딥러닝기법을 활용한 학습성과분석

오정훈*, 유헌창**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 정보대학 컴퓨터학과
e-mail : nal_ra@korea.ac.kr

Learning Performance Analysis Using Deep Learning

Jeong-Hoon Oh*, Heonchang Yu**
*Dept. of Big Data Science, Korea University
**Dept. of Computer Science and Engineering, Korea University

요 약

본 연구의 목적은 교육관리시스템(LMS)에서의 학습활동로그를 바탕으로 학습성과 영향도를 분석하고 이를 예측하기 위한 모델을 개발하는데 있다. 연구방법은 먼저 상관분석을 사용하여 유의미한 변수를 선정하였으며, 딥러닝을 사용하여 예측 모델을 생성하였다. 모델 생성 결과 테스트 데이터 셋에 대해 약 84%의 정확도로 학습성과를 예측할 수 있었다. 본 연구는 온라인 교육환경에서 빅데이터와 인공지능을 적용할 수 있는 새로운 관점을 제공할 것으로 기대한다.

1. 서론

가트너의 '2018 eLEARNING PREDICTIONS HYPE CURVE' 에 따르면 Artificial Intelligence and Predictive Modeling (AI 와 예측모델링) 기술은 도입기를 지나 부풀려진 기대의 정점 단계에 있는 추세이다[1]. 이러한 시대적 흐름에 순응하여 미국, 영국 등 교육 선진국을 중심으로 교육과 첨단 ICT 기술이 접목된 에듀테크 산업이 이러닝 산업의 수요를 대체 중에 있다. 미국의 경우 2015 년 에듀테크 스타트업에 투자된 금액만 18 억 5 천만 불이며 총 198 개의 투자 건이 있었다. CB 인사이츠와 KPMG 가 공동 발행하는 벤처펄스에 의하면 2015 년 4/4 분기에만 10 억 불이 넘게 투자되어 3/4 분기 2 억 9,500 만 불에 비해 300% 가까운 성장을 했다[2]. 이처럼 세계 에듀테크 기업의 성장세에 비해 우리나라 에듀테크 시장은 본격 성장기를 맞지 못했다. 핀테크, O2O 시장 등과 비교했을 때는 아직은 미미한 수준이다. 교육시장이 에듀테크 분야로 변화하는 시대적 흐름에 발맞춰 우리나라 교육현장에도 에듀테크 도입이 필요하다. 매년 개최되는 ATD ICE 에서 최근 자주 등장되고 있는 keyword 또한 빅데이터와 AI 이며, 이를 활용한 교육으로 학습자의 학습경험을 활용한 맞춤형 과정설계와 추천 및 성과예측 등의 실제 적용사례 또한 소개되고 있다. 본 연구에서는 동영상기반 마이크로러닝 LMS (Learning Management System) 에서 학습자의 온라인에서의 수강활동을 수집하고 이를 딥러닝기법을 활용하여 학습자에게는 온라인 교육성과를 보다 정확하게 예측하여 효과적인 학습을 도울 수 있는 시스템을 제안하고자 한다.

2. 관련연구

2.1 LMS (Learning Management System)

사이버 공간에서 학습자가 원하는 학습 진행을 위해서는 교육과정을 개설하고 수강신청을 하는 등 교사와 학생이 학습에 참여하기 위한 준비과정이 필요하다. 준비과정이 끝난 후 실제 학습이 이루어지는 과정에서는 학습자의 학습과정을 추적하고 학습이력을 관리하여 학습자 개인에 대한 맞춤형 학습을 제공하게 된다. 이와 같이 온라인 학습에서 필요한 학급편성 기능, 협동학습 기능, 출결관리 기능, 게시판 기능 등이 LMS 의 주요기능이라 할 수 있다. LMS 의 기능이 고도화될수록 학생의 개별학습을 위한 맞춤형 학습 환경을 효과적으로 구성할 수 있다[3].

<표 1> LMS의 주요 기능

학습지원	교수지원	운영지원
교과학습 시험평가 과제평가 교과상담 부가학습	과정개설 강의실관리 시험평가	교육과정관리 학습운영 수강관리 사용자관리

2.2 딥러닝 기법

딥러닝(Deep Learning)은 기계학습의 하나인 인공신경망이 발전된 형태의 인공 지능이다. 다층 퍼셉트론을 통한 비선형 문제의 해결, 역전파(Back Propagation) 알고리즘을 통한 학습, 복잡한 문제 해결을 위한 신경망의 층수 확대, 최적화 과정에서의

학습 효율성을 개선한 경사감소법(Gradient Descent Algorithm)의 고도화, 경사감소소멸(Vanishing Gradient Descent) 문제에 대응한 활성화 함수(Activation Function)의 개선, 과적합(Overfitting)을 막기 위한 규제화(Regulation) 기법 개발 등 신경망 모델과 학습 알고리즘이 지속적으로 향상되면서 딥러닝이라고 불리게 되었다. 통상 딥러닝은 신경망의 층이 2~3개 이상인 것을 통칭하며, 10개 이상이면 아주 깊은 학습(Very Deep Learning)이라고 한다. 딥러닝에서 깊이, 층의 개수는 은닉층(Hidden Layer) 개수를 의미한다. 또한 뉴런에 해당하는 신경망계층의 노드와 연결되는 가지의 수도 많아야 한다[4].

2.3 피어슨 상관계수

상관계수는 두 변량 사이의 상관관계의 정도를 나타내는 수치이며 다음과 같은 공식으로 계산한다.

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

여기서 μ_X, μ_Y = X, Y의 평균
σ_X, σ_Y = X, Y의 표준편차

이러한 상관계수 중 보편적으로 사용되는 피어슨 상관계수는 변량 x와 y간의 선형 관계성의 정도를 -1에서 1사이로 나타내는 방법이다[5].

<표 2> 피어슨 상관계수 범위와 선형관계

범위	선형관계
-1.0 <= r < -0.7	강한 음적 선형관계
-0.7 <= r < -0.3	뚜렷한 음적 선형관계
-0.3 <= r < -0.1	약한 음적 선형관계
-0.1 <= r < +0.1	거의 무시될 수 있는 선형관계
+0.1 <= r < +0.3	약한 양적 선형관계
+0.3 <= r < +0.7	뚜렷한 양적 선형관계
+0.7 <= r <= +1.0	강한 양적 선형관계

3. 연구 방법

3.1 연구 자료

본 연구는 S사의 2016년 01월 01일부터 2018년 08월 31일까지의 교육관리시스템(LMS)의 자료를 이용하여 분석하였다.

이 자료는 해당기간 동안의 전체 임직원의 온라인시험결과, 사이트접속로그, 자격취득이력, 콘텐츠수료이력, 콘텐츠수강로그, 콘텐츠수강후기, 콘텐츠공유추천이력 등으로 구성되어 있다.

<표 3> 연구자료 구성

Sources	Description	Rows
온라인시험결과	문항별 정답 여부	5,123,339
사용자 정보	개인정보제외 특성정보	22,811
사이트접속로그	로그인, 로그아웃 일시	6,473,503
자격취득이력	자격 취득/만료 일시	91,925
콘텐츠수료이력	콘텐츠 수료 일시	2,867,036
콘텐츠수강로그	콘텐츠 수강 일시	8,443,771
콘텐츠수강후기	콘텐츠 후기 등록 일시	202,427
콘텐츠공유이력	콘텐츠 공유 일시	17,402

딥러닝 알고리즘을 적용하기 전 <표 3> 의 자료를 사용하여 전처리 단계를 거친다. 이를 통해 가공된 변수는 <표 4>와 같다.

<표 4> 전처리 이후 변수 구성

Variable	Label
TC	월간 문제풀이 건수
YC	월간 정답 건수
WM	수검한 사용자의 근속 개월 수
AG	수검한 사용자의 나이
S12	수검 전 12개월 간 사이트 체류시간
S06	수검 전 6개월 간 사이트 체류시간
S03	수검 전 3개월 간 사이트 체류시간
S01	수검 전 1개월 간 사이트 체류시간
P12	수검 전 12개월 간 콘텐츠 수료 건수
P06	수검 전 6개월 간 콘텐츠 수료 건수
P03	수검 전 3개월 간 콘텐츠 수료 건수
P01	수검 전 1개월 간 콘텐츠 수료 건수
V12	수검 전 12개월 간 콘텐츠 수강 건수
V06	수검 전 6개월 간 콘텐츠 수강 건수
V03	수검 전 3개월 간 콘텐츠 수강 건수
V01	수검 전 1개월 간 콘텐츠 수강 건수
QC	수검 전 자격증 취득 건수
RC	수검 전 콘텐츠 후기 등록 건수
SC	수검 전 콘텐츠 공유 건수

3.2 연구 방법

<표 4>의 월간 문제풀이 건수(TC)와 월간 정답 건수(YC)에서 월간 정답율을 계산하였다. 이를 구간별로 나누어 80%이상을 상위, 40%이하를 하위, 그 사이를 중위로 나누어 종속변수로 설정하고 이는 사용자의 학습성과를 나타내는 변수이다. 나머지 교육관리시스템에서의 활동으로부터 추출한 변수를 독립변수로 설정하여 딥러닝 모델을 생성한다. 본 연구에서는 교육관리시스템에서의 활동을 바탕으로 학습자의 학습성과를 예측하고자 한다. 이를 위해 입력한 독립변수 데이터를 바탕으로 성취도 구간의 값을 예측하는 모델을 생성한다.

4. 학습성과 예측모델 개발

4.1 변수 간 상관관계

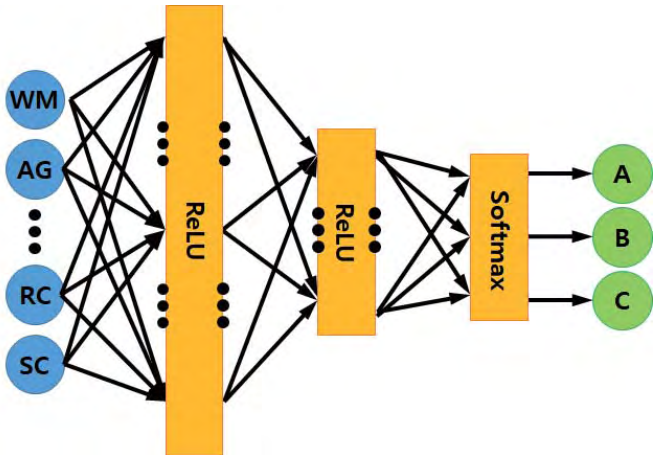
상관관계 분석을 통해 종속변수인 월간 문제풀이 건수(TC), 월간 정답 건수(YC)와 나머지 독립변수들과의 상관관계를 살펴보았다. <표 5>에서 볼 수 있듯이 상관관계 분석 결과 월간 문제풀이 건수(TC), 월간 정답 건수(YC)와 다른 독립변수들사이에 유의한 상관이 나타났다. 특히, 수검 전 1개월 간 사이트 체류시간(S01)이 다른 독립변수들보다 더 높은 상관을 가지고 있는 것으로 나타났다.

4.2 딥러닝 모델 구성

본 연구에서는 데이터 전처리 과정에서 Scikit-Learn 라이브러리에 StandardScaler를 사용하여 숫자형 변수에 선형변환을 적용하였다. 이를 통하여 자료의 오버플로우(overflow)나 언더플로우(underflow)를 방지하고 독립 변수의 공분산 행렬의 조건수(condition number)를 감소시켜 최적화 과정에서의 안정성 및 수렴 속도를 향상시켰다. 또한, LabelEncoder를 사용하여 범주형변수을 숫자형으로 변경해 주었으며 이 변수에 one-hot-encoding을 적용하였다. 시각화를 위해서는 matplotlib 라이브러리를 사용하였다. 딥러닝 모델을 생성하기 위하여 딥러닝 프레임워크인 텐서플로(Tensorflow)와 케라스(Keras)를 사용하였다. 그리고 데이터분석 도구인 numpy와 pandas 라이브러리를 사용하였다. 모델의 과적합(over fitting)을 피하기 위해 데이터 셋을 학습데이터와 테스트 데이터로 7:3의 비율로 구분하여 사용하였다.

트레이닝과 테스트를 통해 구축한 가장 높은 정확도의 딥러닝 layer 는 (그림 1)과 같다. Input은 17개의 독립변수로 구성되고 이를 첫번째 hidden layer에서 input size가 17인 ReLU, 두번째 hidden layer에서 input size가 8인 ReLU, 최종 layer에서는 3개의 구간의 정답률로 classification 되어야 하기에 softmax 로 구성하였다.

Cost Function은 종속변수의 카테고리가 3개이상이므로 categorical_crossentropy를 사용하였으며 Optimizer는 최근 딥러닝에서 많이사용되는 Adam을 사용하였다.



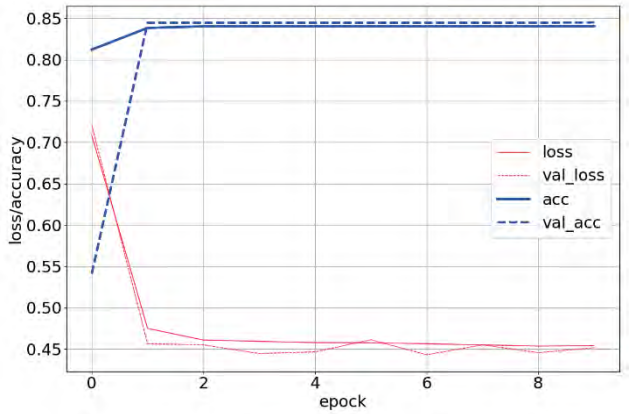
(그림 1) 딥러닝 layer 설계

<표 5> 변수 간 상관관계

	WM	AG	S12	S06	S03	S01	P12	P06	P03	P01	V12	V06	V03	V01	QC	RC	SC
TC	0.05	0.01	0.08	0.19	0.24	0.27	0.15	0.18	0.18	0.17	0.04	0.17	0.18	0.17	-0.06	0.11	0.00
YC	0.06	0.01	0.09	0.20	0.25	0.29	0.17	0.20	0.20	0.19	0.05	0.18	0.19	0.19	0.07	0.12	0.00
WM		0.54	0.26	0.07	-0.15	-0.19	0.02	0.08	0.09	0.09	0.04	-0.15	-0.17	-0.17	0.45	0.01	0.06
AG			0.19	0.02	0.03	0.07	0.09	0.02	0.01	0.01	0.06	0.01	0.03	0.05	0.26	0.05	0.04
S12				0.58	0.39	0.19	0.58	0.26	0.15	0.05	0.27	0.32	0.17	0.04	0.53	0.25	0.11
S06					0.85	0.55	0.46	0.53	0.38	0.20	0.16	0.67	0.49	0.25	0.05	0.25	0.05
S03						0.75	0.36	0.50	0.51	0.31	0.12	0.63	0.65	0.40	0.09	0.19	0.03
S01							0.24	0.37	0.43	0.48	0.07	0.46	0.53	0.61	-0.20	0.13	0.01
P12								0.71	0.54	0.35	0.41	0.61	0.46	0.29	0.17	0.43	0.09
P06									0.83	0.56	0.19	0.87	0.73	0.49	-0.03	0.29	0.04
P03										0.73	0.14	0.71	0.88	0.64	0.08	0.21	0.03
P01											0.09	0.46	0.62	0.88	-0.10	0.13	0.01
V12												0.18	0.14	0.08	0.08	0.18	0.01
V06													0.82	0.53	0.06	0.26	0.02
V03														0.72	-0.14	0.19	0.01
V01															-0.18	0.11	0.00
QC																0.03	0.07
RC																	0.08

4.3 딥러닝 학습 결과

배치 사이즈를 10 으로 설정하고, 10 epock 을 실행한 결과 전체 테스트 데이터와 실제 학습성과값의 오차 평균 값은 0.44976 으로 나타났다. softmax crossentropy loss 를 사용할 경우 오차(loss)값이 의미하는 바는 단순히 i 번째 index 가 정답일 때에, i 번째 index 에 해당하는 softmax 값의 log 평균이므로, 이는 학습성과에 대한 예측이 평균적으로 $e^{-0.44976} = 0.64$ 정도의 softmax 확률값을 가졌다는 의미이다. (그림 2)에서는 Training data set 과 Test data set 의 epock 에 대한 loss 와 accuracy 추이를 볼 수 있다. 1 번째 epock 에서 accuracy 와 loss 모두 특정 값에 수렴하는 결과를 볼 수 있다. 이는 학습시킨 모델이 과적합(overfitting)이 없이 정상적으로 학습되었다는 것을 알 수 있다. 결과적으로 학습시킨 모델의 정확도(accuracy)는 84.463%임을 확인할 수 있다.



(그림 2) Training data set과 Test data set의 loss와 accuracy

5. 결론 및 시사점

본 연구에서는 교육관리시스템(LMS)에서의 학습활동로그를 바탕으로 학습성과 영향도를 분석하고 이를 예측하기 위한 모델을 개발하였다. 상관관계 분석으로 17개의 독립변수를 선정하였으며 종속변수와의 관계를 3개의 hidden layer를 가지는 딥러닝 모델로 학습시켜 84.463%의 정확도를 가지는 모델을 도출하였다.

본 연구의 시사점은 학습자 측면에서 학습성과예측, 교수자 및 운영자 측면에서 학습성과에 효과적인 콘텐츠 개발 및 도입을 가능하게 해줄 수 있는 새로운 관점을 제공할 수 있는 교육관리시스템(LMS)를 제시할 수 있을 것으로 기대한다.

참고문헌

[1] A.Jon H.Andy B.Amber (2017) 2018 eLearning Predictions - Hype Curve
[2] 한상기 (2016) KISA Report - 주요국 에듀테크 산업과 정책 현황
[3] LMS https://ko.wikipedia.org/wiki/학습_관리_시스템
[4] 김의중 (2016) 알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문
[5] 상관분석 https://ko.wikipedia.org/wiki/상관_분석