정보기술응용연구
제 3 권 제 2 호
2 0 0 1 년 6 월

데이터마이닝 분류기법을 이용한 효과적인 연구관리에 관한 연구

황석해* , 문태수** , 이준한***

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 기관별 연구과제의 연구유형 및 연구비에 대한 분석을 통하여 고객유형별로 분류함으로써 향후 대고객관리의 방향을 설정하기 위한 목적으로 시도되었다.

본 연구에서 의사결정나무 알고리듬을 이용하여 자료를 분석한 결과, 17개의 입력변수 중 내외부 계약기관을 분류하는 데 있어서 중요한 변수로는 연구기간, 제경비, 기술개발비의 3개 변수로 나타났다. 연구결과, R사의 고객은 6개월 이상의 연구기간, 3,000만원 이상의 제경비, 그리고 6,075만원의 기술개발비를 기준으로 연구계약기관이 분류되며, 이 연구관련 변수를 이용하여 대고객과의 연구주제 설정, 연구예산 수립 등의 고객관리방안을 수립할 수 있을 것이다.

*) 한국외국어대학교 경영학과 경영학박사
**) 동국대학교 상경대학 정보산업학과 조교수
***) 경주대학교 경영학부 경영정보전공 조교수

1. 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 의미 있는 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 의사결정에까지 이르는 프로세스에 이용되는 주요 정보기술로서 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 규명하고 연구업무의 성격에 따른 계약기관 유형을 파악하여 연구계약기관 선정시 연구특성에 따라 연구기간과 연구비 금액의 책정에 대한 효과적인 평가를 지원하는데 목적을 두고 있다.

2. 고객관계관리와 데이터마이닝

2.1 고객관계관리

상위 20%의 고객이 기업수익의 80%를 기여한다는 파레토 법칙이 시사하듯이 기업에서 우수고객을 확보하는 것은 기업의 계속적인 성장발전을 위해 매우 중요하다. 우수고객의 확보는 의욕만을 가지고 달성할 수 있는 과제는 아니다. 정확한 고객정보를 바탕으로 하여 세분화된 기준으로 분류된 고객의 욕구를 정확히 파악하여 서비스를 제공할 때만이 수익성이 높은 우수고객 확보가 가능한 것이다. 이렇듯 고객에 대한 광범위하고 심층적인 지식을 바탕으로 고객 개개인에게 적합한 차별적인 제품이나 서비스를 제공함으로써 고객과의 접점관리(moment of truth)를 통해 지속적 관계로 강화해 나가는 마케팅 기법이 고객관계관리이다.

기존의 마케팅이 주로 신규고객의 창조에 중점을 둔 반면, 고객관계관리는 신규고객의 창조보다는 오히려 기존고객으로부터 수익기반 유지 및 확대를 위하여 기존고객과의 우호적인 관계구축을 목표로 마케팅 활동을 전개한다. 이러한 고객관계관리가 등장하게 된 배경은 무엇보다도 신규고객의 획득에 비해 기존고객을 유지하는데 투입되는 비용이 적기 때문이다. 관계마케팅의 연구결과에 의하면 기존고객의 유지비용보다 신규고객의 획득비용이 5배 이상 높고, 고객과의 긴밀한 관계를 장기에 걸쳐 유지하는 기업의 이익증대 효과도 높다고 알려져 있다.

2.2 데이터마이닝

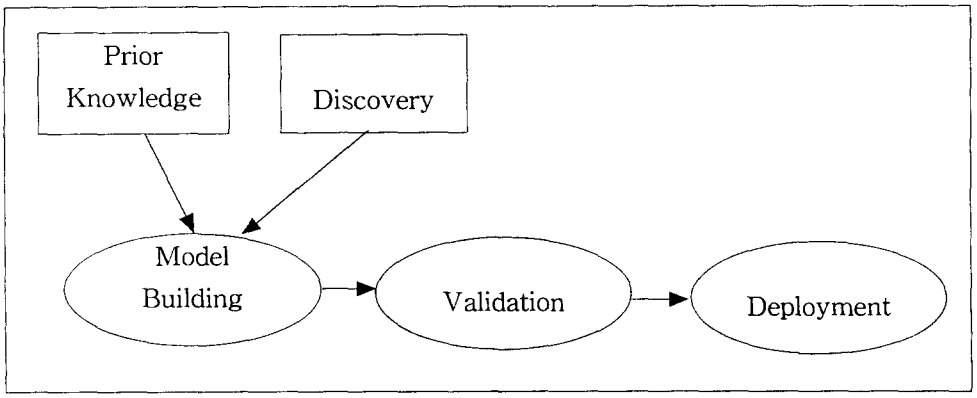
고객관계관리를 수행하기 위한 분석기법으로는 전통적인 통계분석기법 뿐만 아니라 인공신경망, 의사결정트리, 유전자 알고리듬 등을 포함한 다양한 기법들이 데이터마이닝 분석기법으로 통합 활용되고 있다.

데이터마이닝은 데이터에 내재되어 있는 유용한 정보나 변수들간의 관계를 정교한 분석모형을 통하여 찾아내는 작업이다. 즉 데이터마이닝은 데이터로부터 다양한 형태의 유용한 정보를 추출하기 위하여 모델링 방법을 적용하거나, 관측된 패턴의 유용성을 판단하는 일련의 과정이며 수많은 데이터 속에 내재되어 있는 데이터간의 의미 있는 상관관계, 패턴, 경향, 규칙 등을 찾아내어 모형화함으로써 유용한 지식을 추구하는 일련의 과정으로 정의할 수 있다. 여기에는 통계와 인공지능 알고리듬을 비롯한 인공신경망(artificial neural networks)등의 다양한 분석기법이 사용된다[8, 18, 20].

데이터마이닝의 필요성이 부각되는 환경요인들을 보면 다음과 같다. 첫째, 기업들의 운영계에는 데이터마이닝 분석을 수행하기에 충분한 용량의 데이터가 축적되어 있다. 둘째, 자동화된 자료의 수집, 자료저장구조의 기술적 발전 등으로 분석에 적합한 데이터의 저장과 축적이 가능하다. 셋째, 컴퓨터의 성능이 대용량의 자료를 축적하고 분석하기에 충분하다. 넷째, 기업간의 경쟁이 심화됨에 따라 기업의 대용량 데이터에 대하여 고부가가치를 창출할 수 있는 데이터웨어하우징(Data Warehousing)과 시스템구축에 따른 인프라가 이루어지고 있다.

데이터마이닝의 활용 분야는 크게 데이터베이스 마케팅(DB Marketing), 신용평가(Credit Scoring), 품질개선, 부정행위 적발(Fraud Detection), 그리고 이미지 분석 등으로 나눌 수 있다. 데이터베이스 마케팅분야에서는 마케팅 전략수립(Marketing Planning), 목표시장 마케팅(Target Marketing), 시장 세분화(Market Segmentation), 고객성향 변동 분석(Churn Analysis), 교차판매(Cross Selling), 시장바구니 분석(Market Basket Analysis) 등에 주로 활용되고 있으며, 신용평

가 분야에서는 신용평가의 중요한 사안에 대하여 의사결정을 지원하기 위하여 활용되고 있다. 이 외에도 문자 인식, 의료진단, 위험관리, 수요예측 및 판매 관리 등의 다양한 분야에서 활용되고 있다[25].



[그림-1] 데이터마이닝 모델링

모델링 도구는 가설과 데이터 두 가지 형태로 나뉘어 질 수 있다. 가설검증으로 불리우는 가설 모델링은 기존의 현상들에 대해 더 정확히 구체화하며, 가설 모델링 도구는 사용자가 사전적인 지식으로 모델을 세우고 모델의 타당성을 검증하는 방식으로 주로 이용된다. 반면, 데이터 모델링 도구는 자동으로 데이터 상에 존재하는 패턴을 찾는 방식으로 모델을 세운다. 즉 새롭게 발견된 모델이 수용되기 전에 얼마나 타당한지 테스트하는 것이 당연하다. 이런 가설 모델링과 데이터 모델링의 결합은 일반적이고, 최종 모델을 찾기 위한 반복적인 프로세스 모델링이라고 볼 수 있다.

2.3 데이터마이닝기법

데이터마이닝은 사용되는 목적에 따라 '검증(verification)'과 '발견(discovery)'의 두 가지 측면으로 분류해 볼 수 있다. 검증을 위한 데이터마이닝은 사용자가 세운 가설을 증명하기 위해 데이터로부터 정보를 추출하는 것이 목적이며, 발견을 위한 데이터마이닝은 데이터로부터 새롭고 유용한 패턴을 추출하여 사용자에게 제시하는 것이 목적이다. 발견을 위한 데이터마이닝의 기법들은 사용형태에 따라 '예측(prediction)'과 '설명(description)'의 두 형태로 나뉘어진다. 예측 형태에서는 관심 있는 요인들간의 작용을 예측하기 위하여 데이터로부터 관련 패턴

을 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[11, 28].

즉 분류분석을 위해서는 통계학의 다변량 판별분석(multiple discriminant analysis), 신경망의 다계층 퍼셉트론(multi-layered perceptron) 및 기계학습(machine learning)의 ID3 또는 C4.5 등이 사용될 수 있다. IBM Almaden연구소의 Agrawal 등[1993]에 의해 처음 시도된 연관규칙 탐사기술의 개발은 단순한 개념에서 출발하였지만 대용량의 데이터베이스로부터 유용한 지식을 찾아낸다는 점에서 실용성이 매우 큰 기술로 장바구니분석에서 위력을 발휘하였으며 최근에는 순차패턴의 탐색기술도 유용하게 사용되고 있다[11].

2.4 데이터마이닝의 적용사례

최근 마케팅, 금융, 재무, 생산, 보건 분야 등 다양한 영역에서 사용될 수 있는 데이터마이닝 애플리케이션과 프로토타입들이 개발되어 사용됨으로써 국내외에서 많은 적용사례들이 발표되고 있다.

2.4.1 국내 사례

국내에 데이터마이닝이 실제 업무에 적용된 최초의 사례는 1997년 BC카드사의 부정사용자 적발(fraud detection)이라고 할 수 있다. 그리고 데이터마이닝 분석을 시도하고 그 결과를 발표한 첫 사례는 1998년 초에 시도되었던 보험개발원의 사고 다발자 성향분석이다. 이 후 1998년 이동통신업체, 1999년 손해보험사, 그리고 2000년 증권사를 중심으로 데이터마이닝이 도입되었다[6].

BC카드사는 SAS사의 신경망 매크로를 이용하여 타인의 신용카드를 불법으로 사용하는 거래 패턴을 감지하는 시스템 구축을 시도하였다. 사용되었던 신경망 알고리듬은 알파 버전 수준의 것으로 사용자 인터페이스나 산출된 결과물 모두 불안한 상태였다. 뿐만 아니라 데이터마이닝이라는 개념이 최초로 도입되었고, 기존의 통계적 분석방법과 많은 차이가 있는 신경망을 이용한 분석이기에 적용에 있어 많은 오류가 있었다.

보험개발원은 보험사로부터 1년 단위로 모여진 보험가입자 자료 중 '94년 한해에 대하여 자동차 사고 다발자의 성향을 분석하였다. 이 연구는 '94년 자료 중 개인용 자동차보험에 가입한 총 2,176,684개의 가입자에 대하여 증권별 사고건수를 중심으로 진행되었다. 이 사례에서 Friedman과 Fisher(1997)가 개발한 PRIM(Patient Rule Induction Method)을 이용하여 자동차보험 사고 다발자 군을 분석한 결과에 의하면 현재 자동차보험에서 요율 변수로 사용되는 요인 이외의 여러 가지 요소가 조합된 사고 다발자 성향들을 발견할 수 있었다. 그리고 데이터마이닝을 이용한 분석결과의 도출에 있어 제공된 알고리듬만을 이용한 단순 작업보다는 상호 대화적인 분석으로 더욱 합리적인 결과를 얻을 수 있었다.

이동통신업체들의 도입사례는 1998년, PCS 3사가 이동통신 시장에 참여한 이후 신규고객의 확보보다는 기존 고객의 유지관리가 더욱 중요한 기존 이동통신업체의 데이터마이닝 활용에 관한 것으로서, 당시 데이터마이닝을 고려한 CRM 주제는 고객이탈 방지, 이용요금에 대한 수·미납 관리, 대리점 관리와 같은 세 가지였다. 이 중 대리점 관리는 OLAP(On-line Analytic Processing)을 이용한 분석을 사용하였고, 고객이탈 방지 및 수·미납 예측은 데이터마이닝 분석에 의하여 고객관리 시스템을 구축하였다.

이 사례에서는 분류예측의 정확도를 높이기 위하여 혼성모형(hybrid model)을 이용하였다. 혼성모형의 기본 골격은 우선 의사결정트리 모형을 이용하여 분류모형을 형성한 후, 의사결정트리 모형의 종료 노드 번호를 가변수로 형성한 후 신경망과 같이 모형 해석은 불가능하나 정확성이 높은 모형의 설명변수로 사용하는 것이다. 이 과정에서 의사결정트리 모형의 결과는 해지 가능성이 높은 고객에 대한 특성을 파악하여 캠페인 전략을 세우기 위해서 모형의 해석이 필요한 마케팅 부서에 제공되고, 의사결정트리 모형 결과를 신경망 등에 적용하여 정확도가 향상된 해지 예상자 리스트는 콜 센터에 전달되어 각 부서의 요구사항을 모두 만족시켜줄 수 있도록 하였다. 혼성모형을 사용하여 실제로 의사결정트리 모형만을 사용하는 것보다 이득률(gain)이 30% 이상 향상되었고 이를 통해 효율적 해지 이탈방지가 가능하였다.

이외에도 보험회사 이탈고객 관리분석, 카드고객 특성분석, 인터넷 쇼핑몰 고객세분화 등 다양한 영역에 데이터마이닝이 적용되고 있음을 알 수 있다[2][4][7].

보험회사 이탈고객 관리분석 사례에서는 국내 K화재보험회사의 A라는 자동차보험상품에 가입한 고객에 대한 고객 속성 및 해지유무의 변수들로 구성된 자료에 대해 신경망 모형과 의사결정트리 모형을 이용하여 고객이탈에 영향을 미치는 요인과 이탈고객집단을 분류하였다. 이 분석을 통하여 자동차보험에 있어 이탈의 주요 원인으로 가입시기와 지역, 보험료를 수금하는 방법 및 납입방법 등을

발견하였고, 모형 평가의 측면에서는 단순히 오분류표를 이용하여 모형을 결정하는 것보다는 이익도표 등과 같은 여러 도구를 이용하여 분석의 목적에 맞는 평가를 수행하는 것이 실제적용분야에 바람직하다는 결과를 도출하였다.

카드고객 특성분석 사례에서는 A 카드사의 1998년 1월부터 5월까지 서울에 거주하는 고객의 월별 카드사용과 인구통계변수를 포함하는 일반(기업고객 제외) 고객 특성에 대한 자료를 이용하여 의사결정 나무 분석을 실시한 후 고객을 몇 개의 집단으로 구분하여 각 집단별 성향을 파악하고 이를 예측할 수 있는 모형을 구축하였다.

인터넷 쇼핑몰 고객세분화 사례에서는 국내 인터넷 쇼핑몰 기업의 자료를 대상으로 데이터마이닝 방법론 중의 하나인 분류 규칙(Classification)을 이용하여 기존고객을 세분화한 다음 고객 개개인의 특성에 맞는 마케팅 프로모션을 하게 하고 신규고객을 획득할 때는 신규고객의 특성을 미리 예측하여 세분화하였다.

2.4.2 외국 사례

국내에서 뿐만이 아니라 외국에서도 금융업, 유통업, 통신업 등 다양한 영역에서 데이터마이닝을 활용하여 고객 데이터의 분석과 이해를 시도하고 있다. 다음은 SAS사의 Enterprise Miner 도구를 이용한 사례들이다[5].

금융업에서는 은행 계좌의 기록, 자동지급기의 기록, 신용 카드의 거래 기록, 투자 기록 등 날마다 다양한 형태의 데이터가 수집된다. 타 업체와의 경쟁에서 우위를 점하기 위해서는 이러한 자료를 적절하게 제품, 가격, 프로모션 등에 적용을 하여 투자수익률(ROI)를 극대화하기 위한 방안을 모색해야 한다. 전세계의 많은 은행과 카드 업계에서 데이터마이닝 기법을 도입하여 고객의 성향을 예측하며, 사기 감지, 고객 이탈 등의 분야에서 그 효과를 거둔 사례가 발표되고 있다.

미국의 First Union은행은 고객만족도 향상을 위하여 매스마케팅에서 일대일 마케팅으로 전환하고 데이터마이닝 기법을 활용한 결과 캠페인 반응이 60% 향상되었다. 또한 Bank of America는 20%의 은행고객이 150%의 이익에 공헌을 하며, 40~50%에 해당하는 고객이 은행전체의 이익 50%를 감소시킨다는 것에 대하여 데이터마이닝을 실시한 결과 50%의 이익을 감소시키는 고객 중 상위 20% 고객을 탐지해낼 수 있었다. Wells Fargo은행은 기존의 모형에 대한 불만족을 해소하기 위하여 데이터마이닝을 실시하여 기존의 주 단위의 예측 모형에서 일 단위로의 전환을 통해 보다 신속한 파악이 가능해졌다.

독일의 Deutsche Bank는 신용평가 수익성이 높은 고객의 행동예측을 위하여

데이터마이닝을 실시하여 보다 정밀한 모형을 찾게 되었으며, 이에 소비되는 시간도 단축되었다.

Winterthur사는 유럽의 선두 보험회사로서 세계 10대 보험회사에 손꼽히고 있다. 이 회사는 데이터마이닝 프로젝트 실시 당시 스페인에 100만명 이상의 고객을 보유하고 있으며, 해마다 13만건 이상의 계약이 취소되고 있어서, 이윤의 손실과 신규고객 가입비용으로 인해 심각한 재정난을 겪고 있었다. 신경망 네트웍의 조합으로 구성된 마이닝 모형을 이용하여 무작위 테스트의 90%에 대해 계약취소 고객을 예측할 수 있었다[35].

유통업계에서는 고객의 거래 데이터와 인구 통계학적 데이터, 라이프 스타일 등의 관계를 파악하여 그 결과를 고객의 충성도를 향상시키는 방법이나, 신규 고객의 확보나 교차판매, up-sell 등에 주로 마이닝 기법을 이용한다.

미국의 Newport News사는 제한된 카탈로그 분석을 통한 고객세분화를 실시하여 30시간이 걸리던 분석 시간이 30초로 단축이 되었으며, Vermont County Store는 메일링 비용의 증가 등에 따른 세일 기간의 카탈로그 발송의 반응률 향상을 위하여 마이닝을 실시하여 1,182%의 ROI를 거둔 사례가 있다.

통신업계에서는 미국과 유럽을 위시해서 각국마다 다수의 통신업체가 들어서게 되고, 신규 고객의 확보가 어느 정도 한계에 부딪히게 됨에 따라, 그 효율성이나 비용면에서 신규고객의 확보보다는 기존의 고객 유지에 중점을 두고 있는 형편이다.

미국에서는 AT&T가 모델링 프로세스의 비효율성을 개선하기 위하여 마이닝을 실시하여 모형 적합과정에서 50%의 업무처리시간을 단축하였다. 또한 ICG Netcom은 고객의 행동예측을 위하여 마이닝을 실시하여 대중을 대상으로 한 메일 발송 대신 특정의 우수고객을 추출하여 메일을 발송하여 경비를 절감하였다.

프랑스에서는 France Telecom이 고객 이탈 방지, 타겟 마케팅, 마케팅 캠페인의 예측 등을 목적으로 마이닝을 실시하여 이탈고객 방지에 큰 효과를 보았다. 캐나다에서는 MT&T가 자본 투자의 비효율성을 개선하기 위하여 마이닝을 실시하여 시장점유율이 5% 증가하는 효과를 보았다.

2.5 프로젝트관리분야의 데이터마이닝 사례

기업의 정보시스템이 e비즈니스 환경으로 급속하게 변화함에 따라 전사 프로젝트의 체계적인 관리 필요성이 대두되고 있다. 이는 기업들이 업무 관행 혁신을 통해 신상품 개발기간 단축과 기업 경쟁력 제고를 위해 매달 몇 백개씩 추진되고 있는 다양한 단위 프로젝트들의 추진 현황을 종합적으로 파악하고 관리해야

하기 때문이다. 프로젝트 관리는 국방, 건설, 우주항공 등 대형 프로젝트뿐만 아니라 IT 분야에서도 폭 넓게 적용되고 있으며, 다양한 프로젝트 개발방법론과 도구들이 사용되고 있다.

현재 국내에서 추진된 프로젝트 관리 사례를 살펴보면, 국방과학연구소, 한국전력, LG전자, 현대전자, 삼성전자 등이 공관프로테크사의 '오픈플랜(Open Plan)'을 적용하여 공동 개발하였으며, 포항제철, 제일제당, 한빛은행 등이 전사 프로젝트 관리 체제 구축에 PM소프트사의 '리얼타임프로젝트(RTP)'를 적용하여 공동 개발하였다[33].

한국전자통신연구원(ETRI)는 2000년도 초에 연구소의 특성상 빈번하게 발생하는 프로젝트 실시에 따른 프로젝트 실행계획서, 프로젝트 진행시 발생하는 산출물들과 개인 연구원들의 이력관리 등의 문서를 전사차원에서 관리하기 위하여 EDMS(Electronic Document Management System)를 도입하였다. 국내 사이버다임사의 솔루션을 기반으로 구축된 ETRI의 이 프로젝트는 EDMS 백본(Backbone) 위에 사용자 인터페이스를 업무 및 이용자 환경에 맞게 변환하면서 이루어졌다. 이 프로젝트를 통하여 ETRI는 Workflow 엔진에 기반한 프로젝트관리시스템(PMS), 지식관리시스템(KMS)의 문서 리포지토리, 그룹웨어 및 LDAP와의 연동, PDF 자동 변환, 인사·예산관리 기존 시스템과의 연동, Gantt Chart 형식의 프로젝트 일정 관리 등이 이루어질 수 있도록 하였다[32].

삼성전자 네트워크 사업부도 실시간 웹기반의 전사 프로젝트 관리시스템을 개발하였는데, 이 관리시스템은 1,500여명에 달하는 연구개발인력의 업무 배치에 있어 객관적이고 효율적인 기준 마련과, 연구원들이 각각 다중으로 맡고 있는 프로젝트 업무의 진행 상황을 실시간으로 파악해 프로젝트에 투입되는 비용의 객관적 기준 마련과 프로젝트의 질을 높이고 각 연구원들의 업무 하중을 평등하게 배분하는데 그 목적이 있다[34].

포스코개발 압연사업부도 각종 기술 문서를 대상으로 프로젝트별로 문서분류관리, PDF 자동변환 및 PDF Reader와 통합, 포항·광양·서울 분사의 파일 리포지토리를 구축하기 위하여 전자문서관리시스템을 도입하였다. 포항공대도 각종 사무 및 행정 문서를 대상으로 연구정보관리시스템을 구축하고, 사무관리를 위한 문서 리포지토리 기능 및 전자결재 기능이 가능하도록 전자문서관리시스템을 도입하였다[31].

이외에도 한국통신, 한국통신프리텔, 포스코개발 설계본부, LG경제연구원, LG화학, 휴맥스, 국방부 품질관리소 등이 각종 기술문서 관리를 위하여 전자문서관리시스템을 도입하였다[31].

3. 의사결정트리기법

본 연구에서 사용하고 있는 알고리듬인 의사결정트리는 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다. 분류에 관한 연구는 기존의 통계학, 인공신경망, 의사결정트리 등의 분야에서 연구되어 왔다. 의사결정트리는 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단하며, 이해하기 쉬운 규칙으로 전환될 수 있는 장점을 가지고 있다.

의사결정트리분석을 위해서 CHAID, CART, C4.5와 같은 다양한 알고리듬이 제안되어 있으며, 최근에는 이들의 장점을 결합하여 보다 개선된 알고리듬이 제안되어 상용화되고 있다[10]. 사실 많은 연구자들과 소프트웨어들에 의해서 이들 알고리듬이 개선되었고 장점이 서로 결합되었기 때문에 알고리듬의 구별이 모호해지고 있다.

CHAID(Chi-Square Automatic Interaction Detection) 알고리듬이 의사결정트리 분석기법으로 많이 사용되고 있는데, CHAID알고리듬은 카이제곱-검정(이산형 목표변수), 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(Multiway Split)를 수행하는 알고리듬으로 AID(Automatic Interaction Detection) 시스템에서 유래되었다. CHAID는 변수의 성격이 범주형 데이터이고 예측변수와 결과변수간의 관계를 찾아야 할 때 가장 유용하다. 의사결정트리는 분석의 목적과 자료구조에 따라서 적절한 분리기준(Split Criterion)과 정지규칙(Stopping Rule)을 지정하여 의사결정트리를 얻은 후 분류오류(Classification Error)를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거한다. 그리고 이익도표(Gains Chart)나 위험도표(Risk Chart) 또는 시험용 자료에 의한 교차 타당성(Cross Validation)등을 이용하여 결정나무를 평가하는 단계를 거친다.

CHAID 알고리듬은 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 사용되는 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. 분리기준을 카이제곱 통계량으로 한다는 것은 p값이 가장 작은 입력변수와 그때의 최적분리에 의해서 자식마디가 형성된다는 것을 의미한다.

4. 의사결정트리기법을 이용한 분류모델의 개발

4.1 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 연구예산 및 계획에 반영하고 있다. 연구개발(R&D) 활동은 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.

4.2 R사의 데이터 생성 및 사전처리

본 연구는 연구과제의 계약기관 유형을 예측하기 위한 분류모델을 개발하기 위하여 데이터마이닝 기법 가운데 하나인 CHAID 트리 생성 알고리즘을 사용하였다. 의사결정 나무의 분할기준으로는 카이제곱 통계량, 지니지수(Gini Index),

엔트로피 지수(Entropy Index)가 있는데, 세 가지 기준은 큰 예측력의 차이를 보이지 않는다. 세 가지 기준 중 가장 낮은 오분류율을 가진 카이제곱 통계량을 나무의 분할기준으로 활용하였다. 분리기준의 유의수준은 0.20으로 하였다.

정지규칙(Stopping Rule)에 의해서 사전 가지치기(Pruning)를 수행하는 경우에는 일반적으로 모형의 정확도가 다소 감소하게 되지만 너무 적은 개체 수를 가지는 마디들을 포함하고 있는 의사결정트리는 모형을 일반화하기에 무리가 있다. 노드의 최대 깊이(Maximum Depth of Tree)는 4단계까지 분리되도록 정하였으며, 잎에 포함되는 최소 관측치의 수는 15를 지정하여 이 숫자보다 작을 경우 더 이상 분리가 일어나지 않도록 하였다.

데이터마이닝은 많은 변수를 가지고 있는 대규모의 데이터를 대상으로 하며 다양한 분석방법론에 의한 분석을 포함하고 있기 때문에, 모형의 타당성을 평가하고 여러 모형을 비교하는 작업이 필요하다. 이를 위한 하나의 전략으로 데이터를 훈련용(Training), 검증용(Validation), 그리고 평가용(Test)으로 분할하여 분석한다. 즉 훈련용과 검증용 데이터를 이용하여 모형을 구축, 조율(Tuning)한 후 평가용 데이터를 사용하여 도출된 모형의 최종적인 평가를 수행한다. 본 연구에 사용된 데이터는 분류정확도를 높일 수 있도록 반복적으로 자료 구분을 실행하여 최종적으로 훈련용(training)으로 40%, 검증용(validation)으로 30%, 평가용(test)으로 30%로 구분하여 할당하였다.

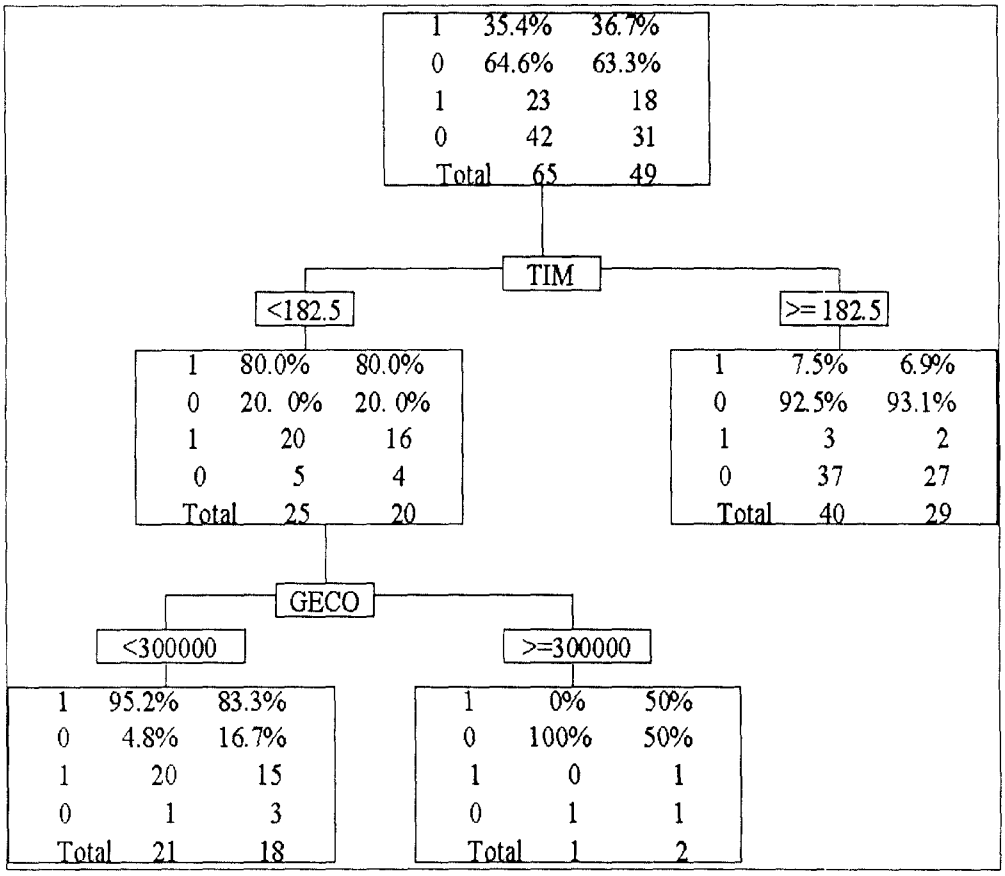
입력변수로는 주로 연구비 관련 비용변수로 17개의 변수를 입력변수로 사용하였다. 사용된 변수는 연구기간(TIM), 인건비(HUCO), 장비사용료(JACO), 컴퓨터사용료(COM), 감가상각비(DEPR), 재료비(MATER), 장치제작비(JAE), 보고서인쇄비(REPO), 국내여비(DOME), 국외여비(FORE), 외주시험비(OUTS), 자료수집비(DATA), 연구자문비(COUN), 위탁과제(OUCO), 기타경비(ETCO), 제경비(GECO), 기술개발비(TECO)를 입력변수로 선정하였으며, 연구계약기관을 목표변수(Target Variables)로 선정하여 그룹내부기관과 외부기관으로 구분하였다.

4.3 분류모델의 개발

[그림-3]은 계약기관 유형변수에 의해서 형성된 의사결정트리 결과이다. 의사결정트리는 뿌리(root)에서 시작하여 가지(node)를 형성해 나가며 의사결정나무가 형성되는데 원래의 의사결정나무는 [그림-3]보다 큰 나무였으나 가지치기를 하여 계약기관 유형에 따라 분류된 의사결정트리이다. 노드 1에서 연구기간(TIM)이 가장 유효한 변수로 사용되었으며 노드 2에서는 제경비(GECO), 노드 3에서는 기술개발비(TECCO)가 사용되었다. [그림-3]에는 분석용 표본의 결과와

평가용 표본의 결과로 구분하여 나타나 있다. 평가용 표본에 대한 분석결과로서 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 크거나 같은 관찰치로서 이 중에서 93.1%인 27개는 외부 계약기관에 해당하며 나머지 6.9%인 2개 관찰치만이 내부 계약기관에 해당한다.



[그림-3] 계약기관 유형 의사결정트리

한편, 연구기간이 182.5미만인 노드는 20개의 관찰치로서 내부 계약기관은 80%인 16개에 해당하며 외부 계약기관은 20%인 4개에 해당된다. 의사결정트리는 이

처럼 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 300,000미만인 계약유형은 내부 계약기관이 83.3%에 해당하는 15개 관찰치이며 외부 계약기관은 16.7%인 3개의 관찰치가 해당된다. 연구기간(TIM)이 182.5미만이면서 제경비(GECO)가 300,000이상인 관찰치는 계약유형으로 내부 계약기관과 외부 계약기관에 각각 1개의 관찰치가 해당되는 것으로 나타났으며, 이러한 하위 노드는 더 이상 분할되지 않으면 터미널 노드(terminal node)가 되면서 잎(leaf)으로 남게 된다. 그림에는 나타나지 않았지만 계약기간이 182.5미만이면서 제경비(GECO)가 300,000미만이고 기술개발비(TECO)가 607,500미만인 계약기관은 내부 계약기관이 83.3%인 15개이며 외부기관이 16.7%인 3개의 관찰치가 해당되며 기술개발비(TECCO)가 607,500 이상인 계약기관은 없는 것으로 나타났다.

의사결정 트리분석의 결과가 제시하는 것은 R사의 연구계약기관이 주로 6개월 이상(TIM 변수값 182.5는 일일단위이며, 월로 환산시 6개월을 의미함)의 연구기간을 근거로 하여 내외부 연구기관으로 분류되며, 분류된 6개월 미만의 연구에 대해서는 3,000만원 이상의 제경비(GECO 변수값 300,000은 3,000만원을 의미함) 예산편성 유무에 따라 연구유형을 분류할 수 있고, 3,000만원 이내의 제경비를 가진 연구유형에 대해서는 6,075만원의 기술개발비(TECO 변수값 607,500은 6,075만원을 의미함)를 기준으로 연구계약기관이 분류될 수 있다는 것을 의미하고 있다. 이 연구계약기관의 분류결과는 곧 연구계약을 체결한 고객기관과의 연구관리에 있어 연구기간, 제경비 및 기술개발비의 편성이 R사의 연구수주와 밀접한 관련성이 있다는 것을 보여주는 결과이며, 향후 연구주제 및 연구예산 수립 등의 업무에 이 결과를 활용할 수 있을 것이다.

4.4 분류모델의 검증

연구기관 유형 의사결정트리에서 선택된 변수는 총 3개의 변수인데 이들의 중요성(Worth)을 보면 [표-1]과 같다. 카이제곱(χ^2)이나 F 통계량을 분리기준으로 사용하는 경우에는 원래의 p-값 대신에 Log Worth를 계산하여 나타낸다. 또

한 지니 지수, 엔트로피 지수, 분산의 감소량을 분리기준으로 사용하는 경우에는 Worth(상대적 감소량)을 계산하여 나타낸다. 분리기준으로 Worth를 사용하였으며 Log Worth나 Worth는 값이 클수록 선호되어진다.

[표-1] 변수의 가치에서 의사결정트리의 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. 의사결정트리는 연구기간(TIM)을 연구모형에서 내부 계약기관과 외부 계약기관 분류하는데 있어 가장 중요한 역할을 하고 있는 변수로 선정하였다. 그 다음으로 제경비(GECO)에 0.4072라는 가치를 부여하여 의사결정트리를 형성하였다. 마지막으로 기술개발비(TECCO)에 0.3425라는 가치를 부여하였다.

[표-1] 변수의 가치

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

이 가치부여의 의미는 상대적 분산감소량이나 상대적 Log 통계량을 의미하는 것으로, 연구기간(TIM)이 1이라는 가치에 대해 제경비(GECO)는 0.4072만큼의 분류율, 그리고 기술개발비(TECCO)는 0.3425 정도의 분류기준을 가지고 있다는 것을 의미하고 있다.

[표-2] 의사결정트리 하위노드 각 노드값

노드구성	분류정확도	
	분석용	평가용
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]는 의사결정트리 하위노드 각 노드값에서 분류계수인 분류기준값이 0.5의 경우 노드 1에서 노드 2로 감에 따라 분석용과 평가용 모두 급격한 분류정확도(예측율)의 향상을 나타내며 노드 2에서 노드 3으로 감에 따라 점진적인 향상을 이루고 있다. 평가용의 경우 노드 2에서 최적의 분류 정확도를 나타내며 분석용에서는 노드 4에서 최적임을 알 수 있다.

[표-3]의 목표변수에 대한 평가 행렬의 평가용 자료에서는 내부 계약기관을 내부 계약기관으로 정 분류한 예측율은 18개 관찰치 중에서 15개로서 83%에 해당되며 외부 계약기관을 외부 계약기관으로 정 분류한 예측율은 31개 관찰치 중 28개인 90%을 나타내고 있다.

[표-3] 목표변수에 대한 평가 행렬

구분		목표변수	내부계약기관	외부계약기관	전체
평가용	관측개수	내부계약기관	15	3	18
		외부계약기관	3	28	31
		합계	18	31	49
	백분율(%)	내부계약기관	83	17	100
		외부계약기관	10	90	100
		합계	37	63	100

[표-4] 의사결정트리 적합 통계량

구 분	사용표본		
	훈련용	검증용	평가용
평균제곱오차	0.043	0.1197	0.1870
제곱오차 합	5.550	11.7263	18.3275
평균제곱오차 근	0.207	0.3459	0.4325
오차의 최대 절대값	0.925	1.0000	1.0000
분류케이스의 빈도	65.000	49.0000	49.0000
오분류율	0.046	0.1224	0.2041
빈도 합	65.000	49.000	4900

[표-4]는 계약기관 유형 변수를 사용한 의사결정트리 모형의 적합 통계량으로서 분석에 사용된 자료는 훈련용(training)이 65개, 분석용(validation)이 49개, 평

가용(test)이 49개로 사용되었으며, 오분류율은 훈련용이 4.6%, 분석용이 12.24%, 평가용이 20.41%로 나타났다. 또한 평균제곱오차는 훈련용이 4.3%, 분석용이 11.97%, 평가용이 18.7%임을 알 수 있다.

[표-5]의 평가용 자료의 의사결정트리 정오분류표 분석에서 사용된 전체 표본수는 49개이며 이 중에서 실제 외부 계약기관과 내부 계약기관 표본은 각각 31개(63.27%), 18개(36.73%)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 관찰치 중에서 정분류율(정확도)은 관찰치가 49/(28+15)로써 87.75%이며 오분류율은 관찰치가 (3+3)/49로서 12.24%로 나타났다.

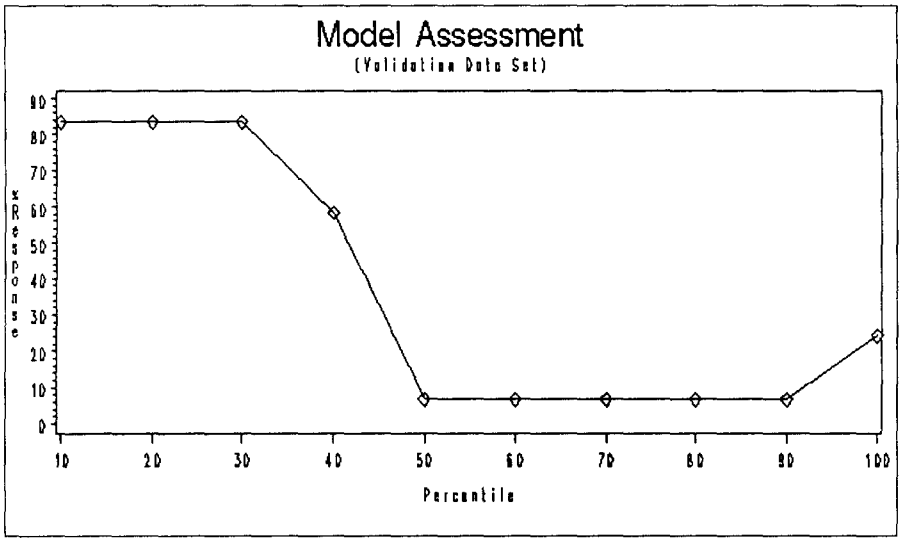
[표-5] 의사결정트리 정오분류표

빈도수 백분율 행 백분율 열 백분율		예측		전체
		외부계약기관	내부계약기관	
실제	외부계약기관	28 57.14 90.32 90.32	3 6.12 9.68 16.67	31 63.27
	내부계약기관	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

4.5 분류성과 측정

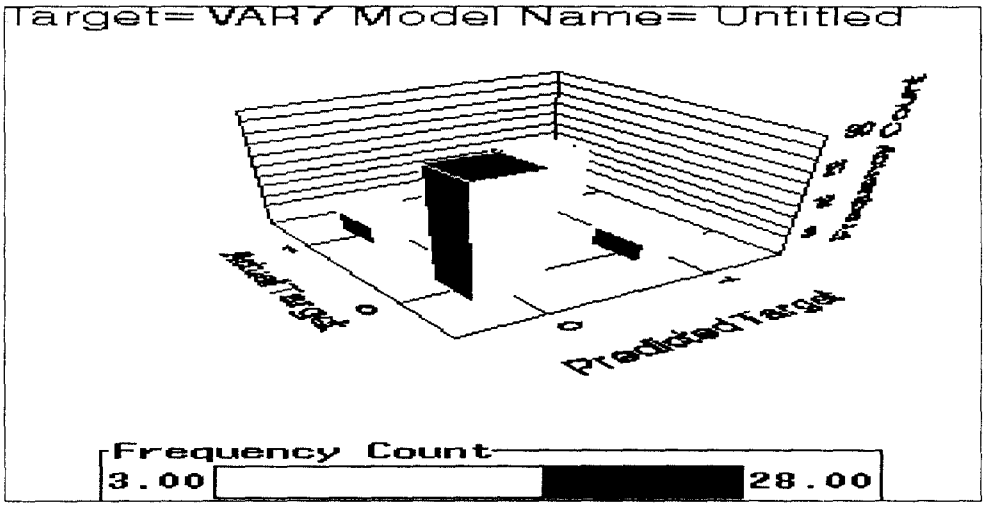
[그림-4]에서 의사결정트리의 누적 빈도 수를 이용한 %Response를 나타내었다. %Response는 해당 집단의 목표변수의 특정 범주 빈도를 해당 집단의 전체 빈도로 나눈 백분율 값이다. 따라서 %Response는 각 집단 내에서 범주 1의 빈도와 집단 내 관찰치의 빈도의 비를 나타내므로, 첫 번째 집단에서 %Response

는 (범주 1의 빈도)/(집단 1의 관찰치의 빈도)의 값이 된다. 즉 %Response는 목표변수 내에서 범주 1의 점유율을 각 집단에 대해 구한 값이라고 해석할 수 있다. [그림-4]에서 수평축은 전체 데이터 세트(Data Set)를 균일하게 N등분한 집단을 나타내며, 수직축은 해당 집단의 전체 빈도에 대한 해당 집단의 목표변수의 특정 범주 빈도에 대한 비율을 나타내고 있다. 즉, 상위 50%의 집단에서 목표변수에 대한 높은 점유율을 나타내고 있음을 알 수 있다.



[그림-4] 의사결정트리의 누적빈도수를 이용한 %Response

[그림-5]에서는 실제 관찰치와 예측치의 개수를 3차원으로 가시적으로 나타내고 있다. 외부 계약기관의 정분류에 해당하는 막대 그래프가 가장 높아 정확률이 뛰어남을 쉽게 알 수 있으며, 빈도 수에서도 그래프 윗면의 색상으로 가장 높은 빈도를 나타내고 있다. 내부 계약기관도 외부 계약기관 보다는 조금 낮지만 높은 정확률을 나타내고 있으며, 빈도 수에서도 외부 계약기관의 정분류에는 미치지 못하지만 높은 것을 알 수 있다.



[그림-5] 의사결정나무의 정오분류행렬

5. 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 내부 계약기관과 외부 계약기관으로 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. 의사결정트리 분석의 결과는 R사의 연구계약기관이 주로 6개월 이상의 연구기간을 근거로 하여 연구고객이 분류되며, 분류된 6개월 미만의 연구에 대해서는 3,000만원 이상의 제경비 예산편성 유무에 따라 연구고객을 분류할 수 있고, 3,000만원 이내의 제경비를 가진 연구고객에 대해서는 6,075만원의 기술개발비를 기준으로 연구계약기관이 분류되었다.

의사결정나무모형에서 계약기관 유형에 사용된 자료는 훈련용이 65개, 분석용이 49개, 평가용이 49개이며 오분류율은 훈련용이 4.6%, 분석용이 12.24%, 평가

용이 20.41%로 나타났다. 의사결정트리는 뿌리마디에서 출발하여 가지를 형성하며 최종 잎을 가진 것으로 형성되었으며, 노드 4에서 더 이상 개선의 효과가 나타나지 않아 훈련용의 분류 정확률은 95.40%를 나타내었다.

분석결과, 실제 내부 계약기관과 외부 계약기관 표본은 각각 18개(36.73%), 31개(63.27%)가 사용되었으며, 정오분류표에서 외부 계약기관을 외부 계약기관으로 예측한 정확률이 90.32%, 오류율이 16.67%로 나타났다. 또한 내부 계약기관을 내부 계약기관으로 예측한 정확률이 83.33%, 잘못 예측한 오류율은 9.68%로 나타났다. 이 결과는 전체 관찰치 중에서 정분류율(정확도) 관찰치가 (28+15)/49로써 87.75%이며 오분류율은 관찰치가 (3+3)/49로서 12.24%인 것으로 나타났다. 이 결과는 외부 계약기관의 정분류가 가장 높은 정확률을 보이고 있으며, 빈도 수에서도 가장 높은 빈도를 보인 것이다. 그리고 내부 계약기관도 비교적 높은 정확률을 나타내며 빈도 수에서도 외부 계약기관의 정분류에는 조금 미치지 못하는 높은 결과를 보였다.

본 연구를 통하여 R사는 내부계약기관과 외부계약기관이 연구과제를 계약시 고려하는 주요 변수로 17가지의 연구항목 중 연구기간, 제경비, 기술개발비 등의 의사결정변수가 사용되고 있다는 점을 파악하게 되었다. 이로써 계약기관들의 연구과제 특성에 따라 연구예산 편성기준을 검토하여 연구비 책정이나 연구예산 수립에 이 결과를 반영함으로써 연구비 항목과 관련한 연구고객의 관리방안을 수립하게 되었다.

향후 연구방향으로는 연구분야별 과제의 특성에 따라 연구비 집행실적을 기준으로 몇 가지 유형을 설정하여 연구예산 편성시의 관리기준을 수립하기 위한 연구가 수행될 수 있으며, 연구계약기관이 지속적으로 추진하고자 하는 분야에 대한 연구주제 선정 및 관리방안 등을 연구가 진행될 수 있을 것이다.

참 고 문 헌

[1] 김기서, 선진금융으로 가는 고객세분화 마케팅, 서울 : 고원, 1999.
[2] 김시환 · 권영식, 데이터마이닝을 이용한 인터넷 쇼핑몰 고객세분화, 2000 SPSS 사용자 사례 논문, http://www.spss.co.kr/cool/userpdf/index.htm
[3] 김재문, e-비즈니스 모델에 맞는 eCRM, , 서울 : 거름, 2000.
[4] 서혜선 · 김미경, Clementine을 이용한 카드고객 특성분석, 1999 SPSS 사용자 사례 논문, http://www.spss.co.kr/cool/userpdf/index.htm
[5]윤지숙, 한국SAS컨설팅교육팀,

http://www.mylab.co.kr/main/html/relevance
/Chemo/chemometrics/chemometrics%20home/pnustat/example.htm
[6] 최대우, 국내 통신업체에서의 데이터마이닝 적용사례, 2001,
http://www.gocrm.net/data /databank.htm
[7] 최종후 외 3인, 클레멘타인을 이용한 보험회사 이탈고객 관리분석, 1999 SPSS 사용자 사례 논문, http://www.spss.co.kr/cool/userpdf/index.htm
[8] Adrianns. P. & D. Zantinge, Data Mining, Addison-Wesley press, 1997.
[9] Agrawal, Rakesh, Ashish Gupta, Sunita Sarawagi, "Research Report : Modeling Multidimensional Databases," IBM Almaden Research Center.
[10] 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.
[11] Agarawal, R & R. Srikant, "Mining Sequential Patterns," Proc. of 11th Int.'.Cong. on Data Engineering, Taipei, Taiwan, march, 1995.
[12] B. de. lalesia, J.C.W. Debuse and V.J. Rayward-Smith, "Discovering Knowledge in Commercial Database Using Modern Heuristic Techniques, "KDD-96, 1996.
[13] 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
[14] Brian R. Gaines, Mark A. Musen, Ramasamy Uthurusamy, Cochairs, "Artificial Intelligence in Knowledge Management," Papers from the 1997 AAAI Symposium.
[15] Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone, "Classification and Regression Trees," Wadsworth, Belmont, 1984.
[16] Fayyad, Usama, "Diving into Databases." Database Programming & Design, March, 1998.
[17] 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
[18] Glymour, Clark, David Madigan, Daryl Pregibon, and Padhraec Smyth, "Statistecal Inference and Data Mining, "Communications of the ACM, Vol. 39, No. 11, 1996.

[19] Gupta, S., "Impact of Sales Promotions on When, What and How Much to Buy," journal of Marketing Research, Vol. 25, 1988. pp. 342-355.
[20] Hogarth, R.M. and S. Makridakis, "Forecasting and Planning: An Evaluation," Management Science, Vol. 27, 1981, pp. 115-138.
[21] Hong, S. J., Data Mining for Decision Support, Working Paper, IBM Watson Research Center, 1996.
[22] Kass, G., " An Exploratory Techniques for Investigating Large Quantities of Categorical Data," Applied Statistics, 1980, 29, 2, pp.119-127.
[23] Mannila, Heikki, Department of Computer Science University of Helsinki. "Methods and Problems in data mining". 1996.
[24] Mehta, Manish, Jorma Rissanen, Rakesh Agrawal, "MDL-based Decision Tree Pruning," IBM, Almadesn Research Center, Mehta, Rissanen, agrawal@almaden.ibm.com.
[25] Michael J. A. Berry and Gordon Linoff, " Data Mining Techniques for Marketing, Sales, and Customer Support, New York : John Wiley & Sons, Inc, 1997.
[26] Parsaye, Kamarn, "OLAP & Data Mining : Bridging the Gap." Database Programming & Design. Februay 1998.
[27] Quinlan, J. R., "Induction of Decision Trees," Machine Learning, Vol. 1, 1989, pp.81-106.
[28] Ronald J.B., Tej Anand, "The Process of Knowledge Discovery in Databases : A First Sketch, "AAAIi-94 Workshop on Knowledge Discovery in Databases, KDD-94, 1994.
[29] Ronald J. B., T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, & E. Simoudis, "Mining Business Databases," Communications of the ACM, Vol. 39, No. 11, 1996, pp. 42-48.
[30] Tumer, K. and Ghosh, J., "A Framework for Estimating Performance Improvements in Hybrid Pattern Classifiers," In Proceedings of the World Congress on Neural Networks, INNS Press, III, 20-225, 1994.
[31] http://millennium.cyberdigm.co.kr/customers/site.htm
[32] http://millennium.cyberdigm.co.kr/news/press20000501.htm
[33] http://www.itbiz.co.kr/news/cio/2000070108.htm
[34] http://www.itbiz.co.kr/news/cio/2000080108.htm
[35] http://www.spss.co.kr/customer/clem_stories/WINAPP-0299.htm

Effective R & D Management using Data Mining Classification Techniques

Hwang, Seok Hae* Moon, Tae Soo** Lee, Jun Han***

Abstract

This purpose of this study is to drive important criteria for improving customer relationship of R institute using data mining techniques. The focus of this research is to consider patterns and interactions of research variables from research management database of R institute, and to classify the outside organizations and the inside organizations for research contract organizations, and to decide the directions of customer relationship management through analyzing the research type and research cost of research topics.

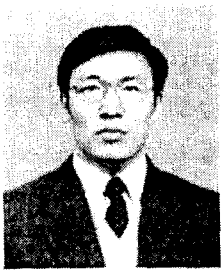
In order to drive criteria variables through pattern analysis of the research database, decision tree algorithm is employed. The results show that determinant variables of 17 input variables are research period, overhead cost, R & D cost as variables to classify the outside and inside contract organization.

* Ph.D. in MIS, Hankuk University of Foreign Studies
** Assistant Professor in MIS, School of Business and Economics, Dongguk University
*** Assistant Professor in MIS, School of Business Administration, Kyongju University

◆ 저자소개 ◆

황석해(Hwang, Seok Hae)

공동저자 황석해는 현재 애드잇 정보기술(주)의 전략사업본부장(이사)으로 재직중이다. 대구대학교 경영학과를 졸업(1991)하고, 한국외국어대학교 경영정보대학원에서 경영학석사(1996)를 취득하였고, 한국외국어대학교 경영학과에서 박사학위(2000)를 취득하였다. 주요 경력으로는 포항제철, 한국생산성본부, 한국외대 MIS연구소에서 연구원 및 컨설턴트로 재직하였다. 주요 관심분야는 데이터마이닝, 고객관계관리(CRM), e-business, 리스크 관리(Risk Management), 정보기술의 전략적 활용 등이다.
Tel: 017-315-0505
E-mail: seokhae@dreamwiz.com)



문태수(Moon, Tae Soo)

공동저자 문태수는 현재 동국대학교 경주캠퍼스 상경대학 정보산업학과에 조교수로 재직중이다. 한국외국어대학교를 졸업(1986)하고, 동대학 경영정보대학원에서 MIS로 경영학 석사학위를 취득(1988)하고, 고려대학교 대학원 경영학과에서 MIS 박사학위(1996)를 취득하였다. RIST 경영경제연구소, 한국전산원, 고려대 기업경영연구소에서 연구원으로 재직하였다. 주요 관심분야는 정보시스템 전략계획, 전사적자원관리(ERP), 전자상거래, 정보기술의 경제성 평가 분야이다.
Tel: 019-538-2344
E-mail: tsmoon@mail.dongguk.ac.kr

이준한(Lee, Jun Han)

공동저자 이준한은 현재 경주대학교 경영학부 경영정보전공 조교수로 재직중이다. 성균관대학교를 졸업(1985)하고, 한국외국어대학교 경영정보대학원에서 MIS로 경영학석사학위(1991)를 취득하였고, 성균관대학교 경영학과에서 경영학박사학위(1998)를 취득하였다. 대웅제약 경영정보실, 정보통신정책연구원 등에서 근무하였다. 주요 관심분야는 데이터웨어하우스(Datawarehouse), 고객관계관리(CRM), 전사적자원관리(ERP) 등이다.
Tel: 016-514-4112
E-mail: leejh@kyongju.ac.kr