한국지능정보시스템학회논문지 제8권 제2호 2002년 12월(pp.15~35)

신용카드 시장에서 데이터 마이닝을 이용한 이탈고객 분석

이건창
성균관대학교 경영학부
(leekc@skku.ac.kr)

정남호
성균관대학교 경영연구소
(nhchung@dragon.skku.ac.kr)

신경식
이화여자대학교 경영대학
(ksshin@ewha.ac.kr)

최근 데이터 마이닝 기법이 주목받고 있는 이유 중의 가장 큰 이유는 자사가 보유하고 있는 고객의 특성을 파악함으로써 기존의 고객을 효과적으로 유지·관리할 수 있도록 지원하기 때문이다. 특히 고객 보유율 5% 신장이 수익률 120% 증대를 가져오는 것으로 보고되고 있는 신용카드 업계에서는 신규고객을 확보하는 것만큼 기존 고객을 유지·관리하는 것이 중요하다. 특히, 신용카드를 발급 받고 거의 사용하지 않은 고객이나 쉽게 이탈하는 고객을 판별하는 것은 신용카드사의 입장에서는 비용절감 차원에서 매우 중요하다. 그러나 아직까지 어떠한 속성을 보유하고 있는 고객이 쉽게 이탈하는지를 판별할 수 있는 연구는 거의 진행되지 않았다. 이에 본 연구에서는 데이터 마이닝 기법 중 널리 알려진 인공신경망, 로지스틱 회귀분석, C5.0 방법을 이용하여 신용카드 시장에서의 고객현황에 대하여 분석하고자 한다. 이를 위하여 본 연구에서는 모 신용카드사의 최근 4년간(97년 3월 이후) 가입고객 및 이탈고객을 대상으로 실증분석을 실시하였다. 분석결과 신용카드 시장에서 카드를 지속적으로 보유하고 있는 고객과 이탈하는 고객을 구분하는 속성이 존재함을 발견하였고, 이를 바탕으로 신용카드사가 수립해야 할 마케팅 전략을 제시하였다.

1. 서론

최근 데이터 마이닝 기법이 주목받고 있는 이유 중의 가장 큰 이유는 자사가 보유하고 있는 고객의 특성을 파악함으로써 기존의 고객을 효과적으로 유지·관리할 수 있도록 지원하기 때문이다. 특히 고객 보유율 5% 신장이 수익률 120% 증대를 가져오는 것으로 보고되고 있는 신용카드 업계에서는 신규고객을 확보하는 것만큼 기존 고객을 유지·관리하는 것이 중요하다. 그러나 국내의 신용카드 업계의 현황을 보면 신규고객 확보에는 많은 노력을 기울이고 있는 반면에 기존 고객을 유지·관리하여 이탈을 방지하는데는 많은 노력을 기울이지 못하고 있는 형편이다.

국내 신용카드의 효시는 지난 69년 신세계백화점이 삼성그룹 임직원을 대상으로 발급한 것이 시초였다. 이어 70년에 조선호텔이 회원제 카드를 발행한 적이 있으며 74년에는 미도파백화점, 79년에 롯데백화점과 코스모스백화점이 카드발급을 시작하면서 부터 확산되기 시작했다. 전문 신용카드사의 첫 등장은 78년 7월 코리안 익스프레스가 처음이며 그 해 9월 한국신용카드가 설립됐다. 그러나 이 당시에는 신용카드 사용이 그다지 활발하지 않았으며 88년 서울올림픽을 전후로

LG, 삼성 등 대기업들이 카드업에 진출하면서부터 시장의 규모가 급속히 확대되었다. 90년대 들어서는 직장인중에 신용카드를 소지한 사람이 없을 정도로 보편화되기 시작했고 소비생활에 깊숙이 뿌리를 내리며 한해 시장규모가 100조원에 이를 정도로 도약하고 있다(매일경제, 2000a). 또한, 인터넷 뱅킹이 활발하게 사용됨에 따라 이제 신용카드의 대금결제 내역 및 신용정보를 인터넷을 통해서도 볼 수 있게 되었다.

또한, 최근 국내 신용카드 이용자들의 현황을 보면 신용카드가 우리 사회에 얼마나 자리잡았는지를 극명히 보여준다(매일경제, 2000b). 2000년 7월 BC 카드가 조사한 국내 신용카드 사용자들의 현황보고를 보면 20대의 카드사용액 증가율은 전 연령대층에서 가장 높아서 99년 대비 124.2%가 증가했으며 전체 평균 증가율보다도 4%가량 늘어났다. 여성층 또한 이용액 증가율이 99년에 비해 121.1%나 신장됐다. 특히, 20대고객층은 물품 구매에 의한 신용판매보다는 현금서비스를 상대적으로 많이 이용하는 것으로 분석되었다. 현금서비스 이용률은 99년보다 158% 늘었으나 신용판매부문은 77% 증가에 머물렀다. 그러나 신용판매부문은 남성이 57.9% 신장한데 비해 여성은 60.1% 늘어 여성의 증가율이 더 높게 나타났다. 이와 같이 이제 신용카드는 연령과 성별에 상관없이 우리 생활의 일부로 자리잡고 있다(매일경제, 2000b).

그러나, 이와 같은 신용카드 시장의 전망이 밝은 것만은 아니다. 그 이유는 카드 산업에 대한 정부의 규제와 가계부분의 빚의 증가, 그리고 카드 시장이 이제 포화상태에 이르렀기 때문이다. 2002년 3월 말 현재 국내에 보급된 신용카드 수는 9,619만 7,000 장으로 성인 1인당 평균 4장의 신용카드를 보유하고 있는 수준이다(한국일보, 2002). 전문가들은 이 정도 보급률이면 신용카드 수는 이미 포화상태라고 전망한다. <표 1>을 보면 국내 카드사의 회원수 증감현황을 통해 신용카드 회원수 증가율이 급격히 감소하고 있음을 파악할 수 있다.

따라서, 이와 같은 상황 하에서 신용카드 사들이 수익을 창출하기 위해서는 신규고객 창출보다는 기존 고객들의 1인당 사용액을 증가시켜야 한다. 결국, 해당 기업에 가입하고 있는 고객의 특성을 파악하고 이들 고객의 특성에 맞는 마케팅 정책을 구사하는 것이 그 어느 때 보다도 중요한 시점이 되었다. 특히 어떠한 고객이 카드 사용이 높고 어떤 고객이 불량률이 높고 해지 가능성이 높은지를 사전적으로 예측할 수 있다면 이는 카드사의 수익창출에 결정적인 기여를 할 것으로 기대된다.

<표 1> 국내 카드사별 회원수 증감현황

(단위: 천명)

업체명	2001년 상반기	2001년 하반기	증가율	2002년 상반기	증가율%
국민카드	9,483	10,587	11.6%	12,013	13.5%
BC카드	19,443	22,825	17.4%	25,640	12.3%
삼성카드	11,760	14,061	19.6%	14,530	3.3%
LG카드	14,240	16,630	16.8%	17,530	5.4%
외환카드	5,701	6,274	10.1%	7,016	11.8%

(자료: 머니 투데이, 2002년 7월 19일)

이에 본 연구에서는 보유고객의 정보를 최대한으로 획득할 수 있는 방법인 데이터 마이닝 기법을 통하여 다음과 같은 연구방법을 제시한다.

첫째, 실제 신용카드 사용자료를 바탕으로 대표적인 데이터 마이닝 기법인 인공신경망, 로짓모델, C5.0을 적용하여 최근 4년간 신용카드보유 유지자와 이탈자의 특성을 파악하고 그 예측력을 비교 검증한다. 둘째, 데이터 마이닝 기법의 각 모형이 제시하는 고객의 특성을 바탕으로 유지고객과 이탈고객에 대한 구체적인 마케팅 전략의 가이드 라인을 제시한다. 특히, 인공신경망과 로짓모델은 유의한 속성을 파악하는 측면에서 많이 사용될 것이며, C5.0은 이들 특성을 이용하여 유지고객과 이탈고객의 특성을 도출하는데 사용될 것이다.

본 연구의 구성은 다음과 같다. 2장에서는 신용카드에 대한 연구 중 신용카드 보유자의 행위분석 측면에서 접근한 기존 연구들을 소개하고 본 연구에서 소개하는 데이터 마이닝이 무엇이며 어떠한 특징이 있는지 간략히 소개한다. 3장에서는 본 연구에서 사용하는 연구방법론인 인공신경망, 로짓모델, C5.0 에 대하여 간략히 소개하고 이들 모형간의 특징과 장·단점을 소개한다.

4장에서는 본 연구에서 사용하는 모형을 제시하는데 먼저 분석자료의 특성을 소개하고 전처리 과정 및 실험과정을 소개한다. 5장에서는 모형별 실험결과를 제시하고, 이를 바탕으로 신용카드 상의 마케팅 전략을 제시한다. 끝으로 6장에서는 본 연구의 시사점 및 공헌도 그리고 한계점 및 향후 연구방향을 제시한다.

2. 이론적 배경

2.1 신용카드에 관한 연구

신용카드는 그 중요성이 점차 증대함에 비해 학문적인 연구는 별로 활발히 이루어지고 있지 않은 형편이다. 국내에서는 80년대 말과 90년대 초에 가정관리학회를 중심으로 가계에서 카드사용의 행위에 대한 연구가 있었으며 최근에는 신용카드를 이용한 범죄유형에 관한 연구들이 주류를 이루고 있다. 한편, 인터넷의 발전과 함께 신용카드가 중요한 결제 수단으로 자리잡고 있으며 따라서, 이에 대한 연구도 진행되고 있다. 본 연구에서는 이러한 신용카드에 대한 다양한 연구 중에서 신용카드 이용자의 행동분석과 관련된 연구를 중심으로 살펴보겠다. 먼저 국내 연구동향을 살펴보면 이재희(1996)는 대학생을 대상으로 신용카드에 대한 인식을 편리성, 유익성, 과시성, 합리성으로 나누어 실증분석을 실시하였다. 분석결과 우리나라 대학생의 신용카드 보유율은 보급단계에 있으며 신용카드에 대한 지식이 낮은 상태에서 편리하다는 인식이 보편적으로 자리잡고 있는 관계로 연체할 가능성이 매우 높다고 하였다. 따라서, 카드사가 카드발급을 남용하지 말고 자격요건을 나이 외에 다양한 요인을 충분히 고려하여 발급해야 함을 강조하였다.

이윤금 등(1998)은 국내의 신용카드에 관한 연구들이 카드발급회사의 이익증대와 신용카드 회원확보를 위한 것이 대부분이라고 지적하였다. 이들은 물론 최근에 들어서 신용카드 사용에 대한 연구가 가정학 측면에서 발전되고는 있지만 신용카드 사용의 급증과 더불어 다수의 신용카드를 사용하게 되는 추세에서 복수신용카드 소지자를 대상으로 한 신용카드 사용행태에 관한 연구

가 부족하다고 강조하였다. 이에 이들은 복수신용카드 소지자를 대상으로 소지자들의 이용행태를 분석하였는데 신용카드 사용빈도에 영향을 미치는 변수들로 고소득층 여부, 전문사카드 보유를 들었으며 월평균 신용카드 사용금액에 영향을 미치는 변수들로 고졸학력, 전문대졸 이상, 취업주부, 전문사카드 보유, 편리성 동기 등으로 분석하였다.

이외에도 신용카드 보유자의 사용행동과 관련된 연구들이 주장한 바를 특징별로 간략히 요약하면 다음과 같다(이윤금 등, 1998). 신용카드 사용에 대한 태도에 따라서 카드사용액 및 카드소지수에 유의한 차이가 있었으며(조의준, 1994), 신용카드에 대한 태도가 신용카드 보유여부와 신용카드 사용경험 여부에 정적인 영향을 미치는 것으로 나타났다(서경의, 1997). 신용카드 사용행동에 영향을 미치는 인구통계학적인 요인을 보면 연령이 낮을수록 신용카드를 사용하는 행위가 높게 나타났으나(이영호 & 지영숙, 1987), 연령이 높을수록 사용액이 많았고(최재복, 1995; 박선태, 1995; 이상영, 1995), 연령과 신용카드 사용빈도 간에 유의한 관계가 없는 것으로 분석한 연구도 있다(조의준, 1994; 박찬실, 1995). 한편, 소득수준이 높을수록 신용카드 사용빈도가 높았고(박찬실, 1995), 소득수준이 높을수록 신용카드를 합리적으로 사용하고 보관을 잘 하였다(박근주, 1990; 이은희, 1992). 또한, 교육수준이 높을수록 신용카드 사용액이 많았으나(최재복, 1995), 교육수준과 월평균 신용카드 사용액과는 유의한 영향관계가 없는 것으로 보고하고 있다(박찬실, 1995).

국외에서도 신용카드 사용행동에 대한 연구가 보고되고 있다. Kinsey(1982)는 신용카드 소지수는 성별에 따라 차이가 있으며 Hirschman (1980)은 성별에 따라 이용패턴이 다르다고 하였다. 즉, 남자는 은행카드를 많이 사용하며 여자는 상점카드를 많이 사용한다는 것이다. 이외에도 은행계좌, 거주지역, 신용카드에 대한 태도, 주택비용, 대부액, 인지된 카드 가격, 소득, 직업, 인종, 성별, 연령, 고용정도 등의 변수들이 신용카드 보유에 영향을 미치는 것으로 나타났다(Awh & Waters, 1974; Curtin & Neubig, 1979, 1980; Hirschman, 1979; Mandell, 1972).

이상의 신용카드 보유자의 행위분석에 대한 국내외의 기존 연구를 보면 나름대로의 의미는 제공하고 있지만, 대부분이 설문지를 작성하여 통계적으로 분석한 연구였기 때문에 여기에서 발생하는 측정오차를 포함하고 있다. 또한, 이러한 방법론상의 문제 외에도 대부분의 연구가 신용카드를 현재 보유하고 있는 고객들만을 대상으로 설문이 실시된 관계로 신용카드 시장에서 이탈한 고객들의 특성은 파악할 수 없다는 단점이 있다.

이에 본 연구에서는 최근 4년간 신용카드 보유고객 및 해지고객을 대상으로 이들의 인구통계적 특성과 신용카드 사용행위와 관련된 실제 자료를 이용하여 이들의 특성을 도출하고자 한다. 이러한 연구는 신용카드사의 고객보유율이 어느 때보다도 중요시 여겨지고 있는 최근의 경영환경에 비추어 볼 때 매우 의미 있는 연구가 될 것으로 기대된다.

2.2 데이터 마이닝에 관한 연구

데이터 마이닝은 데이터로부터 패턴이나 모형을 추출하기 위해 구체적인 알고리즘을 응용하는 과정이다(Fayyad et al., 1996). 또는 대규모 데이터베이스 내에 존재하는, 그러나 대량의 데이터 사이에 숨겨져 있는 상호관련성과 글로벌 패턴에 대한 탐색으로 정의되기도 한다. 데이터 마이닝

과 현재 사용되고 있는 많은 분석 도구들과의 중대한 차이점은 데이터 사이의 상호관계를 찾아내는데 사용하는 방법의 차이이다. 현재 사용되고 있는 많은 분석 도구(SQL 및 리포팅 도구, 통계분석 패키지, OLAP 도구, 시각화 도구 등)들은 사용자가 특정한 데이터 사이의 관계에 대해 가설을 세우고 이러한 가설을 입증하거나 또는 반박할 수 있도록 지원하고 있다. 이 분석방식은 기본 질문을 구성하는데 있어 분석가의 주관을 위주로 하며 잠재적으로 데이터베이스에 대해 매우 복잡한 질의어의 결과를 바탕으로 분석결과를 상세화 한다. 이와 같이 입증을 기반으로 한 분석의 효율성은 분석가가 적절한 질문을 주는지의 여부, 신속한 응답의 여부, 속성 공간의 복잡성을 관리하거나 다른 각도에서 생각을 하는 지의 여부 등 여러 가지 요소에 의해 제한을 받게 된다. 이에 반해 데이터 마이닝은 패턴조화나 다른 알고리즘이 데이터 사이의 핵심적 관계를 결정하도록 하는 발견(Discovery)을 기반으로 하는 어플리케이션을 사용한다.

최근 데이터 마이닝의 중요성이 강조되는 이유는 여러 가지가 있으나 첫째, 방대한 데이터베이스 속에 축적된 많은 양의 데이터를 보다 효율적으로 이용하고 둘째, 데이터 마이닝 알고리즘의 발달과 컴퓨터의 용량 및 성능 향상은 양적으로 증가되고 복잡한 형태를 가진 데이터의 처리 과정을 보다 쉽게 처리 할 수 있도록 함으로서 원하는 정보를 보다 쉽게 얻을 수 있는 환경을 제공한다. 셋째, 데이터 마이닝 기법은 기존의 전문가 시스템이 갖는 한계점인 지식 획득의 병목현상을 유연하게 극복할 수 있는 대안으로 자리잡고 있다. 데이터로부터 유용한 지식을 획득하는 과정인 데이터 마이닝은 언급되는 분야에 따라서 여러 가지 명칭으로 혼재되어 불리고 있다. 예를 들면, 지식추출(Knowledge Extraction), 정보발견(Information Discovery), 데이터 연금술(Data Archeology), 데이터 패턴처리(Data Pattern Process), 정보수확(Information Harvesting), KDD(Knowledge Discovery in Database) 등과 같다. 데이터 마이닝이라는 용어는 주로 통계학자, 데이터베이스 연구개발자, MIS 분야에서 널리 통용되고 있다. 원래 데이터 마이닝은 그 시초가 기계학습(Machine Learning), 패턴 인식(Pattern Recognition), 통계학, 인공지능, 전문가 시스템, 데이터 시각화, 정보조회 등의 다양한 분야이므로 새로운 개념은 아니다(지원철 & 서민수, 1998; Adriaans & Zantinge, 1997; Fayyad et al., 1996; Hong, 1996).

따라서 이와 같은 데이터 마이닝을 위한 구체적인 방법론으로는 여러 가지 도구들이 사용되고 있으며 특히, 데이터유지, 패턴분류, 규칙, 규칙귀납법(Rule Induction), 유전자 알고리즘(Genetic Algorithms), 의사결정 트리(Decision Trees), 교차표(Cross Tabulation), 에이전트(Agents), 신용네트워크(Belief Networks), 방정식접근법(Equational Approaches), 통계학, 인공신경망, 분류(classification), 회귀분석(regression), 카트(CART: Classification and Regression Tree), 체이드(CHAID: Chi Square Automatic Interactive Detection), 퍼지 로직(Fuzzy Logic) 및 고급 시각화와 같은 규칙기반 분석법이 많이 사용되고 있다(Fayyad et al., 1996).

데이터 마이닝이 최근에 더욱 각광 받는 이유는 인터넷이 등장하였기 때문이다. 즉, 인터넷과 같은 클라이언트/서버(Client/Server) 구조 하에서 데이터 마이닝 엔진을 서버에 두고 클라이언트들의 접근기록을 데이터베이스화하여 사후 분석한 후 그 정보를 이용하는 것이다. 물론, 현재

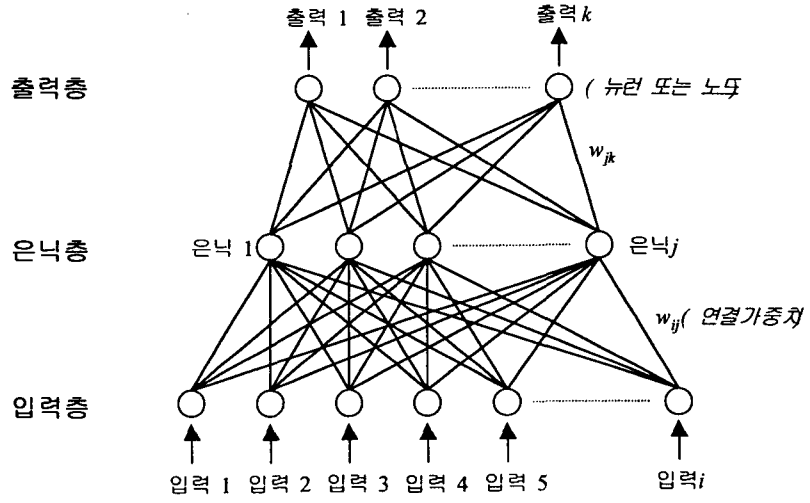
인터넷상의 정보는 너무도 다양한 형식으로 구성되어 있으므로 모든 고객정보를 대상으로 새로운 패턴을 추출하기에는 다소 무리이겠지만 인터넷 문서의 표현 양식의 표준화 및 검색기술의 발전은 가까운 장래에 데이터 마이닝의 적용을 가능하게 할 것으로 보인다(지원철 & 서민수, 1998). 다음 장에서는 본 연구에서 중점적으로 사용하는 인공신경망과 로지스틱 회귀분석, 그리고 규칙귀납법의 한 종류인 C5.0 에 대해 보다 자세하게 소개한다.

3. 연구방법론

3.1 인공신경망

인공신경망은 간단한 계산능력을 가진 처리단위인 뉴런(Neuron) 또는 노드(Node)들이 서로 복잡하게 연결된 컴퓨터 시스템으로서 외부에서 주어진 입력에 대하여 동적인 반응을 할 수 있다. 이러한 특징은 결국 인공신경망을 구성하고 있는 다수의 뉴런간의 상호연결성에 기인한 것이다. 뉴런은 생체내의 신경세포와 비슷한 것으로써 가중치화된 상호연결선으로 서로 연결되어 있다. 가장 일반적으로 많이 사용되고 있는 인공신경망 모형은 Rumelhart 등(1986) 이 제안한 다계층 인공신경망 모형으로서, 입력층(Input layer)에서 은닉층(Hidden layer), 은닉층에서 출력층(Output layer)으로 각 뉴런이 서로 연결되어 있는 것이 특징이다.

각각의 뉴런은 주어진 학습자료를 학습하는 학습기능과 상호연결된 또 다른 뉴런에 그 처리결과를 보내는 전달기능이 있다. 특히 전달기능을 위하여 사용되는 전이함수(Transfer function)는 일반적으로 S자형 함수인 시그모이드(Sigmoid) 함수를 사용한다. 입력층은 외부환경과 상호반응하며 외부입력을 받아 인공신경망에 전달하는 역할을 한다. 또한 출력층은 주어진 외부 입력에



<그림 1> 다계층 인공신경망의 기본 구조
(자료: Jain & Nag, 1997)

대한 적절한 출력을 내보내는 역할을 한다. 한편 입력층과 출력층 사이의 인공신경망층을 은닉층이라고 하며 이는 주어진 입력으로부터 특성을 추출하여 출력층으로 보내는 기능을 한다. 은닉층의 수와 뉴런수는 적용 문제에 따라 달라지며 따라서 그 타당성은 실험을 통해서 확인하여야 한다(Lippmann, 1988). 만약 입력자료가 특성추출이 용이하지 않은 자료로 구성되어 있으면, 그러한 입력자료로부터 고차원의 특성을 추출하기 위해서는 여러개의 은닉층이 요구된다. 반면에, 입력자료가 이미 어느정도 고차원의 특성치를 나타내고 있으면, 하나 또는 두개 정도의 은닉층만 있어도 거의 모든 형태의 문제 해결 공간을 구성할 수 있다. 한편, 서로 다른 층의 뉴런간에 형성되는 연결가중치(Connection Weights)는 역전파 학습과 같은 감독학습(Supervised learning)에 의해서 결정되거나 또는 경쟁학습과 같은 비감독학습(Unsupervised learning)에 의해서 결정된다. 특히, 역전파 학습은 인공신경망 관련 응용에 있어서 가장 많이 이용되고 있는데, 그 이유는 역전파 학습이 갖는 넓은 응용력과 높은 일반화 능력(Generalization effect)으로서 인공신경망의 가장 큰 특징중의 하나이다. 즉, 학습하지 않은 입력자료에 대한 근사추론(Approximation reasoning)을 가능하게 하는 것을 의미하며 처음 대하는 입력자료에 대하여 올바른 결과를 낼 수 있는 추론능력을 의미한다. 이와 같은 다계층 인공신경망을 학습시키기 위한 대표적인 학습방법으로 역전파 학습알고리즘(Back Propagation Algorithm)이 있다. 이 알고리즘의 절차는 다음과 같은데 먼저, 다계층 인공신경망모형에서 처리 단위 j의 역할은 [식 1]으로 표현할 수 있다.

$$o_j = f(net_j) \qquad \text{[식 1]}$$

이때, o_j=j처리 단위의 출력값

$$net_j = \sum_{i=1}^{N} w_{ji} x_i$$

$f(\ )$ = 비선형 전환함수

x_i = 전단계 i처리 단위로부터의 입력값

w_{ij} = 전단계 i처리 단위와 j처리 단위와의 연결강도

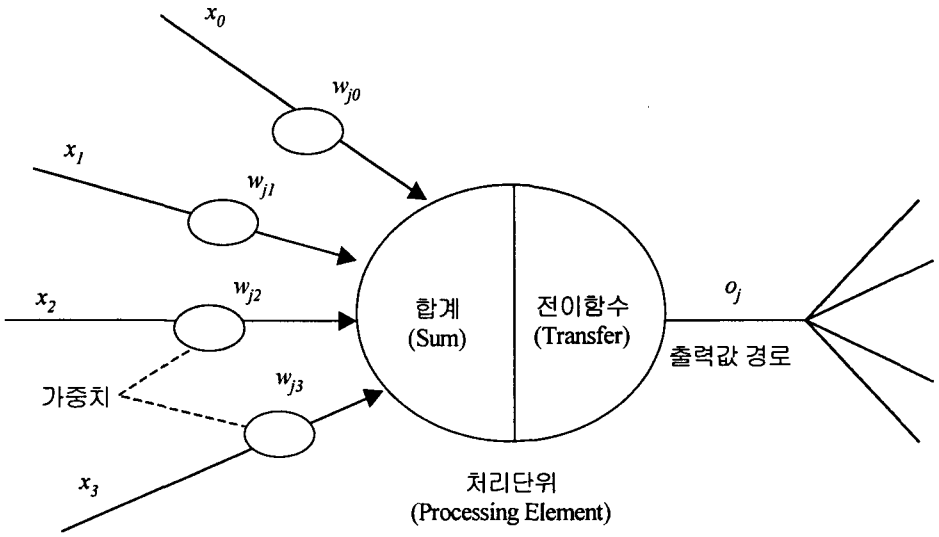
N = j처리 단위와 관련을 가지는 전단계 처리 단위의 수

즉, 처리단위는 전단계 처리 단위에서의 출력을 입력(x_i)으로 하여 연결강도에 의한 가중합(net_j)을 비선형함수로 전환하여 다음 단계의 처리단위로 출력하는 기능을 한다. 예를 들어 <그림 2>의 은닉층의 처리단위는 입력층의 연결된 각 처리단위들로부터 그들의 출력값을 연결강도로 곱하여 받아들이고 이들의 합을 비선형 전환하여 출력층의 처리단위로 출력한다.

이와 같은 입력과 출력이 각 층의 처리단위에서 이루어지고 최종적으로 출력층에서 계산값이 구해진다. 이때 사용되는 것이 [식 2]와 같은 시그모이드 함수인데 이 전환함수는 [-∞, +∞]의 구간에서 나타나는 출력을 [0, 1] 구간으로 제약하기 위해 사용된다.

$$f(net_j) = \frac{1}{(1 + e^{-net_j})} \qquad \text{[식 2]}$$

한편, 이미 언급한 바와 같이 인공신경망이 학습한다고 표현하는 것은 모형의 계산값이 목표값에 가깝도록 연결강도를 조정하는 과정을 의미한다. 즉, 위에서 언급한 다계층 인공신경망의 최종 출력값이 구해지면 이 값을 제시된 실제 목표값



<그림 2> 처리단위 j의 계산 구조
(자료원: Jain & Nag, 1997)

과의 오차를 구한다. 그 다음 이 오차가 최소화되는 방향으로 각 층에서의 연결강도를 조정하는 것이다[2]. 본 연구에서 사용하고자하는 역전파 학습알고리즘에서는 입력과 연결강도를 이용해 구한 출력값과 목표값의 차이인 오차를 하위처리단위로 되돌려 보냄[3]으로써 오차를 감소시키는 방향으로 연결강도를 조정한다. 이와 같은 연결강도의 조정을 오차의 크기로 인정할 수 있을 때까지 앞에서 설명한 모든 과정을 반복함으로써 학습이 이루어진다(Agrawal & Schorling, 1996).

3.2 로짓모형

소비자들은 종종 여러 가지 대안들 중에서 하나를 선택해야 하는 상황에 놓여있다. 예를 들어 고려하고 있는 상표들 중에서 특정 상표를 선택하는 의사결정을 해야하며 접근 가능한 점포들 중에서 특정 점포를 방문하는 선택을 해야만 한다. 신용카드 보유자에 있어서는 신용카드를 계속 보유하고 있을 것인지 아니면 해지할 것인지가 선택의 문제가 된다. 이러한 소비자의 특정 상표나 점포에 대한 선택과정을 파악하고 선택확률을 예측하기 위해 개발된 모형이 이산적 선택모형(Discrete Choice Model)이다. 이산적 선택모형은 효용이론에 근거하여 선택 대안들의 속성들에 대한 평가를 통해 형성된 각 대안들의 효용으로부터 선택확률을 예측한다. 이러한 이산적 선택 모형 중 소비자의 상표선택행위에 대한 실증연구에 많이 사용된 모델이 확률적 효용이론(Random Utility Theory)에 이론적 근거를 둔 로짓모형(Logit Model)이다. 로짓모형은 이산적 선택모형의 기본 가정과 같이 소비자들이 제한된

2) 이미 위에서 언급하였지만 이와 같이 최종 출력값에 비교할 학습의 대상인 목표값이 주어지는 경우를 "감독학습" 이라고 한다.
3) 이것이 역전파 인공신경망, 또는 back-propagation 알고리즘이라고 명명된 이유이다.

범위에서 효용을 극대화한다는 것을 가정하고 있으며 다음과 같이 간단한 수식을 통하여 효용을 도출하고 특정 대안에 대한 선택확률을 예측한다. 소비자 h가 어떤 제품 i에 대하여 특정한 구매시점 t에 느끼는 효용(U_{hit})은 [식 3]과 같이 정의된다.

$$U_{hit} = V_{hit} + \varepsilon_{hit} \qquad [식\ 3]$$

V_{hit}는 소비자 i가 제품 i로부터 특정한 구매시점 t에 느끼는 효용을 구하는 결정적 요소(Deterministic Component)이고 ε_{hit}는 무작위적 요소(Random Component)로서 ε_{hit}는 Type-Ⅱ Extreme 분포를 가정한다. 결정적 요소(Deterministic component)는 [식 4]와 같이 선택에 영향을 미치는 변수들의 선형결합으로 얻어진다(Ben-Akiva & Lerman, 1993).

$$V_{hit} = \alpha_i + \beta X_{hit} \qquad [식\ 4]$$

α_i는 각 제품들만의 독특한 특성을 나타내는 절편(Brand-specific intercept)이고 β는 추정해야할 모수이며 X_{hit}는 제품선택에 영향을 미치는 요인들의 벡터이다. ε_{hit}가 Type-Ⅱ Extreme 분포를 가정하므로 이렇게 해서 구한 각 제품 대안에 대한 효용은 로짓모형에 의해 [식 5]와 같이 제품 선택확률로 나타낼 수 있다.

$$P_{hit} = \frac{\exp(V_{hit})}{\sum_{k=1} \exp(V_{hkt})} \qquad [식\ 5]$$

P_{hit} : 소비자 h 가 특정 대안 i를 구매시점 t에 선택할 확률

V_{hit} : 소비자 h 가 특정 대안 i에 대해 구매시점 t에 느끼는 효용

3.3 C5.0

C5.0은 Quinlan(1996)이 ID3에 이어 개발한 귀납적 학습방법의 하나이다. C5.0 이전에 귀납적 학습방법의 예로는 기계학습(Machine Learning)이라는 이름 하에 발전해온 CLS, ACLS, ID3 등과 통계학에 기반을 둔 CART, CHAID 등이 있다(Berry & Linoff, 2000). C5.0 에서는 학습자료들을 분류하기 위해 사전에 정의된 등급(Class)과 속성(Property) 들간의 관계를 파악하여 단계적으로 의사결정트리(Decision Tree)를 형성한다. 특히, C5.0은 기존의 ID3가 가지고 있었던 문제점인 의사결정트리 생성에서 사용되는 Gain Criterion이 가지는 편향성 문제와 숫자형 변수를 다룰 수 없는 문제점을 개선한 것이다(이상호 & 지원철, 1998). C5.0의 분석과정은 다음과 같은 과정을 통해서 도출이 가능하다(Quinlan & Quinlan, 1997). 임의의 사례로 구성된 집합 S에 C_j 라는 속성이 포함되어 있는 사례의 갯수는 [식 6]과 같이 표시할 수 있다.

$$\frac{freq(C_j, S)}{|S|} \qquad [식\ 6]$$

그리고, [식 6]이 제공하는 정보는 [식 7]과 같이 나타낼 수 있다.

$$-\log_2\left(\frac{freq(C_j, S)}{|S|}\right) \quad bits \qquad [식\ 7]$$

이때 특정한 속성이 제공하는 기대된 정보(Expected Information)를 파악하기 위해서는 전체 사례 S에서 차지하는 속성의 합을 사용하는데 [식 8]과 같이 표현한다.

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times -\log_2\left(\frac{freq(C_j, S)}{|S|}\right) \quad bits$$

[식 8]

<표 2> 세 가지 데이터 마이닝 기법의 특징

	인공신경망	로짓 모형	C5.0 (귀납적 학습방법)
최적화 기준	일반화된 델타규칙	공분산행렬	Gain Ratio
최적화 절차	역전파 학습	오차의 최소화	반복적인 분할
모형의 가정	없음	여러 가지 통계적 가정	상충되는 경우가 없음
지식표현 형태	가중치 네트워크 구조	선형결합 모형	If-Then 규칙(Decision Tree)
지식의 구조	비명제형 지식	반명제형 지식	명제형 지식
주요장점	견고성 (Robustness) 학습력 (Learnability)	설명력, 검증성, 통계적 유의수준 제공	설명력, 쉬운 이해성
주요단점	설명이 난해	가정의 만족이 난해	상충되는 문제의 처리가 난해

(자료: 김광용, 1998 수정)

학습 사례(Training Case)의 집합을 적용할 때는 사례에서 T라는 속성을 식별하는데 필요한 평균적인 정보의 양인 info(T)를 측정해야 한다[4]. 이때 테스트 X에서 n개의 출력으로 분할된 T의 유사한 측정치를 고려하면, 이때 기대되는 정보의 필요량은 부분집합들의 가중치의 합으로 표시될 수 있다.

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \qquad [식\ 9]$$

그리고, 테스트 X에 의해 분할된 T에 의해 획득되는 정보의 양은 [식 10]과 같이 나타낸다.

$$gain(X) = info(T) - info_X(T) \qquad [식\ 10]$$

흔히 Gain Criterion이라 하는 C5.0의 분할 수준을 결정하는 기준은 이러한 Information Gain을 극대화 시키는 테스트를 선택하는 것이다.

한편, C5.0은 생성된 의사결정트리가 지나치게 많은 단계와 리프 (Leap) 노드를 가질 경우에 학습된 의사결정트리의 일반화 능력을 제고하기 위하여 리프노드를 제거하는 방법인 프루닝 (Pruning)을 시행한다. 일반적으로 프루닝을 통하여 예측력은 향상될 수 있으나 오류율도 증가하기 때문에 C5.0 에서는 오류기반 프루닝을 통하여 오류율의 증가를 통제한다. 또한, C5.0에서는 의사결정트리가 해석하기 난해하다는 점을 해결하기 위하여 자동적으로 If-Then 규칙을 생성해 주는 특징이 있다. 규칙은 프루닝 전 단계에서 생성하여 명시적 삭제, Contingency Table, 유의성 검증 등의 방법으로 생성된 규칙들을 간략화 한다(이상호 & 지원철, 1998). 이상과 같이 본 연구에서 사용되는 세 가지 데이터 마이닝의 특징을 표로 정리하면 <표 2>와 같다.

이들 세가지 방법론의 가장 큰 차이점 중의 하나는 지식의 구조이다. 인공신경망은 지식의 표현이 각 계층별로 노드 사이에 연결된 가중치로 표현되는 비명제형 지식(Non-Propositional Knowledge)의 구조를 가지고 있어 이러한 지식을 해석하거나 분석하는 것은 매우 어려운 일이다. 그러나 인공신경망은 견고성과 학습력 측면이 매우 강한 모형이므로 다른 기법과 혼용하여 많이 사용된다(김광용, 1998; 1999). 로짓 모델과 같은 통계적 모형은 반명제형 지식 (Pseudo-Propositional Knowledge) 구조를 가지고 있다. 즉 사

4) 이러한 정보의 양을 집합 S의 엔트로피(Entropy) 라고도 한다.

용된 변수의 상대적 가중치와 그에 대응하는 변수의 값을 곱한 후에 그 결과들의 합에 기인해서 의사결정을 하는 것은 명제형 지식구조이나 각 변수의 통계적 추정은 비명제형 지식구조이기 때문이다(김광용, 1999). 그러나 로짓 모델의 경우 모형이 도출한 결과에 대한 충분한 설명력과 검증이 가능하고 통계적인 유의수준을 제시하기 때문에 인공신경망과 같이 사용할 경우 그 상호보완이 가능하다. 끝으로 귀납적 학습방법의 한 예인 C5.0은 "If-Then" 형태로 표현되는 규칙들의 집합형태인 명제형 지식(Propositional Knowledge)이다(Quinlan, 1986). 이러한 명제형 지식구조는 쉬운 이해력과 분석력을 제공하기 때문에 전문가뿐만 아니라 일반 사용자들도 쉽게 이해가 가능하다.

4. 모형개발

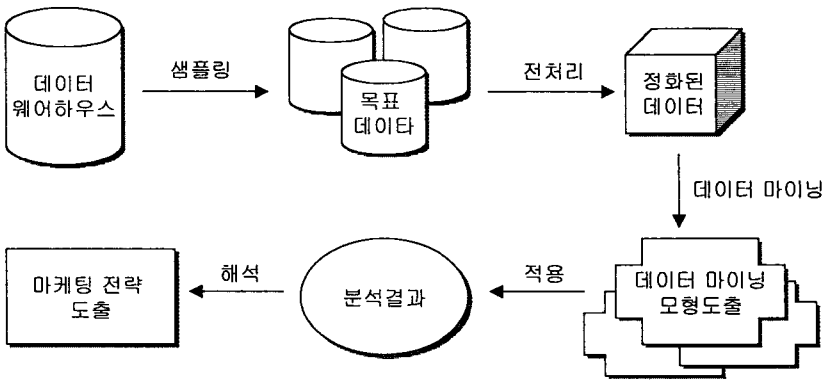
4.1 이탈고객관리 모형

본 연구에서는 데이터 마이닝을 이용하여 신용카드 보유고객의 유지 및 이탈여부를 파악하고 관리하기 위하여 <그림 3>과 같은 연구절차를 모형화 하였다.

이 절차는 기본적인 데이터 마이닝의 전개과정을 이용하였으나 세부 방법론에 있어서는 본 연구에 적합하도록 변형하였다. 그 과정을 살펴보면 첫째, 전체 신용카드 보유자 중에서 목표데이터를 선정하였다. 둘째, 여기에 각 분석기법이 요구하는 형태에 따라서 데이터의 형태를 변화시키고 각 데이터의 정규성을 검정하였다. 셋째, 인공신경망, 로짓모델, C5.0의 데이터 마이닝 기법을 이용하여 정화된 데이터로부터 모형을 도출하고 이를 검증자료에 적용하여 모형의 예측력을 검증하였다. 끝으로 이러한 자료를 분석하고 해석하여 마케팅 전략을 제시하였다.

4.2 자료 및 변수

본 연구에 사용된 자료는 1997년 4월부터 2000년 10월 현재까지 카드를 보유하고 있는 고객과 그 사이에 카드를 해지한 이탈고객에 대한 정보로 구성되어 있다. 이와 같이 최근의 자료를 이용한 이유는 최근의 자료를 이용할수록 현재 시점에서의 정확한 분석이 가능하기 때문이다. 분



<그림 3> 본 연구의 연구절차

석을 위해 먼저 조사된 자료에 대한 기초적인 인구통계학적인 분석과 각각의 이용금액에 관련된 정보를 추출하였다. 이를 바탕으로 모집단에서 이탈하지 않은 고객과 이탈한 고객을 각각 4,605개씩 추출하였다. 자료에 대한 간략한 설명은 <표 3>과 같다.

<표 3> 자료의 구성요소[5)]

변수명	설 명
MON2REN*	신용카드 갱신일 (Month to Renewal)
LS1, LS2, LS3	고객의 이용한도액 (Line Size). 최근 3개월** 자료를 이용
CLS1, CLS2, CLS3	다른 방식에 의해 계산된 Line Size.
AGE	신용카드 보유자의 나이
GENDER	카드 보유자의 성별
AVERAGE	최근 3개월간 이용금액의 평균
AGEING	지불 방법, 할부기간 (Current - 180 Day Past Due)
INTEREST	최근 3개월 간 발생한 이자의 평균
STATUS	신용카드 보유 여부

본 연구에 사용될 자료에 대해 조금 더 자세히 살펴보면 다음과 같다. 본 연구에서 거래금액과 연체금액을 최근 3개월 자료로 국한한 이유는 3개월이 고객의 신용정도를 파악하는 최소 기간이기 때문이다. 즉, 신용카드 사에서는 고객이 3개월 이상 연체를 하면 신용 불량고객으로 분류를 한다 (주석진 등, 1999). 따라서, 3개월 정도의 거래액과 연체금액을 이용하면 고객의 신용불량 여부를 파악할 수 있다. 그러나 신용불량자가 반드시 카드 해지자와 연결되는 것은 아니기 때문에 이 변수의 유의성 여부는 분석을 통해서 파악해야 할 것으로 보인다. 이용한도액도 이러한 측면에서 3개월치 자료를 이용하였다. 이외에 신용카드 보유자의 나이나 성별 (남자:0 여자:1)은 이미 신용카드 관련 연구에서 유의한 변수로 소개하고 있어서 이용하였다. 지불방법은 신용카드일 경우에만 존재하는 독특한 변수로 신용구매를 한 후 그 다음달에 일시불로 갚을 것인지 아니면 그 다음달에 갚을 것인지의 여부를 월별로 최장 6개월까지 두어 총 7개의 카테고리로 구분하였다 (0, 3, 6, 9, 12, 15). 또한 신용카드 보유 여부를 두 개의 카테고리로 나누었다 (유지:0, 이탈 1). 이와 같이 구성된 자료의 표본 수를 <표 4>에 간략히 제시하였다.

<표 4> 본 연구에 사용된 표본의 수

구 분	표본수
이탈하지 않은 고객	4,605
이탈한 고객	4,605
합 계	9,210

한편, 데이터 마이닝에서 사용될 학습용 자료와 검증용 자료를 구분하기 위하여 <표 5>와 같이 세가지 그룹으로 전체자료를 다시 나누었다. 인공신경망의 경우에는 검증용 집합(Validation Set)외에 인공신경망 학습 시 학습수준을 조정하기 위한 테스트용 집합(Test Set)을 따로 두었으나, 로짓모형과 C5.0에서는 그러한 절차가 필요 없는 관계로 테스트용 집합을 학습용 집합(Train-

5) * 신용카드 고객들은 카드를 갱신할 때 수수료를 내야하는데 일반적으로 갱신하기 전월에 해지하는 경향이 있다. 이에 회사에서는 카드의 갱신일을 가지고 있다. 본 연구에서는 이 정보를 해당 고객이 몇 개월 동안 카드를 보유하고 있는지의 정보로도 활용하였다.
** 최근 3개월이란 의미는 이탈하지 않은 고객의 경우 2000년 8월, 9월, 10월의 3달을 의미하고 이탈한 고객의 경우에는 이탈 시점에서 3개월 이전부터의 이용내역을 의미한다.

ing Set) 에 포함시켜 분석에 사용하였다.

<표 5> 학습용 자료와 검증용 자료의 준비

구 분	인공신경망	
	표본수	비 율
학습용 집합	5,526	60%
테스트용 집합	1,842	20%
검증용 집합	1,842	20%
합 계	9,210	100%

구 분	로짓모형 · C5.0	
	표본수	비 율
학습용 집합	7,368	80%
검증용 집합	1,842	20%
합 계	9,210	100%

4.3 자료의 전처리

데이터 마이닝 기법을 적용하기 위해서는 자료에 대한 전처리 과정이 반드시 필요하다. 그 이유는 대상 변수 중에서 최소한의 변수로 모형을 만드는 것의 모형의 간명성(Parsimonious) 측면에서도 좋고, 차후 구축된 모형을 설명하기에도 유리하다. 본 연구에서는 출력변수(종속변수)이 명목(Nominal)척도 이고, 입력변수(독립변수)에는 명목척도와 비율(Scale)척도가 섞여 있는 관계로 각각의 유형에 따라서 T-test와 Chi-Square 검정을 이용하여 유용한 변수를 추출하고자 하였다.

입력변수 중에서 비율척도로 구성된 MON2REN, CLS1, CLS2, CLS3, LS1, LS2, CLS3, AGE, AVERAGE, INTEREST에 대해서는 T-Test를 실시하였고, 명목척도인 GENDER, AGE 에 대해서는 Chi-Square 검정을 실시하였다.

<표 6>에 그 결과가 나와 있다. 분석 결과 본 연구에서 이용하고자 하는 모든 입력변수가 통계적으로 유의함을 알 수 있었다. 이에 본 연구에서는 이들 변수를 대상으로 데이터 마이닝 기법

<표 6> 입력변수의 유의성 검정 결과

• T-Test분석 결과

변수명	이탈하지 않은 고객		이탈한 고객		전체 고객	
	t-value	p-value	t-value	p-value	t-value	p-value
MON2REN	89.951	0.000	55.101	0.000	99.322	0.000
CLS1	63.636	0.000	18.951	0.000	58.365	0.000
CLS2	63.264	0.000	26.609	0.000	62.191	0.000
CLS3	63.428	0.000	25.000	0.000	61.608	0.000
LS1	159.816	0.000	164.434	0.000	225.208	0.000
LS2	159.583	0.000	164.427	0.000	225.126	0.000
LS3	159.906	0.000	164.567	0.000	225.508	0.000
AGE	286.919	0.000	284.607	0.000	394.518	0.000
AVERAGE	47.408	0.000	24.788	0.000	49.494	0.000
INTEREST	62.580	0.000	22.263	0.000	59.550	0.000

• Chi-Square 분석 결과(*p<0.1)

변수명	이탈하지 않은 고객		이탈한 고객		전체 고객	
	Chi-Square value	p-value	Chi-Square value	p-value	Chi-Square value	p-value
GENDER	3.614	0.057*	·	·	3.676	0.055*
AGEING	·	·	·	·	48972.805	0.000

을 적용하기로 하였다.

5. 실험결과 및 분석

5.1 모형별 실험 결과

5.1.1 인공신경망 분석결과

본 연구에서는 인공신경망으로 신용카드 고객의 이용패턴을 분석하기 위하여 다계층 퍼셉트론 인공신경망을 사용하였다. 또한, 인공신경망이 읽을 수 있도록 명목척도는 더미(Dummy)화 시켰다. 이렇게 해서 구성된 인공신경망은 입력층이 18개[6]이고 출력층이 1개인 인공신경망 모형을 구축하였다. 한편, 은닉층 노드의 개수는 반복실험을 통해서 최상의 개수를 결정하고자 하였는데 실험결과 은닉층이 10개 일 때 성과가 가장 좋은 것으로 나타났다. 학습방법은 시그모이드 함수를 이용한 역전파 학습 알고리즘을 이용하였다.

총 9,210 개의 자료 중에서 5,526 개의 자료를 학습용 집합으로 하고 1,842개를 테스트용 집합으로 하였으며 1,842개를 검증용 집합으로 이용하였다. 각 집합에는 신용카드 보유고객과 이탈고객이 동일한 비율로 들어 있도록 하였다. 실험용 소프트웨어로는 Neuro Shell 2를 이용하였다. 이와 같은 조건으로 인공신경망을 학습한 결과 전체 입력변수 중에서 MON2REN, CLS1, AVERAGE, AGEING0가 학습과정에 있어서 가장 큰 영향을 미치는 것으로 나타났다. 이것은 독립변수 중에서 신용카드 보유자의 가입기간, 최근 이용한도, 최근 3개월간 평균사용액, 할부를 사용하지 않는 등의 특징이 신용카드를 보유하고 있는지 이탈했는지를 파악하는데 중요한 변수로 보는 것이다. <표 7>에는 인공신경망에 의한 실험결과가 제시되어 있다.

<표 7> 인공신경망 실험결과

구분	학습용 집합 (5,526)		테스트용 집합 (1,842)		검증용 집합 (1,842)	
	개수	비율	개수	비율	개수	비율
적중률	4,569	82.68%	1,516	82.30%	1,546	83.93%
1종오류	543	9.83%	191	10.37%	161	8.74%
2종오류	414	7.49%	135	7.33%	135	7.33%

분석결과 검증용 집합에 대해서도 83.93%의 적중률을 나타내어 상당히 예측력이 높은 것으로 분석이 되었다. 또한, 이탈한 고객을 이탈하지 않은 것으로 분류한 1종 오류의 경우 8.74%, 이탈하지 않은 고객을 이탈한 것으로 분류한 2종 오류의 경우는 7.33%로 두 오류가 모두 균형있게 낮게 도출되어 모형이 상당히 예측력도 높으면서 신뢰성도 높은 것으로 판단되었다. 일반적으로 1종오류와 2종오류의 편차가 심하게 나는 경우에는 분석에 사용된 자료의 표본 추출에 문제가 있을 수 있기 때문에 1·2종 오류의 차이가 거의 없는 것은 표본 추출이 잘 되었다고 볼 수 있다.

5.1.2 로짓모델 분석결과

로짓모델에서도 입력변수가 명목척도인 경우는 읽을 수 없는 관계로 인공신경망에서 사용한 변수와 동일한 형태의 변수를 사용하였다. 즉, 독립변수로 18개, 종속변수로 1개의 변수가 이용되었다. 단, 로짓모형의 경우 테스트용 집합을 이용하여 모형의 학습 수준을 조정하는 과정이 없는

6) 최초 비율척도 변수 9개, 더미화시킨 성별 2개, 지불유형 7개를 합쳐서 18개이다.

필요없는 관계로 테스트용 집합을 학습용 집합에 포함하여 총 7,368개의 자료를 학습용 집합으로 이용하였다. 실험용 소프트웨어로는 SPSS 10.0을 이용하였고, 분석방법은 분석에 사용되는 변수를 통계량에 의해 검증하는 Stepwise 방법을 사용하였다. 분석결과 8단계에 걸쳐 <표 8>과 같이 총 8개의 변수가 유의수준 0.01에서 유의한 변수로 도출되었다.

<표 8> 로짓모델 분석결과

	β계수	S.E	Wald	P값	EXP(B)
Constant	-17.603	3.898	20.388	.000	.000
CLS1	.000	.000	234.164	.000	1.000
CLS2	.000	.000	53.804	.000	1.000
LS1	.000	.000	33.274	.000	1.000
AGE	-.056	.004	184.943	.000	.945
GENDER	.224	.062	12.892	.000	1.251
AVERAGE	.000	.000	125.721	.000	1.000
AGEING0	18.176	3.893	21.798	.000	78312627.903
INTEREST	.002	.001	7.161	.007	1.002

도출된 변수로는 CLS1, CLS2, LS1, AGE, GENDER, AVERAGE, AGEING0, INTEREST 변수로 MON2REN을 제외하고는 인공신경망에서 유의한 변수가 모두 포함되어 있었다. 이를 8개 변수를 이용한 분석결과가 <표 9>에 제시되어 있다.

<표 9> 로짓모델 실험결과

구분	학습용 집합 (7,368)		검증용 집합 (1,842)	
	개수	비율	개수	비율
적중률	5,952	80.78%	1,539	83.55%
1종오류	1236	16.78%	271	14.71%
2종오류	174	2.36%	31	1.68%

분석결과 검증용 집합의 적중률이 83.55%로 인공신경망 분석결과 보다는 낮았지만 상당히 높은 적중률을 보여 주었다. 그러나 1종오류가 14.71%이고, 2종오류가 1.68%로 1종오류가 과다하게 분석되었으며 1종 오류와 2종 오류의 편차가 너무 심하여 구축된 모형을 신뢰하기가 어려울 것으로 판단되었다.

5.1.3 C5.0 분석결과

C5.0 에서는 입력변수의 경우 비율척도이던, 명목척도이던 상관이 없지만 같은 자료를 사용한다는 측면에서 인공신경망, 로짓모형에서 사용한 변수와 동일한 변수를 사용하였다. 또한, C5.0 역시 테스트용 집합을 이용하여 모형의 학습 수준을 조정하는 과정이 없는 필요없는 관계로 테스트용 집합을 학습용 집합에 포함하여 분석을 실시하였다. C5.0에서는 프루닝의 수준을 어느 정도로 할 지에 따라 의사결정트리의 수준이나 추출되는 규칙의 수가 변하기 때문에 이를 조정해야 한다. 본 연구에서는 SPSS사의 클레멘타인(Clementine) 5.2를 이용하였는데, 푸루닝 수준(Pruning Severity)을 75%로 하고 의사결정나무의 가지(Branch) 당 최소 레코드를 2개로 지정하고 분석을 실시하였다. 분석결과가 <표 10>에 나와 있는데 모형의 전체적인 적중율은 검증용 집합을 기준으로 했을 때 92.24%로 인공신경망이나 로짓모델에 비해 월등히 높았으며 1·2종 오류도 각각 5.41%와 9.86%로 상당히 작으면서 그 편차가 심하지 않아 상당히 잘 예측한 것으로 보인다.

<표 10> C5.0 분석결과

구분	학습용 집합 (7,368)		검증용 집합 (1,842)	
	개수	비율	개수	비율
적중률	6,910	93.78%	1,699	92.24%
1종오류	167	4.69%	47	5.41%
2종오류	291	7.65%	96	9.86%

한편, C5.0으로 학습한 결과 학습용 집합에서 39개의 규칙이 도출되었다. 이들 규칙 중 신용카드 이탈자를 분류할 수 있는 규칙이 17개, 보유자를 분류할 수 있는 규칙이 22개였다. 이때 우리의 주 관심사는 보유자도 중요하지만, 어떠한 고객이 이탈하는지가 가장 관심사이므로 이들 규칙을 면밀히 검토하였다.

이들 규칙을 해석하는 방법은 예를 들어 규칙 #1의 경우, 10월 달의 이용한도 금액이 41,470 이상이고 할부기간이 1개월 이상이면 이탈할 가능성이 높다는 것이다. 그런데 한가지 흥미로운 것은 규칙 #1, 규칙#5, 규칙#6에서 공통으로 나타나고 있듯이 할부기간이 1개월 이상인 사람은 일단 이탈할 가능성이 높은 즉, 신용불량이 될 수 있는 확률이 높음을 알 수 있다. 이와 같은 규칙을 통해 신용카드 업체에서는 어떠한 고객이 신용카드를 해지하고 이탈할 가능성이 높을 지 예측이 가능할 것이다. 또한, C5.0에서 규칙 추출에 사용된 변수들 역시 인공신경망이나 로짓모델에서 도출된 변수와 거의 유사하여 이들 변수를 통제변수로 하여 신용카드 고객들을 관리해야 할 것으로 판단되었다.

<표 11> C5.0에 의한 신용카드 이탈자 분류 규칙

```
Rules for 1.0:
 Rule #1 for 1.0:
   if CLS1 > 41470
   and AGEINGO > 0
   then -> 1.0

 Rule #2 for 1.0:
   if MON2REN > 1
   and MON2REN <= 22
   and CLS1 > -8
   and LS3 > 55000
   and LS3 <= 95000
   and AVERAGE > 1246
   and AVERAGE <= 2173
   and INTREST <= 120
   then -> 1.0

 Rule #3 for 1.0:
   if MON2REN > 1
   and MON2REN <= 22
   and LS3 <= 55000
   and GENDER > 0
   and AVERAGE > 657
   and INTREST <= 10
   then -> 1.0

 Rule #4 for 1.0:
   if MON2REN > 1
   and MON2REN <= 2
   and CLS2 > 4
   and CLS2 <= 1187
   then -> 1.0

 Rule #5 for 1.0:
   if MON2REN > 28
   and CLS1 > 589
   and AGEINGO > 0
   then -> 1.0

 Rule #6 for 1.0:
   if MON2REN <= 26
   and CLS1 > 163
   and AGEINGO > 0
   then -> 1.0

 Rule #7 for 1.0:
   if CLS2 <= 523
   and CLS1 <= -1277
   then -> 1.0

 Rule #8 for 1.0:
   if CLS1 <= -8
   and GENDER > 0
   and AVERAGE > 860
   and INTREST <= 7
   then -> 1.0

 Rule #9 for 1.0:
   if MON2REN > 1
   and MON2REN <= 22
   and CLS2 > 1213
   and LS3 <= 95000
   and GENDER <= 0
   and AVERAGE > 657
   and INTREST <= 120
   then -> 1.0

 Rule #10 for 1.0:
   if MON2REN > 1
   and MON2REN <= 27
   and CLS2 <= 523
   and CLS1 <= -8
   then -> 1.0

 Rule #11 for 1.0:
   if CLS2 > 523
   and CLS1 <= -499
   and GENDER <= 0
   and AVERAGE > 860
   then -> 1.0

 Rule #12 for 1.0:
   if MON2REN <= 25
   and CLS2 <= 1187
   and CLS1 <= 9
   and LS3 > 65000
   and GENDER > 0
   and INTREST > 0
   then -> 1.0

 Rule #13 for 1.0:
   if MON2REN <= 22
   and CLS2 > 1213
   and CLS1 > -8
   and LS3 > 46000
   and LS3 <= 95000
   and AVERAGE > 657
   and INTREST <= 120
   then -> 1.0

 Rule #14 for 1.0:
   if MON2REN > 1
   and MON2REN <= 25
   and CLS3 > -9
   and CLS2 <= 1187
   and LS1 > 46000
   and AVERAGE > 212
   and INTREST <= 0
   then -> 1.0

 Rule #15 for 1.0:
   if MON2REN > 16
   and MON2REN <= 25
   and CLS2 <= 1187
   and CLS1 > -8
   and LS1 > 30000
   and LS1 <= 35000
   and GENDER <= 0
   then -> 1.0

 Rule #16 for 1.0:
   if MON2REN > 2
   and MON2REN <= 11
   and CLS3 <= 373
   and LS1 <= 46000
   then -> 1.0

 Rule #17 for 1.0:
   if MON2REN <= 25
   and CLS2 <= 1187
   then -> 1.0
```

5.2 토의

본 연구는 여러 가지 데이터 마이닝 기법을 신용카드 사례에 적용하여 어떠한 기법이 신용카드 시장에서의 고객이탈을 잘 설명하는지 살펴보는데 주요한 연구목적이 있다. 본 연구결과를 종합하여 보면 신용카드 보유자의 이탈 가능성을 예측하는 변수로는 CLS1, CLS2, LS1, AGE, GENDER, AVERAGE, AGEING0, INTEREST 변수 등이 있었다. 이들 변수는 각각 신용카드 보유자의 이용한도액, 나이, 성별, 평균 이용액, 할부기간, 이자율과 관계된 변수들임을 알 수 있었다. 또한, 개별 분석기법의 입장에서 보면 <표 12>와 같은 결과를 얻을 수 있다.

<표 12> 분석기법간 성과 비교(검증용 자료를 대상으로 할 때)

구분	인공신경망		로짓 모델		C5.0	
	개수	비율	개수	비율	개수	비율
적중률	1,546	83.93%	1,539	83.55%	1,699	92.24%
1종오류	161	8.74%	271	14.71%	47	5.41%
2종오류	135	7.33%	31	1.68%	96	9.86%

<표 12>는 본 연구에서 사용한 3가지 데이터 마이닝 기법의 성과비교를 위하여 이미 분석된 결과를 정리한 것이다. 분석결과를 살펴보면 적중률과 1종오류, 그리고 2종오류의 측면에서 각 기법이 다소 강점과 약점에 차이가 있음을 알 수 있다. 그러나, 전체적으로는 C5.0이 본 자료에서는 인공신경망이나 로짓모델에 비해 더욱 높은 적중률과 비교적 낮은 오류율을 가지고 있다고 볼 수 있다. 따라서, 실무입장에서는 C5.0에 의해 도출된 <표 11>의 이탈자 분류규칙을 활용하여 이탈예상자를 관리한다면 지금보다 효과적인 이탈자 방지 프로그램을 운용할 수 있을 것으로 판단된다.

6. 결론 및 향후 연구방향

본 연구에서는 신용카드 보유고객을 대상으로 인공신경망, 로짓모형, C5.0의 세 가지 데이터 마이닝 기법을 이용하여 이들 고객의 특성을 파악하고자 하였다. 본 연구결과를 간략히 정리하면 다음과 같다.

첫째, 데이터 마이닝을 이용한 신용카드 고객의 이용행태 분석의 가능성을 제시했다는 점이다. 이미 서론에서도 언급한 바와 같이 국내외의 신용카드 관련 이용행위에 대한 연구는 주로 설문방법에 의한 연구들이 많았으며 이것도 주로 현재 카드를 보유하고 있는 고객을 대상으로만 이루어졌다. 따라서, 기존의 연구는 이탈고객의 특성을 전혀 고려하지 못하고 있어 분석의 한계점을 나타내었다. 그러나, 본 연구의 경우 데이터 마이닝 기법을 이용하여 80% 이상 고객의 신용카드 보유여부를 예측할 수 있었으며 여기에 미치는 영향요인과 그 특성을 규칙으로도 도출할 수가 있었다.

둘째, 기법간의 보완성 제공이다. 기존의 통계적 방법을 이용할 경우에는 설문지를 이용하면 분석은 용이하지만 실제 적용이 어렵고, 통계적 가정이 엄격하여 그 결과가 얼마나 신빙성이 있는지 받아들이기가 어려웠다. 그러나 본 연구의 경우 인공지능 기법인 인공신경망과 C5.0의 결과와 출계적 기법인 로짓모형의 결과를 유기적으로 해석에 이용함으로써 기법이 가지고 있는 특성을 이용하여 상호보완이 가능하다. 즉, 인공신경망은 견고하고 학습성이 뛰어나지만 해석하기 어려

운 점을 로짓모형이나 C5.0이 보완한다는 개념이다.

셋째, 이탈고객관리를 위한 마케팅 전략의 제시이다. 본 연구에서는 신용카드 보유자의 행위분석을 위해 실제 거래자료를 이용하였으며 이를 통하여 신용카드 보유자와 이탈자를 규정짓은 영향변수를 도출하였다. 또한, 이를 규칙으로 추출하여 신용카드 이탈자의 특성을 파악하였다. 신용카드사에서는 이러한 영향변수들을 통제변수로 사용하여 이탈예상 고객을 사전에 파악하여 이에 따른 전략적 조치가 가능하다. 또한, 이러한 변수의 도출은 고객을 충성고객 (Loyalty Customer)과 비충성고객으로 분류하여 차별화된 마케팅 전략을 수립할 수 있는 근거를 제공할 수 있을 것이다.

그러나 본 연구에서는 사용한 변수가 한정되어 있는 관계로 신용카드를 이용한 다음에서야 고객의 특성을 파악할 수 있다는 단점이 있다. 즉, 신용카드를 만들기 이전에 해당 고객의 특성을 파악하여 신용카드 이탈고객이 될 가능성을 예측하기는 어렵다는 것이다. 이러한 분석을 위해서는 신용카드 외에 다른 은행거래 실적과 같은 개인 신용에 대한 정보가 추가적으로 요구되는바 실질적으로는 구현하기가 어려울 것으로 생각된다. 하지만 점차 은행권들이 카드사, 보험사들과 결합하여 거대한 지주회사 (Holding Company)로의 변환을 추구하는 작금의 시점에서 이러한 추가적인 연구는 매우 중요할 것으로 생각된다. 또한, 이러한 추가적인 분석을 통해서 보다 신뢰도가 높고 정확성이 있는 결과를 도출할 수 있을 것으로 기대된다.

참고문헌

김광용, "여러가지 Inductive 방법에 대한 통합모형 개발과 그 실증적 유효성에 관한 연구", *한국경영과학회지*, 23권 3호(1998), 185-207.

김광용, "여러가지 Data Mining 기법으로부터 도출된 지식에 관한 전문가의 신뢰도에 관한 실증적 연구", *한국지능정보시스템학회논문지*, 5권 1호(1999), 125-143.

매일경제, 신용카드 고객 연령층 점차 젊어져, 매일경제신문 2000/07/23, 2000a.

매일경제, 국내 신용카드 역사: 69년 신세계카드 효시, 매일경제신문 2000/08/28, 2000b.

박근주, *소비자의 신용카드 사용행동에 관한 연구*, 서울대학교 대학원 석사학위논문, 1990.

박선태, *신용카드 소지자의 카드이용에 관한 실증적 연구*, 계명대학교 무역대학원 석사학위논문, 1995.

박찬실, *은행계 신용카드를 이용한 직장인들의 구매행동에 관한 실증적 연구*, 경남대학교 경영대학원 석사학위논문, 1995.

서경의, *대학생 소비자의 신용카드 사용행동에 관한 연구*, 서울대학교 대학원 석사학위 논문, 1997.

이건창, 정남호, "데이터 마이닝 기법과 지능형 에이전트 기법을 결합한 인터넷 쇼핑몰의 설계 및 구현에 관한 연구", *정보기술응용연구*, 1권 2호 (1999), 113-137.

이상영, *주부들의 신용카드에 대한 지식과 관리행동에 관한 연구*, 부산대학교 대학원 석사학위논문, 1995.

이상호, 지원철, "귀납적 학습방법들의 분류성능 비교: 기업신용평가의 경우", *한국지능정보시스템학회논문지*, 4권 4호, 1-22.

이영호, 지영숙, "도시민의 신용카드 사용패턴에 관한 연구", *한국가정관리학회지*, 5권 1호(1987), 51-68.

이윤금, 김주연, 조향숙, "복수신용카드 소지자들의 신용카드 사용행태와 부채부담에 관한 연구", *대한가정학회지*, 36권 11호(1998), 219-230.

이은희, *신용카드 관리행동의 체계론적 영향요인분석*, 충남대학교 교육대학원 석사학위논문, 1992.

이재희, "대학생들의 신용카드 인식 및 사용에 관한 연구", *한국생활과학회지*, 5권 2호(1996), 99-107.

조의준, *신용카드 이용자의 구매행동에 관한 실증적 연구*, 경남대학교 대학원 석사학위논문, 1994.

주석진, 김재경, 성태경, 김중한, "신용카드 고객의 신용 예측을 위한 지식기반 방법들: 적용 및 비교 연구", *한국지능정보시스템학회논문지*, 5권 1호(1999), 49-64.

지원철, 서민수, "데이터 마이닝을 활용한 공급사슬관리 의사결정지원시스템의 구조에 관한 연구", *경영정보학연구*, 8권 3호(1998), 51-74.

최재복, *은행계 신용카드 고객의 특성과 마아케팅 전략*, 경북대학교 경영대학원 석사학위논문, 1995.

한국일보, 국내 카드산업 성장 한계점, 한국일보, 2002/07/19.

Adriaans, P. and D. Zantinge, *Data Mining*, Addison-Wesely press, 1997.

Agrawal, D. and C. Schorling, "Market share forecasting: An Empirical Comparison of Artificial Neural Networks and Multinomial Logit model", *Journal of retailing*, Vol.72, No.4(1996), 383-407.

Awh, R.Y. and D. Waters, "Discriminant Analysis of Economic, Demographic, and Attitudinal Characteristics of Bank Charge-Card Holders: A Case Study", *Journal of Finance*, Vol.29 (1974), 973-980.

Ben-Akiva, M and S.R. Lerman, *Discrete Choice Analysis : Theory and Application to Travel Demand*, London, The MIT Press., 1993.

Berry, M.J.A., and G., Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, Wiley Computer Publishing, 1997.

Curtin, R.T. and T.S. Neubig, "Changes in Credit Card Use During 1978", *Institute for Social Research*, Working Paper, No.12, 1980.

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

Han, J., and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, 2000.

Hirschman, E.C., "Difference in consumer purchase behavior by credit card payment system", *Journal of Consumer Research*, Vol.6(1979), 58-66.

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

Jain, B.A. and N.B. Nag, "Performance Evaluation of Neural Network Decision Models", *Journal of Management Information Systems*, Vol.14, No.2(1997), 201-216.

Knisey, J., "Determinants of credit cards accounts: An application of tabit analysis", *Journal of Consumer Research*, Vol.9(1982), 179-180.

Lippmann, R.P., "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, Vol.3, No.4(1998), 4-22.

Mandell, L., *Credit Card Use in the United States*, Ann Arbor: Institute for Social Research, The University of Michigan, 1972.

Quinlan, J.R., and Quinlan, J., *C4.5: Programs for Machine Learning*, Morgan Kaufman Publishers, 1997.

Quinlan, R., "Induction of Decision Trees", *Machine Learning*, Vol.1(1986), 81-98.

Rumelhart, D.E., G.E. Hinton, and R.J. Williams, "Learning Internal Representations by Error Propagation", in D.E. Rumelhart and J.L. McClelland (Eds). *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*. Cambridge, MA:MIT Press, 1986.

Abstract

An Artificial Intelligence-based Data Mining Approach to Extracting Strategies for Reducing the Churning Rate in Credit Card Industry

Kun Chang Lee* · Namho Chung** · Kyung-shik Shin***

Data mining has received a lot of attention from practitioners. That is partly because it allows company to extract a set of useful knowledge about customers from database, thereby retaining current customers and magneting potential customers. This logic is especially essential in the field of credit card industry, where just 5% increase of number of customers is known to cause 120% increase in profit. The problem is how to retain current customers and even make them more loyal to company. However, previous studies lacked proposing extensive strategies of reducing the churning rate. In this sense, this study attempts to suggest such strategies by applying neural network, logistic regression, and C5.0 techniques to credit card data. We used a real data set of four years from 1997 to 2000, which were gathered from a credit card company. Experimental results revealed that our approach could yield robust strategies for retaining customers by reducing the churning rate.

Key words: Data Mining, Churning rate, Neural network, Logistic regression, C5.0

* School of Business Administration, Sung Kyun Kwan University
** Management Research Institute, Sung Kyun Kwan University
*** College of Business Administration, Ewha Woman University