강 태 훈 교수 지도

석사학위 청구논문

인지진단평가를 위한 머신러닝의 적용 가능성 탐색

- 데이터 크기와 이상적 문항반응 유형을 중심으로 -

2020

성신여자대학교 대학원

교육학과

허 재 은

인지진단평가를 위한 머신러닝의 적용 가능성 탐색

- 데이터 크기와 이상적 문항반응 유형을 중심으로 -

강 태 훈 교수 지도

이 논문을 석사학위논문으로 제출함

2020년 5월

성신여자대학교 대학원

교육학과

허 재 은

인 준 서

허재은의 석사학위 논문으로 인준함

2020년 5월

심사위원장 ..... (서명 또는 인)

심 사 위 원 ..... (서명 또는 인)

심 사 위 원 ..... (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

본 연구에서는 최근 많은 관심을 받는 연구 방법인 머신러닝을 인지진단평가에 적용해보고자 하였다. 인지진단모형 중 하나인 DINA 모형과 머신러닝과의 비교를 통해 피험자 능력추정 측면에서의 수행을 확인하였으며, 이를 바탕으로 인지진단평가를 위한 머신러닝(machine learning)의 활용 가능성을 탐색해보았다.

연구의 진행을 위하여, 두 가지 자료에 대한 분석을 수행하였다. 첫째, 피험자 3,000명이 응답한 실제 영어 검사자료를 사용하여 DINA 모형과 머신러닝의 피험자 능력 모수 추정의 일치도를 분석하였다. 검사는 총 18문항으로 구성되어 있으며 총 5개의 인지 요소를 가지고 있다. 5개의 인지 요소는 '인식하기', '이해하기', '추론하기', '평가하기', '구성하기'이다. 둘째, 실제 검사자료(real data)에서 추출한 문항 모수 추정치를 생성 모수로 활용하여 모의실험 자료(simulation data)를 생성하여 평가 자료(test data)로 활용하였다. 모의실험 자료는 총 18개의 문항에 대한 응답 반응 3,000개이다.

머신러닝으로 추정되는 피험자 능력 모수는 5개 인지 요소 각각의 숙달 여부와 인지 요소 숙달 패턴을 의미한다. 먼저, 머신러닝의 적용 가능성을 확인하기 위하여 실제 영어 검사 자료를 활용하여 인지진단모형인 DINA 모형으로 피험자 모수를 추정하였다. 그리고 이를 훈련 자료(train data) 및 평가 자료(test data)로 사용하여 머신러닝을 적용하였으며, DINA 모형과 머신러닝 간의 능력 모수 추정에 대한 일치도를 살펴보았다. 실제 검사자료 분석 결과, 자료의 크기가 클수록 분류 일치도가 높은 것으로 나타났으며 머신러닝과 DINA 모형의 일치도가 최소 85%에서 최대 100%로 나타나 머신러닝의 활용 가능성이 높은 것으로 판단할 수 있었다.

그러나 실제 교육 현장에서 인지진단평가를 위한 머신러닝을 적용하고자

할 때 실제 학생들의 문항 반응 자료 및 인지진단모형으로 추정한 인지 요소 숙달 여부를 훈련 자료로 확보하기 힘들다는 어려움이 있기에, 본 연구에서는 Q-행렬을 기반으로 한 이상적 문항 반응 유형으로 훈련 자료를 대체할 수 있는지 살펴보고자 하였다. 이때, 훈련 자료인 이상적 문항 반응 유형의 크기에 따라 피험자 능력 추정의 결과에 차이가 있는지를 확인하기 위하여 실제 검사자료를 평가 자료로 활용하였다. 분석 결과, 훈련 자료의 크기는 능력 모수 추정에 큰 영향을 미치지 않는 것으로 나타나 이상적 문항 반응 유형을 인위적으로 늘릴 필요 없이 이론적인 인지 상태의 조합의 수를 사용하면 되는 것으로 판단하였다.

마지막으로, 머신러닝 방법에 따른 피험자 능력 모수 추정의 정확성을 확인하기 위하여 모의실험 자료를 활용하여 진능력모수, 즉 참값(True)인 진 숙달 여부의 복원을 살펴보았다. 모의실험 자료를 통해 머신러닝 방법에 따른 정확도를 분석해본 결과, k-최근접 이웃(k-nearest neighbor)과 그래디언트 부스팅(gradient boosting)이 가장 적합한 것으로 나타났다.

기존의 인지진단모형은 문항 모수 및 피험자 모수를 추정하기 위해서 큰 표본이 필요하였기에 실제 교육 현장에 적용하는 데 한계점이 있었다. 하지만 타당한 Q-행렬이 존재한다면 기존의 검사 자료가 없더라도 이상적 문항 반응과 머신러닝을 활용하여 개별 학생을 진단할 수 있다는 것을 제안하였다는 점에서 본 연구의 주된 의의가 있다. 머신러닝을 활용한 진단 결과는 DINA 모형을 사용한 피험자 능력 모수 추정 결과와 크게 다르지 않으며 진능력모수 복원력도 비슷하다는 것을 확인하였다. 대부분의 교육 현장이 소표본인 상황을 고려한다면 머신러닝이 인지진단모형의 대안이 될 수 있다는 점에서 다른 인지진단모형과의 비교 및 다양한 인지진단평가 상황에 대한 추가적인 연구가 필요하다.

주요어: 인지진단평가, DINA 모형, 이상적 문항 반응 유형, 능력 모수 추정, 머신러닝, 기계학습, 지도학습

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

1. 연구의 필요성 및 목적

교육에서 이루어지는 평가는 평가 시기와 평가 점수의 해석 기준에 따라 구분이 가능하다. 먼저, 평가가 어느 시기에 시행되는지에 따라 진단평가(diagnostic evaluation), 형성평가(formative evaluation), 총괄평가(summative evaluation)로 구분할 수 있다. 진단평가는 교수·학습을 시작하기 전에 학생의 인지적, 정의적 특성을 파악함으로써 이에 맞는 수업전략과 학습과제를 구성하고자 실시하는 평가이다. 형성평가는 교수·학습 과정 중에 이루어지는 평가로 학생의 부족한 부분을 파악하고 교사의 교수 전략 개선에 도움을 제공하고자 실시하는 평가이다. 형성평가는 교수·학습 과정 중에 이루어지기 때문에 피드백이 즉각적이라는 특징이 있다. 총괄평가는 교수·학습이 완료된 이후에 교육목표가 어느 정도 달성되었는지를 종합적으로 판단하기 위해서 실시하는 평가이며 교육 현장에서는 중간고사, 기말고사라고도 부른다. 다음으로, 평가점수의 해석 기준에 따라 학생들 간의 비교를 기준으로 하는 규준 지향평가(상대평가)와 교육 목표 성취 수준을 기준으로 하는 준거 지향평가(절대평가)로 구분할 수 있다.

이러한 교육 평가의 주된 목적은 교수·학습을 통해 교육 목표가 얼마나 달성되었는지를 확인하고 평가 결과를 바탕으로 교사, 학생, 학부모에게 적절한 피드백을 제공함으로써 교수·학습 과정을 개선하는 것이다(성태제 외, 2013). 그러나 교육 현장에서 이루어지고 있는 대부분의 평가는 결과 중심의 총괄평가이자, 총점을 기반으로 학생의 능력을 추정하고 상대적 위치인 서열을 부여함으로써 학생들의 능력을 비교하고자 하는 규준지향 평가이다. 따라서 교육 현장에서 학생에게 제공하는 정보는 해당 교과목의 원점수, 등급에 대한 정보를 담고 있는 성적표이며, 이는 학생들이 교육과정에서 요구

하고 있는 교육 목표 중에서 무엇을 달성하고 무엇을 달성하지 못했는지에 대한 구체적인 정보를 제공할 수 없다. 또한 학생이나 학부모가 평가 결과를 보고 의미 있는 피드백을 얻기 힘들 뿐만 아니라 교사들에게도 교수·학습 과정을 개선하기 위한 의미 있는 정보를 제공하기 어렵다는 한계가 있다(de la Torre, 2009; Mislevy, 1996).

교육 평가의 주된 목적을 달성하기 위해서는 학생의 능력을 정확하게 추정하고 단순히 원점수와 등급뿐만 아니라 더 많은 정보를 평가 결과로 제공해야 한다. 총점방식의 평가 결과만으로는 교수·학습이 잘 이루어졌는지, 학생들이 교육목표를 달성했는지를 판단하기 어렵다. 이러한 한계를 극복하기 위하여, 최근 교수·학습 과정 중에 즉각적이고 구체적인 피드백을 제공함으로써 학생들의 학습을 돕는 형성평가 관점의 논의가 활발하게 이루어지고 있다. 교육 평가적 측면에서는 평가의 결과로 진단적 정보, 즉 학생의 인지상태에 대한 프로파일을 제공하여 평가 이후에 학생이 학업에 대한 자신의 장단점을 파악할 수 있도록 돕는 인지진단이론(Cognitive Diagnostic Theory, CDT)이 연구되었다(김성훈, 송미영, 2011; 김경리, 2012; 김지효, 2013; 김희경, 박종임, 강태훈, 2013). 인지진단모형은 학생들의 문항 반응을 바탕으로 검사의 정답을 맞히기 위해서 필요한 인지 요소의 숙달 또는 미숙달 여부를 확률적으로 알려준다(Rupp, Templin, & Henson, 2010). 이를 통해 개별 학생의 학업 성취에 관한 정보를 구체적으로 알려줌으로써 학생이 자신의 강점과 약점을 파악할 수 있도록 지원한다(성태제 외, 2013).

그러나 형성 평가적 측면에서 인지진단모형을 교실 혹은 학교 수준에 적용하기에는 어려움이 존재한다. 이를 적용하기 위해서는 신뢰할 수 있는 안정적인 문항 모수 및 피험자 모수 추정을 위한 큰 표본이 요구되기 때문이다(이영주, 2014). 김지효(2013)는 DINA와 DINO 모형에서 동일한 검사 조건에서는 500명 이상의 표본 크기가 필요하다고 주장하였다. 이러한 현실을

반영하듯 인지진단평가에 관한 국내외 선행연구 대부분이 국제수준 또는 국가 수준의 대규모 검사 자료를 활용하여 수천 명의 피험자를 기준으로 연구를 진행하였지만, 대부분의 단위학교에서 고려할 수 있는 피험자의 수는 50명에서 300명 정도이다(은효정, 2017). 이러한 피험자 수의 제약은 인지진단평가가 단위학교에서 활발하게 이루어지지 못하는 가장 큰 원인이 되고 있다. 그러나 은효정(2017)이 고등학교 교사를 대상으로 한 설문조사 결과에 따르면, 많은 교사가 인지진단모형의 적용으로 산출된 평가 결과는 학생의 능력 수준을 파악하고 교수·학습 계획을 수립하는 데 도움이 되는 의미 있는 정보라고 생각하였으며, 인지진단평가 결과를 현장에서 활용할 의사가 있는 것으로 나타났기에 소규모 표본 상황에서의 인지진단 적용에 관한 연구의 필요성이 증대되었다.

이에 검사 자료가 충분하지 못한 상황에서 인지진단평가를 적용하기 위한 방법으로, Q-행렬에서 얻을 수 있는 이상적 문항 반응 유형(ieal response pattern)을 훈련 자료로 사용하는 머신러닝 방법에 관한 연구가 이루어졌다. 머신러닝(machine learning)은 자료(data)에 내재된 규칙을 컴퓨터가 스스로 인식하고 자료를 가장 잘 표현할 수 있는 적합한 함수(fuction)를 산출하는 것을 의미하며, 컴퓨터가 스스로 규칙을 찾아 나가는 과정을 학습(learning) 또는 훈련(training)이라고 부른다(박해선, 2019).

이와 관련하여, Gierl et al.(2007)은 이론적으로 가능한 인지 요소 패턴과 Q-행렬을 바탕으로 기대되는 응답 반응(expected response pattern)을 생성하여 인공신경망의 훈련 자료로 사용하였으며, 요소 간 위계방식(Attribute Hierarchy Method, AHM)에 인공신경망을 적용하여 SAT 검사자료를 분석하였다(Gierl et al., 2008). Shu et al.(2013)은 소규모 사례 수에서의 인공신경망의 활용 가능성을 탐색하였으며, Guo et al.(2017)은 인지진단평가에 인공신경망을 적용한 결과, 피험자 모수 추정의 정확도가 DINA 모형의 정확

도와 비슷하거나 약간 우수하고, 인지 요소 간 관계(skill prerequisite relation) 복원에서는 인공신경망이 더 우수하다고 주장하였다. 우리나라에서는 윤지영(2015)이 Q-행렬을 기반으로 인지 상태에 대한 이론적 데이터인 이상 반응 유형(expected response pattern)과 머신러닝 방법의 한 종류인 인공신경망(Artificial Neural Network, ANN)을 활용한 인지진단평가 방법을 제안하였으며, 이상 반응 유형과 인공신경망을 결합한 인지진단모형은 학생들의 인지상태를 추정하기 위하여 큰 표본이 필요하지 않음으로 학교 현장에서 유용하게 사용할 수 있다고 주장하였다. 이창묵(2019)은 MMPI 심리검사에 대해 이상적인 응답 반응(ideal response pattern)과 심층신경망 모형을 적용한 결과, 심층신경망 모형의 정확도가 양호한 수준으로 산출되었기에 새로운 분석 모형으로써 활용해 볼 수 있다고 주장하였다. 이러한 선행연구를 바탕으로, 이론적 데이터를 활용함으로써 인지진단평가에 머신러닝의 적용이 가능하다는 것을 확인해볼 수 있었다. 그러나 선행 연구는 모두 인공신경망을 활용하고 있음으로 인공신경망 이외의 다른 머신러닝이 적용이 가능한지에 관한 추가적인 연구가 필요하다.

교육 분야에서 학생의 학업 성취와 관련하여 머신러닝을 적용한 연구는 학습관리시스템(learning management system, LMS)에서 생성된 학습자 행동에 대한 자료를 머신러닝 방법으로 분석하는 연구(이혜주, 정의현, 2013; 이혜윤, 2016), 배경 변인을 바탕으로 학생의 학업성취를 예측하고자 하는 연구가 대부분이다(배재호, 2001; 김혜숙, 2006; 오지영, 이수정, 2008; 김완섭, 2012). 최근에는 데이터 마이닝 결과를 활용하여 교수·학습의 성과를 개선하는 학습분석학(Learning Analytics) 측면에서의 연구가 활발하게 진행되고 있으나, 대부분의 연구에서 총괄평가의 결과를 활용하고 있는 것으로 나타났다(조일현, 김윤미, 2013; 조일현, 김정현, 2013; Kim, Park, Yoon, & Jo, 2016). 그러므로 교육 평가의 주요 관심사인 검사 자료, 즉 문항 반응에

머신러닝을 적용하여 피험자 능력 모수를 정확하게 추정할 수 있는지에 대한 연구가 지속적으로 수행될 필요가 있다.

따라서 본 연구에서는 인공신경망 이외의 머신러닝 방법이 인지진단평가에 적용될 수 있는지를 탐색해보고자 한다. 본 연구의 목적은 다음과 같다.

첫째, 머신러닝과 DINA 모형 간의 능력 모수 추정 일치도를 분석해봄으로써 머신러닝의 활용 가능성을 탐색해보고, 다양한 크기의 사례 수에 대해서 머신러닝을 적용하여 검사 자료의 크기에 따라 인지 요소 숙달 여부 분류 일치도에 차이가 있는지를 확인해보고자 한다.

둘째, Q-행렬을 기반으로 생성한 이론적 데이터인 이상적 문항 반응 유형(ideal item response pattern)을 머신러닝의 훈련 자료(train data)로 사용하였을 때, 훈련 자료의 크기에 따라서 인지요소 숙달 여부 분류 일치도에 차이가 있는지를 확인해보고자 한다.

셋째, 머신러닝이 피험자의 능력모수를 정확하게 추정할 수 있는지를 진능력 모수의 복원(recovery) 관점에서 확인해보고자 한다. 이를 위하여, 실제 검사 자료로부터 추정한 문항 모수를 생성 모수(generating parameter)로 하여 가상의 모의실험 자료(simulated data)를 생성한 후에 다양한 머신러닝 방법을 적용하여 피험자 능력 모수인 인지요소 숙달 여부 및 숙달 패턴을 추정하고, 그 결과를 진능력모수와 비교해봄으로써 진능력모수 복원의 정확성을 평가한다. 이를 통해 인지진단평가를 위하여 가장 적합한 머신러닝 방법이 무엇인지 탐색해보고자 한다.

2. 연구 문제

연구의 필요성과 목적에 따른 연구 문제는 아래와 같이 정의된다.

연구문제 1. 인지진단평가에 머신러닝을 적용할 수 있는가?

실제 검사자료(real data)의 문항 반응(item response)과 DINA 모형을 통해서 추정한 피험자 능력모수 추정 자료를 활용하여 머신러닝을 적용하였을 때, 자료 크기(data size)에 따라 DINA 모형과 머신러닝의 능력모수 추정 일치도에 차이가 있는가?

연구문제 2. 이상적 문항 반응 유형의 크기에 따라 차이가 있는가?

이상적 문항 반응 유형(ideal item response pattern)을 머신러닝의 훈련 자료(train data)로 사용하였을 때, 크기에 따라 DINA 모형과 머신러닝의 능력모수 추정 일치도에 차이가 있는가?

연구문제 3. 머신러닝 방법에 따라 분류 정확성에 차이가 있는가?

k-nearest neighbor(KNN), Decision tree, Random forest, AdaBoost, Gradient boosting의 결과를 비교하였을 때, 어떤 방법이 인지진단평가 도구로서 가장 정확한가?

Ⅱ. 이론적 배경

1. 심리측정이론

가. 고전검사이론(Classical Test Theory, CTT)

고전검사이론(Classical Test Theory)은 각 문항의 점수를 합한 검사에 대한 총점에 근거하여 학생들의 잠재적인 능력을 측정하는 이론이다(Crocker & Algina, 1986l). 고전검사이론은 계산과 추정 방법이 쉽고 점수 해석이 간단하다는 장점으로 현재까지도 학교 현장에서 가장 활발하게 사용되고 있는 측정이론이다.

Spearman(1904)은 고전검사이론의 기본 개념인 진점수와 신뢰도, 타당도의 개념을 개발하였다. 고전검사이론에서는 모든 검사에는 오차가 존재한다고 가정하며, 관찰점수는 진점수와 오차점수라는 독립적인 성분으로 이루어진 것으로 보았다. 따라서 고전검사이론에서의 검사와 문항에 대한 분석은 관철점수인 총점에 의존한다(성태제, 2014).

관찰점수 X는 진점수 T와 오차점수 e의 합으로 나타낼 수 있다. 진점수 T(the universe score)는 관찰되지 않은 피험자의 능력을 의미하며, 이는 알 수 없으므로 동일한 검사인 평형 검사를 무한히 반복하여 관찰되는 점수의 평균점수로 추정한다. 오차점수 e(measurement error, e)는 측정상황에서 일관성을 벗어나서 생기는 오차이다. 이는 '무선적 오차(random error)' 또는 '측정의 오차'라고도 불리며 모든 피험자는 동일한 측정의 오차를 가지고 있다고 가정한다. 이때, 진점수 T와 오차점수 e의 상관은 0이다. 이를 식으로 나타내면 아래의 식(1)과 같다.

$$X = T + e \qquad (1)$$

고전검사이론을 통해 이론적으로 정의된 검사의 신뢰도는 진점수 T와 관찰점수 X의 분산의 비로 정의할 수 있다. 이는 식(2)로 나타낼 수 있다.

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_e^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad (2)$$

또한, 고전검사이론 하에서는 문항 난이도(item difficulty), 문항 변별도(item discrimination)를 추정할 수 있다. 문항 난이도(P)는 문항에 정답반응한 피험자 수(R)를 총 피험자 수(N)로 나눈 것이다. 문항 변별도는 문항이 피험자를 변별하는 정도를 나타내는 것으로, 난이도가 높은 문항에 대하여 피험자의 능력이 높을 경우에는 정답 반응이 나타나고, 피험자의 능력이 낮을 경우에는 오답반응이 나타나는 문항이 변별도가 높은 문항이라고 볼 수 있다. 이는 문항점수(X)와 피험자 총점(Y)의 상관계수(r)로 나타낸다.

$$P = \frac{R}{N} \qquad (3)$$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \qquad (4)$$

그러나 고전검사이론에서 추정되는 피험자의 능력인 검사점수는 검사에 종속적이며, 문항모수인 문항 난이도와 문항 변별도의 추정치는 피험자 집단에 종속적이다(Hambleton, Swaminathan, & Rogers, 1991). 쉽게 말해, 피험자의 능력은 측정하는 검사 도구에 따라서 다르게 측정될 수 있으며 문항의 난이도와 변별도는 피험자 집단의 특성에 따라서 결정된다는 것이다. 예를 들어, 동일한 피험자일지라도 난이도가 어려운 문항으로 구성된 검사에

서는 피험자의 능력이 낮게 측정될 수 있으며, 난이도가 쉬운 문항으로 구성된 검사에서는 능력이 높게 측정될 수 있다. 또한 동일한 검사 도구를 사용하더라도 시험을 치르는 피험자 집단의 능력구성에 따라서 문항의 난이도와 변별도가 다르게 추정될 수 있다. 능력이 높은 피험자로 구성된 집단에서는 문항의 난이도가 쉬운 것으로 추정될 수 있으며, 능력이 낮은 피험자로 구성된 집단에서는 문항의 난이도가 어려운 것으로 추정될 수 있다. 따라서 피험자 능력의 불변성과 문항모수의 불변성이 성립하지 않는 것으로 볼 수 있으며, 고전검사이론이 가정하고 있는 모든 검사가 동일하다는 평형검사 가정과 모든 피험자가 동일한 측정의 오차를 가진다는 가정은 성립하기 힘들다는 한계를 가지고 있다.

고전검사이론 하에서는 <표 Ⅱ-1>과 같이 총점이 같으면 두 피험자의 능력은 같은 것으로 추정한다. 그러나 문항1에서부터 문항5까지 난이도가 증가하도록 구성된 검사라고 가정한다면 두 피험자의 능력이 같다고 볼 수 없다. 이러한 고전검사이론의 단점을 보완하기 위한 문항반응이론이 등장하였다.

<표 Ⅱ-1> 문항응답 결과와 검사점수(성태제, 2016)

	피험자1	피험자2
문항1	1	0
문항2	1	1
문항3	1	1
문항4	0	1
문항5	0	0
총점	3	3

나. 문항반응이론(Item Response Theory, IRT)

문항반응이론(Item Response Theory)은 문항에 대한 피험자의 반응을 확률적으로 표현한 모형으로서 Binet와 Simon(1916)이 지능검사의 문항에 대한 정답율을 피험자 집단의 나이 함수로 나타내는 아이디어에서 출발하였다. 각 문항별로 나이(x축)에 따른 정답확률(y축)을 나타내는 점을 찍고 그 점들을 연결하여, 각 연령대에 적합한 문항을 선택하고자 하였는데, 이것이 문항특성곡선(ICC: Item Characteristic Curve)의 출발점이 되었다.

고전검사이론의 분석의 단위는 검사의 총점으로 검사수준에서 분석이 진행되지만 문항반응이론에서는 하나의 문항에 대한 피험자의 반응, 즉 문항단위에서 분석이 이루어진다. 문항반응이론은 고전검사이론의 단점을 보완할 수 있는 능력모수 불변성과 문항모수 불변성을 기반으로 각 문항에 대한 피험자의 반응을 문항특선곡선에 의해 분석한다. 문항특선곡선은 theta(Θ) 척도 위에 나타낼 수 있으며, '피험자의 능력특성'을 나타내는 '피험자 능력모수'와 '문항의 특성'을 나타내는 '문항모수' 모두 이 theta(Θ) 척도 위에서 표현된다. 따라서 <표 Ⅱ-2>와 같이 고전검사이론 하에서는 총점이 같은 피험자1과 피험자2를 능력이 동일한 것으로 추정되지만 문항반응이론 하에서는 피험자1과 피험자2의 능력이 다르게 추정되는 것을 확인할 수 있다.

<표 Ⅱ-2> 총점과 피험자 능력 추정치(Θ) (성태제, 2016)

	총점	능력 추정치(Θ)
피험자1	8	.3773
피험자2	8	.2332
피험자3	5	-.6800
피험자4	18	2.4762
피험자5	6	-.4406
피험자6	7	-.1342
피험자7	6	-.7194

문항반응이론을 적용하기 위해서는 일차원성 가정과, 지역독립성 가정을 모두 충족해야한다. 먼저, 일차원성 가정은 하나의 검사 도구는 하나의 잠재적인 특성, 즉 하나의 능력만을 측정해야한다는 것으로 검사를 구성하고 있는 모든 문항들이 하나의 능력을 측정하여야 함을 의미한다. 다음으로, 지역독립성 가정은 어떠한 한 문항에 정답 반응할 확률과 다른 문항에 정답 반응할 확률이 서로 독립적이라는 것이다. 즉, 어떠한 한 문항에 대한 피험자의 응답이 다른 문항의 응답에 영향을 주지 않는다는 것으로, 한 문항의 내용이 다른 문항에 정답 반응할 수 있는 정보를 제공하지 않는다.

이러한 가정을 하에서, 각 문항을 피험자 능력대별로 문항의 정답을 맞힐 확률로 나타낸 문항특성곡선으로 나타낼 수 있다. 이때, 각 문항은 맞다(1)와 틀리다(0)로 채점되는 이분 문항이며 문항특성곡선은 S자 형태의 곡선으로 나타난다. 이러한 문항특선곡선은 위치모수와 척도모수를 갖는다. 위치모수는 정답률이 .5일 때의 능력 수준, 즉 문항이 theta(Θ) 척도의 어느 지점에 위치하는가를 의미한다. 이를 난이도(a_i)라고 부르며, 그 위치에 따라서 문항을 “쉬운 문항”, “중간 문항”, “어려운 문항” 등으로 구분할 수 있다. 척

도모수는 문항특성곡선이 문항의 위치 모수 아래의 능력을 가진 피험자들과 위치모수 위의 능력에 해당하는 피험자들을 잘 변별할 수 있는가를 의미하며, 이를 문항 변별도(b_i)라고 부른다. 문항 변별도는 문항의 정답률이 .5인 점에서의 곡선의 기울기로 표현한다. 문항 추측도(c_i)는 능력 수준이 음의 무한대에 수렴하는 학생이 정답에 반응할 확률을 의미한다.

한편 문항특선곡선은 각 i번째 문항에 대한 난이도, 변별도, 추측도와 같은 문항모수를 몇 개로 표현할 것인지에 따라 1-모수, 2-모수, 3-모수 모형으로 분류할 수 있다. 1-모수 모형은 식 (5), 난이도(a_i)만을 고려하며, 난이도(a_i)와 변별도(b_i)를 고려하는 2-모수 모형은 식 (6), 난이도(a_i), 변별도(b_i), 추측도(c_i) 모수를 고려하는 3-모수 모형은 식 (7)과 같다.

$$P_i(\theta) = \frac{1}{1+\exp[-(\theta - b_i)]} \quad (5)$$

$$P_i(\theta) = \frac{1}{1+\exp[-a_i(\theta - b_i)]} \quad (6)$$

$$P_i(\theta) = c_i + (1-c_i) \cdot \frac{1}{1+\exp[-a_i(\theta - b_i)]} \quad (7)$$

그러나 문항반응이론은 일차원성 가정과 지역독립성 가정으로 인해 적용에 제약이 존재한다. 대부분의 교육 및 심리검사에 한 문항에 올바르게 응답하기 위해서는 일반적으로 다차원적인 능력을 요구하기 때문에 이러한 가정은 충족되기 어렵다는 단점이 있다(김명연, 2016). 또한 검사자료의 분석결과로 제공되는 정보가 학생들의 인지적 과정이나 인지적 상태에 대한 진단적인 정보를 반영하고 있지 않다는 한계가 있다(de la Torre, 2009;

Mislevy, 1996). 이러한 단점을 보완하기 위해서 인지진단이론(Cognitive Diagnostic Theory, CDT)이 등장하였다.

2. 인지진단이론(Cognitive Diagnostic Theory, CDT)

가. 인지진단이론

인지진단이론(Cognitive Diagnostic Theory)은 검사를 통해 학생들의 지식 및 기능에 영향을 미치는 잠재능력 혹은 인지요소(cognitive attributes)를 파악할 수 있는 이론체계로서 이전의 측정 이론의 대안으로 활발하게 연구되고 있다(강태훈, 2019). 이전의 측정이론인 고전검사이론과 문항반응이론은 피험자의 능력을 단일 차원에서 추정하였다면, 인지진단이론은 검사에 요구되는 하위 인지요소를 기반으로 추정하는 다차원적 평가방법이다. 기존의 측정이론인 고전검사이론과 문항반응이론이 잠재변수인 피험자의 능력을 하나의 점수로 산출하고, 이 점수를 바탕으로 피험자를 비교하고 서열화하기 위한 목적을 가지고 있다. 반면, 인지진단이론은 잠재변수인 피험자의 능력을 다차원적인 인지요소로 구분하고, 이를 숙달(mastery)하였는지를 확인하여 개별 학생의 인지상태에 대한 구체적인 정보를 제공(Leighton & Gierl, 2007)하므로, 단순하게 평가에 그치지지 않고 학생의 학습을 지원하고 돕고자하는 목적을 가지고 있다.

따라서 <표 Ⅱ-3>과 같이 피험자들의 능력을 다차원적인 인지요소로 나타낼 수 있으며, 총점이 동일하더라도 문항반응이 달라질 경우에는 각 인지요소의 숙달 여부 및 인지요소 프로파일이 다르게 나타날 수 있다.

<표 Ⅱ-3> 총점과 인지요소 프로파일

	총점	인지 요소1	인지 요소2	인지 요소3	인지 요소4	인지요소 프로파일
피험자1	10	1	1	1	1	1111
피험자2	10	1	1	0	0	1100
피험자3	10	0	1	0	1	0101
피험자4	10	1	0	0	1	1001

인지진단이론은 학생의 인지구조와 인지적 과정에 대한 관심을 두는 평가 모형으로서 '인지심리학'과 '심리측정학'의 통합을 통해 만들어졌다. 인지진단을 통해 획득한 인지요소 숙달에 대한 정보를 바탕으로, 교수·학습 활동에서 사용할 수 있는 진단적 정보를 제공한다(Nichols, Chipman, & Brennan, 1995). 즉, 각각의 인지요소 숙달에 대한 정보로부터 학생의 인지상태를 진단함으로써, 인지요소 프로파일(attribute profile)을 생성하는 것이다(은효정, 2017). 이를 통해 학생들의 잠재적인 인지상태를 진단할 수 있으며, 인지진단 결과를 바탕으로 적절한 교육적 개입을 시도할 수 있다(윤지영, 2015).

Gierl, Leighton & Hunka(2000)는 인지진단모형을 적용하기 위해서는 세 가지 구성요소가 필요하다고 하였다. '검사를 통해서 측정하고자 하는 인지요소와 인지요소 간의 관계', '각 문항과 인지요소 간의 관계', 그리고 '문항 모수 및 피험자 모수를 추정하기 위한 심리측정 모형의 선택'이다.

나. 인지요소와 Q-행렬

인지요소(cognitive attributes)는 특정한 피험자의 인지 상태를 파악함에 있어서 가장 기본이 되는 요소로써, 피험자들이 각 문항에 정답으로 반응하기 위해서 요구되는 기술(skills), 지식(knowledge), 능력(ability), 인지적 과정(cognitive process) 등을 종합적으로 의미한다(Tatsuoka, 1983). 따라서 인지요소를 숙달하였다는 것의 의미는 해당 지식, 기술, 능력 등을 습득하였다는 것으로, 검사를 통해 측정하고자 하는 잠재적인 특성이나 능력을 갖추고 있다는 것으로 볼 수 있다.

'검사를 통해서 측정하고자 하는 인지요소들 간의 관계'는 보상적(compensatory) 관계, 비보상적(non-compensatory) 관계로 구분할 수 있다. 일반적으로 비보상적 관계를 결합적(conjunctive) 관계, 보상적 관계를 비결합적(disjunctive) 관계라고 표현하기도 한다. 보상적 관계에서는 하나의 인지요소만 숙달하더라도 각 문항에 정답으로 반응할 수 있다. 즉, 각 문항이 요구하는 모든 인지요소를 숙달하고 있지 않더라도 특정 인지요소의 숙달이 다른 인지요소의 미숙달을 보상할 수 있음으로 해당 문항에 정답으로 반응할 수 있다. 비보상적 관계에서 각 문항에 정답하기 위해서는 문항이 요구하는 모든 인지요소를 숙달해야한다. 각 인지요소가 결합적 관계이기 때문에 하나라도 미숙달이 있는 경우에는 해당 문항에 정답으로 반응할 수 없다. 이러한 인지요소와 인지요소 간의 관계는 교과 내용 전문가에 의해 정의된다.

Q-행렬은 '각 문항과 인지요소 간의 관계'를 수리적으로 나타내는 행렬로서 Tatsuoka(1983)에 의해서 도입된 개념이다. 검사 제작 과정에서 개발된 문항과 인지요소를 연결시키는 작업은 Q-행렬을 만드는 것으로 이루어지며(송미영 외, 2011), 각 문항과 이를 해결하기 위해 필요한 인지요소 간의 관계를 정확하게 나타내기 위해서는 Q-행렬이 반드시 필요하다. 따라서 Q-행

렬의 개발은 인지진단모형을 적용하여 학생의 인지상태를 측정하기 위한 기초가 되므로, 가장 주의 깊게 진행되어야 하는 과정이다.(김지효, 2013; Tatsuoka, 1991).

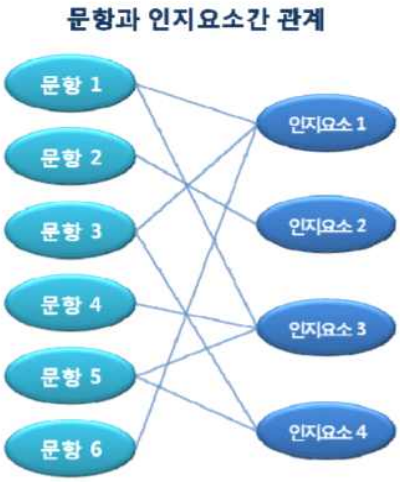
Q-행렬은 행(row)에는 각 문항을 나타내고, 열(column)에는 인지요소를 나타낸다. 따라서 인지요소의 개수가 K개이고, 문항이 J개일 때 K × J 차원을 갖는다. Q-행렬의 원소인 q_{kj}는 0과 1의 값으로 표현된다. 이때, 문항 j에 정답으로 반응하기 위해서 인지요소 k가 필요할 때는 1의 값을 갖고 문항 j에 정답으로 반응하기 위해서 인지요소 k가 필요하지 않을 때는 0의 값을 갖는다. 이는 아래의 식(8)과 같다.

$$q_{kj} = \begin{cases} q_{kj} = 1 \\ q_{kj} = 0 \end{cases} \qquad (8)$$

Q-행렬은 각 문항에 정답하기 위하여 어떠한 인지요소가 필요한지에 대한 정보를 제공한다(김명연, 2016). 예를 들어, [그림 Ⅱ-1]과 같은 문항과 인지요소 간의 관계를 가지고 있는 Q-행렬은 <표 Ⅱ-4>와 같이 나타낼 수 있다. 이를 살펴보면, 문항 1에 정답하기 위해서는 인지요소 1과 인지요소 3이 요구되며 인지요소 2와 인지요소 4, 인지요소 5는 요구되지 않는다는 것을 알 수 있다. 문항 2에 정답하기 위해서 피험자는 인지요소 2만 요구되고, 나머지 인지요소는 요구되지 않는다. 문항 3에 정답하기 위해서는 인지요소 1과 인지요소 4가 요구되며, 문항 4에 정답하기 위해서는 인지요소 3이 요구된다. 문항 5는 인지요소 3과 인지요소 4를, 문항 6은 인지요소 1을 요구한다. 결론적으로 각 문항이 요구하는 인지요소는 각각 다른 것을 알 수 있으며, Q-행렬은 각 문항과 인지요소 사이의 연관성을 나타낼 수 있다.

Q-행렬과 인지진단모형을 통해 피험자의 능력모수인 인지요소 숙달 여부를 추정하고, 추정된 숙달 여부를 기반으로 인지상태를 나타내는 인지요소

프로파일을 생성할 수 있다. <표 Ⅱ-4>와 같이 총 4개의 인지요소를 요구하는 검사를 풀이한 결과, 피험자 A는 인지요소 1, 인지요소 2, 인지요소 3을 숙달하였고 인지요소 4는 숙달하지 못한 것으로 나타났을 경우에는 피험자 A의 인지상태는 [1110]의 벡터로 표현할 수 있으며, 이와 같은 피험자의 인지요소 숙달 패턴은 검사를 통해 측정가능한 모든 인지상태를 표현한다(이현, 2015). 이러한 인지 상태는 인지요소의 개수에 따라 분류되는데, 인지요소의 개수가 K개인 경우, 2^k 개의 인지 상태를 가지게 된다. <표 Ⅱ-3>와 같이 인지요소가 4개인 경우에는 2^4=16개의 인지 상태로 표현된다.



[그림 Ⅱ-1] 문항과 인지요소 간의 관계(김희경, 2013)

<표 Ⅱ-4> Q-행렬의 예(김희경, 2013)

	인지요소1	인지요소2	인지요소3	인지요소4
문항1	1	0	1	0
문항2	0	1	0	0
문항3	1	0	0	1
문항4	0	0	1	0
문항5	0	0	1	1
문항6	1	0	0	0

한편, Q-행렬 작성이 완료되면 인지요소 숙달 여부를 추정하지 않더라도 이론적으로 기대되는 이상적 문항 반응 유형을 작성할 수 있다. 이상적 문항 반응 유형은 인지요소들 간의 관계가 참일 때, 관측 가능한 인지요소 숙달패턴이다(윤지영, 2015). 인지요소의 개수가 K개인 경우, 2^k 개의 인지 상태를 가지게 되는데, 각 인지상태에 따라서 응답할 것으로 기대되는 문항 반응이다. <표 Ⅱ-4>와 같은 Q-행렬을 가졌을 때는 $2^k = 2^6 = 64$ 개의 이상적 문항 반응 유형을 만들 수 있다. 비보상적 관계를 가정하였을 때, 인지요소 1만 숙달한 피험자는 문항 6에만 정답으로 반응할 것으로 기대할 수 있으며, 따라서 이 피험자의 문항 반응은 [000001]로 나타낼 수 있다.

다. 인지진단모형

‘문항 모수 및 피험자 모수를 추정하기 위한 심리측정 모형’을 선택하기 위해서는 인지진단모형의 종류에 대해서 알고 있어야 한다. 인지진단이론 하에서 지난 30여 년간 다양한 인지진단 모형들이 개발되어 수십여 개의 인지진단모형들이 생겨났으며 연구되어 왔다. Rupp & Templin(2010)은 인지진단모형을 <표 Ⅱ-5>와 같이 구분하였다.

<표 Ⅱ-5> 인지진단모형의 분류

관찰된 응답변수	잠재예상변수		모형 유형
	이분 변수	다분 변수	
이분 변수	RSM[1)]		비보상성 모형
	AHM		
	HO-DINA		
	MS-DINA		
	NIDA		
	RERUM		
	BIN	BIN	
	MCLCM	MCLCM	
	Full NC-RUM	Full NC-RUM	
	Recuced NC-RUM	Recuced NC-RUM	
	DINO		보상성 모형
	NIDA		
	BIN		
	MCLCM		
	GDM		
	H-GDM		
	LCDM		
다분 변수	RSM		비보상성 모형
	AHM		
	BIN	BIN	
	MCLCM	MCLCM	
	Full NC-RUM	Full NC-RUM	
	Recuced NC-RUM	Recuced NC-RUM	
	BIN	BIN	보상성 모형
	MCLCM	MCLCM	
	GDM	GDM	
	H-GDM	H-GDM	
	LCDM	LCDM	
	G-DINA	G-DINA	
	G-RUM	G-RUM	
	C-RUM	C-RUM	

출처: Rupp & Templin(2010)

1) RSM, rule-space method; AHM, attribute hierarchy method; BIN; Bayesian inference network; DINA, deterministic inputs, noisy "and" gate; HO-DINA; higher-order DINA';MS-DINA,multistrategy DINA, G-DINA, generalized DINA; DINO, degerministic inputs, noisy "or" gate; NIDA, noisy inputs, deterministic "and" gate; NIDO, noisy inputs, determistic "or" gate; GDM, general diagnostic model; HGDM, hierarchical GDM; MCLCM, multiple classification latent class models; RIM. reparameterized

인지진단모형은 ‘관찰된 응답변수의 척도’, ‘잠재예상변수의 척도’, ‘잠재예상변수의 조합’, 크게 세 가지 기준으로 분류할 수 있다(김명연, 2016; 김성은 외, 2012; Rupp & Templin, 2008). 첫째, 관찰된 응답변수의 척도가 이분변수인지, 다분 변수인지에 따라서 구분된다. 둘째, 잠재예상변수의 척도가 이분 변수인지, 다분 변수인지에 따라서 구분된다. 셋째, 잠재예상변수의 조합이 이루어지는 방식, 즉 측정하고자 하는 인지요소들 간의 관계가 보상적인지, 비보상적인지에 따라서 보상성 모형(compensatory model)과 비보상성 모형(non-compensatory model)으로 구분된다. 예를 들어, 총 5가지 인지요소(‘인식하기’, ‘이해하기’, ‘추론하기’, ‘평가하기’, ‘구성하기’)를 요구하는 영어 검사가 있다. 이 검사의 1번 문항에 정답으로 반응하기 위해서는 ‘이해하기’, ‘추론하기’, ‘구성하기’의 인지요소를 필요로 한다. 보상성 모형에서는 ‘이해하기’와 ‘구성하기’만 숙달하고 있더라도, 두 인지요소가 조합되어 ‘추론하기’의 결핍을 보상할 수 있다고 본다. 따라서 1번 문항이 요구하는 인지요소를 모두 숙달하지 못했더라도 1번 문항에 정답으로 반응할 수 있게 된다. 비결합적 관계가 적용되는 모형은 문항을 풀기 위해서 요구되는 인지요소를 전부 숙달하지 않더라도, 일부 인지요소를 숙달할 경우에는 정답할 확률이 높아지는 특성이 있다(Rupp et al., 2010). 반면, 비보상성 모형에서는 1번 문항에 정답으로 반응하기 위해서는 모든 인지요소를 숙달해야 한다. 즉, ‘이해하기’, ‘추론하기’, ‘구성하기’ 중, 하나의 인지요소라도 미숙달이 발생하였다면 1번 문항에 정답으로 반응할 수 없다. ‘이해하기’와 ‘구성하기’ 인지요소가 ‘추론하기’ 인지요소의 결핍을 보상할 수 없는 것이다. 결합적 관계가 적용된 모형에서는 문항을 풀기 위해서 요구되는 모든 인지요소를 숙달하였을 때 문항에 정답으로 반응할 수 있다(Rupp et al., 2010).

---

unified model/fusion model; C-RUM, compensatory RIM; NC-RUM, noncompensatory RUM; full NC-RUM, NC-RUM with continuous latent interaction term; reduced NC-RUM, NC-RUM without latent interaction term; rERUM, random-effects RUM; LCDM, log-linear cognitive diagnosis model.

다양한 인지진단모형들 중에서 지금까지 가장 많이 연구된 인지진단모형은 DINA 모형이다(은효정, 2017; 김성은, 박윤수, 이영선, 2012; de la Torre, 2009; de la Torre & Douglas, 2004; Huebner & Wang, 2011; Junker & Sijtsma, 2001).

DINA 모형은 다른 인지진단모형에 비해 추정해야 할 모수의 수가 적고, 모수의 해석이 용이하다는 특징이 있어서 여러 연구에서 가장 활발하게 사용되고 있다(반재천, 김선, 2012). DINA 모형은 결합적 규칙을 따르는 대표적인 비보상적 모형이다. 피험자가 각 문항에 정답으로 반응하기 위해서는 문항이 요구하는 모든 인지요소를 숙달해야 한다. 즉, 4개의 인지요소를 요구하는 문항이라면, 4개의 인지요소를 모두 숙달해야 한다. 3개의 인지요소는 숙달하였다고 하더라도 1개의 미숙달된 인지요소로 인해 해당 문항에 오답으로 반응하게 된다. i번째 피험자가 j번째 문항이 요구하는 인지요소 k를 모두 숙달하였으면 η_{ij}=1이 되며, 하나라도 미숙달된 인지요소가 있는 경우에는 η_{ij}=0이 된다. η_{ij}를 추정하기 위한 계산식은 아래의 식(9)와 같다.

$$\eta_{ij} = \prod_{k=1}^{K} (\alpha_{ik})^{q_{jk}} \qquad (9)$$

피험자 I의 인지요소 k에 대한 숙달 여부인 α_{ik}는 피험자가 인지요소 k를 숙달하였으면 1의 값을, 숙달하지 못하였으면 0의 값을 갖는다. 숙달확률을 이용할 경우에는 숙달 여부의 기준이 되는 값인 임계값(threshold)을 설정하는 추가 작업이 필요하다(Huebner & Wnag, 2011; Rupp et al., 2010). q_{jk}는 문항 j에 정답으로 응답하기 위해 필요한 인지요소 k의 필요 여부를 나타낸다. 이는 Q-행렬을 통해 확인할 수 있으며, 인지요소의 숙달을 요구하면 1, 그렇지 않으면 0의 값을 갖는다. X_{ij}는 i번째 피험자의 j번째 문항에 대한 응답 반응으로, X_{ij}는 0 또는 1의 값을 갖는다. 1은 해당 문항에 정답으로 반

응하였음을 의미하고 0은 오답으로 반응하였음을 의미한다. 그러므로 학생의 인지요소 숙달 패턴을 구하기 위한 수식은 아래의 식(10)과 같다.

$$P(X_{ij}=1|\alpha_j,\eta_j) \qquad (10)$$

다음으로, DINA 모형이 추정해야하는 문항모수인 추측(guessing)모수와 부주의 오류(slip)모수를 구하기 위한 식은 아래와 같다.

$$g_j = P(X_{ij}=1|\eta_{ij}=0) \qquad (11)$$

$$s_j = P(X_{ij}=1|\eta_{ij}=1) \qquad (12)$$

g_j는 추측모수로 피험자가 문항 j에 정답으로 반응하기 위해 요구되는 인지요소 모두 숙달하지 못 하였는데도 문항에 정답으로 반응할 확률이며, s_j는 피험자가 문항 j에 정답으로 반응하기 위해 요구되는 인지요소 모두 숙달하였음에도 불구하고 문항 j에 오답으로 반응할 확률을 의미한다. 따라서 피험자 I가 문항 j에 정답할 확률을 구하기 위한 식은 아래와 같다.

$$P(X_{ij}=1|\eta_{ij}) = g_j^{(1-\eta_{ij})}(1-s_j)^{\eta_{ij}} \qquad (13)$$

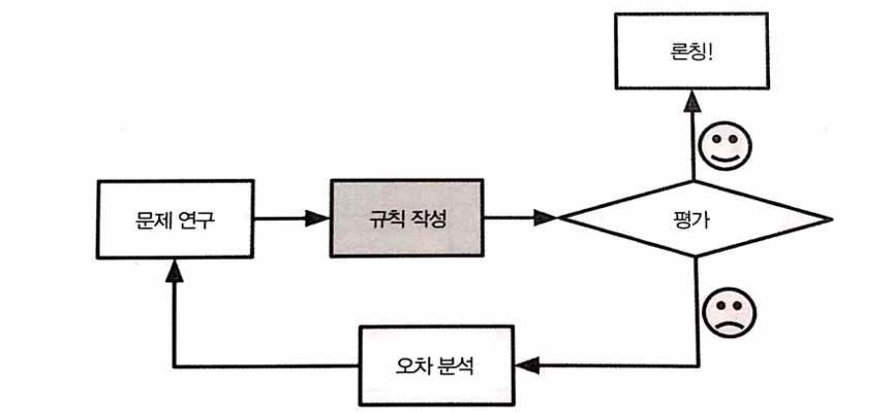
인지진단평가는 위와 같은 추정방식을 통해 검사 자료가 요구하는 인지요소들에 대한 숙달 여부를 진단하고, 학생들의 인지상태에 대한 숙달 패턴을 도출할 수 있다. 교육 현장에서 근무하는 교수자들은 이러한 진단정보를 바탕으로 학생과 교사에게 피드백을 제공함으로써 효과적인 수업활동 및 교육적 처치가 이루어질 수 있도록 도울 수 있다.

3. 머신러닝(Machine Learning)

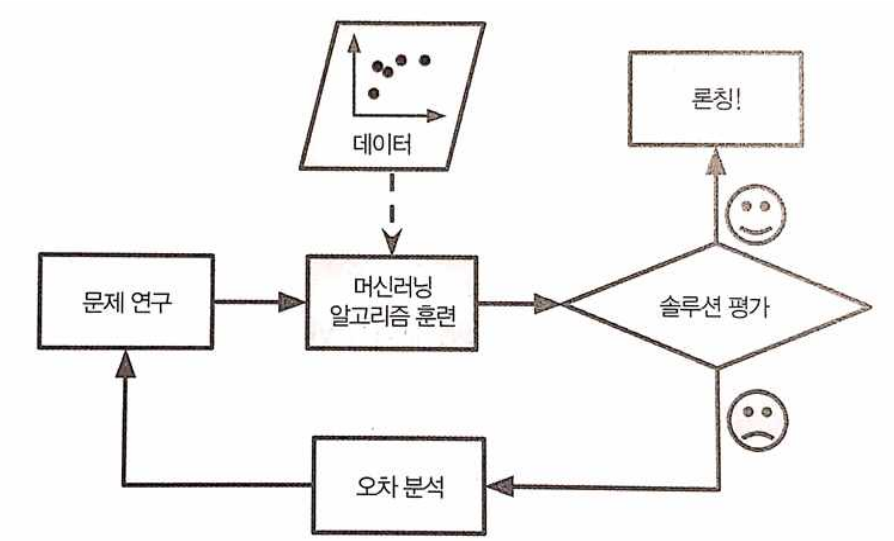
가. 머신러닝

Arthur Lee Samuel(1959)은 머신러닝이라는 단어를 최초로 사용하였으며, 컴퓨터가 스스로 배울 수 있는 능력을 부여하는 방법으로, 데이터를 활용해서 프로그래밍으로 정의되지 않은 동작을 수행하여 결과를 산출하는 방법에 대해 연구하는 분야라고 정의하였다. Tom Mitchell(1997)은 머신러닝의 개념을 작업(task)을 수행하는 성능(performance)이 경험(experience)을 통해 향상되는 것으로 정의하였으며, 기계가 경험으로부터 학습(learning)하면 지능을 가진 것으로 볼 수 있다고 하였다. 예를 들어, 구글(Google)의 딥마인드(DeepMind)가 개발한 바둑 프로그램인 알파고(Alpha Go)의 작업(task)은 바둑을 두는 것이며, 알파고의 성능 측정(performance measure) 방법은 승률을 계산하는 것이다. 알파고가 학습할 수 있는 경험(experience)은 바둑 게임의 기보 혹은 자기 스스로 바둑 게임을 하면서 승리를 경험하는 것이다. 알파고는 이러한 과정을 무수히 반복하면서, 성능인 승률을 증가시킬 수 있도록 학습(learning)한다.

머신러닝이란 데이터를 통해서 스스로 학습하는 프로그램을 의미한다(Géron, A., 2018). 즉, 데이터에 내재된 규칙을 컴퓨터가 스스로 인식하는 것이다. 전통적 프로그램은 사람이 정한 규칙에 따라 결과 값을 산출했다면 머신러닝은 사람이 직접 가르치지 않아도 컴퓨터가 스스로 학습한다. 전통적 프로그래밍 방법과 머신러닝 접근 방법을 그림으로 나타내면 각각 [그림 Ⅱ-2], [그림 Ⅱ-3]와 같이 표현할 수 있다. 컴퓨터는 데이터에서 스스로 규칙을 찾고 이를 수정하는 과정을 반복하면서 데이터를 가장 잘 표현할 수 있는 적합한 모델(함수)을 만든다. 이렇게 기존 데이터를 통한 학습으로 만들어진 프로그램을 모델(model)이라고 부르며, 이러한 모델은 새로운 입력 값이 주어졌을 때 결과를 예측할 수 있다(박해선, 2019).



[그림 Ⅱ-2] 전통적 프로그래밍 방법(Géron, A., 2018)



[그림 Ⅱ-3] 머신러닝 접근 방법(Géron, A., 2018)

이러한 모델을 만드는 과정은 [그림 Ⅱ-4]와 같이 크게 6단계로 구분된다. 1단계는 데이터 수집단계로 분석에 사용할 데이터를 준비하는 단계이다.

필요한 데이터가 무엇인지, 그리고 어떤 데이터를 구할 수 있고 구할 수 없는지를 확인한다. 만약 데이터가 있을 경우에는 데이터베이스에서 필요한 데이터를 추출하여 데이터 병합을 실시하고, 데이터가 없을 경우에는 실험을 통해 수집하거나 구매 또는 오픈 데이터를 활용하게 된다.

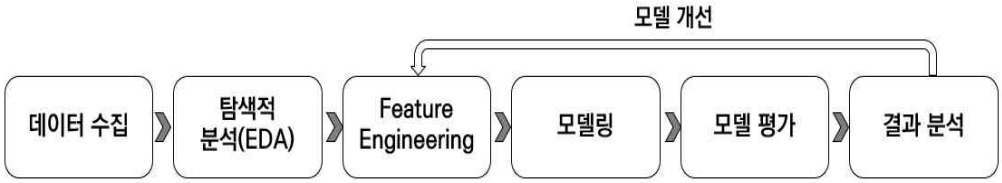
2단계는 탐색적 분석(Exploratory Data Analysis, EDA) 단계로 기본적인 통계분석을 통해 데이터를 이해하는 작업을 수행한다. 종속변수에 영향력이 클 것 같은 변수를 도메인 지식(domain knowledge)을 활용하여 미리 예상하고, 변수 별 기초 통계량 확인, 분포 확인, 변수 간 상관관계 등의 기초 통계분석을 통해 변수의 특징을 파악한다.

3단계는 특성 공학(feature engineering)단계로 데이터 전처리라고도 부른다. 이 단계에서는 이상치 제거, 결측값 처리, 데이터 분포 변환, 데이터 표준화 등의 데이터 가공을 수행함으로써 예측 성능을 높일 수 있다.

4단계는 모델링 단계로 학습용 데이터(train data)와 평가용 데이터(test data)로 분할한다. 학습용 데이터를 이용하여 머신을 훈련시킴으로써 데이터에 적합한 머신러닝 모델을 만든다.

5단계는 모델 평가단계로 학습이 끝난 이후에 모델이 예측한 값과 실제 값을 비교하여 모델의 성능을 평가한다. 이를 통해, 일반화 오차를 평가함으로써 모형이 새로운 데이터에 얼마나 잘 작동하는지 일반화 가능성을 살펴볼 수 있다.

6단계는 결과 분석 단계로 모델의 평가결과를 분석하고 모델을 개선하거나 최종 모델에 대한 결과를 분석하는 단계이다.



[그림 Ⅱ-4] 머신러닝의 모델링 과정

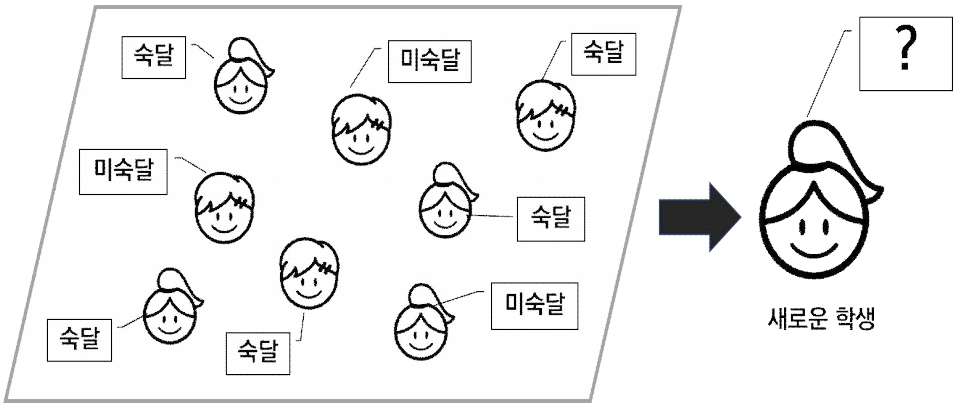
머신러닝은 컴퓨터 프로그램이 훈련하는 방식에 따라 지도 학습(supervised learning), 비지도 학습(unsupervised learning), 강화 학습(reinforcement learning) 세 가지로 구분된다.

1) 지도 학습(Supervised Learning)

지도 학습은 예측에 필요한 변수(독립변수, x)와 예측하고자 하는 변수(종속변수, y)가 모두 존재하는 자료를 사용하여 컴퓨터를 훈련시키는 방법이다(Géron, A., 2018). 예측에 필요한 변수를 입력 특성(input feature), 예측하고자 하는 변수를 출력 특성(output feature) 또는 라벨(label)이라고 한다. 즉, 하나 이상의 입력 특성과 하나의 출력 특성이 서로 쌍을 이루는 자료(paired data)를 사용하는 방법이다. 예를 들어, 공부시간에 따른 기말고사 점수를 예측하는 문제를 해결하고 싶다면 공부시간(x, input feature)과 기말고사 점수(y, output feature)에 대한 자료를 모두 가지고 있어야 한다. 지도 학습에 사용되는 방법은 분류(classification)와 회귀(regression)로 구분된다. 분류는 데이터가 이미 정해진 클래스(class) 중 어느 것에 속하는지를 예측하는 것이고, 회귀는 자료로부터 미래의 특정한 숫자 값을 예측하는 것이다.

인지진단평가 맥락에서 지도학습을 적용해보면 [그림 Ⅱ-5]와 같으며, 이를 분석 가능한 자료의 형태로 나타내면 <표 Ⅱ-6>과 같다. 자료는 입력

특성인 각 학생들의 문항반응과 출력 특성인 인지요소 숙달 여부가 쌍을 이루는 형태로 구성되어 있다.



[그림 Ⅱ-5] 인지진단평가를 위한 지도학습의 적용

<표 Ⅱ-6> 지도학습에 필요한 검사 자료(test data)의 형태

	입력						출력
	문항1	문항2	문항3	문항4	문항5	문항6	인지요소1 숙달 여부
피험자1	1	1	1	1	1	1	1
피험자2	1	1	1	0	0	1	1
피험자3	1	0	1	0	1	0	0
피험자4	1	1	0	0	1	0	1

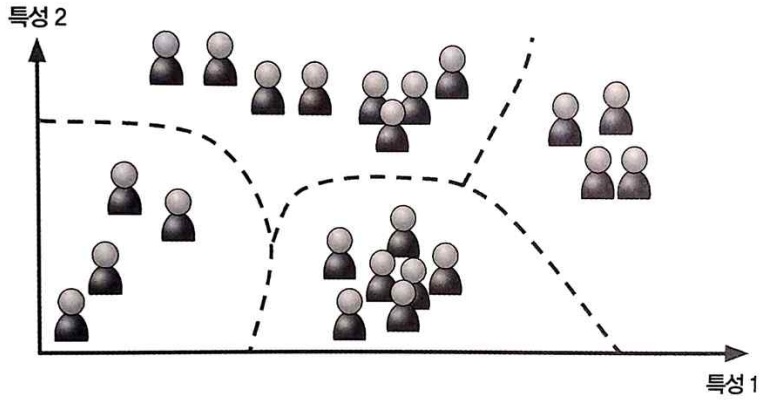
지도 학습을 적용한 교육학 분야의 연구는 학습분석학 측면에서 이루어졌

다. Kim, et al.(2016)은 AOD를 활용한 혼합 학습(blended learning) 환경에서의 학습자 행동로그를 분석하였다. 총 10개의 특성(강좌명, 학생수, 평가기준, 토론 주제 선정 방법, 교수자 역할, 팀 개수, 온라인 토론 방법, 토론 빈도 및 토론기간, AOD구조, 게시글 개수)을 활용하여 고 성취자(high achievers)와 저 성취자(low achievers)를 예측하고자 하였다. 이때, 고 성취자는 최종 학점이 A, B학점을 받은 학생, 저 성취자는 최종 학점이 C인 학생을 의미한다. 고 성취자와 저 성취자의 예측을 위하여 랜덤 포레스트(random forest) 방법을 활용하였으며, 8주차 시점에서의 예측한 학점과 실제 학점을 비교를 통해 랜덤 포레스트 예측의 정확성을 평가하였다. 그 결과, 총 43명의 학생 중 고 성취자로 예측된 28명 모두 고 성취자가 되고 저 성취자로 예측된 11명은 실제로 저 성취자가 되었으므로 랜덤 포레스트의 예측 정확도는 93.6%이라고 보고하였으며, 학습자 행동 로그 데이터만으로도 학습 성과를 비교적 정확히 예측할 수 있다고 주장하였다.

2) 비지도 학습(Unsupervised Learning)

비지도 학습은 정답을 나타내는 라벨(label)이 없는 자료를 사용해서 컴퓨터를 훈련시키는 방법이다(Géron, A., 2018). 예를 들어, 학습 방법에 따라 학습자 유형을 구분하는 것은 비지도 학습이다. 학습분석을 위하여 수집한 자료에는 학습자 유형에 대한 정보가 없기 때문에 군집화(clustering)를 통해 비슷한 학습자를 모아 집단으로 분류해야 하며, 분석을 수행하기 전까지는 어떠한 학습자 유형이 존재하는지 알 수 없다. 훈련 자료에 정답이 존재하지 않기 때문에 지도학습과 같이 정답을 맞히기 위한 패턴을 만드는 것은 불가능하며, 모델의 결과를 평가하기 어렵다는 단점이 있다. 비지도 학습에 사용되는 방법은 군집화(clustering)와 차원 축소(dimensionality reduction)

로 구분할 수 있으며, 군집화는 데이터를 여러 개의 그룹으로 나누는 방법을 의미하고 차원 축소는 자료를 더 적은 차원으로 요약하는 것이다.



[그림 Ⅱ-6] 비지도 학습의 군집(Géron, A., 2018)

이와 관련하여, 이혜윤(2016)은 학습분석학 측면에서 비지도 학습을 활용하였다. 대학 이러닝 환경에서 수집된 학습자 행동로그를 k-평균 군집분석(k-means clustering)방법을 활용하여, 6개의 온라인 행동 참여정도 특성(LMS 접속 빈도, LMS 접속 간격 규칙성, 게시판 활용 시간, 게시판 활용 빈도, 동영상 재생 시간, 동영상 재생 간격 규칙성)에 따라 학습자를 총 3가지 유형으로 구분하였다. 첫 번째 유형은 모든 학습활동에 적극적이고 다른 학습자들과의 상호작용을 중요시하는 '소통형 학습자'이고, 두 번째 유형은 학습영상을 규칙적으로 재생하는 '기본 충실형' 학습자, 세 번째 유형은 가장 저조한 온라인 행동과 비규칙적 학습영상 재생패턴을 보이는 '벼락치기형 학습자'이다.

3) 강화학습(Reinforcement Learning)

강화 학습은 에이전트(컴퓨터)가 수행을 성공 했을 때의 '보상'을 통해 스스로 학습하는 방법이다(Géron, A., 2018). 즉, 에이전트는 환경으로부터의 피드백을 기반으로 행동을 학습한다. 에이전트는 수행을 여러 번 반복함으로써 자신의 수행을 최적화 하고 보상을 최대화 하는 방향으로 학습하며 이러한 최상의 전략을 정책(policy)이라고 부른다.

대표적인 방법으로는 Q-Learning, SARSA, Deep Q Network 등이 있으며, 강화학습을 통해 만들어진 대표적인 인공지능 프로그램으로는 구글 딥마인드가 개발한 바둑 인공지능인 알파고 제로(Alpha Go Zero)가 있다. 알파고 제로는 바둑 규칙이외에 아무 사전지식이 없는 상태에서 바둑 게임(대국)을 통해 승리할 경우 보상을 받고 실패한 경우에는 벌을 받게 된다. 알파고 제로는 스스로 대국을 반복하면서 최고의 보상(maximaized reward), 즉 승률을 높일 수 있는 최선의 수를 학습하게 된다.

나. 머신러닝 방법론

1) k-최근접 이웃(k-Nearest Neighbor, KNN)

k-최근접 이웃은 단순하게 훈련 자료(train data)를 저장하는 방식으로 작동하는 비모수적 패턴인식 방법이다. 새로운 자료가 주어지면 k-최근접 이웃 방법은 새 자료와 가장 가까운 훈련 자료의 포인트를 찾고 이를 새로운 자료의 예측 값으로 산출하는 방식으로 작동한다(Lantz, B., 2017). 즉, 유사하거나 가장 가까운 자료와 동일한 예측값을 부여하는 것이다. k-최근접 이웃은 이해하기 쉽다는 장점 때문에 더 복잡한 방법을 사용하기 전에 시도해 볼 수 있는 머신러닝 방법이다.

k-최근접 이웃에서의 k의 의미는 가장 까가운 이웃의 개수를 의미한다.

훈련 자료(train data)에서 새로운 자료 포인트에 가장 가까운 'k개'의 이웃을 찾는 것이다. 그리고 k개의 이웃들이 분류된 집단(class)을 확인하고, 각각의 집단 중 가장 빈도가 높은 집단을 예측값으로 산출한다.

k-최근접 이웃에서는 크게 두 가지 매개변수(parameter)를 고려한다. 먼저, 자료 사이의 거리이다. 이를 측정하는 방법은 기본적으로 유클리디안 거리 방식을 사용한다. 유클리디안 거리의 공식은 아래의 식(14)와 같다. 다음으로, 이웃의 수는 보통 3개에서 5개 사이를 사용한다. k가 1일 경우에는 각각의 자료 포인트에 대한 영향력이 커지기 때문에 예측의 정확성이 떨어질 수 있으며, k가 너무 많을 경우에는 각각의 자료 포인트에 대한 영향은 감소하지만 중요한 패턴을 무시하게 될 수 있음으로 단순하게 가장 빈도가 높은 클래스를 예측값으로 산출하게 될 가능성이 높아진다.

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)} \qquad (14)$$

2) 의사결정 나무(Decision tree)

의사결정 나무(Decision tree)는 의사결정규칙(decision rule)을 나무구조(tree)로 도표화하여 분류와 예측을 수행하는 방법으로 매우 직관적이며 해석력이 뛰어난 모델을 만들 수 있다는 장점이 있다. 의사결정 나무에서 쓰이는 방법은 분류 기준과 목적함수에 따라 CHAID, CART, C45 등으로 구분된다(Géron, A., 2018).

의사결정 나무에서는 마디(node)라는 용어를 주로 사용하며 마디는 그림의 네모를 의미한다. 뿌리마디(root node)라는 것은 맨 위의 뿌리, 트리 구조가 시작되는 마디를 의미하고 뿌리마디는 전체 자료로 구성되어 있다. 그리고 부모마디(parents node)는 자식마디의 상위마디, 자식마디(child node)

는 하나의 마디로부터 분리되어 나간 2개 이상의 마디들을 의미한다. 끝마디(terminal node) 또는 잎마디(leaf node)는 나무줄기의 끝에 위치하고 있고, 자식 마디가 없는 마디를 의미한다. 가지(branch)는 뿌리마디로부터 끝마디까지 연결된 마디들 의미하며, 깊이(depth)는 뿌리마디로부터 끝마디를 이루는 층의 수를 의미한다.

의사결정 나무의 분석과정은 크게 4단계로 구분된다.

1단계는 의사결정 나무를 만드는 단계로, 분석의 목적과 자료에 따라서 적절한 분리기준(spilt criterion)과 정지규칙(stopping rule)을 찾아서 의사결정 나무를 만드는 것이다. 분리기준이란 변수, 즉 마디를 무엇을 기준으로 선택할 것인지를 의미하며, 목표변수가 범주형 변수일 경우에는 빈도를 기준으로 분리하고 연속형 변수일 경우에는 평균과 표준편차를 기준으로 분리한다. 분리기준으로 가장 많이 쓰이는 지표는 순수도(purity) 또는 불순도(impurity)로 부모마디로부터 자식마디가 형성될 때, 생성된 자식마디에 속하는 자료의 순수도가 가장 많이 증가하도록 의사결정 나무를 형성하게 된다. 즉, 입력변수를 이용해 목표변수의 분포를 얼마나 잘 구별하는지 정도를 순수도와 불순도를 통해서 측정하는 것으로, 목표 변수의 구별 정도를 다양한 범주의 개체들이 포함된 정도인 불순도로 측정하게 된다. 의사결정 나무는 분리 후에 각 영역의 순도(homogeneity)가 증가, 불순도(impurity)가 최대한 감소하도록 하는 방향으로 학습을 진행한다. 다음으로 정지규칙이란 의사결정 나무의 분리를 진행할 때, 더는 분리가 일어나지 않고 현재의 마디가 끝마디가 되게 하는 규칙을 의미한다. 만약 정지규칙이 존재하지 않을 때에는 엔트로피가 0이 될 때까지 분기를 진행하기 때문에 지나치게 많은 마디를 사용하게 되어 학습에 사용하는 자료에만 과적합(over fitting) 될 가능성이 커지기 때문에 적절한 가지치기가 필요하다.

2단계는 가지치기(pruning) 단계로, 분류오류(classification error)를 크게

할 위험이 높거나 부적절한 추론 규칙(induction rule)을 포함하는 불필요한 가지를 제거 및 통합하는 단계이다. 이를 통해 과적합을 방지하고 의사결정 나무의 복잡도를 줄인다.

3단계는 1~2단계를 거쳐서 만들어진 의사결정 나무의 타당성을 평가하는 단계로, 검증용 자료(test data)를 통해 만들어진 의사결정 나무를 평가하게 된다.

마지막 4단계는 의사결정 트리를 해석하고 예측모형을 설정하는 단계이다.

의사결정 나무는 직관적이고 편리하게 해석할 수 있기에 머신러닝에서 많이 사용되지만, 특정 자료에만 잘 작동할 가능성이 높다는 단점이 있다. 이와 같은 문제를 해결하기 위해 등장한 모델이 앙상블을 활용한 랜덤 포레스트 방법이다.

3) 랜덤 포레스트(Random forest)

랜덤 포레스트는 의사결정 나무의 단점을 개선한 방법으로 의사결정 나무를 확장한 것이다. 랜덤 포레스트는 자료의 행(row)을 기반으로 샘플링(sampling)하는 배깅(bagging)과 특징(feature)을 기반으로 샘플링하는 랜덤 패치(random patch), 그리고 의사결정 나무를 혼합하여 사용하는 머신러닝 방법이다. 훈련 자료(train data)를 랜덤하게 추출함으로써 의사결정 나무를 만들 때 각각 다른 자료로 훈련시키는 방법이다. 따라서 랜덤 포레스트는 자료 샘플링과 변수 샘플링을 동시에 적용할 수 있으며, 모든 변수에 대해서 훈련을 하는 일반적인 의사결정 나무와 다르게 자료 및 변수 추출을 통해 랜덤성을 증가시킴으로써 보다 일반화 가능성이 커진다는 장점이 있다(Géron, A., 2018).

랜덤 포레스트가 예측값을 산출하는 분석과정은 크게 4단계이다.

1단계: 전체 자료에서 부트스트래핑(bootstrapping)을 통해 추출된 샘플 자료(sub set)를 생성한다. 부트스트래핑은 중복을 허용하여 복원 추출하는 샘플링 방식을 의미한다.

2단계: 샘플 자료를 통해 의사결정 나무를 생성하고, 1단계와 2단계를 n번 반복해서 진행한다.

3단계: 2단계를 통해서 생성한 n개의 의사결정 나무를 이용해 예측을 진행한다.

4단계: 의사결정 나무의 예측 결과에서 가장 많이 등장하는 결과를 선택하여 최종 예측값으로 선택하거나, 혹은 의사결정 나무의 예측치 평균으로 예측값을 산출한다.

4) 에이다부스트(AdaBoost)

에이다부스트는 적응 부스트(adaptive boost)라는 용어에서 나온 방법으로, 부스팅 방법 중 가장 간단한 모델이다. 비교적 간단한 약한 학습기(weak learner)를 순차적으로 훈련시켜 강한 학습기(strong learner)를 만드는 방법이다. 일반적으로 의사결정 나무를 활용하며, 이전 모델의 오차를 보완하도록 오차에 비례해서 가중치를 부여하여 다음 모델에서 해당 오차를 잘 예측할 수 있도록 한다(Géron, A., 2018).

에이다부스트의 훈련 단계는 크게 4단계로 구분된다.

1단계: 각각의 약한 학습기(weak leaner)에서 학습할 자료를 선택한 후, 모든 자료에 대해 동일한 가중치(D)를 부여한다.

2단계: 학습을 진행한 후에 예측치를 산출하여 오류(ε)를 계산하고 모델에 가중치(α)를 부여한다. 그리고 모델별 가중치를 활용하여 자료 가중치(D)를

부여한다. 즉, 다음 모델이 이전의 잘 예측된 것보다 틀린 것을 잘 예측할 수 있도록 하는 것이다. 각 모델의 오류(ε)와 모델별 가중치(α), 자료 가중치(D)는 아래의 식과 같다. 자료 가중치의 $D_{t,i}$는 t번째 모델에서의 i번째 자료 가중치를 의미하며, $D_{t+1,i}$는 다음 모델에서의 가중치를 의미한다. 만약, 이전 모델의 예측이 맞는 경우에는 가중치를 줄여야하므로 $D_{t,i}e^{-\alpha}$, 예측이 틀린 경우에는 가중치를 높이기 위하여 $D_{t,i}e^{\alpha}$가 된다.

$$\epsilon = \frac{\text{오류 데이터}}{\text{전체 데이터}} \qquad (15)$$

$$\alpha = \frac{1}{2}\ln\left(\frac{1-\epsilon}{\epsilon}\right) \qquad (16)$$

$$D_{t+1,i} = \begin{cases} D_{t,i}e^{-\alpha} & (\text{예측이 맞은 경우}) \\ D_{t,i}e^{\alpha} & (\text{예측이 틀린 경우}) \end{cases} \qquad (17)$$

3단계: 다음 약한 학습기(weak leaner)을 만들 때, 오류가 발생한 자료에 대해 높은 가중치를 부여함으로써 가중치(D)를 갱신한다.

4단계: 손실함수가 0 또는 더 이상 변화하지 않을 때까지, 혹은 모든 모델이 1-3단계 작업을 수행할 때까지 반복한다. 손실함수는 아래의 식과 같으며, $D_{m,i}$는 m번째 모델의 i번째 자료의 가중치를 의미하며,I는 지시함수(indicator function)로써 괄호 안의 조건을 만족하면 1의 값을, 아니면 0의 값을 갖는다. k_m은 개별 모형을 의미하므로, $k_{m(x_i)}$은 m번째 모형에 i번째 자료 x_i를 넣었을 때의 예측값을 의미한다. 이것이 실제값인 y_i와 다르면 지시함수는 1의 값을 가지게 되므로, 해당 자료에 대한 가중치가 손실함수에 더해지게 된다. 따라서 손실함수는 오분류에 대한 가중치를 포함하게 된다.

$$L_m = \sum_{i=1}^{N} D_{m,i} I(k_m(x_i) \neq y_i) \qquad (18)$$

5) 그래디언트 부스팅(Gradient Boosting)

그래디언트 부스팅은 특정 모델을 통해서 예측된 값의 잔차(residual)를 다음 모델을 통해서 예측하여 잔차를 줄여나가는 residual fitting 방법이다. 비교적 간단한 약한 학습기(weak learner) A를 기본 모델(base model)로 사용하여 y_i를 예측하고 남은 잔차를 다음 모델인 B를 통해 예측한다. 이를 반복하며 잔차를 계속 줄여나가게 되며 훈련 자료(train set)를 잘 설명할 수 있는 모델을 만들게 된다(Géron, A., 2018). .

Tree1 예측한 결과값에서 적합선과 자료 간의 잔차를 계산한다. 그 다음 모델인 Tree2는 Tree1에서의 잔차를 자료로 사용한다. 즉, 잔차가 y_i가 되는 것이다. 즉, 앞에서 예측된 모델의 잔차를 다음 모델을 통해서 예측함으로써 잔차를 줄여나가게 된다. Tree1, Tree2, Tree3을 약한 학습기(weak learner), 그리고 이를 결합한 것을 강한 학습기(Strong learner)이라고 부르며, 약한 모델로는 의사결정 나무를 많이 사용한다.

Ⅲ. 연구 방법

1. 연구 자료

가. 실제 검사자료(Real Data)

본 연구에서는 인지진단평가에 머신러닝 방법을 적용하기 위하여, 2014년에 시도교육청 수준에서 중학교 3학년 학생들을 대상으로 실시한 영어 학업성취도 검사 자료를 활용하였다. 본 검사 자료의 피험자 수는 약 22,000명이었고 이 중에서 3,000명의 응답을 무선 표집한 김명연(2016)의 자료를 사용하였다.

검사는 총 18문항으로 구성되어 있으며 검사의 신뢰도는 0.846으로 높게 나타났다. 검사에서 나타난 피험자 원점수인 총점은 최소값 0, 최대값 18, 평균 8.2, 표준편차 4.59로 나타났다.

<표 Ⅲ-1> 고전검사이론 하에서의 문항 모수 및 검사 내용 영역

문항	난이도	변별도	내용	영역
1	0.68	0.60	적절한 의견 말하기	읽기
2	0.50	0.46	적절한 의견 말하기	읽기
3	0.74	0.55	표에 대한 설명하기	읽기
4	0.52	0.48	분위기 찾기	읽기
5	0.48	0.52	적절한 어휘 찾기	읽기
6	0.49	0.60	변화 찾기	읽기
7	0.49	0.50	문장 배열하기	쓰기
8	0.47	0.64	목적 파악	읽기
9	0.29	0.49	문장 넣기	쓰기
10	0.38	0.59	빈칸 넣기	읽기
11	0.35	0.60	글의 요지	읽기
12	0.52	0.58	제목 추론	읽기
13	0.42	0.51	지시대명사의 대상	읽기
14	0.30	0.38	문맥 흐름 파악	쓰기
15	0.40	0.49	제목 추론	읽기
16	0.37	0.47	세부 내용 일치	읽기
17	0.38	0.45	문장 배열하기	쓰기
18	0.43	0.57	본문 내용 요약하기	쓰기

이 검사 자료를 고전검사이론 하에서 분석하였을 때의 문항 모수는 <표 Ⅲ-1>과 같다. 먼저, 문항 난이도를 살펴보면, 쓰기영역의 9번 문항의 난이도가 0.29로 가장 어려운 문항이었으며 읽기영역의 3번 문항의 난이도가 0.74로 가장 쉬운 문항이었다. 쓰기 영역에 해당하는 7번, 9번, 14번, 17번, 18번 문항의 난이도가 0.50을 넘지 않아, 학생들이 대체적으로 쓰기영역을 어려워하는 것을 할 수 있었다. 변별도를 살펴보면 대부분의 문항이 0.3이상으로 양호하게 나타났다. 변별도가 가장 좋은 문항은 8번, 변별도가 가장 낮은 문항은 0.38로 나타난 14번이다.

피험자의 인지요소 숙달 여부를 추정하기 위하여 제작한 영어 검사 자료의 Q-행렬은 아래의 <표 Ⅲ-2>와 같다. 본 연구에서 사용하는 Q-행렬은 김명연(2016)이 내용 전문가(중학교 영어교사 2명, 영어교육 전공 대학원생

2명)의 도움을 받아 작성하였으며, 자카드 계수와 중다회귀 분석을 이용해 타당화한 것이다. 이러한 과정을 바탕으로 도출된 인지요소는 '인식하기', '이해하기', '추론하기', '평가하기', '구성하기'로 총 5개의 인지요소를 가지고 있다.

<표 Ⅲ-2> 영어검사자료의 Q-행렬(김명연, 2016, p.51)

문항	인식하기	이해하기	추론하기	평가하기	구성하기
1	0	1	1	0	1
2	0	1	1	0	1
3	0	1	0	1	0
4	0	1	1	0	0
5	1	0	0	0	0
6	0	1	1	0	0
7	0	1	0	0	1
8	0	1	1	0	0
9	0	1	0	0	1
10	0	1	1	0	0
11	0	1	1	0	0
12	0	1	1	0	0
13	0	1	0	0	0
14	0	1	1	1	0
15	0	1	1	0	0
16	1	1	0	1	0
17	0	1	1	0	1
18	0	1	1	0	1

본 연구에서는 인지진단평가에 머신러닝을 적용하기 위하여, 지도 학습의 분류 기법을 사용하였다. 지도 학습에는 입력(input feature)과 출력(output feature)이 쌍을 이루는 자료가 필요하기 때문에 본 연구에서는 입력으로 피험자들의 문항 반응(item response)을, 출력으로 인지요소 숙달 여부 및 인지요소 패턴을 설정하였다. 또한 머신러닝에서 훈련시키기 위한 자료(train data)로 사용하기 위해 Q-행렬을 참고하여 이상적 문항 반응 유형(ideal

item response pattern)과 인지요소 숙달 여부를 만들었다. 이상적 문항 반응 유형이란 인지상태에 대한 조합이다. 이론적으로 가능한 인지 상태의 조합의 수는 2^k개이며, 본 연구에서는 5개의 인지요소를 사용하였으므로 이에 따른 이상적 문항 반응 유형은 총 32개이다. 아래의 <표 Ⅱ-6>와 같다.

<표 Ⅲ-3> 본 연구의 이상적 문항 반응 유형

	이상적 문항 반응 유형																		인지 상태				
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1
11	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0
12	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	1
13	0	0	0	1	0	1	0	1	0	1	1	1	0	0	1	0	0	0	0	1	1	0	0
14	1	1	0	1	0	1	0	1	0	1	1	1	0	0	1	0	1	1	0	1	1	0	1
15	0	0	0	1	0	1	0	1	0	1	1	1	0	1	1	0	0	0	0	1	1	1	0
16	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1
17	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
18	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
19	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
20	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
21	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
22	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1
23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
24	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
25	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0
26	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	1
27	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1	0
28	0	0	1	0	1	0	1	0	1	0	0	0	1	0	0	1	0	0	1	1	0	1	1
29	0	0	0	1	1	1	0	1	0	1	1	1	1	0	1	0	0	0	1	1	1	0	0
30	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1
31	0	0	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	0	1	1	1	1	0
32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

나. 모의실험 자료

모의실험 자료를 생성하기 위해서 실제 검사자료로부터 추정한 문항모수를 활용하였다. DINA 모형을 적용하여 추정한 문항모수는 <표 Ⅲ-4>와 같으며, 이를 바탕으로 18개 문항에 대한 피험자 3,000명의 문항 반응을 생성하였다. 이를 위하여, WinBUGS를 통한 실제 자료에 대한 모수 추정 과정(문항모수를 고정함)에서 피험자 Burn-In 이후의 반복 자료(replication data)를 생성하여 자료로 보는 전략을 사용하였다. 생성된 모의실험 자료는 <표 Ⅲ-5>와 같다.

<표 Ⅲ-4> DINA 모형을 적용하여 추정한 문항모수

문항	추측모수(g)	부주의 오류 모수(s)
1	0.50	0.01
2	0.37	0.23
3	0.50	0.01
4	0.39	0.20
5	0.20	0.18
6	0.31	0.13
7	0.33	0.21
8	0.26	0.10
9	0.15	0.42
10	0.20	0.24
11	0.17	0.28
12	0.34	0.11
13	0.14	0.29
14	0.21	0.49
15	0.27	0.30
16	0.24	0.36
17	0.26	0.35
18	0.25	0.20

<표 Ⅲ-5> 실제 검사 자료의 문항모수로 생성한 모의실험 자료

	문항 반응 유형																		인지 상태				
1	1	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
2	1	1	0	1	0	0	1	1	1	0	0	0	1	0	0	1	0	0	1	1	0	0	1
3	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	1	0
4	1	1	1	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0	1	1
5	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	0
6	1	1	0	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	1	1	1	0	1
7	1	1	1	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	0	1	1	0
8	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1
9	0	0	1	1	1	0	1	0	1	0	0	1	1	1	0	0	1	0	1	0	1	1	0
10	1	0	0	0	0	0	0	0	1	0	0	1	1	0	1	1	0	0	1	0	1	1	1
11	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0
12	0	0	1	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	1
13	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	1	1	1	0	1	1
14	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
15	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	1	1	0	1
16	1	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	0	1	1	1	1	0
17	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
18	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1
19	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
20	0	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	1
…									…												…		
2995	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1	0
2996	1	1	0	0	0	0	1	0	0	0	0	1	1	1	0	0	1	0	1	1	1	0	1
2997	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0
2998	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0	1
2999	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	1	0
3000	1	1	1	0	1	1	1	0	0	1	1	0	1	0	0	0	0	0	1	1	1	1	1

2. 연구 절차

본 연구에서는 실제 자료 분석과 모의실험 연구를 통해 인지진단평가에 머신러닝 방법을 적용할 수 있는지 그 가능성을 탐색하고자 한다. 다음과 같은 세 종류의 연구 진행을 위해서 김명연(2016)의 5개 인지요소를 갖는 Q-행렬을 사용하였다. 첫째, 머신이 훈련해야할 자료에서 인지숙달 부분을 DINA 모형을 통해 추정한 결과로 사용하였다(연구문제1). 둘째, 이론적 자료인 이상적 문항 반응 유형의 사용가능성을 알아보기 위하여, Q-행렬을 바탕으로 인지 상태에 대한 이상적인 응답 반응을 생성하였다(연구문제2). 셋째, 모의실험 연구로서 검사 자료와 이상적 반응 유형을 각각 훈련 자료로 사용하여 머신러닝을 적용하고, 머신러닝 방법에 따른 분류 정확성을 확인하였다(연구문제3).

각 연구의 구체적인 분석 절차는 다음과 같다.

가. 실제 검사자료 분석 절차

1) 자료 크기에 따른 DINA 모형과 머신러닝 일치도 분석

인지진단평가에 머신러닝 방법의 적용 가능성을 탐구하기 위해서는 머신러닝 방법을 기존의 인지진단모형인 DINA 모형과 분석 결과 측면에서 비교해야할 필요가 있다. 머신러닝을 적용하기 위하여 다음과 같이 입력 자료(input data)를 구성하였다. 첫째, 머신러닝에서의 훈련을 위하여 비보상 모형인 DINA 모형을 바탕으로 피험자 모수인 인지요소 숙달 여부를 추정하였다. 실제 검사자료의 문항 반응과 DINA 모형을 통해서 추정한 피험자 모수를 머신의 훈련 자료(train Data)와 평가 자료(test Data)로 활용하였다. 둘째, 훈련 자료의 크기를 고려하여 입력 자료를 구성하였다. 대부분의 학교 현장에서는 큰 표본을 구하기가 어렵기에, 이러한 점을 고려하여 머신러닝

적용에 어느 정도 크기의 훈련 자료가 필요한지를 알아보고자 전체 자료의 크기와 관련하여 100, 300, 500, 1000, 3000 총 5가지 조건을 설정하였다. 이렇게 구축한 머신러닝으로 추정한 피험자 모수와 DINA 모형이 추정한 피험자 모수 간의 일치도를 분석해봄으로써 머신러닝 방법의 적용가능성을 탐색해보고자 하였다.

1단계: 실제 검사자료와 이에 대한 Q-행렬을 바탕으로 DINA 모형을 적용하여 피험자 능력 모수인 인지요소 숙달 여부를 추정한다.

2단계: 실제 검사자료의 피험자 문항반응을 input feature로 하고, 1단계에서 DINA 모형으로 추정한 피험자 능력모수를 label로 사용하여 머신러닝 방법을 적용하기 위한 훈련 자료(train data)와 평가 자료(test data)로 분할한다.

3단계: 자료를 100, 300, 500, 1000, 3000 총 5가지 조건에 따라 랜덤 추출하여 각 표본 크기에 따라 구분한다.

4단계: 머신러닝을 적용하기 위하여, 각 자료를 훈련 자료(train data)와 평가 자료(test data)로 분할한다(분할비율: 70/30).

5단계: k-nearest neighbor(KNN), Decision tree, Random forest, AdaBoost, Gradient boosting 방법을 활용하여 머신러닝을 적용한다.

본 연구에서 사용한 머신러닝 방법의 하이퍼파라미터(hyper-parameter) 설정은 아래의 <표 Ⅲ-6>과 같으며, 그 외의 하이퍼파라미터는 기본 값(default)을 갖는다.

6단계: 평가용 자료를 사용하여, DINA 모형으로 추정된 피험자 능력모수와 머신러닝 방법을 통해 추정된 피험자 능력모수와의 일치도를 살펴본다.

<표 Ⅲ-6> 하이퍼 파라미터 설정

머신러닝 방법	하이퍼 파라미터
k-nearest neighbor (KNN)	n_neighbors=4
Decision tree (DT)	random_state=0
Random forest (RF)	n_estimators=100, random_state=0
AdaBoost(Ada)	n_estimators=100, random_state=0
Gradient boosting (GB)	n_estimators=100, random_state=0

2) 이상적 문항 반응 유형의 크기에 따른 DINA 모형과 머신러닝 일치도 분석

인지진단평가에 머신러닝 방법의 적용 가능성을 높이기 위하여, 작성된 Q-행렬을 바탕으로 인지요소와 문항 간의 관계를 고려하여 각 인지상태가 요구하는 이상적 문항 반응 유형을 생성하였다. 이를 머신의 훈련 자료(train data)로 사용하여 머신러닝을 적용하고, 실제 검사자료를 평가 자료(test data)로 사용한 머신러닝 방법을 평가하였다. 이때, 이상적 반응 유형의 적용 방식을 살펴보기 위하여 이상적 문항 반응 유형 자료의 크기에 따라 인지요소 숙달 여부의 분류 일치도에 어떠한 차이가 있는지를 살펴보았다. 분석 절차를 각 단계로 나타내면 다음과 같다.

1단계: 검사 자료가 요구하는 5개 인지요소와 문항의 관계를 나타낸 Q-행렬을 바탕으로, 이상적 문항 반응 유형을 Input Feature로 하고, 인지 상태

를 Label로 하는 총 32개의 이상적 문항 반응 유형을 생성한다.

2단계: 32개 인지요소에 대한 이상적 반응유형을 동일 비율로 늘려 총 3가지 조건(32, 3200, 32000)의 훈련 자료(train data)를 생성한다.

3단계: k-nearest neighbor(KNN), Decision tree, Random forest, AdaBoost, Gradient boosting 방법을 활용하여 머신러닝을 적용한다. 하이퍼 파라미터(hyper-parameter) 설정은 이전과 동일하다.

4단계: 실제 검사 자료를 평가 자료(test data)로 사용하여, 이상적 반응유형을 사용하여 훈련한 머신러닝 방법과 DINA 모형으로 추정된 피험자 능력모수와의 일치도를 살펴본다.

나. 모의실험 자료 분석 절차

1) 머신러닝 방법에 따른 분류 정확성 분석

머신러닝의 피험자 모수 추정 정확도를 살펴보기 위하여 모의자료를 활용한다. 모의자료는 실제 검사 자료의 문항 모수(geussing, slip)를 활용하여, 각 문항에 대한 피험자의 문항 반응을 생성하였다. 이 모의실험 자료를 머신러닝 방법으로 분석하여, 피험자의 능력모수인 각 인지요소에 대한 숙달 여부와 인지상태를 추정한다. 이때, 연구문제2의 분석 결과를 활용하여 머신러닝에 적합한 이상적 반응 유형 자료의 크기를 반영한다. 모의실험 자료를 통해 진능력 모수의 복원력을 살펴봄으로써 머신러닝 방법의 분류 정확도를 확인해 볼 수 있다. 정확도를 분석하기 위해 머신러닝 분류모형의 평가방법인 혼동행렬(confusion matrix)을 활용한다.

1단계: 실제 검사자료에 DINA 모형을 적용하여 문항 모수(geussing, slip)

를 추정 한다.

2단계: 1단계에서 추정된 문항모수를 생성모수로 사용하여, 가상의 피험자 문항 반응과 인지상태가 포함된 모의실험 자료를 생성한다.

3단계: 모의실험 자료에 머신러닝 방법을 적용하여 피험자 능력모수를 추정한다.

4단계: 3단계에서 추정한 능력모수와 진능력모수를 비교하여, 추정된 숙달 여부와 진숙달 여부가 얼마나 일치하는지를 살펴봄으로써 능력 모수 복원관점에서 머신러닝 방법을 평가한다.

5단계: 이를 바탕으로 인지진단평가를 위한 머신러닝 방법의 적용 가능성을 탐색해본다.

다. 머신러닝의 능력모수 추정결과에 대한 평가

머신러닝을 적용하여 피험자 능력모수 추정의 정확성을 평가하기 위한 기준으로 아래와 같은 방법을 사용하였다.

머신러닝은 훈련 과정에서 자료에서 패턴을 찾아 스스로 구성되므로 다른 통계적 모형을 평가할 때 사용되는 적합도 지수를 사용할 수 없다. 그 대신에 분류 결과를 바탕으로 성능 준거(performance measure)를 사용한다(이창묵, 2019). 따라서 본 연구에서는 진숙달 여부와 추정된 숙달 여부 간의 분류 정확도를 확인하기 위하여 혼동행렬(confusion matrix)을 사용하였다. 혼동행렬은 분류 모형의 전반적인 성능을 보여주는 지표로써 널리 활용되는 모형 평가 방법이다(권철민, 2019).

<표 Ⅲ-7> 혼동행렬

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual egative	False Positive (FP)	True Negative (TN)

실제 검사자료와 모의실험 자료를 머신러닝으로 분석한 후에 그 결과를 해석한다. 훈련된 머신에 모의실험 자료의 피험자 문항 반응을 입력(input data)으로 넣었을 때 출력(output data, label)으로 나오는 인지요소 숙달 여부를 산출하고, 이를 모의실험 자료의 진숙달 여부(참값)와 비교하여 일치 정도를 확인해본다.

혼동행렬의 의미를 해석해보면 Positive는 머신의 출력 결과가 숙달(1)인 경우를 의미하고, Negative는 미숙달(0)인 경우를 의미한다. True는 머신의 출력 결과와 진숙달 여부가 일치하는 경우를 의미하고, Fasle는 일치하지 않는 경우를 나타낸다. 그러므로 True Positive는 머신이 숙달로 추정하여 진숙달 여부와 일치하는 것이고, True Negative는 머신이 미숙달로 추정하여 일치하는 것이다. 반대로 False Positive는 머신이 숙달로 추정하여 진숙달 여부와 일치하지 않는 것이고, True Negative는 머신이 미숙달로 추정하여 일치하지 않는 것을 의미한다. 이에 본 연구에서는 일치 정도를 확인하기 위하여, 머신이 올바르게 추정한 비율을 살펴보고자 한다. 즉, 혼동행렬에서 True에 해당하는 값인 True Positive와 True Negative를 중심으로 평가하며, 이를 정확도(Accuracy)라고 부른다.

$$\text{정확도(Accuracy)} = (TP + TN) / (TP + TN + FP + FN)$$

Ⅳ. 연구 결과

본 연구는 실제 자료 분석과 모의실험 자료 분석을 바탕으로, 인지진단평가를 위해 머신러닝 방법을 적용할 수 있는지를 확인해보고자 하였다. 이를 위하여 아래와 같이 분석을 실시하였다.

첫째, 실제 검사 자료와 DINA 모형으로 추정한 피험자의 능력모수를 활용하여 머신러닝을 실시하였다. 머신러닝 방법으로 피험자의 능력모수를 추정하고, 이를 DINA 모형으로 추정한 피험자 능력모수와 비교함으로써 인지요소 숙달 여부의 정확 일치도를 살펴보았다. 이를 통해, 머신러닝의 인지진단평가 활용 가능성을 확인하였다. 한편, 피험자 모수 추정을 위하여 머신러닝을 사용하였을 경우에 검사 자료의 크기에 따라 결과에 어떠한 차이가 있는지를 확인해보았다.

둘째, 이상적 반응 유형을 사용하여 머신러닝을 적용하는 것이 가능한지를 탐색해보았으며, 훈련 자료(train data)의 크기에 따라서 피험자 능력모수인 인지요소 숙달 여부 분류 결과에 차이가 있는지를 실제 검사 자료 분석을 통해 확인해보았다.

셋째, 사용한 머신러닝 방법에 따라 피험자 능력 모수 추정의 정확성에 차이가 있는지를 알아보기 위하여 모의실험 자료를 사용하였다. 머신러닝 방법으로는 전통적인 KNN, Decision Tree 뿐만 아니라, 앙상블 방법을 사용한 Random Forest, Ada Boost, Gradient Boost 등을 다양하게 활용하여 머신러닝을 적용하였다. 각각의 모형이 추정한 피험자의 인지요소 숙달 여부 분류 정확도를 비교하여 어떤 방법이 가장 정확한지를 확인해보았다.

1. 실제 검사 자료 분석

가. 자료 크기에 따른 DINA 모형과 머신러닝 일치도 분석

실제 검사자료를 활용하여 100, 300, 500, 1000, 3000명의 총 5개 피험자 조건을 고려하여 학습한 머신러닝 방법의 피험자 모수 추정결과와 DINA 모형의 피험자 모수 추정결과 간의 비교 결과는 <표 Ⅳ-1>에 제시하였다.

분석 결과, 전반적으로 머신러닝으로 추정한 피험자 모수와 DINA 모형으로 추정한 피험자 모수가 상당히 일치하는 것으로 나타났다. 5개 인지요소의 숙달 여부 추정결과는 최소 85%, 최대 100%의 정확 일치도를 보이는 것으로 나타났으며, 인지상태 추정에 있어서는 최소 65%, 최대 95%의 정확 일치도를 보였다.

자료 크기에 따른 일치도는 기존 인지진단모형과 같이 피험자 수가 많아질수록 높은 것으로 나타났다. 인지요소 2, 인지요소 3, 인지요소 5, 인지상태 모두 피험자수가 3,000명일 때 가장 높은 일치도를 보였다. 피험자 수에 따른 일치도의 평균을 살펴보면, k-최근접 이웃(KNN)은 각각 83.2%, 87.3%, 87.5%, 89.2%, 91%로 증가하였으며, 의사결정 나무(DT)도 87.2%, 92.3%, 93.3%, 94.3%, 95.6%으로 나타났다. 앙상블방법을 활용한 랜덤 포레스트형(RF)의 일치도는 각각 98.8%, 94%, 95.1%, 96.4%, 97.2%였으며, 에이다부스트(Ada)의 정확도는 90.8%, 90.9%, 90.7%, 91.3%, 91.1%로 나타났다. 마지막으로 그래디언트 부스팅(GB)의 정확도는 90.7%, 95.3%, 95.9%, 97.2%, 97.9%로 피험자 수가 늘어날수록 일치도도 점점 증가하였다.

머신러닝 방법 중, 평균적으로 가장 좋은 성능을 보인 모형은 그래디언트 부스팅(Gradient Boosting, GB)으로 나타났다. 모든 피험자 수 조건에서 가장 높은 일치도를 보이는 것으로 나타났다.

<표 Ⅳ-1> 자료 크기에 따른 일치도

Data	인지 요소	머신러닝 방법	(n=100)	(n=300)	(n=500)	(n=1,000)	(n=3,000)
			정확도 (Accuracy)				
Real Data	인지 요소1	KNN	0.850	0.867	0.914	0.915	0.949
		DT	1.000	1.000	1.000	1.000	1.000
		RF	1.000	1.000	1.000	1.000	1.000
		Ada	1.000	1.000	1.000	1.000	1.000
		GB	1.000	1.000	1.000	1.000	1.000
	인지 요소2	KNN	0.810	0.880	0.874	0.902	0.909
		DT	0.880	0.907	0.936	0.936	0.960
		RF	0.910	0.957	0.964	0.972	0.977
		Ada	0.940	0.963	0.944	0.951	0.955
		GB	0.920	0.977	0.968	0.982	0.983
	인지 요소3	KNN	0.930	0.957	0.958	0.959	0.945
		DT	0.910	0.913	0.908	0.930	0.937
		RF	0.940	0.957	0.954	0.960	0.965
		Ada	0.960	0.953	0.970	0.976	0.976
		GB	0.920	0.947	0.946	0.969	0.974
	인지 요소4	KNN	0.920	0.937	0.924	0.943	0.960
		DT	0.990	1.000	1.000	1.000	0.999
		RF	0.960	0.977	0.996	0.996	1.000
		Ada	0.990	1.000	1.000	1.000	0.999
		GB	0.990	1.000	1.000	1.000	0.999
	인지 요소5	KNN	0.830	0.883	0.882	0.895	0.909
		DT	0.760	0.900	0.902	0.943	0.964
		RF	0.870	0.903	0.924	0.952	0.969
		Ada	0.880	0.883	0.898	0.920	0.921
		GB	0.900	0.913	0.940	0.951	0.960
	인지 상태	KNN	0.650	0.717	0.700	0.737	0.789
		DT	0.690	0.817	0.854	0.849	0.877

Data	인지 요소	머신러닝 방법	(n=100)	(n=300)	(n=500)	(n=1,000)	(n=3,000)
			정확도 (Accuracy)				
		RF	0.710	0.847	0.868	0.901	0.923
		Ada	0.680	0.657	0.628	0.628	0.617
		GB	0.710	0.880	0.902	0.931	0.956
	평균	KNN	0.832	0.873	0.875	0.892	0.910
		DT	0.872	0.923	0.933	0.943	0.956
		RF	0.898	0.940	0.951	0.964	0.972
		Ada	0.908	0.909	0.907	0.913	0.911
		GB	0.907	0.953	0.959	0.972	0.979

그러나 머신러닝과 DINA 모형의 피험자 모수 추정 결과가 상당히 일치한다고 하더라도, 실제 교육현장에 적용 가능성은 매우 낮은 편이다. 이러한 방식으로 머신러닝을 적용하기 위해서는 학습을 위한 충분한 검사 자료가 존재해야 하며, DINA 모형으로 추정된 인지요소 숙달 여부가 존재해야 한다. 따라서, 본 연구에서는 Study1의 결과를 통해 머신러닝 방법이 DINA 모형과 유사한 결과를 산출한다는 것을 살펴보며 머신러닝의 적용 가능성을 확인하였기에, Q-행렬을 바탕으로 생성한 이론적 자료인 이상적 문항 반응 유형(ideal item response pattern)을 활용하여 머신러닝을 적용하고 이를 실제 검사자료를 바탕으로 평가해보고자 한다.

나. 이상적 문항 반응 유형의 크기에 따른 DINA 모형과 머신러닝 일치도 분석

이상적 문항 반응 유형(ideal item response pattern)를 훈련 자료(train data)로 사용하여 머신러닝을 적용하였으며, 실제 검사 자료를 평가 자료(test data)로 사용하여 모형을 평가하였다. 이때, 훈련 자료(train data)의 크기에 따라서 피험자 능력모수 추정의 정확도에 차이가 있는지를 확인해보기 위하여 훈련 자료의 크기를 32, 3200, 32000 3가지 조건으로 구성하였다. 머신러닝 방법의 인지요소 숙달 여부 추정의 일치도는 <표 Ⅳ-2>에 제시하였다.

분석 결과, 이상적 반응 유형을 사용하였을 때, 훈련 자료의 크기는 큰 영향을 미치지 않는 것으로 나타났으며, 오히려 이론적인 이상적 반응 유형 32개만을 사용하였을 때 가장 높은 일치도를 보이는 것으로 나타났다. 이상적 반응 유형 32개로 학습한 머신러닝 방법의 평균 일치도는 k-최근접 이웃(KNN) 82%, 의사결정 나무(DT) 60%, 랜덤 포레스트(RF) 54%, 에이다부스트(Ada) 53.4%, 그래디언트 부스팅(GB) 54%로 나타났으며, 3,200개로 학습한 머신러닝 방법의 평균 일치도는 k-최근접 이웃(KNN) 62.8%, 의사결정 나무 모형(DT) 59.3%, 랜덤 포레스트(RF) 53.8%, 에이다부스트(Ada) 56.3%, 그래디언트 부스팅(GB) 54.1%로 나타나 일치도가 조금씩 감소한 것으로 나타났다, 마지막으로 32,000개의 이상적 반응 유형으로 학습한 머신러닝 방법의 일치도는 k-최근접 이웃(KNN) 68.4%, 의사결정 나무(DT) 58.8%, 랜덤 포레스트(RF) 54.1%, 에이다부스트(Ada) 58.2%, 그래디언트 부스팅(GB) 54.2%로 나타나 3,200개를 사용하였을 때보다는 소폭 상승하였으나 기본적인 32개와는 일치도가 조금 떨어지거나 비슷한 것으로 보인다.

이는 이론적 응답 유형의 크기는 머신러닝의 피험자 모수 추정의 정확도

에 큰 영향을 미치지 않는 것으로 해석할 수 있다. 따라서 이상적 반응 유형을 추가적으로 늘리는 작업 없이 그대로 사용하면 될 것으로 보인다.

각각의 인지요소 숙달 여부를 분류하기 위해서는 k-최근접 이웃(KNN)이 가장 적합한 것으로 보인다. 그 이유는 k-최근접 이웃(KNN)이 다른 방법보다 비교적 단순한 방법으로 머신러닝을 위해 요구하는 자료의 수가 많지 않기 때문으로 보인다. 반면, 인지상태를 나타내는 것은 각각의 인지요소보다 복잡한 분류 문제이기 때문에 그 일치도가 낮게 나타난 것으로 예상된다. 특히, 랜덤 포레스트는 인지상태를 분류함에 있어 그 일치도가 23.4%로 굉장히 낮게 나타났다. 이는 부트스트랩을 이용하여 다양한 의사결정 나무를 생성하고, 이 의사결정 나무를 종합하여 투표하는 방식으로 작동하는 특성 때문인 것으로 보인다. 이와 관련하여, 그리드 서치(grid search), 랜덤 서치(random search)와 같은 방법을 활용하여 랜덤포레스트 방법의 하이퍼 파라미터(hyper parameter) 탐색을 통해 인지상태 진단에 가장 적절한 최적 하이퍼 파라미터(hyper parameter)를 찾는 것이 필요하다.

<표 Ⅳ-2> 이상적 문항 반응 유형 크기에 따른 일치도

Test Data	인지 요소	머신러닝 방법	train data=32	train data=3,200	train data=32,000
			정확도 (Accuracy)		
Real Data	인지 요소1	KNN	0.784	0.875	0.875
		DT	1.000	1.000	1.000
		RF	1.000	0.979	0.991
		Ada	1.000	1.000	1.000
		GB	1.000	1.000	1.000
	인지 요소2	KNN	0.926	0.740	0.740
		DT	0.786	0.794	0.794
		RF	0.729	0.711	0.724

Test Data	인지 요소	머신러닝 방법	train data=32	train data=3,200	train data=32,000
			정확도 (Accuracy)		
		Ada	0.854	0.847	0.855
		GB	0.861	0.861	0.861
	인지 요소3	KNN	0.974	0.880	0.880
		DT	0.684	0.796	0.767
		RF	0.709	0.741	0.735
		Ada	0.472	0.475	0.635
		GB	0.554	0.554	0.554
	인지 요소4	KNN	0.787	0.590	0.543
		DT	0.418	0.418	0.418
		RF	0.415	0.392	0.391
		Ada	0.344	0.404	0.409
		GB	0.418	0.418	0.418
	인지 요소5	KNN	0.906	0.666	0.879
		DT	0.696	0.712	0.649
		RF	0.611	0.617	0.640
		Ada	0.544	0.603	0.522
		GB	0.563	0.572	0.572
	인지 상태	KNN	0.507	0.262	0.378
		DT	0.437	0.243	0.313
		RF	0.234	0.229	0.216
		Ada	0.456	0.487	0.487
		GB	0.303	0.302	0.305
	평균	KNN	0.820	0.628	0.684
		DT	0.604	0.593	0.588
		RF	0.540	0.538	0.541
		Ada	0.534	0.563	0.582
		GB	0.540	0.541	0.542

2. 모의실험 자료 분석

가. 머신러닝 방법에 따른 분류 정확성 분석

모의실험 자료를 이용하여 머신러닝을 적용한 인지진단평가 방식의 유효성을 살펴보고자 한다. 머신러닝을 적용하기 위한 훈련 자료(train data)로 이론적으로 가능한 이상적 응답 유형(ideal response pattern)과 인지요소 숙달 여부 및 인지상태를 사용하였다. 그리고 모의실험 자료의 문항 반응을 훈련된 머신에 투입하여 피험자 모수인 인지요소 숙달 여부와 인지상태를 추정하였다. 이를 모의실험 자료의 진능력모수와 비교함으로써 머신러닝의 정확도를 살펴보았다. 그 결과는 <표 Ⅳ-3>에 제시하였다.

분석 결과, 인지요소 1의 분류 정확도는 k-최근접 이웃(KNN)은 64.6%, 의사결정 나무(DT)와 랜덤 포레스트(RF), 에이다 부스트(Ada)의 정확도는 모두 83%로 나타났으며, 그래디언트 부스팅(GB)의 정확도는 64.8%로 나타났다. 인지요소 2의 경우에는 그래디언트 부스팅(GB)의 정확도가 60.1%로 가장 낮았으며, k-최근접 이웃(KNN)이 84.9%로 가장 높았다. 인지요소 3의 경우에는 에이다 부스트(Ada)의 정확도가 72.4%로 가장 높았으며, 의사결정 나무(DT)의 정확도가 54.3%로 가장 낮았다. 인지요소 4의 결과는 의사결정 나무(DT)와 그래디언트 부스팅(GB)이 69.7%로 가장 높았고, k-최근접 이웃(KNN)의 정확도가 42,4%로 가장 낮은 것으로 나타났다. 인지요소 5는 k-최근접 이웃(KNN)의 정확도가 66.7%로 가장 높고, 의사결정 나무(DT)가 59.5%로 가장 낮았다. 반면 인지상태에 대한 추정 결과는 k-최근접 이웃(KNN)의 정확도가 50.7%로 가장 높았으며, 그 다음으로 정확도가 높았던 것은 랜덤 포레스트(RF)로 23%였다. 나머지 모형들은 15~18%의 정확도를 보였다.

<표 Ⅳ-3> 머신러닝 방법에 따른 분류 정확도

test data	인지 요소	머신러닝 방법	train data=32
			정확도 (Accuracy)
Simulation Data	인지 요소1	KNN	0.646
		DT	0.803
		RF	0.803
		Ada	0.803
		GB	0.803
	인지 요소2	KNN	0.849
		DT	0.745
		RF	0.687
		Ada	0.824
		GB	0.829
	인지 요소3	KNN	0.616
		DT	0.543
		RF	0.703
		Ada	0.724
		GB	0.721
	인지 요소4	KNN	0.424
		DT	0.697
		RF	0.693
		Ada	0.637
		GB	0.697
	인지 요소5	KNN	0.667
		DT	0.595
		RF	0.664
		Ada	0.640
		GB	0.649
	인지	KNN	0.507
		DT	0.230

test data	인지 요소	머신러닝 방법	train data=32
			정확도 (Accuracy)
	상태	RF	0.151
		Ada	0.158
		GB	0.185
	평균	KNN	0.613
		DT	0.596
		RF	0.546
		Ada	0.597
		GB	0.616

Ⅴ. 결론 및 논의

1. 결론 및 논의

본 연구는 최근 여러 분야에 활발하게 적용되고 있는 머신러닝 방법을 교육 분야, 그중에서도 교육평가 분야에 적용하고자 하는 시도를 목적으로 한다. 교육 측정 및 평가 분야에서 새로운 연구 방법으로 머신러닝의 적용 가능성을 탐색하고 교육평가 분야의 머신러닝 연구에 대한 기초 자료를 제공하고자 하였다.

실제 학교 현장에서 인지진단평가의 활용 가능성을 높일 수 있도록 단위학교 내 소규모 표본 상황에서의 머신러닝 적용에 대해서 탐색해보고자 이상적 문항 반응 유형을 훈련 자료로 고려하였다. 이상적 문항 반응 유형을 사용할 경우에는 각 인지상태가 요구하는 이상적 문항 반응 유형을 사용할 수 있음으로 단위 학교 내에서도 학생의 능력 모수인 인지요소 숙달 또는 미숙달 여부를 추정해볼 수 있다는 장점이 있다. 기존의 인지진단평가의 제한점이었던 대규모 표본에서의 적용이라는 한계를 극복하여, 정부 주도 하의 검사뿐만 아니라 단위학교 내에서 이루어지는 검사에서도 학생의 인지상태를 진단하고 교수·학습을 개선하기 위한 정보를 제공할 수 있다.

연구 문제를 확인하기 위하여 크게 두 가지 자료로 분석을 진행하였다.

실제 검사 자료(real data)인 국가 수준 학업 성취도 평가의 영어 검사 자료를 사용하였다. 영어 검사는 총 18문항으로 구성되어 있으며, 총 5개의 인지요소를 가지고 있다. 인지진단평가를 위한 머신러닝 방법의 적용 가능성을 확인하기 위하여, 인지진단평가 모형인 DINA 모형을 통해 추정한 피험자 모수와 머신러닝을 통해 추정한 피험자 모수와의 일치도를 분석하였다. 피험자 모수는 5개의 인지요소의 숙달 여부를 의미한다. 또한 자료의 크기에 따라 피험자 모수 추정의 일치도에 차이가 있는지를 확인해보기 위하여,

표본 크기를 100, 300, 500, 1000, 3000으로 설정하여 자료 크기에 따른 일치도를 확인하였다. 그 결과, 피험자 수가 증가할수록 피험자 능력 모수 추정 결과의 일치도 높아지는 것으로 나타났다. 그러나 지도학습의 경우에는 머신러닝을 위한 정답 자료인 라벨(label), 즉 검사자료로 추정한 인지요소 숙달 여부가 존재해야 한다. 따라서 기존 DINA 모형을 사용하여 추정한 피험자 모수가 존재하는 자료가 없을 때는 활용하기 어렵다는 제약이 존재한다. 실제 학교 현장에는 DINA 모형을 적용하여 그 결과를 산출한 경우가 거의 존재하지 않음으로, 머신러닝 방법이 DINA 모형과의 일치도가 높다고 하더라도 현실적으로 적용하기 힘들다는 단점이 있다.

이를 고려하여 Q-행렬을 기반으로 생성할 수 있는 이론적인 자료인 이상적 문항 반응 유형을 훈련 자료로 사용하여, 머신러닝을 적용하였다. 그리고 실제 검사 자료인 영어 검사 자료를 사용하여 머신러닝 방법을 평가한 결과, 이상적 문항 반응 유형의 크기는 머신러닝 방법의 인지요소 숙달 여부 추정 일치도에 큰 영향을 미치지 않는 것으로 나타났다.

또한 실제 영어 검사자료의 문항 모수를 활용하여 모의실험 자료를 생성하고, 이를 머신러닝 방법으로 분석하여 피험자의 인지요소 숙달 여부를 추정하였다. 모의실험 자료 분석을 통해 머신러닝 방법의 진능력모수 복원력을 살펴봄으로써 머신러닝 방법의 분류 정확도를 확인해보고자 하였다. 그 결과, k-최근접 이웃과 그래디언트 부스팅 방법이 가장 좋은 결과를 보여줬다.

이상적 문항 반응 유형을 훈련 자료로 사용하여 머신러닝을 적용한 것의 가장 큰 장점은 학생들의 능력모수 추정을 위하여 대규모 표본이 필요하지 않다는 것과 피험자 모수를 추정하기 위하여 문항모수를 추정하지 않아도 된다는 것이다. 인지진단평가에 대한 다양한 연구가 이루어져 왔음에도 학교 현장에서 활발하게 사용되지 못하는 이유는 인지진단평가를 적용을 위하

여 매우 큰 표본이 필요하고, 교사들이 접근하기에 수리적인 어려움이 존재한다는 것이었다. 그러나 머신러닝 모델은 Q-행렬로부터 이론적으로 각 인지요소가 요구하는 이상적 문항반응을 구한 자료만으로도 인지요소 숙달 여부를 추정할 수 있기에 기존의 인지진단모형을 학교 현장에 적용하는 데 가장 큰 한계였던 소표본 문제를 해결하기 위한 단서를 발견할 수 있었다. 다시 말하여, 검사와 Q-행렬만 존재하면 피험자 모수를 추정할 수 있다는 장점이 있음으로 교사들이 학생들의 문항 반응만 입력하면 각 학생의 인지요소 숙달 여부를 추정해주는 프로그램을 고려해 볼 수 있을 것이다. 이를 통해, 단위학교 수준에서도 학생의 인지적 강점과 약점을 진단하여 피드백을 제공함으로써 평가를 통해 학생 스스로 무엇이 부족한지 파악할 수 있을 것으로 기대된다. 또한 교사에게는 학생의 이해도에 따라 수업을 진단하고 개선할 수 있는 효과적 정보를 제공할 수 있을 것이다.

2. 연구의 제한점 및 제언

본 연구는 이론적 데이터인 이상적 문항 반응 유형을 훈련 자료(Train Data)로 사용하여, 다양한 머신러닝 방법을 적용해보고, 어떠한 머신러닝 방법이 가장 적합한지 그 가능성을 탐색해 보는데 그 의미가 있다. 그러나 본 연구는 아래와 같은 제한점을 지니고 있음으로 이에 대한 후속 연구를 제안한다.

첫째, 본 연구는 DINA 모형만을 활용하여 분석을 진행하였기 때문에 인지진단모형 중 비보상적 모형에만 적용할 수 있다는 한계가 존재한다. 더 다양한 연구를 위하여 DINO와 같은 보상적 모형에 머신러닝을 적용한 후속 연구가 필요하다.

둘째, 본 연구에서는 머신러닝의 결과를 개선할 수 있는 여러 가지 방법에 대해서는 적용하지 않았으므로, 하이퍼 파라미터 튜닝과 같은 머신러닝의 성능 향상 방법을 활용하여 모형의 성능을 개선하기 위한 방법을 탐색해 볼 필요가 있다.

셋째, 본 연구는 영어 검사 자료만을 고려하였기 때문에 연구의 결과를 일반화하기에는 한계가 존재한다. 따라서 영어 교과 외의 다른 교과의 검사 자료에도 머신러닝을 적용해볼 필요가 있다. 또한, 모의실험 자료를 10개 이상 생성하여, 피험자 모수 추정의 정확도를 평균으로 제시할 필요가 있다.

넷째, 개별학습, 맞춤형 학습이 요구되는 시대의 흐름에 따라 학습분석학적 관점에서의 인지진단에 대한 연구가 필요하다. 인지진단이론 하에서의 피험자 분석은 인지요소 숙달에 관한 정보를 프로파일 형태로 제공하므로 학생의 능력에 대한 과학적이고 체계적인 진단이 가능하다는 장점이 있으며, 진단의 결과로 제공되는 정보는 즉각적인 피드백을 통해 교수·학습 과정을 개선하는 데 직접적인 도움을 줄 수 있다(김희경 등, 2012). 따라서 학습자 행동을 분석하여 교수·학습 과정을 개선하고 학습자 특성과 수준에 부

합하는 맞춤형 처방 제공을 목표로 하는 학습분석학(조일현 외, 2019)과의 융합을 통해 단순 총괄평가 결과보다 더 많은 정보가 담긴 학생들의 문항 반응 및 인지요소 숙달에 관한 인지진단평가 결과를 활용한다면 교수·학습 과정을 개선을 위한 시너지 효과를 낼 수 있을 것으로 생각된다. 또한 학습분석학의 처방적 관점에서 사용되는 대시보드를 설계하면서 인지진단평가의 결과를 활용함으로써 학생 수준과 스타일에 맞춤화된 처방을 제안하는 적응형 학습(adaptive learning)에 가까워질 수 있을 것이다.

참 고 문 헌

강태훈 (2016). 베이지안 네트워크를 활용한 피험자 능력 추정의 정확성 탐색, **교육평가연구, 29**(1), 1-24

강태훈 (2019). 문항수준 사후예측모형검증을 통한 다중 인지진단모형 활용 가능성 탐색. **교육평가연구, 32**(4), 729-753

권철민 (2019). 파이썬 머신러닝 완벽 가이드. 경기: 위키북스.

김성은, 박윤수, 이영선 (2012). 인지진단모형을 적용한 잠재집단별 분수뺄셈검사 문제해결 전략의 차이 분석. **교육평가연구, 25**(1), 49-68.

김성훈, 송미영 (2011). DINA 모형을 활용한 대규모 학업성취도 결과 분석. **교육과정평가연구, 17**(2), 183-195.

김경리 (2012). **인지진단을 위한 진단적 준거설정과 교육적 활용**. 이화여자 대학교 대학원 박사학위논문.

김명연 (2016). **보상성·비보상성 가정이 심리측정 모형을 통한 다차원적 능력 추정에 미치는 영향**. 성신여자대학교 대학원 박사학위 논문.

김완섭 (2012). 로지스틱 회귀분석과 데이터 마이닝 분석을 이용한 컴퓨터 교양교육 성과의 요인에 대한 연구. **교양교육연구, 6**(3), 743-767.

김지효 (2012). G-DINA 모형에서 인지요소숙달 분류 정확성에 영향을 미치는 요인에 관한 연구. **교육종합연구, 10**(4), 81-101.

김지효 (2013). **DINA와 DINO 모형에서 응시생 분류 정확성에 영향을 미치는 요인 탐구: 응시생 분류방법을 중심으로**. 충남대학교 대학원 박사학위논문.

김혜숙 (2006). **데이터 마이닝을 사용한 방학 중 학습 방법과 학업 성취도의 관계 분석**. 강원대학교 교육대학원 석사학위논문.

김희경, 박종임, 강태훈 (2013). **기초학력 이하 학생의 맞춤형 학습 지도를**

위한 인지진단 프로파일 분석. 한국교육과정평가원 연구보고 RRE 2013-10.

반재천, 김선(2016). 인지진단모형, 인지요소패턴 추정 방법, 사례 수에 따른 인지요소패턴 추정의 분류 일관성 및 분류 정확성 비교. **교육평가연구, 29**(3), 405-431.

박해선 (2019). Do it! 정직하게 코딩하며 배우는 딥러닝 입문. 서울: 이지스퍼블리싱

배재호 (2001). **데이터 마이닝을 이용한 학업성취도 분석**. 경희대학교 교육대학원 석사학위논문.

성태제 외 (2013). 2020 한국 초·중등교육의 향방과 과제 - 교육과정, 교수·학습, 교육평가-. 서울: 학지사

성태제 (2014). 현대교육평가. 서울: 학지사

성태제 (2016). 문항반응이론의 이해와 적용. 서울: 교육과학사

송미영, 이영선, 박윤수 (2011). **인지진단모형을 통한 국가수준 학업성취도 평가 결과 분석 및 성적 보고 방법 탐색.** 한국교육과정평가원 연구 보고 RRE 2011-8.

오지영, 이수정 (2008). 신경망을 이용한 초등학생 컴퓨터 활용 능력 예측. **한국정보교육학회, 12**(3), 267-274.

윤지영 (2015), 요**소 간 위계 방식과 인공신경망을 적용한 수학 인지진단 평가 연구**. 서울대학교 대학원 석사학위논문.

은효정 (2017). **고등학교 수학 교과에 대한 DINA 모형의 학교 수준 적용**. 계명대학교 대학원 박사학위논문.

이명훈 (2017). 신경망 분석을 활용한 학교폭력의 예측요인 분석 및 해결방안 모색. **학습자중심교과교육연구, 17**(22), 537-561.

이영주 (2014). **인공신경망에 근거한 인지진단모형 Q 행렬의 타당성 평**

가. 이화여자대학교 대학원 박사학위논문.

이창묵 (2019). **자기보고식 심리검사에서 심층신경망 모형의 적용 가능성 탐색**. 건국대학교 대학원 박사학위논문.

이현 (2015). **DINA와 DINO모형에 의한 인지진단 결과의 분류 정확성과 일관성 비교**. 동국대학교 대학원 석사학위논문.

이혜윤 (2016). **대학 이러닝 환경에서 학습자 행동 로그에 기반한 군집별 학업성취 예측모형 비교.** 이화여자대학교 대학원 석사학위논문.

이혜주, 정의현 (2013). 컴퓨터활용교육: 청소년의 컴퓨터 오락추구 행동을 예측하기 위한 신경망 활용. **컴퓨터교육학회논문지, 16**(2), 39-48.

조일현, 김윤미 (2013). 이러닝에서 학습자의 시간관리 전략이 학업성취도에 미치는 영향: 학습분석학적 접근. **교육정보미디어연구, 19**(1), 83-107.

조일현, 김정현 (2013). 학습분석학을 활용한 e-러닝 학업성과 추정 모형의 통계적 유의성 확보 시점 규명. **교육공학연구, 29**, 285-306.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3). 210-229.

Binet, A., & Simon, T. (1916). *The development of intelligence in children*: The Binet-Simon Scale (No. 11). Williams & Wilkins Company.

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633-665.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart and Winston.

de La Torre, J. (2009). DINA model parameter estimation: A didactic. *Journal of Educational Behavioral Statistics, 34*, 115-130.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3). 333-353.

Géron, A. (2018). *핸즈온 머신러닝*. 박해선 (번역). 서울: 한빛미디어.

Gierl, M. J., Cui, Y., & Hunka, S. (2007). Using Connectionist Models to Evaluate Examinees' Response Patterns on Tests: An Application of the Attribute Hierarchy Method to Assessment Engineering. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Chicago, Il.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests. *Journal of Modern Applied Statistical Methods, 7*(1), 19.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*(6).

Guo, Q, Cutumisu, M., Cui, Y. (2017). A Neural Network Approach to Estimate Student Skill Mastery in Cognitive Diagnostic Assessments. Paper presented at the Annual Meeting of the 10th International Conference on Educational Data Mining, Wuhan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

Huebner, A., & Wang, B. (2011). A note on comparing examinee

classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407-419.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl. (Eds.), *Cognitively diagnostic assessment for education*: Theory and applications (pp. 19-60). Cambridge, England: Cambridge University Press.

Junker, B. W., & Sijtsma, K.(2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Kim, D., Park, Y., Yoon, M., & Jo, I. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education, 30*, 30-43.

Lantz, B. (2017). *R을 활용한 머신러닝 2/e*. 윤성진 (번역). 서울: 에이콘.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205 - 237.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Nichols, P. D. ,Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational*

Measurement, 33, 379-416.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Mahwah, NJ: Lawrence Erlbaum.

Rupp, A. A., Templin, J., & Henson, R. A. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68,* 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement*: Theory, methods, and applications. Guilford Press.

Shu, Z., Henson, R., & Willse, J. (2013). Using Neural Network Analysis to Define Methods of DINA Model Estimation for Small Sample Sizes. *Journal of Classification, 30*(2), 173-194.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1991). *Item construction and psychometric models appropriate for constructed responses*. ETS Research Report Series, 1991(2), I-38.

ABSTRACT

An Exploratory Study on the Usability of Machine Learning for Cognitive Diagnosis: Focusing on Data Size and Ideal Item Response Pattern

Jae Eun Heo
Department of Education
Graduation School of
Sungshin University

In this study, we wanted to apply Machine Learning, a research method that has received much attention recently, to Cognitive Diagnosis Assessment(CDA), which is part of the educational evaluation. By comparing the DINA, one of the Cognitive Diagnosis Models(CDM), and machine learning, we verified the performance of the examinees' ability parameter estimation, based on which, we explored the possibility of using machine learning for CDM.

To proceed with the research, we organized two data analysis. First, the real data(English test) of 3,000 examinees were used to analyze the consistency between the DINA model and machine learning on the estimation of the examinees' ability parameters. The English test consisted of 18 items and had five cognitive attributes, which are

‘recognition’, ‘understanding’, ‘inference’, ‘assessing’ and ‘configuring’. Second, the estimates of item parameters extracted from the real data were used to generate simulation data for use as test data. The simulated data showed 3,000 responses to 18 items in total.

The examinee's ability parameters, which are assumed by machine learning, refer to the proficiency of each of the five cognitive attributes and the pattern of mastering cognitive attributes. First, to check the applicability of machine learning, we estimated examinees parameters with real data by using the DINA model. And the results were used as train data and test data to apply various machine learning and examine the consistency between the DINA model and the examinees′ ability parameter estimations of machine learning. According to the analysis of real data, the larger the data size was, the higher the consistency between the DINA model and machine learning was found to be at least 85% to 100%, indicating that the machine learning has a high potential for utilization.

However, since it is difficult to secure item responses and estimated examinees′ parameters as training data when trying to apply machine learning for CDM after making tests at real education scenes, we wanted to see if training data could be replaced by an ideal item response pattern based on Q-matrix.

At this time, we studied whether the results of the examinees′ ability estimations differ depending on the size of the ideal item response pattern, which is the training data. As a result, it resulted that the size of the training data does not have a significant impact on accuracy so

that the number of combinations of theoretical cognitive attributes could be used without artificially increasing the ideal item response pattern.

Finally, in order to verify the accuracy of estimating the examinees' ability parameter according to machine learning, the simulation data were used to confirm the restoration of the true ability parameters. After analyzing the accuracy of each machine learning through simulation data, it was found that k-nearest neighbors and gradient boosting are the most suitable.

Existing CDM required large samples to estimate items and examinee parameters, so there was a limitation in applying them to real education scenes. However, the main significance of this study is if when a reasonable Q-Matrix exists, individual students can be diagnosed based on ideal item responses and machine learning without existing test data. It is also confirmed that the results of the diagnosis using machine learning are not much different from the estimates of the examinees' ability parameters using the DINA model and that the resiliency of the true ability parameters is similar. Given that most of real education scenes are small samples, further research on the application of different cognitive diagnostic models and various cognitive diagnostic assessment context is needed in that machine learning models can be an alternative to CDM.

Key Words: Cognitive diagnostic assessment, DINA model, Ideal item response pattern, ability estimation, Machine learning, Supervised learning.