데이터마이닝을 이용한 설문조사의 심층 분석[1)]

김완섭, 이수원
숭실대학교 컴퓨터학부

An In-depth Survey Analysis Applying Data Mining Techniques

Wanseop Kim, Soowon Lee
School of Computing, Soongsil University

국문요약

학과의 교육목표 달성을 위해서는 순환형 자율 개선 구조를 운영하기 위한 시스템이 필요하며, 설문조사 분석을 통한 교육시스템의 개선은 교육목표 달성을 위한 중요한 요소 중의 하나이다. 일반적으로 설문조사 분석에서는 항목별로 통계적인 분포를 조사하거나 두 개의 항목간의 연관성을 조사하는 분석 방법이 주로 사용된다. 그러나 이러한 분석 방법은 다양한 항목들 간의 상호 연관성을 분석하지 못하는 한계가 있으므로 보다 심층적인 분석방법이 필요하다. 본 논문에서는 데이터마이닝 기법을 적용한 심층적인 분석 기법을 제시한다. 데이터마이닝이란 대용량의 데이터에 숨겨져 있는 지식을 추출해 내는 기법으로 설문분석에도 효과적으로 이용될 수 있다. 본 분석에서는 Clementine 데이터마이닝 도구를 사용하여 숭실대학교 컴퓨터학과의 재학생에 대한 설문자료에 대한 심층 분석을 수행하였다. 분석의 결과로 '학점'과 다른 항목들과의 연관성을 계층적으로 분석할 수 있었으며, '학점'에 대한 학생상담과 학과의 교육 프로그램 개선에 실제적으로 사용할 수 있는 유용한 정보들을 획득할 수 있었다.

Abstract

To accomplish the educational objectives of a department, a system for CQI(Continuous Quality Improvement) is necessary. Improving the educational system by survey analysis is one of the most important factors for accomplishing the educational objectives. In general, survey analysis is carried out by using statistical distribution on an attribute or correlation analysis between two attributes. However, these analysis schemes have a limitation that they cannot find relations among various attributes. In this paper, an in-depth survey analysis method applying data mining techniques is presented. Data mining is a technique for extracting interesting knowledges from a large set of data. Survey from undergraduate students in the School of Computing of Soongsil University is analyzed in

1) 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.

this paper by using a data mining tool, called Clementine. Results of Clementine analysis show the relationship between 'grade', and other attributes hierarchically, and provide useful information that can be applied in student consulting and program improvement.

주제어: 설문분석, 데이터마이닝, 분류, 의사결정트리
Keywords: Survey Analysis, Data Mining, Classification, Decision Tree

Ⅰ. 서 론

1. 연구의 목적 및 필요성

학과의 교육목표를 달성하기 위해서는 순환형 자율 개선 구조를 운영하기 위한 시스템이 필요하며, 설문조사 분석을 통한 교육시스템의 개선은 교육목표 달성을 위한 중요한 요소 중의 하나이다. 공학교육인증을 위한 설문조사에서는 교육목표의 중요도 및 달성도, 학습성과의 중요도 및 달성도, 교육과정을 포함한 학교/학과의 교육 시스템에 대한 만족도, 학년/입학년도/성별/군필여부/전형유형/출신계열/거주형태 등에 대한 개인 정보가 수집되며, 신입생/재학생/졸업예정자/졸업생/산업체 전문가/고용주 등으로 설문 대상 그룹을 세분화하여 해당 설문을 실시한다.

일반적으로 설문 분석은 해당 항목에 대한 통계적인 분포를 분석하거나 두 개의 항목간의 연관성을 분석하는 방법을 주로 사용한다. 하나의 항목에 대한 빈도 분석은 기본 통계 정보를 파악하는데 유용하며, 대부분의 기초적인 분석에서는 이러한 빈도 분석을 많이 수행한다. 또한, 항목들 간의 상관관계를 분석하기 위해서는 두 개의 항목에 대하여 빈도 분석을 하거나 Chi-square test와 같은 기법을 사용한다. 그러나 분석 대상 항목(예: 학점, 학과 만족도, 희망진로 등)은 일반적으로 하나의 항목에만 상관성이 있는 것이 아니라 다양한 항목에 의하여 복합적으로 상관성이 있기 때문에 이러한 기존의 분석방법으로는 충분한 분석에 한계가 있다. 따라서 다양한 항목간의 상관성을 파악하기 위한 좀 더 심층적인 분석 방법이 요구된다. 또한 기존의 분석 방법은 대상 학생들의 분포 특성을 이해하는 데는 도움이 될 수 있지만, 교육시스템의 개선을 위하여 상담이나 교육환경 등의 측면에서 어떠한 개선안이 필요한지를 파악하기에는 어려움이 있다. 따라서 상담이나 교육시스템의 개선에 실제 적용이 가능한 분석 결과를 제공해주는 분석 기법이 요구된다.

2. 연구 개요

본 논문에서는 데이터마이닝 기법을 적용하여 단순 통계분석에서 얻을 수 없었던 심층적인 분석 기법을 제시한다. 데이터마이닝 기법은 대용량의 데이터에 내재되어 있는 유용한 지식을 찾아내는 기법으로, 고객 분석 등의 다양한 분야에서 사용되고 있으며 설문분석에서도 효과적으로 이용될 수 있다. 본 연구에서는 데이터마이닝에서 가장 대표적인 분류 분석 방법인 의사결정트리(Decision Tree) 분석을 통하여 데이터에 숨겨져 있는 의미 있는 지식을 찾아내고, 추출된 지식을 학생 상담과 교육 시스템의 개선에 활용하는 방법을 제시한다. 데이터마이닝을 이용한 설문자료의 분석은 기존의 방법으로 찾아내지 못했던 새로운 지식을 발견할 수 있고, 이러한 지식은 교육 프로그램의

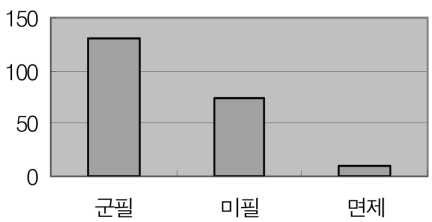
발전방향 및 학생 상담 등에 효과적으로 사용될 수 있다. 특히, 데이터마이닝을 통한 분류 분석은 과거 데이터를 통하여 분류 모델을 생성하고 이를 새로운 학생에 적용하여 학생의 행동 또는 성취도를 예측할 수 있게 한다. 또한 의사결정트리에 의한 분류 모델은 이해하기 쉽고, 학생 상담 및 교육 프로그램 개선에 적용이 가능한 규칙의 형태로 결과를 제공해주는 장점이 있다. 본 연구에서는 SPSS사의 Clementine이라는 데이터마이닝 도구를 사용하여 분석을 수행하였다.

본 논문의 2장에서는 데이터마이닝에 대한 정의와 수행 절차에 대하여 간단히 소개하며 데이터마이닝이 설문 분석에 어떻게 활용될 수 있는가를 기술한다. 3장에서는 설문자료로부터 데이터마이닝 분석을 수행한 결과와 활용방안을 설명하고, 4장에서는 분석 결과에 대한 평가 및 더욱 효율적인 데이터마이닝을 위하여 분석과정에서 파악된 개선 사항을 언급한다.

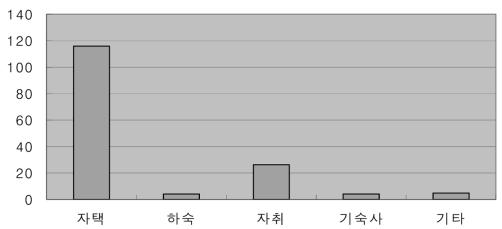
Ⅱ. 데이터마이닝의 소개

데이터마이닝은 대용량의 데이터로부터 의미 있는 정보를 찾는 분석 방법이며, 기존의 통계분석이나 OLAP(On-Line Analytical Processing) 또는 데이터베이스의 질의어로는 찾아내기 힘든 숨겨진 정보들을 추출하기 위하여 사용된다. 기존의 방법들은 일반적으로 가설이 있는 상황에서 정보를 찾아낼 수 있지만, 데이터마이닝의 경우는 가설을 알지 못하는 상황에서 정보를 찾아낼 수 있는 장점이 있다.

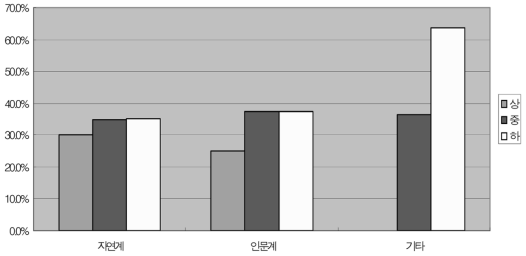
기존의 설문 분석에서는 하나의 항목에 대한 빈도 분석을 많이 사용한다. 예를 들어, [그림 1]은 숭실대학교 컴퓨터학부 2학년 재학생의 군필여부에 대한 빈도 챠트이며, [그림 2]는 거주형태에 대한 빈도 챠트이다. 이러한 분석은 하나의 항목에 대한 통계 분포를 파악할 수 있는 면에서 유용하지만, 다른 항목과의 관계를 파악할 수는 없는 한계가 있다. 따라서 두 개 이상의 항목에 대한 상관관계를 분석하기 위해서는 [그림 3] 또는 [그림 4]와 같은 빈도 분석을 사용하거나 Chi-square test와 같은 기법을 사용한다. [그림 3]은 출신계열과 학점과의 상관성을 분석한 챠트로, 출신계열이 자연계, 인문계, 기타(검정고시 등)인지에 따라 학점의 분포가 어느 정도의 영향이 있는 지를 파악할 수 있다. [그림 4]는 재학생들의 학부 홈페이지 접속 횟수와 학점과의 상관성을 분석한 챠트이다. 이 분석을 통하여 학점이 상, 중, 하의 각 범주값에 속하는 여부가 학생들의 홈페이지 접속 횟수와 영향이 있는지를 파악할 수 있으며, 본 설문조사의 결과는 학점이 높은 학생들일수록 학과 홈페이지에 자주 접속하는 특징을 나타낸다. 이것은 학부 홈페이지의 접속 횟수가 수업 관련 공지



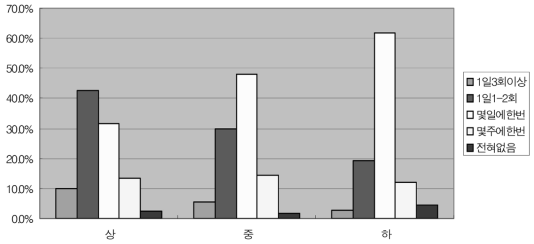
[그림 1] 재학생의 군필 여부



[그림 2] 2학년 재학생의 거주 형태



[그림 3] 출신계열에 따른 학점의 분포



[그림 4] 재학생의 학점에 따른 홈페이지 접속 횟수

사항의 습득과 학습에 유용한 자료의 공유에 긍정적인 영향을 미치는 것으로 해석될 수 있다. 이러한 기존은 분석은 학점과 출신계열 또는 학점과 홈페이지 접속 횟수가 관계가 있을 것이라는 가설이 있는 상태에서 찾아질 수 있는 정보이다. 그러나 이러한 분석 방법은 분석자가 가설을 찾지 못한다면 숨겨져 있는 유용한 지식을 찾아낼 수 없는 한계가 있다. 하나의 항목의 다양한 항목들에 의하여 복합적으로 영향을 받기 때문에 데이터마이닝을 이용한 심층 분석이 요구된다.

데이터마이닝의 분석 방법에는 여러 가지가 있으나 대표적인 분석 방법으로 분류(Classification) 분석, 군집(Clustering) 분석, 연관규칙(Association Rule) 분석 등이 있으며 기타, 추정(Estimation) 및 예측(Prediction), 순차패턴분석(Sequential Pattern Analysis) 등을 들 수 있다.

첫째, 분류 분석은 데이터마이닝에서 가장 많이 사용되는 분석방법으로서 특정 항목에 대한 항목값이 포함된 과거의 데이터로부터의 특성을 파악하여 모델을 생성하고 이를 토대로 새로운 레코드에 대한 항목값을 예측하는 방법이다. 분류분석에서는 다양한 방법들이 사용되나 의사결정트리(Decision Tree)에 의한 분류 기법이 가장 많이 사용된다. 의사결정트리에 의한 분류 분석을 통하여 특정 항목에 영향을 미치는 다른 항목들의 계층적인 구조를 파악할 수 있다. 또한, 의사결정트리 분석의 결과는 트리 형태의 모델로 제공되어 이해하기 쉽고 적용이 가능한 규칙(Rule)의 형태로 변환이 가능하기 때문에 유용하게 적용될 수 있는 것이 장점이다. 예를 들어, 학점모델은 학생들의 학점 상담에 활용될 수 있으며, 학과 만족도 모델은 학교 및 학과 차원에서의 학생 및 산업체 고용주의 만족도 향상을 위한 개선 방안을 도출하는데 사용할 수 있다.

둘째, 군집 분석이란 비슷한 성격을 갖는 레코드들을 그룹핑하는 분석 방법이다. 예를 들어, 학생들이 갖는 다양한 항목들을 모두 고려하여 특성이 비슷한 학생들의 그룹을 생성할 수 있다. 생성된 그룹은 상담에서 활용될 수도 있으며, 모든 학생들을 동일한 형태로 상담하는 것이 아니라 몇 개의 세분화 된 그룹으로 구분한 후 학생들의 특성에 맞게 상담하는 것도 가능하다.

셋째, 연관규칙 분석이란 데이터 안에 존재하는 항목 간의 연관 관계를 찾아내는 작업이다. 연관규칙 분석은 주로 CRM(Customer Relationship Management)의 마케팅에서 사용되며 함께 구매되는 상품들 간의 연관성을 찾아낸다. 이 분석 방법을 설문분석에 적용한다면 과목들 간의 연관성에 대한 규칙을 얻을 수 있다. 예를 들어, “A과목을 수강한 학생들은 B과목을 수강한다”라는 형태의 규칙을 과거수강신청 이력데이터를 통하여 획득할 수 있고, 이 규칙을 다음 학기 학생들의 수강신청 지도에 사용할 수 있다.

데이터마이닝의 주요 분석 방법을 설문 분석에 활용할 경우 획득할 수 있는 정보의 형태와 획득한 정보를 바탕으로 하여 적용이 가능한 개선 방안은 <표 1>과 같이 정리될 수 있다. 본 연구에서는 이 방법들 중에서 분류 분석에 초점을 맞추어 분석을 수행한다.

<표 1> 학생 설문분석에서 데이터마이닝의 방법과 적용방법

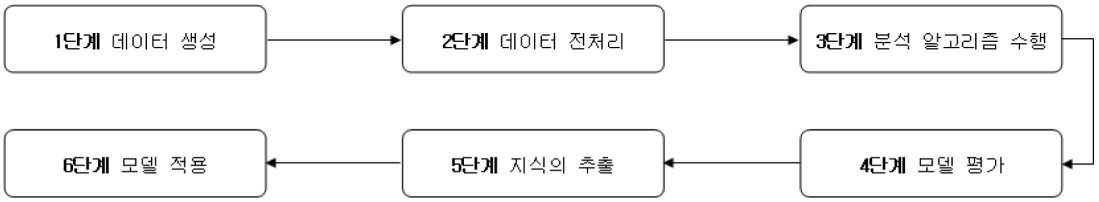
분석 방법	획득 정보의 형태	가능한 적용 방안
분류 (Classification)	학점 모델 학생의 학교/학과 만족도 모델 고용인의 학교/학과 만족도 모델	학점상담, 학교 발전 방안, 학생 만족도 향상 위한 개선 방안, 고용인 만족도 향상 위한 개선 방안
군집 (Clustering)	유사한 특성을 갖는 학생군의 형성	학생특성 분석, 특성이 비슷한 상담 그룹의 형성
연관규칙 (Association Rule)	과목간의 연관성	수강신청 과목의 추천

Ⅲ. 데이터마이닝을 이용한 설문 분석

본 연구에서는 데이터마이닝 분석을 위하여 SPSS사의 Clementine이라는 데이터마이닝 도구에서 지원하는 C5.0 의사결정트리(Decision Tree) 분류 알고리즘을 사용하였다. 1절에서는 데이터마이닝을 수행하기 위해 데이터를 준비한 과정을 기술하며, 2절에서는 '학점' 항목을 중심으로 분류 분석을 수행한 방법과 결과를 설명한다.

1. 연구 방법 – 데이터의 준비

데이터마이닝 분석은 여러 단계를 거쳐 수행되므로 데이터마이닝의 각 단계를 이해하고 올바로 수행하여야만 의미 있는 분석결과를 얻을 수 있다. 데이터마이닝의 일반적인 실행단계는 [그림 5]와 같이 표현된다.



[그림 5] 데이터마이닝 분석의 수행 절차

[그림 5]는 데이터마이닝의 일반적인 실행단계를 6단계로 단순화하여 표현한 그림이다. 1단계인 데이터 생성에서는 데이터마이닝에서 분석하고자 하는 자료를 생성한다. 2단계 데이터 전처리에서는 생성된 데이터를 데이터마이닝 알고리즘에 적용할 수 있는 형태로 변환한다. 이 단계에서

는 데이터 선택과 통합, 데이터 정제, 값변환, 필드삭제, 새 필드생성 등의 작업을 통해 분석에 적합한 데이터로 변환한다. 3단계 분석 알고리즘의 수행에서는 다양한 데이터마이닝의 분석기법 중에서 적절한 분석 방법을 선택하고 알고리즘을 수행하는 단계이다. 이 단계를 통해서 생성된 정보를 모델이라고 한다. 4단계에서는 모델을 평가한다. 모델의 수정이 필요하다면 1, 2단계를 다시 보완한 후 3단계를 수행하여 새로운 모델을 생성한다. 이러한 과정을 통하여 모델이 생성되었으면 5단계 지식의 추출에서는 모델에서 우리가 활용할 수 있는 유용한 지식을 추출한다. 6단계 모델 적용의 단계에서는 추출된 지식을 실제로 활용하는 단계이다. 본 절에서는 본 분석에서 적용한 1단계 데이터 생성과 2단계 데이터의 전처리에 대해서 자세히 설명한다.

1단계, 데이터의 생성 단계는 데이터마이닝을 수행하기 위한 데이터의 수집 및 추출, 통합 과정을 의미한다. 데이터는 데이터베이스에 저장되어 있을 수도 있고, 또는 수작업으로 기록된 설문지에 존재할 수도 있다. 본 연구에서는 설문지를 통하여 데이터를 수집하였고 이를 파일에 저장하여 분석을 수행하였다. 설문조사는 2006년 9월에 수행되었으며 신입생, 재학생, 졸업예정자, 졸업생, 고용주, 교수 및 산업체 산학자문위원을 대상으로 별도의 설문지로 설문을 실시하였다(<표 2>). 본 설문에 응답한 인원은 신입생 156명, 재학생은 343명 (2학년 231명, 3학년 112명), 졸업예정자 175명, 고용주 27명, 교수 17명, 산학자문위원 5명이다.

2단계, 데이터 전처리 단계는 입력된 자료를 데이터마이닝 알고리즘에 입력이 가능한 형태로 변환하는 과정이다. 데이터마이닝을 통해 의미있는 정보를 얻기 위해서는 품질이 좋은 데이터를 입력으로 해야한다. 이 과정에서는 결측치(Missing Value) 처리, 이상치 처리, 새로운 필드의 생성 등의 작업이 필요하다. 결측치 처리는 설문 응답자가 입력하지 않은 항목의 값에 대한 처리를 의미한다. 일반적으로 결측치 값을 갖는 레코드의 경우는 삭제하거나 결측 필드의 값을 평균값 등의 특정 값으로 대치하는 처리를 한다. 본 분석에서는 결측치인 경우 “미입력”이라는 값을 입력한 후 수행하였다. 이상치 처리는 설문 응답자나 입력자의 잘못된 입력값에 대한 처리를 의미한다. 예를 들어, 나이를 묻는 항목에 범위에 맞지 않는 값이 입력된 경우 이상치 처리를 하여야 하며, 이러한 경우 적절한 값으로 이상치를 변경해 주는 것이 필요하다.

2. 연구 결과 – 학점에 대한 분류 분석

본 분석에서는 학점에 관심을 두고 분류 분석을 수행하였다. 어떤 요인들이 학생들의 학점에 영향을 주는가를 파악하고, 이를 토대로 학생들의 학점관리를 지도하고 특히 학점이 저조한 학생들을 위한 상담모델을 추출하는 것을 목표로 하였다. 분류 알고리즘을 수행하기에 앞서 학점에 대한 항목에 대하여 전처리를 수행하였다. 원 설문지의 학점을 묻는 문항에서는 학점을 0.5점 단위로 6단계로 구분하여 입력을 받았으나([그림 6]), 보다 효과적인 분석결과의 해석을 위하여 학점 구분을 상(3.5이상), 중(3.0이상 3.5 미만), 하(3.0 이하) 3단계로 변경하였다.

(11) 지난학기까지의 평점

① 4.0 이상 ② 3.5~4.0 ③ 3.0~3.5

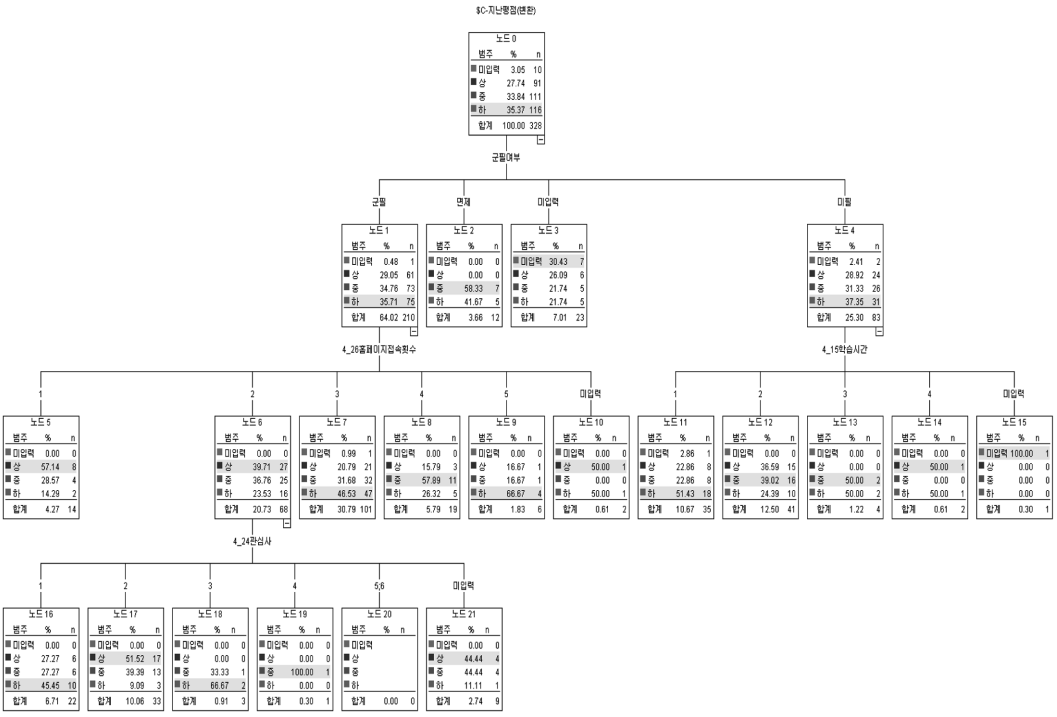
④ 2.5~3.0 ⑤ 2.0~2.5 ⑥ 2.0 이하

[그림 6] 학생들의 학점을 파악하는 문항

<표 2> 대상별 설문 문항 구성

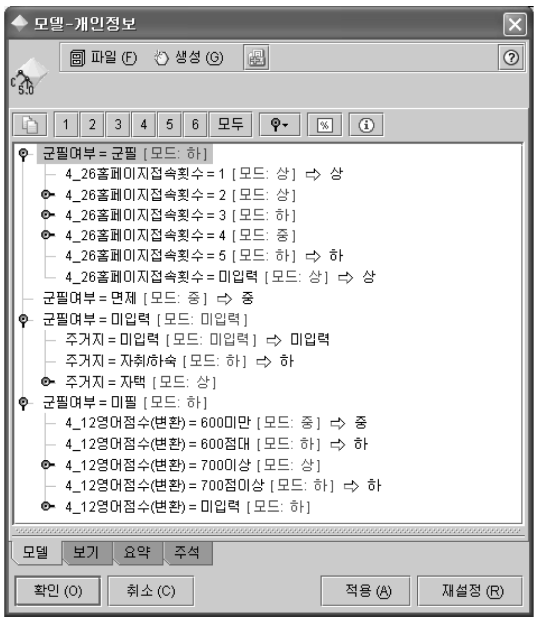
설문지 유형	설문 대상 그룹	설문 문항의 구성
신입생	2006년 입학 신입생	– 교육목표 (객관식 8문항) – 학습성과 (객관식 13문항, 주관식 1문항) – 만족도 조사 (객관식 43문항) – 개인정보 (객관식31문항, 주관식 4문항)
재학생	2학년 및 3학년 재학생	
졸업예정자	4학년 재학생 중 졸업예정자	– 교육목표 (객관식 8문항) – 학습성과 (객관식 13문항, 주관식 1문항) – 교양수업 관련 (객관식 15문항, 주관식 2문항) – 수학 및 기초과목 관련 (객관식 9문항, 주관식 2문항) – 전공과목 관련 (객관식 44문항, 주관식 4문항) – 만족도 조사 (객관식 43문항) – 개인정보 (총 37문항)
졸업생	졸업생	– 교육목표 (객관식 8문항) – 학습성과 (객관식 13문항, 주관식 1문항) – 전문교양 (객관식 15문항, 주관식 2문항) – 수학 및 기초과학 (객관식 9문항, 주관식 2문항) – 전공 (객관식 44문항, 주관식 4문항) – 중요한 툴 및 기술 (주관식 1문항) – 만족도 조사 (객관식 11문항) – 학부에 대한 건의사항 (주관식 1문항) – 개인정보 (객관식 12문항, 주관식 4문항)
고용주	졸업생의 직장상사	– 교육목표 (객관식 8문항) – 학습성과 (객관식 13문항, 주관식 1문항) – 전문교양 (객관식 15문항, 주관식 2문항) – 수학 및 기초과학 (객관식 9문항, 주관식 2문항) – 전공 (객관식 44문항, 주관식 2문항) – 중요한 툴 및 기술 (주관식 1문항) – 학과에 대한 만족도 점수 (주관식 1문항) – 학부에 대한 건의사항 (주관식 1문항) – 개인정보 (객관식 7문항, 주관식 3문항)
산업체	산업체자문위원	– 개인정보 (객관식 2문항, 주관식 4문항) – 기업정보 (주관식 3문항) – 교육목표 (객관식 4문항) – 학습성과 (객관식 13문항, 주관식 1문항) – 전문교양 (객관식 15문항, 주관식 2문항) – 수학 및 기초과학 (객관식 9문항, 주관식 2문항) – 전공 (객관식 46문항) – 중요한 툴 및 기술 (주관식 1문항) – 학과에 대한 만족도 점수 (주관식 1문항) – 학부교육 발전을 위한 제안사항 (주관식 1문항)

전처리 과정을 거친 후, '학점' 항목을 목표 필드로 하고, 입력필드로는 성별, 나이, 입학유형, 수능성적 등의 30개의 개인 정보 항목을 입력으로 하여 의사결정트리 알고리즘을 수행하였다. 의사결정트리 분석의 결과는 트리 형태의 모델로 출력되며, [그림 7]은 Clementine을 이용한 분석 결과를 트리 모델로 나타낸 것이다. 트리 모델은 노드(Node)와 이음선(Edge)로 구성되어 있으며 각 노드는 레코드(학생)들의 집합을 의미하고, 이음선은 분류 기준을 의미한다. 맨 위에 있는 노드는 루트노드로서 전체 학생들의 집합을 의미하며, 하위 노드들은 상위 노드에 속하는 레코드(학생)들 중에서 이음선의 조건에 해당하는 레코드(학생)들의 집합을 의미한다. 실제 Clementine 분석결과에서는 트리의 형태가 더욱 복잡하기 때문에 식별을 위하여 노드의 해당 레코드(학생)수가 20명 이상인 것만 표시되도록 하였다. [그림 8]은 같은 결과를 텍스트 보기 형식으로 표시한 것으로 깊이가 2인 노드까지 표시되도록 한 그림이다.



[그림 7] 인적사항을 입력으로 한 분류 결과의 그림

[그림 7]과 [그림 8]의 분석 결과를 보면 학생들의 학점에 영향을 미치는 요인으로는 군필여부, 주거지, 학부 홈페이지 접속횟수 등인 것으로 나타났다. 의사결정트리 분류 분석은 해당 요인들이 그림과 같이 계층적인 형태로 학점에 영향을 주기 때문에 루트노드로부터 트리의 경로를 추적하면 규칙을 추출할 수 있다. <표 3>은 추출 가능한 규칙들 중 유용한 몇 가지의 규칙들을 정리한 것이며 노드에 해당하는 레코드의 개수가 너무 적거나 의미가 적은 것들을 제외하였다. 다음은 <표 3>에서 언급된 규칙을 정리한 것이다.



[그림 8] 인적사항을 입력으로 한 분류 결과의 텍스트 보기 그림

<표 3> 학점에 대한 분류 분석을 통해 추출된 규칙 리스트

규칙 번호	규칙 내용	참고
규칙 1	군필여부가 '군필'이고 학부 홈페이지 접속횟수가 1일 '3회 이상' 또는 '1-2회'로 높으면 학점이 상위권에 속하는 비율이 높다.	[그림 7]의 노드5
규칙 2	군필여부가 '미필'인 학생들의 경우 대체로 중위권, 하위권에 속하는 비율이 높다.	[그림 7]의 노드4
규칙 3	군필여부가 '미필'이고 학습시간이 '1일 1시간 미만'인 경우 하위권에 속하는 비율이 높다.	[그림 7]의 노드11
규칙 4	군필여부가 '미필'이고 홈페이지 접속횟수가 '1일 1-2회'이며 관심사가 '학비조달'인 경우 중하위권에 속하는 비율이 높다.	[그림 7]의 노드18
규칙 5	군필여부가 '미입력'(여학생)이고 주거지가 '자취/하숙'인 경우 학점이 중위권, 하위권인 비율이 높다.	[그림 8]
규칙 6	군필여부가 '미입력'(여학생)이고 주거지가 '자택'인 경우 학점이 중위권, 상위권이 비율이 높다.	[그림 8]

(1) '군 미필'인 경우 학점이 하(下)에 속하는 학생들이 많았다. 군입대를 앞둔 경우 학업에 소홀한 경우가 많으며 이것이 학점에 미치는 영향이 매우 큰 것으로 파악된다. 따라서 '군 미필' 학생들에 대한 학점관련 상담지도가 필요한 것으로 사료된다.

(2) '군면제' 학생들의 경우 중위권으로 대체로 학점이 양호한 것을 볼 수 있다.

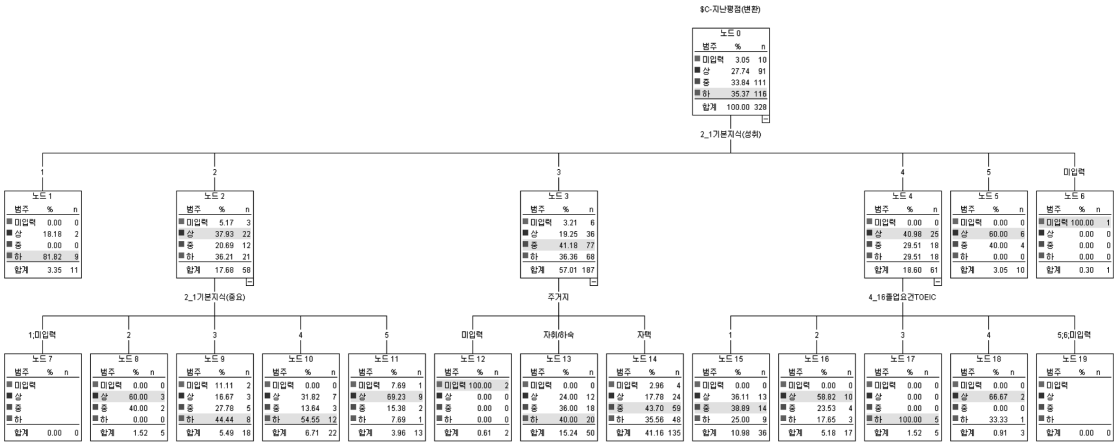
(3) '군필'인 학생들의 경우 홈페이지 이용횟수가 많을수록 학점이 좋은 것으로 파악되었다.

(4) 군필여부에서 '미입력'은 여학생들이며, 여학생들의 경우 주거지가 영향을 많이 주는 것으로

나타났다. 주거지가 '자택'인 경우 학점이 좋은 반면, '자취/하숙'인 경우 학점이 저조함이 파악되었다.

다음으로 개인 정보이외 교육목표, 학습성과, 만족도에 대한 항목들을 입력 항목에 추가하여 의사결정트리 분류 분석을 수행하였다. 이 분석은 설문을 통해 입력 받은 다양한 항목들이 학점에 어떻게 영향을 주는가를 파악하기 위한 분석이다. [그림 9]는 분석 수행 결과 모델을 그래프 보기를 했을 때의 화면이며, [그림 9]에 대한 분석을 통하여 얻을 수 있는 의미있는 규칙들은 <표 4>와 같다.

이 분석에서는 학점에 가장 큰 영향을 미치는 요인은 학습성과 1번 '<기본지식 : 수학, 기초과학, 공학의 지식과 정보기술을 응용할 수 있는 능력>에 대한 성취도'인 것으로 파악되었다. '<기본지식>에 대한 성취도'는 학습성과와 교과목 간의 연관성 매트릭스에 따라 1학년 과목에서 인증필수 과목으로 정해져 있는 MSC(수학, 기초과학) 과목들을 잘 이수하고 이해하고 있는가에 대한 학생 본인의 성취도를 의미한다. 그 외에 '<기본지식>에 대한 중요도 인식정도', '주거지' 등이 영향을 많이 주는 것으로 파악되었다. 이 분석 모델에서는 많은 요인들이 선택되지는 않았지만 여러 학습성과 중 <기본지식>이 가장 중요하게 인식된다는 점이 유의할 만한 점이다. 즉, MSC 과목에 대한 성취도가 높은 학생들이 고학년이 되었을 때도 학점이 양호하고, MSC 과목에 대한 성취도가



[그림 9] 모든 속성을 사용한 분류 결과

<표 4> 학점에 대한 분류 분석을 통해 얻은 규칙 리스트

규칙 번호	규칙 내용	참고
규칙 1	기본지식에 대한 성취도가 '매우낮음'인 학생들의 경우 학점이 하위권에 속하는 비율이 매우 높다.	[그림 9]의 노드1
규칙 2	기본지식에 대한 성취도가 '낮음'이지만 기본지식에 대한 중요도를 '매우높음'으로 한 학생들의 경우 학점이 상위권인 비율이 높다.	[그림 9]의 노드11
규칙 3	기본지식에 대한 성취도가 '높음' 또는 '매우높음'인 학생들의 경우 상위권 및 중위권에 속하는 비율이 매우 높다.	[그림 9]의 노드4, 노드5

스스로 낮다고 여기는 학생들의 경우 학점이 저조하다는 점이다. 따라서 MSC 과목에 대한 중요성을 강조하고 성실한 수강을 유도해야 할 것으로 판단된다. 또한 모델을 자세하게 분석하여 보면, '<기본지식>의 성취도'가 '보통'인 경우 '주거지'에 따른 영향이 있음을 파악할 수 있다. '자택'인 경우 학점이 양호하나 '자취/하숙'인 경우 학점이 낮은 것으로 나타나 '자취/하숙'인 학생들의 경우 상담이 더욱 필요함을 알 수 있다.

<표 3> 및 <표 4>의 규칙들은 학생들에 대한 상담에 활용할 수 있다. 본 연구에서는 이 규칙들을 통하여 학점 상담이 필요한 그룹을 선별하였다. <표 5>와 같이 세분화된 상담 그룹의 형성을 통하여 전체 학생들 중 상담이 필요한 학생들을 파악할 수 있고 각 그룹의 특성에 맞는 상담에 유용한 지침으로 활용될 수 있다.

<표 5> 재학생 분석을 통한 세분화된 상담 그룹

상담 그룹	그룹 특성	중점 상담 사항
세분화 그룹1	• 군필이고 학부 홈페이지 접속이 적은 학생	• 복학 후의 학과 생활 적응 • 관련 정보 습득에 대한 상담 (학과 홈페이지 이용, 교수 상담 요청)
세분화 그룹2	• 군미필인 남학생들	• 군입대 계획 • 군입대 전의 대학생활 상담
세분화 그룹3	• 여학생이고 주거지가 자취/하숙	• 자취/하숙 생활에 대한 상담
세분화 그룹4	• 기본지식에 대한 성취도가 낮은 학생들	• 기초필수 과목에 대한 보충학습 • 개인 학습시간을 확보하도록 상담
세분화 그룹5	• 군미필이고 관심사항이 '학비조달' 인 학생들	• 효과적인 학비마련에 대한 상담 • 학교 및 학과 차원의 장학금 안내

Ⅳ. 결 론

본 논문에서는 데이터마이닝을 이용한 심층적 설문 분석 방법에 대하여 기술하였다. 데이터마이닝 분석을 통하여 기본의 통계적 빈도 분석이나 두 개의 항목에 대한 상관성 분석에서 얻을 수 없었던 의미있는 정보들을 발견할 수 있었다. 이 정보들은 분석에 사용될 뿐 아니라 학생상담 등에 있어서 교육 시스템을 개선하는데 실제적으로 활용될 수 있는 장점이 있다. 본 연구에서는 '학점'에 대한 분류 분석을 통하여 주어진 입력 항목에 따라 각각의 모델을 생성할 수 있었고, 이 모델을 이용하여 상담 등에 활용할 수 있는 규칙을 추출할 수 있었다.

보다 발전적인 데이터마이닝 분석을 위해서는 더욱 다양한 설문을 통하여 풍부한 항목의 데이터를 준비하는 과정이 필요하다. 졸업생의 경우에도 과목에 대한 중요성을 묻는 항목이 있었는데 이들의 직장이나 직종 정보 외에 평점 평균, TOEIC 점수, 입학유형, 고등학교 내신 등의 정보가 추가될 경우 학점이나 대하여 더욱 의미 있는 분석 결과를 얻을 수 있을 것이다. 아울러 학점 뿐만 아니라 취업희망 기업이나 희망 직종에 영향을 미치는 요인 분석 등 추가적인 분석이 지속적으로

요구된다. 또한, 항목이 많은 경우 효율적인 데이터마이닝 분석을 위하여 OLAP 다차원 분석 등과의 상호보완적인 분석도 필요하다.

〔 참 고 문 헌 〕

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

알렉스 버슨 외 지음, 홍성완 외 옮김(2000). CRM을 위한 데이터마이닝. 대청미디어.

장남식, 홍성완, 장재호(1999). 성공적인 지식경영을 위한 핵심 정보기술, 데이터마이닝. 대청미디어.

한경식・이수원(2005). 대용량 데이터를 위한 전역적 범주화를 이용한 결정 트리의 순차적 생성. 한국정보처리학회지 2005년 8월, Vol. 12, 487-498.

허명회・이용구(2004). 데이터마이닝 모델링과 사례, 데이터솔루션.

허준 외(2003). Clementine 7 매뉴얼. 데이터솔루션.

Richard J.Roiger & Michael W. Geatz(2003). *Data Mining: A Tutorial-Based Primer.* Addison-Wesley.

Ian H. Witten & Eibe Frank(2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann.

Jiawei Han & Micheline Kamber(2006). *Data Mining Concepts and Techniques.* Morgan Kaufmann.