의사결정트리 알고리즘을 이용한 학생진로 예측 프로그램의 설계

김근호 · 정종인 · 김창석 · 강신천 · 김의정

공주대학교 컴퓨터교육과

Design of the student Career prediction program using the decision tree algorithm

Geun-Ho Kim · Chong-In Jeong · Chang-Seok Kim · Shin-Chun Kang · Eui-Jeong Kim

Dept. of Computer Educaion, Kongju National University

E-mail : rmsh3@kongju.ac.kr

요 약

최근 IT분야에서는 빅 데이터를 활용한 인공지능이 크게 대두되고 있다. 이와 같이 빅 데이터를 효과적으로 처리하기 위한 서비스 또는 기술에 대하여 다양한 연구가 진행되고 있다. 교육분야에 있어서도 학생들에 대한 빅 데이터가 존재하나 이러한 데이터를 단순히 수집, 조회, 저장하는 단순처리과정을 거칠 뿐이다. 향후 인공지능이나 기계학습, 통계분석 등을 폭 넓게 이용하여 교육분야의 빅 데이터에서 의미 있는 규칙이나 패턴 및 관계를 찾아내어, 실제 학생들에게 도움이 되는 데이터를 생산 지능적인 활용이 요구되고 있다. 이에 따라서 본 연구에서는 학생들의 수업 관찰을 통한 데이터를 의사결정트리 알고리즘을 이용하여 학생들의 진로를 예측하는 프로그램을 설계하고자 한다. 진로예측 프로그램을 통하여 학생들의 상담에 활용 진로를 제시하고 또한 희망 진로에 따른 수업 태도 및 방향을 제시하는데 도움이 될 것으로 사료된다.

ABSTRACT

In recent years, artificial intelligence using big data has become a big issue in IT. Various studies are being conducted on services or technologies to effectively handle big data. The educational field, there is big data about students, but it is only a simple process to collect, lookup and store such data. In the future, it makes extensive use of artificial intelligence, machine learning, and statistical analysis to find meaningful rules, patterns, and relationships in the big data of the educational field, and to produce intelligent and useful data for the actual students. Accordingly, this study aims to design a program to predict the career of students using a decision tree algorithm based on the data from the student's classroom observations. Through a career prediction program, it is believed to be helpful to present application paths to students ' counseling and to also provide classroom behavior and direction based on the desired courses.

키워드

의사결정트리, 진로예측, 인공지능, 빅데이터

I. 서 론

최근 IT분야에서는 빅 데이터를 활용한 인공지능이 크게 대두되고 있다. 이와 같이 빅 데이터를 효과적으로 처리하기 위한 서비스 또는 기술에 대하여 다양한 연구가 진행되고 있다. 교육분야에 있어서도 학생들에 대한 빅 데이터가 존재하나 이러한 데이터를 단순히 수집, 조회, 저장하는 단순처리과정을 거칠 뿐이다.

향후 인공지능이나 기계학습, 통계분석 등을 폭 넓게 이용하여 교육분야의 빅 데이터에서 의미 있는 규칙이나 패턴 및 관계를 찾아내어, 실제 학생들에게 도움이 되는 데이터를 생산 지능적인 활용이 요구되고 있다.

이에 따라서 본 연구에서는 학생들의 수업 관찰을 통한 데이터를 의사결정트리 알고리즘을 이용하여 학생들의 진로를 예측하는 프로그램을 설계하고자 한다.

II. 의사결정트리 알고리즘

2.1 의사결정트리 알고리즘 개념

의사결정트리는 주어진 데이터를 분류하고 규칙을 생성하는 모형이다. 플로우 차트와 유사하며, 루트 노드와 리프 노드로 구성되어 있다. 루트 노드는 입력된 데이터의 속성을 분류하여 결정한다. 리프 노드는 결정의 결과로 더 이상 분리되지 않는 노드를 의미한다. 가장 첫 단계의 결정 노드는 뿌리 노드이며 다음 그림의 루트 노드1에 해당 된다. 하나의 대안이 여러 개의 리프 노드를 통해 결정될 수 있으며, 이를 규칙으로 정리 할 수 있다. 뿌리 노드에 가까운 단계의 결정 노드일수록 목표한 대안을 설명하기 용이한 변수이다.[1]

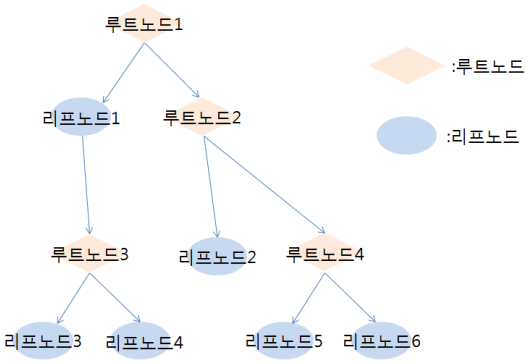

그림 1. 의사결정트리 예시.

의사결정트리모델은 모델의 해석이 쉽고 높은 효율성을 가진다.[1] 비모수적인 분석 방법으로 분석 자료의 선형성, 정규성, 등분산성의 가정을 필요로 하지 않는다. 데이터의 노이즈와 이상치에 큰 영향을 받지 않는다.

의사결정트리는 CAHID, CART, ID3, C4.5, C5.0 등의 다양한 알고리즘이 존재한다. 각각의 의사결정트리 알고리즘에 따라 결정 노드의 처리 할 수 있는 데이터의 종류와, 분류기준, 분류방법이 달라진다. CAHID와 CART 알고리즘은 통계적 기법을 기반으로 지니 계수(Gini index), 카이제곱(Chi-Squared statistics), 이득 비율(Gain rate)의 개념을 사용한다.[2]

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 (p_i)$$

S = 데이터 샘플
π = 범주 i에 속할 비율

수식 35 엔트로피 및 지니계수 계산식

ID3, C4.5, C5.0알고리즘은 순차적으로 개발된 알고리즘으로, 엔트로피와 정보이득(information gain)의 평가지수를 적용한다. 따라서 인공지능과 머신러닝의 접근 방법에 속한다.[3]

본 연구는 학생들의 수업행동 관찰 표준안을 만들고 이 표준안에 따라 학생들의 행동 및 진로를 예측하고 추천 하는 목적을 가지므로, 본 연구에서는 가장 보편적인 CHAID알고리즘을 사용하고자 한다.

2.2 의사결정트리 알고리즘별 비교

표 1. 의사결정트리 알고리즘의 비교.

알고리즘	평가지수	비고
ID3	엔트로피	다지분리
C4.5	정보이득	다지분리 및 이진분리
C5.0	정보이득	통계적접근
CHAID	카이제곱	
CART	분상의차이	항상 2진분리

III. 의사결정트리를 이용한 학생진로 예측 알고리즘

학생의 진로예측 및 추천 전체 처리과정은 다음 그림과 같다.

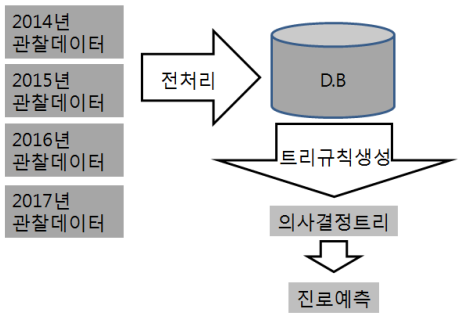

그림 2. 순서도.

첫 번째 관찰데이터를 의사결정 트리로 제작을 위한 데이터 전처리 과정 두 번째는 데이터를 입력받고 저장하는 데이터 처리 과정, 세 번째는 데이터를 변환하여 의사결정 트리를 제작하는 과정 그리고 마지막으로 의사결정트리에서 규칙을 추출하여 진로를 예측 하는 과정을 거친다.

3.1 전처리 과정

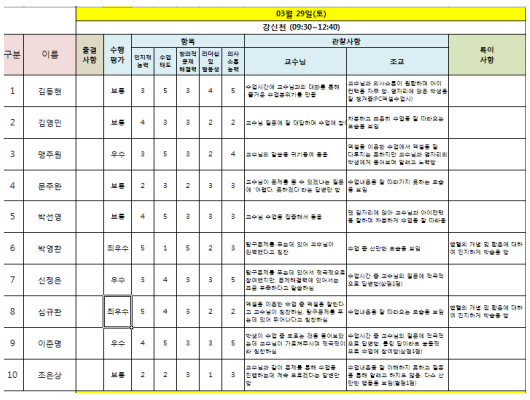

그림 3. 학생관찰 기록지 데이터.

위의 그림은 학생 관찰일지의 일부이다 관찰일지는 교사 및 학생들의 수업중의 행동 및 내용을 관찰교사가 기록하여 남겨둔 데이터로 관찰교사들이 학생들의 태도 및 행동들을 자세히 기술해 놓았다. 이는 데이터로서는 훌륭하지만 의사결정트리로 만들기 위해서는 다음과 같이 몇가지 항목을 정리하여 데이터를 변환 전처리 하였다.

표 2. 전처리 데이터 항목.

항목	변수
학생ID	학생별 분류
수행평가	척도 1~5(취우수~저조)
수업	척도 1~5(취우수~저조)
학습활동	척도 1~5(취우수~저조)
인지적능력	척도 1~5(취우수~저조)
수업태도	척도 1~5(취우수~저조)
창의적문제해결력	척도 1~5(취우수~저조)
리더쉽및의사소통	척도 1~5(취우수~저조)
오전오후	척도 1~5(취우수~저조)
희망진로	직업분류표

희망진로는 학생들의 상담일지를 바방으로 학생들이 기록한 희망진로를 2018년 취업알선 직업 분류표에 따라서 아래와 같이 크게 10가지로 분류하였다.

표 3. 직업군 분류표.

분류번호	직업군
0	경영・사무・금융・보험직
1	연구직 및 공학 기술직
2	교육・법률・사회복지・경찰・소방직 및 군인
3	보건・의료직
4	예술・디자인・방송・스포츠직
5	미용・여행・숙박・음식・경비・청소직
6	영업・판매・운전・운송직
7	건설・채굴직
8	설치・정비・생산직
9	농림어업직

전처리된 데이터는 아래와 같다.

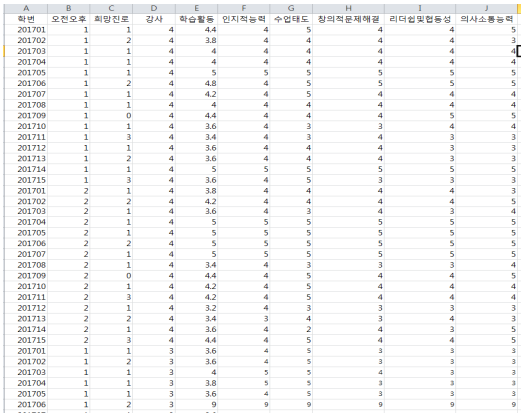

그림 4. 전처리 데이터.

3.2 의사결정 트리 과정

우리가 예측하고자하는 데이터는 학생들이 제시한 희망진로를 바탕으로 관찰된 학생들의 행동 및 수업 내용을 가지고 진로를 추천 및 예측하고자 하기 때문에 입력된 데이터 중에 희망진로를 변수로 잡고 다양한 항목으로 이진 트리를 구성한다.

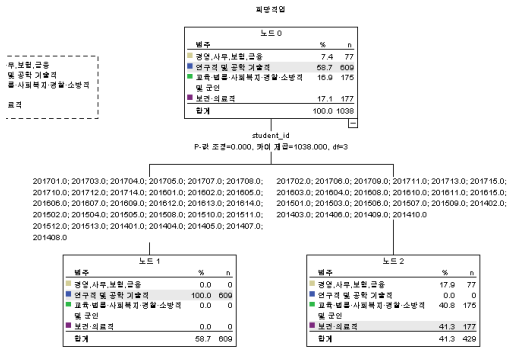

그림 5. 학생id를 기준으로 의사결정트리.

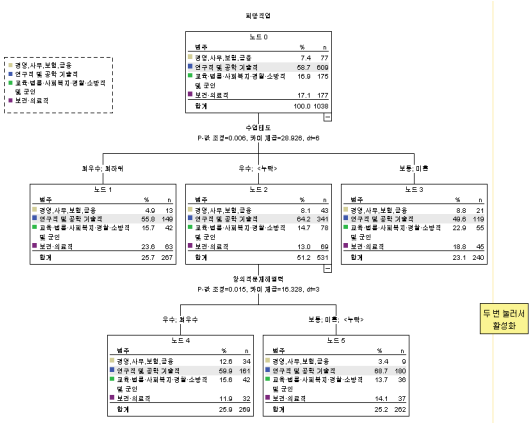

그림 6. 수업 태도 및 창의적문제해결력을 바탕으로 의사결정트리.

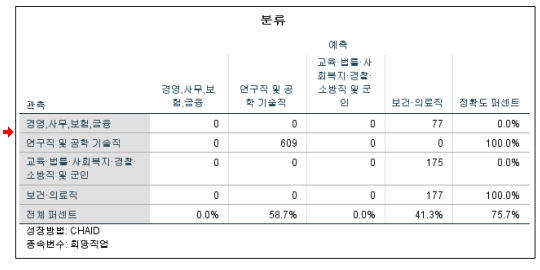

그림 7. 의사결정트리를 바탕으로 분류 및 예측.

위의 의사결정트리들을 분석해보면 학생들의 희망 진로와 그 희망진로에[따른 진로추천이 확률로서 제시됨을 볼 수 있다.

그러나 입력된 데이터의 수에 비하여 현저하

게 낮은 노드수와 한쪽으로 편중된 예측값을 보여주고 있다.

이는 학생들의 데이터가 과학영재교육원의 소프트웨어 영재반 학생들이라는 한쪽으로 편중된 데이터이고 이에 따른 학생들의 희망 진로 및 수업내용도 소프트웨어 영재에 맞는 내용이기에 편중되어 나타날 수 밖에 없다.

실제로 직업군은 10개로 분류하였으나 학생들이 희망하는 진로는 네 가지 영역으로 한정되어 있었고 이 네 가지 영역중에서도 연구직 및 과학기술직에 편중되어있었다.

3.3 규칙추출 과정

표 4. 규칙 추출 커리문.

```
/* Node 1 */.
DO IF (VALUE(수업태도) EQ 5 OR VALUE(수업태도) EQ 1).
COMPUTE nod_001 = 1.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.558052.
END IF.
EXECUTE.

/* Node 4 */.
DO IF (SYSMIS(수업태도) OR VALUE(수업태도) EQ 9 OR VALUE(수업태도) NE 5 AND VALUE(수업태도) NE 3 AND VALUE(수업태도) NE 2 AND VALUE(수업태도) NE 1) AND (VALUE(창의적문제해결력) EQ 4 OR VALUE
 (창의적문제해결력) EQ 5).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.598513.
END IF.
EXECUTE.

/* Node 5 */.
DO IF (SYSMIS(수업태도) OR VALUE(수업태도) EQ 9 OR VALUE(수업태도) NE 5 AND VALUE(수업태도) NE 3 AND VALUE(수업태도) NE 2 AND VALUE(수업태도) NE 1) AND (SYSMIS(창의적문제해결력) OR VALUE(창의적문제해결력)
EQ 9 OR VALUE(창의적문제해결력) NE 4 AND VALUE(창의적문제해결력) NE 5).
COMPUTE nod_001 = 5.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.687023.
END IF.
EXECUTE.

/* Node 3 */.
DO IF (VALUE(수업태도) EQ 3 OR VALUE(수업태도) EQ 2).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.495833.
END IF.
EXECUTE.
```

IV. 결 론

학생들의 진로를 예측하는 것은 학생들의 학생 및 미래를 위하여 아주 중요한 일이다. 기존에는 학생들의 희망 진로를 전해들은 다음 학생들에게 희망진로를 위한 학습 방법 및 노력해야 할 점들을 추천해주는 수준에 진로상담이 그쳐 있었다.

더불어 학생들이 기술하고 제시한 진로의 진실성 또한 의문성을 지니고 있다. 또한 해당 학생이 해당 학생이 희망하는 진로를 향해서 제대로된 학습이나 생활을 하고 있는지는 세세한 관찰이 힘들기에 학생들의 서술이나 기술에 의존하고 있기에 적성이나 학생태도에 따른 진로 추천에도 힘겨움이 있었다.

본 시스템의 희망진로 예측에 따르면 학생들의 학습태도 및 석적 등의 다양한 관찰 데이터를 바탕으로 희망진로를 바탕으로 예측진로를 추천할 수가 있고, 희망진로에 따른 학습 태도 및 공부 방향까지 결정해 줄 수 있다고 고려된다.

다만 본 연구에서 사용된 데이터가 과학영재교육원 S/W반에 한정된 데이터기에 희망 진로 및 예측진로가 한쪽 방향으로 편중되어 있기에 향후에는 일단 학생들을 대상으로 더욱 다양한 데이터를 바탕으로 실험을 진행해 볼 필요가 있다. 더불어 현제 시스템의 구축보다는 의사결정 트리를 만들어 해석하고 검증하는데 그쳐있기에 실제 DB를 구축하고 시스템을 구축하여 많은 사람들이 이용할 수 있는 자동화 시스템을 만드는 노력이 필요 할 것이다.

참고문헌

[1] Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

[2] Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. Machine Learning, 3(4), 319-342.

[3] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

[4] 황태건. (2017). 의사결정트리를 활용한 관광지선택과정에 대한 연구

[5] 윤혜성. (2000). 결정 트리 알고리즘을 이용한 데이터 분류 및 예측

[6] 김진철. (2000). 데이터마이닝의 소개와 의사결정나무를 이용한 실제 자료의 분석

[7] 문유형. (2003). 의사결정트리 알고리즘을 이용한 학생취업상황예측연구