데이터마이닝 기법을 이용한 전공이탈자 분류를 위한 성능평가

Evaluation on Performance for Classification of Students Leaving Their Majors Using Data Mining Technique

임 영 문[1)]
Leem Young Moon
유 창 현[2)]
Ryu Chang Hyun

Abstract

Recently most universities are suffering from students leaving their majors. In order to make a countermeasure for reducing major separation rate, many universities are trying to find a proper solution. As a similar endeavor, this paper uses decision tree algorithm which is one of the data mining techniques which conduct grouping or prediction into several sub-groups from interested groups. This technique can analyze a feature of type on students leaving their majors. The dataset consists of 5,115 features through data selection from total data of 13,346 collected from a university in Kangwon-Do during seven years (2000.3.1 ~ 2006. 6.30). The main objective of this study is to evaluate performance of algorithms including CHAID, CART and C4.5 for classification of students leaving their majors with ROC Chart, Lift Chart and Gains Chart. Also, this study provides values about accuracy, sensitivity, specificity using classification table. According to the analysis result, CART showed the best performance for classification of students leaving their majors.

Keywords : Decision Tree, ROC Chart, Lift Chart, Major Separation Rate, Gains Chart

1) 강릉대학교 산업공학과 교수
2) 강릉대학교 산업공학과 석사과정

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 있다. 이러한 현상들은 학문의 전문성을 저하 시키고, 대학들의 경쟁력을 약화시키는 요인이 작용하고 있다. 이로 인해 학생들이 전공을 쉽게 바꾸고 각 학과에서 다른 학과로 이동하는 전공이탈자들이 발생하고 있다.

데이터마이닝을 적용하여 알고리즘을 비교한 기존 연구들을 살펴보면 신용카드 고객의 이탈고객 분석[4], 통신회사의 고객정보 데이터를 통한 해지 고객 예측 모형[1] 등이 있으며, 전공이탈에 관련된 기존연구들에는 회귀모형을 통한 전공이탈자 예측모형[8], 의사결정나무를 이용한 전공이탈자 예측모형[2] 등이 있다.

본 연구에서는 학생들의 전공 이탈현상에 따른 특성을 파악하기 위하여 데이터마이닝 기법중 하나인 의사결정나무 기법을 적용하였다. 각 전공이탈자를 분류할 수 있는 특성치들을 비교하여 최적의 모형을 구축할 수 있는 기초자료를 제시하고자 한다.

2. 연구내용 및 방법

본 연구에서 사용된 데이터는 강원도 소재 4년제 대학의 2000년부터 ~ 2006년까지 재학생 및 졸업생에 관련된 자료 13,346명의 자료중 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 나누어서 분류표를 이용한 정분류율, 민감도, 특이도 값을 비교 제시하였다. 그리고 이를 토대로 이탈자를 예측할 수 있는 최적 알고리즘을 제시하고자한다. 분석도구로는 SAS Enterprise Miner 4.3을 사용하였다.

3. 분석결과

3.1 변수선택

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[3].

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(Accuracy), 오분류율(Error Rate), 민감도(Sensitivity), 특이도(Specificity)를 분류표(Classification Tables)를 이용하여 계산하였다. 분류표란 목표변수의 실제 범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 것으로 다음과 같이 정의할 수 있다[5].

$$정분류율(Accuracy) = \frac{(실제0, 예측0)의 빈도 + (실제1, 예측1)의 빈도}{전체빈도}$$

$$오분류율(Error\ Rate) = \frac{(실제0, 예측1)의 빈도 + (실제1, 예측0)의 빈도}{전체빈도}$$

$$민감도(Sensitivity) = \frac{(실제1, 예측1)인 관찰치의 빈도}{실제1인 관찰치의 빈도}$$

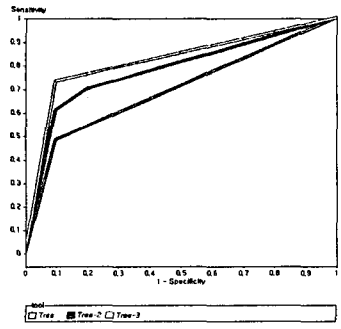
$$특이도(Specificity) = \frac{(실제0, 예측0)인 관찰치의 빈도}{실제0인 관찰치의 빈도}$$

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[6]. 분류표를 통하여 얻은 분류값 들을 분석용 데이터와 평가용 데이터로 비교한 값은 <표 1>과 같다.

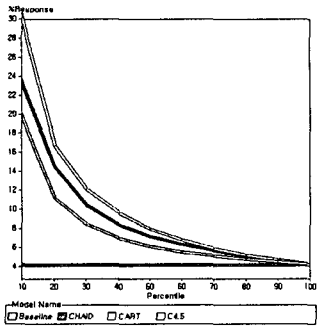
<표 1> 알고리즘별 분류값 비교

		정분류율 (%)	민감도 (%)	특이도 (%)
CHAID	분석용	96.56%	31.73%	99.31%
	평가용	96.28%	30.77%	99.43%
C4.5	분석용	96.68%	25.96%	99.67%
	평가용	95.85%	18.80%	99.55%
CART	분석용	97.34%	36.54%	99.92%
	평가용	96.99%	36.75%	99.88%

<표 1>에서 볼 수 있듯이 정분류율은 분석용 데이터에서 CART 97.34%가 가장 뛰어났으며 평가용 데이터에서도 CART가 96.99%로 높은 정확도를 보였다. 민감도에서는 분석용 데이터에서 CART가 36.54%로 가장 높았고, 평가용 데이터에서도 CART가 36.75%로 가장 높은 값을 보였다. 특이도에서는 CART를 포함한 3가지모형 모두 상당히 높은 값들을 보였다.



<그림 1> ROC Chart



<그림 2> Lift Chart

<표 2> 누적이익도표

모델명	백분위수	Response Rate(%)
CHAID	10	23.0994
	20	14.4445
	30	10.3491
C4.5	10	19.5258
	20	11.1541
	30	8.3442
CART	10	29.6713
	20	16.6483
	30	12.0588

ROC Chart는 각각의 관측치에서 사후확률을 구한 후 분류 기준값에 따른 오분류표를 만들어 (1-특이도)와 민감도를 이용하여 ROC곡선을 표현한 것이다. <그림 1>은 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 되며[6], <그림 1>에서는 CART(Tree-3)의 민감도가 가장 높게 나타났다.

Lift Chart는 사후확률을 이용하여 예측의 정확성을 알 수 있다. 누적 %Response 도표의 수평축은 사후확률 값으로 전체 데이터를 정렬하여 10%씩 나눈 각 집단을 나타내고, 맨 좌측의 10은 사후확률이 가장 높은 10% 집단을 나타낸다[7]. 위의 <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.

4. 결론 및 추후연구사항

본 연구의 주된 목적은 이탈학생들의 특성을 파악하여, 이탈 모형구축의 기반이 되는 자료를 제공하고자 함이다. 의사결정나무의 3가지 알고리즘 중 최적의 예측력을 보이는 알고리즘 선정을 위하여 분류표, ROC Chart, Lift Chart를 이용하여 특성치를 비교하여 본 결과 정확도, 민감도, 특이도에서 CART가 가장 우수한 분류성능을 보였다.

분석에 사용된 데이터는 강원도 내 4년제 대학의 학생 및 졸업생들의 데이터를 분석하였다. 하지만 자료의 충실성이 의심되는 데이터들이 많았기 때문에 데이터의 활용이 분석에 있어서 완벽하게 적용되었다고는 생각하지 않는다. 추후 연구로 충실성 있는 데이터에 의한 연구가 필요할 것이며, 전과이탈자를 예측할 수 있는 예측모형에 대한 연구가 필요할 것으로 사료된다.

5. 참 고 문 헌

[1] 문정호, 사례연구를 통한 데이터마이닝 수행과정 연구, 서울대학교 석사학위논문, 2002.

[2] 박철용, Analysis of Students Leaving Their Majors Using Decision Tree, 한국데이터정보과학회지 제13권 제2호, pp. 157~165, 2002.

[3] 배화수, 조대현, 석경하, 김병수, 최국렬, 이종언, 노세원, 이승철, 손용희, SAS Enterprise Miner를 이용한 데이터마이닝, 교우사, 2005.

[4] 이건창, 정남호, 신경식, 신용카드 시장에서 데이터마이닝을 이용한 이탈고객 분석, 한국지능정보시스템학회 2001년도 춘계정기학술대회, 2001, pp. 421~444.

[5] 이석호, 데이터베이스 시스템, 정익사, 1995.

[6] 조윤정, "데이터마이닝을 이용한 종합건강진단센터의 데이터베이스 마케팅에 관한 연구", 서울대학교 보건대학원 보건학석사학위논문, pp. 53~56, 2001.

[7] 최종우, 한상태, 강현철, 김은석, 김미경, 이성건, "SAS Enterprise Miner 4.0을 이용한 데이터마이닝 기능과 사용법", 자유아카데미, 2001.

[8] 최재성, Logistic regression model for major separation rate, 한국데이터정보과학회지 제13권 제2호, pp. 129~138, 2002.