JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 22, No. 8: 1055~1060, Aug. 2018

진로교육을 위한 희망진로 예측프로그램 설계

김근호[1] · 김의정[1*]

Design of a Hopeful Career Forecasting Program for the Career Education

Geun-Ho, Kim[1] · Eui-Jeong, Kim[1*]

[1*]Department of Computer Education, Kongju National University, Gongju, 32588 Korea

요 약

4차 산업혁명을 맞이하여 학교 교육에 있어서 진로교육의 문제가 크게 대두되고 있다. 일선 현장에서도 인공지능 및 빅 데이터들을 효과적으로 처리하기 위한 서비스 또는 기술에 대하여 다양한 연구가 진행되고 있으나, 교육분야에 있어서는 학생들에 대한 데이터들을 단순처리과정을 거칠 뿐이다.

이에 본 논문에서는 인공지능 및 빅데이터를 활용한 학생들의 진로교육을 위한 진로 예측 프로그램을 설계 제시하고자 한다. 영재교육원 학생들의 관찰데이터를 이용하여 의사결정 트리중 가장 인공지능에 가깝고 효과적이라고 알려진 C4.5알고리즘으로 의사결정 트리를 구성하고 학생들의 희망 진로를 예측하는 것이다. 판별결과 카파계수는 0.7을 넘어 상당한 일치도를 보였고 평균절대오차도 0.1정도로 상당히 낮은 수치를 보였다.

이에 따라서 본 연구에서 보이듯이 많은 연구 및 데이터를 구축하여 학생들의 상담에 활용 진로를 제시하고 수업태도 및 방향을 제시하는데 도움이 될 것으로 사료된다.

ABSTRACT

In the wake of the 4th Industrial Revolution, the problem of career education in schools has become a big issue. While various studies are being conducted on services or technologies to effectively handle artificial intelligence and big data, in the field of education, data on students is simply processed.

Therefore, in this paper, we are going to design and present career prediction programs for students using artificial intelligence and big data. Using observational data from students at the institute, the decision tree is constructed with the C4.5 algorithm known to be most intelligent and effective in the decision tree and is used to predict students' path of hope. As a result, the coefficient of kappa exceeded 0.7 and showed a fairly low average error of 0.1 degrees.

As shown in this study, a number of studies and data will be deployed to help guide students in their consultation and to provide them with classroom attitudes and directions.

키워드 : 진로교육, 의사결정트리, 예측시스템, 인공지능, C4.5

Keywords : career education, Decision tree, Forecasting system, Artificial intelligence, C4.5

Received 15 April 2018, **Revised** 1 May 2018, **Accepted** 16 June 2018

*** Corresponding Author** Eui-Jeong, Kim(E-mail:ejkim@kongju.ac.kr, Tel:+82-41-850-8823)
Department of Computer Education, Kongju National University, Gongju, 32588 Korea

Open Access http://doi.org/10.6109/jkiice.2018.22.8.1055 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

4차 산업혁명으로 인한 미래 직업 생태계의 변화, 공유 경제로의 전환, 탈도시화 및 분산화, 인간성 상실의 위기 등의 변화 속에서 미래 인재상은 물론 교육 시스템에도 많은 변화가 필요하다.[1] 특히 2020년까지 710만 개의 일자리가 사라지고 210만개의 새로운 직업이 생성될 것으로 보고된 바, 이를 대비하여 진로를 준비라고 탐색하는 진로교육 시스템의 변화가 필요하다.[1]

특히 4차 산업혁명의 시대에 IT분야에서는 빅 데이터를 활용한 인공지능이 크게 대두되고 있다. 이와 같이 빅 데이터를 효과적으로 처리하기 위한 서비스 또는 기술에 대하여 다양한 연구가 진행되고 있다.[2] 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.[3]

이에 따라서 본 연구에서는 학생들의 수업 관찰을 통한 데이터를 활용하여 학생들의 희망진로를 바탕으로 학생들의 진로를 예측하여 진로교육에 도움을 주는 프로그램을 설계하고자 한다.

Ⅱ. 의사결정트리

2.1. 의사결정트리 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. 그림 1은 일반적인 의사결정트리의 분리과정을 설명한 그림이다. 가장 첫 단계의 결정 노드는 뿌리 노드이며 그림 1의 루트 노드1에 해당 된다. 하나의 대안이 여러 개의 리프 노드를 통해 결정될 수 있으며, 이를 규칙으로 정리할 수 있다. 뿌리 노드에 가까운 단계의 결정 노드일수록 목표한 대안을 설명하기 용이한 변수이다.[4]

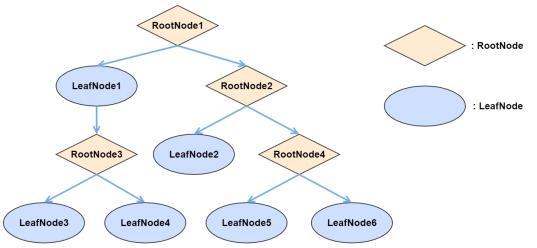

Fig. 1 Example of a decision tree

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.[5]

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

$S =$ 데이터 샘플
$\pi =$ 범주 i에 속할 비율

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

본 연구는 학생들의 수업행동 관찰 표준안을 만들고 이 표준안에 따라 학생들의 행동 및 진로를 예측하고 추천 하는 목적을 가지므로, 본 연구에서는 인공지능에 유사한 기반을 두고 있는 C4.5 알고리즘을 사용하고자 한다.

2.2. 의사결정트리 알고리즘별 비교

의사결정트리 알고리즘 종류는 ID3 알고리즘,C4.5 알고리즘,C5.0 알고리즘, CART ,CHAID 알고리즘이 있으며, 데이터 마이닝에서 가장 많이 사용되는 알고리즘은 C4.5또는 C5.0이다.ID3알고리즘의 수치형 데이터를 분류 할 수 없어 이를 보완한 알고리즘이 C4.5가 개발되었으며 다시 이를 보완한 알고리즘이 C5.0이다. 이들 알고리즘들은 크게 인공지능,기계학습 분야에서 발전된 ID3,C4.5, C5.0과 통계학 분야에서 개발된 CART, CHAID 알고리즘으로 분류 된다.[6]

표 1은 의사결정트리의 각 알고리즘의 특징을 구분해놓았다.[7]

Table. 1 Separation of decision tree algorithms

Algorithm	Mmethod	Relative
ID3	Entropy	Poly separation
C4.5	Information Gain	Poly separation & Binary separation
C5.0	Information Gain	Similar C4.5
CHAID	Chi-Square	Statistical Approach
CART	Gini Index	Statistical Approach & only Binary separation

Ⅲ. 연구방법

3.1. 연구대상

본 연구는 공주대학교 과학영재교육원 소속의 정보반 학생들의 관찰 데이터를 이용하였다. 2014년부터 2017년까지의 영재교육원의 학생들의 수업을 대상으로 관찰교사가 학생들의 수업태도 및 다양한 행동들을 기록한 데이터를 바탕으로 상담일지에 기록된 학생들의 희망진로를 데이터로 사용하였다.

Table. 2 Number of students per year

Year	Number of students
2014	10
2015	13
2016	15
2017	15

표 2에서 보이듯이 4년간의 학생 53명학생의 데이터를 활용하였으나 좀 더 다양한 데이터를 얻기 위하여 학생당 수강한 강의 별 관찰데이터를 오전 오후로 나누어서 데이터를 총 1038개의 데이터로 늘려서 연구를 실시하였다.

3.2. 데이터 전처리 및 수치화

	3/25						
	computing thinking(Professor Kang)						
ID	competence	attitude	question	leadership	communication	Ptofessor	observation
2017207	4	5	4	4	5	Unplugged computer accident 1. Present the situation 2. Creative experiment design	in the presentation class There's an active look at him trying to solve the problem, have a good attitude in class.
2017032	4	4	4	4	3	Unplugged computer accident 1. Present the situation 2. Creative experiment design	Lessons are coming soon, but I can't see you participate. He is not active when he sees it.
2017033	4	4	4	4	4	Unplugged computer accident 1. Present the situation 2. Creative experiment design	Follow me well soon and have excellent attitude in class.
2017038	4	4	4	4	4	Unplugged computer accident 1. Present the situation 2. Creative experiment design	The attitude of the class was excellent. With my friends We communicate together to teach He showed good followings.
2017048	5	5	5	5	5	Unplugged computer accident 1. Present the situation 2. Creative experiment design	Despite the individual activity, Solve problems while communicating He shows a good attitude in class.
2017054	4	5	5	5	5	Unplugged computer accident 1. Present the situation 2. Creative experiment design	The ability to observe when conditions are presented I am good at communicating with children He has a very good attitude in class because of his excellence.
2017062	4	5	4	4	4	Unplugged computer accident 1. Present the situation 2. Creative experiment design	Working with friends, Well done, and the classroom attitude is very Outstanding.
2017068	4	4	4	4	4	Unplugged computer accident 1. Present the situation 2. Creative experiment design	Student follows class well soon with calm appearance Communicate with friends next to me and do well in problem Solve.

Fig. 2 Student observation data

그림 2는 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 다음 표 3과 같이 몇 가지 항목을 정리하여 데이터를 변환 전처리 및 수치화 하였다.

Table. 3 Data item

Attributes	Variable
student	student_id
APM	1, 2(1=AM, 2=PM)
hopejob	0 ~ 9(Job category table)
teacher	1 ~ 4
competence	1 ~ 5(unsatisfactory~best)
attitude	1 ~ 5(unsatisfactory~best)
question	1 ~ 5(unsatisfactory~best)
leadership	1 ~ 5(unsatisfactory~best)
communication	1 ~ 5(unsatisfactory~best)

희망진로는 학생들의 상담일지를 바탕으로 학생들이 기록한 희망진로를 2018년 취업알선 직업 분류표에 따라서 아래 표 4와 같이 크게 10가지로 분류하였다.

Table. 4 Job category table

Variable	Job category
0	management, office, finance, insurance
1	a research and engineering profession
2	Education, law, social welfare, police, firefighting and military personnel
3	a health/medical position
4	art, design, broadcasting and sports jobs
5	beauty, travel, lodging, food, expenses, and cleaning jobs
6	Sales, sales, driving and transport jobs
7	Construction and mining jobs
8	Installation, maintenance, and production jobs
9	agricultural and fisheries jobs

위와 같은 과정을 거쳐서 최종적으로 데이터를 전처리 및 수치화 시킨 데이터는 아래 그림 3과 같다.

student_id	APM	hopejob	teacher	competence	attitude	question	leadership	communication
201701	1	1	4	4	5	4	4	5
201702	1	2	4	4	4	4	4	3
201703	1	1	4	4	4	4	4	4
201704	1	1	4	4	4	4	4	4
201705	1	1	4	5	5	5	5	5
201706	1	2	4	4	5	5	5	5
201707	1	1	4	4	5	4	4	4
201708	1	1	4	4	4	4	4	4
201709	1	0	4	4	4	4	5	5
201710	1	1	4	4	3	3	4	4
201711	1	3	4	4	3	4	3	3
201712	1	1	4	4	4	4	3	3
201713	1	2	4	4	4	4	3	3
201714	1	1	4	5	5	5	5	5
201715	1	3	4	4	5	3	3	3
201701	2	1	4	4	4	4	4	3
201702	2	2	4	4	4	4	4	5
201703	2	1	4	4	3	4	3	4
201704	2	1	4	5	5	5	5	5
201705	2	1	4	5	5	5	5	5
201706	2	2	4	5	5	5	5	5
201707	2	1	4	5	5	5	5	5
201708	2	1	4	4	3	3	3	4
201709	2	0	4	4	5	4	4	5
201710	2	1	4	4	5	4	4	4
201711	2	3	4	4	5	4	4	4
201712	2	1	4	4	3	3	3	3
201713	2	2	4	3	4	3	4	3
201714	2	1	4	4	2	4	3	5
201715	2	3	4	4	5	4	4	5
201701	1	1	3	4	5	3	3	3
201702	1	2	3	4	5	3	3	3
201703	1	1	3	5	5	4	3	3
201704	1	1	3	5	5	3	3	3
201705	1	1	3	4	5	3	3	3
201706	1	2	3	9	9	9	9	9

Fig. 3 Pre-treatment data

Ⅳ. 의사결정 트리를 이용한 학생 진로 예측 프로그램 설계

4.1. 학생진로 예측 프로그램 순서도

학생의 진로예측 및 추천 전체 처리과정은 다음 그림 4와 같다.

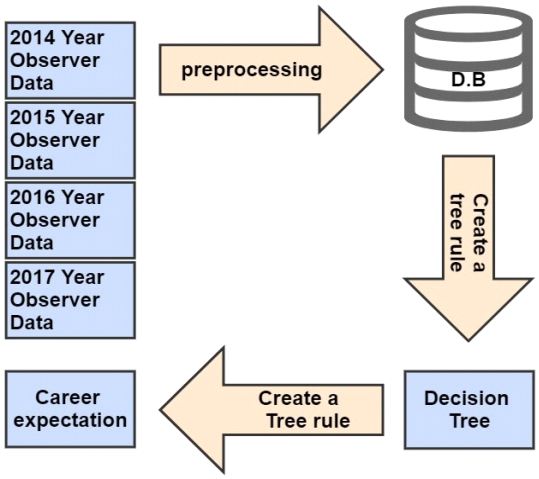

Fig. 4 Flow chart of a career prediction system

그림 4에서 보이듯 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.

4.2. 의사결정 트리 및 결과

우리가 예측하고자하는 데이터는 학생들이 제시한 희망진로를 바탕으로 관찰된 학생들의 행동 및 수업 내용 등을 가지고 의사결정트리 중 C4.5 알고리즘을 이용하여 트리를 구성하면 아래 그림 5와 같이 트리가 구성된다.

Number of Leaves : 165

Size of the tree : 209

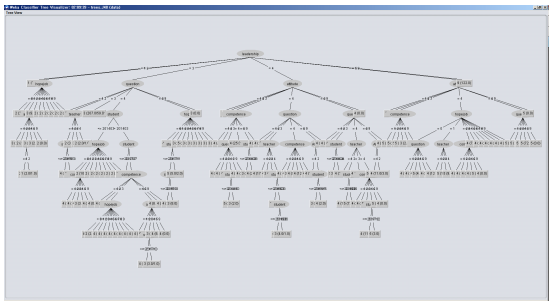

Fig. 5 C4.5 Decision tree

위 그림 5에서 보이듯이 사이즈는 209 총 노드는 165개의 크기를 가지는 의사결정 트리가 생성되었다. 그림 6은 그림 5의 트리중 일부분을 확대한 모습이다.

그림 6은 루트 노드에서 리더쉽 항목에 따라 분기했고 다음단계에서 학습태도에 따라 분기한 그림이다. 수업태도가 4등급인 학생을 다시 학습활동에 따른 구분으로 분리된 부분을 확대해 보았다.

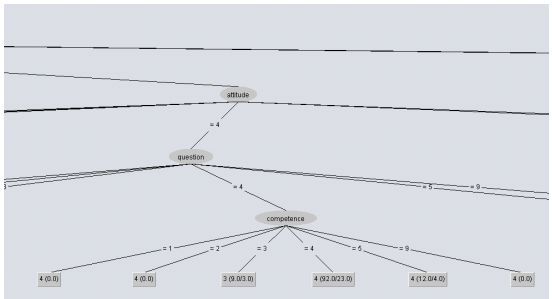

Fig. 6 Part of C4.5 Decision tree

Table. 5 Summary of C4.5 Decision tree

Correctly Classified Instances	828	79.7688 %
Incorrectly Classified Instances	210	20.2312 %
Kappa statistic		0.7293
Mean absolute error		0.1021
Total Number of Instances		1038

Table. 6 Confusion Matrixmary

a	b	c	d	e	f	<-classified as
24	2	1	0	0	0	a = 1
1	**36**	14	0	1	0	b = 2
2	10	**277**	30	4	0	c = 3
0	2	80	**242**	25	0	d = 4
0	0	10	28	**127**	0	e = 5
0	0	0	0	0	**122**	f = 9

표 5는 그림5의 의사결정트리를 요약 분석한 표이다. 위의 의사결정트리들을 분석해보면 총 1038개의 데이터 중에 79.7688%정도 되는 828개의 데이터가 유의미하게 사용되었고 20.2312%에 해당하는 210개의 데이터가 잘못 분류된 데이터로 사용되었다. 이는 데이터의 입력 할 때 결석이나 지각으로 인하여 학생관찰기록이 없는 경우 별도의 분류로 표시하였으나 데이터 전처리 및 수치화할 때 적용시키지 못하였기에 이렇게 나타났다.

관찰자 간의 일치도를 측정하는 카파계수는 0.7293으로 상당한 일치도를 보였다. 평균절대오차는 0.1021로 상당히 낮게 나타났다.

표 6의 매트릭스로 학생들의 진로에 대한 분류로 보게되면 학생들의 희망 진로와 그 희망진로에 따른 진로추천이 확률로서 제시됨을 볼 수 있다.

그러나 제시되어있는 직업군은에서 한쪽으로 편중된 예측 값을 보여주고 있다.

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

실제로 직업군은 10개로 분류하였으나 학생들이 희망하는 진로는 6가지 영역으로 한정되어있었고 이 6 가지 영역중에서도 연구직 및 과학기술직에 편중되어있었다.

Ⅴ. 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 노력해야할 점들을 추천해주는 수준에 진로상담이 그쳐 있었다.

본 시스템의 희망진로 예측에 따르면 학생들의 학습태도 및 성적 등의 다양한 관찰 데이터를 바탕으로 희망진로를 바탕으로 예측진로를 추천할 수가 있고, 희망진로에 따른 학습 태도 및 공부 방향까지 결정해 줄 수 있다. 카파계수가 0.7이상의 수치로 상당한 일치도를 보였고 평균절대오차도 0.1의 수치를 보이는 등 높은 수준의 적합성을 나다내고 있다고 할 수 있다.

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 의사결정트리를 만들어 해석하고 검증하는데 그쳐있기에 실제 DB를 구축하고 시스템을 구축하여 많은 사람들이 이용할 수 있는 자동화 시스템을 만드는 노력이 필요 할 것이다.

REFERENCES

[1] M. S. Lee, "The Effect of Reasons of College Major Selection and Stresses of College Life on Career Confidence according to Experience of Gifted Education," *Secondary Education Research*, vol. 66, no. 1, pp. 229-255, Mar. 2011.

[2] D. J. Kim, D. Sharma, "Implementation of Decision Based Fruits Protection System Using Classification and Clustering Techniques," *Asia-pacific Journal of Convergent Research Interchange*, vol. 2, no. 4, pp. 23-31, Dec. 2016.

[3] S. H. Song, E. J. Kim, "The Recognition of Cyber Education and Development Plan of Chungcheongnam-do Civil Servants," *Journal of the Korea Institute Of Information and Communication Engineering*, vol. 21, no. 11, pp. 2184-2190, Nov. 2017.

[4] L. Brett, *Machine learning with R,* 2th ed. Seoul, Seoul: Acornpub, 2017.

[5] J. Ramos, D. C. Avila, and J. Morales "Induction of Decision Trees Using an Internal Control of Induction," *Lecture Notes in Computer Science*, vol. 3512, pp. 795-803, Jun. 2005.

[6] H. W. Yim, "Security education and research in accordance with the paradigm shift in the industry Security," *Journal of Security Engineering,* vol. 12, no. 6 pp. 597-608, Dec. 2015.

[7] J. H. Seo, "A Comparative Study on the Classification of the Imbalanced Intrusion Detection Dataset Based on Deep Learning," *Journal of Korean Institute of Intelligent Systems*, vol. 28, no. 2, pp. 152-159, Apr. 2018.

김근호(Geun-Ho Kim)

2009년 공주대학교 컴퓨터과학과 석사
2016년 공주대학교 컴퓨터교육학과 박사 수료
2011년~현재 공주대학교 컴퓨터교육과 조교
※관심분야 : 컴퓨터교육, 인공지능, 가상현실, 패턴인식

김의정(Eui-jeong Kim)

1997년 충남대학교 컴퓨터공학과 공학박사
1997년~1998년 ETRI 연구원
1998년~현재 공주대학교 컴퓨터교육과 교수
※관심분야 : 컴퓨터비전, 패턴인식, 가상현실, 컴퓨터교육