데이터마이닝 기법을 이용한 전공이탈자 예측모형

- Predicting Model of Students Leaving Their Majors Using Data Mining Technique -

임 영 문 *
Leem Young Moon
유 창 현 **
Ryu Chang Hyun

Abstract

Nowadays most colleges are confronting with a serious problem because many students have left their majors at the colleges. In order to make a countermeasure for reducing major separation rate, many universities are trying to find a proper solution. As a similar endeavor, the objective of this paper is to find a predicting model of students leaving their majors. The sample for this study was chosen from a university in Kangwon-Do during seven years (2000.3.1 ~ 2006. 6.30). In this study, the ratio of training sample versus testing sample among partition data was controlled as 50% : 50% for a validation test of data division. Also, this study provides values about accuracy, sensitivity, specificity about three kinds of algorithms including CHAID, CART and C4.5. In addition, ROC chart and gains chart were used for classification of students leaving their majors. The analysis results were very informative since those enable us to know the most important factors such as semester taking a course, grade on cultural subjects, scholarship, grade on majors, and total completion of courses which can affect students leaving their majors.

Keywords : Data Mining, Predicting Model, Major Separation Rate, ROC-Chart

* 강릉대학교 산업공학과 교수
** 강릉대학교 산업공학과 석사과정
2006년 9월접수; 2006년 10월 수정본 접수; 2006년 10월 게재확정

1. 서 론

사회적, 경제적인 상황에 따라 교육기관의 구성원인 학생들의 인식과 목표도 그 현실성에 맞게 변화하고 있다. 현재 대학생들에게 개인의 재능과 특성을 살리는 교육보다는 사회의 요구에 따라 전공을 맞춰가는 현상들이 발생하고 있으며, 이러한 현상들은 대학 내에서 전공이탈로 이어지고 있다. 이러한 문제점들은 "학문의 다양화"라는 이점을 살리지 못하고, 특정 학문의 과잉현상과 학문적 균형을 무너트리는 병폐를 낳고 있으며, 전문적인 지식을 습득하여 전문적인 인력으로 양성되어야 하는 대학생들의 전문적인 가치를 떨어트리는 요인으로 작용되고 있다.

데이터마이닝을 적용하여 알고리즘을 비교한 기존 연구들을 살펴보면 신용카드 고객의 이탈고객 분석[5], 통신회사의 고객정보 데이터를 통한 해지 고객 예측 모형[2] 등이 있으며, 전공이탈에 관련된 기존연구들에는 회귀모형을 통한 전공이탈자 예측모형[8], 의사결정나무를 이용한 전공이탈자 예측모형[3] 등이 있다. 회귀모형은 결과에 대한 해석이 어려워 현실에서 적용하기 어렵다는 단점이 있으며, 의사결정나무를 이용한 전공이탈자 예측모형은 그 자료가 미비할 뿐만 아니라 의사결정나무의 타당성을 검증할 수 있는 여러 가지 검증방법들이 적용되지 않았다.

본 연구에서는 전공이탈 문제점에 착안하여 데이터마이닝 기법 중 하나인 의사결정나무를 이용하여 전공이탈 학생에 대한 통계적 모형을 구축하고자 한다. 과거의 데이터의 분석을 통하여 보다 객관적이고 정량화된 데이터를 얻을 수 있으며, 또한 이를 통하여 이탈자들의 특성을 파악하여 이탈방지를 위한 특성을 제시 하였다.

2. 연구내용 및 방법

본 연구에서 사용된 데이터는 강원도 소재 4년제 대학 1곳의 2000년부터~2006년까지 재학생 및 졸업생에 관련된 자료 13,346명의 자료 중 더 이상 전과가 가능하지 않은 5학기 이상 학점 이수자 5,115명을 대상으로 하고 있다.

재학, 휴학, 졸업자중 전과기록이 있는 자들을 (전공)이탈로 정의하였다.

의사결정나무의 3가지 알고리즘별(CHAID, CART, C4.5)로 각각의 특성치들을 비교하기 위하여, 분석용 데이터와 평가용 데이터를 각각 50:50 비율로 나누어 분석한후 타당성 평가를 위하여 알고리즘별 정분류율(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 오분류율(Error Rate), ROC Chart, 이익도표 등을 이용하여 비교 분석하였다.

그리고 이를 토대로 이탈자를 예측할 수 있는 최적 알고리즘을 제시 하였다. 분석도구로는 SAS Enterprise Miner 4.3을 사용하였다.

3. 분 석 결 과

3.1 변 수 선 택

대용량 데이터에서 하나의 목표변수에 후보가 될 만한 입력변수는 대단히 많이 존재하는 것이 일반적이다. 이중 목표변수와 관련성이 높은 변수군을 선발한 후 이를 이용하여 모형구축을 시도하는 것이 모든 가능한 변수를 바로 모형구축에 이용하는 것보다는 훨씬 효율적이다[4].

본 연구에서는 χ^2값 3.84를 기준으로 사용하였으며, 카이제곱 통계량 3.84의 의미는 95% 신뢰구간을 의미한다. Chi-Square 값이 작거나 Missing Value값들은 변수에서 제외되고 나머지 목표변수를 제외한 15개의 변수(생년, 출신고교주소, 2학년1학기 장학금수여내역, 1번째학기 평점, 1번째학기 이수학점, 2번째학기 평점, 3번째학기 평점, 4번째학기 이수학점, 1번째학기 계열기초이수학점, 2번째학기 계열기초평점, 3번째학기 전공평점, 3번째학기 교양이수학점, 4번째학기 전공이수이수학점, 4번째학기 교양이수학점, 4번째학기 지정이수평점)들이 분석에 사용되었다.

3.2 모델별 결과 비교

데이터마이닝 모델들을 비교 분석하기 위하여 모델별로 정분류율, 민감도, 특이도, 오분류율 값을 구하여 비교 분석하였다. 분류표(Classification Tables)란 목표변수의 실제범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 것으로 다음의 <그림 1>과 같다.

		예측된 변수 0	예측된 변수 1	
원래 목표변수	0	실제0 예측0	실제0 예측1	실제0
	1	실제1 예측0	실제1 예측1	실제1
		예측0	예측1	

<그림 1> 분류표의 구성

위의 그림을 참조로 하여 정분류율 또는 정확도, 민감도, 특이도, 오분류율의 개념을 다음과 같이 정의할 수 있다[6].

$$\text{정분류율}(Accuracy) = \frac{(\text{실제}0, \text{예측}0)\text{의빈도} + (\text{실제}1, \text{예측}1)\text{의빈도}}{\text{전체빈도}}$$

$$\text{오분류율}(Error\ Rate) = \frac{(\text{실제}0, \text{예측}1)\text{의빈도} + (\text{실제}1, \text{예측}0)\text{의빈도}}{\text{전체빈도}}$$

$$\text{민감도}(Sensitivity) = \frac{(\text{실제}1, \text{예측}1)\text{인 관찰치의 빈도}}{\text{실제}1\text{인 관찰치의 빈도}}$$

$$\text{특이도}(Specificity) = \frac{(\text{실제}0, \text{예측}0)\text{인 관찰치의 빈도}}{\text{실제}0\text{인 관찰치의 빈도}}$$

위의 정의를 해석해 보면 정분류율 또는 정확도는 트리가 얼마나 잘 분리되었는가에 대한 능력, 민감도는 참(True)인 것을 참이라고 선언하는 능력, 특이도는 거짓(False)인 것을 거짓이라 선언하는 능력 또는 거짓인 것을 배제할 수 있는 능력으로 정리될 수 있다. 이들 특성치에 대한 우선순위를 열거하면 정분류율 또는 정확도, 민감도, 특이도의 순서로 표현될 수 있다[7].

분류값 들을 분석용 데이터와 평가용 데이터로 비교한 값은 <표 1>과 같다.

<표 1> 알고리즘별 분류값 비교

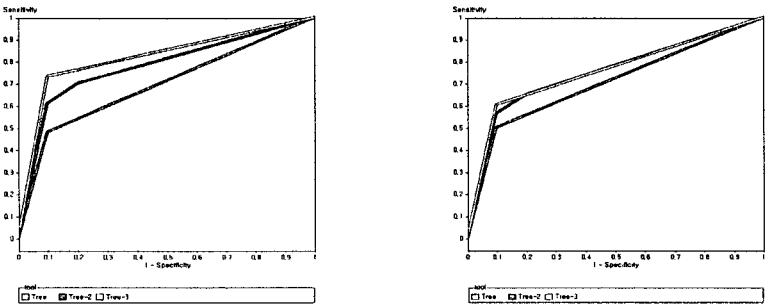
		정분류율(%)	오분류율(%)	민감도(%)	특이도(%)
CHAID	Train	96.56%	3.44%	31.73%	99.31%
	Validation	96.28%	3.72%	30.77%	99.43%
C4.5	Train	96.68%	3.32%	25.96%	99.67%
	Validation	95.85%	4.15%	18.80%	99.55%
CART	Train	97.34%	2.66%	36.54%	99.92%
	Validation	96.99%	3.01%	36.75%	99.88%

<표 1>에서 볼 수 있듯이 정분류율은 분석용 데이터에서 CHAID 96.56%, C4.5 96.68%, CART 97.34%로 CART가 가장 뛰어났으며 평가용 데이터에서도 CHAID 96.28%, C4.5 95.85%, CART 96.99%로 CART가 높은 정확도를 보였다.

민감도에서는 분석용 데이터에서 CART가 36.54%로 가장 높은 민감도를 보였고 C4.5가 25.96%로 가장 낮았다. 평가용 데이터에서는 CART 36.75%, C4.5 18.80%로 CART가 제일 높은 결과값을 보였으며, C4.5는 10%대의 가장 낮은 민감도를 보였다.

특이도에서는 3가지모형 모두 상당히 높은 값들을 보였다. 분석용 데이터에서는 CART 99.92%, CHAID 99.31%로 CART가 제일 높았고, CHAID가 가장 낮았다. 평가용 데이터에서는 CART 99.88%, CHAID 99.43%로 분석용 데이터와 마찬가지로 CART가 제일 높고, CHAID가 가장 낮았다.

분류표를 통한 정분류율, 민감도, 특이도 들의 값들을 비교하여 보았을 때 CART가 분석용 데이터 및 평가용 데이터에서 뛰어난 분류율을 보이고 있으므로, 학생들의 전공이탈 자료에는 CART모형이 최적의 모형이라고 말할 수 있다.



Train Validation

<그림 2> ROC Chart

ROC Chart는 이진형의 목표변수를 가지는 모형들의 성능을 비교·평가하는데 매우 유용한 도표로 사용되며, 각각의 관측치에서 사후확률을 구한 후 분류 기준값에 따른 오분류표를 만들어 (1-특이도)와 민감도를 이용하여 ROC곡선을 표현한 것이다. <그림 2>는 ROC Chart로 수평축은 (1-특이도)이고 수직축은 모형의 민감도를 나타내고 있으며, 이러한 결과에 따라 그래프가 도표의 왼쪽 상단으로 더 가까운 모형을 성능 면에서 우수한 모형으로 판단하게 된다[7].

CART(Tree-3)의 민감도가 분석용 데이터나 평가용 데이터에서 가장 높기 때문에 다른 모형들 보다 성능이 좋은 모형이라고 할 수 있다.

Lift Chart는 사후확률을 이용하여 예측의 정확성을 알 수 있다. Lift Chart는 각각의 관측치에서 사후확률을 구한 후 사후확률의 크기순서에 따라 전체 자료를 균일하게 N등분한 후 각 집단에서의 %Captured Response, %Response 그리고 Lift를 계산한다. 각각의 의미는 다음과 같다[1].

$$\%\text{Captured Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{전체에서 목표변수의 특정범주 빈도}} \times 100$$

$$\%\text{Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{해당 등급에서 전체 빈도}} \times 100$$

$$\text{Base Line \%Response} = \frac{\text{해당 등급에서 목표변수의 특정범주 빈도}}{\text{전체 빈도}} \times 100$$

$$\text{Lift} = \frac{\text{해당등급의 \%Response}}{\text{Base Line Lift}} \times 100$$

누적 %Response 도표의 수평축은 사후확률 값으로 전체 데이터를 정렬하여 10%씩 나눈 각 집단을 나타내고, 맨 좌측의 10은 사후확률이 가장 높은 10% 집단을 나타낸다. %Captured Response는 특정범주 내에서 특정 등급이 차지하고 있는 점유율로 해석할 수 있으며, Lift Value는 전체 집단에 비해서 해당 등급에서 예측력이 향상된 정도를 나타낸다. 이탈 가능성이 높은 상위집단 10%, 20%, 30%의 누적이익도표는 <표 2>와 같다.

<표 2> 누적이익도표

모델명	백분위수	Captured Response Rate(%)	Response Rate(%)	Lift Value
CHAID	10	56.8156	23.0994	5.6816
	20	71.0556	14.4445	3.5528
	30	76.3641	10.3491	2.5455
C4.5	10	48.0259	19.5258	4.8026
	20	54.8694	11.1541	2.7435
	30	61.5709	8.3442	2.0524
CART	10	72.9799	29.6713	7.2980
	20	81.8969	16.6483	4.0948
	30	88.9901	12.0588	2.9660

<표 2>를 살펴보면 응답률(Response Rate)이 상위 10%집단에서 CART가 29.67%를 나타내고 있다. 이것은 상위 10% 집단에서 CART가 29.67%의 이탈률이 높은 학생들을 포함한다는 것을 의미한다. 이것은 CART가 다른 모형 CHAID(23.10%), C4.5(19.53%)보다 높은 예측의 정확성을 가진다는 것을 나타낸다. 상위 10%뿐만이 아니라 상위 20%, 30% 부분에서도 CART가 가장 높은 정확성을 나타낸다.

의사결정나무의 세 가지 알고리즘의 최적예측모형을 선정하기 위하여 정분류율, 오분류율, 민감도, 특이도, ROC Chart, 이익도표를 이용하여 비교한 결과 CART가 예측력에 있어서 성능이 가장 좋은 모형이라고 판단되기 때문에 CART를 최종예측모형으로 선택하였다.

3.3 최적 예측 모형

CART를 이용하여 구성된 이탈예측 모형화 모델은 <그림 3, 4>와 같다. 전체모형의 분포가 넓기 때문에 3~6 Depth의 그림은 <그림 4>와 같이 ①, ②로 나누어 표현하였다. CART 이탈예측 모형화 모델의 노드들을 분석해 보면 총 20가지의 유형을 파악할 수 있으며, 이 노드들 중에 가장 많은 발생빈도를 나타내는 상위 3유형을 살펴보면 <표 3>에서와 같이 유형2, 유형14, 유형1순으로 나타났다.

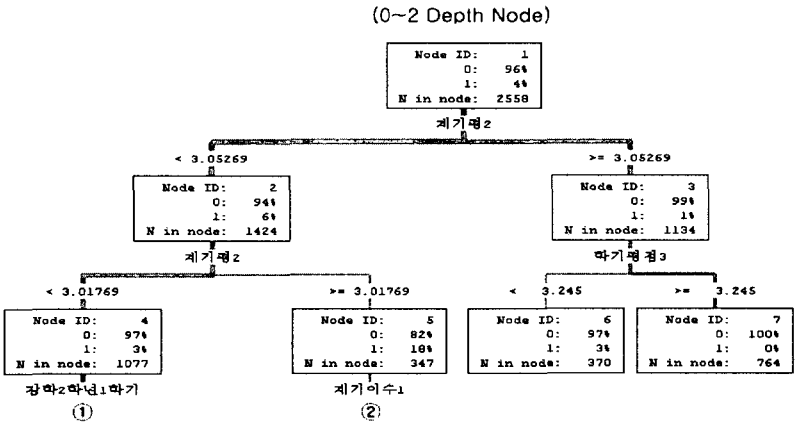
먼저 유형2는 2번째 학기 계열기초평점이 3.05보다 높고 3번째 학기 평점이 3.25보다 높은 경우 764건으로서 가장 높은 이탈유형을 보였다. 유형14는 2번째 학기 계열기초평점이 3.05보다 높고 2번째 학기 계열기초평점이 3.02보다 작고 2학년1학기 장학금수여내역이 미수여 또는 A급, 3번째 학기 전공평점이 0.43이상, 1번째 학기 이수학점이 16학점초과, 4번째 학기 전공이수학점이 8학점 이상인 경우 571건으로 나타났다. 유형1은 2번째 학기 계열기초평점이 3.05보다 높고, 3번째 학기 평점이 3.25보다 적은경우로 370건으로 나타났다.

유형1, 유형2의 결과는 2번째 학기 계열기초평점과 3번째 학기 평점이 전과에 가장

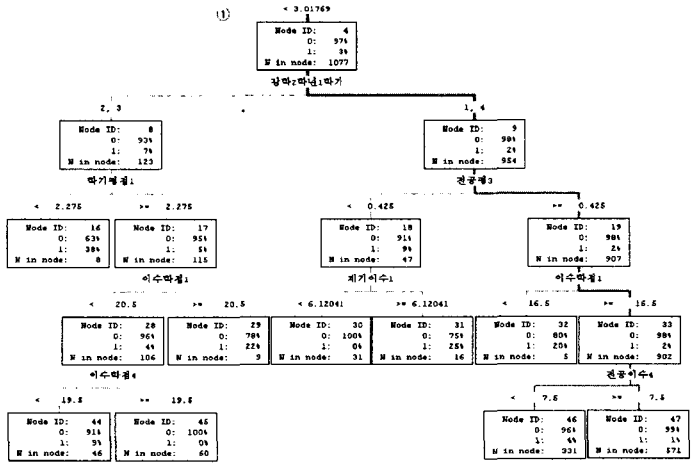
큰 영향을 주는 요인이라 할 수 있다.

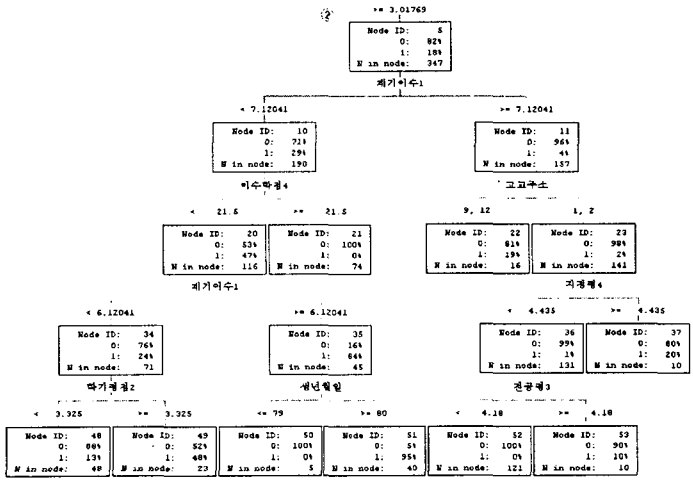
<표 3> 상위 3유형 노드

노드 유형	노드 경로	N	depth
유형1	node1 → node3 → node6 계기평2(>3.05) → 학기평점3(<3.25)	370	2
유형2	node1 → node3 → node7 계기평2(>3.05) → 학기평점3(>=3.25)	764	2
유형14	node1→node2→node4→node9→node19→node33→node47 계기평2(<3.05)→계기평2(<3.02)→장학2학년1학기(1,4)→전공평3(>=0.43)→이수학점1(>=16.5)→전공이수4(>=7.5)	571	6



<그림 3> 이탈예측 모형화 모델(0~2 Depth)





<그림 4> 이탈예측 모형화 모델(3~6 Depth)

4. 결론 및 추후 연구사항

본 연구의 주된 목적은 이탈학생들의 특성을 파악하고, 그 원인을 규명하여 대학은 물론 각 학과의 전공이탈 학생들을 관리할 수 있는 정책에 도움이 되는 방안을 제시하는 것이다. 의사결정나무의 3가지 알고리즘 중 최적의 예측력을 보이는 알고리즘 선정을 위하여 분류표, ROC Chart, 이익도표를 이용하여 특성치를 비교하여 본 결과 정확도, 민감도, 특이도에서 CART가 가장 우수한 분류성능을 보였다.

CART를 이용한 이탈예측 모형화 모델의 노드들을 분석해 보면 총 20가지의 유형을 파악할 수 있으며, 이 노드들 중에 가장 많은 발생을 보이고 있는 상위 3유형은 유형2, 유형14, 유형1로 나타났다.

이 결과는 2번째 학기 계열기초평점과 3번째 학기 평점이 이탈에 영향을 주는 요인이라 평가 할 수 있다. 2학년1학기 장학금수여내역, 3번째 학기 전공평점, 1번째 학기 이수학점, 4번째 학기 전공이수학점 등이 부수적인 영향을 주는 요인이라고 평가할 수 있었다.

본 연구에서 사용된 데이터는 학교 내 인트라넷에 구성된 신상자료와 학적자료에 기반을 두고 있다. 이러한 데이터는 개인의 프라이버시를 중시하는 세대에 맞물려 자료의 분석 · 활용에 있어서 어려움이 많았으며, 자료의 충실성이 의심되는 데이터들이 많았기 때문에 데이터의 활용이 분석에 있어서 완벽하게 적용되었다고는 생각하지 않는다. 또한 이탈에 영향을 주는 요인들 중에는 분석에서 사용되지 않은 사회적 분위기, 가정형편, 학과의 취업률 등의 제 3요인들이 제외되어 있기 때문에 완벽한 결과라고 판단하기는 힘들다. 추후 연구로 충실성 있는 데이터에 의한 연구가 필요할 것이다.

5. 참 고 문 헌

[1] 강현철, 한상태, 최종우, 김은석, 김미경, "SAS Enterprise Miner 4.0을 이용한 데이터마이닝 방법론 및 활용", 자유아카데미, (2001)
[2] 문정호, "사례연구를 통한 데이터마이닝 수행과정 연구", 서울대학교 석사학위논문, (2002)
[3] 박철용, "Analysis of Students Leaving Their Majors Using Decision Tree", 한국데이터정보과학회지 제13권 제2호, (2002):157-165
[4] 배화수, 조대현, 석경하, 김병수, 최국렬, 이종언, 노세원, 이승철, 손용희, "SAS Enterprise Miner를 이용한 데이터마이닝", 교우사, (2005)
[5] 이건창, 정남호, 신경식, "신용카드 시장에서 데이터마이닝을 이용한 이탈고객 분석", 한국지능정보시스템학회 2001년도 춘계정기학술대회, (2001):421-444
[6] 이석호, "데이터베이스 시스템", 정익사, (1995)
[7] 조윤정, "데이터마이닝을 이용한 종합건강진단센터의 데이터베이스 마케팅에 관한 연구", 서울대학교 보건대학원 보건학석사학위논문(2001)::53-56
[8] 최재성, "Logistic regression model for major separation rate", 한국데이터정보과학회지 제13권 제2호, (2002):129-138

저 자 소 개

임 영 문 : 연세대학교에서 학사, 석사학위를 취득하였고, 미국 텍사스주립대학교 산업시스템공학과에서 공학박사를 취득하였으며, 미국 ARRI (Automation and Robotics Research Institute) 연구소에서 선임연구원 및 연구교수를 거쳐 현재는 강릉대학교 산업공학과 부교수로 재직 중이다.

유 창 현 : 현재 강릉대학교 산업공학과 대학원 석사과정에 재학 중이며, 관심분야는 데이타마이닝, 알고리즘 분석 및 활용 등이다.

저 자 주 소

임 영 문 : 서울시 서초구 서초4동 아크로비스타 C동 910호

유 창 현 : 경기도 연천군 전곡읍 전곡4리 9반 310-30