碩士學位論文

데이터 마이닝을 이용한 은행의 데이터베이스 마케팅에 관한 연구

濟州大學校 經營大學院

經營情報學科 經營情報學專攻

李 東 勳

碩士學位論文

데이터 마이닝을 이용한 은행의 데이터베이스 마케팅에 관한 연구

指導敎授 崔 炳 吉

濟州大學校 經營大學院

經營情報學科 經營情報學專攻

李 東 勳

2000年度

데이터 마이닝을 이용한 은행의 데이터베이스 마케팅에 관한 연구

指導敎授 崔 炳 吉

이 論文을 經營學 碩士學位 論文으로 提出함

2000年 6 月 日

濟州大學校 經營大學院
經營情報學科 經營情報學專攻

李 東 勳

李東勳의 經營學 碩士學位 論文을 認准함

2000年 6 月 日

委員長 徐 賢 珠

委 員 金 斗 京

委 員 崔 炳 吉

차 례

[표차례]

[그림차례]

제 1 장 서 론

제1절 문제의 제기

최근 금융관련 각종 규제 완화 및 금융시장이 개방됨에 따라 우리나라 은행들은 경쟁력을 확보하기 위해 여러 가지 방안을 본격적으로 논의하게 되었으며 각 금융기관의 업무 영역이 확대됨에 따라 고객중심의 영업문화, 새롭고 효과적인 영업 경로, 다양한 상품 개발 및 고품질 서비스의 대처 방안이 필요하게 되었고, 이러한 변화는 은행업 활동의 중심 축을 고객으로 바꾸는 결과를 초래하였다.

1999년 3월에 발표한 한국은행의 "미국 은행 소매금융부문의 리엔지니어링"이라는 보고서에 따르면 '금융상품은 쉽게 모방될 수 있으나 고객에 대한 서비스를 개발하고 유지하는 데는 수년이 소요되므로 소매금융부문에서 상업은행이 생존하기 위하여 보다 더 고객과 긴밀한 관계를 유지하는 고객 지향적이 될 수밖에 없을 것으로 전망' 하고 '보다 고객 지향적이 되기 위하여 소비자 계층별 유통채널과 금융상품의 이용 형태 및 고객 수익성 기여도 등에 대한 분석이 더욱 정교화 되어야 할 것' 이라고 보고하고 있다.[1] 이는 은행의 정교한 고객 분석에 따른 고객관계관리가 선택이 아닌 생존일 수밖에 없음을 시사하고 있는 것이다.

특히 은행에 있어서 가장 중심에 있고 중요한 것은 고객이다. 각 은행들이 고객의 정보를 자산으로 수익을 창출하기 위해서는 각종 대고객 서비스 채널을 통해 수집된 데이터를 기반으로 고객의 금융상품 구매 성향과 금융 서비스의 결함 등을 파악하고 이를 분석해 마케팅에 활용하며, 마케팅 활동의 결과로 새로운 금융상품 등을 기획하여 영업에 반영하는 등 각 부문에 상호 피드백을 주는 새로운 형태의 마케팅 전략이 요구되고 있다. 즉, 시장경제 원리에 기반을 둔 경쟁이 치열한 금융환경 하에서 우리나라 은행은 경쟁우위 확보를 위해 새로운 형태의 데이터베이스

1) 한국은행, "미국 은행 소매금융부문의 리엔지니어링", 한국은행 은행부 경영분석2실, 1999. 3, p.22.

마케팅(Database Marketing; 이하 DB마케팅) 능력을 배양하지 않으면 생존이 어렵게 되었다.

따라서 금융환경 변화에 따른 은행의 전략적 변화가 새로운 DB마케팅의 필요성을 증대시키고 있다. 규제완화 및 금융개방과 함께 은행간은 물론 비은행 금융기관과의 경쟁이 심화되고 있고 경기침체에 따른 은행들의 부실여신이 증가하면서 외형위주의 성장보다는 수익성 등 내실을 기하는 은행경영이 필요하게 되었으며 DB마케팅은 이러한 변화를 실현시키는 수단으로 각광받고 있다.

최근 들어 한국의 은행들이 DB마케팅에 대한 관심을 갖고 이의 실행을 위해 투자를 하거나 투자계획을 하고 있는 것은 바로 새로운 시장환경에 적응하면서 수익성을 제고시키려는 노력의 일환으로 이해할 수 있다. 즉 각 은행에서 개별고객과의 릴레이션십의 강화, 고객 유지를 위한 체계적 고객관리, 분산된 데이터베이스의 통합 및 데이터 웨어하우징의 구축, 텔레마케팅이나 이와 같은 직접유통수단의 활용, 우량고객 우대 프로그램의 실행 등 일대 일 커뮤니케이션을 중심으로 하는 마케팅 활동에 대해 많은 논의가 이루어지고 있으며, 이러한 과정에서 구체적인 실행 대안으로서의 DB마케팅의 중요성이 부각되고 있다.[2)]

수익성 위주의 경영과 함께 은행의 영업활동은 기존의 상품과 수신 중심에서 고객 중심으로 변하고 있다. 또한 신규고객획득 중심의 대고객 영업행위가 이제는 기존 고객을 얼마만큼 유지하느냐 하는 기존 고객 중심의 대고객 영업행위로 변화해 가고 있다. 고객 중심영업의 핵심은 고객과 은행간의 일대 일의 밀착된 관계를 형성하여 이를 바탕으로 고객의 수요 변화에 적극적으로 대처하는 소위 고객관계관리(Customer Relationship Management, CRM)의 도입이다. 이는 은행의 영업활동이 불특정다수를 대상으로 하는 대중영업(mass marketing)에서 고객 각각에 대한 차별화된 영업(individual marketing)으로 전환되고 있음을 시사한다.

이러한 추세에 부합하는 새로운 DB마케팅 시스템의 구축은 대형의 개선된 고객서비스 영역을 유지하여 새로운 영업 채널(channel)을 제공하며 특화된 기능과 개선된 처리과정 및 현재의 효과적인 기술사용 등을 포함해야 한다. 다시 말해 보다

---

2) 박찬욱, "데이터베이스 마케팅의 실행 수준에 영향을 미치는 요인들에 대한 연구 : 한국 은행을 중심으로", 마케팅연구 제14권 제2호, 한국마케팅학회, 1999.6, pp.45~46.

신속한 서비스, 보다 편리한 서비스, 넓은 선택의 기회 등 급증하는 고객의 요구사항을 만족시키기 위해서는 서비스의 품질 및 정보의 제공, 편리함의 척도로 마케팅 시스템의 업무를 재정의 하지 않으면 안된다.

정확한 고객정보를 바탕으로 하여, 세분화된 기준으로 분류된 고객의 욕구를 정확히 파악하여 서비스를 제공할 때만이 수익성이 높은 우수고객 확보가 가능한 일이다. 최신 정보기술(Information Technology)을 이용하여 이러한 고객정보를 분석하여 고객 세분화를 추진함으로써 수익성을 제고시킬 수 있는 해결책(solution) 중의 하나가 바로 데이터 마이닝(Data Mining)이다. 증대된 고객들의 요구사항을 수용하면서, 날로 더해 가는 은행간 경쟁에 효과적으로 대처하고, 경영혁신으로 인한 은행 내 계층구조의 축소에도 불구하고 적시에 정확한 의사결정을 내릴 수 있는 능력을 보유하기 위해서는 고도의 정보활용 능력제고가 필수적이며 이는 곧 경쟁력 강화로 이어진다는 점 때문에 데이터 마이닝의 역할이 더욱 중요해지고 있다.

개별 은행의 입장에서는 은행에 기여도가 큰 고객집단을 찾아내고, 이들을 중심으로 고객만족을 실현시킬 수 있는 '목표 마케팅(target marketing) 전략' 을 수립해 나가야 할 것이다. 이러한 전략수립을 위해서는 현재 각 은행마다 데이터베이스화되어 있는 고객정보파일 내에서 의사결정나무(decision tree), 인공신경망(neural network), 유전자 알고리즘(genetic algorithm) 등 데이터 마이닝 방법을 적극 활용할 필요성이 있다.

제2절 연구의 목적

과거 전통적인 은행의 마케팅 시스템은 주로 표준화된 금융상품을 대다수의 고객들에게 판매하는 대중마케팅(mass marketing)에 국한되었으나, 정보기술의 급속한 발달 및 금융기관간 상호협력체제의 가속이라는 새로운 환경 하에서, 고객의 금융거래 패턴 분석을 위한 마케팅 등의 구체적이고 종합적인 고려가 필요하게 되었다. 이는 은행 경쟁력 강화라는 총체적인 결과를 달성하기 위해서 은행의 단위조직이나 개별 업무흐름만이 아니라 조직간 상호협력을 추구하는 과정에서 필연적으로

당면하게 되는 현상이다.

DB마케팅이 강조되는 이유는 정보기술 및 인터넷의 발달로 인한 전자금융의 급속한 보급 및 확대 덕택으로, 고객의 요구가 즉각적으로 은행의 고객정보로서 활용되어지고, 고객과 은행간 정보교환이 보다 신속하게 이루어지고 있다는 것이다. 이와 같은 정보기술의 발달은 과거 단순히 직관이나 추정에 의존했던 마케팅에서 벗어나서, 대량의 고객 정보를 기반으로 보다 과학적이고 정교한 분석을 통해 효과적인 마케팅이 이루어질 수 있는 기반을 제공하고 있다. 즉, 정보의 부족이 아니라 정보로부터 은행의 마케팅에 가치있고 유용한 정보를 어떻게 추출하느냐가 더욱 중요한 관건이 되었다.

최근 데이터베이스 및 인공지능 분야를 중심으로, 정보의 과학적 분석을 통한 새로운 지식을 창출하는 데이터 마이닝(Data Mining) 기법이 활발하게 개발됨에 따라, 은행 DB마케팅 분야에서도 데이터 마이닝 기법의 활용을 통한 보다 세밀하고 정교한 마케팅을 위한 기반이 이루어 졌다고 할 수 있다.

이에 본 연구에서는 대용량의 데이터베이스로부터 의미 있는 지식(knowledge)을 찾아내는 데이터 마이닝에 대해 문헌을 토대로 개념을 정립하고 DB마케팅과 데이터 마이닝이 서로 어떻게 관련되어 있으며, 금융과 같은 분야에서는 어떻게 관련되어 있는지를 명확히 함으로써 이 분야에 대한 전체적 개념을 제공하고자 한다.

본 논문은 고객 데이터베이스로부터 높은 수익 기여도가 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 목표 마케팅(target marketing)의 수행 대상이 되는 우수고객을 선별하는 모델의 개발에 관한 것이다. 즉, 마케팅 담당자에게 이 모델에 의하여 선별된 고객의 리스트를 제공함으로써 마케팅의 성공률을 높이는 효과적인 DB마케팅에 관한 연구라고 할 수 있다. 본 연구는 고객 데이터베이스로부터 마케팅 대상 고객을 무작위로 추출하는 것이 아니라 데이터 마이닝을 통해 발견된 패턴에 따라 추출된 대상 고객을 마케팅 담당자에게 할당함으로써 마케팅 업무의 성과를 향상시키는 것을 목표로 하고 있다. 또한 A은행의 신용카드 고객 데이터베이스에 데이터 마이닝을 적용한 실증적 연구를 통해 데이터베이스를 효과적으로 은행 마케팅에 활용할 수 있도록 방향을 제시하고 현재와 미래에 대한 연구 방향을 제시하고자 한다.

제3절 연구의 방법 및 논문의 구성

본 연구는 이상과 같은 연구의 목적을 살펴보고자 문헌연구와 실증적 실험을 병행하였다. 문헌연구는 국내·외 저서 및 논문 등을 참고로 은행의 DB마케팅, 그리고 데이터 마이닝에 대한 이론적 고찰을 하였다. 실증적 실험은 A은행의 신용카드 고객 데이터베이스에 데이터 마이닝 기법중의 하나인 의사결정나무 알고리즘을 적용하여 모델을 개발하고 개발된 모델은 정오분류행렬(Confusion Matrix)표를 통하여 검증하였다. 그리고 모델의 성과는 리프트(LIFT)를 사용하여 평가하였다.

본 연구는 5개의 장으로 구성되어 있으며 그 내용은 다음과 같다.

제1장에서는 본 연구를 제기하게 된 상황 및 연구목적, 연구방법, 논문의 구성에 대하여 언급함으로써 본 논문이 연구하고 있는 것이 무엇인가에 대한 관점을 제시하고 있으며 연구의 접근 방법을 나타내고 있다.

제2장에서는 문헌연구를 통한 이론적 고찰을 하였는바, DB마케팅과 데이터 마이닝에 대한 이론적 설명과 데이터 마이닝의 활용 및 적용분야에 대한 논의, 그리고 데이터 마이닝의 적절한 적용을 위한 데이터 마이닝 기법과 방법론을 살펴보면서 각각의 상황에 맞는 방법과 장단점을 제시해 주고 있다. 또한 은행 DB마케팅과 데이터 마이닝에 대한 선행 연구내용에 대해 살펴보고 있다.

제3장에서는 본 연구의 실증적 실험을 위한 준비작업과 사전처리작업 등에 관한 사항을 기술하였다.

제4장에서는 이전 장에서의 이론과 방법론을 실증적 실험을 통하여 검증하고 있다. 이 실험은 A은행의 신용카드 고객에 대한 마케팅에 데이터 마이닝 기법을 적용하여 실험에 대한 결과를 분석함으로써 데이터 마이닝의 적절한 사용방향을 제시하고 있다.

마지막으로 제5장에서는 본 연구의 성과와 시사점을 제시하고 연구의 한계 및 앞으로의 연구방향에 대해 기술하였다.

제 2 장 은행의 데이터베이스 마케팅과 데이터 마이닝

제1절 데이터베이스 마케팅과 마케팅정보 관리

1. 데이터베이스 마케팅

데이터베이스는 관계형 데이터베이스 관리시스템의 발전에 따라 1970년 후반에 OLAP(On-Line Analytical Processing)의 보편화가 이루어졌으며, OLAP은 1980년대 중반 데이터 웨어하우스(Data Warehouse)로 발전되게 되었고, 1990년대에는 DB마케팅으로 활용되어지고 있다.

DB마케팅에 대해서는 많은 학자나 실무가들이 다양한 정의를 내리고 있다. 예를 들어, 마케팅 컨설턴트인 홀츠(Holtz)는 DB마케팅을 '고객에 대한 접근법, 마케팅 전략, 방법론 등의 마케팅에 대한 제요소가 단순한 구매자 리스트가 아닌 잠재고객에 대한 풍부한 정보에 근거한 마케팅'이라고 정의하고 있다.[3] 또한 DB 마케팅 컨설턴트인 휴지스(Hughes)는 DB마케팅을 '잠재고객과 기존고객에 대한 적합한 정보가 수록되어 있는 컴퓨터화된 관계형 데이터베이스 시스템을 고객에게 보다 질 높은 서비스를 제공하고 이들과 장기적인 관계를 구축할 수단으로 운용하는 것'으로 정의하고 있다.[4] 그리고 박찬욱은 DB마케팅을 '고객에 대한 여러 가지 정보를 컴퓨터를 이용하여 데이터베이스화하고 구축된 데이터베이스를 바탕으로 고객 개개인과의 장기적인 릴레이션십 구축을 위한 마케팅 전략을 수립하고 집행하는 제활동'이라고 정의하고 있다.[5]

이처럼 DB마케팅은 여러 가지 측면에 따라 다양하게 정의할 수 있지만 공통적

---

3) Holtz, Herman, *Databased Marketing*, John Wiley & Sons, Inc., 1992, p.5.
4) Hughes, Arthur M., "Strategic Database Marketing : The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program", Irwin, 1994, p.9.
5) 박찬욱, 「금융기관의 데이터베이스 마케팅」, 시그마인사이트그룹, 1999, p.18.

인 특징은 '고객에 대한 여러 가지 정보가 다양한 경로를 통하여 축적된 데이터베이스를 기반으로 고객 개개인의 릴레이션 구축을 위한 고객특성에 적합한 마케팅 전략을 수립·집행하는 모든 활동'이라 할 수 있다. 이는 산재되어 있는 기업 내·외부의 데이터들을 고객중심의 데이터베이스로 구축하고, 과학적(통계적)인 방법을 이용하여 이를 효율적으로 탐색, 새로운 특성을 발견하는 것이다. 또한 이를 근거로 마케팅 전략과 프로그램을 수립하여, 기업의 수익으로 연결시키는 과정이다.

2. 금융 마케팅정보 관리와 데이터베이스 마케팅[6]

1990년대 들어 미국은행을 중심으로 데이터베이스 마케팅이 본격 도입되기 시작한 이래 이제 데이터베이스 마케팅은 금융기관 경영에 없어서는 안되는 중요한 수단으로 자리잡고 있다.[7] 특히, 최근 은행업계에서는 수익성위주의 경영과 다양한 고객 욕구(needs)에 적극적으로 대응할 수 있는 차별적 전략의 추진수단으로 DB 마케팅이 크게 주목을 받고 있다. 종래의 상품중심, 불특정다수의 고객을 대상으로 한 매스 마케팅(mass marketing) 개념은 개별고객위주의 개별 마케팅(individual marketing), 일대 일 마케팅(one-to-one marketing), 고객과의 지속적인 관계 유지를 도모하기 위한 관계 마케팅(relationship marketing), 기존고객의 이탈을 방지하고 은행 이용도를 높이기 위한 고객유지 마케팅(retention marketing) 등 새로운 마케팅 패러다임(marketing paradigm)으로 변환되고 있다.

DB마케팅은 단순히 금융상품의 판매촉진을 위한 고객정보관리 뿐만 아니라 상품개발, 영업점관리, 직원교육, 점포정책 등 경영전략을 결정하는 수단으로 광범위하게 활용되고 있다.

은행의 DB마케팅은 정보기술을 이용하여 개별고객에 대한 각종 정보를 MCIF(Maketing Customer Information File)와 같은 형태로 데이터베이스화하고 이를 분석하여 고객을 세분화하며, 이들 각각의 특징에 맞는 금융상품 및 서비스를 영업점 창구, DM(Direct Mail), TM(Tele-Marketing) 등을 통해 권유함으로써 판매기회를 확대해 나가는 과학적 마케팅 방식을 말한다.

---

6) 조태현, 「금융마케팅 2」, 한국금융연수원, 2000, pp.40~52.
7) 김기서, 「선진금융으로 가는 고객세분화 마케팅」, 도서출판 고원, 1999, p.101.

여기서 MCIF(마케팅고객정보 파일)란 금융기업을 이용하는 고객들의 정보를 가구단위로 통합 구축한 데이터베이스로서 크게 내부데이터와 외부데이터로 구성되어 있다. 내부데이터는 성명, 연소득, 주소, 직업 등 고객속성관련 정보와 거래 개시 및 종료일, 예금·대출잔액, 거래상품의 수 등 거래관련 정보를 말한다. 외부데이터는 개인의 총금융자산, 주택 및 자동차소유여부, 가족구성 등 주로 외부전문기관을 통해 제공받는 데이터를 말한다. 이는 필요한 고객정보, 상품판매 정보 등을 실시간으로 검색 가공할 수 있고, 통계적인 분석도 용이하게 함으로써 보다 정밀한 DB마케팅을 실천할 수 있는 시스템이다.

DB마케팅은 고객과의 일대 일(one-to-one) 커뮤니케이션을 통해 지속적 관계를 구축하고 고객 만족도를 극대화함으로써 고객 이탈률 감소, 평균 구매량 증대도모 및 이에 따른 마케팅비용 절감으로 고객의 생애가치를 극대화 할 수 있다. 또한 DM, TM 등을 이용하여 직접판매 및 고객응대의 다양한 수단을 제공함으로써 고객의 요구에 신속하게 반응할 수 있는 효율적인 유통채널과 서비스 수행체제를 구축할 수 있다.

한편 상품의 구매행위, 문의응답, 촉진반응 등 고객과 접촉하는 전 과정을 통해 고객에 대한 데이터를 효율적으로 획득, 관리할 수 있고, 고객의 개별적인 욕구를 실시간(real time)으로 파악할 수 있으므로 상시적인 마케팅조사 체제의 구축이 가능하게 된다.

3. 마케팅 데이터베이스의 분석과 활용

1) 마케팅 데이터베이스의 분석

마케팅 데이터베이스는 그 가치를 발휘하기 위해서는 마케팅 정보(marketing information)로 전환되어 유효하게 사용될 수 있어야 한다. 따라서 고객 데이터베이스를 효율적으로 사용하기 위해서는 데이터베이스에 담겨있는 자료들을 함축적으로 요약할 수 있는 분석방법이 이용되어야 한다.

이러한 분석방법으로써 통계분석은 기술적인 단순분석이나(예를 들어, 합계, 평균, 분산 교차분석 등) 간단한 이변량분석(예를 들어, T-test, Chi-Square test 등)

을 이용할 수 있고, 보다 복잡한 분석방법인 다변량분석(예를 들어, 회귀분석, 군집분석, AID 등)이 사용될 수 있다. DB마케팅에서 가장 많이 쓰이는 단순분석방법 중의 하나로 이변량분석 방법인 교차분석(cross-tabulation)을 들 수 있는데, 이기법은 변수와 변수간에 관계를 보기 위해 어떤 변수가 특정 값을 가질 때의 빈도가 또 다른 변수의 값이 달라짐에 따라 어떻게 분포되는가를 나타내준다. 이 기법을 이용하면 카이스퀘어 테스트(chi-square test)를 통한 유의성 검정(예를 들어, 두 집단간에 유의한 차이가 있는가에 대한 검정)도 가능하지만 주로 시장상황을 개관하거나 시장특성에 대한 직관을 얻는데 유용하게 사용되고 있다. 보다 복잡한 형태의 다변량분석 방법으로는 회귀분석, 판별분석, 군집분석, AID(Automation Interaction Detection), CART(Classification and Regression Trees), CHAID(Chi-Square Automation Interaction Detection) 등과 같은 분석방법들이 유용하게 사용된다. 최근에 와서는 인공지능(Artificial Intelligence)분야에서 개발된 인공신경망(Neural Network) 모델과 같은 분석들도 적용되고 있다.[8)]

이외에도 DB마케팅에서 가장 많이 쓰이는 분석방법으로 RFM(Recency, Frequency, Monetary) 점수분석방법이 있다.

2) RFM 분석

수익성이 가장높은 고객을 찾아내기 위해 데이터베이스를 이용하는 방법과 가장 좋은 고객들을 구별해내는 기본적인 방법은 구매시기/빈도/구매량(RFM)공식이다. 각각의 고객들의 기록에 날짜, 구매량, 구매성격들을 포함함으로써 고객 각각의 성과기록을 알 수 있다.[9)]

RFM분석은 데이터베이스 마케팅에서 고객의 가치를 평가하는데 가장 널리 사용되는 대표적인 점수부여(scoring) 시스템이다. RFM분석은 최근 구매월(Recency), 이용빈도(Frequency), 이용금액(Monetary) 등 고객의 수익 기여도를 나타낼 수 있는 3가지 지표로 고객의 가치를 평가하는 방법이다.[10)]

---

8) 박찬욱, 「데이터베이스 마케팅」, 연암사, 1996, pp.170~171.
9) Bob Stone 著·금강기획마케팅전략연구소 譯, 「데이터베이스 마케팅」, 한국언론자료간행회, 1999, p.52.
10) 조태현, 전게서, p.44.

(1) RFM의 개념

RFM분석은 1950년대 카탈로그 마케팅회사에 의해 개발되었으며 특히 다이렉트 메일이나 카탈로그 우송 리스트 추출에 빈번히 사용되어 왔다.

RFM의 구성요소를 보면 결국 RFM은 가장 최근에, 가장 자주, 많은 액수의 상품을 구매한 고객을 가장 가치있는 고객으로 평가하는 기법이다. RFM을 고객의 가치를 평가하는 지표로 삼기 위해서는 각 요소별로 점수화에 필요한 구간설정과 함께 요소들간의 가중치 산정이 필요하다. 이러한 가중치산정은 상품특성이나 시장환경을 감안하여 합리적으로 조정되어야 하기 때문에 적절한 가중치 선정이 RFM 분석법의 실행을 위한 주요 노하우가 된다.[11]

① 최근 구매월

최근 구매월은 특정시점을 기준으로 고객의 마지막 구매로부터 경과된 기간에 따라 결정된다. 고객들의 이탈가능성을 사전에 파악하여 실제 이탈률을 최소화하는 한편 반복구매가 가능하도록 활성화전략을 실행할 수 있다는 측면에서 고객관리에 유용한 지표가 될 수 있다. 반복구매주기가 지났는데도 불구하고 재구매를 하지 않는 고객, 즉 최종 구매월이 지나치게 경과한 고객들은 이탈가능성이 상대적으로 높은 고객으로 판단할 수 있다.

② 구매빈도

구매빈도는 일정기간 동안 고객이 특정기업의 제품 또는 서비스를 구입한 횟수를 의미하는 것으로 고객의 상표충성도 또는 호감도를 반영한다는 측면에서 유용한 고객관리기준이 될 수 있다. 구매빈도는 주로 특정기업에 대한 충성도 또는 호감도 등과 같은 태도에 의해 결정되는 고객의 행동변수이다. 일정수준이상의 구매빈도는 건당 거래비용체감 등으로 수익성과 비례관계를 가지므로 유용한 고객평가기준이 될 수 있다.

---

11) 김기서, 전게서, pp.123~125.

③ 이용금액

RFM의 구성요소간의 가중치 산정은 업종, 목표고객 등에 따라 탄력적인 적용이 필요하나 일반적으로 이용금액이 흔히 수익성과 직결된다는 점에서 가장 중요한 고객평가 기준이 될 수 있다. RFM의 구성요소 가운데 다이렉트 마케팅 회사가 아닌 서비스 업체나 제조업체의 경우에는 구매액의 가중치중에서 가장 큰 비중을 차지하는 것이 일반적이다.

RFM은 최근에 구입한 고객일수록, 자주 구입한 고객일수록, 또한 많은 금액을 지불한 고객일수록 추가적인 구입을 할 가능성이 크다는 경험법칙에 그 근거를 두고 있다. 예를 들어 많은 금액을 은행에 예치하고 있는 고객은 일반적으로 여러 은행에 예금을 분산시켜 놓기 때문에 이를 대상으로 추가적인 예금을 유지하는 것이 예치금액이 적은 고객으로부터 똑같은 금액을 추가로 유지하는 것 보다 훨씬 수월하다. RFM에 근거하여 고객의 점수를 산출하기 위해서는 위의 3가지 요소에 가중치를 주어 합산해야 한다.[12]

RFM 모델은 그 원리가 매우 단순하지만 실제로 높은 반응률을 가져오기 때문에 지금까지도 다이렉트메일을 활용한 마케팅 분야에서 폭넓게 활용되고 있다. 실례로 대표적인 카탈로그마케팅 회사인 슈피겔(Spiegel)사의 RFM 모델 보다 더 좋은 반응률을 가져다주는 데이터분석방법은 아직까지 존재하지 않는다고 주장할 정도로 각광을 받고 있다.[13]

(2) RFM의 특징[14]

RFM 분석은 신규고객의 획득과 기존고객의 유지라는 2가지 마케팅 목표 중에서 특히 기존고객의 유지에 비중을 두는 전략적 사고에 기초하고 있다. 시장확대기 또는 경영환경이 안정적이고 성숙기에 접어들거나 경쟁이 심화될 때는 고객유지전략이 보다 합리적인 대안이 될 수 있다. 특히 고객유지가 코스트측면이나 효율의

12) 박찬욱, 전게서, 1999, p.117.
13) 김기서, 전게서, p.129.
14) 상게서, pp.131~134.

관점에서 볼 때 신규고객개척에 비해 상대적으로 우위에 있는 경우가 많다.

또한 RFM 분석은 신규고객보다는 기존고객을 중시하는 기본적 사고에 입각하고 있으나 기존고객들 가운데도 고객가치나 기여도를 기준으로 차별화해서 관리하는 우수고객 차별화에 큰 비중을 부여하고 있다. 즉 모든 기존고객이 다 가치있는 고객이 될 수 없다는 사실을 감안할 때 자사에 공헌하는 고객과 그렇지 않은 고객으로 구분해 우수고객과의 관계를 심화시키는 것이 보다 현명하다는 사고에 입각한 것이다.

RFM 분석은 특히 고객평가에 있어 구매행동을 기준으로 하는 분석방법이다. 고객을 선별하는데 있어서 고객의 속성과 구매행동이라는 2가지 측면에서 접근하는 것이 일반적이다. 그러나 RFM 분석에서는 고객의 구매행동을 보다 유의한 기준으로 파악하는데 이는 고객의 속성보다는 구매행동이 예측의 불확실성이 작고, 자사 상품의 구입을 예측하는 적중률이 더 높다고 가정하기 때문이다. 이것은 사람이 과거에 행한 행동과 동일한 행동을 이후에도 채택할 가능성이 높기 때문에 일단 나타난 구매행동이 차후의 구매행동을 예측하는데 가장 유용한 지표가 될 수 있기 때문이다.

RFM 분석은 고객세분화를 통해 기업의 수익성 제고에 기여할 수 있다는 사고에 기초하고 있다. RFM 분석을 통해 RFM 스코어가 높은 고객, 중간인 고객, 낮은 고객으로 나누는 것이 가능할 뿐만 아니라 개별고객의 RFM 점수와 기업경영 채산간에 긴밀한 상호관계가 있다는 것을 확인할 수 있다. 즉 RFM 점수가 높은 고객은 기업의 수익성에 기여를 하는 고객인 반면 RFM 점수가 낮은 고객은 수익성을 저하시킬수 있는 고객이라는 대응관계가 성립한다. 그러므로 RFM 분석을 활용, 개별고객의 수익성을 평가하여 일정수준 이상의 고객들만을 표적시장으로 선정하여 차별적인 마케팅 전략을 실행한다면 기업은 무리 없이 큰 수익을 획득할 수 있다.

따라서 RFM 분석은 구매가능성이 높은 우수고객에게 집중한 마케팅 전략을 실행하고 불필요한 자원낭비를 방지함으로써 근본적으로 매출액 증대보다는 이익을 증진시켜주기 위해서 고안된 모델이라 할 수 있다. 즉 RFM 분석은 고객가치측면에서 낮은 평가를 받거나 기여도가 미미한 고객들을 마케팅 소구 대상에서 제외함으로써 매출액 확대로 직결되지는 않더라도 불필요한 비용을 최소화함으로써 이익

을 증가시킬 수 있다는데 착안하고 있다.

RFM 분석의 특징을 요약하면 다음과 같다. RFM 분석은 리스트에 있는 모든 고객을 대상으로 마케팅을 하는 것이 아니라 고객을 세분화하고 채산성 있는 고객만을 대상으로 마케팅을 하는 것이 바람직하다는 사고에 기초하고 있다. 즉 고객을 채산집단과 비채산 집단으로 나누기 위해 테스트마케팅을 실시하여 여기서 판명된 반응률을 기준으로 손익분기점을 도출한다는 것이다.

RFM 분석은 채산성 있는 고객에게만 카달로그를 발송함으로써 이익실현을 확보한다는 접근법이다. RFM 분석은 본격적인 마케팅의 실시를 제창하는 것이 아니라 사전에 채산집단을 대상으로 하는 비즈니스를 구상하고, 그 구상 중에 매출액과 경비, 이익을 계산하여 확실히 이익을 얻을 수 있다고 판정된 후에 채산집단을 대상으로 본격적인 마케팅을 제공해야 한다는 접근법이다.

RFM 분석이 가능하기 위해서는 고객 1인 1인을 대상으로 삼아 각각의 고객이 언제, 무엇을 얼마나 구입하였는가에 관한 데이터베이스를 파악하여야 하며 고객 데이터베이스는 R, F, M의 관점에서 포착하고 고객 각각의 반응률 성향과 구매행동 성향을 수치화 할 수 있어야 한다.

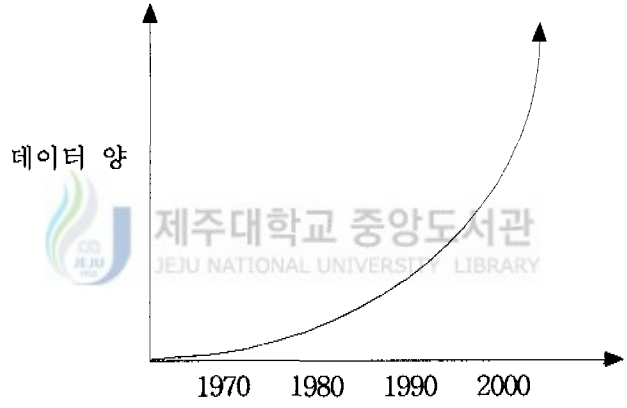
결론적으로 기업이 고객데이터베이스를 가지고 RFM 분석에 의해 고객을 세분화하여 유의한 고객세분계층에 대한 마케팅활동을 기획한다면 그 기업의 이익도 보장 가능하다. 세분화는 단순히 고객을 복수의 집단으로 세분화하는 것이 아니라 채산성에 따라 고객집단을 구분하고 채산성 있는 고객에게만 마케팅노력을 집중함으로써 수익성 증대를 도모한다는 발상에 입각한 마케팅기법이다.

제2절 데이터 마이닝(Data Mining)의 개념

1. 지식발견(Knowledge discovery)과 데이터 마이닝

1998년 6월에 발표한 팔로알토 매니지먼트 그룹의 보고서에 따르면 1997년의 데이터 웨어하우스 규모는 평균 272 기가 바이트(Giga Byte)였으며, 향후 3년간 24배 가량 성장하여 2000년말이 되면 6.5 테라 바이트(Tera Byte) 규모로 늘어날 것이라고 한다.[15]

<그림 2-1> 데이터 양의 증가 추세



자료 : Ruth Dilly, "Data Mining An Introduction Student Note", The Queen's University of Belfast, Version 2.0, 1995.12. [http://www.qub.ac.uk/]

기존의 데이터베이스 관리 시스템이 취급하고 이에 저장되는 데이터의 양이 증가함에 따라서 저장된 자료로부터 원하는 정보를 효과적으로 추출하거나 데이터에 내재된 규칙들을 찾는 일은 대단히 어렵게 되고 있다.[16] 몇 기가 바이트 혹은 테

15) 한국썬마이크로시스템즈, "한국썬소식", 1998.11월호
16) Ming-Syan Chen, Jiawei Han and Philip S. Yu, "Data Mining : An Overview from Database Perspective", IEEE Trans. on Knowledge and Data Engineering, VOL. 8, No. 6, December 1996, pp.866~883.

라 바이트에 이르는 거대한 양의 데이터를 분석하기란 쉽지 않기 때문에 우리나라의 많은 은행들은 <표 2-1>에 제시되어 있는 바와 같이 단순한 통계 정보만을 얻는데 그치거나 아예 분석할 엄두조차 내지 못하고, 이 귀중한 자산을 제대로 이용하지 못하고 있다.

<표 2-1> 우리나라 은행의 정보계 DB의 활용 형태

활용 형태	은행 수	비 고
단순조회 및 리스트 추출	17개	정보계의 DB를 보유하고 있는 20개 은행의 중복응답 분석결과
우량고객의 선별	16개	
시장세분화	6개	
고객구매행동 분석	2개	

자료 : 박찬욱, 전게서, p.56.

2000년대가 되어 데이터 양이 더욱 방대해지면, 사용자에 의해 일일이 데이터가 분석되어질 수 없다. 이 때가 되면 데이터 마이닝 기술을 도입하여 컴퓨터에 의한 지식발견(데이터베이스 안에 내재되어 있는 패턴이나 추세, 또는 데이터간의 상관관계)으로 분석방식의 일대 전환이 예상된다.[17]

한 조직이 데이터를 아무리 잘 수집하고 조직화하여 고객 데이터베이스나 데이터 웨어하우스를 구축하였다 하더라도 단순히 이를 저장하는 수준으로는 조직의 경쟁력 강화와 이익 창출에 아무런 도움이 되지 않는다. 구축되어 있는 고객 데이터베이스나 데이터 웨어하우스에서 유용한 정보나 새로운 지식을 발굴하여 활용할 수 있는 수단이 제공되어야만 다양한 이익을 창출하는 완벽한 고객 데이터베이스라 할 수 있다. 이런 상황에서 데이터 마이닝은 방대한 규모의 데이터베이스로부터 숨겨진 지식, 예상치 않았던 패턴 및 규칙 등을 추출하는 가장 좋은 수단 가운데 하나로 인식되고 있다.[18]

17) 조재희, "데이터웨어하우징 기술을 이용한 DB마케팅 전략에 관한 연구", 정보기술과 데이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4, p.107.
18) 김신곤, "데이터마이닝 기법(CHAID)을 이용한 효과적인 데이터베이스 마케팅에 관한 연구", 정보기술과 데이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4, p.90.

1) 지식발견 프로세스

KDD(Knowledge Discovery in Databases)와 데이터 마이닝(Data Mining)은 명확하게 구분되지 않으며, 일반적으로 같은 의미로 사용된다. 여기서는 하위수준의 데이터를 상위수준의 지식(knowledge)으로 변환시키는 전반적인 프로세스를 KDD라고 간주한다. KDD는 "타당하고, 새롭고, 잠재적으로 유용한 그리고 궁극적으로 이해하기 쉬운 패턴을 데이터에서 확인해 내는 프로세스"로 정의할 수 있다. 그러므로 지식은 새롭고, 쉽게 알 수 없으며 인간이 활용할 수 있어야 한다. 흔히 사용되는 데이터 마이닝의 정의, 관찰된 데이터로부터 패턴이나 모델을 추출하는 것 역시 KDD의 정의로 볼 수 있다. 그러나, 이러한 데이터 마이닝의 정의는 지식발견 프로세스의 핵심적 과정들에서 수행되는 노력의 작은 부분(15 ~ 25%로 추정됨)에 불과하다. 즉, 데이터 마이닝을 전반적인 지식발견 프로세스의 한 단계로 보는 것이 더욱 타당하다. KDD의 다른 단계들을 예로 들면 다음과 같다:[19]

- 데이터 마이닝 프로세스를 적용할 수 있는 영역과 목적에 대한 이해 증진
- 대상이 되는 데이터 집합(data set)의 선정과 수집
- 데이터 집합(data set)의 통합(integrating)과 검토(checking)
- 데이터의 정제(cleaning), 사전처리(preprocessing) 및 변환
- 모델 개발 및 가설 구축
- 적절한 데이터 마이닝 알고리즘 선정
- 결과 해석(interpretation) 및 시각적(visualization) 제시
- 결과 검증(testing) 및 타당화(verification)
- 발견된 지식의 활용(using)과 유지보수(maintaining)

지식발견 프로세스는 비선형이며, 반복적이고 상호 연관을 갖고 있다. 어떤 한 단계는 이전 단계들의 변형이므로 다양한 피드백(feedback) 순환고리를 형성할 수

19) Michael Goebel, Le Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", ACM SIGKDD, June 1999, p.21.

있다. 따라서 핵심적인 데이터 마이닝 단계뿐만 아니라 전체 지식발견 프로세스를 지원할 수 있는 도구의 개발이 촉진된다. 그러한 도구는 데이터 선정, 사전처리, 통합, 변환 등을 위해 데이터베이스 시스템이나 데이터 웨어하우스와의 밀접한 통합을 필요로 한다.

2) 지식발견과 데이터베이스

현재 이용 가능한 많은 지식발견 도구들은 대체로 인공지능(AI)이나 통계학 분야로부터 개발된 것들이다. 그러한 도구들은 일반적으로 데이터 소스와 분리되어 작동하며, 데이터를 내보내고 불러들이는데, 사전 및 사후 처리에 그리고 데이터 변환에 엄청나게 많은 시간이 소요된다. 따라서 현재의 DBMS 지원체계를 이용할 수 있도록 지식발견 도구와 데이터베이스간의 밀접한 통합이 요구된다.

지식발견 도구를 검토하고자 할 때는 다음의 특성들을 고려할 필요가 있다.[20]

① 다양한 데이터 소스의 채택 가능성 : 많은 경우, 분석될 데이터는 회사 여기저기에 흩어져 있으며, 의미 있는 분석이 이루어지기 전에 데이터는 취합되고, 검토되고, 통합되어져야 한다. 서로 다른 데이터 소스를 곧바로 채택할 수 있는 기능은 데이터 변환에 소요되는 시간을 엄청나게 줄여줄 수 있다.

② 온라인/오프라인 데이터 접근(Online/Offline data access) : 온라인/오프라인 데이터 접근이 의미하는 것은 실행조건들이 데이터베이스에서 곧바로 처리되며, 다른 업무들과 동시에 처리될 수 있다는 것이다. 오프라인 데이터 접근(Offline data access)에서는 데이터 소스의 일부분으로 분석이 이루어지며, 많은 경우 데이터 소스로부터 지식발견 도구에 의해 요구되는 데이터 포맷으로의 내보내기/불러오기 과정을 거친다. 온라인(Online) 혹은 오프라인(Offline)여부는 계속 변하는 지식과 데이터를 다룰 때 특히 중요해 진다. 예를 들어 금융시장처럼 급속하게 변하는 시장 환경에서는 이전에 발견된 규칙이나 패턴들의 타당성은 낮을 것이므로 온라인 접근(Online access)이 더욱 중요하게 될 것이다.

---

20) *Ibid*, pp.21~22.

③ 데이터 모델 : 오늘날 이용 가능한 많은 도구들은 입력 값을 하나의 테이블 형태로 취하며, 테이블에서 각 사례(record)에는 일정한 개수의 속성 값들이 있다. 다른 도구들은 관계 모델(relational model)에 근거하며, 데이터베이스에 직접 접근하는 것을 허용한다. 객체 지향적이고 비표준적 데이터 모델은 대부분 현재 KDD 기법의 범위밖에 있다.

④ 테이블/행/속성의 최대 개수 : 이는 지식발견 도구의 처리 능력에 관한 이론적 제한점이다.

⑤ 도구가 처리할 수 있는 데이터베이스 크기 : 분석될 데이터의 예상 크기는 분석 도구를 선택하는 데 있어서 중요한 요인이다. 테이블/행/속성들의 최대 개수는 이론적인 제한점인 반면, 계산시간, 메모리, 구현능력 등에 의해 제기되는 실제적인 제한점도 있다. 예를 들면 모든 데이터를 주 메모리에 보유하는 도구는, 비록 이론적으로 최대 행의 개수는 제한되지 않을 지라도 매우 큰 데이터 소스에는 부적절할 지도 모른다.

⑥ 도구가 처리할 수 있는 속성 유형 : 어떤 분석 도구는 입력 데이터의 속성 유형에 있어서 한계가 있다. 예를 들면, 신경망에 기초한 도구는 대개 모든 속성들이 숫자형(numeric)일 것을 요구한다. 다른 접근 방법들은 연속(continuous) 데이터를 처리하지 못할 수도 있다. 따라서 분석 도구를 선택하는데 있어서, 데이터 소스에서 나타나는 속성 유형들이 고려되어야 한다.

⑦ 질의어(Query language) : 질의어(Query language)는 사용자와 데이터베이스 간의 인터페이스로 작용한다. 사용자가 데이터를 처리할 수 있게 해 주고, 지식발견 프로세스의 절차를 지시할 수 있게 해 준다. 어떤 도구들은 질의어를 갖고 있지 않으며 사용자와 도구와의 상호작용은 프로세스 파라미터를 지정하는 것에 제한된다. 다른 도구들은 SQL이나 다른 질의어로 형식화된 질문들을 통해 데이터나 지식에 대해 실행조건을 설정하도록 해 준다. 데이터와 지식에 대한 조건 설정은 그

래픽 사용자 인터페이스(GUI)를 통해 이루어질 수도 있다.

지금까지 언급된 도구들은 특성들에서 서로 다를 수 있다. 따라서 이용 가능한 데이터의 형식과 크기, 지식발견 프로세스의 목적, 최종 사용자의 욕구와 훈련 정도 등을 고려하여 도구를 선택해야 한다.

2. 데이터 마이닝의 정의 및 목적

1) 데이터 마이닝의 정의

우선 데이터 마이닝이 무엇인지 이해하기 위해 대표적인 정의들을 살펴보면 다음과 같다.[21]

- 대량의 실제 데이터로부터 묵시적이고 전에는 알려지지 않았지만 잠재적으로 유용한 정보를 추출하는 것
- 대규모 데이터베이스 내에 존재하는, 그러나 대량의 데이터 사이에 숨겨져 있는 상호 관련성(relationship)과 글로벌 패턴(pattern)에 대한 탐색
- 대량의 데이터로부터 패턴 인식기술과 통계기법, 수학적 기법을 이용하여 의미있는 새로운 상관관계(correlation), 패턴 그리고 추세(trends)를 발견하는 과정

데이터 마이닝이란 자동화된 지능을 갖춘(automated and intelligent) 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD: Knowledge Discovery in Databases), 정보발견(information discovery), 정보수확(information harvesting) 등의 이름으로도 소개되어 왔는데 일반적으로 '대량의 데이터로부터 새롭고 의미있는 정보를 추출하여 의사결정에 활용하는 작업' 이라 정의된다. 용어에 '채굴하다'라는 의미의 'mining'을 포함시킨 이유는 데이터로부터 정보를 찾아내는 작업이 마치 금

21) 한국컴퓨터연구조합, 주전산기산학연협의회, "데이터웨어하우스", 타이컴월드 98/11 - 12 제22호, 1998, p.34.

이나 다이아몬드를 발견하기 전에 수많은 양의 흙과 잡석들을 파헤치고 제거하는 것과 유사하다는 데에 기인한다.[22]

1995년 캐나다 몬트리얼에서 개최된 지식발견과 데이터 마이닝에 관한 국제 학술대회(The first international conference on knowledge discovery & datamining)에서 지식발견은 데이터로부터 유용한 정보를 발견하는 프로세스의 전 과정이라 정의하였고, 데이터 마이닝은 지식발견 프로세스 중에서 데이터로부터 정보를 추출하기 위해서 기법을 적용하는 특정 단계라 제안하였다.

기업체는 최신 통계 분석과 정교한 고객 데이터베이스, 병렬처리 등을 종합적으로 활용하여 고객과의 관계를 최적화 시키고 사업을 확장한다. 이러한 기술의 집합을, 즉 핵심적인 전략적 가치로서 정보를 사용하여 이익을 증대시키는 하나의 과정을 데이터 마이닝이라고 데이터베이스 마케팅 회사인 테세라 엔터프라이즈 시스템사(Tessera Enterprise Systems)의 기술부장인 폴바스(Paul Barth)는 정의하였다.[23] 또한 박찬욱 교수는 '수집된 대용량의 데이터베이스를 가공하여 기업경영에 결정적인 영향을 미칠 수 있는 지식을 발견하기 위한 일련의 작업과정'으로 정의하였다.[24]

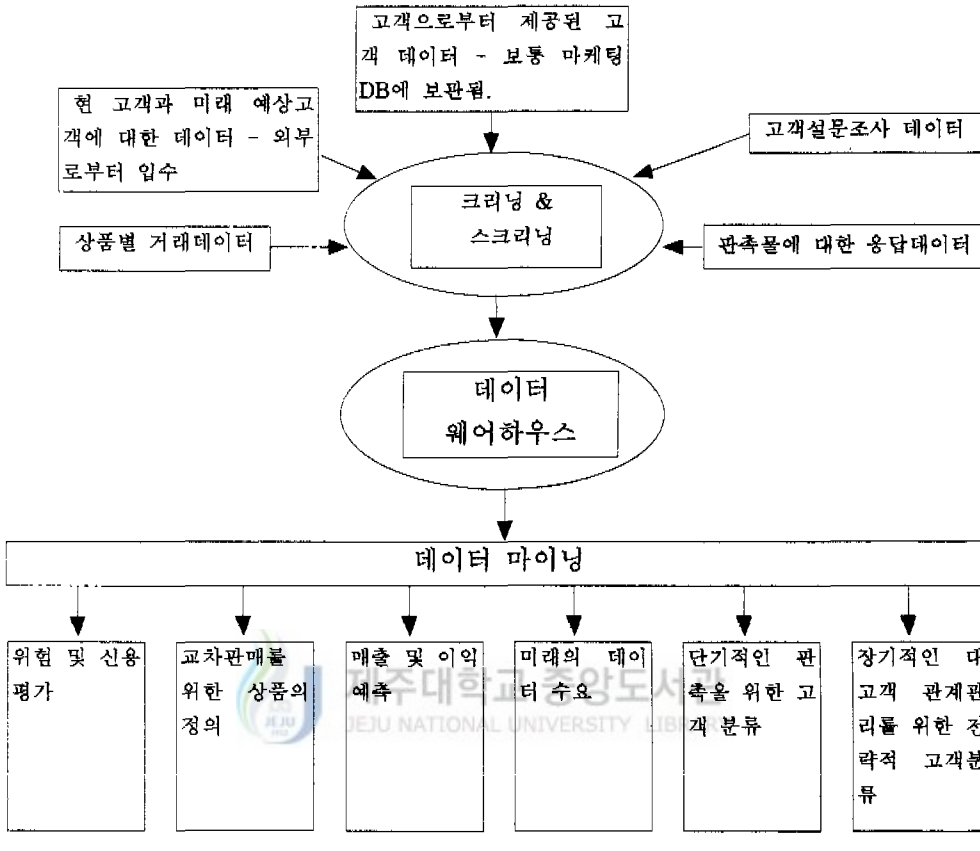
이처럼 데이터 마이닝은 다양하게 정의할 수 있지만 일반적인 개념으로는 기업이 보유하고 있는 일일 거래데이터, 고객 데이터, 상품 데이터 혹은 각종 마케팅 활동의 고객 반응 데이터 등과 기타 외부 데이터를 포함하는 모든 사용 가능한 근원 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙(Rule) 등을 발견하고, 이를 실제 비즈니스 의사결정 등을 위한 정보의 활용 및 분석을 지원하는 일련의 과정이라 할 수 있다.

일반질의나 OLAP 도구 등의 기존 조회 방식과는 달리 데이터 마이닝은 데이터에 숨겨져 있는 정보를 찾아내는데 사용된다. 잘 설계된 데이터 마이닝 도구는 데이터로부터 짧은 시간 내에 가능한 한 다수의 유용한 가설을 산출해내는 방식으로 정보를 발견하도록 설계되어 있다.[25]

---

22) 장남식 · 홍성완 · 장재호, 「데이터 마이닝」, 대청미디어, 1999. 10, p.26.
23) Ramon C. Barquin외 14 著, 함문성, 김석호 譯, 「데이터웨어하우스(Ⅱ)」, 도서출판 니드, 1999, p.127.
24) 박찬욱, 전게서, 1999, p.120
25) 장남식 · 홍성완 · 장재호, 전게서, p.30.

<그림 2-2> 데이터 마이닝의 흐름



자료 : 정철용 · 함유근, 「고객정보시스템 구축 및 활용 전략」, 한국금융연구원, 1999, p.12

데이터베이스로부터 찾고자 하는 것이 무엇인가를 정확히 알고 있는 경우에는 SQL 등의 조회도구를 사용하지만 오직 찾고자 하는 것을 윤곽만 알고 있는 경우에는 데이터 마이닝을 활용해야 한다. 데이터 마이닝은 기존의 조회 도구를 대처하는 것이 아니라 보완하는 기능을 제공하는 것이다. 하지만 데이터 마이닝을 통해 추가적으로 얻게되는 정보의 가치는 그야말로 무한하다.

데이터 마이닝은 데이터 웨어하우징이라는 또 다른 중요한 개발과 밀접하게 연관되어 있다. 데이터 마이닝을 위해서는 잘 정리된 많은 양의 자료가 필요하며, 따라서 데이터 웨어하우스는 최적이라고 말할 수 있다. 분석가는 데이터 웨어하우스

가 구축되어 있으면, 데이터를 정제·통합·보충하는데 신경 쓸 필요 없이 마이닝 작업에 전념할 수 있다.[26] 실제적으로 데이터 마이닝을 위하여 데이터베이스를 재정리하고 조작하는 작업은 데이터 웨어하우스에서 행하는 작업과 거의 유사하기 때문에 두 가지 프로젝트를 동시에 수행할 수도 있을 것이다.

데이터 마이닝을 통해 얻을 수 있는 정보의 형태는 매우 다양하며, 이에 따라 다양한 기법이 존재한다. 가장 대표적인 데이터 마이닝 기법으로는 사건들의 연관성(associations) 탐사, 연속성(sequence) 탐사, 분류(classifications) 규칙 탐사와 군집구분(clustering)을 들 수 있으며, 핵심 알고리즘으로 귀납적 추론(rule induction)이나 신경망(neural network)과 같은 인공지능(artificial intelligence)등을 응용한 방법이 소개되면서 데이터 마이닝이 더욱 활성화되기 시작하였다.

2) 데이터 마이닝의 목적과 기본적 기능

데이터 마이닝은 데이터로부터 패턴을 추출하기 위한 KDD 프로세스의 핵심적 과정이며, KDD 프로세스에 따라 다른 목적을 갖는 데이터 마이닝 방법들이 연속적으로 적용될 수 있다. 예를 들면 신제품을 구입할 가능성이 큰 고객들을 확인하기 위해, 고객 데이터베이스를 세분화하는 클러스터링 방법을 먼저 적용하고, 각 클러스터의 구매행동을 예측하기 위해 회귀분석(regression)을 적용할 수 있다.

데이터 마이닝의 목적은 사용자에게 유용한 새로운 지식을 생성하는 것으로 현실세계에 적합한 모형의 수립이 전제되어야 한다. 따라서 모형수립 과정에서는 기업의 거래처리 자료, 고객 이력 및 신상 자료, 및 신용정보기관에 의해 제공되는 고객신용정보와 같은 외부자료 등 다양한 자료들을 사용되어 진다. 수립된 모형은 데이터에 내재된 패턴이나 관계를 설명해 주는 것으로 다음의 두 가지 목적으로 사용된다.[27]

---

26) Inmon, W. H., "The Data Warehouse and Data Mining", Communications of ACM, Vol. 39, No. 11, November 1996.

27) 정보통신부, "데이터 웨어하우스 기반의 Data Mining 소프트웨어 개발", 1997, p.48.

(1) 묘사(Description) : DB에 내재된 패턴이나 관계를 묘사한 것만으로도 의사결정에 필요한 정보를 제공할 수 있다. 예를 들어 슈퍼마켓의 취급상품들에 대한 소비자의 구매패턴은 연관규칙에 의해 묘사될 수 있는데 이는 상품의 진열장소를 결정하는데 매우 유용하다.

(2) 예측(Prediction) : 발견된 패턴들은 예측에 이용될 수 있다. 만약, 우편광고시에 특정유형의 홍보물에 반응을 잘 보이는 고객을 예측할 수 있는 패턴을 알고 있다면 우편물을 해당 고객들에게만 집중시킴으로써 적은 비용으로 높은 광고 효과를 얻을 수 있을 것이다.

따라서, 묘사는 데이터베이스에 내재된 고객의 거래패턴이나 관계를 나타내는 것으로 예를 들면, 고객만족에 있어 어떠한 속성이 중요한 영향을 미치며, 고객의 서비스 이용이 어떠한 패턴으로 작용하는지를 묘사해 주는 기능을 말한다. 예측은 데이터 마이닝을 통해 발견된 지식 또는 패턴으로 향후 고객들에 대한 성향을 미리 알아볼 수 있음을 말한다. 즉 수익성이 높은 고객집단과 낮은 고객집단의 예측결과를 통해 이들 고객의 욕구(needs)를 파악함으로써 개별집단에 맞는 마케팅 전략을 수립할 수 있게 된다.

이와 같은 데이터 마이닝의 목적은 대부분 다음의 범주에 포함된다:[28]

① 데이터처리(Data Processing) : KDD 프로세스의 목적과 요구사항에 따라 분석가는 데이터를 선택하고, 여과하고, 통합하고, 표집하고, 정제 그리고 변환한다. 몇 가지 가장 전형적인 데이터 처리 과정을 자동화하거나 그러한 과정들을 전체적인 프로세스로 통합 처리함으로써 업무의 양을 상당부분 줄일 수 있다.

② 예측(Prediction) : 데이터와 예측 모델이 있을 때, 데이터의 특정 속성의 값을 예측한다. 예를 들면, 신용카드 거래의 예측 모델에서 특정한 거래가 사기일 가능성을 예측한다. 예측은 발견된 가설을 검증하는데 사용되기도 한다.

---

28) Michael Goebel, Le Gruenwald, *op. cit.*, p.22.

③ 회귀분석(Regression) : 데이터 항목들의 집합이 있을 때 회귀분석(regression)은 어떤 속성이 다른 속성들에 종속해서 발생할 수 있는 관계성을 분석하여, 새로운 레코드에 대해서 이러한 속성 값을 예측할 수 있는 모델을 산출한다. 예를 들면, 신용카드 거래에서 새롭게 이루어지는 거래가 사기일 가능성을 예측할 수 있는 모델을 생성한다.

④ 분류(Classification) : 일련의 범주들이 사전에 분류되어 있을 때, 특정한 데이터 항목이 이러한 분류체계 중 어디에 속하는 가를 결정한다. 예를 들면, 의료처치의 반응에 따라 환자들이 분류되어 있을 때, 새로운 환자가 가장 잘 반응할 것 같은 처치의 형태를 밝혀 낸다.

⑤ 군집(Clustering) : 일련의 데이터 항목들이 있을 때, 유사한 특성을 갖는 항목들을 함께 묶는다. 클러스터링은 유사한 항목들의 집단을 찾아내는데 가장 잘 사용된다. 예를 들어, 고객들의 데이터 집합에서 유사한 구매행동을 갖는 고객들의 하위집단을 밝혀 낸다.

⑥ 연관분석(Link Analysis(Associations)) : 한 패턴의 출현이 다른 패턴의 출현을 암시하는 속성이나 항목들간의 관계성을 밝혀 낸다. 이러한 관계는 동일한 데이터 항목 내에서 속성들간의 연관성(밀크 구매자 중 64%는 빵을 함께 구입)일 수도 있으며, 다른 데이터 항목들간의 연관성(어떤 주식이 5% 하락할 때마다 다른 주식은 2~6주 후에 13% 오른다)일 수도 있다. 일정 기간에 걸쳐 발생하는 항목들간의 관계성은 '순차 패턴 분석(sequential pattern analysis)'으로도 불린다.

⑦ 모델 시각화(Model Visualization) : 시각화는 발견된 지식을 이해 가능하게 하고 해석 가능하게 만드는 데 있어서 중요한 역할을 한다. 게다가, 인간의 시각-뇌 시스템은 지금까지 알려진 가장 훌륭한 형태인식 도구이다. 시각화 기법은 단순한 산포도와 막대그래프로부터 3차원 영상에 이르기까지 다양하다.

⑧ 탐색적 데이터 분석(Exploratory Data Analysis : EDA) : 탐색적 데이터 분석(EDA)은 사전에 설정된 가정이나 모델에 크게 의존하지 않고 데이터를 탐색하여 흥미 있는 패턴들을 밝히려고 시도한다. 시각과 직관을 위해 데이터의 그래픽 표현이 흔히 사용된다. 데이터 탐색을 지원하기 위해 개발된 수많은 소프트웨어들이 있지만 시각화를 전체적인 KDD환경으로 통합하는 것이 바람직할 수 있다.

데이터 마이닝의 기본 기능을 사용되는 모형의 형태에 따라 다시 구분하면, 묘사의 경우는 군집화(clustering), 연관규칙 및 순차패턴 탐사(association and sequence pattern discovery)로 나누어지며, 예측의 경우는 분류(classification), 회귀분석(regression) 및 시계열 분석(time series analysis)등이 대표적 모형들이다.

데이터 마이닝에서 하나의 모형수립을 수립하기 위해 사용될 수 있는 기술 또는 알고리즘은 다양하게 존재하며 이러한 기술들은 여러 학문 분야로부터 발전되어 왔다. 대부분의 경영문제에 있어서는 하나의 모형에 대해서도 다양한 알고리즘들을 사용하는 것이 필요하다는 것이다. 이는 주어진 문제의 유형과 데이터의 성격에 따라 모형이나 알고리즘의 성과가 달라지며, 이들을 적용해 보기 전에는 최선의 모형이나 알고리즘을 알기 어렵기 때문이다.[29]

3) 데이터 마이닝의 특징

데이터 마이닝은 몇 가지 관점에서 특징을 가지고 있는데, 이를 간단히 요약하면 다음과 같다.[30]

① 데이터 마이닝은 대용량(massive)의 관측 가능한 자료(observational data)를 다룬다. 실험자료는 가설검정 등의 구체적인 문제에 답하기 위해 여러 요인들이 통제되고 조작된 가운데 만들어진다. 그러나 관측자료는 시간의 흐름에 따라서 비계획적으로 축적되며, 자료분석을 염두에 두고 수집되지 않는 것이 일반적이다.

---

29) 정보통신부, 전게서. pp.48~49.
30) 강현철 외5, 「SAS Enterprise Miner를 이용한 데이터마이닝 방법론 및 활용」, 자유아카데미, 1999, pp.6~7.

② 데이터 마이닝은 컴퓨터 중심적 기법(computer-intensive method)이다. 현대의 컴퓨터 중심적 기법들은 기존의 기법들로서는 해결하기 곤란한 경우에 있어, 이를 해결하기 위하여 컴퓨터의 강력한 처리속도와 능력을 활용할 수 있도록 해주고 있다.

③ 데이터 마이닝은 경험적 방법(adhockery method)에 근거하고 있다. 많은 데이터 마이닝 기법들은 이론적 원리에 기초하여 개발되었다기보다는 경험에 기초하여 개발되었다. 이러한 기법들은 그 특성이 수리적으로 밝혀지지 않은 것들이 많다.

④ 데이터 마이닝은 일반화(generalization)에 초점을 두고 있다. 일반화는 예측모형(prediction model)이 새로운 자료(new data)에 얼마나 잘 적용되도록 하는 것인가를 의미한다. 따라서 일반화는 데이터 마이닝 기법의 비정형성을 어느 정도 해결 또는 보완하여 주는데 도움을 주고 있다. 추론(inference)이 통계적 모형들의 초점이라고 하면, 일반화는 데이터 마이닝의 초점이라 할 수 있다.

⑤ 데이터 마이닝은 기업의 다양한 경영상황 하에서 경쟁력 확보를 위한 의사결정을 지원하기 위해서 활용(business applications)될 수 있다.

⑥ 데이터 마이닝 기법들은 통계학(statistics), 컴퓨터과학(computer science), 인공지능(artificial intelligence : AI), 공학(engineering)과 같은 분야에서 개발된 특징을 가지고 있다. 그러나 실제로 이를 활용하는 전문가들은 경영, 경제, 정보기술(information technology : IT)분야에 있는 사람들이다.

제3절 데이터 마이닝(Data Mining)의 활용

1. 데이터 마이닝 방법론

데이터 마이닝은 하나의 기법이 아니라 유용한 데이터로부터 더 많은 정보를 추출하기 위한 여러 가지 방법들을 지칭한다. 목적에 따라 서로 다른 방법들이 적용되며 각 방법은 고유한 장단점을 갖고 있다. 데이터 마이닝에 사용되는 대부분의 방법은 일반적으로 다음과 같이 분류될 수 있다.[31]

① 통계적 방법(Statistical Methods) : 역사적으로 통계작업은 주로 사전에 설정된 가설을 검증하고 모델을 데이터에 적합화 시키는 데 초점을 두어 왔다. 통계적 접근은 일반적으로 명백한 확률모델에 의존한다. 또한 이러한 방법들은 통계학자에 의해 사용될 것이라고 가정되며, 따라서 사람이 수행하는 역할은 가설을 설정하고 모델을 추출하는 것으로 가정되어 왔다.

② 사례기반추론(Case-Based Reasoning) : 사례기반추론(CBR)은 과거 경험과 해법(solutions)을 직접적으로 활용함으로써 주어진 문제를 해결하고자 시도하는 기법이다. 하나의 사례(case)는 대개 과거에 부닥쳤고 해결된 특정한 문제점이다. 새로운 문제점이 나타나면 CBR은 저장된 사례들을 점검하고 유사한 것들을 찾아낸다. 만일 유사한 사례들이 존재한다면, 그것들의 해법이 새로운 문제에 적용된다. 그리고 그 문제는 향후의 참조를 위해 사례집에 추가된다.

③ 신경망(Neural Networks) : 신경망(NN)은 인간의 대뇌를 모방한 일종의 시스템이다. 인간의 대뇌가 시냅스에 의해 상호 연결된 수백만의 뉴런들로 구성된 것처럼, 신경망은 수많은 유사 뉴런들로 형성되며, 이러한 뉴런들은 대뇌 뉴런들과 유사한 방식으로 서로 연결된다. 인간의 대뇌에서처럼, 뉴런 상호연결의 강도는 제시된 자극 혹은 획득된 출력에 대한 반응으로 변할 수도 있다(혹은 학습 알고리즘에 의해 변할 수도 있다). 이러한 변화가 네트웍으로 하여금 학습하게 한다.

④ 의사결정나무(Decision Trees) : 의사결정나무에서 각 마디는 고려되고 있는 데이터 항목에 대한 검증 혹은 결정을 나타낸다. 검증의 결과에 따라 특정한 가지

31) Michael Goebel, Le Gruenwald, *op. cit.*, p.23.

를 선택한다. 특별한 데이터 항목을 분류하기 위해서는 뿌리노드(root node)에서 출발하여 종단노드(terminal node)에 이를 때까지 진행한다. 종단노드에 이르면 결정이 이루어진다. 의사결정나무는 특별한 형태의 규칙 집합-위계적인 규칙들의 조직으로 특징되는-으로 해석될 수도 있다.

⑤ 추론규칙(Rule Induction) : 규칙(rules)은 하나의 데이터 항목에서 발생하는 속성들간의 통계적 상관성 혹은 하나의 데이터 집합에서 데이터 항목들간의 통계적 상관성을 의미한다. 일반적인 형태의 연관성 규칙(association rule)은 $Xl \wedge \cdots \wedge Xn \Rightarrow Y[C,S]$이며, 이는 속성 $Xl,\ldots,Xn$은 신뢰도(confidence) C와 유의도(significance) S로 Y를 예측한다는 것을 의미한다.

⑥ 유전자알고리즘/진화프로그래밍(Genetic algorithm/Evolutionary programming) : 유전자알고리즘/진화프로그래밍은 자연진화에서 관찰되는 원리에 의해 촉발된 알고리즘적 최적화 전략이다. 서로 경쟁하는 잠재적인 문제 솔루션들 중에서 가장 좋은 솔루션들이 선정되고 서로 조합된다. 그렇게 함으로써 솔루션들의 전반적인 유익성이 더 좋아질 것이라고 기대한다. 이는 유기체들의 진화과정과 유사하다. 유전자알고리즘/진화프로그래밍은 변인들간의 종속성에 대한 가설을 형식화하기 위해 연관성규칙이나 다른 내적 형식화의 형태로 데이터 마이닝에서 사용된다.

⑦ 퍼지집합(Fuzzy Sets) : 퍼지집합은 불확실성을 표현하고 처리하는데 핵심적인 방법론이다. 불확실성은 오늘날 데이터베이스에서 부정확, 비구체성, 비일관성, 모호함 등 다양한 형태로 발생한다. 퍼지집합(Fuzzy sets)은 시스템의 복잡성을 다루기 위해 불확실성을 탐색하며, 불완전하고 깨끗하지 않은 혹은 부정확한 데이터를 다룰 뿐만 아니라, 전통적 시스템보다 더 현명하고 유연한 모델을 개발할 수 있다. 즉, 퍼지시스템은 정확한 입력이 불가능하거나 비경제적인 상황에서 강력한 예측모델을 제공할 수도 있다.

2. 데이터 마이닝의 활용 및 관련 분야

1) 데이터 마이닝의 활용분야

데이터 마이닝은 <표 2-2>에서와 같이 다양한 산업에서 적용되고 있다.

<표 2-2> 데이터 마이닝의 응용분야 및 적용사례

분야	적용사례
소매/마케팅	- 고객의 구매패턴과 선호도 발견 - DM(Direct Mail)에 응답할 가능성이 높은 고객예측 - 제품/서비스 교차 판매 - 판매실적에 영향을 미치는 요소발견 - 고객분류, 그룹별 특성발견 - 광고, 프로모션, 이벤트의 효과 측정
은행, 카드	- 신용카드 도용패턴 추적 - 이탈 예상고객 선정 및 특성분석 - 우수고객 선정 및 특성분석 - 서비스별 홍보 대상고객 선정 - 신용평가 모형 개발 - 주식 거래규칙 발견
보험	- 고객분류를 통한 보험료 가격 정책 수립 - 보험료 청구 사기 패턴 추적 - 클래임 처리시간에 영향을 미치는 요소발견
통신	- 장거리/무선 전화의 부정한 이용패턴 추적 - 이탈 예상고객 선정 및 특성분석 - 서비스간의 연관관계 발견 - 우수고객 선정 및 특성분석
제조	- 최종 생산품의 품질에 영향을 미치는 요인 발견 - 경쟁사의 입찰액 예측 - 제품의 수요 예측 - 대리점 여신평가 모형 개발
유통	- 매장진열 전략 수립 - 상품 카달로그 디자인 - 상품 교차판매
의료	- 환자의 질병 진단이나 질병의 예후 분석 - 환자의 특성에 따른 의약품의 부작용 분석

자료 : 장남식외3, 전게서, p.40.

국내의 경우 데이터 마이닝은 아직 개념이나 기법, 제품 등이 소개되는 초기 도입 단계이다. 그러나 외국의 많은 선진 기업들을 통해 활용사례들을 접할 수 있는데, 재고유지부터 복합적인 기술을 요구하는 주식예측까지 매우 다양한 비즈니스 영역에서 사용할 수 있다. 기존의 의사결정 시스템이 기업의 내부 문제를 풀어왔다면, 데이터 마이닝은 판매점에서 고객의 구매성향을 파악하는데 쓰이거나 고객의 일반적인 성향을 파악해 고객 이탈 방지 프로그램 등에 직접 활용되면서 고객 행동 분석이 데이터 마이닝의 주요 적용대상으로 자리잡고 있다. 데이터 마이닝은 대용량의 데이터 웨어하우스를 기반으로 실행되어야 한다. 현재 데이터 마이닝이 제 능력을 발휘하기 쉬운 분야로 마케팅, 은행, 보험, 소매, 통신분야 등이 꼽히는 이유도 대용량의 자료가 구축되어 있기 때문이다.

미국은 이러한 측면에서 분명히 유럽에 앞서 있는 데 아메리칸 익스프레스나 AT&T와 같은 대규모 회사는 이미 고객의 파일을 분석하는 데 KDD를 사용하고 있다. 영국의 경우는 BBC에서 시청률 조사를 위해 데이터 마이닝 기법을 응용하고 있다. 대부분의 유럽국가에서는 대규모 은행이나 보험회사들이 잠정적으로 KDD에 대한 기초 연구를 하고 있는 중이다.[32]

데이터 마이닝은 위에서 언급한 특정 업종에만 국한된 것이 아닌 모든 분야에 적용할 수 있는 수평적 응용기술이다. 데이터 마이닝이 제공하는 정보들은 효과적인 광고전략 개발이나 상품배치, 목표고객 선정, 프로세스의 개선 등에 활용됨으로써 전체적인 비즈니스 효율 향상과 비용 절감을 꾀할 수 있다. 또한 데이터 마이닝에서 발굴되는 정보를 통해 기업은 선점 효과를 누릴 수 있는 경우가 많다.[33]

데이터 마이닝은 정의의 다양성만큼 활용분야도 매우 다양하다고 할 수 있다. 특히 기업의 의사결정 문제에서 많이 활용되고 있는데, 주요 활용분야를 나열하면 다음과 같다.[34]

(1) 데이터베이스 마케팅

데이터베이스 마케팅(Database Marketing)은 데이터 마이닝이 가장 성공적으로

32) Pieter Adriaans, Dolf Zantinge 著・용환승 譯, 「데이터 마이닝」, 그린, 1998, p.14.
33) 장남식・홍성완・장재호, 전게서, pp.39~41.
34) 강현철 외5, 전게서, pp.5~6.

적용되고 있는 분야 중의 하나이며, 목표 마케팅(Target Marketing), 고객세분화(Segmentation), 고객성향변동분석(Churn Analysis), 교차 판매(Cross Selling), 시장바구니 분석(Basket Market Analysis)등에서 주로 이용된다. 데이터 베이스 마케팅은 소매, 통신판매, 금융서비스, 건강, 보험, 통신, 운송, 제약 등 다양한 분야에서 활용되고 있다.

(2) 신용평가

신용평가(Credit Scoring)는 특정인의 신용거래 대출한도를 결정하는 것이 주 업무로써 목적은 불량채권과 대손을 추정하여 이를 최소화하기 위한 것이다. 신용거래 확대를 위한 의사결정 적용분야로서는 신용카드, 주택할부금융, 소비자 대출, 상업 대출 등을 들 수 있다. 신용평가의 중요한 사안은 현재의 대출한도액을 유지·관리하면서 불량채권에 대한 최선의 대응책을 결정하는 것이다. 신용관리는 은행, 금융서비스, 저당권보험(담보부 보험), 소매(할부 판매) 등 다양한 분야에 적용되고 있다.

(3) 품질개선

품질개선의 목적은 불량품을 찾고, 그 원인을 밝혀서 궁극적으로 이를 예방하는 것이다. 병원과 의료보험조합 등에서는 병원에서 발생하는 사망, 불필요한 장기입원 및 의료비의 과다청구에 초점을 맞추고 있으며, 제조업체에서는 제품보증청구를 유발시키는 불량품 감소를 통한 이윤 증가에 중점을 두고 있다.

(4) 부정행위의 적발

부정행위적발의 목적은 고도의 사기행위를 발견할 수 있는 패턴을 알아내는 것이다. 은행에서는 발견된 패턴을 이용하여 신용카드 거래사기 및 불량수표를 적발할 수 있고, 통신회사에서는 전화카드거래사기를 방지하며, 보험회사에서는 허위 및 과다 청구를 예방할 수 있다.

(5) 이미지분석

이미지분석은 디지털화 된 사진으로부터 패턴을 추출하는 기법이며, 천문학, 문

자인식, 의료진단, 방위산업 등 다양한 분야에서 활용되고 있다

이외에도 위험관리(Risk Management), 망 관리(Network Management), 수요 및 판매 예측(Forecasting)등 데이터 마이닝은 다양한 분야에서 활용되고 있다.

이처럼 데이터 마이닝에 대한 관심이 급격히 부상하게 된 것은 다음과 같은 요인에 의해서 부분적으로 설명될 수 있다.[35]

① 1980년대에 모든 주요 조직들은 하부 구조로서 자신의 고객, 경쟁 업체 및 생산 제품 등에 대한 데이터를 가지는 데이터베이스를 구축하였다. 이 데이터베이스는 잠재적으로 금광의 역할을 할 수 있게 되었다. 데이터베이스에는 SQL이나 다른 표층만을 탐사할 수 있는 질의 도구들을 사용해서는 추적할 수 없는 많은 '숨겨진' 정보들을 가지는 수 기가 바이트의 데이터가 포함되어 있는 것이다. 데이터 마이닝 알고리즘은 데이터베이스 내에서 '최적으로' 분류하는 방식이나 의미있는 관련성들을 찾을 수 있게 한다. SQL은 단지 질의어일 뿐이다. 이것은 이미 알고 있는 데이터를 특정한 조건을 사용하여 찾도록 도와줄 뿐이다. 데이터 마이닝 알고리즘은 보통 데이터베이스에서 더욱 관심있는 부분으로의 축소를 가능하게 한다. 대부분 이 과정에서 여러 번 반복되는 SQL 질의문을 사용하거나 중간 결과를 저장하기도 한다. 물론 이 과정을 하나하나 수동으로 할 수도 있지만 이러한 경우 이 과정은 대부분 매우 번거로운 일이다.

② 네트워크의 활용이 증대되면서 점차적으로 데이터베이스들을 연결하는 일이 쉬어지게 되었다. 그래서 고객의 파일과 인구 조사 데이터와 연결함으로써 특정 인구 집단에 속하는 사람들의 소비 패턴에 대한 예기치 못했던 사실들을 찾을 수 있다.

③ 과거 수 년 동안 기계-학습 기술은 급속도로 발전해왔다. 신경망, 유전자 알고리즘 그리고 단순하면서도 광범위하게 응용될 수 있는 학습 기법들은 데이터베이스 내에서 의미있는 관련성을 찾는 작업을 용이하게 한다.

---

35) Pieter Adriaans, Dolf Zantinge 著, 용환승 譯, 전게서, pp.9~10.

④ 클라이언트/서버의 변혁은 개별적인 지식인이 자신의 터미널로부터 중앙 정보 시스템에 접근하도록 하여 준다. 또한 영업 담당자나 정책 입안자들은 이와 같이 새롭게 얻어진 기술의 활용으로 인한 새로운 활용 가능성을 기대하고 있다.

이외에도 다양한 요인들이 있겠으나, 특히 위의 요인들로 인하여 데이터 마이닝은 광범위한 분야로 관심이 확산되고 있다.

2) 데이터 마이닝의 관련분야

많은 데이터 마이닝의 기법들은 매우 다양한 분야에서 개발되었는데 대표적인 몇 가지 분야들을 소개해 보면 다음과 같다.[36]

① KDD(Knowledge Discovery in Databases)

데이터베이스 안에서의 지식(데이터간의 연관성이나 패턴) 발견이라는 표현은 일반적으로 데이터 마이닝과 가장 유사한 의미로 사용되고 있다. 그러나 보다 정확하게 표현한다면, KDD는 지식을 추출하는 전 과정을 의미하고, 데이터 마이닝은 온라인분석처리(OLAP: On-Line Analytical Processing)나 데이터 웨어하우징(data warehousing)등과 마찬가지로 전체적인 KDD과정 중 한 과정인 탐사 단계를 의미한다고 할 수 있다.

② 기계학습(Machine learning)

인공지능(AI)의 한 분야로서 자동적인 학습기법을 설계하고 구현하는 분야이다.

③ 패턴인식(Pattern recognition)

공학에서 출발하였으며, 이미지 분류와 깊은 관련을 가지고 있다. 패턴인식은 데이터베이스에서 유용한 패턴을 찾아내는 다양한 기법들을 제공한다.

④ 통계학(Statistics)

---

36) 강현철 외5, 전게서, pp.7~8.

데이터 마이닝의 대부분은 통계학의 한 분야라고 할 수 있다. 한 가지 예를 들어 본다면, 데이터 마이닝의 모형화 부문에서 가장 많이 사용되는 기법 중의 하나인 판별분석(discriminant analysis)은 1936년에 R. A. Fisher에 의해 시작된 다변량 통계분석의 한 분야라 할 수 있다.

⑤ 뉴로컴퓨팅(Neurocomputing)
신경망 등과 관련된 다양한 학문적 배경을 가진 분야이다.

3. 데이터 마이닝의 사용기법

데이터 마이닝이란 표층에 드러난 것보다 더 많은 지식이 데이터에 은닉되어 있는 측면과 마찬가지로 하나의 기법으로 이루어진 것이 아니다. 이러한 관점에서 보면 데이터 마이닝은 실제 '어떠한 기법도 관련될 수 있는' 것이다. 데이터에서 드러난 것 이상을 추출할 수 있는 기법들은 모두 적용 가능하므로 데이터 마이닝 기법은 이질적인 그룹을 형성한다. 여러 가지 기법들이 서로 다른 목적으로 사용될 수 있으나 현재 주목되는 기법들은 다음과 같다.[37]

- 질의 도구(Query Tools)
- 통계적 기법(Statistics Technique)
- 가시화(Visualization)
- 온라인 분석 처리(OLAP : Online Analytical Processing)
- 사례-기반 학습(Case-Based Learning, 최단인접 이웃(K-nearest neighbor))
- 의사결정나무(Decision Tree)
- 연관 규칙(Association Rule)
- 신경망(Neural Networks)
- 유전자 알고리즘(Genetic Algorithm)

데이터 마이닝 기법이란 대량의 데이터로부터 새롭고 의미있는 정보를 추출하는

37) Pieter Adriaans, Dolf Zantinge 著, 용환승 譯, 전게서, pp.72~73.

기술이다. 따라서 위에서 나열한 기법들은 공히 데이터로부터 정보를 추출하는 기능을 제공하기 때문에 데이터 마이닝 기법이라 해석할 수 있다. 그러나 이 중에서도 의사결정나무와 인공신경망 기법 등과 같이 인공지능(artificial intelligence)에 기반을 둔 기법들이 대표적인 데이터 마이닝 기법이라고 하는 의견과 이들 모두를 기존의 통계기법 범주에 포함시키는 의견이 학자들간에 분분하다. 실제로 의사결정나무나 인공신경망과 같이 기계학습(machine learning)에 근거한 기법들에 대한 활발한 연구가 시작된 원인 중의 하나가 전통적인 통계기법을 통한 데이터 분석의 한계이다.

본 절에서는 위에 열거한 여러 가지 데이터 마이닝기법들 가운데 최근 정보기술의 발달로 인하여 급격히 관심이 부각된 기법들 즉, 의사결정나무, 인공신경망, 유전자알고리즘에 대하여 살펴보고자 한다.

1) 의사결정나무(Decision Tree)

의사결정나무는 데이터 마이닝의 분류 작업에 주로 사용되는 기법으로, 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 부류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 그리고 이렇게 만들어진 분류모형은 새로운 레코드를 분류하고 해당 부류의 값을 예측하는 데 사용된다.[38]

의사결정나무는 의사결정 규칙(decision rule)을 나무구조로 도표화하여 분류(classification)와 예측(prediction)을 수행하는 분석방법이다. 이 방법은 분류 또는 예측의 과정이 나무구조에 의한 추론규칙(induction rule)에 의해서 표현되기 때문에 다른 방법들(예를 들면, 신경망, 판별분석, 회귀분석 등)에 비해서 연구자가 그 과정을 쉽게 이해하고 설명할 수 있다.[39]

의사결정나무는 새로운 레코드의 부류값을 예측하기 위해 이미 만들어진 분류모형(의사결정나무)이 지시하는 바에 따라 레코드의 속성값을 질문하는 작업을 반복적으로 수행한다. 특히 결정적인 질문을 던지게 되면 다른 모든 속성의 값을 묻지

38) 장남식 · 홍성완 · 장재호, 전게서, p.54.
39) 강현철 외5, 전게서, p.205.

않고도 레코드의 부류값을 정확히 예측할 수 있다. 따라서 레코드를 분류하고 예측할 수 있는 나무(모형)를 얼마나 잘 만드느냐가 의사결정나무 기법의 핵심이다.

의사결정나무는 판별분석 또는 회귀분석 등과 같은 모수적(parametric) 모형을 분석하기 위해서 사전에 이상치(outlier)를 검색하거나, 분석이 필요한 변수를 찾아내고 모형에 포함되어야 할 교호효과를 찾아내는 데 사용될 수고 있고, 그 자체가 분류 또는 예측 모형으로 활용될 수도 있다.[40]

의사결정나무는 순환적 분할(recursive partitioning)방식을 이용하여 나무를 구축하는 기법으로, 나무의 가장 상단에 위치하는 뿌리마디(root node), 속성의 분리기준을 포함하는 중간마디(internal nodes), 마디와 마디를 이어주는 가지(branch), 그리고 최종분류를 의미하는 잎(leaf)들로 구성된다. 마디는 그 기능에 따라서 <표 2-3>과 같이 여러 가지로 분류할 수 있다.

의사결정나무 기법은 먼저 각 속성들이 고객들을 분류하는데 영향을 미치는 정도를 측정한 후, 그 중에서 가장 영향력이 있는 속성을 선정하여 나무의 뿌리마디에 지정한다.

의사결정나무는 첫째 항목에서 시작하여 특정한 기준값을 찾은 후 계속해서 또 다른 기준값을 찾는 방식으로 자료들이 정확하게 분류될 때까지 이 과정을 반복함으로써 데이터베이스에 대한 의사결정 나무를 생성하게 된다. 이러한 의사결정 나무를 자동적으로 생성할 수 있는 알고리즘에는 여러 가지 있는데 n(log n)의 복잡도를 가지므로 매우 효율적이다.[41] 또한 나무 유도 알고리즘은 대규모 데이터 집합에 대해 쉽게 확장되며, 의사결정 과정의 원리를 직관적으로 제공한다는 것이다.

---

40) 최종후 외3, 「Answer Tree를 이용한 데이터마이닝 의사결정나무 분석」, SPSS아카데미, 1998, p.19.

41) Pieter Adriaans, Dolf Zantinge 著・용환승 譯, 전게서, p.91.

<표 2-3> 의사결정나무의 구성요소

구 분	기 능
뿌리마디(root node)	나무구조가 시작되는 마디로써 전체자료로 구성
자식마디(child node)	하나의 마디로부터 분리되어 나간 2개 이상의 마디들을 의미
부모마디(parent node)	자식마디의 상위마디를 의미
끝마디(terminal node) 또는 잎(leaf)	각 나무줄기의 끝에 위치하고 있는 마디를 의미
중간마디(internal node)	나무구조의 중간에 있는 끝마디가 아닌 마디들을 의미
가지(branch)	하나의 마디로부터 끝마디까지 연결된 일련의 마디들을 의미, 이 때 가지를 이루고 있는 마디의 개수를 깊이(depth)라고 한다.

이러한 알고리즘의 결과로 만들어진 의사결정 나무는 사람들에 의해 바로 활용될 수 있으며 이와 대조적으로 예를 들어 신경망의 경우는 자신이 결론에 어떻게 도달하게 되었는가에 대한 정보를 갖지 못하는 블랙박스만을 제시한다.[42]

일반적으로 의사결정나무 분석은 <표 2-4>와 같은 과정을 거쳐서 수행된다.

<표 2-4> 의사결정나무 형성 단계

구 분	처 리 과 정
의사결정나무의 형성	분석의 목적과 자료구조에 따라서 적절한 분리기준과 정 지규칙을 지정하여 의사결정나무를 얻는다.
가지치기	분류오류(classification error)를 크게 할 위험(risk)이 높거나 부적절한 추론규칙(induction rule)을 가지고 있는 가지(branch)를 제거한다.
타당성평가	이익도표(gains chart)나 위험도표(risk chart) 또는 검증용 자료(test data)에 의한 교차타당성(cross validation) 등을 이용하여 의사결정나무를 평가한다.
해석 및 예측	의사결정나무를 해석하고 예측모형을 설정한다.

자료 : 최종후 외3, 전게서, p.20.

42) 상게서, p.95.

이상과 같은 과정에서 분리기준, 정지규칙, 평가기준 등을 어떻게 지정하느냐에 따라서 서로 다른 의사결정나무가 형성된다

의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되어 있으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무 형성방법이 만들어진다. 의사결정나무 형성을 위한 알고리즘으로는 CHAID, CART, QUEST, C4.5 등이 알려져 있는데, 그 개념은 다음과 같다.

(1) CHAID 알고리즘

1975년 J. A. Hartigan에 의해 처음 발표된 CHAID(Chi-squared Automated Interaction Detection) 알고리즘은 본 연구에서 언급된 알고리즘 중에서 가장 오래된 알고리즘이다. 이것은 또한 SPSS와 SAS와 같은 유명한 통계 패키지의 일부로서 상용화되어 있다. CHAID는 일찍이 1963년 J. A. Morgan과 J. N. Sonquist에 의해 발표된 AID(Automatic Interaction Detection)시스템에서 유래되었다. 'AID'에서 암시하는 것과 같이 CHAID에 대한 원래의 동기는 변수들 간의 통계적 관계를 찾기 위한 것이다. 이것은 다시 의사결정 나무를 통해 변수들간의 통계적 관계를 찾기 때문에 이 방법은 분류도구(classification tool)로 사용하게 되어왔다.[43]

CHAID는 범주형 변수(성별, 고용상태, 국적 등)들과 범주형으로 분류되어진 연속형 변수(수입, 교육, 나이 등)들을 사용하기 위하여 설계된 시장세분화 기법이다. CHAID는 하나의 개체를 어떤 지정된 기준과 다른 적어도 2개의 구분들로 나눈다. CHAID는 복잡한 처리절차는 아니지만 거의 어떤 시장세분화 질의에도 사용될 수 있는 상당히 용도가 다양한 도구이다.[44]

이러한 CHAID 알고리즘에 대한 좀 더 구체적인 설명을 하면 다음과 같다.[45]

CHAID는 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이

43) Berry, Michael. J. A., Gordon Linoff, "Data Mining Techniques for Marketing, Sales, and Customer Support", Wiley Computer Publishing-John Wiley & Sons, Inc., 1997, p.265.

44) To Market Strategies' Information & Technology Home Page, market Strategies, Inc., 1998. [http://www.marketstrategies.com/itmethod/chaid.htm]

45) 최종후 외3, 전게서, 1998, pp.32~41.

용하여 다지분리(multiway split)를 수행하는 알고리즘이다. 여기서 다지분리란 부모마디에서 자식마디들이 생성될 때, 2개 이상의 분리가 일어나는 것을 허용함을 의미한다.

CHAID는 목표변수가 이산형일 때, Pearson의 카이제곱 통계량 또는 우도비 카이제곱 통계량(likelihood ratio Chi-square statistic)을 분리기준으로 사용한다. 여기서 목표변수가 순서형 또는 사전그룹화된 연속형인 경우에는 우도비 카이제곱 통계량이 사용된다.

카이제곱 통계량은 관측도수(frequency, f_{ij})로 이루어진 r×c 분할표(contingency table)로부터 계산된다. 분할표로부터, Pearson의 카이제곱 통계량은

$$\chi^2 = \sum_{i,j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

과 같이 정의되고, 우도비 카이제곱 통계량은

$$\chi^2 = 2\sum_{i,j} f_{ij} \times \log_e\left(\frac{f_{ij}}{e_{ij}}\right)$$

로 정의된다. 이때 두 통계량의 자유도(degrees of freedom)는 (r-1)(c-1)로서 동일하다. 여기서 e_{ij}는 분포의 동일성 또는 독립성의 가설 하에서 계산된 기대도수(expected frequency)를 말하며, 아래에 주어진 식

$$e_{ij} = \frac{f_{i.} \times f_{.j}}{f_{..}}$$

와 같이 계산된다.

카이제곱 통계량이 자유도에 비해서 매우 작다는 것은, 예측변수의 각 범주에 따른 목표변수의 분포가 서로 동일하다는 것을 의미하며, 따라서 예측변수가 목표변수의 분류에 영향을 주지 않는다고 결론지을 수 있다. 자유도에 대한 카이제곱 통계량 값의 크고 작음은 p-값으로 표현될 수 있는데, 카이제곱 통계량 값이 자유도에 비해서 작으면 p-값은 커지게 된다. 결국, 분리기준을 카이제곱 통계량 값으로 한다는 것은, p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디를 형성시킨다는 것을 의미한다.

CHAID는 목표변수가 연속형인 경우에는 두 개 이상의 그룹에 대해서 평균치 차를 검정하는 분산분석표(ANOVA Table: Analysis of Variance Table)의 F통계

량을 분리기준으로 이용한다. F통계량이 자유도에 비해 매우 작다는 것은 예측변수의 각 범주에 따른 목표변수의 평균치 차가 존재하지 않다는 것을 의미하며, 따라서 예측변수가 목표변수의 예측에 영향을 주지 않는다고 결론지을 수 있다. 카이제곱 통계량과 마찬가지로 자유도에 대한 F통계량의 크고 작음은 p-값으로 표현될 수 있는데 F통계량이 자유도에 비해서 작으면 p-값은 커지게 된다.

CHAID에서는 계산된 F통계량의 p-값을 기준으로 명목형 목표변수인 경우와 유사하게 병합과 분리를 계속하여, p-값이 가장 작은 예측변수와 그 때의 최적분리에 의해서 자식마디가 형성된다.

(2) CART 알고리즘

CART 알고리즘은 의사결정나무 방법론 중 가장 알려진 것 중에 하나이다. 1984년 Breiman et al 이 CART(Classification and Regression Trees)기법을 발표한 이래로 기계학습 실험의 필수기법이 되어 왔다.[46]

CART는 지니 지수(Gini Index : 이산형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)를 이용하여 이지분리(binary split)를 수행하는 알고리즘이다. 여기서 이지분리란 부모마디로부터 자식마디가 2개만 형성되게 한다는 것을 의미한다.

카이제곱 통계량과 마찬가지로 지니 지수도 불순도(impurity)를 측정하는 하나의 지수이다. 여기서 지니 지수의 의미와 특징에 대해서 간단히 설명하면 다음과 같다.[47]

각 마디에 속하는 개체를 그 마디에서 도수가 가장 많은 목표변수의 한 범주에만 모두 할당하는 분류규칙을 고려해 보자. 임의의 한 개체가 목표변수의 i번째 범주로부터 추출되었고, 그 개체를 목표변수의 j번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기서 $P(i)$는 각 마디에서 한 개체가 목표변수의 i번째 범주에 속할 확률이다. 이러한 오분류 확률을 모두 더하여

46) Berry, Michael J. A., Gordon Linoff, *op. cit.*, p.252.
47) 최종후 외3, 전게서, 1998, pp.42~48

$$G = \sum_{j=1}^{c} \sum_{i \neq j} P(i)P(j)$$

를 얻을 수 있고, 이는 위와 같은 분류규칙 하에 오분류 확률의 추정치(estimate)가 된다. 여기서 c는 목표변수의 범주수를 말한다.

일반적으로 CART는 이산형 목표변수에 대해서는 지니 지수(Gini index)를 분리기준으로 사용한다. 지니 지수는 각 마디에서의 불순도(impurity) 또는 다양도(diversity)를 재는 측도 중의 하나이다.

지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속해있을 확률을 의미하며 Simpson의 다양도 지수(diversity index)로도 알려져 있다. 목표변수의 범주가 2개인 경우에는 지니 지수는 다음과 같이 표현될 수 있으며,

$$G = 2P(1)P(2) = 2\left(\frac{n_1}{n}\right)\left(\frac{n_2}{n}\right)$$

이는 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다.

CART는 이 지니 지수를 가장 감소시켜주는 예측변수와 그 변수의 최적분리를 자식마디로 선택하는데, 지니 계수의 감소량은 다음과 같이 계산된다.

$$\Delta G = G - \frac{n_L}{n} G_L - \frac{n_R}{n} G_R$$

여기서 n은 부모마디의 관측치 수를 말하고, n_R 과 n_L 는 각각 자식마디의 관찰치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록(순수도가 가장 크도록) 자식마디를 형성하는 것이며, 이는 다음과 같은 자식마디에서의 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)\, G_L + P(R)\, G_R = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R$$

목표변수가 연속형일 때는 각 마디의 다양도의 측도로써 다음과 같은 분산(variance)를 고려할 수 있다.

$$V=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

이는 마디의 목표변수의 평균을 그 마디에 속하는 모든 개체의 예측값으로 사용하는 것을 고려한다면, 예측오차를 최소화하는 것과 동일하다 할 수 있다.

명목형 목표변수인 경우와 마찬가지로 분산의 감소량을 최대화하는 것은 다음과 같은 기준

$$\Delta V= V-\frac{n_L}{n}V_L-\frac{n_R}{V_R}$$

를 최대화하는 것과 같다. 여기서 n 은 부모마디의 관측치 수를 말하고, n_R 과 n_L 는 각각 자식마디의 관측치 수를 의미한다. 이는 자식마디에서의 집단내 분산(within variance) $P(L)\ V_L+P(R)\ V_R$ 을 최소화하는 것과 동일하다.

불순도를 분산으로 측정하는 경우에 있어서 예측위험은 나무구조의 모든 끝마디에서 집단내 분산들의 가중합을 계산한 것과 같다.

(3) QUEST 알고리즘[48]

QUEST(Quick, Unbiased, Efficient, Statistical Tree)는 CART에서와 같이 이지분리(binary split)를 수행하는 알고리즘이다.

QUEST는 명목형 목표변수에 대해서만 분석을 수행할 수 있으며, 예측변수의 측도에 따라서 서로 다른 분리규칙을 사용한다. 예측변수가 순서형 또는 연속형인 경우에는 분리규칙으로 ANOVA F검정 또는 Levene의 검정(Levene's robust test of homogeneity of variance)을 사용하며, 예측변수가 명목형인 경우에는 Pearson의 카이제곱 검정을 사용한다.

목표변수의 범주가 3개 이상인 경우에는 CART에서와 유사하게 2-평균 군집분석(tow-means clustering)을 수행하여 두 개의 그룹을 만든 후 분석을 수행한다. 또한 각 예측변수의 최적분리를 찾기 위하여 2차 판별분석(quadratic discriminant

48) 상게서, pp.48~49.

analysis)을 수행하고, 목표변수를 가장 잘 분류하는 예측 변수의 최적분리를 이용하여 자식마디를 형성한다.

일반적으로 CART는 자식마디를 형성할 때 보다 많은 이산값을 가지는 예측변수를 선택하는 경향이 있기 때문에, 계산시간이 다소 많이 걸리고 분류 또는 예측오차가 커질 가능성이 있다. QUEST는 위와 같은 알고리즘을 이용하여 변수선택 편의(bias)나 계산시간을 줄이고자 하는 방법이다. QUEST는 관측치의 수가 많거나 복잡한 자료에 대해서는 효율적이지만, 이 방법 역시 모든 관점에서 다른 알고리즘보다 항상 좋은 결과를 주는 것은 아니라는 것이다.

(4) C4.5 알고리즘

C4.5는 J. Ross Quinlan이 몇 년에 걸쳐 진화시키고 정제시킨 의사결정 알고리즘으로 가장 최근의 알고리즘이다.[49]

C4.5를 언급할 때 따라 다니는 지수는 엔트로피이다. CART와 마찬가지로 C4.5도 마디의 순수함을 재는데 이번에는 비트(bit)개념을 이용한다. 비트(bit)의 개념을 다음의 예로 이해해보자. 만약 8개의 카테고리로 이루어진 마디가 있다면 그 각각의 카테고리를 표시할 비트는 총 $\log_2 8 = 3$개가 필요하게 된다(000은 제1 카테고리, 001은 제2 카테고리,…, 111은 제8 카테고리로 표현 가능하다).

그 마디에 가지가 쳐져서 4개의 카테고리가 있는 자식마디가 생기면 그 자식마디는 $\log_2 4 = 2$개의 비트만 있으면 표현이 가능하다. 결국 가지가 쳐지면서 3개가 필요했던 비트가 2개로 줄어들어 엔트로피 1의 이득을 본 것이다. C4.5에서는 이런 비트개념이 수리적으로 보다 더 정교하게 지수화되는데 그 식은 다음과 같다.

$$Entropy = -\sum P(i) \log_2 P(i)$$

결국 위의 식은 각 카테고리가 차지하는 비로 가중치를 준, 필요한 비트들의 합이 된다(식 맨 앞에 붙은 음수기호는 전체 값을 양수화시키기 위함이다). 마디가 순수할수록 엔트로피는 줄어드는데 응답자로만 구성된 마디의 경우, 엔트로피는 $1 * \log_2 1 = 0$임을 알 수 있다. 결국 나무는 엔트로피가 감소하는 방향으로 자라

49) Berry, Michael J. A., Gordon Linoff, *op. cit.*, p.259.

는 것이며 엔트로피를 가장 많이 줄여주는 변수를 기준으로 가지는 쳐진다.[50]

위에서 의사결정나무 알고리즘들에 대한 차이를 간략히 살펴보았는데 어떤 상황에 어떤 알고리즘이 더 좋다는 법칙을 말하기는 어렵다. 단, 가지고 있는 데이터에 다양한 알고리즘을 적용하여 보다 더 타당하게 해석이 가능하고 의미있는 결론은 유출해주는 알고리즘을 선택하여야 할 것이다.

위에서와 같은 의사결정나무 기법의 장점과 단점은 다음과 같다.[51]

① 장점

- 분류나 예측의 근거를 알려주기 때문에 이해하기가 쉽다.
- 데이터를 구성하는 속성의 수가 불필요하게 많을 경우에도 모형 구축시 분류에 영향을 미치지 않는 속성들을 자동으로 제외시키기 때문에 데이터 선정이 용이하다.
- 연속형이나 명목형 데이터 값들을 기록된 그대로 처리할 수 있기 때문에 지식발견 프로세스 중 데이터의 변환단계에서 소요되는 기간과 노력을 단축시킨다.
- 어떠한 속성들이 각각의 부류값에 결정적인 영향을 주는가를 쉽게 파악할 수 있다.
- 모형구축에 소요되는 시간이 짧다.

② 단점

- 나이나 소득 등과 같은 연속형 데이터를 처리하는 능력이 신경망이나 통계기법에 비해 떨어지며, 결과적으로 예측력도 감소한다. 따라서 데이터에 다수의 연속형 변수(속성)가 포함되어 있을 경우 값들을 그룹화하여 이산형(discrete)이나 범주형 값으로 변환시킬 필요가 있는데, 그룹화하는 과정에서 발생하는 치우침을 배제할 수 없다.
- 부류가 주가나 주택의 가격 등과 같은 연속형 변수의 형태를 취하며, 이것을

50) 한국SAS소프트웨어 DB 마케팅팀, "마이닝의 접근방법론과 기법", 데이터베이스 월드, 1998년 3월호. p.83.
51) 장남식외2, 전게서, pp.58~59.

예측하는 모형을 구축하는 것이 목적일 경우에는 적합하지 않다.

- 모형을 구축하는데 사용되는 표본의 크기에 지나치게 민감하다. 따라서 보다 정확한 모형을 만들기 위해서는 서로 상이한 값을 갖는 레코드들을 가능한 한 많이 포함하는 데이터가 필요하다.

2) 인공 신경망(Neural Networks)

신경망은 인간이나 동물들이 가지고 있는 생물학적인 뇌의 신경세포(Neuron)을 모델화하여 인공적으로 지능을 만드는 것이다. 즉 인간의 뇌에 존재하는 생물학적 신경세포와 이들의 연결 관계를 단순화시켜 수학적으로 모델링하여 인간의 두뇌가 나타내는 지능적 형태를 구현하는 것[52]으로 마디(node)와 고리(link)로 구성된 망 구조를 모형화하고, 의사결정나무와 마찬가지로 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다. 신경망은 분류, 군집, 연관규칙 발견과 같은 작업에 널리 사용되는 데이터 마이닝 기법으로 신용평가, 카드 도용패턴 분석, 수요 및 판매예측, 고객세분화(customer segmentation)등 여러 가지 목적으로 다양한 산업분야에 폭 넓게 적용되고 있다.[53]

(1) 인공 신경망의 구조

신경조직의 기본적인 구성요소는 신경세포 즉, 뉴런(Neuron)이다. 인간의 두뇌속에는 이러한 뉴런이 10^{10}에서 10^{11}개까지 존재하며 연산의 기본단위를 이룬다.[54] 인공신경망은 인간두뇌의 정보처리과정과 그 구조를 모형화한 것이다. 이는 뇌의 신경회로망같이 신경세포에 해당하는 다수의 처리 프로세서들을 네트워크 형태로 접속하여 상호간에 신호를 주고받으면서, 병렬로 작동하여 하나의 인공신경망(Neural Network)으로 자료를 처리한다. 그 구조는 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer)의 삼층구조로 나누어지며, 각 층마다 정보

52) 임영도 · 이상부, 「퍼지 · 신경망 · 유전진화」, 영과일, 1998, p.107.
53) 장남식외2, 전게서, pp.59~60.
54) 임영도 · 이상부, 전게서, p.108.

처리단위(Processing Unit)인 다수의 뉴론을 가지고 있다. 입력층의 처리단위는 외부로부터 입력을 받아들이는 처리단위이며, 출력층의 처리단위는 외부에 출력값을 내보내는 처리단위이다. 은닉층의 처리단위는 외부환경과 상호작용을 하지 않는 처리단위로서 외부로부터 은닉되어 있다.

입력계층은 결과변수를 설명하는데 이용하고자 하는 입력변수들(예를 들어, 카드의 부정거래를 색출하려는 문제에 있어서 승인시간간격, 승인금액, 누적승인건수 등이 있을 수 있다)이고 출력계층은 예측치를 얻고자 하는 결과변수(부정거래인지 정상거래인지의 여부)이다. 두 개 이상 놓여질 수 있는 은닉계층은 인간의 신경망을 모형화 한 몇 개의 은닉마디(hidden unit)로 이루어져 있는데, 각 은닉마디는 연결함수(combination function)를 통해 입력변수들과 연결되어 있고 이러한 각 연결에 사용되는 계수는 연결가중치(synaptic weights)라고 불린다.

각 은닉마디와 은닉계층에는 대개 비선형함수인 활성함수(activation function)를 지정하게 되는데 입력으로 들어오는 데이터는 이 함수에 의해 변환되어져 다음 마디나 계층으로 전달된다.

데이터를 가지고 신경망을 학습 혹은 훈련(training)시키는 것은 여러 가지 입력변수의 정보로부터 은닉계층 내에서의 다소 복잡한 내부작업을 통해 가장 정확한 결과를 주도록 연결가중치의 값을 찾아가는 것이다.

가장 일반적인 훈련방법은 후진전파(back propagation)이다. 데이터의 각 레코드별로 그 값을 읽어 신경망을 훈련하고 이에 의한 출력 결과와 이미 알고 있는 실제값을 비교하여 그 차이가 충분히 작아져 오차가 일정한 수준에 수렴할 때까지 반복적으로 가중치를 조정해간다.

여기에는 레벤버그-마쿼르트(Levenberg-Marquardt), 준뉴턴(Quasi-Newton), 결합기울기(conjugate gradient)를 비롯한 몇 가지 반복적인 최적화 알고리즘이 사용될 수가 있다.

일련의 훈련과정이 끝나고 연결가중치의 값이 정해지면 이 가중치들을 아직 결과변수의 값을 알지 못하는 새로운 데이터에 적용하여 예측치를 얻게 된다.[55]

---

55) 한국SAS소프트웨어 DB 마케팅팀, 전게자료, pp.84~85.

(2) 인공신경망의 특성 및 장·단점

신경망 분석의 특성으로는 다음과 같다.[56]

① 병렬 처리(Parallel processing)

기존의 컴퓨터는 하나의 CPU로 모든 행동에 대해 순차적으로 명령하는 직렬 형태였다. 이와는 대조적으로 신경망은 다수의 뉴런이 모여서 동시에 서로 다른 처리를 한다. 이때 뉴런 각각은 그 처리 속도가 느리지만 여러 개에 의한 병렬 분산 처리로 직렬 처리 보다 빠른 정보처리를 할 수 있다.

② 학습과 훈련을 통한 적응성(Adaptive by learning & training)

신경망은 인간의 두뇌처럼 경험을 통하여 학습하는 기능이 있는데 이것이 신경망의 핵심적인 점이다. 이러한 경험을 통한 학습 기능으로 알고리즘이나 프로그램이 어려워 기피되어 온 응용 분야도 컴퓨터 처리가 가능해지게 되었다. 학습 방법으로는 관리학습(supervised learning)과 비관리학습(unsupervised learning)의 두 가지가 있다. 우선 관리 학습은 모형의 분석에서 입력 자료와 출력 결과에 대한 정보를 가지고 있을 때 선호되는 방법이다. 얻어진 자료를 가지고 입력층에서 마지막으로 출력층에 이를 때까지 다음 층에 가중치를 부과하는 과정을 반복해 가면서, 모형을 만들어 분석을 한다. 즉 관리학습은 실제값과 관측 값의 차이로 발생하는 값을 이용하여 학습하고 새로운 모형을 구축하게 된다. 관리학습에 대표적인 방법으로는 다층 인식모형(multi-layer perceptron)이 있다. 이것은 오류-역전파망 학습 알고리즘의 근간을 이루고 있다. 비관리학습은 주로 자료의 유사성을 측정하여 같은 군집으로 분류하는데 사용되며, 관리학습처럼 사전에 어떠한 정보를 요구하거나 처리요소들을 그 중요도에 따라 가중치를 주지 않는다. 대표적인 모델로서는 투에보 코호넨의 자가구성지도(self-organizing map)가 있다.

---

56) 조용준·허준·최인규, 「Neural Connection을 이용한 데이터마이닝 신경망분석」, 자유아카데미, 1999, pp.7~8.

③ 뉴런간의 고도의 상호 연결성(Highly interconnected)

신경망은 정보를 기억장치의 정해진 장소에 일괄 저장해 두지 않고, 노드(혹은 뉴런)이라 불리는 많은 처리장치에 저장한다. 이는 노드간에 고도의 연결성을 의미한다. 예를 들어 '아기'라는 음성을 인식할 때, 전통적인 음성 인식 시스템에서는 음성의 한 단어나 음절을 저장한 뒤, 탐색표를 가지고 있어 입력된 음성을 이 표와 일일이 대응시켜 비교하여 결국에는 '아기'라는 음성을 인식하게 된다. 하지만 신경망에서는 여러 개의 처리장치들을 동시에 사용하여 그와 같은 음성을 인식하게 된다. 그러므로 하나의 정보가 네트워크 전체나 어느 부분에 걸쳐 분산되어 있을 수도 있으며, 여러 개의 정보가 하나의 네트워크에 동시에 저장될 수도 있다.

위에서 언급한 분산저장법은 정보의 표현이 풍부하며 일부 정보가 파손된다 하더라도, 작업을 계속 수행할 수 있다. 이러한 점에서 신경망은 결함 허용 시스템(fault-tolerant system)이라고 할 수 있다.

이와 같은 인공신경망은 다른 통계적인 방법에 비해 고객집단 분류, 카드사기 적발, 라이프사이클 예측관리 등과 같은 예측분야에서 좋은 성과를 보이고 있는데, 이는 인공신경망이 가지고 있는 학습성(learnability)과 견고성(robustness)때문이다.[57]

첫째, 학습성은 주어진 입·출력 데이터로부터 학습을 통해 숨겨진 규칙성을 찾아내는 것을 말한다. 이러한 규칙성은 인공신경망 내의 분산된 처리단위에 각각 저장되는데, 이는 기계자동학습의 지식베이스에 해당한다. 둘째, 일반적으로 잘못된 데이터는 오류를 가지게 되고, 결과추론을 잘못하게 한다. 그러나 인공신경망 모형은 특정의 몇몇 처리단위에 오류가 발생해도 인공신경망의 전체적인 기능이 크게 영향을 받지 않는 견고성을 가지고 있다. 따라서 인공신경망의 성과는 오류정도가 증가됨에 따라 점차적으로 감소하는 추세를 보이며, 급격한 환경변화나 예측하지 못한 환경에 대한 예측에서 안정적으로 평가하는 기능을 가지고 있다.

이러한 인공신경망의 장점으로는 비교적 쉽게 비선형적 특성을 갖는 다양한 문

57) 이용회, "Data Mining을 이용한 리테일 뱅킹 전략에 관한 실증적 연구", 전국은행연합회 논문집, 1998, p.17.

제를 모형화하여 예측할 수 있다는 것과 자료가 불확실하고, 불완전해도 효과적으로 모형을 생성하고, 또한 학습능력 및 지식발견에서도 성과가 월등하기 때문에 데이터 마이닝의 도구로써 사용되고 있다. 따라서 인공신경망은 패턴인식(pattern recognition), 예측(prediction)을 필요로 하는 다양한 분야에서 기존의 방법으로는 해결하기 어려웠던 부분들을 해결하거나 성과를 제고시키는 방법으로 적용되고 있다. 그러나 인공신경망은 'Black Box' 의 개념이 들어있어 사용자에게 어떻게 해서 신경망 내에서 그런 결정이 나왔는가에 대한 설명을 하지 못한다는 단점이 있다. 그래서 근래에는 이러한 인공신경망의 단점을 극복하기 위해 통계학, 의사결정나무, 사례기반추론, 유전자 알고리즘 등의 데이터 마이닝 기법들과 통합하는 연구가 이루어지고 있다.[58]

신경망은 점점 더 많이 이용되고 있지만, 이 방법은 계산에 많은 비용이 들며 이해하기 어렵다. 그러나 성능이 가장 중요한 우선 순위라면 신경망의 정확성은 선형적인 통계 방법보다 모든 점에서 우수하므로 마케팅에 중요한 기여를 할 수 있다.[59]

3) 유전자 알고리즘(Genetic Algorithm)

유전자 알고리즘은 자연 선택의 원리와 자연계의 생물 유전학에 기본이론을 두며 병렬적이고 전역적(global)인 탐색 알고리즘으로서, 모든 생물은 주어진 다양한 환경속에 적응함으로써 살아남는다는 Darwin의 적자생존(survival of the fittest)의 이론을 기본 개념으로 한다. 유전자 알고리즘은 풀고자하는 문제에 대한 가능한 해들을 정해진 형태의 자료구조로 표현한 다음 이들을 점차적으로 변형함으로써 점점 더 좋은 해들을 만들어 낸다. 다시말하면 미지의 함수 $Y=f(x)$를 최적화하는 해 x를 찾는 모의 진화(simulated evolution)형의 탐색 알고리즘이다.[60]

인공지능을 단적으로 말하자면 컴퓨터가 스스로 판단하도록 학습하는 것으로 볼 수 있다. 이러한 것들로부터 컴퓨터는 어떤 문제의 해결책을 찾을 것이다. 인공지

58) 상게논문, p.18.
59) Ramon C. Barquin et al. 著, 함문성 · 김석호 譯, 전게서, p.133.
60) 임영도 · 이상부, 전게서, p.193.

능에서 문제를 해결하기 위한 해결책을 찾는데 탐색은 상당히 중요하다. 특히 인공지능을 이용하는 문제들은 대부분 탐색해야 할 공간이 엄청나기 때문에 그 중요성은 더욱 부각된다. 유전자 알고리즘은 탐색 공간이 크거나 분석적으로 해를 찾을 수 없는 문제에 대해 해결책을 제시할 것이다.

유전자 알고리즘은 선택 도태나 돌연변이같은 생물 진화의 원리로부터 착안된 알고리즘으로서 확률적 탐색이나 학습 그리고 최적화를 위한 한 가지 기법이라고 간주할 수 있다.

역사적으로 유전자 알고리즘을 살펴보면 홀랜드(Holland)는 1975년에 "Adaption in Natural and Artificial System" 이라는 저서에서 처음으로 유전자 알고리즘을 소개하였다. 또한 포겔(fogel) 등도 진화방식의 모형화를 시도하여 간단한 유한 상태 시스템의 최적화를 수행하기도 하였다. 1985년에는 미국의 카네기멜론 대학에서 제 1회 유전자 알고리즘에 관한 국제회의가 개최되기도 하였으며 그 이후에도 이 회의는 2년마다 개최되고 있다. 개략적으로 유전자 알고리즘을 설명한다면 다음과 같다.[61]

- 문제에 맞는 염색체(chromosome)의 구조를 정의하고 그 유전자의 적합도를 따지는 적합도 함수를 설계한다. 또 문제에 맞는 유전자 조작을 정의한다.
- 초기 모집단을 생성한다. 이때 모집단의 숫자도 제한한다.
- 모집단을 가지고 유전자 조작을 한다.
- 조작된 유전자들의 염색체를 적합도 함수를 써서 평가한다.
- 설정된 확률대로 점수가 높은 유전자(gene)들의 모집단에서의 비율을 높이고 낮은 점수들은 없앤다.
- 확률에 의해 점수가 높은 순서대로 복제(reproduction)과정을 통해 더욱 많은 자가복제를 하고 그것들끼리의 교배(crossover)를 통해 새로운 모집단을 생성한다.

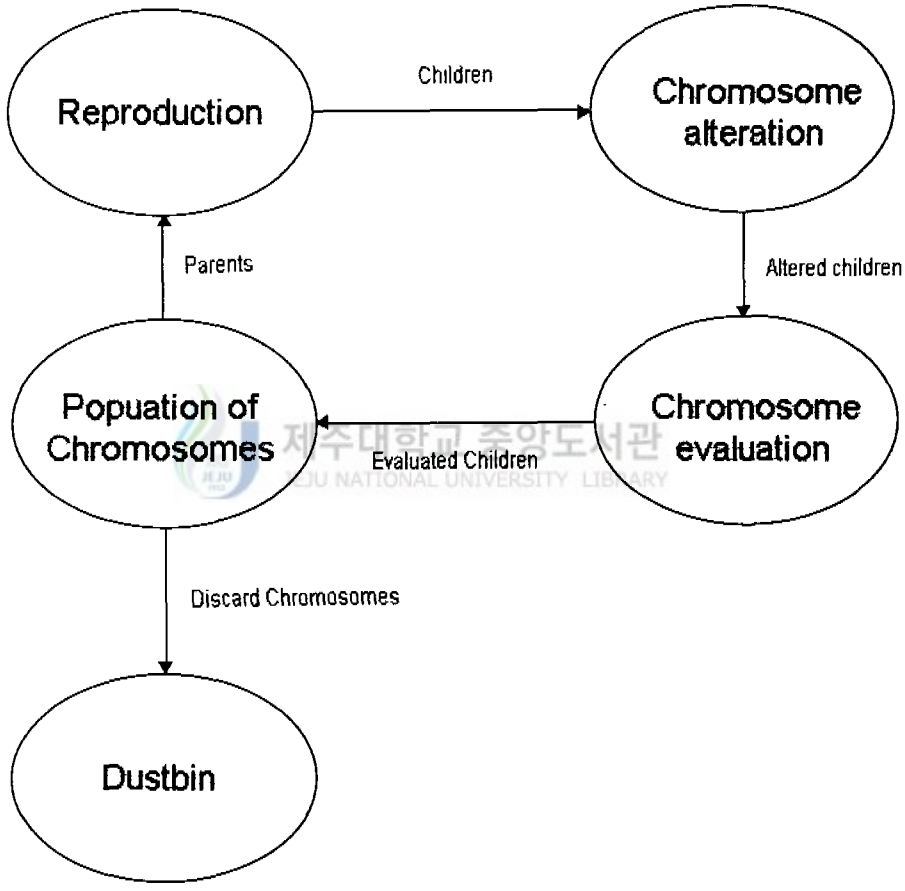
이러한 과정을 반복하면서 높은 점수를 가지는 유전자들을 계속 확장시켜 나가

61) 김남훈 · 문성광 · 박세진 · 이인, "유전자 알고리즘의 이해와 구현", 프로그램세계, 신영미디어, 1997.7, p.122.

는 것이다. 이러한 과정을 <그림 2-3>로 보이고 있다.

하지만 이 방법대로 진행하면 어떤 경우 국지 최적해(local minima)에 빠질 위험이 있으므로 변이(mutation)를 만들어 임의의 영역을 탐색한다. 교배를 통해 원하는 해에 수렴하고, 또한 변이를 통해 임의의 영역으로 탐색 공간을 넓혀 적합도를 계산한다. 이러한 과정을 반복함으로 해서 원하는 해에 수렴하고자 하는 것이다.

<그림 2-3> 유전자들의 확장과정



사료 : 상게자료, p.123.

유전자 알고리즘에서 중요한 유전 연산자의 세 가지 기능의 재생산, 교배, 돌연변이에 대해서 살펴보도록 하겠다.

(1) 재생산(Reproduction)

재생산은 각각의 스트링이 가지는 적합도에 따라 그 스트링을 복제하는 과정이며 이 때 적합도 함수는 사용자가 최대화하기를 원하는 어떠한 형태의 함수도 가능하다. 이 과정은 적합도가 높은 개체일수록 다음 세대에 더 많은 자손을 가질 확률이 높음을 의미하며 이는 주어진 환경에 더 잘 적응하는 개체만이 살아남는다는 자연 선택의 원리를 담고 있다.

이러한 과정으로 집단의 수만큼 재생산함으로써 현재 세대(t)에서 다음 세대(t+1)로 가는 중간 과정인 중간 세대를 만든다. 그리고 현재 세대에서 적합도가 가장 높은 개체는 다음 세대에서 최소한 하나 이상의 자손을 보장해 주는 엘리트 규칙을 사용한다.[62]

(2) 교배(Crossover)

교배는 두 부모의 염색체를 조합하여 바꾸어 자식의 염색체를 만드는 조작이다. 가장 단순한 방법은 교배하는 위치를 하나 결정하고 그 앞과 뒤에서 어느 쪽 부모의 유전자형을 받을 것인가를 변경시키는 방법이다. 이것을 1점 교배(one-point crossover)라고 한다. 아래는 1점 교배의 예이다.

개체 A　1 0 0 1 | 1 1 1 ⇒ 1 0 0 1 0 0 0
개체 B　0 0 1 1 | 0 0 0 ⇒ 0 0 1 1 1 1 1

또 다른 방법으로는 복수점 교배가 있다. 복수점 교배는 교배위치가 복수인 방식이다. 예를 들어서 교배 위치가 2와 5라면 새로운 개체의 하나는 개체 A의 선두로부터 두 번째까지, 개체 B의 세 번째로부터 다섯 번째까지, 개체 A의 여섯 번째로부터 마지막까지로 유전자가 만들어진다. 동시에 그 반대의 조합에 의해서 또 다른 하나의 새로운 개체의 유전자가 만들어진다. 아래는 복수점 교배의 예이다.[63]

62) 임영도 · 이상부, 전게서, p.197.
63) 김남훈 · 문성광 · 박세진 · 이인, 전게자료, p.123.

개체 A 10 | 011 | 11 ⇒ 1011011

개체 B 00 | 110 | 00 ⇒ 0001100

(3) 변이(Mutation)

변이는 유전자를 일정한 확률로 변화시키는 조작이다. 돌연변이를 너무 큰 변이 확률로 설정하면 적합도가 떨어지는 경우도 생기게 되지만 오히려 적합도가 높아지는 경우도 생긴다. 돌연변이가 없는 경우에는 초기 유전자의 조합 이외의 공간을 탐색할 수 없어 초기 조합에 적절한 해가 없을 경우 원하는 해를 구할 수 없게 된다. 일반적으로 돌연변이는 고정된 확률로 각 유전자가 변화하도록 설정하지만 변이율을 동적으로 변화시키는 기법도 있다. 이러한 유전자 조작을 문제에 맞게 정의하는 것 이외에 고려해야 할 사항들이 많다. 적합도 함수의 설계나 모집단의 크기를 정하는 일들도 유전자 조작 못지 않게 중요하다. 유전자 알고리즘의 가장 어려운 점은 바로 수렴되는 시기를 결정하는 일이다. 즉 어느 정도가 지나야 우리가 만족할 수 있는 해를 구할 수 있느냐는 것을 결정하는 일이다. 이것은 다음과 같은 다양한 방법이 존재할 수 있다.64)

- 반복 수를 정의한다. 즉 전체 과정이 몇 번 반복되었느냐에 따라 수렴을 결정하는 방법이다.

- 확률적으로 정의한 교배나 변이의 발생 수 또는 두 가지 모두에 대한 발생 횟수를 정의한 다음 그 수에 도달하면 끝낸다.

- 어떤 유전자들이 일정한 적합도(fitness)값 이상을 가지면 끝낸다.

- 전체 모집단의 적합도 평균이 일정값 이상이면 끝낸다.

위에 예시된 것과 같은 다양한 방법이 존재할 수 있는데 문제에 맞게 하나의 방법을 찾아내는 것이 중요하다.

---

64) 김남훈 · 문성광 · 박세진 · 이인, 상게자료, p.123.

문제 해결을 위해 유전자 알고리즘을 적용하는 절차는 다음과 같다.[65]

① 문제를 제한된 알파벳의 스트링으로 우수하며 훌륭한 코딩 방법을 설계한다.

② 컴퓨터 내에서 해답들이 투쟁과정에서 서로 결합할 수 있는 인공적인 환경을 만든다. 전문적인 용어로 '적합 함수(fitness function)'라고 부르는 성공과 실패를 측정하는 객관적인 기준을 제공한다.

③ 후보 해답들이 결합될 수 있는 방법을 개발한다. 부모의 스트링을 간단히 잘라서 교환한 후에 다시 접합하는 '교잡(cross-over)' 연산 방법이 많이 사용된다. 복제에서는 모든 종류의 돌연변이 연산자가 적용될 수 있다.

④ 잘 분포된 초기 개체들을 만들고 각 세대에서 부실한 해답들은 제거한 후 우수한 해답의 자손이나 변이들로 대체함으로써 컴퓨터에서 '진화'를 할 수 있도록 한다. 성공적인 일련의 해답이 만들어지면 종료한다.

이와 같은 방법을 살펴 볼 때 유전자 알고리즘의 응용은 어려워 보이지 않으며 실제로도 그렇다. 어느 프로그래머도 기본적인 구조를 쉽게 작성할 수 있으며 부분적으로 이미 성공한 것과 마찬가지로 볼 수도 있다. 인간의 창조성은 특히 코딩(소위 표현 공학(representational engineering))을 우수하게 해서 형식화하는 것과 효과적인 돌연변이 연산자를 발견하는 데 필요하다.

일반적으로 유전자 알고리즘의 장점과 단점은 자연선택의 경우와 부분적으로 일치한다. 두 가지 단점은 개체의 대규모 과잉 발생이고 탐색 과정에서의 우연적 특성이다. 보통 중요한 결과를 얻기 위해서는 상당한 규모의 컴퓨팅 능력이 요구된다. 반면에 이 기법은 견고하다. 해답이 있으면 유전자 알고리즘은 아마 발견할 것이다. 그러나 오퍼레이션 리서치와 같은 특정 분야의 문제에서는 유전자 알고리즘이 종종 특수 설계된 알고리즘에 상대가 되지 못하지만 이것의 실용성은 주로 응용의 범위가 넓다는 것과 개념적으로 명료하다는 것에서 비롯된다. 사촌간의 신경

65) Pieter Adriaans, Dolf Zantinge 著 · 용환승 譯, 전게서, p.115.

망과 더불어 자가-학습(self-learning)시스템의 '팔방미인(jacks-of-all-trades)'인 것이다. 유전자 알고리즘에 의해 발견된 해답은 기호로 코딩되어 있으므로 쉽게 읽을 수 있어서 블랙박스의 기능만을 하는 신경망에 비교할 때 장점을 갖는다.

기존에 다른 학습 방식을 사용하고 있는 개체에도 유전적 접근 방식을 적용할 수 있다는 의미에서 유전자 알고리즘은 메타-학습 전략의 일종으로 볼 수 있다. 최근 수년 동안 여러 가지 혼성 접근방식이 개발되었는데 그 중에는 신경망을 유전자 알고리즘의 입력을 만드는 데 사용하거나 반대로 유전자 알고리즘을 신경망의 출력을 최적화하는 데 사용하는 것이 있다. 현재 유전자 프로그래밍은 금융 영업이나 보험 분야의 응용에 폭넓게 사용되고 있다.[66]

제4절 데이터 마이닝의 선행연구 및 연구 배경

DB마케팅과 데이터 마이닝에 관한 연구는 여러 분야에서 찾을 수 있지만, 본 연구에서의 주제와 관련 있는 연구들은 드물게 이루어 졌다. DB마케팅 수행을 위한 고객 분석과 같은 기업활동 사례는 다양한 연구를 기반으로 수행되었다 할 수 있기 때문에 여러 가지 시사점을 줄 수 있으며, 이러한 의미에서 본 절에서는 기업활동 사례와 선행연구를 살펴볼 필요가 있다.

먼저 효율적 DB마케팅을 위해 데이터 마이닝 기법을 적용한 기업활동 사례(Merrill Lynch, HNC software Inc., People Bank)와 연구(김신곤, 1999), 그리고 데이터 마이닝을 의사결정지원시스템에 활용한 연구(조성진·정인정, 1999, 지원철·서민수, 1998.)는 본 연구의 목적을 위하여 살펴볼 필요가 있다. 또한 더 나아가 데이터 마이닝 기법의 예측력을 비교한 연구(이용희, 1998)도 본 연구의 주제와 연관이 있다고 할 수 있다.

우선 Merrill Lynch사의 판별함수를 이용한 잠재고객의 발견 사례는 은행의 예는 아니지만 투자신탁 상품도 은행의 주력상품 가운데 하나라는 점을 감안하면 판별함수와 같은 다변량 분석기법이 은행의 사업활성화를 위해서도 효과적으로 사용

66) 상게서, pp.116~117.

될 수 있음을 보여 주는 좋은 예라고 할 수 있다.[67]

금융서비스 산업에서는 가능성이 큰 잠재고객의 발견이 기업 성장의 큰 관건이 되고 있는데, 미국의 유명한 금융회사인 Merrill Lynch사는 판별함수가 잠재고객을 발견해 냄에 있어 어떻게 기여할 수 있는가에 대한 좋은 사례를 보여 주고 있다. Merrill Lynch사의 내부 컨설팅 조직이라고 할 수 있는 Management Science Group은 미국 가구의 인구통계변수를 이용하여 반응 가능성이 큰 잠재고객을 식별해 낼 수 있는 판별함수를 개발했다. Management Science Group은 먼저 Merrill Lynch사의 우수고객과 평균적인 소비자를 구분하기 위해 자사의 데이터베이스에서 우수고객을, 그리고 평균적인 미국 가구에서 우수고객과 비슷한 숫자의 표본을 임의추출(random sampling)한 다음, 이 두 집단을 구분할 수 있는 판별함수를 구하기 위해 100개가 넘는 변수들을 분석과정에 포함시켰다. 판별분석 결과 100여 개의 변수 가운데 4개의 변수가 두 집단의 구분에 통계적으로 유의한 영향을 주고 있는 것으로 나타났다.

판별함수를 통한 잠재고객 추출모델이 개발되기 전까지 이 회사는 추정 가구소득과 사회경제적인 소속집단의 2개 변수를 기준으로 잠재고객을 추출해 내는 모델을 사용하고 있었는데, 잠재고객의 선별에 판별함수를 이용한 결과, 종전보다 고객의 평균자산에서는 167%, 기업의 수입에서는 39%, 잠재고객 가운데 실제로 Merrill Lynch사에 계좌를 개설한 고객의 비율(conversion rate)에서는 43%의 증가를 가져오는 성과를 거두었다. 이러한 성공을 바탕으로 Merrill Lynch사는 잠재고객을 식별함에 있어 판별함수모델을 공식적으로 채택하게 되었다. 이 모델은 연간 35~60억 달러에 이르는 고객자산액의 증가와 20만~45만 달러에 이르는 기업수입의 증가를 가져올 것으로 예상되고 있다.

다음으로는 HNC software Inc.의 인공 신경망 분석 연구사례 이다.[68]

HNC software Inc.라는 회사는 신경망 분석기법을 이용한 예측모델을 패키지화하여 판매하고 있는 회사인데, 1년에 두 번씩 DM(Direct Mailing)캠페인을 실시하는 어떤 은행이 대출상품 판매를 위한 DM캠페인을 실시하면서 자사의 신경망 분

67) Labe Jr., Russell P., "Database Marketing Increases Prospecting Effectiveness at Merrill Lynch" , Interfaces, 24(5), September/October, 1994, pp.1~12.
68) 박찬욱, 전게서, 1999, p.120.

석모델을 적용하여 발송 리스트를 추출함으로써 DM을 종전보다 27%나 적게 발송하고도 반응 수는 종전보다 뒤지지 않은 결과를 얻었다고 소개하고 있다. 이러한 결과는 실제로 반응률이 종전보다 50%가 향상된 것이며, DM발송 수를 줄임으로써 DM 제작비용을 7만 달러 정도 절감할 수 있었음을 설명하고 있다.

또한 미국 Peoples Bank의 사례를 보면 다음과 같다.[69] 미국 Peoples Bank의 영국 신용카드업계 진출 사례는 선진카드사 해외진출의 전형을 보여 주고 있다. 이 은행이 1996년 5월에 영국에서 카드사업을 전개하기 시작하면서 가장 먼저 수행한 업무는 잠재고객의 데이터베이스를 구축하는 것이다. 이를 위해 먼저 5백만 명의 소비자를 대상으로 DM을 발송하여 소비자의 카드사용형태나 인식 등을 조사하였으며, 그 다음으로 선거인 명부에 수록된 4,400만 명을 근간으로 DB를 구축하고 1997년에는 이 가운데 1,400만 명을 선별하여 잠재고객 DB 구축을 위한 DM을 발송하였다. 대상고객 4,400만 명 가운데 DM발송 대상고객 1,400만 명을 추출하기 위해 여러 단계를 거쳤다. 먼저 파산 선고자 등 신용상태가 극히 불량한 소비자들을 제외하였으며, 그 다음으로 리스트 공급사업자들이 보유한 정보들을 구입하여 회귀분석, CHAID(chi-square automatic interaction detection), 신경망분석(neural network analysis) 등의 여러 가지 다변량분석 기법들을 적용하였다. 이러한 기법을 이용하여 우수 신용카드 사용자이면서 DM캠페인에 반응할 확률이 높은 고객들을 선별하고 이들을 중심으로 고객 DB를 구축하고 있다. 즉, Peoples Bank는 영국의 3천만 카드 소지자 가운데 약 8~9백 만 정도로 추산되는 우수고객을 카드사업의 주요 타켓으로 삼고 있다. 이 은행은 향후에도 1997년과 같은 공격적인 DM캠페인을 지속적으로 실시할 계획을 갖고 있다.

국내에서의 한 연구사례[70]는 고객 데이터베이스로부터 텔레마케팅을 수행하였을 경우 높은 응답율이 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 텔레마케팅의 수행 대상이 되는 고객을 선별하는 모델의 개발에 관한 연구에서 데이터 마이닝 기법 중 의사결정나무 알고리즘(CHAID)을 적용하여 모델을 개발하였고 개발된 모델의 성과는 리프트(Lift)를 이용하여 평가하였다. 이 연구는 인구통계적 정보, 직접우편 발송 여부, 쿠폰발송 여부, 자동이체신청 여부와 카드유치에 관한

69) "Letters from America", *Marketing*, February 5, 1998, p.25.
70) 김신곤, 전게논문, 1999.

정보 등의 변수를 가지고 텔레마케팅을 수행할 경우 가능성이 높은 고객을 선별하기 위한 분류 모델을 개발하였는데, 이 분류모델을 사용할 경우 추출 대상 고객의 수에 제한이 없다면 무작위 추출에 의한 경우와 비교하여 텔레마케팅의 효율성을 높일 수 있음을 확인하였다.

또한 의사결정에 도움이 될 수 있는 지식을 획득하는 데이터 마이닝 기법을 사용하여 의사결정지원 시스템을 구성한 연구[71]에서는 데이터 마이닝 기법 중에서도 특성화 규칙을 사용하여 규칙을 생성하고 분석 트리를 구성하였고 계층적으로 분석이 용이한 트리구조를 선택하였다. 제안하는 트리는 기존의 의사결정 트리와 다르게 다각도로 분석한 결과를 시각화시킴으로써 사용자가 원하는 만큼 세분화하여 각 요소별로 다양한 분석을 할 수 있음을 제시하였다.

그리고 공급사슬관리(Supply Chain Management, SCM)에 데이터 마이닝을 활용한 연구[72]에서는 SCM의 마케팅, 생산 및 정보시스템 측면에서 데이터 마이닝을 활용한 의사결정지원시스템의 구조를 제시하였다.

또 다른 연구[73]로는 은행의 자료를 이용하여 신경망 모델의 예측력과 판별분석의 예측력을 비교평가한 자료를 보고하고 있다. 이 연구는 신경망 모델과 판별분석 기법을 이용하여 거래내역, 거래년수, 월평균소득, 총수신금액 등의 고객정보를 가지고 고객의 만족/불만족 및 수익성의 고/저를 예측하는 작업을 수행하였는데, 고객집단을 만족/불만족으로 구분해내는 적중률에 있어서 판별분석이 63.4%, 신경망 모델이 78.8%로 나타났으며, 고객의 수익성의 고/저를 구분해 내는 적중률에 있어서도 판별분석이 71.9%, 인공신경망 75.1%로 나타나 신경망 모델이 판별분석에 비해 예측력에 있어 보다 우수한 것으로 조사되었다.

위와 같이 살펴본 사례와 선행연구들은 데이터 마이닝이 다양하게 활용되고 있으며 데이터 마이닝을 이용한 DB마케팅 수행에는 상당한 효과를 가져다주고 있음을 시사하고 있다. 그러나, 개별 데이터 마이닝 기법들의 상대적인 영향력에 대해서는 일관성 있는 결과를 얻어 졌다기 보다는 비즈니스 목표나 연구 목적에 따라

---

71) 조성진 · 정인정, "데이터 마이닝을 이용한 의사결정지원 시스템", 가을 학술발표논문집 Vol. 26. No. 2., 한국정보과학회, 1999, pp.45~47.

72) 지원철 · 서민수, "데이터 마이닝을 활용한 공급사슬관리 의사결정시스템의 구조에 관한 연구", 경영정보학연구, 경영정보학회, 제8권 3호, 1998.12, pp.52~73.

73) 이용희, 전게논문, pp.1~36.

적용 기법이 서로 다른 방법들을 적용하고 있음을 보여주고 있다. 즉, 특정 데이터 마이닝 기법이 개별 기업의 특수한 환경에 따른 고객 데이터의 특성에 영향을 받기 때문에 일반화 될 수 있다고 할 수 없다. 따라서 개별 기업의 데이터 분석목적, 수집된 데이터의 성격, 산출되는 정보의 설명력, 사용의 용이성 등 여러 가지 기준을 고려하여 적절한 기법을 선정할 필요가 있다.

또한, DB마케팅의 대표적인 전략 실행이라고 할 수 있는 우수고객 우대프로그램의 경우 우수고객을 대상으로 DM, TM 등을 활용하는 일대 일 커뮤니케이션 활동을 시행할 수가 있는데, 우수고객 즉, 수익기여도가 높은 고객을 선별하기 위한 연구가 필요하다고 할 수 있다.

따라서, 본 연구에서는 수익기여도가 높은 고객을 선별하기 위해 DB마케팅 분석기법 가운데 RFM점수분석법을 이용하여 우수고객 분석을 위한 목표변수를 도출하였으며, 그 도출된 목표변수를 가지고 위에서 언급된 한 연구(김신곤, 1999)에서 제시한 의사결정나무 기법 가운데 CHAID 알고리즘을 적용하여 수익기여도가 높은 고객의 패턴을 발견하는 분류모델 개발하고, 개발된 모델의 성과측정은 리프트(Lift)를 이용하여 평가하였다.

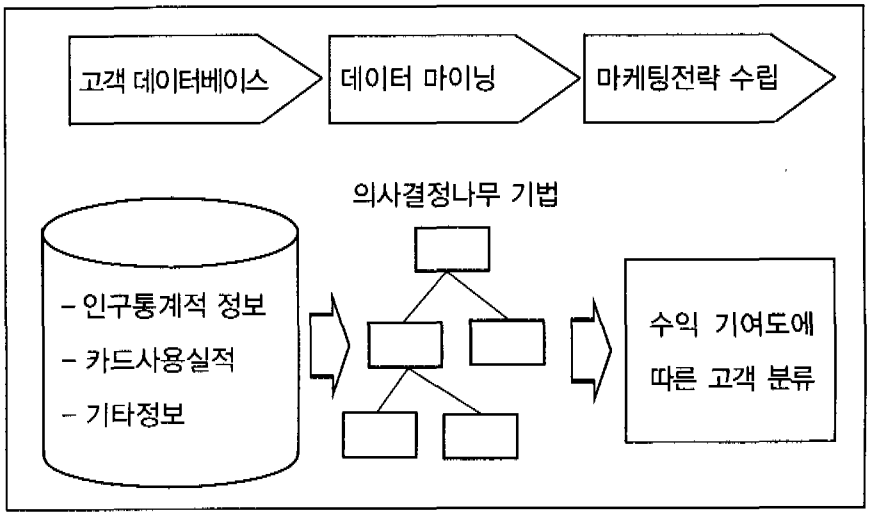
제 3 장 연구의 모델 및 실험 계획

제1절 연구모델

우리나라 은행의 경우 고객에게 매월 보내는 신용카드 이용대금 청구서를 보면, 이용내역, 신용한도 등 기본적 정보를 제외하고는 회원별 안내내용이나 통신판매 안내 팜플렛의 내용이 천편일률적으로 같다. 이러한 비차별적 매스 마케팅 접근은 고객의 다양한 욕구를 충족시키기 어려울 뿐 아니라 비용 측면에서도 효율적이지 못하다. 이러한 문제점을 극복하기 위해서는 고객유형별 욕구에 부합하는 DM(Direct Marketing), TM(Tele-Marketing) 등의 차별적 마케팅 전개가 필요하다.

따라서, 본 연구는 고객 데이터베이스로부터 수익 기여도가 높은 우량고객으로 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 차별화된 DM, TM 등의 다양한 마케팅 수행 대상이 되는 고객을 선별하는 모델의 개발에 관한 것이다. 마케팅 담당자에게 이 모델에 의하여 선별된 고객의 리스트를 제공함으로써 마케팅의 성공률을 높이는 효과적인 데이터베이스 마케팅에 관한 연구이다. 즉, 그 동안 마케팅 담당자나 은행의 의사결정자의 경험 및 직관에 의해 수행되었던 마케팅을 실제 고객 데이터베이스로부터 데이터 수집과 과학적인 분석에 입각해 이루어지게 함으로써 불필요한 비용을 줄이고 마케팅 효율을 더욱 높이는 등의 마케팅 업무의 성과를 향상시키고, 새로운 마케팅 전략 수립의 과학화를 목표로 하고 있다.

<그림 3-1> 연구의 모델



본 연구를 위한 실증적 실험에서는 A은행이 확보한 신용카드 고객 DB에서 수익 기여도가 높은 고객의 패턴을 발견하기 위해서 최근 1년간의 신용카드 사용실적이 있는 고객 레코드 중 결측치가 50% 이상인 레코드를 제외하는 등의 사전처리 작업을 선행하여 14,978명의 표본을 임의 추출하여 분석하였다.

실험에 사용된 고객 DB는 14,978개의 레코드였지만, 실제 한 은행이 구축한 고객 DB는 수 십만 레코드에서 수백만 레코드에 이르는 방대한 규모의 자료로서 수시로 기존고객이 이탈하거나 거래실적이 바뀌며 또한 신규고객이 유입되는 상황이기 때문에 모든 자료를 일괄적으로 분석하기에는 어려움이 많다. 따라서 자료분석의 효율성을 기하기 위해서는 일부 표본만을 대상으로 수익 기여도가 높은 고객을 분류하고 적용시키는 것이 바람직할 것이다.

만약, 지역, 연령, 성별 그리고 고객의 라이프 스타일에 따라 수익 기여도가 다르다면 이는 고객을 세분화할 수 있는 주요한 변수로 사용할 수 있다. 즉, 개인의 거래성격은 그 사람의 인구사회학적 특징이 반영되어 있기 때문에 이에 대한 마케팅 활동도 다르게 전개할 수 있을 것이다. 또한 특정요인에 의해 은행에 이익을 기여할 수 있는 특정 고객을 선별하여 마케팅 투자 노력을 집중할 수 있다면 투자 효율성은 극대화 될 수 있을 것이다.

따라서, 표본으로 추출된 고객 개인별 데이터의 항목은 예측변수인 인구통계적 변수, 거래실적 변수와 목표변수로 구성하였다. 목표변수(수익 기여도)는 제2장에서 언급한 RFM 분석방식을 이용하여 고객별 수익 기여도를 점수화(scoring)하고 이것을 기준으로 1,000점 이상의 고객 중 최근 1년간 거래정지 사실이 없는 고객을 수익 기여도가 높은 고객으로 분류하고, 나머지는 수익 기여도가 낮은 고객으로 이진분류 하였다.

연구에 사용된 항목 외에도 구축된 고객 DB의 여러 가지 항목에 따라 다양한 분석을 시도할 수 있겠지만, 여기서는 통계적 자료분석에 적절한 항목만을 선정하여 분석하였다. 그리고, 데이터의 형태에 따라 가장 적절한 데이터 마이닝 기법을 사용코자 하였다.

본 연구에서는 데이터 마이닝 기법 중에서 의사결정나무를 사용하였다. 의사결정나무는 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려대상이 되는지를 구별하는데 도움을 준다.[74] 그리고 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단하며, 이해하기 쉬운 규칙으로 전환될 수 있기 때문에 데이터 마이닝 기법으로서 의사결정나무를 사용하였다.

제2절 실험계획

본 실험의 계획은 다음과 같이 실행된다.

1) 준비작업

데이터를 처리하고 분석하기에 앞서 얻고자하는 정보를 정의하고, 이의 활용방안을 수립하며, 정보의 원천이 되는 데이터의 종류 등을 파악한다.

74) Metha, Manish, Jorma Rissanen. Rakesh Agrawal, "MDL-based Decision Tree Pruning", IBM, Almaden Research Center.

2) 데이터의 선택(Data Selection) 및 생성

구체적인 데이터 항목과 소재 등을 파악하고, 수집하여 통합하는 작업이다.

3) 데이터의 사전처리

① 데이터의 정제(Data Cleaning)

정확도를 높이기 위해 데이터에 존재하는 오류값이나 특이값을 보정하고, 결손값을 처리하며, 중복데이터를 제거하는 작업이다.

② 데이터의 보완(Data Enrichment)

정확도를 높이기 위해 분석하고자 하는 데이터의 양과 깊이를 늘이는 단계이다.

③ 데이터의 변환(Data Transformation)

데이터에 포함된 불필요한 레코드와 항목을 삭제하는 작업이다.

4) 데이터 마이닝 기법 선택 및 적용(Selection & Application of Data Mining Technique)

분석의 목적에 적합한 데이터 마이닝 기법을 선택하고 적용하는 단계이다. 데이터 마이닝 기법을 선택하기 전에 추가적으로 고려해야 할 점들은 다음과 같다.[75]

① 모형의 설명력 : 데이터 마이닝 모형이 제공하는 결과를 설명할 수 있는 능력
② 모형구축의 효율성 : 선택된 기법을 이용하여 데이터 마이닝 모형을 구축하는데 소요되는 노력과 시간
③ 모형의 정확성 : 모형이 제공하는 결과의 신뢰성

75) 장남식 · 홍성완 · 장재호, 전게서, p.100.

④ 기법의 보편성 : 선택된 기법을 다양한 데이터 마이닝 작업이나 데이터의 성격에 관계없이 적용할 수 있는 정도

⑤ 기법의 가용성 : 상용화된 제품의 종류 및 이들이 지원하는 O/S와 하드웨어 플렛폼의 다양성

5) 모형의 평가(Model Evaluation)

마지막 단계로서 데이터 마이닝 기법을 이용하여 구축한 모형이 과연 실제로 현장업무에 적용하기 적절한가를 판단하는 단계이다.

제3절 데이터의 사전처리(Data Preprocessing)

지식발견 프로세스를 통해 산출되는 정보의 품질은 정보의 원천이 되는 통합 데이터의 충실도에 달려 있다고 해도 과언이 아니다. 대개 데이터의 충실도란 데이터의 정확도, 데이터의 양(레코드의 수), 그리고 데이터의 깊이(항목의 수)에 의해 평가되는데, 데이터의 사전처리 단계에서는 이 중에서도 정확도를 높이기 위해 데이터에 존재하는 오류값이나 특이값을 보정하고, 결손값을 처리하며, 중복데이터를 제거하는 작업이 수행된다. 이처럼 양질의 데이터를 가지고 의미있는 데이터 분석을 위한 데이터의 사전처리에 대해 구체적으로 설명하면 다음과 같다.[76]

데이터 사전처리 과정은 실제적인 데이터 분석이 처리되기 전에 취해지는 모든 작업들로 구성되어 있다. 실제 원시 데이터 벡터 X_{ik}을 새로운 데이터 벡터 Y_{ij}로 전환시키는 변환작업은 필수이다.

$$Y_{ij} = T(X_{ik})$$

76) A. Famili, Wei-Min Shen, Richard Weber, Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis", Intelligent Data Analysis : Elsevier Science Inc., 1996.

이때 (i) Y_{ij}는 X_{ik}에서 "가치있는 정보"를 보존한다.

(ii) Y_{ij}는 X_{ik} 내 문제들 중 최소 하나의 문제점을 제거한다.

(iii) Y_{ij}는 X_{ik} 보다 더 유용하다.

위 관계식에서 i, j, k는 다음의 관계를 만족한다.($m \neq l$)

$i=1,\cdots,n$, n : 객체(objects) 수

$j=1,\cdots,m$, m : 사전처리 후의 특징(features)들의 개수

$k=1,\cdots,l$, l : 사전처리 전의 속성(attributes)/특징(features)의 수

가치있는 정보는 데이터 내에 존재하는 지식의 구성요소들을 포함하고 있으며 데이터 분석의 목표는 의미있는 방식으로 그것들을 발견하고, 명시하는 것이다. Fayyad, Piatetsky-Shapiro, 그리고 Smyth는 가치 있는 정보를 네 가지 속성으로 정의하였다. 타당성(valid), 혁신적(novel), 잠재적인 유용성(potentially useful), 그리고 이해 용이성(ultimately understandable)이 필요하다. 데이터 문제들은 데이터 분석 도구의 효율적 사용을 방해하거나 허용할 수 없는 결과들을 발생하는 결과를 발생시킬 수 있는 상황을 말한다.

데이터의 사전 처리는 다음과 같은 이유로 처리 되야 한다.

- 일정 데이터 타입의 분석 수행을 방해하는 데이터 문제의 해결
- 데이터의 특성 이해와 더욱 의미 있는 데이터 분석 수행
- 데이터 셋에서 더욱 의미 있는 지식 추출

1) 데이터의 문제점(Problems with the Data)

실제 데이터는 항상 문제들을 갖고 있다. 데이터의 문제점들은 다음과 같이 분류될 수 있다.

(1) 너무 많은 데이터 (Too Much Data)

① 충돌과 잡음 데이터(colision and noisy data)

데이터 충돌은 센서의 미작동, 데이터 전송 실패, 또는 부적절한 입력과 같은 이유 때문에 일어날 수 있고 이들 중 다수가 데이터 수집 시에는 알 수 없는 경우가 많다. 데이터의 잡음은 여러 가지 이유가 있다. 즉, 데이터 측정 또는 전송 오류, 데이터를 수집하는 시스템 또는 처리과정의 특성과 같은 이유에서 기인되기도 한다.

원인에 상관없이 충돌과 잡음은 반드시 찾아야 하고 문제해결을 위한 해결책 또한 찾아야 한다. 일반적으로 데이터의 잡음은 특성의 예측성을 약화시킨다.

② 특성 추출(Feature Extraction)

많은 측정치들이 있을지라도 이벤트가 거의 일어나지 않을 수 있다. 이들 측정치로부터 데이터는 이벤트의 의미있는 설명과 일치되어야한다. 이것은 적절한 데이터 사전처리가 없으면 매우 어려운 작업이다. 특성 추출은 프로세스 처리과정으로부터의 수치 데이터는 유용한 레이블과 일치시킬 수 있는 수치형-기호(numeric-symbolic) 해석을 예로 들 수 있다.

③ 방대한 데이터의 크기(Very large data sizes)

때때로, 방대한 데이터는 데이터 분석에 사용되는 하드웨어와 소프트웨어의 처리능력을 벗어날 수가 있다.

(2) 너무 적은 데이터(Too little data)

① 속성의 상실(Missing attributes)

상실(Missing) 또는 충분하지 못한 속성(attributes)은 대부분의 분석 시스템에서 학습과 같은 데이터 분석 업무를 복잡하게 할 수 있는 문제를 야기시킬 수 있다. 충돌과 결측치(missing attribute)는 여러 문제를 야기시킬 수 있다. 다음은 귀납적 데이터 분석 처리과정에서 일어날 수 있는 예이다.

- 의사결정 트리에서 속성의 부재는 동일하지 않은 길이가 되는 벡터를 야기시킨다. 두 가지 속성을 나타내는 두 벡터의 정보 값이 비교되거나 검증이 한 속성의 값에서 실행될 때 오차(bias)가 만들어질 수 있다.

- 많은 데이터 분석 애플리케이션은 트레이닝 데이터 셋과 테스트 셋으로 데이터를 쪼갤 수 있다. 비록 쪼갬(splitting) 처리가 여러 번 반복될 지라도, 결측치(missing attribute)는 결과의 부정확한 측정을 야기시킬 수 있다.

② 속성값의 상실(Missing attribute values)

이러한 경우 데이터 레코드는 완전하지 못하다. 즉, 결측치를 갖고 있다. 전통적으로 결측치의 20%이상이라면, 전체 레코드는 삭제시킨다.

(3) 잘못된 데이터 (Fractured data)

① 비호환 데이터(incompatible data)

데이터 호환(data compatibility)은 데이터가 여러 그룹에서 수집될 때 매우 중요하게 된다. 감지장치 데이터(sensor data)가 수집되고 분석되는 분야에서 특히 중요하다. 감지장치 데이터는 많은 텍스트(text)와 심볼(symbolic)속성들이 존재한다. 비호환 문제는 데이터 수집과정에서 텍스트 또는 자연어를 사용하여 표현하는 인간의 방법 때문에 야기된다.

② 데이터의 다양한 원천 (multiple sources of data)

거대한 회사에서 데이터는 여러 부서에 있고 서로 다른 형태로 흩어져 있다. 대부분의 경우 데이터는 상이한 소프트웨어 시스템을 사용하여 유지되고 획득된다. 데이터 수집의 목적, 깊이 그리고 표준은 전 회사에 걸쳐 다를 수 있다. 그 결과 하나 이상의 그룹으로부터 수집된 데이터가 분석이 필요할 때, 다양한 원천으로부터 데이터의 사용과 관련된 문제는 일어날 수 있다.

2) 데이터 분석을 위한 사전 준비(Preparation for data Analysis)

데이터가 가지고 있는 문제들이 해결되고 데이터가 준비되었을 때, 실제 데이터 분석을 시작하기 전에 여전히 처리해야 할 과정이 있다.

(1) 데이터 특성의 이해(Understanding the Nature of Data)

데이터가 가지고 있는 모든 알려진 문제들이 해결되었을 때, 데이터의 특성을 이해하는 것은 많은 방법에서 유용하다.

- 데이터 분석 도구의 적절한 사용과 크고 복잡한 데이터 셋의 적절한 해석은 사람의 사고 능력으로는 어렵다. 그러므로 데이터를 잘 이해하기 위한 데이터 사전처리를 수행하는 것은 유용하다. 예로 데이터 시각화와 주요 구성요소 분석을 들 수 있다.

- 대부분의 데이터 분석 도구들은 데이터 특성과 관련하여 한계점을 갖고 있다. 데이터 분석의 선택과 결정을 위해 데이터의 특성들을 이해하는 것은 중요하다.

(2) 세밀한 데이터 분석을 위한 데이터 사전 처리(Data Preprocessing for In-depth Data Analysis)

일반 데이터 분석 도구와 기법들은 모든 애플리케이션에서 충분하지 못한 수준으로 분석을 위한 수단을 제공하고 있다. 세밀한 데이터 분석은 실제 데이터 분석이 시작되기 전에 적합하게 사용되어야 하는 데이터 사전 처리를 위한 추가적 지원 도구를 필요로 한다. 예를 들어, 데이터가 귀납 법칙을 통해 분석되고 단일 이벤트를 표현하는 레코드의 데이터가 구성된다면, 데이터의 일시적이고 다른 형태의 트렌드(trends)는 귀납 처리를 통해 적합하게 인식될 수 없을 것이다. 그러나, 레코드가 단일 이벤트라기 보다는 일반적인 트렌드를 나타내기 위해 전환된다면, 데이터 분석의 결과는 더욱 의미 있을 것이다. 세밀한 데이터 분석을 위한 사전 처리

과정의 다른 형태는 다음과 같다.

① 새로운 특성(new feature)의 수작업 추가 또는 자동 추가

② 일반적으로 측정될 수 없는 매개변수 생성을 위한 데이터 시뮬레이션

③ 다양한 원천(source)으로부터 추출된 데이터를 통합하기 위해 수행되는 데이터 융합

④ 양적인 데이터 보다는 질적인 데이터를 사용하고 생성의 지원을 제공하는 다차원 분석

(3) 데이터 사전 처리과정의 고려사항

다음은 데이터 마이닝을 위해 사전 처리되어야 할 때 고려되어야 할 중요한 사항들이다.

① 비록 데이터 사전 처리과정이 의미 있는 데이터 분석을 수행하기 위해 많은 애플리케이션에서 필요하고 유용하다 할 지라도 적합한 데이터 마이닝 기법이 선택되지 않았다면, 분석에서 발전되어야 할 유용한 정보의 손실 또는 변질된 결과를 낳게 될 것이다.

② 의미 있는 데이터 사전 처리과정을 수행하기 위해, 현업 담당자 또는 전문가와 함께 데이터 사전 처리과정이 이루어져야 한다. 이들의 참여는 데이터 사전처리 과정 기법들의 타당성을 검증하기 위한 유용한 피드백의 결과를 가져올 수 있다.

③ 대부분의 데이터 사전 처리과정은 반복적이다. 이것은 데이터 제거 또는 데이터 선택과 같은 데이터 사전 처리기법이 최상의 분석 결과를 얻을 수 있을 때까지 수많은 반복에서 이용된다는 것을 의미한다.

이상과 같이 데이터의 문제와 처리 방법 등을 살펴보았다. 이와 같이 데이터 사전 처리는 다방면에서 도움이 된다. 신경망을 사용하는 분류(classification)에서, 부적절한 데이터 제거는 더 적은 데이터 셋과 부적절한 데이터에 의해 야기되는 혼돈을 감소시킴으로써 더 빠른 학습이 가능하다. 신경망과 같은 애플리케이션에서 일반적으로 결과의 정확성 대 단순성이 상관(trade-off)관계에 있다. Fayyad, Piatetsky-Shapiro, 그리고 Smyth는 데이터베이스 프로젝트로부터 어떤 지식발견의 필수 사항으로써 데이터 사전 처리 기술들의 사용을 강조하였다.

제 4 장 실험결과의 분석

제1절 자료의 생성 및 사전처리

1. 실험 준비작업

본 실험의 목표는 A은행의 고객 데이터베이스로부터 수익 기여도가 높은 우량 고객으로 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 효율적 마케팅 수행을 위한 고객 분류모델 개발이다. <표 4-1>는 실험 목표와 주요성공요소 외에 필요정보, 정보활용방안, 그리고 이러한 정보를 도출하는데 필요한 데이터를 보여주고 있다.

<표 4-1> 지식발견 프로세스 준비작업의 산출물

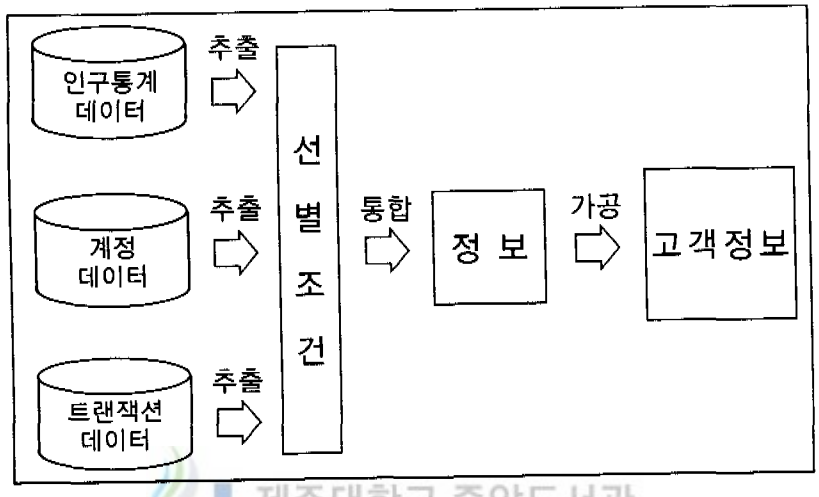
실험 목표	주요성공요소	필요정보	활용방안	필요데이터
효율적 마케팅 수행을 위한 고객 분류 모델 개발	-수익기여도에 따른 고객분류 -우량고객의 예측력 강화 -차별적 마케팅 전략 수립 -우량고객 유지 전략 강화	수익 기여도 높은 고객의 특성 정보	-캠페인 수행 대상 리스트 추출 -우량고객에 대한 우대 프로그램 개발	-인구통계적 데이터 -계정 데이터 -거래내역 데이터

2. 실험자료의 생성 및 분할

정보분석의 중요성을 역설하는 과정에서 두 가지 대응을 경험하게 된다. 첫째는 "분석할 데이터가 없다."고 걱정하는 경우이다. 가장 우려하는 부분은 고객원장의

데이터이다. 직업, 수입, 거주지 등 고객에 대한 인구·사회학적 데이터는 당연히 부실하다. 대안으로 거래내역(Transaction Log) 데이터에 관심을 집중하는 것이 하나의 방법일 수 있다. 개인의 거래성격은 그 사람의 인구·사회학적 특징이 반영되어 있기 때문이다.[77]

<그림 4-1> 고객정보 생성



따라서, 본 연구에서는 A은행의 신용카드 고객DB로부터 최근 1년간 사용실적이 있는 고객 데이터를 대상으로 <표 4-2>과 같은 데이터 속성을 갖는 인구통계적인 데이터 및 고객의 기본적 데이터를 추출하였고, 고객별 점수화를 위하여 <표 4-3>과 같이 연구에 필요한 레코드를 거래내역원장(Transaction Log)로부터 데이터를 추출하여 실험의 입력 데이터로 생성하였다.

<표 4-3>에서 입력 레코드의 항목 중 RFM항목은 2장에서 서술한 RFM 점수분석법을 고객의 거래내역 데이터에 적용하였으며, RFM 점수 계산방식으로는 <그림 4-2>와 같다.

신용카드업무에서 일시불거래와 할부거래, 현금서비스의 대표적 3가지 부문을 중심으로 RFM 스코어링(Scoring)을 적용하여 <그림 4-2>에서와 같이 일시불거래(MS)와 할부거래(IS), 현금서비스(CS)별로 각각의 최근 구매월, 구매회수, 구매금

77) 박태원, "성공적인 정보분석을 위한 전략방안", IT BUSINESS, 1999. 5.

액의 가중치를 주었다. 이러한 가중치를 기준으로 각 부문별 RFM점수를 산출할 수 있으며, 이를 모두 합하면 개별고객의 총 RFM점수가 유도되는 것이다. 고객 개개인에 대해서 RFM점수가 계산되면, 이 점수에 따라 모든 고객을 순서대로 배열할 수 있게 된다. 이는 결국 고객을 수익 기여도에 따라 수익 기여도가 가장 높은 사람부터 가장 낮은 사람까지 순서를 매기는 것이나 마찬가지이다.[78]

이렇게 유도된 RFM 점수를 기준으로 1,000점 이상의 고객 중 최근 1년간 연체 등의 사유에 의한 거래정지 횟수가 없는 고객을 수익 기여도가 높은 고객으로 <표 4-2>의 목표변수(Loyalty)를 '1'로 처리하고, 나머지 레코드는 수익 기여도가 낮은 고객으로 목표변수(Loyalty)를 '0'로 처리하였다.

<표 4-2> 실험 데이터의 레코드 속성

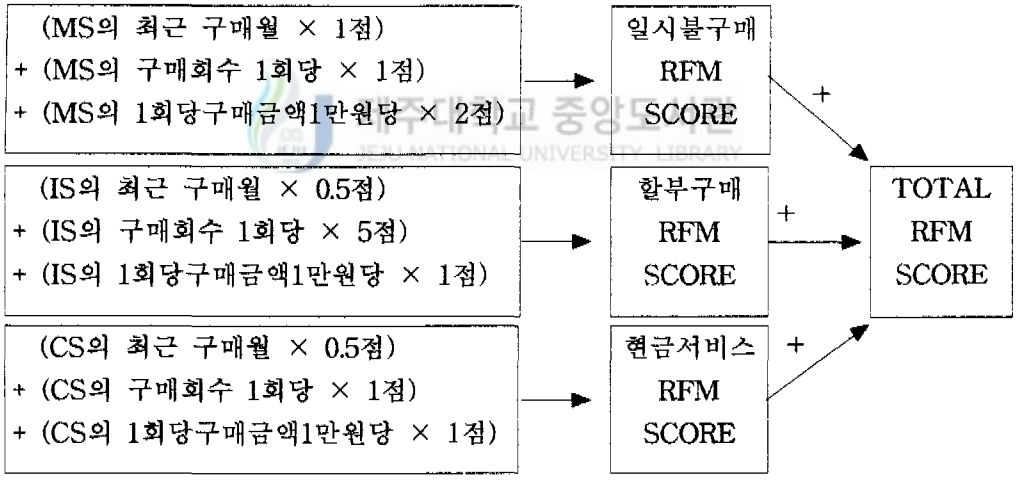
Variable	Label	Description	Type	Measurement	Model Role
ID	고객ID	고객ID	char	nominal	id
JOB	직업코드	A:공무원, B:교육기관종사자 C:금융기관종사자 D:회사원, E:자영업자, F:기타	char	nominal	input
AGE	나이		num	interval	input
GENDER	성별	M : 남성, F : 여성	char	binary	input
KIND	카드종류	1:일반 2:특별	num	binary	input
TXDAY	결제일	10일, 17일, 25일	num	nominal	input
LOCATION	주거지역	A: A지역, B: B지역, C: C지역, D: D지역, E: E지역, F: F지역, G: G지역 H: H지역	char	nominal	input
CARDLOAN	카드론 사용여부	1 : 사용 0 : 미사용	num	binary	input
LOYALTY	수익 기여도	1:고 0 :저	num	binary	target

78) 김기서, 전게서, p.138.

<표 4-3> RFM 점수계산을 위한 필요 데이터 항목

Variable	Description	Type
ID	고객ID	numeric
BLCNT	거래정지횟수	numeric
MSAMT	일시불(MS) 1년간 사용액	numeric
MSCNT	일시불(MS) 1년간 사용빈도	numeric
MSMONTH	일시불(MS) 최근 사용월	numeric
ISAMT	할부(IS) 1년간 사용액	numeric
ISCNT	할부(IS) 1년간 사용빈도	numeric
ISMONTH	할부(IS) 최근 사용월	numeric
CSAMT	현금서비스(CS) 1년간 사용액	numeric
CSCNT	현금서비스(CS) 1년간 사용빈도	numeric
CSMONTH	현금서비스(CS) 최근 사용월	numeric
RFM	RFM Score	numeric

<그림 4-2> 신용카드업의 고객별 RFM점수 계산



자료 : 상게서, p.139.

데이터 마이닝은 많은 변수를 가지고 있는 대규모의 데이터를 대상으로 하며 다양한 방법론에 의한 분석을 포함하고 있기 때문에, 모형의 타당성을 평가하고 여러 모형을 비교 검토하는 작업이 필요하다. 이를 위한 한 가지 전략은 데이터를 분석용(training), 평가용(validation), 검증용(test) 데이터로 분할하여, 분석용 데이터를 이용하여 모형을 구축하고 평가용 또는 검증용 데이터를 이용하여 모형의 비교와

최종적인 평가를 수행하는 것이다.[79]

따라서 본 연구에서는 모형 적합의 예비단계에 사용되는 데이터 셋(data set)을 분석용(training) 데이터 셋으로 70%, 모형의 적합도를 평가하기 위해 평가용(validation) 데이터 셋을 30%로 분할하였다.

데이터의 분할은 상호 배반적인(mutually exclusive) 데이터 셋을 제공한다. 여기서 상호 배반적이라는 것은 구분된 데이터 셋들은 관찰치를 공유하지 않는다는 말이다. 입력 데이터를 분할하게 되면 모형 적합의 예비단계에서 시간을 줄일 수 있다. 그러나, 크기가 작은 데이터 셋을 분할하게 되면 각 역할별 데이터 셋의 크기도 작아져 모형의 적합도와 일반성을 감소시킬 수도 있다.[80]

본 연구에서 데이터를 분할하기 위한 표본 추출 방법은 입력 데이터로부터 전체 모집단의 층(Strata)을 형성하는 변수를 기준으로 표본 내에서 전체자료의 층간 비율을 지키기 위해 층화임의추출(stratified random sampling) 방법을 사용하였다.

이러한 방법은 적합시킬 모형의 분류정확도(classification precision)를 향상시킬 수 있다.[81]

3. 실험자료의 사전처리

본 연구에서 사용된 원천 데이터(raw data)는 A은행의 신용카드 고객 데이터베이스로부터 추출한 데이터로써 이중에는 자세한 수준까지 기록된 데이터가 있는 반면 일부 항목이 시스템간 연동성 및 표준화 결여로 인해 동일 개체의 코드구분이 다른 경우와 결측치, 이상치 등이 다소 발견되었다. 또한 고객의 세부 정보를 추출할 수 없다는 점 때문에 데이터를 확보하였다고 하더라도 업무 특성상 그 값을 알 수 없는 경우가 많았다. 또한 소수 몇 개의 관측치가 일반적인 다른 관찰치보다 극단적으로 크거나 작은 값을 가지기 때문에, 구축하고자 하는 모형을 크게 왜곡시킴으로서 구축된 모형 자체를 무의미하게 할 수 도 있다.

이러한 이유들로 인하여 본 연구에서는 이 원시 데이터를 데이터 마이닝 작업에

79) 강현철 외5, 전게서, p.32.
80) 최종후 외4, 전게서, 1999, p .44.
81) 상게서, 1999, p.45.

앞서 모형에 영향을 미칠 수 있는 이상치를 제거하였다. 그러나 이상치로 간주되는 값들을 그냥 없애버릴 경우, 이 값들은 결측치로 간주되어 정보의 손실 문제가 발생하므로 이를 방지하기 위하여 이상치를 적절한 다른 값으로 대체하였다. 이상치 뿐만 아니라 데이터 셋 내의 결측치 또한 적절한 값으로 대체하는 등의 분석용 데이터 셋에 대한 보완 하였다. 즉, 분석에 불필요한 항목을 제거하고, 보완하는 사전처리 작업을 데이터 마이닝 실행에 앞서 수행하였다.

제2절 데이터 마이닝 실행

1. 데이터 마이닝 기법 선택

본 연구에서는 데이터의 형태에 따라 가장 적절한 데이터 마이닝 기법을 사용코자 하였다.

모델은 어떤 고객이 수익 기여도가 높고 어떤 고객이 수익 기여도가 낮은 고객일 것이라는 분류 기능을 제공해야 하며, 이와 더불어 수익 기여도가 높고 낮은 사유를 의사결정자에게 합당한 근거를 제시할 수 있어야 한다. 따라서, 데이터 마이닝 기법 가운데 모델에 대한 설명력이 뛰어난 의사결정나무를 데이터 마이닝 기법으로 선택하였다. 또한 실험에 사용된 변수들의 대부분이 범주형 변수이므로 의사결정나무 기법 가운데 CHAID 알고리즘을 사용하였다.

CHAID는 변수의 성격이 범주형 데이터이고 예측변수(predictor variable)와 결과변수간의 관계를 찾아야 할 때 가장 유용하다.[82]

의사결정나무는 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다.[83] 그리고, 의사결정나무는 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단하며, 이해하기 쉬운 규칙으로

82) Pyle, Dorian, "Putting Data Mining In Its Place", Database Programing & Design, March 1998.
83) Metha, Manish, Jorma Rissanen, Rakesh Agrawal, *op. cit.*

전환될 수 있기 때문에 본 연구는 데이터 마이닝 기법으로서 의사결정나무를 사용하고 있다.

2. 분류모델의 개발

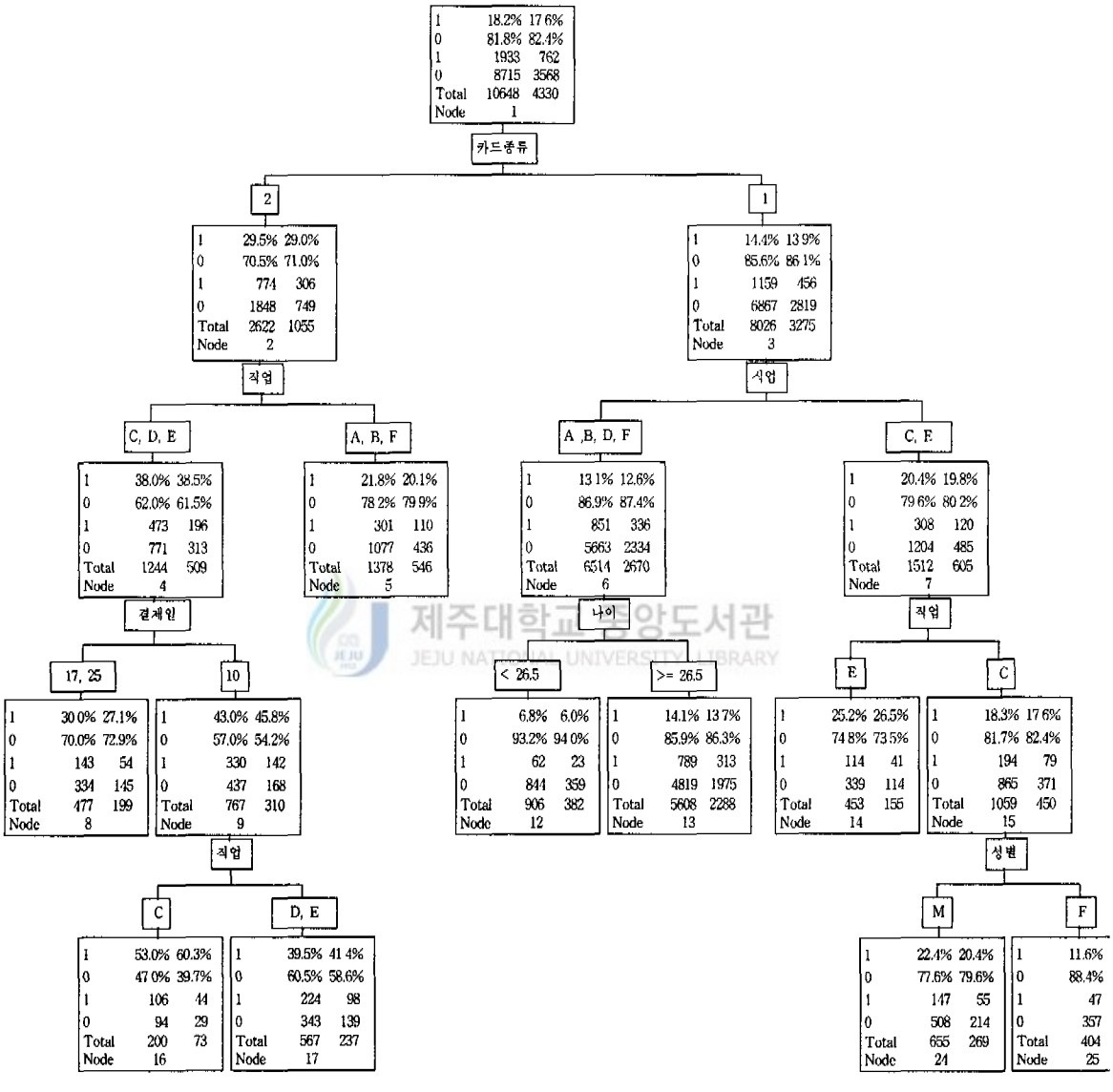
추출된 데이터는 총 14,978건이며, 이 가운데 수익 기여도가 높은 고객, 즉 최근 1년간 거래정지 횟수가 없으면서 RFM점수가 1,000점 이상인 고객 레코드는 2,695건으로 전체 데이터 셋(data set)의 반응 비율은 18.0%(2,695 / 14,978 × 100 = 18.0%)이다. 모델을 개발하고 검증하기 위하여 14,978건의 전체 데이터 셋을 모델 개발을 위한 트레이닝 데이터 셋(training data set)과 개발되어진 모델을 평가하기 위한 평가용 데이터 셋(validation data set)으로 나누고, 모델을 이용하여 평가용(validation) 데이터 셋의 부류 값을 예측한 결과를 각 레코드의 실제 부류 값과 비교하는 정오분류행렬(Confusion Matrix)표를 이용하여 검증하였다.

고객 데이터의 트레이닝 데이터 셋은 전체 데이터 셋의 70%로 총 10,648건(14,978 × 0.7 = 10,648)이다. 이 가운데 수익 기여도가 높은 레코드 건수는 1,933건, 수익 기여도가 낮은 레코드 건수는 8,715건으로, 트레이닝 데이터 셋의 반응 비율은 18.2%(1,933 / 10,648 × 100 = 18.2%)이다. 평가용 데이터 셋은 전체 데이터 셋의 30%를 차지하는 4,330건(14,978 × 0.3 = 4,330)이다. 이 가운데 수익 기여도가 높은 레코드 건수는 762건, 수익 기여도가 낮은 레코드 건수는 3,568건으로, 평가용 데이터 셋의 수익 기여도가 높은 고객의 비율은 17.6%(762 / 4,330 × 100 = 17.6%)이다.

<표 4-4> 데이터 셋의 분할

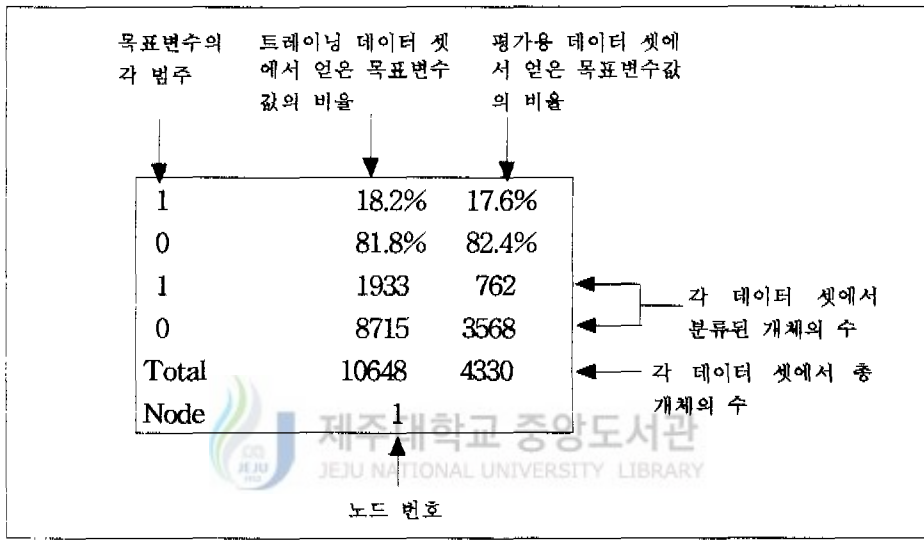
구 분	수익기여도 높은 고객수	수익기여도 높은 고객(%)	수익기여도 낮은 고객수	수익기여도 낮은 고객(%)	합계	전체 비율 (%)
전체 데이터 셋	2,695	18.0	12,283	82.0	14,978	100
트레이닝 데이터 셋	1,933	18.2	8,715	81.8	10,648	70
평가용 데이터 셋	762	17.6	3,568	82.4	4,330	30

<그림 4-3> 의사결정나무의 분류모델



수익 기여도가 높은 고객을 분류하기 위한 분류모델(classification model)을 개발하기 위하여 의사결정나무 기법 가운데 하나인 CHAID 알고리즘을 사용하였다. CHAID 알고리즘을 데이터 셋에 적용한 결과 <그림 4-3>과 같은 나무 구조를 얻을 수 있다. 나무 구조도를 해석하고 설명의 편의를 위하여 각각의 노드 하단에 노드번호를 부여하였다. 각 마디에 나타나는 통계량들은 <그림 4-4>와 같다.

<그림 4-4> 각 마디에서의 통계량



트레이닝 데이터 셋으로부터 나무를 생성하기 위하여 의사결정나무 알고리즘의 분리 수준은 카이제곱 검정(Chi-square test)을 사용했으며, 정지 규칙(stopping rules)을 적용할 때 노드 깊이(node depth)는 4로 설정하였다. 부모 노드(parent node)와 자식 노드를 형성하기 위한 필요조건으로 최소 데이터 수를 부모 노드는 400건, 자식 노드는 200건으로 각각 설정하였다. 즉, 하나의 노드로부터 나올 수 있는 가지의 수를 최대 2로 이진(binary) 분류로 정하였기 때문에 잎(leaf)의 최소 관측수는 200(400 / 2 = 200)이다.

(잎의 최소 관측수) ≤ (가지를 치기위해 요구되는 관측치의 개수) / (하나의 노드로부터 나올 수 있는 가지의 수)

또한 의사결정나무 노드의 최적화를 위하여 입력 변수(데이터 속성)와 결과 변수(목표 변수)간의 상관관계가 높은 변수를 뿌리노드에 가장 가까운 입력변수로 선정하였다. 의사결정 나무를 형성하기 위한 유의수준은 0.05로 하였다.

제3절 데이터 마이닝 실험결과 분석

1. 실험결과의 분석

개발된 분류모델은 <그림 4-3>에서 보는 바와 같이 뿌리 노드를 포함하여 17개의 노드로 표현되었으며 나무의 모든 노드는 최근 1년간 사용실적이 있는 고객 중에서 수익 기여도가 높은 고객과 수익 기여도가 낮은 고객으로 구분하여 각각의 분포를 보여주고 있다.

생성된 나무 구조도에서 알 수 있듯이, 첫 번째 분리변수는 카드종류(Kind)임을 알 수 있다. 분리된 결과를 보면 카드종류가 2(특별카드)인 경우가 수익 기여도가 높다는 것(29.5%)을 알 수 있고, 카드종류가 1(일반카드)은 수익 기여도가 낮은(14.4%)것으로 나타나고 있다.

다음으로 카드종류가 2(특별카드)에 해당하는 마디(node 2)를 보면, 직업으로 분리가 이루어졌는데, 여기서 특이한 점은 어떤 범주의 경우와 함께 그룹화가 되어 있음을 볼 수가 있다. 즉, 직업이 C, D, E와, A, B, F가 그룹화 되었다. 이러한 경우가 발생하는 이유는 범주간 통계적 차이는 없는 경우로서, 병합된 범주는 통계적 관점에서 근본적으로 동일하다고 볼 수 있다. 마찬가지로 카드종류가 1인 마디는 직업이 A, B, D, F와 직업이 C, E인 마디로 분리되었는데, 여기서 직업이 C, E인 마디(node 7)가 수익 기여도가 높은 고객 비율(20.4%)이 높은 것으로 나타났다. 그 다음으로 직업이 C, D, E로 병합된 마디(node 4)에서는 결제일이 17일과 25일로 병합된 마디(node 8)와 결제일이 10일인 마디(node 9)로 다시 분리가 이루어졌다. 결제일이 10일인 범주가 다른 범주들 보다 수익 기여도가 높은 고객 비율(43.0%)이 높음을 알 수 있다. 마찬가지로 카드종류가 1이면서 직업이 A, B, D, F인 마디

(node 6)는 다시 나이로 분리가 이루어졌는데, 26.5세 이상인 경우가 수익 기여도가 높은 고객 비율(14.1%)로 다소 높게 나타났다. 그리고, 카드종류가 1이면서 직업이 C, E인 마디(node 7)는 다시 직업이 C와 E로 분리되었다. 그 중에서 직업이 E인 마디(node 15)가 수익 기여도가 높은 고객 비율(25.2%)로 병합된 마디(node 7)보다 다소 높게 나타난 것을 알 수 있다. 다음으로 카드종류가 2이고 직업이 C, D, E면서 결제일이 10일인 마디(node 9)는 다시 직업이 C인 마디(node 16)와 직업이 D, E인 마디(node 17)로 분리되었는데, 직업이 C인 마디(node 16)가 수익 기여도가 높은 고객 비율(53.0%)로 마디들 중에서 가장 높게 나타났다. 마찬가지로 카드종류가 1이고 직업이 C인 마디는 다시 성별로 분리가 이루어 졌는데, 남성이 여성보다 수익 기여도가 높은 고객 비율(22.4%)이 더 높게 나타난 것을 알 수 있다.

<그림 4-3>의 분류모델을 살펴보면 뿌리노드로부터 트레이닝 데이터 셋의 전체 10,648건 중 수익 기여도가 높은 고객의 비율은 18.2%임을 알 수 있다. 다음으로 16번, 17번 노드의 반응율은 각각 53.0%와 39.5%로서 다른 노드 보다 높은 것을 알 수 있다. 이 노드들은 전체집단 가운데 직업이 금융기관 종사자(직업코드 : C), 회사원(직업코드 : D), 자영업자(직업코드 : E) 이면서 결제일이 10일인 특별카드를 발급 받은 집단으로서 분류되어 있다. 이 와는 반대로 나이가 26.5세 미만이고 직업이 공무원, 교육기관 종사자, 회사원 등인 일반카드를 발급 받은 집단은 수익 기여도 낮은 고객 비율(93.2%)이 상당히 높게 나타났다.

2. 분류모델의 검증

의사결정나무 알고리즘을 이용하여 개발된 분류모델이 얼마나 타당성을 가지고 있는지를 평가하는 것은 매우 중요하다. 개발된 분류모델을 평가용 데이터 셋에 적용하여 봄으로써 모델의 적합도, 타당성 및 그 성능을 평가할 수 있다. 트레이닝 데이터 셋에서 만들어진 모델을 평가용 데이터 셋에 적용하여 동일하거나 거의 유사한 결과를 나타낸다면, 이것은 모델이 데이터를 잘 표현하고 있다는 것을 의미하며 현재 데이터와 유사한 성격을 가진 다른 데이터에 대해서도 유사한 결과를 나타낼 것이라고 유추할 수 있다.

정보 이익 요약표(Information Gain Summary)는 분류모델이 형성한 각 노드의

타당성을 평가하기 위해 사용될 수 있다.[84]

<표 4-5>는 트레이닝 데이터 셋의 정보 이익 요약표이다. 정보 이익 요약표는 목표변수의 각 개체들이 각 마디에서 어떻게 분포되고 있는지를 알려주며, 이를 통해 기존 마디의 병합과 새로운 마디의 분리에 대한 정보를 제공하여 주고 있다.

- Node : 나무 구조도에 나타난 노드의 번호
- Node n : 노드에 속하는 개체의 빈도
- %Node : 노드에 속하는 개체의 빈도 / 전체 개체의 빈도
- Resp n : 노드에 속하는 목표 변수의 특정 범주(수익기여도 높음)의 빈도
- %Resp : 노드에 속하는 목표 변수의 특정 범주의 빈도 / 노드에서 전체 빈도
- %Captured Resp : 노드에 속하는 목표변수의 특정 범주의 빈도 / 전체에서 목표변수의 특정 범주의 빈도
- Lift : 노드에서의 목표변수의 비율 / 전체에서의 목표변수의 비율

<표 4-5> 트레이닝 데이터 셋에 대한 정보이익 요약표

Training data set												
Non-Cumulative							Cumulative					
node	node n	%node	resp n	%captured resp	%resp	lift	node n	%node	resp n	%captured resp	%resp	lift
16	200	1.88	106	5.48	53.00	2.91	200	1.88	106	5.48	53.00	2.91
17	567	5.32	224	11.59	39.51	2.17	767	7.20	330	17.07	43.02	2.36
8	477	4.48	143	7.40	29.98	1.65	1244	11.68	473	24.47	38.02	2.09
14	453	4.25	114	5.90	25.17	1.38	1697	15.94	587	30.37	34.59	1.90
24	655	6.15	147	7.60	22.44	1.23	2352	22.09	734	37.97	31.21	1.71
5	1378	12.94	301	15.57	21.84	1.20	3730	35.03	1035	53.54	27.75	1.52
13	5608	52.67	789	40.82	14.07	0.77	9338	87.70	1824	94.36	19.53	1.07
25	404	3.79	47	2.43	11.63	0.64	9742	91.49	1871	96.79	19.21	1.06
12	906	8.51	62	3.21	6.84	0.38	10648	100.00	1933	100.00	18.15	1.00
계	10648		1933									

트레이닝 데이터 셋의 정보 이익 요약표 <표 4-5>과 평가용 데이터 셋의 정보 이익 요약표 <표 4-6>에 나타난 %Captured Resp열과 %Resp 열의 수치를 비교

84) 김신곤, 전게논문, 1999, p.96.

하여 보면 매우 유사한 결과를 보여주고 있다.

<표 4-6> 평가용 데이터 셋에 대한 정보이익 요약표

Validation data set												
Non-Cumulative							Cumulative					
node	node n	%node	resp n	%captured resp	%resp	lift	node n	%node	resp n	%captured resp	%resp	lift
16	73	1.69	44	5.77	60.27	3.42	73	1.69	44	5.77	60.27	3.42
17	237	5.47	98	12.86	41.35	2.35	310	7.16	142	18.64	45.81	2.60
8	199	4.60	54	7.09	27.14	1.54	509	11.76	196	25.72	38.51	2.19
14	155	3.58	41	5.38	26.45	1.50	664	15.33	237	31.10	35.69	2.03
24	269	6.21	55	7.22	20.45	1.16	933	21.55	292	38.32	31.30	1.78
5	546	12.61	110	14.44	20.15	1.14	1479	34.16	402	52.76	27.18	1.54
13	2288	52.84	313	41.08	13.68	0.78	3767	87.00	715	93.83	18.98	1.08
25	181	4.18	24	3.15	13.26	0.75	3948	91.18	739	96.98	18.72	1.06
12	382	8.82	23	3.02	6.02	0.34	4330	100.00	762	100.00	17.60	1.00
계	4330		762									

정보 이익 요약표는 비누적(Non-cumulative) 통계량과 누적(Cumulative) 통계량으로 구성되어 있고, 마디들은 %Resp 열에 의하여 높은 순으로 정렬되어 있다.

먼저 비누적 통계량에 관해서 살펴보면, <표 4-6>에서 먼저 첫 번째 열에 나와 있는 Node는 의사결정나무 그림에서 보여주는 마디번호(node number)를 의미하는데, 마디번호가 16인 것은 직업이 C(금융기관 종사자)이면서 결제일이 10일인 특별카드를 발급 받은 고객들로 구성되어 있음을 알 수 있다. 다음으로, 두 번째와 세 번째 열에 나와있는 Node n과 %Node는 해당 마디번호에서의 자료수와 백분율을 나타낸다. 마디번호가 16인 경우를 보면 자료수가 73이고 백분율은 1.69%(73 / 4330 ×100 = 1.69)임을 알 수 있다. 또한, 네 번째와 다섯 번째 열인 Resp n과 %Captured Resp는 해당 마디번호에서 목표변수의 원래범주와 같게 올바로 분류된 자료수와 백분율을 의미하는데, 마디번호 17로 분류된 관찰치 개수가 237개인데 그 중에 98개가 목적범주(수익 기여도 높은 고객)와 같게 분류되어 12.86%(98 / 762 × 100 = 12.86)로서 전체 수익 기여도가 높은 고객 데이터(762건)의 12.86%가 17번 마디에 할당되어 있음을 알 수 있다. 다음으로 여섯 번째인 %Resp는 (Resp n) / (Node n)의 비율을 뜻하며, 마지막으로 일곱 번째인 Lift는 해당 마디에서 목

적범주가 제대로 분리된 자료수의 비율(%Resp)이 전체 자료에서의 목표범주의 비율과 얼마나 비교되는지의 측도로 주어진다. 마디 16은 해당마디에서의 수익 기여도가 높은 고객 비율이 60.27%이고, 전체자료에서의 수익 기여도 높은 고객 비율이 17.6%이므로 Lift는 3.42(60.27 / 17.6 = 3.42)가 되는 것이다. 따라서 은행에서 마케팅 업무 담당자는 마디 16과 같은 특성을 갖는 고객들을 찾아내어 고객과의 지속적 관계를 유지해야 할 것이다. 반대로, 마디 12(나이가 26.5세 미만이고 직업이 A, B, D, F(공무원, 교육기관종사자, 회사원, 기타)인 일반카드를 발급 받은 고객)와 같이 마디 내의 수익 기여도 높은 비율(6.02%)이 전체자료의 비율(17.6%)보다 훨씬 작아 Lift가 0.34(6.02% / 17.6% = 0.34)인 경우 은행 마케팅 담당자들은 마디 12와 같은 특성을 갖는 고객들에게는 역마케팅(de-marketing) 등을 하는 의사결정을 내려야 할 것이다.

누적통계량에 대한 결과를 살펴보면, 누적통계량은 최적의 고객 세분화를 이룸으로써, 수익 기여도가 높은 고객들의 집단을 얼마나 잘 찾아주는지에 대한 정보를 알려준다. <표 4-6>에서 가장 수익 기여도에 대한 반응률이 좋은 마디(node 16)만을 취한다면, 전체자료의 1.69%(73 / 4330 × 100 = 1.69%)의 자료를 접촉하여 전체 수익 기여도 높은 고객의 5.77%(44 / 762 × 100 = 5.77%)의 비율을 얻을 수 있게 된다. 여기서 추가적으로 다음 마디(node 17)를 포함한다면, 전체자료의 7.16%(310 / 4330 ×100 = 7.16%)를 접촉하여 전체 수익 기여도 높은 고객의 18.64%(142 / 762 × 100 = 18.64%)의 비율을 얻게 되는 것이다. 다음으로 마디 8을 포함하면, 11.76%의 접촉으로부터 전체 수익기여도 높은 고객의 25.72%까지로 비율을 증가시키게 되고, 이러한 방식으로 마디 5까지를 포함하면, 34.16%의 접촉으로부터 전체 수익 기여도가 높은 고객의 52.76%까지로 반응률을 증가시키게 된다. 그러나 이 단계는 교차점을 지나면서부터 전체자료의 접촉 비율과 포함마디까지의 반응률과의 차이가 줄어들게 된다. 실제로 마디 13을 포함하면 수익 기여도가 높은 고객의 비율은 93.83%이지만, 전체자료의 87.00%나 접촉을 해야 하는 비효율적 문제가 발생함을 알 수 있다.

이상과 같이 정보 이익 요약표는 의사결정에 대해 매우 유익한 정보를 주고 있다. 물론, 얼마나 많은 세분화 그룹(마디)이 필요할 것인가에 대한 결정이 필요할 수 있을 것인데, 그러한 결정은 경영자가 시장에서 마케팅을 목표로 하는 고객의

비율이나 공략하기 원하는 고객의 비율 등을 고려하여 결정하면 된다. 즉, 전체 시장에서 고객의 몇 %를 공략하기 원하는지를 먼저 정한 후 정해진 결정에 맞추어 정보 이익 요약표를 활용하는 것이 바람직할 것이다. <표 4-6>에서 만약 최소한 45%의 수익 기여도가 높은 고객의 비율을 예측하기를 원한다면, 첫 두 마디(node 16(60.27%)와 node 17(45.81%))를 타겟으로 삼으면 될 것이다.

분류모델의 적합도를 판단하고 예측력을 쉽게 파악할 수 있는 다른 방안으로 정오분류행렬(Confusion Matrix)표가 사용될 수 있다. 정오분류행렬표란 목표변수의 실제범주와 모형에 의해 예측된 분류범주 사이의 관계를 나타내는 표라고 할 수 있다. 즉, 목표변수의 범주별로 이를 제대로 분류한 빈도와 그렇지 못한 빈도를 함께 제시한 표이다. 이는 목표변수의 범주가 c개인 경우 c × c 개의 셀(cell)로 이루어진 표 형식을 취한다.[85] 분류모델의 정오분류행렬표에서 대각(diagonal)에 존재하는 도수(frequency)는 실제 범주와 예측범주가 일치하는 즉, 제대로 예측한 개체의 수이고 비대각(off-diagonal)에 존재하는 도수는 예측이 어긋난 개체의 수라고 할 수 있다.[86]

<표 4-7>에서 정오분류행렬표는 트레이닝 데이터 셋과 평가용 데이터 셋의 각각에 대하여 실제 데이터 수와 예측된 데이터 수의 관계를 보여주고 있다.

정오분류행렬표는 목표변수가 얼마나 정확하게 분류되었는가를 보여주는데, 이를 위해 목표변수의 실제범주(actual category)와 예측범주(predicted category)의 패턴을 표로 나타내었다. 이 표는 목표변수의 실제범주가 의사결정나무의 추론규칙(이 표에서는 각 끝마디에서 그 마디에 해당하는 모든 개체를 가장 높은 비율을 가지는 목표변수의 범주에 할당하고 있다.)에 의해 예측된 범주와 대비시켜 보여주고 있다.

85) 강현철 외5, 전게서, p.78.
86) 최종후 외3, 전게서, 1998, p.59.

<표 4-7> 정오분류행렬(Confusion Matrix)표

Confusion Matrix									
Training data set					Validation data set				
		Predicted					Predicted		
		Target 0	Target 1	Total			Target 0	Target 1	Total
Actual	Target 0	8621	94	8715	Actual	Target 0	3539	29	3568
	Target 1	1827	106	1933		Target 1	718	44	762
	Total	10448	200	10648		Total	4257	73	4330
Risk Estimate			0.18040946		Risk Estimate			0.17251832	
Accuracy(%)			81.96		Accuracy(%)			82.75	

<표 4-7>의 정오분류행렬표에서 보는 것과 같이 트레이닝 데이터 셋에 대한 위험 추정치(risk estimate)는 실제로 수익 기여도 낮은 고객의 건수를 수익기여도가 높은 고객으로 잘못 예측한 94건의 데이터와 실제 수익 기여도가 높은 데이터를 수익 기여도가 낮은 데이터로 잘못 분류한 1,827건의 합계 1,921건을 트레이닝 데이터 셋의 총 데이터 건수인 10,648로 나눈 0.18040946이다.

트레이닝 데이터 셋의 총 10,648건 중 올바르게 분류된 데이터는 8,727건(106 + 8621 = 8,727)으로 트레이닝 데이터 셋에 대한 분류정확성은 81.96%(8,727 / 10,648 × 100 = 81.96%)를 나타내고 있다. 즉, 트레이닝 데이터 셋의 정오분류행렬표는 이 분류모델이 트레이닝 데이터 셋의 81.96%를 올바르게 분류하고 있는 것을 보여주고 있다. 따라서 이 분류모델에 의하면 수익 기여도가 높은 고객을 수익 기여도가 낮은 고객으로 분류하거나 수익 기여도가 낮은 고객을 수익 기여도가 높은 고객으로 분류할 가능성은 적은 것으로 나타나 있다. 평가용 데이터 셋에 대한 위험 추정치는 실제로 수익 기여도 낮은 고객을 수익 기여도 높은 고객으로 잘못 예측한 29건의 데이터와 실제 수익 기여도 높은 데이터를 수익 기여도 낮은 데이터로 잘못 분류한 718건의 합계 747건을 트레이닝 데이터 셋의 총 데이터 건수인 4,330로 나눈 0.17251832로서 트레이닝 데이터 셋의 위험 추정치와 큰 차이는 없다.

또한 평가용 데이터 셋 4,330건 중 올바르게 분류된 데이터는 3,583건(44 + 3,539 = 3,583)으로 평가용 데이터 셋에 대한 분류 정확성은 82.75%(3,583 / 4,330 × 100

= 82.75%)이므로 트레이닝 데이터 셋에 대한 분류 정확성 81.96%와 거의 차이를 보이지 않고 있다. 즉, 평가용 데이터 셋의 정오분류행렬표는 이 모델이 평가용 데이터 셋에 대하여 82.75%의 높은 정확성을 나타내고 있으며, 정확도(Accuracy)가 트레이닝 데이터 셋의 정확도 81.96%와 비슷하여 매우 안정적임을 보여주고 있다.

3. 분류모델의 평가

분류모델의 성과을 비교하기 위한 대부분의 일반적 방법은 LIFT라 불리는 측정치를 사용하는 것이다. 이 기법은 다른 작업을 위해 설계된 모델을 비교하는데 적합할 수 있다. LIFT가 실제 측정하는 것은 분류모델을 사용하여 모집단에서 목적에 의해 표본을 추출할 때 특정 부류(class)가 차지하는 비율의 변화이다.

$$LIFT = P(\ class_t \mid Sample) / P(\ class_t \mid Population)$$

LIFT는 직접 마케팅(direct marketing)산업에서 유래된 것으로, 예를 들면 쉽게 이해할 수 있다. 고객에게 직접우편(direct mailing)을 보냈을 때 누가 반응할 가능성이 높은가를 예측하는 분류모델을 개발한다면, 통상 사전 분류된 트레이닝 데이터 셋(training data set)과 필요하다면 사전 분류된 테스트 데이터 셋(test data set)을 사용하여 모델을 개발할 수 있다. 개발된 분류모델은 각각의 대상고객에 대하여 응답 또는 무응답으로 예측한다. 물론 이러한 예측이 실제 결과와 일치하는 것은 아니다. 그러나 이 모델이 좋은 모델이라면 이 모델에 의하여 추출한 표본(bias sample)에 포함되어 있는 응답건수의 비율은 전체 평가 모집단에 포함되어 있는 응답건수 의 비율보다 높을 것이다. 만약 평가 모집단이 5%의 응답비율을 보이고 있는데 반해 분류모델에 의하여 선택된 표본은 50% 응답비율을 나타내고 있다면 그 모델의 LIFT는 10(50/5=10)이다.[87]

즉, LIFT란 비모델링 데이터를 사용하여 얻은 정확성의 수준과 모델링 기술에 의해 각 집단 세그먼트에 대해 정확하게 예측된 응답자를 비교하여 얻은 밀도의

87) Berry, Michael J. A., Gordon Linoff, *op. cit.*, 1997, p.107

비를 나타낸다. 리프트 비(ratio)는 모델이 모델링이 존재하지 않는 경우 보다 얼마나 정확한지를 수량화한 것이다.[88]

<표 4-6>은 평가용 데이터 셋의 분류모델의 마지막 노드(leaf node) 가운데 16, 17, 8, 14, 24, 5번의 6개 노드가 전체 평가용 데이터 셋의 수익 기여도가 높은 고객의 비율 즉, 뿌리 노드의 수익 기여도가 높은 고객의 비율인 17.6% 이상의 비율을 나타내고 있음을 보여주고 있다.

개발된 분류모델에 의하여 수익 기여도가 높은 고객을 선별하여 마케팅 수행할 경우, 정보 이익(%Resp)이 가장 큰 노드의 고객부터 우선적으로 마케팅의 대상고객에 포함시켜야 성공 가능성이 가장 높으며 이때의 반응율은 60.27%에 이르고 있다. 마케팅 대상고객의 수가 늘어남에 따라 정보이익이 그 다음으로 큰 노드가 차례로 그 대상에 포함되어야 한다.

<표 4-6>에 의하면 마케팅의 대상고객으로는 정보이익이 60.27%로 가장 높은 노드 16번이 가장 먼저 포함되어야 하며, 이 노드의 73건의 대상고객에는 수익 기여도가 높은 반응을 보인 고객이 44건이 포함되어 있어 노드 16번의 반응율은 60.27%임을 의미한다. 또한 노드 16의 리프트는 3.42(60.27 / 17.6 = 3.42)임을 나타내고 있다. 이것은 이 분류모델에 의하여 선택된 노드 16의 고객에게 마케팅을 실시할 경우 대상고객을 무작위로 추출하여 실시하였을 때 보다 3.42배 이상의 성공 가능성이 높다는 것을 의미한다.

노드 16에 포함되어 있는 73건 보다 많은 마케팅 대상 고객수가 필요하다면 당연히 그 다음으로 정보이익이 큰 노드 17에 속해 있는 고객 237건을 추가적으로 포함시켜야 할 것이다. 이때 대상고객의 수는 310건(73 + 237 = 310)으로서 그 가운데 수익 기여도가 높은 고객의 수는 142건(44 + 98 = 142)이 포함되어 있어 반응율은 45.81%(142 / 310 × 100 = 45.81%)이다. 또한 리프트는 2.60(45.81 / 17.6 = 2.60)로서 노드 16과 노드 17의 고객에게 마케팅을 실시할 경우 대상고객을 무작위로 추출하여 실시하였을 때 보다 2.60배 이상의 성공 가능성이 높다는 것을 의미한다.

이와 같이 대상이 되는 노드는 <표 4-6>에서 알 수 있는 바와 같이 반응율이 17.6% 이상을 나타내는 6개 노드 즉, 16, 17, 8, 14, 24, 5번째의 6개 노드만이 의미

88) Ramon Barquin et al. 著, 함문성 · 김석호 譯, 전게서, pp.134~135.

있는 정보로 해석할 수 있다. 6개 노드의 전체 데이터 수는 1,479건(6개 노드의 데이터 누적 건수)이고, 이 가운데 수익 기여도가 높은 건수는 402건(6개 노드의 수익 기여도 높은 데이터 누적 건수)이므로 6개 노드의 총 반응율은 27.18%(402 / 1,479 × 100 = 27.18%), 리프트는 1.54이다. 평가용 데이터 셋에 대하여 분류모델에 의한 목표 마케팅(target marketing)을 수행할 경우 마케팅의 효율성을 높일 수 있다.

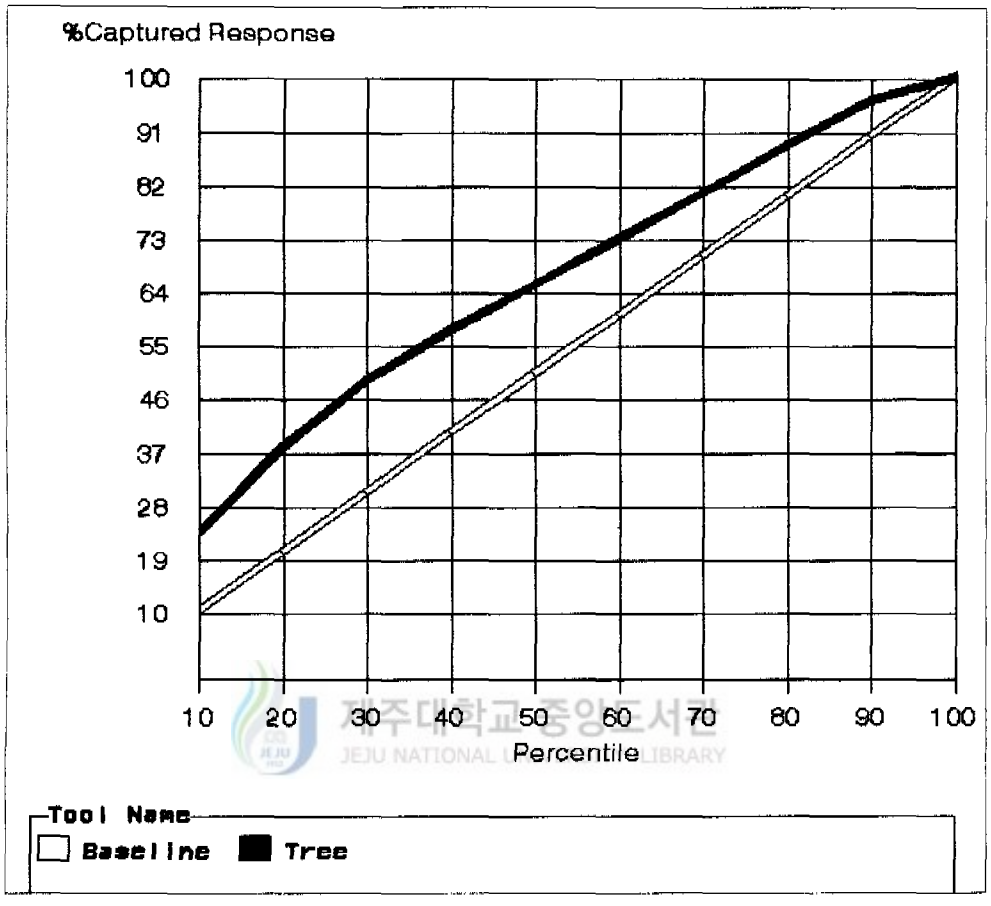
평가용 데이터 셋에 대하여 분류모델에 의한 목표 마케팅을 수행할 경우의 마케팅의 효율성을 그래프로 나타낸 것이 <그림 4-5>이다.

<그림 4-5>의 대각선은 무작위로 평가용 데이터 셋에서 추출하여 마케팅을 실시하였을 경우 예상되는 반응율을 나타내는 것이고, 그 위쪽에 선은 분류모델을 사용하였을 경우의 효율성을 나타내 주고 있다. 즉, 대각선과 그 위쪽선의 차이가 나는 부분은 분류모델에 의한 목표 마케팅을 실시함으로써 얻어지는 정보 이익, 또는 효율성의 차이라고 볼 수 있다. 따라서 마케팅 담당자는 <그림 4-5>로부터 얻을 수 있는 정보 이익에 관한 정보를 대상고객의 크기를 결정하는 한가지의 요소로 고려할 수 있다.

평가용 데이터 셋에 대한 분류모델의 효용성은 <표 4-6>의 정보 이익 요약표와 <그림 4-5>을 통하여 실제적인 의미를 살펴볼 수 있다.

<표 4-6>에 의하면 정보 이익(%resp)이 높은 노드의 순서대로 대상고객에 포함되어 전체 데이터의 반응율인 17.6%보다 높은 노드 5 까지 대상고객에 포함될 경우, 총 대상고객 수는 1,479건으로 전체 테스트 셋 4,330건의 34.16%에 불과하다. 그러나 이들을 대상으로 마케팅을 실시할 경우 평가용 데이터 셋에 포함되어 있는 762건의 총 반응건수 가운데 52.76%에 해당하는 402건의 반응을 기대할 수 있다. 즉, <표 4-6>의 결과로부터 정보 이익(%resp)이 높은 상위 34.16%에 해당하는 고객을 관리하는 것이 고객 전체를 관리하는 것에 비해서 1.54배의 효율을 얻을 수 있다는 것을 알 수 있으며, 상위 34.16%인 1,479건 중 전체 반응을 보인 762건의 52.76%인 402건으로부터 반응을 얻어낸 것을 알 수 있다.

<그림 4-5> 이익도표 : %Captured Resp

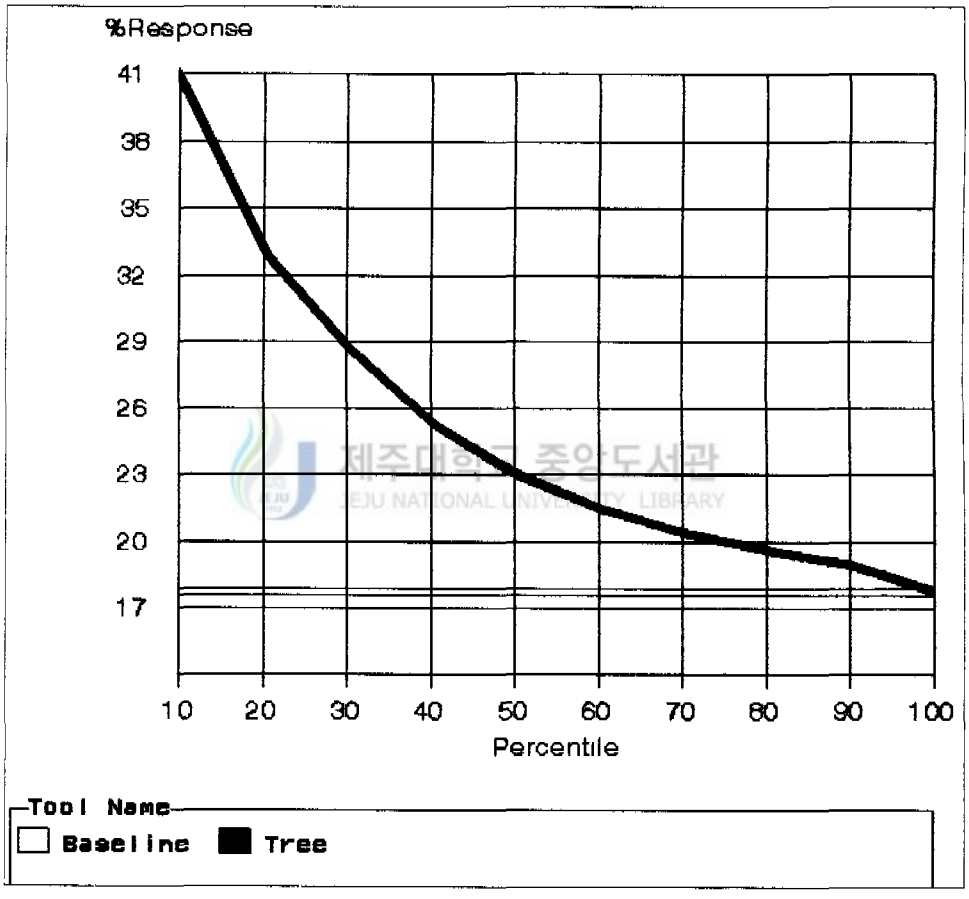


예측력이 우수한 모형을 구축한 후 사후확률 순으로 정렬하는 경우, 상위 집단에는 목표범주의 빈도가 높고 하위집단에는 낮은 양상이 나타나게 된다. 이러한 양상은 모형의 성능이 떨어질수록 상하위 집단간의 차이가 없어지게 되며, 극단적으로 모형의 효과가 전혀 없는 경우 모든 집단에서 특정범주의 빈도가 비슷한 양상을 띄게 된다. 이는 전체 데이터 셋을 임의로 10등분하는 의미와 동일하다.[89] <그림 4-6>은 평가용 데이터(validation data)에 대한 이익도표를 보여주고 있다. 의사결정나무 모형에 의해서 수익 기여도가 높을 확률이 큰 상위 10%에 대해서 타켓 마

89) 강현철 외5, 전게서, p.101.

케팅을 하면 기대 반응율(expected response)이 약 41%정도가 된다는 것을 나타내고 있다. 이는 전체적인 수익 기여도 높은 고객의 비율이 17.6%임을 고려한다면 <그림 4-7>에서와 같이 수익기여도가 높을 확률이 큰 상위 10%에 대해서 약 2.3배의 효율을 얻을 수 있음을 의미한다.

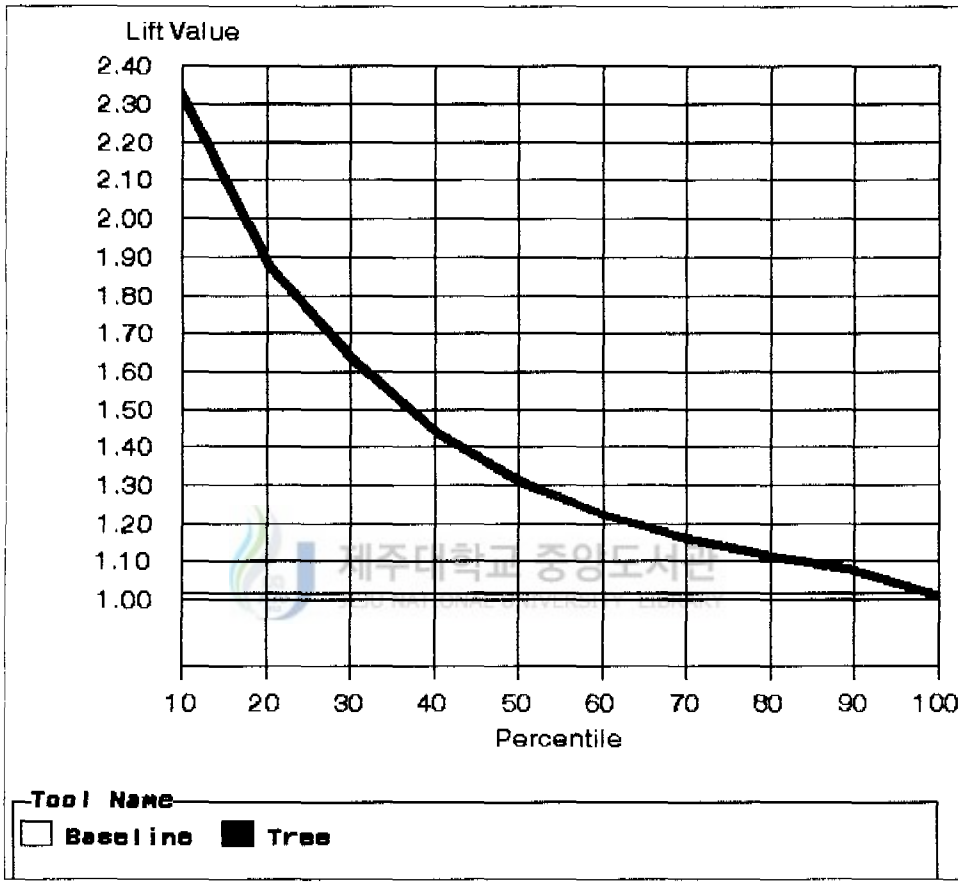
<그림 4-6> 이익도표 : %Resp



그리고, 모형이 평가용 데이터 셋으로부터 대상고객의 추출 표본수가 증가함에 따라 리프트의 값이 적어지는 현상을 나타내고 있다. 따라서 노드 17의 대상고객이 노드 16에 추가적으로 포함되면 대상고객의 표본수가 늘어나게 되고, 이때 리프트는 3.42에서 2.60로 떨어진다. 또한 모형구축의 의미가 전혀 없는 사후확률에 따라

데이터 셋을 10등분하면 17.6%의 수익 기여도가 높은 고객의 비율로 하단의 일직선(baseline)으로 나타나고 있다.

<그림 4-7> 리프트 도표 : Lift Value



4. 모델의 활용

의사결정나무 모델은 간단한 "IF-THEN-ELSE" 구문과 같은 효율적인 기록 방법을 사용하므로, 이해하기 쉬운 예측을 제공한다.[90] 즉, 개발된 모델에서 도출된

90) Ramon Barquin et al. 著, 함문성 · 김석호 譯, 전게서, p.133.

정보는 'IF-THEN-ELSE' 형식으로 전환하여 신용카드 고객 데이터베이스에 적용함으로써 수익 기여도가 높은 고객을 추출하여 마케팅 전략 수립에 활용할 수 있다. 다음은 실험 결과 도출된 정보를 'IF-THEN' 형식으로 나타낸 것이다.

IF 직업 IS ONE OF: A B F
AND 카드종류 EQUALS 2
THEN
NODE : 5
N : 1378
1 : 21.8%
0 : 78.2%

IF 결제일 IS ONE OF: 17 25
AND 직업 IS ONE OF: C D E
AND 카드종류 EQUALS 2
THEN
NODE : 8
N : 477
1 : 30.0%
0 : 70.0%

IF 나이 < 26.5
AND 직업 IS ONE OF: A B D F
AND 카드종류 EQUALS 1
THEN
NODE : 12
N : 906
1 : 6.8%
0 : 93.2%

IF 26.5 <= 나이

AND 직업 IS ONE OF: A B D F

AND 카드종류 EQUALS 1

THEN

NODE : 13

N : 5608

1 : 14.1%

0 : 85.9%

IF 직업 EQUALS E

AND 카드종류 EQUALS 1

THEN

NODE : 14

N : 453

1 : 25.2%

0 : 74.8%

IF 직업 EQUALS C

AND 결제일 EQUALS 10

AND 카드종류 EQUALS 2

THEN

NODE : 16

N : 200

1 : 53.0%

0 : 47.0%

IF 직업 IS ONE OF: D E
AND 결제일 EQUALS 10
AND 카드종류 EQUALS 2
THEN

NODE : 17
N : 567
1 : 39.5%
0 : 60.5%

IF 성별 EQUALS M
AND 직업 EQUALS C
AND 카드종류 EQUALS 1
THEN

NODE : 24
N : 655
1 : 22.4%
0 : 77.6%

IF 성별 EQUALS F
AND 직업 EQUALS C
AND 카드종류 EQUALS 1
THEN

NODE : 25
N : 404
1 : 11.6%
0 : 88.4%

제 5 장 결 론

제1절 연구의 성과 및 시사점

최근 급변하는 금융환경 속에서 은행들은 경쟁력 강화와 안정적 수익 확보 및 지속적 성장을 위해 DB마케팅을 적극적으로 활용하고 있다. DB마케팅 가운데 최신 정보기술을 이용하여 고객정보를 분석하고 고객을 세분화하는 방안 중의 하나가 바로 데이터 마이닝이다.

데이터 마이닝은 사용자의 질의나 보고서가 효과적으로 밝혀낼 수 없었던 데이터베이스내의 패턴을 발견하고 규칙을 추론한다. 이러한 패턴과 추론은 의사결정을 지원하고 은행 환경의 변화를 예측하는데 사용될 수 있다. 예를 들어 데이터 마이닝의 결과로 특정 유형의 고객이 다른 고객에 비해 수익 기여도가 높은 경향이 있다는 사실을 알아낼 수 있다. 은행은 이 두 유형의 고객에 대한 차이를 알게 됨으로써 불특정 대중이 아닌 목표 고객에 집중된 마케팅을 수행할 수 있을 것이다.

본 연구에서는 고객 데이터베이스로부터 높은 수익 기여도가 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 목표 마케팅의 수행 대상이 되는 우수고객을 선별하는 모델 개발을 위해 A은행이 확보한 신용카드 고객 DB에서 최근 1년간의 신용카드 사용실적이 있는 고객 데이터에 데이터 마이닝 기법을 적용하여 실증적 실험을 하였다. 실험에서는 DB마케팅의 분석기법 가운데 수익 기여도가 높은 우량고객의 파악을 위한 유용한 방법의 하나로 RFM점수분석법을 이용하였으며, 또한 여기서 도출된 점수를 기준으로 데이터 마이닝 기법 가운데 의사결정나무 알고리즘(CHAID)을 이용하여 고객 분류모델을 개발하였다.

실험결과 A은행의 전체 신용카드 고객집단 가운데 직업이 금융기관 종사자이면서 결제일이 10일인 특별카드를 발급 받은 집단이 수익 기여도가 가장 높은 집단으로 분류되었고, 이와는 반대로 나이가 26.5세 미만이면서 직업이 공무원, 교육기관 종사자, 회사원 등인 일반카드를 발급 받은 집단은 수익 기여도가 가장 낮은 집단으로 분류되었다. 따라서 A은행의 마케팅 담당자는 전자와 같은 특성을 갖는 고

객들을 찾아내어 고객과의 지속적 관계를 유지해야 할 것이다. 반대로, 후자와 같은 특성을 갖는 고객들에게는 역마케팅(de-marketing) 등의 의사결정을 내려야 할 것이다.

그리고 이 분류모델을 마케팅에 적용할 경우 기존의 무작위 추출에 의한 매스 마케팅의 경우와 비교하여 볼 때 최고 3.42배 이상, 최소 1.54배 이상의 효율성을 높일 수 있음을 확인하였다. 또한 이 실험은 거래실적이 있는 고객 데이터만을 사용했기 때문에 실제 거래실적이 없는 고객 데이터까지 포함하여 이 분류모델에 적용할 경우 그 효율성은 상당히 높을 것으로 예상된다. 이는 고객을 분류모델에 의해 반응률이 높은(가치있는) 우량고객을 선별하고, 분류된 집단에서 수익 기여도가 높은 고객층을 밝혀냄으로써 향후 차별적 마케팅 혹은 고객 유지 마케팅의 기준 자료를 제시할 수 있을 것이다.

이와 같은 연구결과를 살펴 볼 때, 은행의 고객 데이터베이스로부터 획득한 전체의 고객을 대상으로 마케팅을 전개하는 것이 아니라, 데이터 마이닝을 통하여 고객을 세분화하고 채산성 있는 고객을 대상으로 차별적 마케팅을 전개하는 것이 바람직하다고 할 수 있다. 또한 세분화는 단순히 고객을 복수의 집단으로 세분화하는 것이 아니라 채산성에 따라 고객집단을 구분하고 채산성 있는 고객에게만 마케팅 노력을 집중함으로써 수익성 증대를 도모할 수 있다. 즉, 고객을 채산집단과 비채산집단으로 나누기 위해 RFM점수분석법과 의사결정나무 등의 데이터 마이닝 기법을 이용하여 여기서 판단된 반응률을 기준으로 손익분기점 또는 ROI(Return On Investment)를 도출할 수 있다면 은행의 수익적인 면에서 상당한 효과를 볼 수 있을 것이다. 또한 향후 개별은행의 DB마케팅 전략을 다음과 같이 생각해 볼 필요가 있다.

첫째, 데이터베이스를 정보수집과 고객명단 세분화에만 이용하지 말고 마케팅이나 전반적인 은행 경영전략에 대한 의사결정을 하는 데도 이용해야 한다. 즉, 캠페인 수행을 위해 시장 및 고객 성향을 분석하고 반응예측 모델을 만들어 수익성을 예측, 분석해야 한다. 또한 각종 데이터 분석을 통해 은행의 업무와 관련된 의사결정 및 경영정책을 수립할 수 있어야 한다. 이를 효과적으로 수행하기 위해서는 데이터 웨어하우스 및 데이터 마이닝, 캠페인 관리기술이 필수적이다. 데이터베이스를 분석적 도구로 활용하게 되면 테스트의 결과뿐만 아니라 조사의 통계적 기법이

나 발견점들을 구체화 할 수 있다. 또한 마케팅에 대한 반응을 측정하고 기록을 보관할 수 있으며, 모든 마케팅 의사결정에 대한 효과를 분석하고 해석하며 평가할 수 있다. 특히 미래의 반응을 예측할 수가 있을 것이다.

둘째, 고객 DB를 활용한 차별적 가격전략을 구사해야 한다. 본 실험에서 고객 DB를 활용하여 어떠한 유형의 고객이 수익 기여도가 높고, 낮은지를 밝혀낼 수 있음을 확인하였다. 이는 고객별로 차별적 가격전략의 구사가 가능하다고 할 수 있다. 따라서 수익 기여도가 낮은 고객에게는 엄격하게 수수료를 징수하고, 수익 기여도가 높은 고객에게는 이익이나 수수료를 환원해주거나 신용한도를 늘려줌으로써 추가적인 수수료 수입을 증대시킬 수 있는 차별적 가격전략이 바람직하다 할 수 있다. 이를 위해서는 지속적으로 축적된 방대한 양의 데이터베이스로부터 데이터 마이닝하여 수익 기여도가 높은 고객을 파악하고 이들과의 개별적 커뮤니케이션에 의해 은행의 수익성을 제고시키려는 노력이 필요하다.

셋째, 신규고객을 늘리는 것도 중요하지만 신규거래 이후부터 고객들의 만족도를 높이고 은행 이용률을 높이는데 초점을 두는 애프터 마케팅(after marketing) 활동이 더욱 중요하다고 볼 수 있다. 수익에 결정적 기여를 하는 것은 신규고객이 아니라 거래기간이 긴 기존고객이기 때문이다. 기존고객의 이용률을 높이고 거래 관계를 지속해 나가기 위해서는 고객의 기여도에 따라 차별적 보상이나 인센티브를 제공함으로써 장기적 관계를 강화 할 수 있는 고객관계관리(CRM) 시스템에 대한 도입이 필요하다. 따라서 고객별 수익 기여도에 상응하는 서비스 제공이 가능한 시스템을 갖춤으로써, 일반적으로 20%의 고객이 80%의 이익에 기여한다는 '파레토 법칙'이 주는 의미에 정합성을 갖는 마케팅을 전개함으로써 주요고객의 이탈을 방지하고 기존 고정고객의 충성도(loyalty)를 유지하는 전략 필요하다.

넷째, 지금은 은행경쟁력의 원천이 예·대마진 확보를 통한 이익창출에서 기존 고객 정보의 가공 및 활용을 통한 부가가치 창출로 바뀌어가고 있는 시점에 있다. 따라서 은행은 그 동안 축적해 온 방대한 양의 고객 정보를 효율적인 데이터 마이닝을 위한 고객 데이터베이스(Marketing Customer Information File : MCIF) 정비, 데이터 웨어하우스 구축 등을 통해 다양한 마케팅 활동을 전개할 수 있는 인프라(infrastructure)를 조속히 구축하여야 할 것이다.

제2절 연구의 한계 및 미래 연구의 방향

본 연구에서 데이터 마이닝 실험을 통해 발견된 지식은 앞으로 A은행의 마케팅 활동을 전개하는데 있어서 여러 가지 불확실성을 해소하는데 필요한 정보를 제공할 수 있지만, A은행의 신용카드 마케팅 업무에 직접 적용함으로써 모델의 사후적 검증을 하지 못하였다. 또한, 고객 데이터의 분석적 측면에서는 데이터의 형태나 연구의 목적에 따라 적절한 데이터 마이닝 기법을 통하여 유용한 결과를 도출하였으나, 분석한 고객 데이터베이스는 최근 1년간 축적된 초보적 자료이기 때문에 실제 심층적 분석을 하는데는 한계가 있었다.

향후 연구에 부언하고 싶은 점은 전체 고객정보를 수익 기여도가 높고 낮음 등으로 분류할 요소에 대해 가중치와 비중을 적용할 때 데이터 마이닝 기법을 이용할 필요성이 있다. 또한 분류된 특정 고객집단의 구매액 변화를 시계열적으로 분석하여 고객의 생애가치(Life Time Value) 증대를 위한 다양한 데이터 마이닝 기법에 대한 연구가 필요하다.

본 연구가 지니고 있는 이론적, 방법론적 한계점에도 불구하고 본 연구의 유용성은 데이터 마이닝으로 의사결정나무 기법을 이용하여 고객들을 분류해보았다는 것과 본 실험으로 은행의 실제 데이터로부터 데이터 마이닝의 효과가 어느 정도인지를 확인한 것에 그 의의를 가진다. 또한 실험자료로서 고객 데이터베이스에 다양한 항목이 구축되어 있었다면 고객 세분화에 따른 좀 더 다양한 인구사회적 특성 및 라이프 스타일 등에 따른 차이를 밝힐 수 있었겠지만, 본 연구가 향후 마케팅 활동을 전개하는데 참고적인 자료가 되기를 기대해 본다.

參 考 文 獻

<國內 文獻>

강현철 · 한상태 · 최종후 · 김차용 · 김은석 · 김미경, 「SAS Enterprise Miner를 이용한 데이터 마이닝 방법론 및 활용」, 자유아카데미, 1999.

김기서, 「선진금융으로 가는 고객세분화 마케팅」, 도서출판 고원, 1999.

김남훈 · 문성광 · 박세진 · 이인, "유전자 알고리즘의 이해와 구현", 프로그램세계 7월호, 신영미디어, 1997.

김신곤, "데이터 마이닝 기법(CHAID)을 이용한 효과적인 데이터베이스 마케팅에 관한 연구", 정보기술과 데이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4.

김신곤, "데이터 마이닝과 지식발견", 춘계학술대회 논문집,한국 전문가시스템학회, 1997.

박민식, "데이터 마이닝", 데이터베이스월드, 1998년 8월호.

박찬욱, 「고객정보를 활용한 은행 데이터베이스 마케팅 전략에 관한 연구」, 한국금융연구원, 1998.

박찬욱, 「금융기관의 데이터베이스 마케팅」, 시그마인사이트그룹, 1999.

박찬욱, 「데이터베이스 마케팅」, 연암사, 1996.

박찬욱, "데이터베이스 마케팅의 실행 수준에 영향을 미치는 요인들에 대한 연구 : 한국 은행을 중심으로", 마케팅연구 제14권 제2호, 한국마케팅학회, 1999. 6

박태원, "성공적인 정보분석을 위한 전략방안", IT BUSINESS, 1999. 5.

이용희, "Data Mining을 이용한 리테일 뱅킹 전략에 관한 실증적 연구", 전국은행연합회 논문집, 1998.

임영도 · 이상부, 「퍼지 · 신경망 · 유전진화」, 도서출판 영과 일, 1998. 1.

장남식 · 홍성완 · 장재호, 「데이터 마이닝」, 대청미디어, 1999. 10.

정보통신부, 「데이터 웨어하우스 기반의 Data Mining 소프트웨어 개발」, 1997.

정철용 · 함유근, 「고객정보시스템 구축 및 활용 전략」, 한국금융연구원, 1999.

조성진 · 정인정, "데이터 마이닝을 이용한 의사결정지원 시스템", 가을 학술발표논

문집 Vol. 26. No. 2., 한국정보과학회, 1999.

조용준 · 허준 · 최인규, 「Neural Connection을 이용한 데이터마이닝 신경망분석」, 자유아카데미,1999.

조재희, "데이터웨어하우징 기술을 이용한 DB마케팅 전략에 관한 연구", 정보기술과 테이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4.

조재희 · 박성진, "애경백화점 DB마케팅시스템 구축사례", 정보기술과 테이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4.

조태현, 「금융마케팅2」, 한국금융연수원, 2000.

지원철 · 서민수, "데이터 마이닝을 활용한 공급사슬관리 의사결정시스템의 구조에 관한 연구", 경영정보학연구, 경영정보학회, 제8권 3호, 1998. 12.

최종후 · 한상태 · 강현철 · 김은석, 「Answer Tree를 이용한 데이터마이닝 의사결정나무분석」, SPSS아카데미, 1998.

최종후 · 한상태 · 강현철 · 김은석 · 김미경, 「SAS Enterprise Miner를 이용한 데이터마이닝 기능과 사용법」, 자유아카데미, 1999.

최홍국 · 한성렬, "유통업계 데이터베이스 마케팅의 해외 사례", 정보기술과 테이터베이스 저널 제6권 1호, 한국 데이터베이스학회, 1999. 4.

한국은행, "미국 은행 소매금융부문의 리엔지니어링", 한국은행 은행부 경영분석2실, 1999. 3.

한국컴퓨터연구조합, 주전산기산학연협의회, "데이터웨어하우스", 타이컴월드 98/11-12, 제22호, 1998.

한국SAS소프트웨어 DB 마케팅팀, "마이닝의 접근방법론과 기법", 데이터베이스 월드, 1998년 3월호.

한국SAS소프트웨어 DB 마케팅팀, "데이터베이스 마케팅 적용 사례", 데이터베이스월드, 1998년 5월호.

한국SAS소프트웨어 DB 마케팅팀, "데이터 마이닝의 금융권 활용방안", 데이터베이스월드, 1998년 6월호.

한국SAS소프트웨어 DB 마케팅팀, "DB마케팅을 위한 접근 방법론", 데이터베이스 월드, 1999년 7월호.

한국썬마이크로시스템즈, "한국썬소식", 1998. 11.

한재홍 · 전용준, "전략적 고객관리(CRM)특강", 데이터베이스월드, 1999년 11월호.

함유근, "DB마케팅의 의의와 은행의 도입", 금융 11월호, 1997.

Bob Stone 著 · 금강기획마케팅전략연구소 譯, 「데이터베이스 마케팅」, 한국언론자료간행회, 1999.

Pieter Adriaans, Dolf Zantinge 著 · 용환승 譯, 「데이터 마이닝」, 그린, 1998.

Ramon Barquin et al. 著, 함문성 · 김석호 譯, 「데이터웨어하우스(II)」, 도서출판 니드, 1999.

<海外 文獻>

A. Famili, Wei-Min Shen, Richard Weber, Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis", *Intelligent Data Analysis* : Elsevier Science Inc., 1996.

Arthur Middleton Hughes, "How to Succeed with RFM Analysis - A Case Study", Database Marketing Institute, February 2, 2000. [http://www.dbmarketing.com/articles/Art106.htm]

Berry, Michael J. A., Gordon Linoff, "Data Mining Techniques for Marketing, Sales, and Customer Support", Wiley Computer Publishing-John Wiley & Sons. Inc. 1997.

Edelstein, H., "Mining Data Warehouse", *Information Week*, Jan 1996.

Holtz, Herman, "Databased Marketing", John Wiley & Sons, Inc. 1992.

Hughes, Arthur M., "Strategic Database Marketing : The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program", Irwin, 1994.

Inmon. W. H., "The Data Warehouse and Data Mining", *Communications of ACM*, Vol. 39, No. 11, November 1996.

Inmon. W. H., "What is Data Mart?", *White Paper*, D2K Inc, 1996. [http://www.d2k.com/d2k/library2.htm]

John F. Roddick, Noel G. Craske, and Thomas J. Richards, "Handling Discovered Structure in Database Systems", *IEEE Trans. on Knowledge and Data Engineering*, VOL. 8, No. 2, April 1996.

Jong Soo Park, Ming-Syan Chen, Philip S. Yu, Fellow, "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules", *IEEE Trans. on Knowledge and Data Engineering*, VOL. 9, No. 5, October 1997.

Labe Jr., Russell P., "Database Marketing Increases Prospecting Effectiveness at Merrill Lynch" , Interfaces, 24(5), September/October, 1994.

"Letters from America", *Marketing*, February 5, 1998.

Metha, Manish, Jorma Rissanen. Rakesh Agrawal, "MDL-based Decision Tree Pruning", IBM, Almaden Research Center, 1998.

Michael Goebel and Le Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", *ACM SIGKDD*, June 1999. [http://research.microsoft.com/datamine/sigkdd/]

Ming-Syan Chen, Jiawei Han and Philip S. Yu, "Data Mining : An Overview from Database Perspective", *IEEE Trans. on Knowledge and Data Engineering*, VOL. 8, No. 6, December 1996.

Nikolay I. Nikolaev, Vanio Slavov, "Inductive Genetic Programming with Decision Trees", *Intelligent Data Analysis* : Elsevier Science Inc., 1997.

Nitin Indurkhya, Sholom M. Weiss, "Estimating Performance Gains for Voted Decision Trees", *Intelligent Data Analysis* : Elsevier Science Inc., 1998.

Pieter Adriaans, Dolf Zantinge, "Data Mining", Addtion Wesely, 1996.

Pyle, Dorian, "Putting Data Mining In Its Place", Database Programing & Design, March 1998.

Ruth Dilly, "Data Mining An Introduction Student Note", *The Queen's University of Belfast*, Version 2.0, 1995.12. [http://www.qub.ac.uk/]

ABSTRACT

A STUDY ON THE DATABASE MARKETING IN DOMESTIC BANK USING DATA MINING

Dong-Hun Lee
Department of Management Information
Graduate School of Business Administration
Cheju National University
Supervised by Professor Byoung-Gill Choi

With the rapidly changing financial situation, banks came to discuss the various methods to ensure their competitiveness. As a result, the importance of Database Marketing has been raised as the specific executive solution. In the process, they have realized that the customer relationship management resulted from the accurate customer analysis about customer profitability contribution measurement would be not their choice but their inevitability.

One of the solutions that can promote customer profitability, analyzing the customer information and carrying out the customer segmentation using the current information technology, is the Data Mining.

Data Mining is a series of process that finds hidden knowledge, unexpected trends or new rule that based on all the usable raw data including daily transaction data, customer data, customer response data in various marketing activities, or other external data, and supports the practical use and analysis of the information for the real business decision.

This thesis is about finding the customer pattern expected high profitability from the customer Database, and , using the discovered pattern, developing the model that selects superior customers who will be the subject of the target marketing performance. Also, this is the search for the effective Database marketing by offering the marketers the list of selected customers, and then raising the prospects of success.

This study carried out literature investigation and the practical experiment at the same time to make a vigorous examination of the purpose of this search. When it comes to the literature investigation, It was inquired into the Database marketing of a bank and Data Mining theoretically referring to internal and external works and theses. In the practical experiment, the model was developed by applying the Data Mining technique at the credit-card customer Database of 'A' bank, and verified the developed model using Confusion Matrix Table. Then the outcome of the model was evaluated through LIFT.

In this study, the RFM scoring analysis was used as the useful method for selection of the superior customers, which was set by the standard on the basis of the extracted scores. Moreover, the customer classification model was developed, using decision tree algorithm among many Data Mining techniques.

Applying this classification model to marketing, It was obvious that it could raise the efficiency by 3.42 times at the largest, 1.54 times at the smallest, compared with the case of the mass marketing by the existing random sampling.

According to the results of this study, it is desirable to do the differential marketing to choose the profitable customers by segmenting the customers

through Data Mining, instead of conducting the marketing to all the customers in the bank's customer Database.