https://doi.org/10.5392/JKCA.2020.20.06.695

기계학습을 활용한 대학생 학습결과 예측 연구

A Study on the Prediction of Learning Results Using Machine Learning

김연희*, 임수진**
호서대학교 전자디스플레이공학부*, 청주대학교 교육혁신원**

Yeon-Hee Kim(kimyh@hoseo.edu)*, Soo-Jin Lim(soojinlom@gmail.com)**

요약

최근 교육분야에 IT의 활용이 증가하고 이를 통한 학습결과 예측에 대한 연구가 진행되고 있다. 본 연구에서는 학습분석을 참고하여 학습결과에 영향을 미칠 수 있는 학습활동 데이터를 수집하였다. 조사에 참여한 학생은 1062명으로, 조사는 2018년 10월부터 12월까지 충청남도 소재의 4년제 종합 사립대학인 A대학에서 진행되었다. 먼저 기계 학습의 예측 변인들의 타당성 확보를 위하여 학습결과에 대한 개인·학업·행동요인으로 모형을 구성하여 위계적 회귀 분석을 실시하였다. 위계적 회귀 분석의 모형이 유의하였고, 단계별로 설명력(R^2)이 증가하는 것으로 나타나 투입된 변수들이 적절한 것으로 나타났다. 또한 기계학습의 선형 회귀분석방법을 통해 투입한 학습활동 변수가 학습 결과를 얼마나 예측할 수 있는지 확인하였으며, 오차율은 약 8.4%로 수집되었다.

■ 중심어 : | 학습분석 | 기계학습 | 텐서플로우 | 위계적 회귀분석 |

Abstract

Recently, There has been an increasing of utilization IT, and studies have been conducted on predicting learning results. In this study, Learning activity data were collected that could affect learning outcomes by using learning analysis. The survey was conducted at a university in South Chung-Cheong Province from October to December 2018, with 1,062 students taking part in the survey. First, A Hierarchical regression analysis was conducted by organizing a model of individual, academic, and behavioral factors for learning results to ensure the validity of predictors in machine learning. The model of hierarchical regression was significant, and the explanatory power (R^2) was shown to increase step by step, so the variables injected were appropriate. In addition, The linear regression analysis method of machine learning was used to determine how predictable learning outcomes are, and its error rate was collected at about 8.4%.

■ keyword : | Learning Analytics | Machine Learning | Tensorflow | Hierarchical Regression Analysis |

I. 서 론

4차 산업혁명이라고 말하는 현실과 가상공간의 융합은 산업과 인간 활동 전 분야에 변화를 가져오고 있다. 교육 현장에서도 OCW와 MOOC등에 의해 학습의 형태가 바뀌고 있으며 교수설계 영역에서 학습자들의 학습 효과를 강화하는 학습과정이나 다양한 교수전략에 IT가 접목되고 있다. 국내 교육 분야에서도 스마트교육

* 이 논문은 2018년도 호서대학교의 재원의 학술연구비 지원을 받아 수행된 연구이며(과제번호: 20180315) 이에 감사드립니다.
접수일자 : 2020년 06월 02일
수정일자 : 2020년 06월 16일
심사완료일 : 2020년 06월 16일
교신저자 : 김연희, e-mail : kimyh@hoseo.edu

이라는 정책 아래 IT의 접목이 이루어졌고, 빅데이터의 처리 기술의 발전은 디지털 공간에서의 데이터뿐만이 아니라 학습과정에서 사용되는 모든 도구, 개별적 교수·학습환경을 제공하는 플랫폼, 그리고 분석기법까지를 포괄하는 용어로 정립되어 가고 있다[1]. 또한, 교육이 이루어지는 모든 조직에서는 구성원의 학습을 보다 더 체계적으로 관리하기 위한 목적으로 학습관리 시스템(Learning Management System, 이하 LMS)을 개발하여 사용하게 되었다.

이와 같은 IT가 실제 교수학습 환경에 적용되는 사례를 보면 오스틴 커뮤니티 컬리지는 Civitas Learning과 제휴하여 학위 맵을 구현하여 어드바이저와 학생에게 학위 요건에 쉽게 접근하고 비교할 수 있는 방법을 제공함으로써 개인화된 학습 경로를 가능하게 했다. 또 다른 주목할 만한 사례는 위스콘신 대학의 Insights Student Success System으로 교육에서 기존의 데이터 보고 방식을 뛰어넘고 예측 분석 및 데이터 모델링을 활용하여 주요 성과, 참여 및 완료 데이터를 심층적으로 검토하여 학습자의 진행 상황과 성공에 대한 상세한 관점을 제공한다. 또한, Delgado Community College 와 글래스고 대학은 학생들이 원하는 학습 성과를 목표로 진행 상황을 평가하고 안내하는 GPS와 같은 도구를 제공하고 있다. Southern New Hampshire University (SNHU)는 Blackboard와 제휴하여 자동화된 코스 제공을 포함하는 MHE(Megraw Hill Education) 라는 자동화 학점 이수 블록을 개발하여 코스 설정 프로세스에 교수자 개입을 최소화하였다[2-4].

학습자에 대한 데이터 양의 급격한 증가로 이를 분석하고 결과를 활용하는 것이 교수자가 학습자를 이해하는데 중요한 분야가 되고 있다. 2000년대 이후 IT기술을 기반으로 발생된 학습분석학(Learning Analytics)은 학습과 학습환경을 이해하고 최적화하기 위해 학습자와 학습자의 상황과 관련된 데이터를 측정하고 모으고 분석하고 보고하는 것이다. 이것은 학습자의 성적뿐만 아니라 행동, 성격 등 다양한 데이터를 수집해 풍부한 프로파일을 제공함으로써 개별화 학습이 가능하다. ICT의 발전은 이와 같이 다양한 데이터의 활용을 가능하게 하였다. 따라서 본 연구에서는 현재 진행중인 학습분석 시스템 개발을 목표로 대학 학습자에 대하여 학습분석학에서 다루어지고 있는 학습데이터를 중심으로 기계학습을 활용한 학습 결과 예측 가능성을 검토해보는데 목적이 있으며, 투입된 변수의 타당성 확인을 위하여 위계적 회귀분석을 실시하였다.

II. 이론적 배경

1. 학습분석(Learning Analytics)

IT 기반기술의 발전은 학습테크놀로지의 발전을 의미하며 이러닝이나 다양한 컴퓨터를 도구로 하는 교육방법이 나타났고 이러한 교육용 애플리케이션의 생태계가 확대되면서 교사와 학생을 위한 혁신적인 학습 도구의 통합이 가능해지고 있다[5].

학습분석은 학습활동 데이터를 분석함으로써 그동안 알지 못했던 학습에 대한 통찰을 얻기 위한 것이다. 분석학은 IT 기술의 발전과 함께 급격히 많아진 데이터 분석을 위하여 함께 성장하여 왔다고 해도 과언이 아니다. 기존의 설문조사 등 면대면에 의한 방식으로 수집된 소규모 데이터 분석에 비하여 인터넷 활용으로 생성되는 방대한 양의 학습활동 데이터의 수집, 처리에는 보다 고도의 분석방법이 필요하기 때문이다. 교육학에서 분석학의 발전은 학습관련 데이터의 분석을 통해 개인별 학습활동의 특성 파악, 학습성과 예측 및 개별화 학습 등 다양한 통찰을 얻을 수 있게 되었다. 2000년대 초반부터 지속된 이러한 연구를 학습분석이라 하며, 연구자나 연구단체에 따라 다양한 정의가 있으나 대체적으로 '학습이 이루어지는 환경과 학습을 이해하고 최적화된 학습 환경을 제공해 주기 위해 학습자와 학습 맥락에 대한 데이터를 수집, 측정, 분석하는 일련의 과정'이라고 정의한다[6]. 학습분석을 위한 세부 기술로는 데이터 추출, 저장, 분석, 시각화, 예측, 결과적용이 있으며, 이를 연구하고 발전시킨 연구자 중에서는 여러 가지 모델과 프레임워크를 비교 및 결합함으로써 선택, 수집, 종합 및 보고, 예측, 활용, 개량, 공유의 Tenya의 7단계 학습 분석 프로세스가 있다[7][8]. 기존의 연구에서 다루어진 것처럼 학습의 결과는 학습자의 학습활동과 관련한 다양한 요인에 의해 영향을 받는다. Hellas

등은 잠재 디리클레 할당 알고리즘등을 사용하여 학습결과 예측을 위해 활용할 수 있는 고수준의 예측 요인으로 [표 1]과 같이 제시한바 있다[9-11].

표 1. 학습성과 예측을 위한 주요 요인

Research	Factor
Hu, Cheong, Ding & Woo(2017)	Activity and course features, Demographic features, Learning behavior features, Self-reported features, Student history record and performance, Student record and performance in current course, Others / unclear features
Kumar, Singh & anda(2017)	Academic, Family, Institutional, Personal, Social
Lei & Li (2015)	Academic performance, Socio-economic, Personal information
Na & Tair(2017)	Learning behaviour data, Learning network data, Learning level data, Learning emotional data, Other
Shahiri, et al(2015)	Cumulative Grade Point Average, Engage time, External assessments, Extra-curricular activities, Family support, High-school background, Internal Assessment, Social interaction network, Study behavior, Student demographic, Student interest

2. 기계학습

최근에 이러한 학습분석에 필요한 데이터분석이나 예측등의 기술을 구현하기 위하여 데이터마이닝이나 머신러닝 기법이 많이 사용되고 있다. 머신러닝 즉, 기계학습은 인공지능의 한 분야로, 컴퓨터가 데이터를 통해 학습하고 그 데이터의 함의를 도출하는 것으로서 “환경과의 상호작용에 기반한 경험적인 데이터로부터 스스로 성능을 향상시키는 시스템을 연구하는 과학과 기술”로 정의 할 수 있다. 즉, 기계학습은 빅데이터를 활용하여 데이터의 의미를 파악하고, 결과를 제공하는 귀납적 방식으로 일련의 학습과정을 지속적으로 반복하여 결과를 향상시키고 이를 통해 미래를 예측하는 특성이 있다[12].

기계학습을 통한 예측 연구의 예를 보면 동영상을 기반으로 한 학습에서 기계학습의 서포트 벡터 머신을 통해 85.71%의 정확도로 학업성취를 예측하였다[13]. 또한, 동계전력수요 예측에 기계학습을 활용하여 사용 방법에 따라 다양한 결과를 도출하였으나 98~99%로 수요를 예측한 바 있다[14]. 대표적인 자료분석 방법인 통계분석과는 다르게 기계학습은 자료분석을 통해 추론과정과 예측까지를 목적으로 한다. 데이터의 패턴을 검색하고 식별하는 데이터마이닝은 예측을 위한 기계학습과는 조금 다른 방향이라고 볼 수 있다.

Ⅲ. 연구방법

1. 연구목적 및 설계

본 연구의 목적은 기계학습을 통한 학습분석 시스템을 구축하기 위한 사전 연구로서 필요한 학습데이터를 분석하는 것이다. 또한 학습 결과 예측 시스템을 함께 구축하기 위하여 학습결과와 관련된 항목을 조사하고 이를 기계학습을 통해 예측하여 예측 정확도를 고찰해 보는 것이다. 아래 [그림 1]은 현재 구축중인 학습분석 시스템의 랜딩페이지이다.

그림 1. 학습분석시스템 랜딩페이지

학습데이터에 대한 연구는 여러 연구자에 의해 진행되고 있으며 최근의 그 대상은 대부분 학교에서 사용하고 있는 LMS에 존재하는 학생들의 학습활동 및 학습결과 정보를 대상으로 한다. 학생들의 학습을 지원하기 위한 LMS는 사용 후 그 흔적이 남아 그 기록을 분석하여 결과를 도출하는 것인데, 본 연구에서는 학습분석 시스템을 LMS에 연계하는 것을 목적으로 충청남도 소재의 4년제 종합사립대학의 LMS 시스템상의 학습활동 데이터 항목을 참고하여 변수를 선정하고 이를 설문조사 방법을 통해 조사하였다. 학습활동 데이터는 학습분석을 기반으로 LMS에 존재하는 것 중에 기존 연구를 참고하여 변수를 정하였으며 LMS에 존재하지는 않지만 학습결과에 중요한 영향을 줄 수 있는 변수를 추가 선정하였다. 또한, 조사된 각 변인들의 타당성을 확인하

기 위하여 통계분석방법의 하나인 위계적 회귀분석을 실시하여 회귀식의 유효성과 설명력을 확인하였다. 학습결과 예측을 위하여 본 연구에서는 아래 [그림 2]와 같은 절차를 통해 본 연구의 데이터를 분석하고 그 결과를 도출하였다.

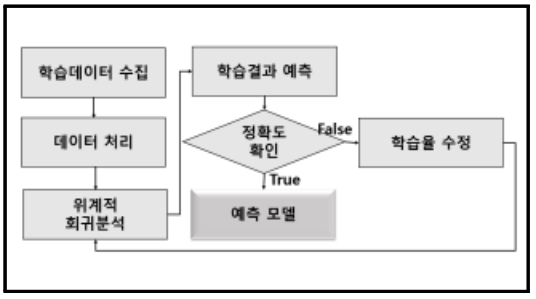


그림 2. 연구절차

2. 주요변인의 설정 및 측정

본 연구는 향후 LMS와 연동한 학습분석시스템을 구축하기 위한 사전 검증 과정으로써 제시한 모형의 성능과 변인들의 타당성을 검증하고 예측을 진행하였다. 예측모델을 세우기 위한 학습데이터는 학사시스템에 쌓여 있는 데이터를 직접 추출하여야 하나 개인정보와 관련한 데이터 접근의 어려움으로 설문조사로 대체하였다. 설문 문항은 학습성과를 예측하는 주요 변인으로 개인요인, 학업요인, 행동요인으로 설정하였다. 학생 개인요인으로는 성별과 함께 부모에게서 완전히 경제적 독립을 이루지 못하는 우리나라의 특성상 가족소득수준에 영향을 받는 용돈 월총액[15] 및 중요한 개인적 속성으로 학습에 대한 흥미요인을 도출한 관심교과과정 문항을 추가하였다[10]. 또한, 학생들이 좋은 학점을 받기위하여 들이는 노력정도를 총 100%라 할 때 레포트, 퀴즈 및 중간·기말고사, 기타에 어느 정도의 노력을 들이는지를 학업요인으로

표 2. 학습데이터 조사 항목

구분	항목	내용	서버추출 가능여부
종속변수	1. 전학년 평균 학점		○
개인요인	2. 성별		○
	3. 관심교과과정		×
	4. 용돈 월 총액		×
학습데이터	5. 중간, 기말고사	직전학기	○
	6. 퀴즈, 레포트	직전학기	○
	7. 기타 노력 정도	직전학기	×
행동요인	6. 수업 좌석 위치	전, 중, 후	×
	7. 수업 외 평균 학습 시간		×

설정하였으며. 마지막으로 행동요인으로는 학습시간과 함께 학습패턴을 묻는 문항으로 구성하였는데, 우리나라 상위권 대학생들은 중간 줄을 선호한다는 선행 연구결과에 따라 수업시간에 학생들이 주로 선호하는 좌석의 위치를 답변하도록 하였다[16].

추후 학교 LMS 서버에서 데이터 추출을 고려하여 서버추출 가능여부와 함께 정리한 내용은 위의 [표 2]와 같다. 설문은 충청남도 소재의 4년제 종합사립대학인 A대학에 재학 중인 남녀 학부생을 성별과 단과대학을 고려한 임의할당 방식으로 4개 단과대학에 고르게 할당하였으며, 각 학과의 협조를 얻어 연구 참여에 동의한 학생을 대상으로 조사를 실시하여 1062명의 데이터를 수집하였다. 이중 다수의 문항에 무응답을 하거나 불성실한 응답을 한 203명을 제외한 859명의 조사 결과를 분석에 사용하였다. 분석에 사용한 연구 참여자의 개인적 특성은 [표 3]과 같다.

표 3. 개인적 특성 **N=859**

분 류		빈도	범위(%)
성별	남성	634	73.8
	여성	225	26.2
단과대학	사회과학대학	198	23.1
	공과대학	238	27.7
	자연과학대학	221	25.7
	예체능대학	202	23.6

Ⅳ. 연구결과

학습분석학에서 정의하는 학습데이터는 학습자들의 물리적 행동을 포함하는 학습활동과 관련된 모든 데이터를 수집해야 한다. 그러나 학습을 위한 모든 학습 활동이란 학습 결과에 영향을 미치는 학습자의 다양한 내·외부 요인이 있을 수 있으므로 가능한 관련성을 검

토한 후 학습 활동 데이터를 선정하여야 한다. 또한 본 논문은 학습데이터를 통한 대학생의 학습결과 예측과 관련한 연구로서 학습데이터 요인보다는 예측에 중요성을 두었다. 추가적으로, 학교 서버 추출가능 데이터와 함께 선행연구를 참고하여 학습과 관련된 추가 데이터 즉, 학습패턴, 평균학습시간의 영향력을 검증하기 위하여 위계적 회귀분석을 실시하였다. 위계적 회귀분석은 이론적으로 설정된 단계에 따라 독립변수를 누적하면서 회귀분석을 수행하는 방법으로서 이전에 투입된 독립변수들의 효과를 통제하고 새롭게 추가되는 독립변수 군이 종속변수를 설명하는 데 얼마나 기여하는지를 비교하기 위한 것으로 3단계 모델을 설정하여 첫 번째 단계는 개인요인으로 설정하고 성별, 용돈월총액, 관심교과과정을 투입하였다. 2단계는 학업요인으로 직전학기 학생이 학습에 들인노력으로 중간고사, 기말고사, 레포트, 퀴즈, 기타 학습에 들인 노력을 투입하였고, 3단계는 행동요인으로 주당 평균학습시간, 평소 수업시간에 앉는 좌석을 투입하였다[17]. 단계별 변수의 평균과 표준편차를 구한 기술통계는 다음의 [표 4]와 같다. 대학생의 학업성취도에 미치는 영향력을 탐색하기 위해 개인요인, 학업요인, 행동요인으로 3단계 회귀모형을 설정한 후 관련 변수들을 차례대로 투입하여 다음의 [표 5]와 같이 위계적 회귀분석을 시행하였다[18][19].

단계별 회귀식이 1단계(F=24.127 p<.001), 2단계(F=14.653, p<.001), 3단계(F=19.221, p<.001)에서 모두 유의하게 나타났고, 설명력 또한 차이가 크진 않으나 증가하는 경향을 보였다. 단계별 회귀모형을 살펴보면, 1단계 개인요인에서 학업성취도에 영향을 미치는 변수로는 가정수준에 영향을 받는 용돈 월총액(t=5.20, p<.001)과 학생의 주 관심 교과과정(t=-6.02, p<.001)이 유의한 영향을 미치는 것으로 나타났고, 성별(t=-1.81, p>.05)은 유의한 영향을 미치지 않는 것으로 나타났다. 이러한 영향력은 2, 3단계까지 유지되었으며, 2단계 투입한 학업요인에서는 기말고사(t=3.19, p<.001)와 레포트(t=-2.84, p<.001)가 통계적으로 유의하였다. 또한, 3단계 학생의 평소 학습 행동요인을 의미하는 변수들을 투입하였을 때, 평소 학생이 수업시간

표 4. 기술통계

변수명		변수설명	M	SD	N
종속변수	학업성취도	학생의 평균학점	3.23	0.57	862
개인요인	성별	남자=0, 여자=1			
	용돈월총액	① 10~20만원 ② 20~30만원 ③ 30~40만원 ④ 40만원 이상	2.68	0.95	865
	관심교과과정	전공=0, 교양과정=1			
학업요인	중간고사	중간고사에 들인 노력	0.31	0.09	868
	기말고사	기말고사에 들인 노력	0.31	0.09	868
	레포트	레포트에 들인 노력	0.22	0.12	868
	퀴즈	퀴즈에 들인 노력 (중간, 기말고사 외 시험)	0.15	0.09	868
	기타	기타학습에 들인 노력	0.02	0.07	868
행동요인	수업시간좌석	수업시간시 칠판을 기준으로 앉는 위치 ①앞 ②중간 ③ 뒤	2.06	0.73	868
	평균학습시간(주)	학생의 주 단위 학습 평균시간	2.04	1.45	868

표 5. 위계적 회귀분석 결과

학습 활동	개인요인				학업요인				행동요인			
	B	β	t	p	B	β	t	p	B	β	t	p
(상수)	3.37		49.52	0.00	3.15		15.50	0.00	2.73		13.49	0.00
성별	-0.04	-0.06	-1.81	0.07	-0.04	-0.06	-1.83	0.07	-0.03	-0.04	-1.40	0.16
용돈 월총액	0.23	0.17	5.20	0.00	0.25	0.18	5.70	0.00	0.23	0.17	5.53	0.00
관심교과과정	-0.27	-0.20	-6.02	0.00	-0.23	-0.17	-5.21	0.00	-0.20	-0.15	-4.70	0.00
직전학기 중간고사					-0.11	-0.02	-0.36	0.72	-0.10	-0.01	-0.35	0.73
직전학기 기말고사					0.99	0.13	3.19	0.00	0.89	0.12	2.97	0.00
직전학기 레포트					-0.66	-0.12	-2.84	0.00	-0.56	-0.10	-2.49	0.01
직전학기 퀴즈 노력 정도					0.48	0.07	1.80	0.07	0.36	0.05	1.38	0.17
기타 학습 노력 정도					0.13	0.01	0.22	0.82	0.15	0.01	0.26	0.80
수업시간 좌석									0.11	0.15	4.80	0.00
수업외 주당 평균 학습시간									0.10	0.18	5.71	0.00
R2(ade R2)			.078	(.075)			.121	(.113)			.185	(.175)
F			24.127	(.000)			14.653	(.000)			19.221	(.000)

에 앉는 좌석의 위치(t=4.80, p<.001), 평균학습시간(t=5.71, p<.001) 모두 유의한 영향을 미치는 것으로 나타났다. 위계적 회귀분석의 결과 대학생들의 개인, 학업, 행동요인이 학업성취를 예측할 수 있는 변수임을 확인하였다.

변수에 대한 검토와 함께 예측을 위한 기계학습의 Training과 Test data 구분을 위하여 Murat Pojon(2017)의 다수결의 정확도를 적용하여 859명 중 704명이 성적이 좋거나 만족스럽게 나타나 다수결의 정확도가 82%로 나타났다[20]. 따라서, 수집된 최종 데이터의 80%인 690명을 Training data로 20%인 169명의 데이터를 Test data로 결정하여 사용하였으며 직전학기 평점을 결과값으로 예측하는 작업을 수행하고 그 정확도를 확인하였다[21].

본 연구에서는 Google에서 제공하는 Tensorflow를 사용하여 기계학습의 회귀분석 방법으로 데이터를 분석하였으며 선형회귀분석은 종속변수 y와 한 개 이상의 독립변수 x와의 상관관계를 모델링하는 기법이다. 기계학습을 위한 프로그램은 구글에서 2011년에 개발을 시작하여 2015년에 오픈소스로 공개한 Tensor- flow를 사용하였으며[15] 이 프로그램에서 사용하는 Linear regression에서의 가설식과 실제값과의 차이를 최소화하는 Cost Function은 최소제곱법으로 아래와 같이 설정하였다.

$$H_{(x)} = W \times x + b \quad (1)$$

$$Cost(W,b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2 \quad (2)$$

여기서 W는 가중치, x는 독립변수, H(x)는 가설값, y는 실제값으로서 가장 작은 W, b를 찾아서 실제와 가설의 차이를 작게 만들고 이를 통해 정확한 예측을 할 수 있게 된다. 이러한 학습 데이터를 기반으로 기계학습 알고리즘 또는 Function F를 만들어 내고 이후 학습을 하지 않은 입력벡터 $\hat{x}$가 입력되면 자동으로 출력값 $\hat{Y}$를 추론하게 된다. 연구를 위한 위의 가설식을 근거로 학습데이터를 분석하고 예측하기 위하여 Training Data와 Test Data를 아래와 같이 적용하였다.

```
xy = np.genfromtxt('training_data.csv', delimiter=',', dtype=np.float32)
test = np.genfromtxt('test_data.csv', delimiter=',', dtype=np.float32)
```

또한 학습분석을 위한 학습데이터 항목은 기본사항과 학습데이터 부분으로 구분하여 학점을 제외한 9개의 독립변수를 구성하였다. 여기서 데이터 전체의 예측값과 실제 목표값인 Y 와의 차이 평균을 구하는 함수가 cost값으로 가중치 W와 계산된 예측값이 목표값 Y와 연관되도록 오차 최소화(minimize) 과정을 거쳤다. 사전에 분류한 80%의 Training Data를 이용하여 학습을 실시하였으며 학습율은 0.0001로 설정하였다. 20%의 Test Data를 학습이 완료된 모델에 입력하여 입력된 모델에 적용하여 예측값과 실제 학점을 비교하여 오차율을 측정하였다.

```
W = tf.Variable(tf.random_normal([9, 1]), name='weight')
b = tf.Variable(tf.random_normal([1]), name='bias')

hypothesis = tf.matmul(X, W) + b
cost = tf.reduce_mean(tf.square(hypothesis - Y))
optimizer = tf.train.GradientDescentOptimizer(learning_rate =1e-5)
train = optimizer.minimize(cost)

sess = tf.Session()
sess.run(tf.global_variables_initializer())

for step in range(1000001):
    cost_val, hy_val, _ = sess.run([cost, hypothesis, train], feed_dict={X: x_data, Y: y_data})
```

cost 함수 값과 W로 표현된 그래프에서 기울기를 구해 cost가 더 작은 값으로 수렴할 수 있도록 다음 W 값을 설정해 준다. 수집한 표본 데이터중 859명에 대한 데이터를 사용해 [표 2]의 학습데이터 조사 항목에서 직전학기 평점을 Y에 그 외 데이터항목을 X에 넣고 training 시켰다. 위 과정 중 초기 Hypothesis 식을

설정하는 과정을 제외한 과정은 반복되는 과정으로써, 주기가 지날수록 minimize 과정에 의해 cost(W)값이 줄어들며 W 값은 임의의 값에서 점점 실제 목적 값과 연관된 배열로 바뀌어 나간다. 여기서 가중치가 적용된 가설값을 찾아 실제값과 가설값의 차이를 찾고 차이값을 모두 제곱하여 총합을 도출하고 이를 전체 조사 횟수와 나누어 평균값을 구하였다. 제곱하는 이유는 음수를 모두 제거하기 위한 것이며 제거된 음수값과 함께 cost값을 찾게 된다. 따라서, 코스트값이 0에 가까울수록 가설의 정답에 가까워진다. 즉, 미분에 의해 그래프의 기울기를 구하고 오차가 작을수록 기울기가 0에 가까워지는 것을 말한다. 또한, Cost 값을 0에 수렴하도록 하는 함수가 Grandient Descent Optimizer 이며 이때 rate를 전달하기 위하여 0.0001만큼 감소되도록 하였으며 minimize함수를 통하여 W와 b값을 구하였다. 기계학습을 통한 예측모델을 위해 Training data의 패턴을 Test data에 적용하여 실제 학점을 비교한 결과는 아래 [그림 3]과 같다. 859명중 169명의 Test Data를 학습한 결과와 실제 학점을 비교한 결과 최초 학습할 때의 평균 오차는 약 18.4%로 나타났으나 데이터의 학습 횟수를 증가시킬수록 오차율은 감소하였다.

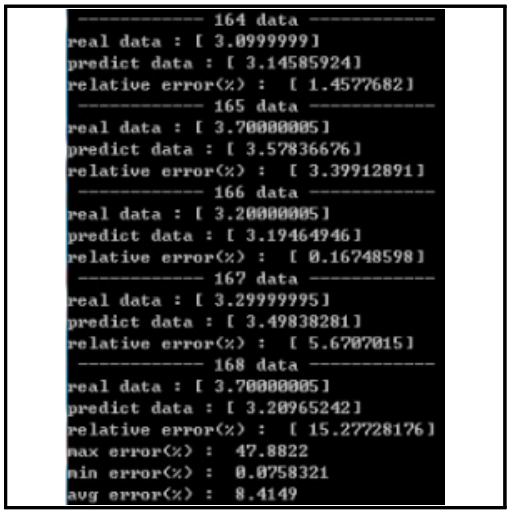
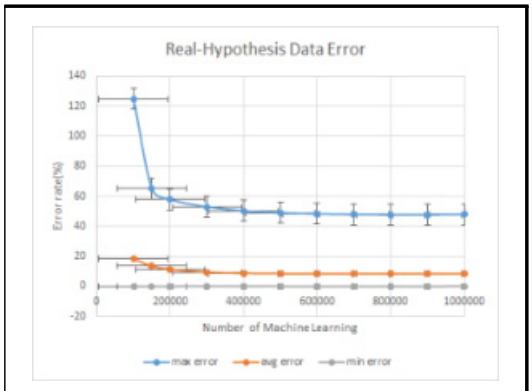


그림 3. 기계학습 결과

학습횟수가 최초 10만번에서 18.37%이던 평균오차율이 40만번에서는 8.68%로 급격히 줄어들었고 이후에는 아래 [그림 4]와 같이 평균 오차율에 큰 변화가 없었다. 이와 같이, 학습데이터에 기계학습의 선형회귀분석 알고리즘을 사용하여 학습을 시키고 결과를 확인해 본 결과 40만번 이후에서 평균 오차율은 약 8.41~8.67%에 수렴됨을 알 수 있었다.



Learning Num. / Error	100,000	150,000	200,000	400,000	600,000	800,000	1,000,000
Max Error	124.879	65.0897	57.8722	50.3141	48.2326	47.7652	47.8822
Avg Error	18.3669	13.4384	11.3799	8.67916	8.42804	8.43927	8.4149
Min Error	0.00752	0.00194	0.01780	0.04947	0.00645	0.04020	0.07583

그림 4. Test Data 와 가설 비교

V. 결론

학습자의 학습 성취가 성공적이냐에 따라 학습자의 동기부여는 더욱 강화될 것이고 학습자가 현재의 학습에 대한 자기 상태를 인지한다면 학업지속과 강화에 유의미한 영향이 있다는 점에서, 학습결과 예측은 결과인지에 대한 동기부여 및 학습 계획 차원에서 매우 중요하다고 할 수 있다. IT의 발전과 함께 학습분석의 개념이 정립된 후 학습데이터를 분석해서 활용하는 방안이 다양하게 연구되고 있다. 이에 본 연구에서는 학습분석을 참고하여 2018년도 10월부터 12월까지 충청남도 소재의 4년제 종합사립대학인 A대학에 재학중인 학생 1062명을 대상으로 학습결과에 영향을 미칠 수 있는 학습활동 데이터를 수집하였으며 기계학습의 선형

회귀분석 방법을 통하여 이를 분석하고 학습활동이 학습결과를 얼마나 예측할 수 있는지 확인하고자 하였다. 기계학습을 통한 학습활동 데이터를 분석하여 예측하는 연구는 현재 이론적인 연구가 주로 이루어지고 있어 기초적 방법의 적용이기는 하나 기계학습을 적용한 사례로서 본 논문의 의미가 있다. 예측을 위해 사용한 학습활동 데이터의 타당성을 검토하기 위하여 통계분석의 위계적 회귀분석을 사용하였다. 위계적 회귀분석에서는 회귀식이 유의하고, 분석단계에 따라 설명력(R^2)이 증가하는 것으로 나타나 예측을 위한 적정한 요인으로 구성되었음을 의미하였다. 다만 설명력이 낮은 경향이 있는데, 학습성과에는 보다 다양한 영향요인들이 있을 수 있으므로, 학습자 및 학습환경 등의 요인발굴을 위한 추가적인 연구를 통해 의미 있는 결과 도출이 필요하다. 학습결과 예측을 위하여 본 연구에서는 구글에서 제공한 오픈소스 라이브러리인 Tensorflow를 사용하여 현재 완료된 학습활동과 학습결과를 비교·분석하여 학습자의 학습활동에 따른 학습결과 예측의 가능성을 확인하고자 하였다. Tensorflow는 데이터 플로우 그래프방식을 사용하며 노드는 수학적 계산, 데이터 입/출력, 그리고 데이터의 읽기/저장 등의 작업을 수행한다. 특히, 엣지는 노드들 간 데이터의 입출력 관계를 나타내며 동적 사이즈의 다차원 데이터 배열(=텐서)을 실어 나르는데, 여기에서 Tensorflow라는 이름이 지어졌다. 학습자의 학습활동에 관한 주요 예측 변수를 바탕으로 다차원 데이터로 학습활동 데이터를 수집하여 Training Data와 Test Data를 통해 예측여부를 확인해본 결과 오차율이 약 8.4%로 수렴되었다. 오차율은 예측정확도라고도 표현되고 선행연구를 통해 동영상을 기반으로 한 학습에서 기계학습의 서포트 벡터 머신을 통해 85.71%의 정확도로 학업성취를 예측한 연구와 동계전력수요 예측에 기계학습을 활용하여 98~99%로 예측하는 연구 등의 기계학습의 사용 방법에 따라 다양한 예측결과를 도출해 내는 사례가 있으나 현재 교육분야에서는 이론적 연구가 주로 이루어지고 있다

Test Data에 대한 오차율이라는 것은 Training Data의 학습활동과 결과를 비교 분석한 패턴에 의해 Test Data로 분류한 학습자의 학습활동데이터를 분석하고 예측한 학점의 결과값과 실제 그들의 학점을 비교한 것이다. 이는 다음 학기에 있을 최종 학점이 나오기 전 어느 시점에서 학습자의 학습활동 데이터를 수집하여 분석하면 예측학점이 나오는 결과를 도출할 수 있다. 물론 좀 더 정교한 접근이 필요하지만 본 연구에서 그 가능성을 확인하였다. 그러나 본 연구에서는 학습활동 데이터 수집에 있어 몇 개 단과대학만을 대상으로 조사를 진행하여 결과에 대한 일반화가 부족하며 또한 예측을 위하여 적용한 기계학습의 선형회귀분석 방법 외에도 다양한 분석 방법이 있어 이를 적용하고 그 결과를 비교한다면 더 정확한 예측결과를 도출할 수 있다. 따라서, 향후 후속 연구에서예측 정확도를 높이기 위한 데이터 수집 방법과 기계학습에 대한 몇가지 제언을 하면, 첫째 학습활동 데이터 항목 선정에 있어 보다 적극적인 관련성 검토가 필요하다. 현재 학습활동 데이터를 통한 연구의 대부분이 학습분석에서 다루어지고 있는 항목들이 사용되고 있으나 국내외 선행연구를 적용하여 보다 확장된 학습 데이터를 고려해야 한다. 다음 연구에서는 현재 적용한 학습데이터 외에도 더 깊은 관련성을 갖는 데이터를 발굴하여 기계학습을 수행한다면 오차율을 좀 더 줄일 수 있을 것이다. 이는 예측의 정확도를 높이고 의미 있는 예측 변인을 추출할 수 있는 데이터 수집의 다각화가 필요함을 의미한다. 비록 본 연구에서는 학습분석에서 의미하는 학습데이터를 LMS 서버에서 직접 추출하지 못하였고 또한 충분한 학습 데이터 항목을 수집하지는 못하였지만, 향후 LMS와 연결된 학습예측이 가능해진다면 주요 학습활동 데이터에 대한 다양한 선택과 수집이 가능할 것이다. 현재 대학이나 기관의 서버에서 학습데이터를 불러온다는 것은 개인정보와 관련되어 여러 가지 어려운 점은 있으나 가까운 미래에 가능할 것으로 예상되며 그 외 서버상에 존재하지 않는 학습과 관련된 데이터를 어떻게 수집할 것인가도 고민해 보아야 할 것이다. 둘째, 본 논문은 기계학습을 적용한 학습 분석의 예측 가능성을 검토하기 위한 것으로 기계학습의 가장 기본적 분석방법인 Linear Regression을 사용하였다. 그러나 데이터의 형태와 결과도출의 방향을 고려하여 다음 연구에는 다양한 기계학습 방법을 사용할 필요가 있다. 즉, 예측의 정확도를 고려하여 정형데이터뿐만 아니라 비정형 데이터까지 포함하여야 할 것이며 이러한 데이터에 대한

LR(Linear Regression), ANN (Artificial Neural Netwerk), SVM(Support Vector Machine) 등을 적용하여 데이터 부족 및 불균형 문제 등의 제한적 환경을 극복하여야 할 것이다. 또한, 기계학습에 많이 사용되고 있는 Tensorflow는 그 기능들이 함수화되어 있어 도출하는 과정이 명확히 보이지 않으므로 좀 더 구체적인 프로세스를 알기 위해서는 Python 등을 고려하여야 할 것이다.

참 고 문 헌

[1] 한국교육학술정보원 연구보고 KR 2014-10. "교육 빅데이터를 활용한 아젠다 개발" 한국교육학술정보원. 2014.

[2] Romero, C., & Ventura, S. " Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol.40, No.6, pp. 601-618. 2010.

[3] Baker ,R .S .J. D., & Yacef ,K.. "The state of educational data mining in 2009: A review and future visions" Journal of Educational Data Mining, Vol.1, No.1, pp. 3-17. 2009.

[4] Mostow,J.,& Beck,J. "Some useful tactics to modify, map and mine data from intelligent tutors." Natural Language Engineering, Vol.12, No.2, pp. 195-208. 2006.

[5] Learning Impact Report. "2014 Learning Impact Report of Effective and Sustainable Technology Innovations." IMS Global Learning Consortium, 2004.

[6] Siemens. G, Long. P, "Penerrating the fog-analytics in learning and educations." Education Review, Vol.46, No.5, pp.30-32, 2011.

[7] Thiago, M. Barros., Placido A. Souza, Neto., & Luiz. Guedes.,(2019), "Predictive Models for Imbalanced Data: A School Dropout Perspective." Education Science, 9(275) : 3-17.

[8] Elias, T. "Learning Analytics: Definitions, Processes and Potential." *https://pdfs.semantic-scholar.org/732e/452659685fe3950b0e515a28ce89d9 c5592a.pdf,* pp.9-10, 2020.04.24.

[9] Arto, Hellas. et al. "Predicting Academic Performance: A Systematic Literature Review." Innovation and Technology in Computer Science Education, July 2-4, pp.177-178, 2018.

[10] Kumar. M, Singh. A. J, Handa. D, "Literature survey on student's performance predicton in education using data mining techniques." International Journal of Education and Management Engineering. Vol.6, No.6, pp.40-49, 2018.

[11] Xiao Hu, Christy WL Cheong, Wenwen Ding, and Michelle Woo. "A systematic review of studies on predicting student learning outcomes using learning analytics." In Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 528-529, 2017.

[12] 권신혜, 박 경우, 장병펄, 장병희, "기계학습의 미디어산업 적용:콘텐츠 평가 및 제작 자원을 중심으로." 한국콘텐츠학회논회지, 제19권, 제7호, pp.526-537, 2019.

[13] 이정은, 김다솜, 조일현. "동영상 기반 학습 환경에서 머신러닝을 활용한 행동로그의 학업성취 예측 모형 연구" 컴퓨터교육학회 논문지, 제23권, 제2호, pp.53-64, 2020.

[14] 안준용, 박상민, 김창복. "동계 전력수요예측을 위한 신경망 모델에 관한 연구." 한국정보기술학회논문지. 제15권, 제9호, pp.1-9, 2017.

[15] 손일락, 김연선. "청주지역 대학생들의 식생활라이프스타일에 따른 외식행동연구." 한국콘텐츠학회논문지, 제8권, 제11호, pp.347-355. 2008.

[16] H. Benli, "Performance prediction between horizontal and vertical source heat pump systems for greenhouse heating with the use of artificial neural networks," Heat and Mass Transfer, vol. 52, no.8, pp. 1707-1724, 2016.

[17] 이성희, 강지영. 한국인 대학생의 강의실 좌석 선택과 학업 성적의 연관성 연구. 한국교육문제연구, 제30권, 제2호, 215-233, 2012.

[18] 하만석, 안현철, “정형데이터와 비정형데이터를 동시에 고려하는 기계학습 기반의 직업훈련 중도탈락 예측 모형.” 한국콘텐츠학회논문지, 제19권, 제1호, pp.6-7, 2019.
[19] 윤지영, 유지윤, 이장석, “유튜브 브이로그 이용 동기 및 이용자 특성이 이용 만족 및 지속이용의도에 미치는 영향.” 한국콘텐츠학회논문지, 제20권, 제4호, pp.189-201, 2020.
[20] Murat, P. “Using Machine Learning to Predict Student Performance.” M. Sc. Thesis, pp 35. 2017
[21] 신종호, 최재원, “학습분석기반 대학 신입생 대상 학습부진 위험학생조기 예측 모델 개발 및 군집별 특성 분석.” 교육공학연구. 제35권, 제2호, pp.430- 431, 2019.
[22] 오영환, “선형회귀분석 기법을 이용한 고교야구투수의 투구속도 예측.” 한국지식정보기술논문지, 제14권, 제4호, pp.381-390, 2019.
[23] 유진은, “기계학습: 대용량/패널자료와 학습분석학 자료 분석으로의 활용.” 교육공학연구, 제35권, 제2호, pp.313-338, 2017.
[24] 김승환, 전성해, “딥러닝의 변수 중요도를 이용한 인공지능 기술 분석.” 한국지능시스템학회, 제29권, 제1호, pp.70-75, 2019.
[25] 조명희, 김은진, 이현우, “학사경고자 예측을 위한 학습분석학적 모형 탐색.” 제34권, 제4호, pp.877-900, 2018.

저 자 소 개

김 연 희(Yeon-Hee Kim) 정회원

- 2005년 : 호서대학교 안전시스템공학과(박사)
- 2006년 3월 ~ 현재 : 호서대학교 전자디스플레이공학부 교수
- 2018년 8월 : 충북대학교 교육학과 박사 수료

〈관심분야〉 : 학습분석, 빅데이터, 인공지능

임 수 진(Soo-Jin Lim) 정회원

- 2016년 : 충북대학교 교육학과(교육학석사)
- 2018년 : 충북대학교 교육학과(박사 수료)
- 2018년 ~ 현재 : 청주대학교 교육혁신원

〈관심분야〉 : 교육성과, 학습분석, 빅데이터