의사결정트리를 이용한 교육성과 요인에 관한 연구

A Study on Factors of Education's Outcome using Decision Trees

김완섭†
숭실대학교 베어드학부대학 조교수

Wan Seop Kim†
Assistant Professor, Baird University College, Soongsil Univ.

요 약

대학에서 운영되는 강좌를 효과적으로 관리하고 교육성과를 향상시키기 위해서는 각 클래스의 현재의 교육성과를 진단하고 교육성과에 영향을 미치는 요인들을 파악하는 과정이 요구된다. 요인을 발견하는 연구에는 연관성 분석, 회귀분석 등의 통계기법들이 많이 사용되고 있으며 최근에는 데이터마이닝의 결정트리 분석도 사용되고 있다. 결정트리 분석은 결과 모델을 이해하기 쉽고 의사결정에 적용하기 쉽다는 장점이 있지만, 다중공선성 등의 입력 데이터의 특성에 견고하지 못한 문제점이 있다. 본 연구에서는 기존의 결정트리 분석의 문제점들을 정리하고, 이 문제점들을 보완하기 위한 하나의 실험적 해결책으로 다중 결정트리를 이용한 요인의 발견 방법을 제안한다. 실험을 통해 다중 결정트리를 수행이 다중 결정트리를 적용할 때보다 신뢰할 수 있는 요인을 발견하고 각 변수의 중요성을 발견할 수 있음을 보였다.

주제어: 교육성과, 요인, C5.0, CHAID, CART, 결정트리

Abstract

In order to manage the lectures efficiently in the university and improve the educational outcome, the process is needed that make diagnosis of the present educational outcome of each classes on a lecture and find factors of educational outcome. In most studies for finding the factors of the efficient lecture, statistical methods such as association analysis, regression analysis are used usually, and recently decision tree analysis is employed, too. The decision tree analysis have the merits that is easy to understand a result model, and to be easy to apply for the decision making, but have the weaknesses that is not strong for characteristic of input data such as multicollinearity. This paper indicates the weaknesses of decision tree analysis, and suggests the experimental solution using multiple decision tree algorithm to supplement these problems. The experimental result shows that the suggested method is more effective in finding the reliable factors of the educational outcome.

Keywords: Education outcome, Factor, C5.0, CHAID, CART, Decision tree

Ⅰ. 서론

대학에서 동일 강좌에 해당하는 여러 클래스가 개설될 때, 여러 가지 요인에 의하여 클래스별로 교육성과에 차이가 발생한다. 개설 강좌를 효과적으로 운영하기 위해서는 각 클래스의 교육성과를 비교 평가하고, 강좌의 교육성과에 영향을 미치는 요인을 발견하는 과정이 요구된다. 이에 관련하여 대학의 강의 만족도 평가 또는 재학생 및 졸업생을 대상으로 한 설문지 평가를 토대로 하여 효과적인 강의의 요인을 발견하고자 한 많은 연구들이 이루어진바 있다(한송엽·서경덕, 2002; 류춘호·이정호, 2003; 한경석·노미현, 2004; 소연희, 2006, 2009).

요인을 발견하는 연구는 교육 분야에서 뿐 아니라 사회과학, 의학/보건학, 경영학 등 다양한 분야에서 접근하고 있는 연구 방식이다. 요인을 발견하기 위한 분석의 방법으로는 통계학의 연관성 분석, 다중회귀 분석 및 구조방정식모형 분석이 전통적으로 많이 사용된다. 연관성 분석은 두 변수 간의 연관성을 측정하는 방법으로, 변수의 형태에 따라 수치형의 경우에는 피어슨 상관계수를 사용하고, 범주형의 경우에는 카이제곱 검정이 주로 사용된다(박동준·이수영, 2009; 한경석·노미현, 2004). 대부분의 연구에서는 연관성 분석만을 사용하지 않고 다중회귀분석도 수행하여 요인을 발견한다.

논문접수일: 2010년 3월 10일
최종수정일: 2010년 7월 12일
논문완료일: 2010년 7월 20일
† 교신저자: 김완섭

연관성 분석이 두 변수 간의 관계만을 평가하는 반면, 회귀분석은 여러 입력변수들이 종속변수에 미치는 영향을 회귀식으로 표현해주기 때문에 각 입력변수의 종속변수에 대한 영향력을 상대적으로 비교할 수 있는 장점이 있다. 그 외에도 구조방정식모형(SEM: Structural Equation Model) 분석을 사용한 요인 연구들도 있다(배정이, 2008; 이중섭·이용교, 2009). 구조방정식모형 분석은 단순히 입력변수의 영향력의 정도의 크기만을 측정하고 비교하는 것이 아니라 여러 입력변수들의 인과관계를 고려하여 모형을 발견하는 점에서 차이가 있다.

최근에는 요인을 발견하기 위해 기존의 통계적 방법들과 더불어 데이터마이닝의 분류 분석 방법 중 하나인 의사결정트리(Decision tree)가 많이 사용되고 있다(이수원·김완섭, 2006; 이주리, 2009). 결정트리는 내부적으로 정보이득(Information Gain) 개념이나 카이제곱 검정(Chi-square test) 등의 평가기준을 사용하여 중요한 변수를 발견한다. 중요한 변수를 선택하고, 선택된 변수를 기준으로 데이터집합을 분리하는 과정을 반복적으로 수행하여 나무 구조의 계층적인 모델로 결과를 제공한다(Han J. & Kamber M., 2006; Roiger, 2003; 김신곤·박성용, 1999). 의사결정트리 분석은 여러 입력변수들과 종속변수와의 관계를 복합적으로 분석하며, 결과를 나무 모형으로 시각화하여 결과를 제공하여 분석자가 결과를 쉽게 이해하고 의사결정에 적용할 수 있는 장점을 갖는다. 그러나 의사결정트리 분석은 모수적 분석(Parametric Analysis)에 속하여 입력데이터의 특성에 대하여 견고하지 못한 문제점들을 갖고 있으나, 대부분의 의사결정트리를 사용하여 요인을 발견하는 연구에서 이 문제는 간과되고 있다(Selwyn, 2008).

본 연구에서는 의사결정트리 분석이 갖는 몇 가지 한계점들을 제시하고 이를 개선하기 위하여 다중의 의사결정트리를 사용한 분석방법을 제시하고 실험에 적용하고자 한다. 실험에서는 숭실대학교의 1학년 교양필수 과목인 '컴퓨터활용1'의 2009년 시험 결과 데이터를 사용하였다. 일반적으로 동일 강좌에 대해 여러 클래스가 개설되었을 때 각 클래스의 교육성과를 객관적으로 비교할 수 있는 척도의 확보가 어렵기 때문에 교육성과의 요인을 발견하는 연구에 한계가 있다. 일반적으로 학기 말에 이루어진 강의평가 결과나 재학생 및 졸업생을 대상으로 한 설문조사 자료를 실험에 사용되는데 이 정보는 엄밀하게는 학생의 교육성과의 정도와는 차이가 있으며 주관적 평가이므로 교육성과의 요인에 관한 연구 실험에는 적합하지 못한 면이 있다. 그러나 본 연구의 실험 에 적용된 '컴퓨터활용1' 강좌의 경우 MOS(Microsoft Office Specialist) 자격증 시험으로 평가를 하기 때문에 학생들의 '합격/불합격' 여부는 학생들의 객관적 교육성과의 지표로 볼 수 있으며, 또한 이 정보를 토대로 각 클래스 간의 교육성과를 비교할 수도 있기 때문에 교육성과의 요인을 발견하는데 적합한 데이터로 평가되었다. 본 연구에서는 이 데이터를 토대로 C5.0, CART 및 CHAID 의사결정트리 알고리즘을 적용하여 요인을 발견하는 실험을 진행하였다. 단지 각 알고리즘을 적용하고 해석하는데 목적을 두지 않고 각 알고리즘별로 차이가 발생하는 원인을 살펴보고, 여러 의사결정트리 알고리즘의 수행결과를 종합하여 중요한 요인들을 발견하고, 입력변수들의 중요성을 측정하는 방법을 제시하였다.

본 논문의 2장에서는 교육성과에 관련된 연구에서 제시하고 있는 요인의 분류들을 정리하고, 3장에서는 단일 의사결정트리를 적용하고 그 결과를 해석한 내용을 설명하였다. 그리고 4장에서는 단일 의사결정트리 적용의 문제점을 제시하고, 이 문제를 해결할 수 있는 다중 의사결정트리 적용방법을 제시하였다. 마지막으로 5장에서는 연구의 내용을 정리하고 향후 연구를 제시하였다.

Ⅱ. 교육성과 관련 연구

교육성과에 영향을 미치는 요인들은 매우 다양하게 존재하는데, 효과적인 수업의 요인은 일반적으로 다음의 네 가지로 구분될 수 있다(소연희 2006). 첫째는 학습자의 특성, 두 번째는 교육자의 특성, 세 번째는 교육의 내용, 마지막으로 네 번째는 수업환경이다. 본 장에서는 이 네 가지 요인에 대한 기존의 연구들을 정리하고, 본 연구의 실험에서 사용된 각 요소에 해당하는 속성들을 설명하였다.

1. 학습자 특성

학습자의 특성에는 학습자의 인지능력, 사전지식, 적성, 인구통계학적 특성 및 흥미 등이 포함된다. 인지능력은 학생 개인의 학습능력을 의미하며 선천적인 요소에 해당하기 때문에, 수업환경을 개선해도 교육성과에 큰 차이가 나지 않는다는 연구도 있다(김성숙 1999). 또한 사전지식은 교육의 성과와 밀접한 관계가 있으며, 사전지식의 양이 적을수록 교육자의 역할이 커진다는 연구가 있다(소연희, 2006). 본 연구에서는 학습자의 특성으로 학생의 소속대학, 소속학부(학과), 학년의 정

보를 고려하였다. 학생이 IT기술에 관련성이 높은 대학, 학과일수록 '컴퓨터활용1' 강좌에 대한 사전지식 및 학습능력이 상대적으로 우수할 것이며, 이러한 학습자 특성이 교육성과에 영향을 줄 수 있음을 고려하였다.

2. 교육자 특성

교육자 특성에는 담당교수의 강의방법, 강의능력, 인구통계학적 속성 등이 해당된다. 학습자의 학습능력의 차이가 존재할 지라도 교육자의 역할에 따라 교육성과가 좌우될 수 있기 때문에, 학습자의 특성보다는 교육자의 교수법 개선이 교육성과에 더 중요하다는 주장이 있다(허인숙, 2002). 강의방식으로는 개별화학습, 문제중심학습, 협동학습 및 컴퓨터 보조학습 등의 강의방법들이 제시되고 있다. 개별화학습은 교수 당 학생수를 고려할 때 오프라인 수업에서는 현실적으로 어렵지만 담당교수가 온라인 강의를 활용할 경우 접근할 수 있는 방식을 의미한다. 특히 최근에는 컴퓨터와 인터넷의 사용이 일반화되면서 컴퓨터 시스템을 활용한 보조 수업을 통해 교육성과의 향상을 꾀하고 있다. 교육성과를 최대화하기 위해 강좌에 적합한 강의방식을 채택하거나 컴퓨터시스템을 활용하는 것은 담당교수의 교육 능력에 해당한다. 본 연구에서 대상으로 한 '컴퓨터활용1' 강좌의 경우 담당교수 별로 교육방법에는 큰 차이가 존재하기 않으며, 강의능력을 평가할만한 속성의 확보가 어려웠기 때문에, 담당교수의 나이, 성별 등의 인구통계학적 속성 정보만을 고려하여 분석을 수행하였다.

3. 교육 내용

교육 내용은 강의에서 포함하고 있는 교과내용, 과제내용 등을 의미한다. 학습자가 배워야할 내용이 포함되지 않거나 또는 지나치게 많은 학습 내용을 포함하는 경우 교육성과에 부정적인 영향을 미칠 수 있다. 그 외로 학습자에게 흥미와 동기를 유발하기 위한 자기관련성, 신기성, 모호성의 내용을 포함하는 것이 교육성과에 긍정적인 영향을 준다는 연구가 있다. 자기관련성은 교과내용이 학습자 자신과 얼마나 관련이 있고 실제적으로 유용한가에 대한 내용이며, 모호성은 내용이 명확한 정보 전달보다는 모호함을 제공함을 통해 학습자로 하여금 비판 및 추론을 통한 문제해결능력을 높여준다는 내용이다. 이러한 클래스 간 교육 내용의 차이는 교육성과에 영향을 미칠 수 있다. 그러나 각 클래스의 교육 내용을 발견하여 분석에 적용하기 어려운 측면이 있다. 특히, 숭실대학교 '컴퓨터활용1' 강좌의 경우 강의내용 및 교재가 공통적으로 선정되어 진행되기 때문에 각 클래스 별로 교육내용의 두드러진 차이가 없었다. 따라서 본 연구의 실험에서는 교육 내용에 관련된 요인들을 고려되지 않았다.

4. 수업 환경

수업 환경은 교육성과에 영향을 주는 또 하나의 중요한 요소이다. 수업 환경은 강의실 시설 등의 유형적인 요소 뿐 아니라 성적평가 방법, 자율성의 정도 등의 무형적인 요소를 포함한다. 학습자가 편안하고 자유롭게 참여할 수 있는 수업의 환경 조성이 요구되며, 다양한 콘텐츠의 활용이 가능하도록 교육시설의 현대화가 이루어져야 한다. 또한 피드백의 여부가 학습의 효과를 높이기 위한 결정적인 요인이라고 주장되기도 한다(허인숙, 2002). 피드백은 정보전달의 역할 뿐 아니라 학습동기를 유발하거나 유지시키는 역할을 하기 때문이다. 평가의 방법으로는 경쟁(상대평가) 또는 비경쟁(절대평가)에 대한 다양한 관점의 연구들이 있다. 본 연구에서는 강의실 번호, 강의실 크기의 강의실에 관련된 정보와 9시 수업 유무의 정보로 강의시간에 대한 정보를 포함하여 분석하였다. 일반적으로 아침 9시 수업의 경우 결석 및 지각으로 학생들의 출석률 및 참여도가 상대적으로 저조한 상황을 고려한 것이다. 또한 강의실의 크기는 대형(70인), 중형(55인), 소형(30인)으로 구분하여 사용하였다. 대형 강의실의 경우 수업의 집중도가 떨어질 수 있는 문제를 고려하였다.

Ⅲ. 의사결정트리를 이용한 요인의 발견

본 장에서는 데이터마이닝의 대표적인 분류 기법인 의사결정트리(Decision Tree)를 사용하여 교육성과의 요인을 발견하는 실험을 수행하고 그 결과를 설명하고자 한다. 일반적으로 결정트리를 사용하는 연구에서 하나의 의사결정트리 알고리즘만을 적용하여 실험하고 결론을 내리는 경우가 많다. 그러나 적용한 알고리즘에 따라 상이한 결과가 나타나는 경우가 많기 때문에 본 연구에서는 대표적인 알고리즘인 C5.0, CART 그리고 CHAID 를 모두 적용하여 교육성과의 요인을 발견하는 분석을 수행하였다.

1. 실험 데이터

본 연구의 실험에서는 숭실대학교의 1학년 교양필수 과목인 '컴퓨터활용1'의 2009년의 결과 데이터를 사용하였다. 본 과목은 64개의 클래스가 개설되었으며, 12

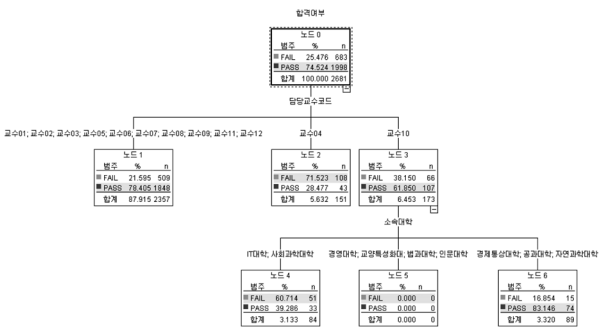
명의 교수가 강의를 담당하였고, 총 2,681명의 학생이 수강한 후 시험까지 응시하였다. 데이터에는 담당교수 정보(ID, 성별, 나이), 학생 정보(학생ID, 소속대학, 소속학과, 학년) 그리고 강의환경 정보(강의실번호, 강의실크기, 강의시간, 강의기간) 및 시험결과(합격여부)의 내용들이 포함되어 있다.

2. 의사결정트리 수행 결과

분석에서는 SPSS Modeler 13.0 프로그램을 사용하였다. 분석 시 '합격여부' 속성을 종속변수로 설정하였고, 나머지 모든 속성들을 입력변수로 설정한 후 분석을 수행하였다. 아래에 C5.0, CART 그리고 CHAID 알고리즘 순으로 수행 결과와 그에 대한 해석 내용을 기술하였다.

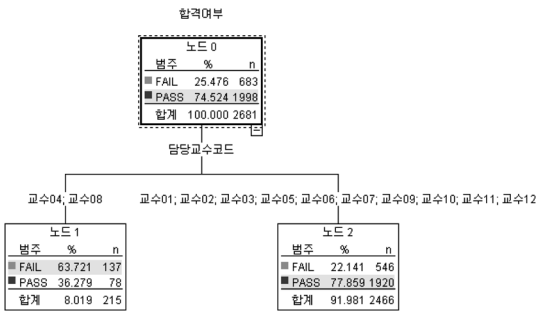
[그림 1]은 C5.0 알고리즘을 수행한 결과이다. 담당교수가 가장 중요한 요인으로 선택되었으며, '교수10'의 경우 학생의 소속대학에 따라 교육성과에 차이가 있는 것으로 결과가 도출되었다.

[그림 2]는 CART 알고리즘을 수행한 결과이다. 동일하게 '담당교수'가 중요한 속성으로 선별된 것을 확인할 수 있다. C5.0의 경우 '담당교수' 속성이 선택된



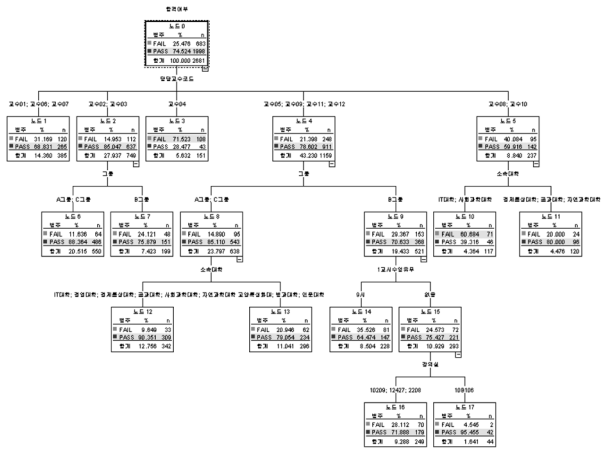
[그림 1] C5.0 의사결정트리 분석 결과

[Fig. 1] Result of C5.0 Decision Tree Analysis



[그림 2] CART 의사결정트리 분석 결과

[Fig. 2] Result of CART Decision Tree Analysis



[그림 3] CHAID 의사결정트리 분석 결과

[Fig. 3] Result of CHAID Decision Tree Analysis

후 3개의 하위 노드로 분리되었으나 CART 알고리즘은 '담당교수' 속성에 의해 2개의 하위 노드로 분리된 차이를 볼 수 있다. 이것은 CART 알고리즘은 항상 이진(Binary) 분기 방식만을 사용하기 때문에 나타나는 차이이다. 이진 분기는 결과를 단순하게 표시하여 가독성을 높이는 장점이 있지만, 정확도 측면에서는 오히려 단점이 될 수 있다. CART 분석 결과에서는 '담당교수'만이 선택되었고, 다른 모든 변수들이 선택되지 않은 점이 C5.0 및 CHAID의 수행 결과와 비교된다.

[그림 3]은 CHAID 알고리즘을 수행한 결과이다. C5.0 및 CART 와 마찬가지로 '담당교수'가 가장 중요한 속성으로 파악되었으며, 하위 노드에서는 그룹, '소속대학'이 교육성과에 영향을 미치는 중요한 속성으로 파악되었다.

3. 결정트리 결과 해석

분석된 결과 트리에 나타난 교육성과에 영향을 미치는 요인으로 파악된 변수들을 <표 1>에 정리하였다. '담당교수'가 선택된 점은 동일하지만, 발견된 요인의 개수 및 내용에 상당한 차이가 존재함을 확인할 수 있었다.

<표 1>에서 보여주는 바와 같이, 적용하는 알고리즘에 따라 발견되는 요인에 큰 차이가 발생할 수 있으며, 다른 말로 하면 하나의 의사결정트리만을 사용할 경우 중요한 요인을 놓칠 수 있다는 점을 고려해야 한다. 따라서 본 연구에서는 3개의 알고리즘을 통한 분석 결과를 모두 고려하여 분석하였다. 분석 결과를 아래에 네 가지 내용으로 구분하여 설명하였다.

첫 번째로, '담당교수'가 교육 효과에 영향을 미치는

<표 1> 의사결정트리 알고리즘 별로 발견된 요인들
<Table 1> Discovered factors of Decision Tree Algorithms

종류	요인개수	변수 리스트(중요도 순)
C5.0	2개	담당교수, 소속대학
CART	1개	담당교수
CHAID	5개	담당교수, 소속대학, 그룹, 9시수업유무, 강의실

<표 2> 평균 합격률에 의한 담당교수의 레벨 구분
<Table 2> Group of Teacher on Average Pass Rate

구분	해당 담당교수	평균합격률
수준1	교수2, 교수3	85%
수준2	교수5, 교수9, 교수11, 교수12	79%
수준3	교수1, 교수6, 교수7	68%
수준4	교수8, 교수10	59%
수준5	교수4	30%

가장 중요한 요인으로 파악되었다. C5.0, CART 그리고 CHAID의 모든 결과에서 1순위로 '담당교수'가 선택된 것을 통해 분명히 알 수 있다. 즉, 각 교수의 교육능력이 소속대학/학과에 의해 발생하는 학생들의 컴퓨터 사용 능력보다 중요함을 알 수 있다. CHAID 의사결정트리 분석 결과를 통해 담당교수들의 교육 능력은 담당한 학생들의 합격률에 따라 아래와 같이 몇 개의 수준으로 구분될 수 있다.

두 번째로, 학생의 '소속대학'이 교육성과에 영향을 미치는 것으로 파악되었다. C5.0, CHAID의 분석 결과에서 '소속대학'이 '담당교수' 다음으로 중요한 속성임을 보여주고 있다. CHAID 분석결과를 토대로 해석하면 '수준2'에 속하는 담당교수에게 수강한 학생들의 경우 IT대학, 공과대학 등에 소속된 학생들이 인문대학 등 상대적으로 컴퓨터에 연관이 적은 학생들에 비해 합격률이 높은 것을 볼 수 있다. 반면, '수준4'에 속하는 담당교수에게 수강한 학생들의 경우 IT대학과 사회대학의 학생들이 다른 대학보다 상대적으로 낮은 교육성과가 나타남을 볼 수 있다. 이것은 IT대학과 사회대학의 학생들의 학업 능력이 부족해서라기보다는 IT대학의 경우 B그룹(중간고사) 기간에 강의가 이루어졌고, 사회과학대학의 경우 대부분의 학생들이 '교수6'의 수업을 들었기 때문으로 파악되었다. '담당교수', '소속대학', '그룹(강의기간)'의 세 변수가 서로 간에 상관관계가 높기 때문에 트리의 모양으로만 해석하기보다 연관된 다른 변수를 고려하여 해석해야 한다. 의사결정트리 분석은 매개효과, 상호작용 등 변수들 간의 상관관계에 견고하지 못한 문제가 있기 때문에, 본 연구에서는 변수들 간의 상관관계를 고려하여 의사결정트리의 결과를 해석하였다.

<표 3> 그룹별 강의 기간 구분
<Table 3> Teaching Period on Group

구분	강의기간	비고
A그룹	2009/3/2 ~ 2009/4/4	학기초
B그룹	2009/4/6 ~ 2009/5/9	중간고사 기간 중
C그룹	2009/5/11 ~ 2009/6/13	기말고사 시작 전

세 번째로, '그룹(강의기간)'이 교육성과에 영향을 미치는 것으로 파악되었다. '컴퓨터활용1' 강좌는 1학점 1시간 강좌이다. 보통 매주 1시간 씩 수업하는 것이 일반적이지만, 본 강좌에서는 강의의 집중도를 위하여 5주 동안 주 3시간 수업으로 운영된다. 첫 5주는 A그룹, 다음 5주는 B그룹, 마지막 5주는 C그룹에 해당된다. B그룹의 경우 중간고사 기간에 수업이 진행되기 때문에 다른 과목들의 시험과 동시에 수업이 진행되기 때문에 학생들의 수업 참여도가 상대적으로 부족한 것을 이유로 분석할 수 있었다.

네 번째로, 9시 수업의 유무가 교육 결과에 영향을 미치는 것으로 파악되었다. 9시 수업의 경우 지각하는 학생들이 많아 수업의 참여도가 떨어지는 경향이 있게 때문으로 이해할 수 있다. 특히, 이러한 특성은 B그룹(중간고사 기간)에 분명하게 나타난다. 다른 과목들을 위한 시험 준비로 지각 및 결석하는 경우가 많이 발생하기 때문이다. 중간고사 기간 특히 9시에 수업이 진행되는 클래스의 경우 학생들의 성실한 수업 참여를 지도해야할 것으로 판단된다.

그 외에 강의실이 교육성과에 영향을 미치는 요인으로 파악되었다. 그러나 10B106 강의실이 02208, 10209, 12427 강의실보다 교육 여건이 좋다고 해석할 수 있지만, 실제 강의실 여건을 고려할 때 10B106의 강의 여건이 좋다고 판단할 수 는 없다. 변수가 다중공선성 점검에서 강의실과 담당교수는 매우 높은 연관성을 갖는 것으로 파악되었다. 즉, 강의실이 선택된 것은 담당교수별로 강의실이 구분되기 때문에 발생하는 것으로 이해할 수 있다. 따라서 의사결정트리에 강의실이 하나의 요인으로 도출되었으나 본 연구에서는 강의실에 의한 교육 효과의 차이는 없다고 판단하였다.

Ⅳ. 다중 의사결정트리를 통한 요인의 발견

1. 결정트리 분석의 취약점

의사결정트리 분석은 데이터마이닝의 대표적인 분

석 기법으로서 요인의 발견과 의사결정을 위해 다양한 분야에서 활용되고 있다. 그러나 의사결정트리 분석의 특성을 제대로 이해하지 못하고 데이터 분석에 적용할 경우 잘못된 해석을 내릴 수 있는 위험이 있다. 이에 관련하여, 결정트리는 모수적 분석방법에 속하며, 입력 데이터의 특성에 따라 부정적인 영향을 받는 문제가 제시되고 있다(Selwyn, 2008).

가. 중요한 요인의 미발견

실제로 영향력이 큰 중요한 변수이지만 결정트리에는 표시되지 않을 수 있다. 이러한 경우는 크게 두 가지 원인에 의해 발생된다. 첫째는 원인은 입력변수들 간의 높은 연관성이다. 수치형 변수에서는 다중공선성(Multicollinearity)이라는 용어로 문제점이 제시되어 있다. 예를 들어, 두 변수 A, B가 종속변수 T에 높은 영향을 주는 경우에도 두 변수가 A, B가 높은 연관성을 갖는다면 그 중 하나의 변수만 트리에 나타나게 되는 것이다. 두 번째 원인은 2개 이상의 입력변수들이 비슷한 변별력을 갖는 상태이다. 예를 들어, 두 변수 C, D가 종속변수 T에 미치는 영향이 비슷하다고 하자. 그러면 평가기준에 따라 C5.0에서는 C변수가 선택되고, CHAID에서는 D변수가 선택되는 차이가 발생할 수 있다. 상위 레벨에서의 서로의 경쟁에 의해서 선택되는 변수에 차이가 발생하면 그 하위 트리에서 발견되는 변수들에 차이가 나타나기 때문에 최종적인 결정트리를 통해 발견되는 요인에 큰 차이가 나타날 수 있다(박현경·송문섭, 2003).

나. 알고리즘에 따른 상이한 결과 도출

분석에 적용한 알고리즘에 따라 상이한 결과가 도출될 수 있다. 물론 분석법의 특성에 차이가 있기 때문에 어느 정도 차이가 발생할 것은 고려할 수 있으나 발견되는 요인이 다를 경우 해석 결과가 달라질 수 있음으로 적용에 유의가 필요하다. 알고리즘에 따른 결과의 차이는 주로 변수 선택을 위한 평가기준의 차이에서 발생한다. 가장 많이 사용되는 C5.0, CART 그리고 CHAID 알고리즘에 대해서 비교해보면, C5.0의 경우 Entropy Gain을 분기 기준으로 사용하며, CART의 경우 Gini Index와 분산의 차이를 사용하고, CHAID의 경우 t검정과 F검정을 사용하는 점에서 차이가 있다. 그 외에도 분기 방법에 따라 모형이 변경될 수 있다. CART는 하위노드로 분기할 때 단순화된 트리를 생성하기 위하여 항상 이진(Binary) 분기만을 사용하는 것이 특징이다. 또한 C5.0의 경우 상업화된 프로그램에 따라 범주형 변수의 분기 방식에 차이가 있다. 기본적으로 알고리즘에서는 모든 범주값에 따라 하위 노드를 분리하지만, 트리가 지나치게 커지고 복잡해지는 문제를 해결하기 위하여 유사성이 높은 범주값들을 묶어 소수의 노드로 분할하는 방법을 사용하기도 한다. 이러한 차이에 따라 기본적인 의사결정트리 알고리즘의 방식은 같다고 하더라도 결과가 상이할 수 있다. 특히 이 원인은 위에서 설명한 변수들 간의 경쟁적 관계에서 나타난다. 본 연구에서는 이러한 문제를 해결하기 위한 하나의 방법으로 여러 알고리즘들을 수행한 후 그 결과들을 종합하여 해석함을 통해 변수의 중요성을 산출하는 방법을 다음 절에서 제시한다.

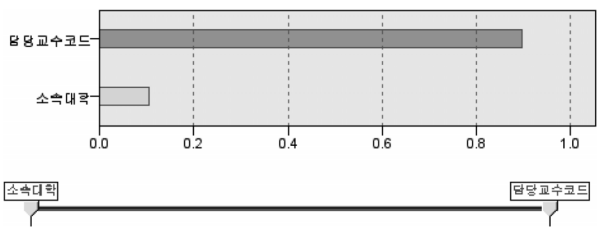
다. 변수의 중요도의 측정과 비교

의사결정트리를 통해서 변수의 중요도를 발견하고자 하는 경우가 있다. 물론 결정트리를 통해서 중요한 규칙(Rule)을 발견할 수도 있지만, 분석가는 의사결정트리를 통해 중요한 변수를 또한 발견하고, 그 변수들 간의 상대적 중요성을 비교하기를 원한다. 이러한 필요에 따라 SPSS Modeler 13.0 프로그램의 경우 의사결정트리를 분석하면 [그림 1]의 결정트리에 대하여 [그림 4]의 변수의 중요도 차트를 제공해준다.

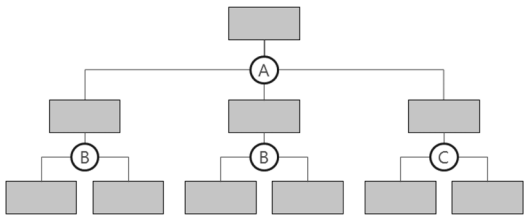
[그림 4]에서는 '담당교수'가 0.95의 중요도로 0.05의 중요도를 갖는 '소속대학'보다 20배 정도 높은 중요성을 갖는 것으로 보여준다. 그러나 [그림 1]의 의사결정트리에서 [그림 4]의 중요도가 산출된 근거는 설명되지 않는 문제가 있다. 본 연구에서는 결정트리의 모양으로부터 변수의 중요도를 산출하는 수식을 다음 절에서 제안한다.

2. 결정트리를 통한 변수의 중요도 계산

결정트리를 사용하는 대부분의 연구에서는 분석을 통해 얻은 트리로부터 중요한 변수(요인)을 발견하고, 이 변수들의 중요도를 비교하고 있다. 물론, 결정트리 분석이 모든 요인들을 다 발견하지 못할 수 있으며, 원래의 종속변수에 대한 주효과(main effect)와 일치하



[그림 4] C5.0에 대한 변수중요도 차트
[Fig. 4] Chart on Variables Importance



[그림 5] 간단한 결정트리 예시

[Fig. 5] A Example of Simple Decision Tree

지 않지만, 그렇다 하더라도 결정트리 분석의 결과로부터 변수의 중요도를 요약하는 것은 의미가 있다.

일반적으로 결정트리로부터 중요성을 비교할 때 트리에서 변수가 사용된 레벨(깊이)이나 횟수를 고려하여 중요성을 비교한다. 예를 들어, [그림 5]에서 레벨 1의 분기점에서 사용된 A가 가장 중요한 요인이며, 그 다음으로 B, C의 순서로 중요하다고 본다. 같은 레벨에 발견되었으나 표시 횟수를 고려하여 B가 C보다 2배 더 중요하다고 볼 수 있다.

그러나 트리가 더욱 복잡해 질 경우 트리로부터 변수의 중요성을 직관적으로 판단하기는 어렵다. 본 연구에서는 결정트리로부터 변수의 중요도를 산출하는 수식으로 식 (1)을 제안한다.

$$IMP_X = \frac{\sum_{All\ Branch\ b\ with\ X} \left(\frac{1}{Branch\,Count(b)}\right)\left(\frac{1}{2}\right)^{depth(b)-1}}{\sum_{n=1}^{depth}\left(\frac{1}{2}\right)^{n-1}} \quad (1)$$

식 (1)에서 IMP_X는 변수 X의 중요도를 의미하며, 0에서 1사이의 값으로 계산된다. $depth$는 의사결정트리의 깊이이며, $Branch\,Count(b)$는 결정트리에서 분기점 b가 속한 레벨의 전체 분기점의 개수이고, 마지막으로 $depth(b)$는 분기 b가 해당된 레벨(깊이)를 의미한다. 식 (1)에 의해 산출된 모든 변수의 중요도의 합계는 1.0이 된다. 예를 들어, [그림 5]의 의사결정트리에서 변수 B의 중요도는 식 (2)에 의해 0.22로 계산된다.

$$IMP_B = \frac{\left(\frac{1}{3}\right)\left(\frac{1}{2}\right)^{2-1} + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)^{2-1}}{\left(\frac{1}{2}\right)^{1-1} + \left(\frac{1}{2}\right)^{2-1}} = 0.22 \quad (2)$$

3. 다중 결정트리를 통한 요인의 발견

1절에서 언급한 바와 같이 적용하는 의사결정트리 알고리즘에 따라 선택되는 변수 그리고 변수들의 중요도에 차이가 발생하는 문제가 있다. 따라서 하나의 결정트리 알고리즘을 사용하기보다는 여러 의사결정트리 알고리즘을 사용하여 그 결과들을 종합함을 통해 이 문제를 보완할 수 있다. 본 연구에서는 다중 결정트리 통해 중요도를 산출하는 수식으로 식 (3)을 제안한다.

$$CIMP_x = \frac{1}{TreeCount}\sum_{all\_used\_Tree} IMP_{t,x} \quad (3)$$

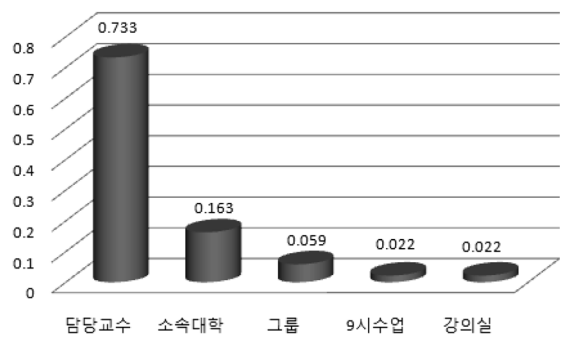
식 (3)에서 $CIMP_X$는 여러 알고리즘의 결과를 종합한 변수 X의 종합적 중요도를 의미한다. 그리고 $Tree\,Count$는 분석에 사용한 의사결정트리 알고리즘의 개수를 의미하고, $IMP_{t,x}$는 알고리즘 t를 사용한 분석 결과 트리를 식 (1)에 적용하여 계산된 변수 X의 중요도이다.

3장의 C5.0, CART 및 CHAID 알고리즘을 적용한 실험 결과에 식 (3)을 적용하여 3개의 의사결정트리를 종합한 변수요도를 산출한 결과를 <표 2>에 정리하였다. '담당교수'가 0.733으로 종합적으로 가장 중요한 변수이고, 다음으로 '소속대학'이 0.163으로 중요한 변수로 발견되었다. 그 외에 그룹, '9시수업유무' 그리고 '강의실'이 중요한 변수로 발견되었다. 하나의 알고리즘만 적용할 경우 '소속대학', '그룹(강의기간)', '9시수업

<표 4> 그룹별 강의 기간 구분

<Table 4> Teaching Period on Group

알고리즘	담당교수	소속대학	그룹	9시수업	강의실
C5.0	0.667	0.333	0.0	0.0	0.0
CART	1.000	0.0	0.0	0.0	0.0
CHAID	0.533	0.156	0.178	0.067	0.067
CIMPx	0.733	0.163	0.059	0.022	0.022



[그림 6] 변수 중요도 막대 차트

[Fig. 6] Bar Chart on Variable Importance

유무', '강의실'의 요인들은 발견되지 않을 수도 있었지만 다중 의사결정트리의 적용과 해석을 통해 충분한 요인들을 발견할 수 있었다. 최종적으로 산출된 변수의 중요도를 <그림 6>의 차트로 표시하였다.

Ⅴ. 결론

본 연구에서는 컴퓨터활용 교육에 영향을 미치는 요인을 발견하기 위하여 데이터마이닝의 의사결정트리 알고리즘을 중심으로 분석을 수행하였다. 데이터의 특성과 분석 알고리즘의 분할 기준에 따라서 상이한 분석 결과를 얻을 수 있기 때문에 본 연구에서는 C5.0, CART 그리고 CHAID 3개의 의사결정트리에 대하여 분석을 수행하였고, 모든 결과 트리 모델을 종합적으로 고려하여 교육 효과에 영향을 미치는 중요한 요인들을 발견하였다.

본 연구에서는 분석된 의사결정트리에서 변수의 중요도를 산출하는 수식과 다중 결정트리를 통하여 결과들을 종합하여 변수의 중요도를 해석하는 방법을 제안하였다. 결론적으로 '담당교수'가 가장 중요한 속성으로 판단되었으며, 그 외에 '소속대학', '그룹(강의기간)', '9시수업유무' 등이 중요한 속성으로 판단되었다.

본 연구에서 제안한 다중결중트리를 이용한 요인의 발견 방법은 기존의 단일 의사결정트리를 이용한 해석보다 신뢰성있는 분석을 할 수 있고 변수들 간의 중요성을 비교할 수 있는 장점이 있다. 그러나 중요한 요인을 놓칠 수 있는 문제를 완전히 해결하지는 못한다. 따라서 향후연구에서는 변수들 간의 높은 연관성, 매개효과, 상호작용 등이 존재할 경우 이러한 변수들 간의 관계를 고려하여 의사결정트리를 해석하는 연구나 다양한 변수들 간의 관계에 견고한 결정트리 알고리즘의 연구가 필요하다.

참고문헌

류춘호·이정호(2003). 대학의 강의평가에 영향을 미치는 학생관련 요인에 관한 연구. 경영교육연구, 32(3): 789-807.

이주리(2009). Data Mining을 이용한 초등학생의 삶의 만족도에 대한 보호요인 및 위험요인 탐색. 아동학회지, 30(1): 11-25.

박동준·최수영(2009). 수학 학업성취도의 영향 요인 분석 연구. 한국수학교육학회, 23(2): 383-398.

김완섭·이수원(2006). 데이터 마이닝을 이용한 설문조사의 심층분석. 공학교육연구, 9(4): 71-82.

김완섭·이수원(2003). 상품별 구매 패턴을 이용한 프로파일 기반 추천과 협력적 추천과의 결합. 데이터마이닝학회 2003 추계학술대회 논문집, 172-176.

한경식·이수원(2005). 대용량 데이터를 위한 전역적 범주화를 이용한 결정 트리의 순차적 생성. 한국정보처리학회지, 12-B(4): 487-498.

한경석·노미현(2004). 데이터마이닝 가상강의 효과와 만족도에 영향을 미치는 주요 요인 분석, 경영교육연구, 19(1): 309-318.

소연희(2006), 효과적인 교실수업에 영향을 미치는 요인 탐색, 교육방법연구, 18(1): 1-22.

한송엽·서경덕(2002). 공학교육 성과를 위한 졸업생 설문조사 사례연구, 공학교육연구, 5(1): 34-49.

허인숙(2002), 사회과 교육에서 사전지식을 고려한 학습과 개념도의 활용, 시민교육연구, 34(2): 235-254.

김신곤·박성용(1999). 의사결정트리 알고리즘의 성과 비교에 관한 연구, 한국경영정보학회 춘계학술대회 Vol. 1999: 371-383.

배정이(2008). 중학생의 학교적응 관련요인의 인과적 구조분석, 대한간호학회지, 38(3): 454-464.

이중섭·이용교(2009). 부모의 교육수준이 자녀의 학업성취 수준에 영향을 미치는 경로, 한국가족복지학, 26: 159-192.

소연희(2009). 수업에 대한 자율성 지각과 다중지능 및 교과흥미가 학업성취에 미치는 영향, 한국아동교육학회, 18(2): 5-18.

박현경·송문섭(2003). 데이터 마이닝에서 변수중요도에 관한 연구, 서울대학교 석사학위논문.

권혜숙·송문섭(2002). 데이터마이닝 패키지에서 분류나무 알고리즘의 비교 연구, 서울대학교 석사학위논문.

Selwyn Piramuthu(2008). Input Data for Decision Tree, *Expert System with Applications, 34*(2): 1220-1226.

Roiger, R. J. & Geatz, M. W.(2003). Data Mining: A TUTORIAL-BASED PRIMER. Addison-Wesley.

Han, J. & Kamber, M.(2006). Data Mining Concept and Techniques. Morgan Kaufmann.

저 자 소 개

김완섭 (Kim, Wan Seop)
2000년: 숭실대학교 컴퓨터학부 졸업
2003년: 동 대학원 컴퓨터학과 석사
2006년: 동 대학원 컴퓨터학과 박사수료
2007년: 숭실대학교 IT대학 컴퓨터학부 전임강사
2008년~현재: 숭실대학교 베어드학부대학 조교수
관심분야: 데이터마이닝, 인공지능, 개인화서비스, 공학교육 등
Phone: 02-828-7169
Fax: 02-6280-7169
E-mail: wskim92@ssu.ac.kr