설명 가능한 AI 학습 지원 시스템 개발

김성훈† · 김우진† · 장연주† · 김현철††

요 약

대부분의 온라인 교육 시스템은 많은 학습자에게 서비스를 제공하는 데 효율적이나, 개별 학습자를 진단하고 필요한 처방을 내리는 맞춤형 학습을 하기에는 부족한 실정이다. 이에 본 논문에서는 교수자를 위한 설명 가능한 AI 학습 지원 시스템을 제안한다. 시스템은 학습자의 인지적, 환경적 영역의 데이터를 이용해 개별 학습자를 진단한다. 제안한 시스템은 인공지능 모델인 DKT와 XGBoost를 이용해 학습자의 지식 상태를 모델링하고, 그 결과를 설명 가능한 인공지능 기법인 LRP와 SHAP으로 분석해 학습자의 지식 상태를 해석 가능한 형태로 교수자에게 제공한다. 교수자는 이 정보를 통해 학습자의 교과 이해 정도와 학습에 영향을 미치는 환경적 요소를 파악하여, 맞춤형 학습을 제공하는 데 활용할 수 있다. 이상의 연구를 통해 인공지능 모델과 XAI 분야의 기법을 교육 도메인에 적용하여 맞춤형 학습이 가능한 AI 학습 지원 시스템을 개발하였다.

주제어 : Knowledge Tracing(KT), eXplainableAI(XAI), 학습자 진단, 맞춤형 학습, 딥러닝

Development of Explainable AI-Based Learning Support System

Seonghun Kim† · Woojin Kim† · Yeonju Jang† · Hyeoncheol Kim††

ABSTRACT

The majority of online education platforms are efficient in providing its service to a large number of students. However, these online platforms hardly achieve individualized learning. This paper proposes a framework of an explainable intelligent tutoring system for teachers. The system uses an individual student's cognitive and environmental factors to assesses the student. The proposed framework uses DKT and XGBoost to model a student's knowledge state and analyzes the model's prediction with LRP and SHAP. Teachers receive the predicted knowledge state of a student as an interpretable form. Teachers may identify a student's understanding of the subject and environmental factors that impact a student's learning and provide individualized feedback. This paper contributes to an application of AI models and XAI techniques on education to achieve individualized learning.

Keywords : Knowledge Tracing(KT), eXplainableAI(XAI), Student Assessment, Individualized Learning, Deep Learning

†정 회 원: 고려대학교 일반대학원 컴퓨터학과
††종신회원: 고려대학교 교수(교신저자)
논문접수: 2020년 10월 30일, 심사완료: 2020년 11월 11일, 게재확정: 2020년 11월 19일
* 본 논문은 2018년 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2018-0-01405)

1. 서론

온라인 교육 시스템을 개발하기 위한 연구는 지속되어 왔다[1][2]. 온라인 교육 시스템으로 교수자의 교수법을 지원하기 위한 CAI(Computer Assisted/Aided Instruction), 학습자의 학습을 지원하기 위한 CBT(Computer-Based Test), 온라인에서 학습 콘텐츠를 제공하는 e-Learning 등과 같은 분야가 있다. 그러나 이러한 시스템은 많은 학습자를 효율적으로 교육할 수 있으나, 학습자의 수준을 진단하고 각 학습자에게 필요한 피드백을 적절히 제공하기는 어렵다[3].

학습자에게 교육적으로 유의미한 피드백을 제공하기 위해서는, 먼저 학습자의 지식 상태(Knowledge State)에 대한 진단이 선행되어야 한다[4]. 학습자의 지식 상태를 진단하기 위한 연구로 학습자와 관련된 데이터를 이용하여 학습자를 모델링하는 KT(Knowledge Tracing)가 있다[5]. 최근에는 KT에 딥러닝 모델을 적용하여 예측 정확도가 크게 증가하였다[6]. 그러나 기존의 통계기반 KT 모델에 비해 모델의 복잡도가 높아져 모델의 예측을 해석하는 데 어려움이 있다[7]. 사용자, 즉 교수자가 학습자 모델링의 결과를 해석하지 못하면, 학습 지원 시스템이 높은 성능의 모델을 사용하더라도 예상된 역할을 수행하기에는 한계가 있다. 따라서 모델의 예측 과정을 교수자가 이해할 수 있는 형태로 제공하는 것은 모델에 대한 신뢰성을 높이고, 시스템이 교수·학습에 적극적으로 활용되도록 하는 데 필수적이다[8].

본 논문에서는 교수자가 개별 학습자의 지식 상태를 진단하고, 맞춤형 처방을 할 수 있는 설명 가능한 AI 학습 지원 시스템을 제안한다. 설명 가능한 AI 학습 지원 시스템은 기계학습 모델과 딥러닝 모델을 사용하여 학습자의 지식 상태를 모델링하고 개별 학습자의 지식 상태를 진단한다. 그리고 설명 가능한 인공지능 기법을 적용하여 교수자에게 이해 가능한 형태로 학습자 진단 결과를 제공한다. 교수자는 진단 결과를 이용하여 학습자 맞춤형 처방을 할 수 있다.

시스템 적용 예를 제시하기 위해, 학습자 정보 중 인지적, 환경적 영역의 데이터를 시스템 구현에 사용하였다. 이렇게 본 시스템은 학습자의 지식 상태를 모델링할 때 인지적 영역의 정보뿐만 아니라 환경적 영역의 정보까지 반영함으로써, 학습자의 학업 성취에 영향을 미치는 학습자 내·외적 요인을 모두 고려한다[9].

인지적 영역의 데이터로는 Junyi Academy Math Practice Log 데이터의 학습자 513명의 6가지 수학 개념의 문제풀이 데이터[10]를 사용하고, 환경적 영역의 데이터로는 30가지 환경 정보와 학년 말 최종 수학 성적을 포함한 총 31가지 항목의 UCI Student Performance 데이터[11]를 사용하였다.

2. 연구 배경

2.1 Knowledge Tracing(KT)

Knowledge Tracing(KT)이란 어떤 기능을 습득하는 동안 변화하는 학습자의 지식 상태를 모델링하는 분야이다[5]. KT는 학습 이력 데이터를 활용하여 학습자의 지식 상태를 진단한다[6]. 여기서 학습 이력 데이터란 개별 학습자 별로 푼 문항과 정답여부를 나타낸다. 이를 통해 학습자의 학습 개념에 대한 숙련도를 정량적으로 진단할 수 있으며, 나아가 학습자 맞춤형 추천을 할 수 있다[6][7].

KT 모델은 특정 인지적 영역에 대한 학습자의 현재 숙달 수준이나 지식 상태를 확률로 예측한다. 예측된 지식 상태를 바탕으로 교수자는 효율적으로 학습자의 학업 성취를 높이는 데 적절한 피드백을 제공할 수 있다[12].

2.1.1 통계 기반 Knowledge Tracing

통계 기반 KT 모델로는 대표적으로 Bayesian Knowledge Tracing(BKT)가 있다. BKT는 은닉 마르코프 모델(Hidden Markov Model, HMM)와 Bayesian 추론을 사용하여 학습자의 지식 상태를 모델링하는 접근법이다[5].

BKT는 다음과 같은 한계를 가지고 있다[6]. 먼저 BKT는 길이가 긴 학습 이력에 대한 모델링이 어려워 MOOC 플랫폼 사용의 증가로 인해 생성된 복잡한 데이터를 처리하기에는 적합하지 않다. 또한 BKT는 학습자 지식 상태를 이진 표현으로 나타내는데 이는 다양한 학습자의 지식 상태를 표현하

기에 제약이 있다. 또한, BKT는 문항 간 관계를 독립적으로 가정하는데, 현실의 학습 개념은 위계적으로 관계가 있어 적용에 한계가 있다.

2.1.2 Deep Knowledge Tracing(DKT)

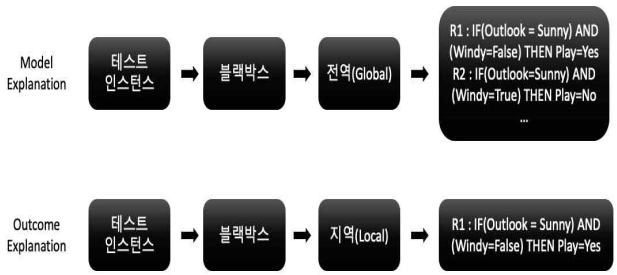
학습자의 지식 상태를 모델링하고 학습 성취를 예측하는 데 딥러닝 기반의 접근방법이 좋은 성능을 보이고 있다[6][13]. 한계점이 있는 BKT 보다는 딥러닝의 순환 신경망(Recurrent Neural Network, RNN) 모델이 복잡하고 긴 학습 이력에 대한 유의미한 정보를 추출할 수 있다. 또한, 학습 이력 데이터는 시계열 형태를 띠며, 이전 시간의 정보와 이후 시간의 정보 간의 관계성이 존재한다. 따라서, 이런 유형의 학습자 데이터에는 RNN 모델을 사용하는 것이 적합하다.

Deep Knowledge Tracing(DKT)은 그 중 하나로, RNN 계열인 Long Shot-Term Memory(LSTM) 구조를 이용하여 학습자 응답을 예측한다[6]. DKT의 AUC(Area Under the ROC Curve)는 0.86로 기존 BKT의 AUC가 0.65 정도였던 것에 비해 큰 성능 향상을 이루어 내었다[6].

2.2 설명 가능한 인공지능(eXplainableAI)

대부분의 머신러닝 모델과 딥러닝 모델은 사용자에게 내부 결정 과정을 숨기는 블랙박스(Black Box)이다. 군사, 의료, 교육 등 사람을 직접적으로 상대하는 분야에서 결정에 대한 설명을 제공하지 않은 모델을 사용하는 것은 실용적, 윤리적 문제를 야기할 수 있다. 설명 가능한 인공지능(eXplainableAI, 이하 XAI)은 블랙박스 모델의 예측값을 사람이 해석 가능한 형태로 의미를 제공하는 연구 분야이다. XAI 연구 문제인 블랙박스 설명(Black Box Explanation)은 모델 설명(Model Explanation), 출력 결과 설명(Outcome Explanation), 모델 검사(Model Inspection)의 3가지로 나뉜다[14].

모델 설명, 출력 결과 설명의 공통점은 블랙박스 모델의 숨겨진 내부 논리를 추출하여 사람이 이해할 수 있는 형태로 표현한다. 다만, 모델 설명은 데이터 전체에 대해 전역적(global) 설명을 하고, 출력 결과 설명은 데이터의 특정 인스턴스(instance)에 대해 지역적(local) 설명을 한다. 예를 들어, 날씨에 따라 외출할 지 아닐 지를 분류하는 모델에서 규칙들을 추출하여 사용자에게 제공할 때 [그림 1]과 같은 차이점이 있다.



[그림1] 모델 설명(Model Explanation)과 출력 결과 설명(Outcome Explanation) 비교

출력 결과 설명 방법에는 Layer-wise Relevance Propagation(LRP)[15], Local Interpretable Model-Agnostic Explanations(LIME)[16], SHapley Additive exPlanations(SHAP)[17] 등이 있다.

모델 검사는 특정 예측값 또는 블랙박스 모델의 특성을 이해하고자, 시각적 표현으로 설명을 제공한다. 제시된 시각적인 근거의 예로는 데이터 속성의 변화에 따른 모델 민감도나 모델의 특정한 예측에 영향도가 높은 뉴런의 발견 등이 있다.

2.2.1 Layer-wise Relevance Propagation(LRP)

Layer-wise Relevance Propagation(LRP)은 한 클래스의 우도(likelihood)가 입력 요소들이나 개별 레이어에 관한 노드들을 역(backward)방향으로 추적될 수 있다는 아이디어에 기반 한다[15]. LRP는 딥러닝 모델의 각 뉴런에 local redistribution rule을 적용하고 이 룰들을 역(backward)방향에 적용하여 모델 입력 값에 대하여 decomposition을 생산한다. 즉 LRP는 기존 모델을 이용하여 모델의 출력결과를 편미분 하여 출력에 영향을 주는 정도를 계산한다.

2.2.2 Shapley Value

Shapley Value는 다수의 플레이어가 있는 협력 게임 상황에서 각 플레이어의 중요도 순서를 결정하는 XAI 기법 중 하나이다[18]. 즉 Shapley Value

는 데이터 특성(feature)들의 모든 가능한 조합들 중에서 각 특성이 갖는 중요도의 정도를 나타낸다. 각 특성의 중요도는 특정 특성을 제외했을 때 해당 특성이 전체에 미치는 변화 정도를 계산하여, 각 특성이 모델의 예측에 기여한 정도를 구한다.

Shapley Value는 전체성과를 창출하는 데 각 특성이 얼마나 공헌했는지를 수치로 표현한다. 각 특성의 기여도는 그 특성의 기여도를 제외하였을 때 전체성과의 변화 정도로 나타낼 수 있다.

Shapley Value 기법을 회귀 문제에 확장하여 적용하면, 모든 특성들의 조합 가운데 각 특성이 기여하는 정도를 찾아내는 작업에 활용될 수 있다[18].

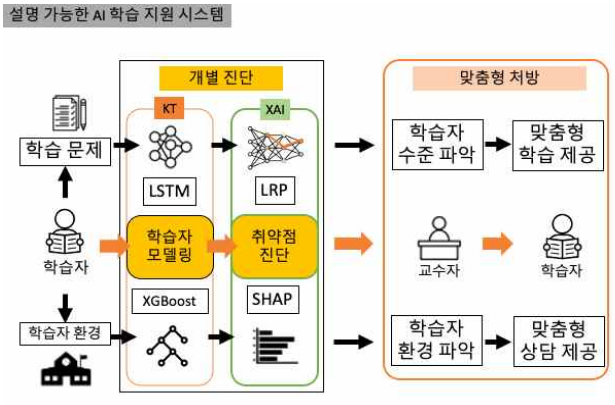
2.3 교육 데이터의 XGBoost 적용

교육 데이터 활용에는 KT 이외에도 머신러닝 모델을 사용하여 학습자 데이터 분석 및 교육 현장의 문제를 해결하려는 다양한 시도가 있다. 그중 XGBoost를 교육 현장에 적용한 사례들은 다음과 같다[19][20]. 중국 Beihang Shoue College에서는 XGBoost를 사용하여 'Four Pin' 교육 시스템에서의 학습자의 행동과 성적 사이의 관계를 모델링함으로써, 학습자의 행동과 학업 성취에 정적 상관관계가 있음을 확인하였다[19]. University of Miami의 Cengiz Zopluoglu는 시험을 정직하게 치루는 학습자와 그렇지 않은 학습자를 분류하는 작업에 XGBoost를 활용하였다[20].

XGBoost는 회귀 트리 앙상블 모델을 엑스트라 경사 부스팅 (eXtra Gradient Boosting) 프레임워크로 학습시키는 방식이다[21]. 트리 구조는 분류에서 주로 사용되나 회귀 문제 또한 해결할 수 있다. 트리 앙상블 모델은 성능 향상을 위해 개별 회귀 모델의 예측 결과 값을 혼합하여 이를 기반으로 최종 회귀 값을 예측한다. XGBoost는 이전에 주로 사용되었던 경사 부스팅(Gradient Boosting)의 시간적 제한을 극복한 알고리즘이며, 상대적으로 뛰어난 예측 성능을 발휘하기 때문에 트리 기반의 앙상블 학습에서 각광받고 있다[21].

하지만 트리 앙상블 모델은 여러 결정 트리 모델을 혼합하였기 때문에 예측값을 설명할 수 없는 블랙박스 모델이다. 앙상블 모델에 설명 가능한 인공지능 기법의 적용은 높은 예측 성능과 더불어 예측값에 대한 설명까지 제공함으로써 교육 데이터에서 유의미한 분석을 도출할 수 있다.

3. 설명 가능한 AI 학습 지원 시스템



[그림 2] 설명 가능한 AI 학습 지원 시스템

제안하고자 하는 설명 가능한 AI 학습 지원 시스템의 전체적인 구조는 [그림 2]와 같다. 설명 가능한 AI 학습 지원 시스템은 학습자와 관련된 학습 문제 데이터와 학습 환경 데이터를 활용한다. 시스템은 개별 진단 후 맞춤형 처방의 2단계로 구성되어 있다. KT와 설명 가능한 인공지능 기법을 사용하여, 학습자 개별 진단을 하고, 이를 바탕으로 맞춤형 처방을 제공한다. 개별 진단은 데이터 및 과업 특성에 알맞은 KT 모델로 학습자의 지식 상태를 예측하고, 설명 가능한 인공지능 기법으로 취약점을 진단한다. 맞춤형 진단 단계에서 교수자는 시스템의 결과를 활용하여, 학습자 수준 또는 환경을 파악한다. 그리고 맞춤형 학습 또는 상담을 학습자에게 제공할 수 있다.

학습자의 지식 상태에 대한 진단을 딥러닝 및 머신러닝 모델을 사용할 경우 설명 가능성의 역할이 중요하다. 학습자는 이러한 시스템에서 주어지는 진단을 자기 평가(Self-assessment)에 대한 일종의 피드백으로 여겨 학습에 긍정적인 영향을 끼친다[22].

3.1 인지적 영역 모델링

3.1.1 문제 상황

본 시스템은 학습자의 인지적 지식 상태를 모델링하기 위하여 DKT 모델을 사용하였다. 학습자의

학습 이력 데이터를 이용한 인지적 지식 상태의 모델링을 다음과 같이 규정한다. 데이터의 한 인스턴스는 학습 지원 시스템에서 제공된 문항 q에 대하여 t번째 시간에서의 학습자의 답변 a로 구성된다 (t: 시간, q: 문항, a: 학습자의 답변). t번째 시간까지의 문항에 대한 학습자의 답변 정보가 주어졌을 때, DKT는 t+1번째 시간에 주어진 문제를 학습자가 맞힐 확률을 예측한다. 이는 이진 분류(binary classification) 문제에 해당한다. DKT의 예측을 통해 미래에 맞힐 확률이 가장 큰 문항을 선별하고 학습자의 지식 상태를 가늠할 수 있다.

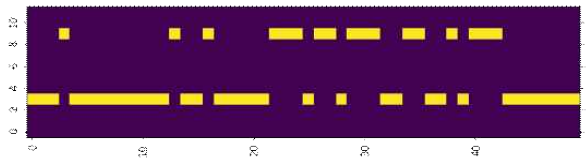
3.1.2 데이터 설명

학습이 일어날 때 학습자에게 요구되는 인지적 영역은 특정 개념에 대하여 가지는 추론 능력, 문제 해결력, 단순 이해력, 통합적 이해력 등을 포함한다[23]. 본 연구에서는 특정 개념에 대한 단순 이해력을 확인할 수 있는 것으로 가정하였다. 학습자의 인지적 영역 모델링은 Junyi Academy Math Practice Log 데이터를 사용하여 시도하였다. Junyi Academy는 타이완의 최대 온라인 학습 플랫폼이며 수학을 포함한 다양한 과목들에 대해 학습 자료와 퀴즈들을 제공하고 있다. Junyi Academy Math Practice Log 데이터는 이 학습 플랫폼에서의 학습자 기록을 담고 있다. 데이터는 수학 교과의 덧셈과 뺄셈 등의 기초 개념부터 통계와 선형대수의 심화 개념까지 포함한다. 해당 데이터는 카네기 멜론 대학의 교육 데이터 저장소인 DataShop에 공개되어 있다[10].

3.1.3 데이터 전처리

Junyi Academy Math Practice Log 데이터에는 199,549명 학습자의 25,628,935개 문제풀이 기록이다. 본 시스템은 이중에서 기초 수준의 수학 개념을 선택하여 그 개념에 대한 학습자의 인지적 영역을 모델링하고자 하였다. 선택된 수학 개념은 수세기, 수비교, 수순서, 덧셈, 뺄셈, 곱셈의 총 6개이다. 데이터의 모든 학습자 중에서 선택된 개념을 하나라도 푼 학습자의 수는 513명이며, 고유한 문항의 개수는 5,783개이다. 위 데이터를 328개의 훈련 데이터, 82개의 검증 데이터, 103개의 실험 데이터로 나누었다.

모델이 학습하는 데이터는 [그림 3]과 같이 전처리하였다. X축은 학습자가 문제를 푼 시간 순서이고, Y축은 학습 개념의 종류이며, 0~5까지는 데이터의 6가지 개념 중 맞힌 개념이고 6~11까지는 틀린 개념을 의미한다. [그림 3]은 학생 A가 3번 개념만을 풀다가 맞히고 틀린 학습 데이터를 보여준다. 학생 A가 3번 개념을 맞힌 경우에는 Y축 3에 표시가 되며 틀린 경우에는 Y축 9에 표시가 된다.



[그림 3] 시각화한 학생 A의 학습 문제 데이터

3.1.4 모델 구현 및 평가

모델 구현은 파이토치(PyTorch)를 사용하였으며, 구현된 DKT는 하나의 LSTM 층을 가지며, 입력층과 은닉층의 차원은 각각 100이다. 모델의 손실 함수는 이진 교차 엔트로피(binary cross entropy)를 사용하였으며 학습률(learning rate)은 0.001, 배치(batch) 크기는 128이다. 또한 분산 병렬 처리(Distributed Data Parallel) 기법을 적용하여 2개의 GPU를 가진 하나의 머신에서 훈련하였다. 모델의 AUC는 0.9981이다.

3.1.5 모델의 활용 및 적용 예시

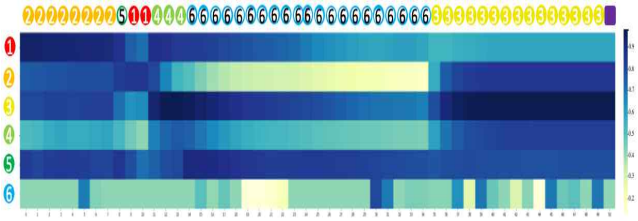
DKT는 [그림 3]의 학생 A가 다음 단계, 즉 51번째 시간에 각 학습 개념을 맞힐 확률을 예측한다. 이에 대한 예시는 <표 1>과 같다.

<표 1> 학생 A에 대한 DKT 예측 결과

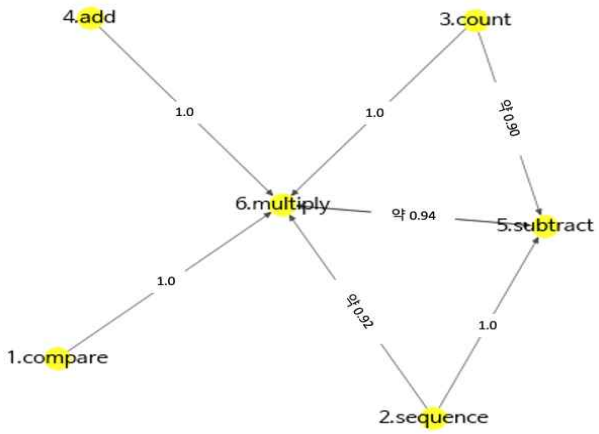
학습 개념 종류	해당 개념을 맞힐 확률
addition	0.9986
multiplication	0.990
number_sequence	0.9843
count_numbers	0.9834
comparison_between_numbers	0.9824
subtraction	0.3295

[그림 4] 50번까지의 문제를 풀 동안 모델이 예측한 학생 B의 지식 상태

또한, 모델의 예측값을 통해 학습자의 변화하는

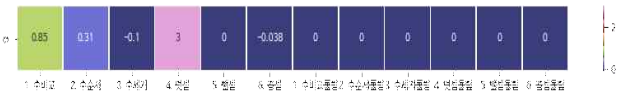


지식 상태를 시각화 할 수 있다. [그림 4]는 학생 B가 푼 50 문항의 기록이 주어졌을 때, 매 문항마다 DKT가 예측한 학습자 지식 상태를 시각화한 것이다. X축은 시간축이고 Y축은 위에서부터 1~6까지의 학습 개념이며, 색깔이 진할수록 맞힐 확률이 1에 가까운 것을 의미한다. 예를 들어, [그림 4]는 학생 B가 2번 유형의 문제를 맞히고 5번을 틀렸으며 1번, 4번을 맞힌 후 6번을 연속적으로 틀린 후에 3번을 연속적으로 맞힌 경우이다. 문제를 푼 순서 별로 학생 정답과 오답에 따라 예측값이 변화하는 것을 볼 수 있다. 6번 유형의 문항을 연속적으로 틀려 예측된 지식 상태가 계속 낮게 유지가 되며, 3번 유형의 문항을 연속적으로 맞혀 확률이 상승하는 것을 확인할 수 있다.



[그림 5] Junyi 데이터의 개념 간 상관관계

또한 DKT를 활용하여 개념 간 상관관계를 나타낼 수 있다. Junyi 데이터를 바탕으로 그려진 상관관계는 [그림 5]와 같다. 이는 학습자가 1번 개념을 맞혔을 때, 1번 개념을 제외한 다른 모든 문항들을 맞힐 확률이 0.9 임계치를 초과하는 개념 간 관계만을 그래프로 시각화한 것이다. [그림 5]는 다음과 같이 해석될 수 있다. 학습자가 4번 개념인 덧셈(addition) 개념의 문항을 맞힐 경우 대체로 6번 개념인 곱셈(multiply) 또한 맞힐 확률이 높다는 뜻이다.



[그림 6] 학생 A의 LRP 결과

DKT는 LSTM을 기반으로 한 블랙박스 모델이기 때문에 모델 예측 과정이 불투명하다. 앞서 언급된 LRP 기법을 사용하면 DKT 모델의 예측에 대하여 설명을 제공할 수 있다. 모델은 학생 A가 '덧셈' 문제를 맞힐 것이라고 예측하였고, 이 예측에 대해 설명을 제공하고자 LRP를 사용하여 도출된 결과를 시각화하면 [그림 6]과 같다. [그림 6]은 모델의 예측에 대한 각 문항의 영향도이다. LRP의 절댓값이 클수록 모델의 예측에 영향을 많이 끼친 것이고, 0에 가까울수록 영향도가 없는 것이다. LRP 값으로 예측 결과를 설명하면, 학생 A가 덧셈, 수비교 순으로 문제를 잘 해결했기 때문이며 학생 A가 틀린 문항은 예측값에 영향도가 없다고 설명할 수 있다. 이와 같이 LRP는 예측에 대하여 입력 중 예측에 대해 영향도가 큰 부분을 판단할 수 있다. 따라서 이러한 설명을 통해 학습자의 강점과 약점을 유추하고 학습자의 학업 성취를 높이기 위한 맞춤형 처방을 제공할 수 있다.

3.2 환경적 영역 모델링

3.2.1 문제 상황

학습자의 환경에는 가족 구성원의 크기, 부모의 직업 등 가정환경과 수업 결손여부, 교우 관계 등 학업 실태에 대한 정보를 포함한다. 이러한 환경은 학습자의 학업 성취에 중요한 영향을 미친다[9]. 따라서 학습자의 환경적 영역 정보를 사용하여 학습자 성적에 커다란 영향을 미친 학습자 환경 요소를 파악하고자 한다. 학습자 환경 요소를 바탕으로 학습자의 성취 정도를 예측하고, 그 예측에 학습자의 환경적 영역 중 영향도가 큰 영역 분석이 가능한 모델을 제시한다.

학습자의 환경적 요소에 따른 학습 성취 예측은 회귀 문제로 표현될 수 있다. 학습 문제 데이터와 달리 특성 순서 간의 관련성을 반드시 고려할 필요가 없기에 RNN에 비해 가볍고 예측과 분류 성능이 뛰어난 XGBoost[22]모델을 사용하여 학습자 모델링을 하였다. XGBoost 모델로도 특성들 간의 중요도를 계산할 수 있으나, 긍정적인 영향도 만을 계산하는 한계점이

있으므로, 각 특성의 기여도가 긍정적이거나 부정적인 영향을 끼치는 모든 경우를 확인할 수 있는 Shapley Value[17] 기법을 이용하여 취약점을 진단한다.

3.2.2 데이터 설명

학습자의 환경적 영역 모델링은 UCI Student Performance 데이터를 이용하였다[11]. 해당 데이터는 포르투갈의 두 학교 학생들의 개인 정보, 가족 정보, 학교생활 등의 학습자 환경 데이터와 더불어 수학 과목의 최종 성적을 담고 있다. 데이터는 UCI Machine Learning Repository에 공개되어 있으며, 총 395명의 학습자 정보로 이루어져 있다. 본 연구에서는 학습자의 개인 정보, 가족 정보, 학교생활 등의 학습자 환경 데이터를 통해 최종 수학 성적(0~20점)을 예측하였다. 데이터 셋이 제공하는 환경 특성은 <표 2>와 같다.

<표 2> Student Performance 데이터 특성

학습자 환경 특성	설명
failures	수업 불합격/미달 횟수
goout	친구들과 어울리는 시간
age	나이
studytime	공부 시간
romantic	이성교제 여부
famsize	가족 구성원 크기
grade	최종성적
Fedu / Medu	아버지 / 어머니의 최종 학력
Pstatus	부모님과 동거 여부
paid	추가 유료 수업 여부
internet	가정에서 인터넷 접속 여부
freetime	방과 후 여가 시간
absences	결석 횟수
famrel	가족관계 친밀도 정도
Fjob / Mjob	아버지 / 어머니의 직업

3.2.3 데이터 전처리

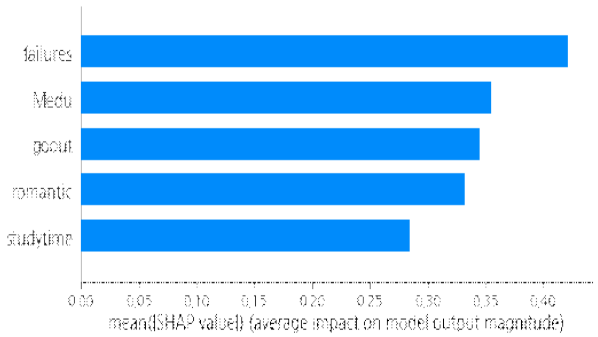
전체 395개의 데이터를 316개의 훈련 데이터와 79개의 실험 데이터로 나누었다. 인스턴스 갯수가 적어 교차 검증(Cross-validation)을 하였다. 0~20 사이의 정수인 grade와 이진 값인 Pstatus, internet을 제외한 나머지 특성들은 전처리 과정에서 각 특성 별로 전체 값의 평균을 기준으로 크면 1, 작으면 0으로 변환하였다.

3.2.4 모델 구현 및 평가

XGBoost모델의 학습률은 0.01이며 사용된 트리의 개수는 1,000개이다. 모델의 RMSE (Root Mean Squared Error)은 4.2552이다. Shapley Value의 구현은 SHAP 패키지를 이용하였다[17].

3.2.5 모델의 활용 및 적용 예시

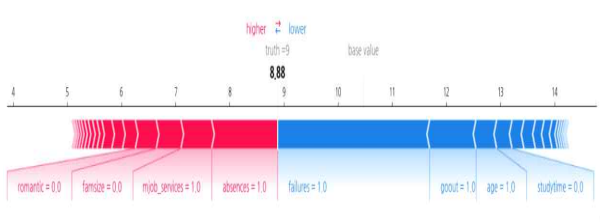
파악된 학습자 환경에 따라 맞춤형 상담을 제공하는 과정은 다음과 같다. Shapley Value 기법을 통하여 전체 학습자의 수학 성적에 영향도가 가장 큰 특성들을 확인할 수 있다. [그림 7]은 모델의 예측에 대한 변수들의 평균 영향도를 나타낸다. 수학 성적에 큰 영향을 미치는 특성으로는 failures (학습자가 수업에 fail한 횟수), Medu(어머니의 학력), goout(학습자가 친구들과 어울리는 시간), romantic(이성교제 여부), 그리고 studytime(평균 공부 시간)임을 알 수 있다.



[그림 7] 모델 출력 규모에 대한 평균 영향

그리고 학생 C의 예측된 성적을 바탕으로 취약점 진단을 하기위해 Shapley Value 기법을 적용하여 최종 성적에 긍정적인 또는 부정적인 영향을 끼친 환경적 요소들을 [그림 8]과 같이 시각화 하였다. 각 특성들은 전처리를 통해 0.0 또는 1.0으로 변환되어 표시되었고, 막대의 크기는 각 특성의 최종 성적에 대한 영향도를 의미한다. 실제 성적은 9점이나 모델의 예측은 8.88로 예측하였다. 어머니 직업의 직종이 서비스인 것(mjob_services=1)이 학생의 최종 성적 예측값을 높이는 데 영향을 미쳤다. 또한, 수업 불합격/미달 횟수가 많은 것(failures=1.0)과 친구들과 어울리는 시간이 많은 것(goout=1.0)이 학생의 최종 성적 예측 값을 낮추는 데 영향을 미친다고 설명할 수 있다.

[그림 8] 학생 C에 대한 예측 및 SHAP 결과



4. 교육적 의의 및 활용 방안

제안된 시스템의 교육적 의의 및 활용 방안은 다음과 같다.

첫째. [그림 5]와 같이 인지적 영역 모델링의 결과물을 시각화하여 나타낸 개념 간 상관관계는 상위 개념의 이해를 위하여 어떠한 하위 개념들이 선행되어야 하는지를 해석할 수 있다. 따라서 이를 활용하여 교과 개념의 위계를 재정립하거나 수정하는데 참고로 사용될 수 있다.

둘째. 환경적 영역 모델링의 결과물인 [그림 8]을 통해 교수자는 교수 학습자의 학업 성취에 영향을 미치는 주요 환경적 요인을 파악하고 맞춤형 상담을 제공할 수 있다. 또한, 환경적 영역 모델링 이외에도 SHAP을 이용한 기법은 학습의 다른 요소로도 확장할 수 있다. 학습자의 환경적 영역에 대한 정보 대신에 학습자의 각 과목별 또는 개념별 성취 정보를 바탕으로 최종 성적을 예측하는 모델을 생성할 수도 있을 것이다. 이 모델에 SHAP 기법을 적용한다면 한 학습자의 최종 성적 예측값에 어떠한 과목이 긍정적인 또는 부정적인 영향을 주었는지 분석할 수 있다.

셋째, 기존의 DKT 분야의 기법은 주로 학습자의 교과 이해 정도만을 보았다면, 본 연구에서는 학습자 환경 정보와 같은 환경적 영역에 대한 진단과 처방도 제시하였다.

넷째, 시스템은 local explanation 기법들을 사용한다. 전체 분포에서 학습자의 상대적 위치나 수준 정보가 아닌 학생 개개인에 대한 분석을 제공하므로, 예상되는 학습자의 심리적 부담 경감을 기대할 수 있다.

다섯째, 교과에 한정되지 않고, DKT와 XAI 분야의 기법을 교육에 적용 할 수 있는 범용적인 시스템을 제안하였다. 인지적 영역의 경우 수학 교과의 데이터를 사용하였지만, 제안하고자 하는 시스템은 어느 교과로도 확장이 가능하다.

5. 결론 및 후속연구

본 연구는 온라인 교육 시스템에서 부재한 맞춤형 학습을 해결하기 위해, 설명 가능한 AI 학습 지원 시스템을 개발하였다. 시스템은 인공지능 모델로 학습자의 지식상태를 예측하고, 그 결과를 설명 가능한 인공지능 기법으로 분석해 학습자의 지식상태를 해석 가능한 형태로 교수자에게 제공한다. 교수자는 이 정보를 바탕으로 맞춤형 학습을 제공하는데 활용할 수 있다.

이상의 연구를 바탕으로 다음의 향후과제를 진행하고자 한다.

첫째, 학습자 환경 데이터는 성별이나 가정환경 등의 민감한 정보를 담고 있기 때문에 모델이 학습자의 성별이나 경제적 환경에 편향되지 않은 결정을 내리는지에 대한 검증이 필요하다.

둘째, 본 연구에서 제안한 시스템의 효과성 및 신뢰성을 검증하기 위한 전문가 검토가 필요하다. XAI의 경우 일반적으로 사람 전문가의 검토를 통해 시스템의 신뢰성을 검증한다.

셋째, 맞춤형 학습과 상담을 동시에 제공할 수 있는 범용적인 기계학습 모델을 개발하여 시스템 효율성 제고가 필요하다.

참고문헌

[1] 김경미, 김현숙. (2017). 디지털시대에 플립드러닝을 활용한 학습자 맞춤형 소프트웨어 교육 방안 연구. Journal of Digital Convergence, 15(7), 55-64.

[2] 김경아, 문남미. (2010). 수준별 프로그래밍 교육을 위한 단계별 클러스터링 기반 추천시스템. 한국컴퓨터정보학회논문지, 15(8), 51-58.

[3] 안진현. (2018). 프로그래밍 교육을 위한 학습자 모델링 및 맞춤형 피드백 기법 연구. 석사학위 논문. 고려대학교 대학원. 서울.

[4] Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. Journal of Educational Measurement, 44(4), 341-359.

[5] Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.

[6] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Advances in neural information processing systems (pp. 505-513).

[7] Lu, Y., Wang, D., Meng, Q., & Chen, P. (2020). Towards Interpretable Deep Learning Models for Knowledge Tracing. In International Conference on Artificial Intelligence in Education (pp. 185-190).

[8] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

[9] 박현린, 김누리. (2015). 학교와 가정의 사회심리적 환경 변인이 저성취 초등학생의 학업성취에 미치는 효과. 아동교육, 24(2), 39-52.

[10] 미국 카네기 멜론 대학의 DataShop 사이트. 2020.10.28. 검색 https://pslcdatashop.web.cmu.edu/Project?id=244

[11] 미국 UCI Machine Learning Repository 사이트. 2020.10.28. 검색 https://archive.ics.uci.edu/ml/datasets/student+performance

[12] Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models. In International Conference on Artificial Intelligence in Education (pp. 171-180).

[13] Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017, April). Dynamic key-value memory networks for knowledge tracing. In Proceedings of the 26th international conference on World Wide Web (pp. 765-774).

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[15] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).

[18] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28), 307-317.

[19] Guang-yu, L., & Geng, H. (2019). The Behavior Analysis and Achievement Prediction Research of College Students Based on XGBoost Gradient Lifting Decision Tree Algorithm. In Proceedings of the 2019 7th International Conference on Information and Education Technology (pp. 289-294).

[20] Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). Educational and psychological measurement, 79(5), 931-961.

[21] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm

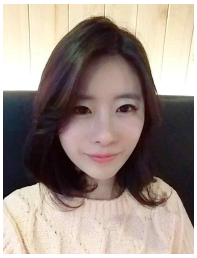
sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
[22] Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. arXiv preprint arXiv:1807.00154.
[23] 김현주, 김원경. (2017). 수학 교과에 대한 정의적 영역과 인지적 영역의 연관성에 대한 종단 분석. 교원교육, 33(2), 67-88.

김 성 훈

2011년 청주교육대학교(교육학 학사)
2020년 고려대학교 일반대학원 컴퓨터학과(박사 수료)
2020년 ~ 고려대학교 대학원 연구원
관심분야: SW/AI교육, 기계학습
E-Mail: ryankim0409@korea.ac.kr

김 우 진

2018 Macalester College
B.A Computer Science
2019 ~ 현재 고려대학교 일반대학원 컴퓨터학과 석박통합과정
관심분야: 딥러닝, 학습자모델링
E-Mail: woojinkim1021@korea.ac.kr



장 연 주

2013~ 초등학교 교사
2013 서울교육대학교 초등컴퓨터교육과(교육학학사)
2019 서울교육대학교 초등컴퓨터교육과(교육학석사)
2020 ~ 현재 고려대학교 일반대학원 컴퓨터학과 박사과정
관심분야: 컴퓨터교육, 인공지능 교육, 머신러닝, 딥러닝
E-Mail: spring0425@korea.ac.kr

김 현 철

1988년 고려대학교 전산과학과(학사)
1990년 Univ of Missouri-Rolla 전산학석사
1998년 Univ of Florida 전산정보학 박사
1999년~현재 고려대학교 컴퓨터학과 교수
2014년 ~ 2018년 한국컴퓨터교육학회 회장
관심분야 : SW/AI교육, 기계학습
E-Mail: harrykim@korea.ac.kr