데이터마이닝 기법을 이용한 효과적인 연구관리에 관한 연구

황석해(한국외국어대학교 경영학과 박사과정)
문태수(동국대학교 상경대학 정보산업학과)
이준한(경주대학교 경영광고학부 경영정보학과)

seokhae10@hotmail.com
tsmoon@mail.dongguk.ac.kr
leejh@tour.kyongju.ac.kr

요 약 본 연구는 R사의 대고객 만족도 향상을 위하여 고객관계관리(customer relationship management, CRM)를 수행하기 위한 목적으로 추진되었다. 연구의 주안점은 연구관리 데이터베이스로부터 연구관련 변수들의 패턴 및 상호작용을 고려하여 연구계약기관을 그룹내부기관과 외부기관으로 분류함으로써 기관별 연구과제의 연구유형 및 연구비에 대한 분석을 통하여 향후 대고객관리의 방향을 설정하기 위한 목적으로 시도되었다.

Ⅰ. 서론

기업들은 다양한 환경변화 요인에 의해 도전을 받고 있으며, 고객의 요구에 부응하기 위해 많은 노력을 기울이고 있다. 또한 고객관리에 대한 관심과 데이터의 양이 많아짐에 따라 이를 위한 효율적인 관리에 투자를 증대시키고 있다. 데이터마이닝(data mining)은 방대한 양의 데이터로부터 의미있는 패턴, 규칙들을 발견하기 위하여 자동적인 혹은 반자동적인 방법으로 데이터를 분석하고 탐색하는 것을 말한다.

향상된 데이터 분석의 궁극적인 목적은 가능한 한 직접적인 이익을 얻을 수 있는 의사결정을 하는 것이며, 데이터분석을 통하여 "데이터에서 정보로, 정보에서 지식으로, 지식에서 의사결정"에 이르는 가치사슬 경로를 설명하는 것이다. 여기서 데이터에서 의사결정에 까지 이르는 프로세스에 이용되는 주요 정보기술로서는 데이터베이스, 정보기술 인프라, 데이터마이닝 등을 들 수 있다.

본 연구는 R사의 대고객 만족도 향상을 위하여 고객관계관리(customer relationship management, CRM)를 수행하기 위한 목적으로 추진되었다. 연구의 주안점은 연구관리 데이터베이스로부터 관련 변수들간의 패턴 및 상호작용을 고려하여 연구과제를 수행하였을 경우 연구계약기관을 그룹내부기관과 외부기관으로 분류하여 이들을 분류하는 유의변수를 규명하여 연구업무의 성격에 따른 계약기관 유형을 파악하여 연구계약기관 선정시 연구특성에 따라 연구기간과 연구비 금액의 책정에 대한 효과적인 평가를 지원하는데 목적을 두고 있다.

Ⅱ. 데이터마이닝과 기법

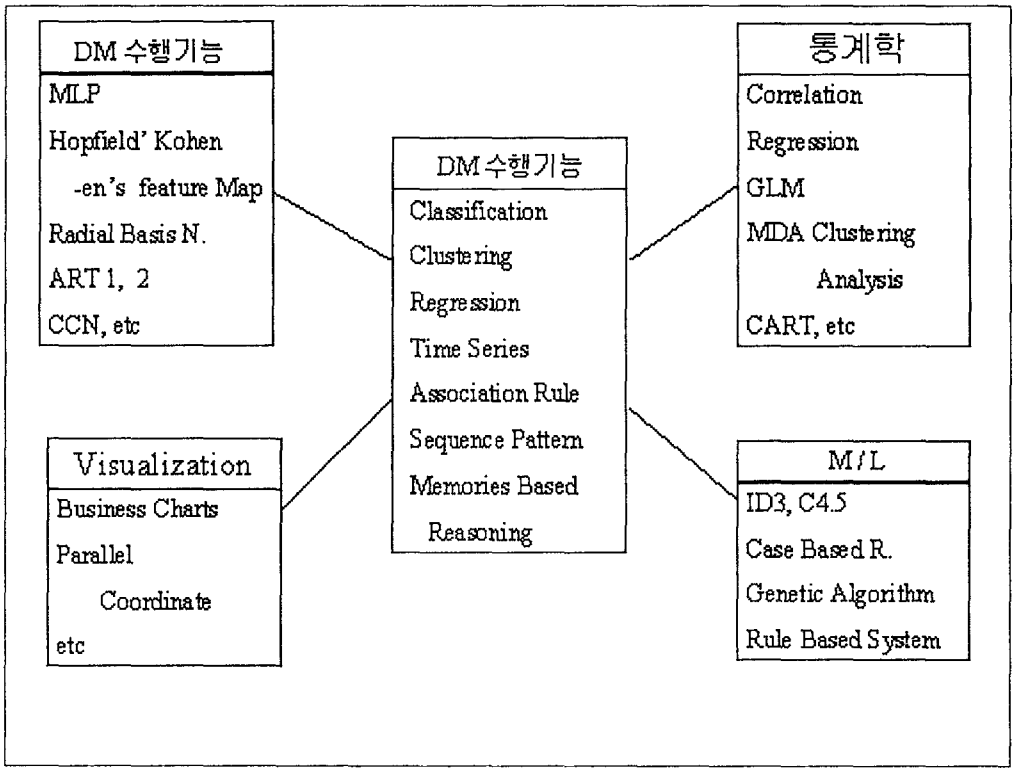
데이터 마이닝(data mining technique)은 데이터로부터 다양한 형태의 유용한 정보를 추출하기 위하여 모델링 방법을 적용하거나 관측된 패턴의 유용성을 판단하는 일련의 과정이다. 최근 데이터 마이닝의 중요성이 강조되는 이유는 다음과 같다. 첫째, 방대한 데이터베이스 속에 축적된 많은 양의 데이터를 보다 효율적으로 이용한다. 둘째, 데이터 마이닝 알고리즘의 발달과 컴퓨터 용량의 양적 증가 및 성능의 질적 향상으로 복잡한 형태를 가진 데이터의 처리과정을 보다 쉽게 처리할 수 있게 하여 원하는 정보를 얻을 수 있는 환경을 제공한다. 셋째, 대용량 데이터베이스로부터 얻어진 정보를 의사결정에 활용함으로써 시장에서의 경쟁적인 우위를 차지한 사례들이 늘고 있다. 넷째, 데이터 마이닝 기법은 기존의 전문가 시스템이 갖는 한계점인 지식획득의 병목현상을 극복하는데 효과적이다.

데이터로부터 유용한 패턴 또는 지식을 발견하는 작업, 즉 데이터 마이닝 과정은 거론되는 분야에 따라 지식 추출(knowledge extraction), 정보발견(information discovery), 데이터 연금술(data archeology), 정보 수확(information harvesting), KDD(Knowledge Discovery in Database) 등의 서로 다른 이름으로 혼용되고 있다. 데이터 마이닝이라는 용어는 주로 통계학자, 데이터베이스 연구개발자, MIS 연구자들에 의해 사용되고 있다. 데이터 마이닝은 데이터베이스, 기계학습(machine learning), 패턴 인식(pattern recognition), 통계학, 인공지능, 전문가시스템, 데이터가시화(data visualization), 정보조회(information retrieval) 등의 다양한 분야로부터 발달해 왔다. 따라서, 데이터 마이닝 시스템이란 다양한 분야의 이론 및 알고리즘들을 통합하여 데이터로부터 유용한 지식을 추출해내는 총체적인 시스템을 의미한다[Fayyad, et al., 1996; Hogh, 1996; Adriaans & Zantinge, 1997].

데이터 마이닝은 사용되는 목적에 따라 "검증(verification)"과 "발견(discovery)"의 두 가지 측면으로 분류해 볼 수 있다. 검증을 위한 데이터 마이닝은 사용자가 세운 가설을 증명하기 위해 데이터로부터 정보를 추출하는 것이 목적이며, 발견을 위한 데이터 마이닝은 데이터로부터 새롭고 유용한 패턴을 추출하여 사용자에게 제시하는 것이 목적이다. 발견을 위한 데이터 마이닝의 기법들은 사용형태에 따라 "예측(prediction)"과 "설명(description)"의 두 형태로 나뉘어진다. 예측 형태에서는 관심 있는 요인들간의 작용을 예측하기 위하여 데이터로부터 관련 패턴을 찾고, 설명 형태에서는 특정 사실을 사용자에게 좀 더 쉽게 이해시키기 위하여 관련 패턴들을 찾는다. 그러나 대개의 경우, 패턴 발견에 대한 두 가지 형태는 서로 결합되어 이용되는 것이 일반적이다. 예측을 위해서는 회귀분석(regression), 시계열 분석(time series analysis), 분류(classification)분석이 주로 이용되고 있으며, 설명을 위해서는 군집화(clustering), 연관규칙(association rule) 및 순차패턴(sequence pattern)탐사, 요약기법(summarization), 가시화 기법(visualization), 변화와 편차의 탐지 등이 이용된다[Ronald, et al., 1996]. 데이터 마이닝에 의해 수행될 수 있는 대표적 기능들과 기능수행에 사용될 수 잇는 분석도구(tools)들을 정리하면 [그림 1]과 같다.

데이터마이닝의 각 기능을 수행할 수 있는 분석도구들은 여러 학문 분야들로부터 발전되어 왔다. 즉 분류분석을 위해서는 통계학의 다변량 판별분석(multiple discriminant analysis), 신경망의 다계층 퍼셉트론(multi-layered perceptron) 및 기계학습의 ID3 또는 C4.5 등이 사용될 수 있다. 하지만 각 학문 분야가 데이터 마이닝의 모든 기능을 지원하는 것은 아니다. 정보활용방법의 발전에 대한 데이터베이스 학계의 최근 공헌은 데이터웨어하우스 및 OLAP(online analytical processing)기술의 개발 외에도 연관성 분석(affinity analysis)라는 데이터 마이닝의 기술을 개발한 것이다. IBM Almaden연구소의 Agrawal, et al.[1993]에 의해 처음 시도된 연관규칙 탐사기술의 개발은 단순한 개념에서 출발하였지만 대용량의 데이터베이스로부터 유용한 지식을 찾아낸다

는 점에서 실용성이 매우 큰 기술로 장바구니분석(market basket analysis)에서 위력을 발휘하였으며 최근에는 순차패턴(sequence pattern)의 탐색기술도 개발되고 있다[Agrawal & Srikant, 1995].



[그림 1] 데이터마이닝 수행기능과 도구들

[그림 1]에는 표현되어 있지 않지만, 최근의 기술적 동향과 관련하여 지적할 중요한 사항은 데이터마이닝과 인터넷과의 관계이다. 현재 데이터마이닝에서 인터넷 관련기술을 활용하는 방안은 client/server 구조하에서 데이터마이닝 엔진을 서버에 두고 client들이 intranet을 통하여 실행하는 것이다. 이러한 접근방식은 client/sever체제 구축에 따른 비용절감 효과 및 데이터마이닝의 결과를 조직내에 폭넓게 확산시킬 수 있다는 장점이 있다. 앞으로 World Wide Web(WWW)은 다양한 정보의 원천이므로 데이터마이닝의 좋은 적용대상이 될 것이다. 현재 대상 정보들이 너무도 다양한 형식에 의해 표현되어 있으므로 WWW상에서의 직접적인 데이터마이닝의 적용에는 아직 한계가 있다. 하지만 최근 원천정보를 가진 데이터베이스에의 직접 접근이 가능한 검색엔진의 개발 및 인터넷 문서의 표현양식 표준화 및 발전은 가까운 장래에 데이터마이닝의 적용을 가능하게 할 것이다. 특히 인터넷의 응용기술과 관련하여 인공지능의 에이전트 기술이 많이 활용되고 있는 점도 데이터마이닝과 관련하여 주목하여야 한다.

III. 의사결정트리기법

본 연구에서 사용하고 있는 알고리즘인 의사결정트리는 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다[Mehta et al., 1996]. 분류에 관한 연구는 과거 통계(Statistics), 인공신경망(Neural Network), 의사결정트리(Decision

Tree)등의 분야에서 연구되어 왔다. 의사결정트리는 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단하며, 이해하기 쉬운 규칙으로 전환될 수 있기 때문에 본 논문은 데이터마이닝 기법으로서 의사결정트리 가운데 하나인 CHAID(Chi-Square Automatic Interaction Detection) 알고리즘을 사용하고 있다[Imielinski and Mannila, 1996].

1975년 J.A. Hatigan에 ,의해 처음 발표된 CHAID알고리즘은 카이제곱-검정(이산형 목표변수), 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(Multiway Split)를 수행하는 알고리즘으로 1963년 J.A. Morgan과 J.N. Sonquist이 발표한 AID(Automatic Interaction Detection)시스템에서 유래되었다. AID에서 암시하고 있는 것과 같이 CHAID는 원래 변수들 간의 통계적 관계를 찾는 것이 목적이었다. 변수들간의 통계적인 관계는 다시 의사결정트리를 통해 표현될 수 있었으므로 이 방법은 분류기법(Classification Technique)으로 사용할 수 있다[Thearling, 1995].

CHAID는 변수의 성격이 범주형 데이터이고 예측변수와 결과변수간의 관계를 찾아야 할 때 가장 유용하다[Pyle, 1998]. 다른 의사결정트리와 마찬가지로 CHAID알고리즘은 두 개 이상의 자식노드(Child Node)로 트레이닝 데이터를 쪼개기 위한 입력변수를 찾는다. 즉 CHAID는 분리기준(Split)을 찾는 것을 시발점으로 하여 자식노드는 특정 변수가 갖고 있는 결과변수의 확률이 각 노드마다 다르게 하는 방식으로 선택된다. CHAID는 데이터의 집합을 검색하여 예측변수의 예측치로서 가장 유의성이 높은 변수를 결정한다. CHAID알고리즘은 카이제곱 통계량을 통해 비율이 유지되는 정도를 파악하는데, 여러 변수 중 비율을 가장 많이 깨뜨리는 변수가 결국 결과변수에 영향을 가장 많이 미치는 변수가 된다. 비율이 깨진 정도는 카이제곱에서 r × c 분할표(Contingency Table)로 계산된다. 이 때 Pearson의 카이제곱 통계량은 다음과 같다.

$$x^2 = \sum \frac{(fo - fe)^2}{fe}$$

fo : 관찰치, fe : 예측치

이 통계량은 자유도가 (r-1)(c-1)인 카이제곱 분포를 따른다. 카이제곱 통계량이 자유도에 비해 매우 작다는 것은 입력변수의 각 범주에 따른 결과변수의 분포가 동질적이라는 것을 의미하여, 입력변수가 결과변수의 분류에 영향을 주지 않는다고 말할 수 있다. 자유도에 대한 카이제곱 통계량의 크고 작음은 p-값으로 표현될 수 있는데, 카이제곱 통계량이 자유도에 비해서 작으면 p-값은 커지게 된다. 결국 노드는 p-값이 가장 작은 변수를 기준으로 가지가 형성되는 것이다[이성근 등, 1996]

Ⅳ. 의사결정트리기법을 이용한 분류모델의 개발

1. R사의 연구관리시스템

R사는 1987년에 설립된 P그룹사의 산하 연구기관이다. R사의 주요 연구분야는 철강, 신소재, 에너지, 환경, 전기전자 등이며, 연간 250여건의 연구과제를 수행하고 있다. R사는 최근 P 그룹사의 의존도를 줄이면서 자생력을 키우기 위한 장기적인 전략을 수립 중에 있다. 또한 최근에는 정부 및 관련산업 기업체들의 연구개발에 대한 요구가 증가하면서 대고객관계관리에 대한 필요성이 증가하고 있음을 파악하고 이에 대한 대응책을 수립하고자 노력하고 있다.

R사의 연구관리시스템은 연구과제의 발굴, 연구계약, 연구진행관리, 연구결과보고, 연구사후관리 등의 단계로 연구개발활동을 관리하고 있으며, 그 중에서도 연구사후관리를 통하여 향후 발생할

연구계약금액을 예측하여 다음 연도의 R사 연구계획에 반영하고 있다. 연구개발활동(R&D)은 경기의 변동에 민감하여 미래에 대한 예측이 불분명하며, 연구원들의 독특한 업무특성에 따라 고객에 대한 관계관리가 쉽지 않은 편이다.

그리하여 R사는 연구과제의 기관별 유형에 따라 연구관리의 방향을 설정하고 고객관계관리 및 대고객 만족도 제고를 위한 방안 마련에 부심하고 있다. 본 연구도 이의 일환으로 R사의 과거 연구수행실적을 바탕으로 기관별 연구과제의 연구유형 및 연구비에 대한 분석을 통하여 향후 대고객관리의 방향을 설정하기 위한 목적으로 시도되었다.

2. R사의 데이터 생성 및 사전처리

본 연구는 연구과제의 계약기관 유형을 예측하기 위한 분류모델(classification)을 개발하기 위하여 데이터마이닝 기법 가운데 하나인 CHAID트리 생성 알고리즘을 사용하였다. 의사결정 나무의 분할기준으로는 카이제곱 통계량에 의한 지니지수, 엔트로피 기준이 있는데, 세가지 기준은 큰 예측력의 차이를 보이지 않는다. 세 가지 기준 중 가장 낮은 오분류율을 가진 카이제곱 통계량을 나무의 분할기준으로 활용하였다. 두 개 이상의 마디에 대해서 평균의 차이를 검정하는 F통계량을 분리기준으로 설정하였으며 유의수준은 0.20으로 하였다.

각 마디에서의 불순도(impurity)를 재는 특도인 분사의 감소량을 분리기준으로 설정하였으며 분산의 상대적 감소량은 0으로 설정하였다. 입력변수로는 재무비율변수로 47개의 변수를 입력변수로 사용하였다. 사용된 데이터는 훈련용(training)으로 40%, 검증용(validation)으로 30%, 시험용(test)으로 30%로 구분하여 할당하였다

사용된 변수는 연구기간, 인건비, 장비사용료, 컴퓨터사용료, 감가상각비, 재료비, 장치제작비, 보고서 인쇄비, 국내여비, 국외여비, 외주시험비, 자료수집비, 연구자문비, 위탁과제, 기타경비, 제경비, 기술개발비를 입력변수로 선정하였으며, 연구계약기관을 목표변수로 선정하여 그룹내부기관과 외부기관으로 구분하였다.

3. 분류모델의 개발

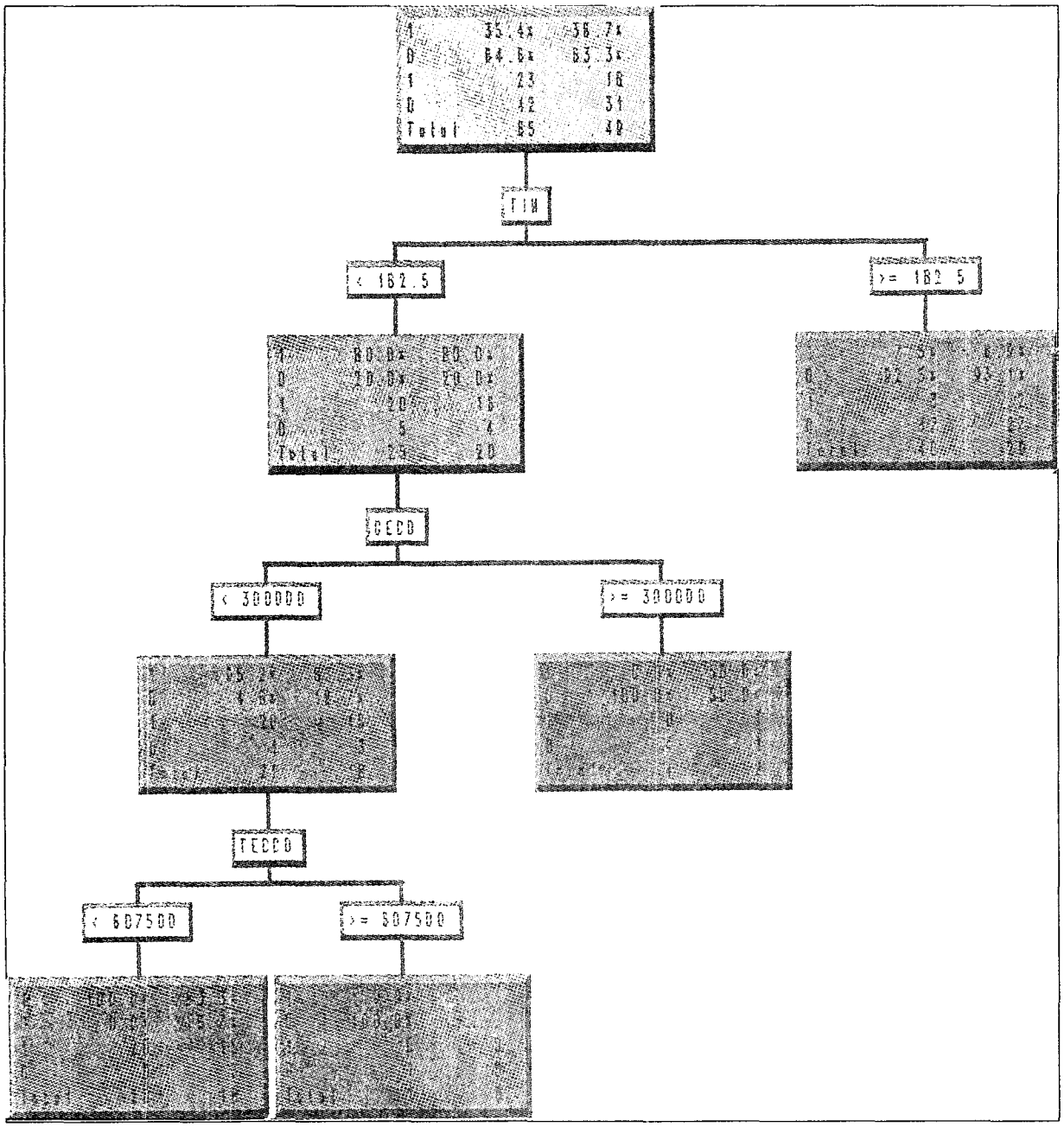
[그림 2]는 계약기관 변수에 의해서 형성된 의사결정나무 결과이다. 의사결정나무는 뿌리마디(root)에서 시작하여 가지(node)를 형성해 나간다. 노드 1에서 연구기간(TIM)이 가장 유효한 변수로 사용되었으며 노드 2에서는 제경비(GECO), 노드 3에서는 기술개발비(TECCO)가 사용되었다. [그림 2]의 의사결정 노드에는 분석용 표본의 결과와 평가용 표본의 결과로 구분하여 나타나 있다. 평가용 표본에 대한 분석결과로는 뿌리마디(root node)의 총 49개의 관찰치는 그룹내부기관과 외부기관이 각각 18, 31개로 구성되어 있다.

먼저 연구기간(TIM)에 의하여 의사결정나무가 형성되며 연구기간(TIM)의 관측값 182.5를 기준으로 두 개의 하위 노드(sub node)로 구분된다. 총 49개의 관찰치중 29개는 연구기간이 182.5보다 크며 93.1%인 27개가 외부계약 계약기관이며 연구기간이 182.5미만인 노드는 20개의 계약기관으로서 그룹내부기관은 80%로서 16개의 계약기관이다. 의사결정 나무는 이처럼 선택된 결정변수의 기준값을 이용하여 하위노드로 분할하고 이들 하위노드 중에서 다시 표본을 가장 잘 구분할 수 있는 변수를 선정하여 또 다른 하위노드를 형성해 나가면서 자신은 부모노드(parent node)가 되는 과정(recursive partitioning)을 가진다.

계약기간(TIM)이 182.5미만인 하위 노드 1은 제경비(GECO)에 의해 부모노드가 됨과 동시에 또 다른 하위노드를 형성한다. 연구기간(TIM)이 182.5미만이면서 제경비(GECO)가 300000미만인 기업은 그룹내부기관이 83.3%에 해당하는 15개이며 외부기관은 16.7%인 3개가 해당된다. 연구기간

(TIM)이 182.5%미만이면서 제경비(GECO)가 300000이상인 기업은 계약기관인 그룹내부기관과 외부기관에 각각 1개의 기업이 나타났으며, 이러한 하위 노드는 더 이상 분할되지 않으면 터미널 노드(terminal node)가 되면서 잎(leaf)으로 남게 된다.

계약기간이 182.5미만이면서 제경비(GECO)가 300000미만이고 기술개발비(TECCO)가 607500미만인 계약기관은 그룹내부기관이 83.3%인 15개이며 외부기관이 16.7%인 3개이며 기술개발비(TECCO)가 607500 이상인 계약기관은 없는 것으로 나타났다.



[그림 2] 계약기관 유형 의사결정수

이와 같은 과정을 거치면서 의사결정나무가 형성되는데 원래의 의사결정나무는 이보다 큰 나무이었으나 가지치기를 하여 만든 의사결정나무가 [그림 2]의 결과이다.

4. 분류모델의 검증

사용변수명		분석지표	중요성(worth)
연구유형	연구기간	TIM	1.0000
	제경비	GECO	0.4072
	기술개발비	TECCO	0.3425

[표 1] 변수의 가치

연구기관유형 의사결정트리에서 선택된 변수는 총 3개의 변수인데 이들의 중요성(worth)을 보면 위의 [표 1]과 같다. 의사결정나무의 뿌리마디를 분류하는 변수가 가장 중요한 가치를 가지고 있는 변수이다. 1차적으로 이 변수의 가치로 1.00을 부여하면서 하위노드를 형성하고 2차적으로 분할되는 변수에 두 번째 가치를 부여한다. 의사결정나무는 연구기관(TIM)을 연구모형에서 그룹내부기관과 외부기관 분류하는데 있어 가장 중요한 역할을 하고 있는 변수로 선정하였다. 그 다음으로 제경비(GECO)에 0.4072라는 가치를 부여하여 나무를 형성하였다. 마지막으로 기술개발비(TECCO)에 0.3425라는 가치를 부여하였다.

분석용 자료에서 1(그룹내부기관)을 0(외부기관)으로 정확하게 예측한 개수가 32이고 1(그룹내부기관)을 0(외부기관)으로 잘못 예측한 계약이 1개이며, 0(외부기관)을 0(외부기관)으로 정확하게 예측한 개수가 26이며 0(외부기관)을 1(그룹내부기관)로 잘못 예측한 개수가 2개 계약으로 나타나 있다.

노드구성	분류정확도	
	훈련용	검증용
노드 1	0.6462	0.6327
노드 2	0.8769	0.8776
노드 3	0.9385	0.8776
노드 4	0.9538	0.8776

[표 2] 의사결정나무 하위노드 각 노드값

분류기준값 0.5의 경우 노드 2에서 급격한 향상을 나타내며 노드 3에서부터 점진적인 향상을 이루고 있지만 분류기준값의 변동에 따른 분류정확도는 큰 차이를 나타내고 있지는 않다.

표본구분	통계량	목표변수	1 (그룹내부기관)	0 (외부기관)	TOTAL
훈련용	N	1	20	3	23
훈련용	N	0	0	42	42
훈련용	N	+	20	45	65
훈련용	행%	1	87	13	100
훈련용	행%	0	0	100	100
훈련용	행%	+	31	69	100
훈련용	열%	1	100	7	35
훈련용	열%	0	0	93	65
훈련용	열%	+	100	100	100
훈련용	%	1	31	5	35
훈련용	%	0	0	65	65
훈련용	%	+	31	69	100
분석용	N	1	15	3	18
분석용	N	0	3	28	31
분석용	N	+	18	31	49
분석용	행%	1	83	17	100
분석용	행%	0	10	90	100
분석용	행%	+	37	63	100
분석용	열%	1	83	10	37
분석용	열%	0	17	90	63
분석용	열%	+	100	100	100
분석용	%	1	31	6	37
분석용	%	0	6	57	63
분석용	%	+	37	63	100

*) N : 관측개수, % : 전체퍼센트, + : 목표변수의 총계

[표 3] 이산형 목표변수에 대한 평가행렬

구분			평가용		분석용	
잎(leaf) 노드	관측 개수	평가용 수	내부계약 기관 (%)	외부기관 (%)	내부계약 기관 (%)	외부기관 (%)
8	20	18	83.33	16.67	100.00	0.00
6	1	0	0.00	0.00	0.00	100.00
5	4	2	50.00	50.00	0.00	100.00
6	10	3	0.00	100.00	30.00	70.00
7	30	26	7.69	92.31	0.00	100.00

[표 4] 잎(leaf) 노드 통계량

최종 잎(Leaf)에서 20개의 관측치 중 18개의 평가대상과제가 사용되었으며 평가용 자료에서의 그룹내부기관과 외부기관은 각각 83.33%, 16.675%이고 분석용에서는 100%, 0.00%로 나타났다

구 분	사용표본		
적합통계량	Training	Validation	Test
Average Squared Error	0.043	0.1197	0.1870
Sum of Squared Errors	5.550	11.7263	18.3275
Root Average Squared Error	0.207	0.3459	0.4325
Maximum Absolute Error	0.925	1.0000	1.0000
Divisor for ASE	130.000	98.0000	98.0000
Total Degrees of Freedom	65.00	.	.
Frequency of Classified Cases	65.000	49.0000	49.0000
Misclassification Rate	0.046	0.1224	0.2041
Number of Estimated Weights	4.000	.	.
Sum of Frequencies	65.000	49.000	4900
Sum Case Weights * Frequencies	130.000	98.0000	98.0000

[표 5] 의사결정나무 적합통계량

[표 5]는 계약기관 유형변수를 사용한 의사결정나무모형의 적합통계량으로서 분석에 사용된 자료는 훈련용(training)이 65개, 분석용(validation)이 49개, 평가용(test)이 49개로 사용되었으며 오분류율은 훈련용이 4.6%, 분석용이 12.24%, 평가용이 20.41%로 나타났다.

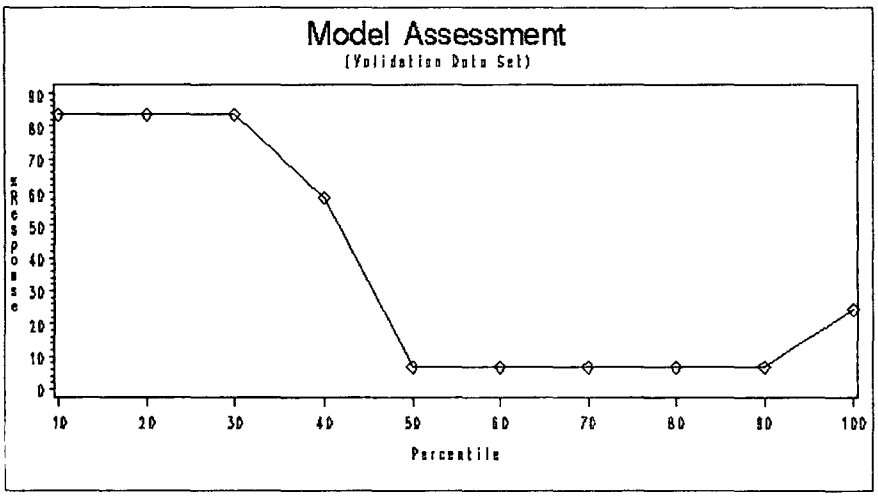
의사결정나무는 뿌리마디에서 출발하여 가지를 형성하며 최종 잎을 가진 것으로 형성된다. 노드 4에서 더 이상 개선의 효과가 나타나지 않으며, 훈련용의 분류정확률은 95.40%를 나타낸다. [표

2]에서 노드 1과 2에서 현격한 분류정확률 향상을 나타내며 노드 4에 이르기까지 점진적인 향상을 보이고 있다. 노드 4에서 최적의 의사결정 분류기준을 나타내는 것으로 나타났다.

빈도수 백분율 행 백분율 열 백분율		예측		전체
		외부기관	그룹내부기관	
실제	외부기관	28 57.14 90.32 90.32	3 6.12 9.68 16.67	31 63.27
	그룹내부기관	3 6.12 16.67 9.68	15 30.61 83.33 83.33	18 36.73
전체		31 63.27	18 36.73	49 100

[표 6] 의사결정트리 정오분류표

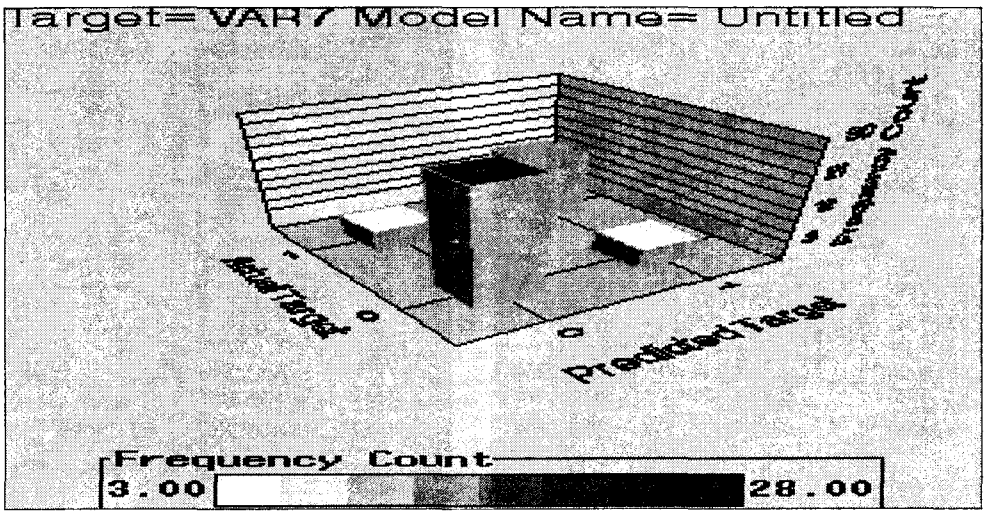
분석에 사용된 전체 표본수는 49개이며 이중 실제 그룹내부기관과 외부기관 표본은 각각 18개(36.73%), 31개(63.27%)가 사용되었다. 정오분류표에서 외부기관을 외부기관으로 예측한 정확률이 90.32%이며 외부기관을 그룹내부기관으로 잘못 예측한 오류율이 16.67%로 나타났다. 또한 그룹내부기관을 그룹내부기관으로 예측한 정확률이 83.33%이며 그룹내부기관을 외부기관으로 잘못 예측한 오류율은 9.68%로 나타났다. 전체 관찰치중에서 정분류율(정확도)는 관찰치가 (28+15)/49로써 87.75%이며 오분류율은 관찰치가 (3+3)/49로서 12.24%로 나타났다.



[그림 3] 의사결정트리의 누적빈도수를 이용한 %Response

5. 분류성과 측정

[그림 3]에서의 %Response는 각 집단 내에서 범주 1의 빈도와 집단 내 관찰치의 빈도의 비를 나타내므로, 첫 번째 집단에서 %Response는 (범주 1의 빈도)/(집단 1의 관찰치의 빈도값이 된다. 즉, 목표변수 내에서 범주 1의 점유율을 각 집단에 대해 구한 값이라고 해석할 수 있다. 상위 50%의 집단에서 높은 점유율을 나타내고 있다.



[그림 4] 의사결정나무의 정오분류행렬

[그림 4]에서는 실제관찰치와 예측치의 개수를 3차원에서 나타내고 있다. 외부기관의 정분류가 가장 높아 정확률이 뛰어나며, 빈도수에서도 가장 높은 빈도를 보이고 있다. 그룹내부기관도 높은 정확률을 나타내어 빈도수에서도 외부기관의 정분류 다음으로 높은 것을 알 수 있다.

V. 결론

본 연구는 R사의 대고객 만족도 향상을 위하여 고객관계관리(customer relationship management, CRM)를 수행하기 위한 목적으로 추진되었다. 연구의 주안점은 연구관리 데이터베이스로부터 연구관련 변수들의 패턴 및 상호작용을 고려하여 연구계약기관을 그룹내부기관과 외부기관으로 분류함으로써 기관별 연구과제의 연구유형 및 연구비에 대한 분석을 통하여 향후 대고객관리의 방향을 설정하기 위한 목적으로 시도되었다.

연구기관유형 의사결정트리에서 선택된 변수는 총 3개의 변수로 연구기간, 제경비, 기술개발비 등이 선정되었으며, 연구과제관리의 의사결정나무 뿌리마디를 분류하는 가장 중요한 가치를 가진 변수들이다. 의사결정나무모형에서 계약기관 유형에 사용된 자료는 훈련용(training)이 65개, 분석용(validation)이 49개, 평가용(test)이 49개이며 오분류율은 훈련용이 4.6%, 분석용이 12.24%, 평가용이 20.41%로 나타났다. 의사결정나무는 뿌리마디에서 출발하여 가지를 형성하며 최종 잎을 가진 것으로 형성되었으며, 노드 4에서 더 이상 개선의 효과가 나타나지 않아 훈련용의 분류정확률은 95.40%를 나타내었다.

분석결과, 실제 그룹내부기관과 외부기관 표본은 각각 18개(36.73%), 31개(63.27%)가 사용되었으며, 정오분류표에서 외부기관을 외부기관으로 예측한 정확률이 90.32%, 오류율이 16.67%로 나타났다. 또한 그룹내부기관을 그룹내부기관으로 예측한 정확률이 83.33%, 잘못 예측한 오류율은 9.68%

로 나타났다. 이 결과는 전체 관찰치중에서 정분류율(정확도) 관찰치가 (28+15)/49로써 87.75%이며 오분류율은 관찰치가 (3+3)/49로서 12.24%인 것으로 나타났다. 이 결과는 외부기관의 정분류가 가장 높은 정확률을 보이고 있으며, 빈도수에서도 가장 높은 빈도를 보인 것이다. 그리고 그룹내 부기관도 높은 정확률을 나타내어 빈도수에서도 외부기관의 정분류 다음으로 높은 결과를 보였다.

참 고 문 헌

Adrianns. P. & D. Zantinge, Data Mining, Addison-Weslery press, 1997.

Agrawal, Rakesh, Ashish Gupta, Sunita Sarawagi, "Research Report : Modeling Multidimensional Databases," IBM Almaden Research Center.

Agrawal, R., T. Lmielniski & A. Swami, "Mining Association Rules in Large Database", Proc. of ACM SIGMOD Conf. on Management of Data. Washington D.C., 1993, pp. 207-216.

Agarawal, R & R. Srikant, "Mining Sequential Patterns," Proc. of 11th Int.'.Cong. on Data Engineering, Taipei, Taiwan, march, 1995.

B. de. lalesia, J.C.W. Debuse and V.J. Rayward-Smith, "Discovering Knowledge in Commercial Database Using Modern Heuristic Techniques, "KDD-96, 1996.

Brachman, Ronald J. Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis, "Mining Business Databases," Communications of the ACM, November, Vol 39, No. 1996

Brachman, Ronald J., Tej Anand, "The Process of Knowledge Discovery in Databases : A First Sketch, "AAAIi-94 Workshop on Knowledge Discovery in Databases, KDD-94, 1994.

Fayyad, Usama, "Diving into Databases." Database Programming & Design, March, 1998.

Fayyad, U., G. Piatetsky-Shapiro, & P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM. Vol. 39, No. 11, 1996. pp. 27-34

Glymour, Clark, David Madigan, Daryl Pregibon, and Padhraec Smyth, "Statistecal Inference and Data Mining, "Communications of the ACM, Vol. 39, No. 11, 1996.

Gupta, S., "Impact of Sales Promotions on When, What and How Much to Buy," journal of Marketing Research, Vol. 25, 1988. pp. 342-355.

Hogarth, R.M. and S. Makridakis, "Forecasting and Planning: An Evaluation," Management Science, Vol. 27, 1981, pp. 115-138.

Hong, S. J., Data Mining for Decision Support, Working Paper, IBM Watson Research Center, 1996.

Mannila, Heikki, Department of Computer Science University of Helsinki. "Methods and Problems in data mining".

Mehta, Manish, Jorma Rissanen, Rissanen, Rakesh Agrawal, "MDL-based Decision Tree Pruning," IBM, Almadesn Research Center, mmehta, rssanen, a grawal@almaden.ibm.com.

Parsaye, Kamarn, "OLAP & Data Mining : Bridging the Gap." Database Programming & Design. Februay 1998.