (공학박사)
2007년～2014년: NTT Software Innovation Center 근무
2018년～2020년: NAVER AiRS 근무
2020년～현　재: 동아대학교 컴퓨터 AI공학부 조교수
※관심분야 : 추천 시스템, 정보 검색, 데이터 마이닝

김현석(Hyunseok Kim)

2001년 : 동아대학교 전자공학과 (공학학사)
2005년 : 한국과학기술원 대학원 (공학석사)
2014년 : 한국과학기술원 대학원 (공학박사)
2001년～2003년 삼성전자 연구원
2005년～2009년: LG전자 선임연구원
2011년～2022년: 한국전자통신연구원 책임연구원
2022년～현　재: 동아대학교 컴퓨터공학과 조교수
※관심분야 : 강화학습, 로봇, 인공지능, 군집지능