교육데이터마이닝을 이용한 학부모 학교 만족도 예측에 관한 연구

양영보, 유헌창
고려대학교 컴퓨터정보통신대학원
e-mail : {diddudqh79, yuhc}@korea.ac.kr

A Study on Prediction of Parent School Satisfaction Using Educational Data Mining

YoungBo Yang, Heonchang Yu
Graduate School of Computer & Information Technology, Korea University

요 약

학습관리시스템의 도입으로 학습자들은 다양한 형태로 학습하게 되고 데이터를 남기게 된다. 교육데이터마이닝은 다양한 형태로 기록되는 교육 데이터를 분석해서 유의미한 정보를 찾아 내는 방법이다. 교육데이터마이님을 활용하면 학생 개인의 학습성과 향상에 도움을 주거나 학습성과 예측 결과를 참고하여 부족한 부분을 지원해 줄 수도 있다. 기존 연구에서는 학습자의 행동 영역 특징이 학습성과에 영향을 끼친다는 것을 검증하기 위하여 나이브 베이즈, 의사결정트리, 신경망 기계학습 알고리즘으로 데이터를 분석했다. 따라서 본 연구에서는 기존 연구를 확장하여 학습자의 행동 영역 특징이 학부모 학교 만족도에 영향을 끼치는지 여부를 확인하는 실험을 수행했으며 kNN, 의사결정트리, SVM 기계학습 알고리즘으로 데이터를 분석하였다. 분석결과 학습자의 행동 영역 특징이 학부모 학교 만족도에 영향을 미치는 것을 확인했다.

1. 서론

교육데이터마이닝은 오프라인수업 방식에서 온라인 시스템이 결합되면서 나타난 분야이며 온라인 수업에 사용되는 학습관리시스템을 이용하여 학습자의 다양한 데이터를 수집하고 데이터를 분석해서 학습활동에 도움이 되는 유의미한 정보를 얻어 내는 것이다. 수집되는 데이터는 학습자의 로그인 횟수, 학습동영상 시청시간, 온라인 토론참여 횟수 등 시스템을 이용하는 전반적인 행동에 관한 것들이 수집된다. 이러한 데이터를 수집하기 위해서는 학습관리시스템을 구축해야 하는데 신규로 구축하는 것은 많은 비용을 필요로 하기 때문에 moodle, kalboard 360 등 오픈소스를 활용해서 학습관리시스템을 구축한다.

교육데이터마이님을 활용한 연구는 다양하며 학습자 개인에 학습성과 향상에 도움을 주거나 또는 학습성과 예측 결과를 참고하여 부족한 부분을 지원해 주는 연구 등이 많이 이루어진다. [1]에서는 kalboard 360 을 활용하여 학생들 480 명에 데이터를 4 개의 특징 영역(총 16 개의 세부 특징)으로 학습성과에 영향을 끼치는 특징 영역을 분석했다. 그리고, 4 개의 특징 영역 중 행동 특징 영역 4 개의 특징(학습자가 교실에서 손을 드는 횟수, 온라인 수업자료를 열어 본 횟수, 공지사항을 읽어 본 횟수, 그룹 토론에 참여한 횟수)이 학습성과에 많은 영향을 끼친다고 말했다.

본 연구는 [1]에 연장으로 행동 특징 영역 4 개의 특징이 학부모 학교 만족도에 얼마나 영향을 끼치는지를 알아보기 위한 연구이다. 데이터 분석에는 R 을 사용하였으며 기존 데이터에서 행동 특징 영역과 학부모 학교 만족도 이외에 데이터를 삭제시키는 전처리 작업 후 kNN, 의사결정트리, SVM 기계학습 알고리즘을 활용하여 학부모 학교 만족도를 예측해 보았다.

2. 관련연구

온라인 시스템이 오프라인 수업에 접목이 되면서 수업시간에 전자칠판, 전자교과서를 사용하고 학습관리시스템으로 출결관리, 성적관리, 동영상강좌 학습이 이루어지며 이런 모든 것들이 데이터로 저장된다. 교육데이터마이닝은 이렇게 만들어진 학습자 데이터를 분석해서 학습에 도움이 되는 정보를 찾아 내는 것이다. [2]에서는 kalboard 360 학습관리시스템에 xAPI[3]를 활용하여 학생들에 데이터를 수집하였다. 수집된 데이터는 인구 통계 학적 영역, 학문적 배경 영역, 행동적 특징 영역 총 14 개의 특징으로 데이터 전처리 작업을 하였고, 학생들에 성적은 3 개의 영역(Low Level : 0~69, Middle Level : 70~89, High Level : 90~100)으로 분류하였다. [2]에서 학생들 성적에 영향을 끼치는

특징을 분류하기 위하여 나이브 베이즈, 의사결정트리, 신경망 기계학습 알고리즘을 활용하여 분석했다. 분석시 행동적 특징 영역과 비행동적 특징 영역으로 나누어서 실험했으며 결과는 행동적 특징 영역이 학습성과에 영향력이 높다고 나왔다. 행동적 특징 영역을 나이브 베이즈, 의사결정트리, 신경망 분석 시 72.5%, 61.3%, 73.8% 예측 정확도(Accuracy)를 나타냈다. [1]에서는 [2] 연구에 연장으로 학부모 참여 영역 특징 2 개(설문지 참여 여부, 학부모 학교 만족도)를 추가하였고 특징 분석 시 앙상블 알고리즘을 추가하여 나이브 베이즈, 의사결정트리, 신경망 분석 정확도를 72.2%, 77.7%, 79.1%로 향상시켰다.

3. 설계 및 구현

3.1 전처리

[1]에서 사용한 데이터를 [3]에서 csv 파일로 제공받았다. R 로 csv 데이터를 불러온 후 행동적 특징 영역 4 가지와 학부모 학교 만족도 외에 데이터는 필요없기 때문에 R 에서 제공하는 데이터프레임 메소드를 사용하여 필요없는 데이터를 삭제시키는 전처리 작업을 하였다.

3.2 행동 영역 특징간에 연광성

[2]에서 행동 영역 특징 중 학습자가 교실에서 손을 드는 횟수가 가장 중요한 특징으로 판단했다. 학습자가 교실에서 손을 드는 횟수를 기준으로 나머지 3 개 특징간에 연광성은 없는지 확인해 보았다.

<표 1> 카이제곱 분석 결과

	카이제곱 분석
	p-value
온라인 수업자료를 열어 본 횟수	2.2e-16
공지사항을 읽어 본 횟수	9.134e-11
그룹토론에 참여 한 횟수	2.453e-07

<표 1> 결과를 보면 4 개의 특징간에 카이제곱 결과 값은 모두 p-value 0.05 이하로 서로 연관성은 없는 것을 확인했다.

3.3 학부모 학교 만족도 예측

독립변수로 학부모 학교 만족도를 종속변수로 행동 영역 특징 4 가지를 선택하고 kNN, 의사결정트리, SVM 기계학습 알고리즘으로 데이터를 분석하였다. 전체 데이터는 480 개이며 80%인 384 개의 데이터를 트레이닝 데이터로 사용 했으며 20%인 96 개의 데이터를 테스트 데이터로 사용했다.

4. 분석

380 개의 트레이닝 데이터를 kNN, 의사결정트리, SVM 기계학습 알고리즘으로 각각 트레이닝 시킨 후 96 개의 테스트 데이터로 예측 확률을 검증해 보았고 예측 분석 결과는 [5]에서 제시한 방법을 기준으로 했다. <표 2>와 <표 3>은 [5]에서 제시한 분류 결과표와 평가방법이다.

<표 2> 분류 결과표

		예측	
		Positive	Negative
실제 결과	Positive	TP	FN
	Negative	FP	TN

분류는 TP, FN, FP, TN 으로 분류하며 실제 결과와 예측에 일치 여부에 따라 나뉘어진다. TP, TN 은 예측이 맞은 경우이며 FN, FP 는 예측이 틀린 경우이다.

<표 3> 평가방법

Accuracy	TP+TN / TP+FN+FP+TN
Precision	TP / TP+FP
Recall	TP / TP+FN
F-measure	2*(Precision* Recall/ Precision+ Recall)

평가방법은 Accuracy, Precision, Recall, F-measure 을 사용하며 Accuracy 은 전체 결과 중에 예측이 맞은 비율, Precision 은 총 Positive 예측 중에 실제 Positive 인 비율, Recall 은 실제 결과값 중 총 Positive 중에 예측이 Positive 인 비율, F-measure 은 Precision 과 Recall 을 기반으로 두 결과값에 관계를 판단하는 수치이다.

R 에서 kNN, 의사결정트리, SVM 알고리즘에 따라 각각 나온 예측 결과 값을 <표 2> 분류표 기준으로 정리한 후 <표 3> 에 평가방법을 적용해서 <표 4>에 정리하였다.

<표 4> 학부모 학교 만족도 예측 분석 결과

평가방법	kNN	의사결정트리	SVM
Accuracy	84.3	63.5	**85.4**
Precision	86.7	72.7	**87.8**
Recall	94.7	66.6	**94.7**
F-measure	90.5	69.5	**91.1**

<표 4>에 분석결과를 보면 SVM 이 가장 좋은 예측 정확도(Accuracy)를 나타냈으며 kNN 도 SVM 보다 조금 낮지만 좋은 정확도를 나타냈다. 상대적으로 의사결정트리는 정확도가 낮게 측정이 됐다.

5. 결론 및 향후 연구

본 연구는 학습자의 행동 특징이 학부모의 학교 만족도에 영향을 끼치는지 알아 보기 위한 실험이다. SVM 예측 정확도가 85% 이상으로 학습자의 행동 특징이 학부모의 학교 만족도에 영향일 끼친다는 것을 확인했다.

본 연구의 향후 연구는 신경망, 딥러닝 알고리즘을 사용했을 때 예측 정확도 변화를 확인해 볼 것이며, 해외 데이터가 아닌 국내 교육기관에 패널 데이터를 분석하여 학습성과 예측에 관한 연구를 수행할 계획이다.

참고문헌

[1] Elaf Abu Amrieh, Thair Hamtini, Ibrahim Aljarah,. “Mining Educational Data to Predict Student's academic Performance using Ensemble Methods”. 2016

[2] Elaf Abu Amrieh, Thair Hamtini, Ibrahim Aljarah,. “Preprocessing and Analyzing Educational Data Set Using X-API for Improving Student’s Performance”. 2015

[3] experience api. https://www.xapi.com

[4] kaggel. https://www.kaggle.com

[5] POWERS, D.M.W. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. 2011