한국지능정보시스템학회논문지 제7권 제1호 2001년 6월(pp. 177~197)

은행고객 세분화를 통한 이탈고객 관리분석
-가계성 예금을 중심으로-

이건창
성균관대학교 경영학부 교수
(leekc@skku.ac.kr)
권순재
성균관대학교 경영학부 박사과정
(sjkwon@dragon.skku.ac.kr)
신경식
이화여자대학교 경영대학 교수
(ksshin@ewha.ac.kr)

IMF이후로 우리나라의 은행들은 현재 큰 구조조정을 맞이하고 있으며 이 속에서 살아남기 위하여 나름대로의 전략을 수립하고 있다. 예를 들어 모 은행의 경우에는 평균 잔액이 일정수준을 넘지 아니하는 경우에는 고객들에게 이자를 지급하지 않는 전략을 수립하고 있다. 이에 기존의 고객의 유형을 분석하고 이를 전략에 활용하는 연구의 필요성이 높아지고 있다. 기존의 연구를 살펴보면 은행 고객들의 유형을 설문지 분석방법에 의존하여 몇 개의 군집으로 분류하고 이들의 집단별 특성을 연구하고자 하였다. 하지만 설문데이터의 경우 고객들의 실제적인 행동이 반영되지 못하는 한계점을 가지고 있다. 이에 본 연구에서는 C은행의 실제 고객 자료를 통하여 다양한 데이터 마이닝 기법을 적용하여 가계성 예금 고객을 세분화하였다. 또한 세분화된 고객을 중심으로 이들이 가계성 예금을 해지하고 다른 은행으로 이탈하는 집단의 특성을 분류하고 규칙을 도출하였다. 또한 이들을 관리하는 전략을 제시하였다.

1. 개 요

IMF이후로 우리나라의 은행들은 현재 큰 구조조정을 맞이하고 있으며 기존의 기업을 대상으로 전략을 추구해오던 기업들이 상대적으로 소규모의 투자이지만 안정성이 높은 개인을 중심으로 전략을 전환하고 있다. 따라서 기존의 개인고객들을 중심으로 이들의 고객의 유형을 분석하고 이를 마케팅 전략에 활용하려는 연구의 필요성이 높아지고 있다.

시장을 세분화하는 것은 시장내의 이질성을 분석하여 비교적 동질적인 하위시장을 파악하고 이 정보를 표적마케팅에 활용하려는 것이라고 할 수 있다. 박찬욱(1996)은 "동질"의 의미를 물건을 소비하는 소비자의 반응이 같음을 의미한다고 언급하고 시장세분화를 다음과 같이 크게 두 가지로 나누고 있다. 첫째, 상품의 구입, 정보의 요청, 특정 캠페인의 참여 등과 같은 고객 행동에 대한 정보를 바탕으로 고객을 세분화하는 것이고, 두 번째는 고객 행동에 대한 정보와는 상관없이 단순히 고객의 특성을 바탕으로 고객을 세분화한다는 것이다.

본 연구는 위의 두 가지 방법 중에서 전자에 해당하는 것으로 보다 구체적으로 살펴보면, 고객의 반응정보와 이에 영향을 미치는 변수들을 이용한 분석을 통하여 고객들의 반응확률을 계산해낼 수 있는 반응함수를 도출해내고, 도출해낸 반응함수를 통하여 고객 개개인에 대한 반응점수를 부여한 후 이 점수를 바탕으로 고객을 몇 개의 집단으로 분류할 수 있다. 이와 같은 고객 반응 정보 중 고객들의 자사 제품에 대한 지속적 사용여부, 즉 이탈여부에 초점을 두어 고객 세분화를 시도한 것이 이탈고객방지 분석이라고 할 수 있다. 따라서 고객이탈방지 분석의 궁극적인 목적은 고객 이탈의 원인을 파악하고 이를 이용하여 이탈예상 집단에 대한 관리 활동을 강화하여 이탈고객을 최소화시키는데 있다고 할 수 있다.

원래 이탈고객 및 유지고객의 유형을 세분화하고 이들의 특징을 분석하는 연구는 통신산업에서 기존의 통신수단을 해지하고 다른 업체의 통신수단으로 이동하는 고객들의 특징을 분석하기 위하여 연구가 시작되었다. 이에 본 연구에서는 이러한 방법론을 활용하여 가계성 예금의 분야에 적용시켜 보았다는데에 그 공헌점이 있다고 하겠다. 특히 IMF이후로 우리나라의 은행들은 현재 큰 구조조정을 맞이하고 있으며 이 속에서 살아남기 위하여 기존의 고객의 유형을 분석하고 이를 마케팅 전략에 활용하는 본 연구는 매우 시의 적절한 연구라고 사료된다. 이에 본 연구에서는 C은행의 실제 고객 자료를 통하여 데이터마이닝 기법을 적용하여 고객을 세분화한 다음 고객이 예금을 해지하고 다른 은행으로 이탈하는 집단을 분류하고 이들을 관리하는데에 연구의 목적이 있다.

본 연구의 구성은 다음과 같다. 2장에서는 금융시장의 세분화, 데이터마이닝, 데이터마이닝 관점에서의 고객의 세분화, 고객의 구매패턴에 대한 기존연구를 고찰한다. 3장에서는 본 연구에서 사용되는 연구방법론인 로짓분석, 인공신경망, C5.0에 대하여 살펴본다. 4장에서는 데이터마이닝 모형을 구축하기 위하여 분석에 사용되는 자료를 결정하고 입출력 변수를 정의한다. 5장에서는 실험 및 분석을 실시하고 이어지는 6장에서는 이 분석결과를 중심으로 마케팅 전략을 제시하고자 한다. 마지막 장에서는 결론 및 향후 연구방향에 대하여 언급한다.

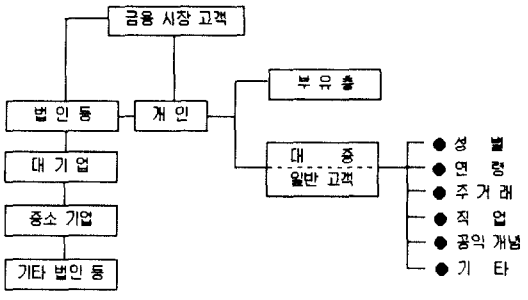
2. 기존문헌 연구

본 절에서는 금융시장의 세분화, 데이터마이닝, 데이터마이닝 관점에서의 고객의 세분화, 고객의 구매패턴에 대한 기존연구를 중심으로 분석을 실시한다.

2.1 금융시장

금융시장에 있어 고객은 대기업, 중소기업, 기타 법인, 개인 등으로 크게 나눌 수 있다. 그 중에서도 전략적으로 개인 고객의 경우 우선 부유층(VIP)고객과 일반 대중(mass)으로 구분된다. 기존의 많은 연구에서는 일반 대중은 또한 거주지와 직업, 연령, 성별 등에 따라 세분화하고 있으며, 이렇게 세분화가 되면 세분화된 대상별로 라이프 사이클이 다르고 금융거래 형태가 다르게 된다. 금융시장 시장세분화는 우선 마케팅 전략의 기초가 되는 시장조사를 통하여 고객들의 금융시장에 대한 지각과 선호를 파악하고 이를 기초로 시장 세분화 전략을 수립하고 표적시장을

선정하게 된다. 또한 표적시장 내에서 한 은행의 금융상품을 고객의 인식 속에 어떤 위치에 부각시켜야 할 것인지를 결정한다. 금융시장의 세분화에 관한 대표적인 연구로써 하영원(1996)은 가계성 예금에 대한 세분화를 실시하기 위하여 전국의 은행을 대상으로 1000여부의 설문을 실시하였다. 그는 설문조사의 결과를 바탕으로 고객을 세 가지 그룹으로 세분화하여 각 그룹에 적절한 시장전략을 제시하였다. 하지만 전통적인 마케팅 조사방법에 근거하여 설문을 중심으로 시장을 세분화하였으므로 실제 데이터가 가지고 있는 충분한 의미를 귀납적으로 도출할 수 없다는 한계점을 가지고 있다. Marsh(1998)은 단일 상품이 모든 고객들의 욕구를 만족시켜줄 수 있는 것이 아니며, 모든 고객들이 은행에게 이익을 가져다 줄 수 없으므로 은행에서는 가치 있는 고객 시장을 찾아내는 것이 금융시장 세분화에서 중요한 과정이라고 언급하였다. 아래의 <그림 1>에는 금융시장의 고객을 구분하고 있다.



<그림 1> 금융시장의 고객 구조

(자료원: Marsh, 1998)

2.2 데이터마이닝

데이터마이닝이라고 하는 것은 대량의 데이터로부터 묵시적이고 잠재적인 알려지지 않은 정보를 찾아내는 것이다. 이러한 데이터마이닝 기법을 이용함으로써 기업은 대용량의 데이터베이스에 숨겨져 있는 데이터간의 관계, 패턴 등을 찾아내고 이를 의미 있는 정보로 변환하여 기업의 의사결정과정을 지원하고 그 결과를 예측할 수 있는 것으로 기대한다.

데이터마이닝을 위한 구체적인 방법론으로는 여러 가지 기법들이 사용되고 있으며, 특히 통계 및 인공신경망(Artificial Neural Networks), 카트(CART: Classification and Regression Tree), 체이드(CHAID: Chi Square Automatic Interactive Detection) 등의 의사결정나무(Decision Free) 기법, 유전자 알고리즘(Genetic Algorithms) 기법 등이 많이 사용되고 있다(Fayyad et al., 1996).

또한 데이터마이닝의 유형을 살펴보면, 연관규칙(Association), 연속규칙(Sequence), 분류규칙(Classification), 데이터군집화(Clustering) 등 4가지 유형으로 나뉘어진다. 그리고 이들의 각 유형에 적합한 데이터마이닝 기법이 많은 연구자들에 의하여 연구되어 왔다. 이러한 기법과 문제분야를 정리하면 다음 <표 1>과 같다.

<표 1> 다양한 기법, 분석내용, 알고리즘

	분류내용	알고리즘
연관규칙	다수의 사건 중에서 2가지 이상의 사건이 동시에 일어날 가능성 발견	Apriori, GRI
연속규칙	어떠한 추세를 분석하고 이를 가지고 행후를 예측	회귀분석, 추세분석
분류규칙	몇 가지 기준 점을 이용하여 분류	CART, C5.0, QUEST
데이터 군집화	데이터의 특성을 몇 개의 집단별로 분류	K-Means

2.3 데이터마이닝(데이터베이스) 관점에서의 고객세분화

데이터베이스 마케팅(Database Marketing)이란 기업의 기존 고객 혹은 가망 고객에 대한 데이터를 사내 전산 시스템에 축적해놓고, 고객에 대한 개별적인 정보속성을 고려하여 마케팅 활동에 적용하는 것을 의미한다. 그리고 데이터베이스 마케팅의 대전제는 "모든 고객이 똑같은 구매행동을 나타내지 않는다" 라는 것이다(송현수, 1998).

정보기술의 발전으로 기업의 고객정보 처리능력이 급속도로 향상되어 고객의 특성에 대한 정보를 기업내부에 축적해놓고 개별고객에 대한 차별적인 마케팅을 실시하게되는 것이다(김정수, 1997) 이러한 차별적인 마케팅을 실시하게 하는 데이터베이스 마케팅의 궁극적인 목적은 고객의 이탈율을 최소화하고 이들의 반복구매를 촉진시켜 거래관계를 심화시켜 나가는 고정 고객화 전략을 통해 매출을 촉진시키고 기업이익을 극대화하는 것이다.

기존의 데이터베이스마케팅에서는 흔히 소비자 행동모델을 만들어 이를 적용하여왔다. 여기서 이야기하는 소비자 행동모델링이란 소비자의 미래행동을 예측해내는 것으로 일반적인 통계확률 이론을 이용한 샘플분석으로 파악된 소비자의 속성정보를 가지고 미래 행동을 추론하는 것이다. 과거 행동에 대한 데이터가 분석의 출발점으로서 작용한다 즉 소비자 행동을 예측하게 만들어주는 가장 강력한 변수는 특정고객의 과거 구매 행동이라는 것이 모델링의 대전제인 것이다. 그리고 이런 모델링 과정을 통하여 샘플자료에서 고객속성과 소비자의 행동에 상관관계를 가지고 전체소비자(모집단)에서 이 모델링을 적용하여 각 개인의 특정행동을 미리 예측한다는 것이다(송현수, 1998).

이중에서 가장 알려진 고객세분화 방법은 R-F-M분석방법으로 여기서 (Recency:거래의 최근성), (Frequency: 거래빈도), (Monetary: 거래규모) 등으로 단순화시킴으로써 고객이 얼마나 최근에 구입했는가, 고객이 얼마나 빈번하게 우리제품을 구매했는가, 고객이 구입했던 총 금액은 어느 정도인가 등에 관한 정보를 축약하여 구입가능성이 높은 고객들을 세분화할 수 있는 분석방법이다. 하지만 R-F-M분석방법을 이를 구축하는데 많은 시간이 소요되며, 많은 누적 데이터가 있어야 좋은 R-F-M분석 모형이 구축할 수 있다는 단점이 있다. 즉 적어도 2~3년 정도의 데이터가 구축되어야 한다. 그리고 구축시간에 비하여 시장은 매우 빠른 속도로 변하고 있다는 점 또한 한계점이다. 또한 완성한 모델링을 활용하여 분석을 실시하였을 때 신뢰성이 많이 떨어진다는 사실이 있다. 따라서 고객의 요구사항을 적절히 반영하는 정교한 모델으로써는 한계점을 가지게 된다(http://www.spss.com).

이러한 한계점등으로 본 연구에서는 대량의 데이터를 정교하게 모델링할 수 있는 데이터마이닝 기법을 이용하여 가계성 예금에 대한 고객의 이탈여부를 분석하고 이를 통하여 데이터베이스에 근거한 마케팅 전략을 제시하고자 한다.

2.4 고객의 구매패턴 분석

고객의 구매 패턴을 분석하고 이를 의사결정에 이용하려는 연구는 전통적인 마케팅에서 로짓(Logit) 함수를 이용하여 예전부터 많은 기존 연구가 이루어져 왔다. 특히 최근에는 다양한 데이터마이닝 툴을 이용하여 고객의 특성 및 성향을

분석하여 구매 패턴을 인식하는 연구가 이루어졌다. 더욱이 인터넷 시대로 경제가 개편되면서 온라인 쇼핑몰을 통한 매출이 급증하게되고 이들에 대한 관심이 높아지자 인터넷 쇼핑몰을 이용하는 고객들의 구매 패턴을 분석, 활용하려는 연구가 태동하고 있다. 하지만 웹로그를 통하여 고객의 구매 패턴을 분석하는 연구 결과는 아직 제시된 바가 없다. 이에 본 연구에서는 마케팅에서 기존의 전통적인 연구 방법을 통하여 일반적인 사용자의 패턴을 예측하기 위한 연구를 중심으로 기존문헌 연구를 고찰하기로 한다.

일반적으로 소비자의 구매를 예측하기 위해서 전통적인 통계방법인 회귀분석이나 판별분석, 로짓모형이 이용되어 왔으나 1980년대 후반부터 등장하기 시작한 귀납적 학습방법, 인공신경망(Neural Networks)등의 인공지능 기법이 기업의 신용평가 및 도산 예측에 활용되면서 인공지능 기법들이 많이 활용되었다.

로짓모형의 경우 대안 선택을 예측하는데 널리 사용되어 왔으며 인공신경망의 예측능력을 이용한 문제해결은 주로 도산예측, 신용평가, 그리고 주가예측 등의 문제에 활용되어 왔다. 먼저 로짓모형을 대안의 선택을 예측하는데 이용한 연구는 점포선택(안광호와 채서일 1993), 제품선택(한충민과 이한구 1996), 아파트 구매(안광호와 임영균 1996)등 여러 분야에서 적용되고 있다. 안광호와 채서일(1993)은 소비자 선택행위에 대한 설명과 예측에 유용한 모델로서 MNL(Multinomial Logit)모형을 소개하고 백화점에 대한 소비자의 선택행위를 실증 분석하였다. 그들은 점포크기와 근접성외에 점포이미지를 점포선택에 중요한 속성으로 파악하고 MNL 모형으로 소비자의 점포선택행위를 실증 분석한 결과, 추정된 모수의 결과가 소비자의 점포선택행위를 잘 설명하고 예측하는 결과를 나타내었다. 안광호와 임영균(1996)은 우리나라 소비자의 아파트 선택행위를 Nested Logit 모형과 MNL(Multinomial Logit)모형으로 분석하여 소비자의 대안선택과정과 시장경쟁구조를 파악하였다. 연구결과 Nested Logit 모형이 MNL 모형보다 소비자의 아파트 선택과정을 더 잘 설명하는 것으로 나타나 소비자의 대안선택이 계층적 의사결정을 따를 경우 Nested Logit 모형이 MNL 모형보다 더 적절함을 제시하였다. 한충민과 이한구(1998)는 우리나라 소비자의 수입담배 구매 행위에 영향을 미치는 요인들을 로짓 모델을 이용하여 실증하였는데, 제품속성의 우위, 소비자 외향성과 사회참여도, 소비자 애국심이 수입담배의 채택과 구매에 유의한 영향을 미치는 것으로 나타났다.

3. 연구방법론

가계성 예금을 분석하기 위한 방법론으로 전통적인 통계분석에서 많이 사용한 로짓 분석과 데이터마이닝 기법 중 가장 일반화된 인공신경망 분석, 그리고 규칙을 추출하는 대표적인 방법론으로 C5.0에 대하여 살펴보도록 한다.

3.1 전통적인 통계분석-로짓분석

로짓 모델의 기본 개념은 한 개인이 특정대안(예를 들면 제품 또는 점포에 대하여 느끼는 효용(Utility)을 그 대안에 대한 선택확률로 연결시켜준다는 데에 있다. 어떤 소비자가 특정제품에 대하여 느끼는 효용이 클수록 그 제품을 선택할 확률이 높아진다. 예를 들어, A, B, C 라는 세 가

지 제품 중에서 소비자가 A를 선택하였다면, 제품 A의 상대적인 효용이 B, C에 비하여 그만큼 높기 때문이라고 할 수 있다. 이렇듯 효용은 선택행동에 중요한 요소로 작용하며, 로짓 분석에서도 모델 추정의 근간을 이룬다(채서일, 2000).

로짓 모델은 다음 [식-1] 같은 과정을 통하여 도출된다. 소비자가 느끼는 효용은 아래와 같이 관찰이 가능한 부분과 관찰할 수 없는 부분(오차)로 구성된다. 이때 소비자는 효용의 극대화라는 기준 하에서 대안을 선택하는데 즉 고려중인 대안들 중에서 가장 효용이 높은 대안을 선택하게 된다.

$$U_{ij} = V_{ij} + \varepsilon_{ij} \qquad [식-1]$$

여기서 U_{ij} = 소비자 i가 제품 j에 대하여 느끼는 효용
V_{ij} = 독립변수들에 의하여 설명되어지는 부분
ε_{ij} = 오차

여기서 V_{ij}는 소비자 i가 제품 j로부터 느끼는 효용을 나타내는 요소로써 결정요소(Deterministic Component)라고도 한다. ε_{ij}는 오차로써 독립적이고 동일한 Weibull 분포를 따른 다고 가정한다. 결정적 요소에 영향을 미치는 변수들을 선형으로 표현하면 다음 [식-2]와 같다.

$$V_{ij} = \alpha_i + \beta X_{ij} \qquad [식-2]$$

α_i는 각 제품들이 고유하게 보유하고 있는 특성이고 β는 추정해야할 모수이며 X_{ij}는 제품선택에 영향을 미치는 요인들의 벡터이다. 앞서 무작위오차 ε_{ij}는 서로 독립적이며 Weibull 분포를 따른다고 가정하였다. 따라서 Weibull 분포를 따른다고 가정하였을 때 소비자가 제품을 선택할 확률은 [식-3]과 같이 나타낼 수 있다.

$$P^{*}_{ij} = \frac{e^{v^{*}_{ij}}}{\sum_{j}^{N} e^{v^{*}_{ij}}} \qquad [식-3]$$

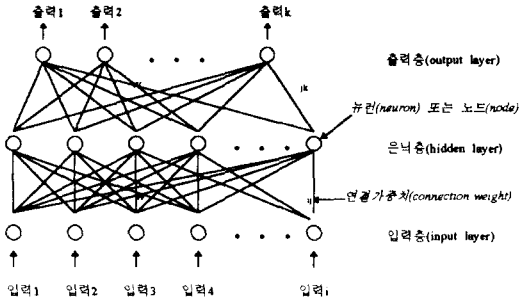
P_{ij} : 소비자 i가 제품 j를 확률
V_{hit} : 소비자 i가 제품 j에 대해 구매시점에 느끼는 효용
N : 고려대안수

3.2 인공신경망

신경망 모형은 신경생리학 분야에서 두뇌의 활동을 이해하고자 하는 목적하에 신경의 작업을 설명하려는 시도에서 출발하여 생물학적인 프로세스를 컴퓨터를 이용하여 모형화 하려는 노력에서 비롯된 것으로, 80년대 이후 컴퓨터 성능의 진보에 따라 빠르게 진보되고 있다(Bishop, 1995)

신경망은 은닉마디와 은닉계층을 몇 개로 하느냐에 따라 다양한 아키텍처를 구성할 수 있다. 특정문제나 데이터에 대해 최적의 아키텍처가 무엇인지 정해져 있지는 않다. 적용하고자 하는 문제 혹은 데이터가 얼마나 복잡한 가에 따라 달라질 수 있다. 충분치 못한 아키텍처의 구성은 데이터에 내재된 복잡한 비선형적인 구조를 찾아내는데 불충분한적합(Underfitting)이 될 수 있고, 반면 너무 많은 은닉계층이나 은닉마디는 과도적합(Overfitting)이 될 수 있다 (Haykin, 1994). 그러므로 적절한 아키텍처를 구성하기 위해서는 어느 정도 시행착오가 수반되어진다. 다음 <그림 2>는 다층신경망의 구조를 나타낸 것이다.

많은 경우에 있어서 몇 개의 은닉계층을 두느냐 보다는 몇 개의 은닉노드를 두느냐 즉 가중치의 수를 몇 개로 할 것인가가 더 중요하게 고려



<그림 2> 신경망 구조
(자료원: Jain & Nag, 1997)

되어야 할 사항이다. 일단 설정한 아키텍처에 대해 트레이닝 데이터를 이용하여 트레이닝하고, 그 결과를 테스트 데이터에 평가해보고 적당한 아키텍처를 수정하여 다시 트레이닝, 그 결과를 평가하고 앞 단계와 비교하는 다소 반복적인 과정을 거치면서 만족할 만한 아키텍처를 찾아나가게 된다. 접근방법으로는 우선 간단한 구성에서 시작하여 점차 복잡하게 구성해나가는 방법과 반대로 크고 복잡한 구성에서부터 점차 줄여나가는 방법이 있다. 그리고 신경망을 지원하는 툴에서 기본적으로 제공되는 아키텍처를 중심으로 은닉노드(혹은 은닉계층)를 증가 혹은 감소시키면서 그 예측력을 비교해보고 적당한 아키텍처를 구성해 나갈 수가 있다(Haykin, 1994).

본 연구에서는 이러한 신경망 기법 중 다층 퍼셉트론에서 가장 일반적인 방법인 역전파알고리즘(Backpropagation)을 사용하였다. 본 연구에서는 여러 가지 변수들로 구성된 복잡한 데이터 구조로부터 고객세분화 문제를 해결하기 위해 일반화 능력이 높은 신경망 알고리즘을 이용하였다. 이 알고리즘의 절차는 다음과 같은데 먼저, 다계층 인공신경망모형에서 처리 단위 j의 역할은 [식 4]로 표현할 수 있다.

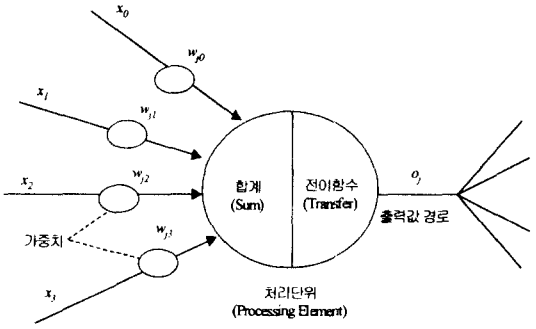
$$o_j = f(net_j) \qquad [식\ 4]$$

이때, o_j=j처리 단위의 출력값

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

$f(x)$ = 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 3>의 은닉층의 처리단위는 입력층의 연결된 각 처리단위들로부터 그들의 출력 값을 연결강도로 곱하여 받아들이고 이들의 합을 비선형전환하여 출력 층의 처리단위로 출력한다.



<그림 3> 처리단위 j의 계산 구조
(자료원: Jain & Nag, 1997)

이와 같은 입력과 출력이 각 층의 처리단위에서 이루어지고 최종적으로 출력 층에서 계산값이 구해진다. 이때 사용되는 것이 [식 5]와 같은 시그모이드 함수인데 이 전환함수는 [-∞, +∞]의 구간에서 나타나는 출력을 [0, 1] 구간으로 제약하기 위해 사용된다.

$$f(net_j) = \frac{1}{(1+e^{-net_j})} \qquad [식\ 5]$$

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 목표 값과의 오차를 구한다. 그 다음 이 오차가 최소화되는 방향으로 각 층에서의 연결강도를 조정하는 것이다. 본 연구에서 사용하고자하는 역전파 학습알고리듬에서는 입력과 연결강도를 이용해 구한 출력 값과 목표 값의 차이인 오차를 하위처리단위로 되돌려 보냄[1]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).

3.3 C 5.0

고객의 특성에 대한 예측을 위해 데이터마이닝 기법 중에서 널리 사용되는 방법중의 하나가 의사결정 나무 분석이다. 대표적인 의사결정나무 분석의 기법 중에 C 5.0이 있는데 본 연구에서는 이 방법을 적용하여 가계성예금 고객에 대한 특성을 도출하고 이를 바탕으로 고객의 유형을 세분화하고자 한다. C5.0은 Quinlan (1996)이 ID3에 이어 개발한 귀납적 학습방법의 하나이다. C5.0의 분석과정은 다음과 같은 과정을 통해서 도출이 가능하다 (Quinlan & Quinlan, 1997). S개의 집합 (Set)으로 구성된 사례에서 C_j 라는 속성이 포함되어 있으므로 이를 [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.

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times -\log_2\left(\frac{freq(C_j, S)}{|S|}\right) \quad bits \qquad [식\ 8]$$

훈련 사례 (Training Case)의 집합을 적용할 때는 사례에서 T라는 속성을 식별하는데 필요한 평균적인 정보의 양인 info(T)를 측정해야 한다. 이때 엔트로피는 [식 9]와 같다.

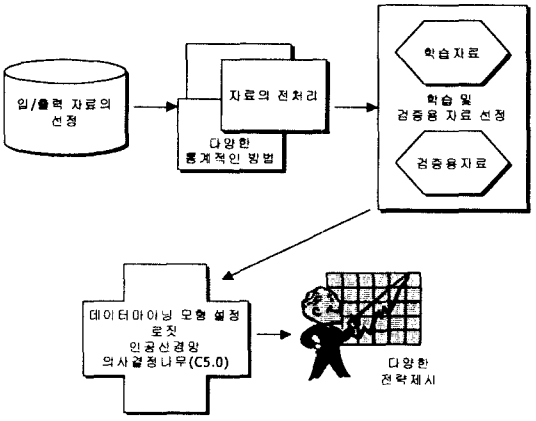
1) 이것이 역전파 인공신경망, 또는 back-propagation 알고리즘이라고 명명된 이유이다.

$$entropy = -\sum_{i=1}^{C} p_i \log p_i \qquad [식\ 9]$$

여기서 C는 목표변수의 범주의 수를 의미하며, P_i는 목표범주의 모비율을 말한다. 의사결정나무 분석의 여러 알고리즘에 대한 비교를 하면 다음 <표 2>와 같다.

4. 데이터마이닝 모형 설정

본 절에서는 전통적인 통계방법에서 고객의 구매여부를 분석하는데 많이 사용된 로짓분석을 적용하여 가계성 예금에 대한 고객의 유형을 세분화하고 이탈고객의 특성을 분석해본다. 다음으로는 인공신경망을 활용하여 이탈고객의 특성을 일반화하고 이를 검증해본다. 마지막으로 규칙 추출의 대표적인 메커니즘인 C5.0을 활용하여 이탈고객의 패턴을 분석하고 규칙을 도출한다. 이에 본 연구에서 이루어지는 금융회사의 이탈고객에 대한 분석단계를 도시하면 다음 <그림 4>와 같다.



<그림 4> 데이터마이닝 모형

4.1 입출력 자료(Data)의 선정

분석을 위하여 사용된 데이터는 1999년 10월부터 2000년 11월 동안의 1년 간의 자료를 분석대상으로 선정하였다. 현재 C은행 자료로써 본 자료는 고객정보와 예금정보로 나뉘어져 있다.

첫째, 고객정보란 은행에서 처음으로 통장을 개설할 때의 정보를 말한다. 이는 고객 ID, 거래개시일, 직업, 통장개수, 연령, 자동이체건수, 자사 신용카드 보유 개수, 예금월평잔, 종합통장 대

<표 2> 의사결정나무분석에 사용되는 알고리즘의 비교

기준 \ 알고리즘	CHAID	C5.0	CART	QUEST
입력변수에 의한 분리형식	다중분할	다중분할	이진분할	이진분할
연속형 목표변수	처리불가	처리가능	처리가능	처리가능
나무구조 생성시 오류분류 비용사용 여부	사용안함	사용	사용	사용안함
결측치 처리	하나의 범주로 처리	하나의 범주로 처리	결측치 대체	결측치대체
사정확률 사용여부	사용안함	사용안함	사용	사용

출 유무, 성별 등으로 구성되어 있다.

둘째, 예금정보란 가계성 예금에 대한 정보로써 신규일, 해지일, 만기일 금액(적금), 중도해지 여부, 예금금액 등으로 구성되어 있다.

이상의 고객정보와 예금정보를 모두 데이터마이닝에 사용할 수 없으므로 본 연구에서는 연구자의 경험과 가계성 예금에 유용하다고 생각되는 입력변수를 선정하였다. 본 연구에서 데이터마이닝에 사용한 입력변수는 아래의 <표 3>과 같다.

<표 3>에 제시되어 있는 입력자료에 대하여 구체적으로 살펴보자. "GEN"은 성별을 나타내고 남자는 1이며 여자는 0을 나타낸다. 이 변수는 명목척도이므로 분석의 정확성을 높이기 위하여 인공신경망과 로짓 분석을 할 때에는 0과 1로 카테고리화하여 분석을 실시하였다. "AGE"는 가계성 예금을 가입한 고객의 나이를 뜻한다. "TERM"은 가계성예금을 거래한 기간을 뜻하는 것으로 최초 가입일로부터 해지하거나 현재시점까지의 기간을 월로 구분하였다. "LOAN"은 종합대출 여부를 나타낸다. 대출을 받으면 1이고 받지 않으면 0으로 구분하였다. 이 변수 역시 명목척도이므로 인공신경망과 로짓분석에서는 더미화하여 분석을 실시하였다. "TRANS_NO"는 C은행에서의 자동이체건수를 나타내는 변수이다. "BOOK_NO"는 고객이 가지고 있는 가계성 예금통장의 개수이다. 고객이 한 거래은행과 여러 개의 거래계좌를 개설하는 경우도 많으므로 이를 입력변수로 선정하였다. "CARD"는 고객이 사용하는 월평균 카드금액을 나타낸다. "MON_AVER"는 가계성 예금의 월평균 잔액으로 여러 개의 가계성 예금을 가지고 있는 고객의 경우에는 이들의 평균값을 사용하였다. "INTEREST"는 가계성예금의 이자율로써 다수

<표 3> 자료의 구성 필드

변수명	설 명
GEN	성별을 나타내고 남자는 1이며 여자는 0으로 표현
AGE	가계성 예금을 가입한 고객의 나이
TERM	가계성예금을 거래한 기간을 뜻하는 것으로 최초 가입일로부터 해지하거나 현재시점까지의 기간을 월로 구분
LOAN	종합대출 여부로 대출을 받으면 1이고 받지 않으면 0으로 구분
TRANS_NO	C은행에서의 자동이체건수
BOOK_NO	고객이 가지고 있는 가계성 예금 통장의 개수
CARD	고객이 사용하는 월평균 카드금액
MON_AVER	가계성 예금의 월평균 잔액으로 여러 개의 가계성 예금을 가지고 있는 고객의 경우에는 이들의 평균값
INTEREST	가계성 예금의 이자율로써 다수의 통장의 경우 평균값
DEPOSIT	정기예금의 거래액

의 통장을 가지고 있는 고객의 경우에는 이들의 평균 이자율을 말한다. "DEPOSIT"은 정기예금의 거래액을 나타낸다. 가계성 예금에는 정기예금도 포함된다.

<표 3>에 제시된 입력자료가 통계적으로 유의한지를 검증하기 위하여 가계성 예금 유지고객과 해지고객을 대상으로 비율척도와 명목척도로 나누어서 명목척도에서는 카이테스트 검증을 비율척도에서는 T-Test 검증을 실시하였다. 분석결과는 <표 4>에 제시되어 있는데 검증결과 모든 자료가 유의한 것으로 분석되었다.

4.2 학습데이터 및 검증데이터의 선정

입력변수를 결정하였으면 데이터마이닝에 사용될 자료를 결정하는 과정이 이루어져야 한다. 본 연구에서 전체 관찰자료의 개수는 약 9,000개인데 이를 데이터마이닝에 적용하기 위해서는 학습용 자료와 검증용 자료를 구분하여야 한다. 다음 <표 5>에는 학습용 자료와 검증용 자료를 정리하였다.

<표 4> 입력자료의 통계적인 분석결과

(a) Chi-Square Test

변수명	유지고객		해지고객		전체	
	Chi-S	P-Value	Chi-S	P-Value	Chi-S	P-Value
LOAN	2191.5	0.000	2314.1	0.000	4505.1	0.000
GEND	395.88	0.000	39.5	0.000	786.3	0.000

(b) T-Test

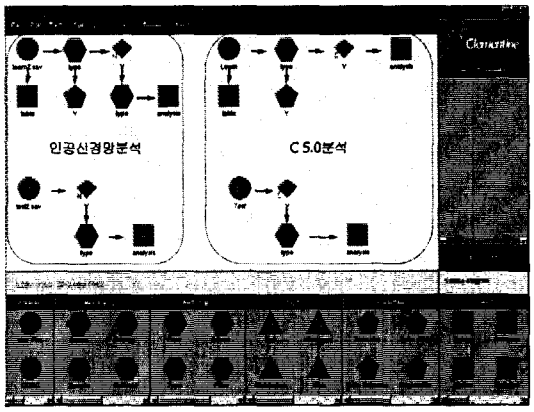
변수명	유지고객		해지고객		전체	
	T-Value	P-Value	T-Value	P-Value	T-Value	P-Value
AGE	495.206	0.000	529.415	0.000	723.196	0.000
TERM	226.204	0.000	301.848	0.000	343.717	0.000
TRANS_NO	69.160	0.000	69.370	0.000	95.888	0.000
BOOK_NO	1536.7	0.000	1435.2	0.000	1672.2	0.000
CARD	59.611	0.000	61.370	0.000	85.542	0.000
MON_AVER	43.996	0.000	31.806	0.000	54.084	0.000
DEPOSIT	78.62	0.000	84.45	0.000	95.52	0.000
INTEREST	844.935	0.000	347.757	0.000	655.390	0.000
Y	186.24	0.000	157.993	0.000	200.991	0.000

<표 5> 데이터마이닝에서 사용된 학습 및 검증용 자료

구분	표본수		비율
	이탈(0)	유지(1)	
학습자료	3,045	3,044	67%
검증용자료	1,502	1,502	33%
합 계	4,547	4,546	100%

5. 실험 및 결과분석

본 연구에서는 데이터마이닝 툴로써는 클레멘타인(Clementine 5.2)을 활용하여 데이터마이닝을 실시하였다. 클레멘타인은 데이터 마이닝에 용이한 기술을 통합한 팩키지로써 현재 산업계에서 많이 활용하고 있는 프로그램으로 개체지향적이며 초보자가 사용하기 용이하다는 장점이 있다. 따라서 본 연구에서는 클레멘타인을 활용하여 인공신경망과 C5.0 분석을 실시하였다. 또한 SPSS10.0을 활용하여 입력자료의 기초적인 통계분석과 로짓 분석을 실시하였다. 아래의 <그림 5>에는 이러한 분석과정이 제시되어 있다.



<그림 5> 분석 과정

5.1 로짓 분석결과

정기성 예금의 가입여부를 전통적인 통계 분석방법 중에서 제품에 대한 고객의 반응여부가 구입 혹은 비구입과 같이 두 가지 유형으로 나뉘어질 때 사용하는 로짓 모델을 활용하여 분석을 실시하였다. 로짓 모델의 경우 입력변수가 명목척도로 구분된 경우에는 분석이 불가능하다. 이에 대출여부변수와 성별 변수를 각 2개씩 더미화하여 분석을 실시하였다. 따라서 분석에 사용된 독립변수는 12개이며, 종속변수는 이탈여부 1개였다.

로짓 분석에서 사용된 방법은 일반적인 분석방법인 변수진입방법(Enter)을 사용하여 분석을 실시하였으며 분석결과가 다음 <표 6>에 제시되어 있다. 분석결과 입력에 사용된 변수 중에서 성별과 관련된 "GEN1,2" 두 변수를 제외하고는 모든 변수가 유의한 것으로 통계적인 방법으로 분석되었다. 가계성 예금의 분석을 위하여 사용된 자료가 통계적인 가정이 강한 분석방법중의 하나인 로짓 분석결과 모두 유용하다는 것은 앞으로 추가적으로 분석을 실시할 인공신경망과 의사결정나무 분석(C5.0)에서도 그 유용성이 높을 것이라고 예상할 수 있다. 분석결과를 로짓 모델로 제시하면 아래와 같다.

분석결과 검증용 자료의 적중률이 77.5%로 학습자료보다는 예측율이 조금 떨어지는 것을 알 수 있다. 이는 로짓 모델이 강한 통계적인 가정 때문에 테스트 자료에 영향을 받기 때문으로 분석할 수 있다. 또한 1종 및 2종 오류가 각각 12%, 10.5%로 1종 오류가 조금 높게 분석되었다.

<표 6> 로짓 분석결과

(a) 로짓 분석시 유의한 변수

	B	S.E.	Wald	Sig.	Exp(B)
GEN1	.021	.053	1.126	0.23*	1.004
GEN2	.015	.064	0.624	0.430*	1.015
LOAN1	.088	.015	34.373	.000	1.092
LOAN2	.088	.015	34.373	.000	1.092
AGE	.008	.004	4.117	.042	1.008
TERM	-.024	.013	3.647	.046	.996
BOOK_NO	.088	.015	34.373	.000	1.092
TRANS_NO	.243	.046	27.780	.001	1.275
CARD	.000	.000	374.806	.000	1.000
MON_AVER	.000	.000	33.570	.000	1.000
DEPOSIT	.000	.000	13.550	.000	1.000
INTERES	.578	.021	751.948	.000	1.782
Constant	-5.382	.223	580.846	.000	.005

* 유의수준 5%에서 유의하지 않음

(b) 로짓분석의 예측율

			예측치		1종 오류	2종 오류	예측율
			0(해지)	1(유지)			
학습 자료	실제값	0(해지)	2311	732	12%	7.6%	80.4%
		1(유지)	461	2585			
검증용 자료		0(해지)	1189	315	12%	10.5%	77.5%
		1(유지)	360	1140			

5.2 인공신경망 분석결과

본 연구에서는 인공신경망으로 가계성예금 고객의 해지패턴을 분석하기 위하여 다계층 퍼셉트론 인공신경망을 사용하였다. 또한, 인공신경망이 읽을 수 있도록 명목척도는 더미(Dummy)화 시켰다. 이렇게 해서 구성된 인공신경망은 입력층이 12개이고 출력 층이 1개인 인공신경망 모형을 구축하였다. 이때, 은닉층 노드의 개수는 반복실험을 통해서 최상의 개수를 결정하고자 하였는데 실험결과 은닉 층이 6개 일 때 성과가 가장 좋은 것으로 나타났다. 학습방법은 시그모이드 함수를 이용한 역전파 학습 알고리즘을 이용하였다. 학습자료의 선정은 총 8,993개의 자료 중에서

6,089개의 자료를 학습자료로 하고 나머지 3,004개의 자료를 검증용 자료로 분류하였다. 학습자료와 검증용 자료에서는 가계성예금의 유지고객과 해지고객이 동일한 비율로 구성되어 있다. 가계성예금의 해지여부에 영향을 많이 미치는 입력변수를 분석하였는데 카드 월평균 사용금액(CARD)이 가장 많은 영향을 미쳤으며, 다음으로 통장의 개수(BOOK_NO), 이자율(INTEREST), 월평균 가계성예금잔액(MON_AVER) 등의 순서로 분석되었다. <표 7>에는 인공신경망에 의한 실험결과가 제시되어 있다.

인공신경망 분석결과 학습자료의 경우 예측율이 88.7% 였으며, 검증자료의 경우에는 예측율이 88.9%로써 유사하게 분석되었다. 1종 오류 및 2종 오류의 경우 학습자료 검증용 자료 모두 1종 오류의 값이 2종 오류보다 적었으며 특히 검증용자료의 1종 오류가 더 좋은 것으로 분석되었다. 일반적으로 학습자료의 1종 오류가 더 적게 분석되지만 본 연구에서는 검증용 자료 1종 오류가 더 적게 분석되었다. 이는 본 자료의 학습이 잘되어 일반화가 잘 이루어졌으므로 학습자료와 검증용 자료에 상관없이 가계성예금의 고객들의 해지 및 유지를 잘 예측하는 것으로 분석된다.

5.3 C 5.0 분석결과

C5.0을 사용하여 가계성예금의 해지여부를 분석한 결과가 다음 <표 8>에 제시되어 있다. 학습자료의 경우 예측율이 93.9%이며, 검증용 자료의 경우 94.2%로써 매우 높게 나타났다. 1종 오류와 2종 오류를 살펴보면 전체적으로 1종 오류가 2종 오류가 약간 높게 나타났다. 하지만 전체적인 모형의 예측율이 높아서 1종 및 2종 오류의 차이가

<표 7> 인공신경망 분석의 예측율

			예측치		1종 오류	2종 오류	예측율
			0(해지)	1(유지)			
학습자료	실제값	0(해지)	2552	341	5.6%	5.7%	88.7%
		1(유지)	348	2848			
검증용자료		0(해지)	1269	135	4.5%	6.6%	88.9%
		1(유지)	199	1401			

<표 8> C5.0의 예측율

			예측치		1종 오류	2종 오류	예측율
			0(해지)	1(유지)			
학습자료	실제값	0(해지)	2743	200	3.3%	2.8%	93.9%
		1(유지)	173	2973			
검증용자료		0(해지)	1395	99	3.3%	2.5%	94.2%
		1(유지)	74	1436			

별다른 의미를 가지지 못할 것으로 분석된다.

C5.0 분석결과 가계성 예금을 유지하고 있는 고객에 대해서는 8개의 규칙을 도출할 수 있었으며 해지하고 이탈한 고객에 대하서는 10개의 규칙을 도출할 수 있었다. 아래의 <표 9>에는 분석결과가 요약되어 있다.

분석결과 통장의 개수(BOOK_NO), 자동이체 건수(TRANS_NO), 일정액 이상의 카드사용액(CARD) 및 월평균 잔액(MON_AVER) 등에 많은 영향을 받고 있는 것으로 분석되었다. 이는 인공신경망 분석의 결과시 고객의 유지 및 해지에 가장 많은 영향을 미치는 변수로 카드 월평균 사용금액(CARD), 통장의 개수(BOOK_NO), 이자율(INTEREST), 월평균 가계성예금잔액(MON_AVER) 등의 순서와 유사함을 알 수 있다. 이는 인공신경망과 C5.0의 분석결과가 서로 비슷하다는 뜻이다. 따라서 이러한 결과를 중심으로 전략을 제시하고자 한다.

5.4 로짓, 인공신경망, C 5.0 결과 비교

로짓 분석, 인공신경망, C5.0 분석의 결과를 비교하여 다음 <표 10>에 요약·정리하였다. 구체적으로 살펴보면 예측율에서는 학습자료 검증용 자료 모두에서 C5.0>인공신경망>로짓 순으로 분석되었으며, 1종 오류에서는 로짓>인공신경망

<표 9> C5.0을 통한 유지 및 해지 고객의 규칙

유지고객 규칙		해지고객규칙	
Rule #1 for 1:(386, 0.951)* if TRANS_NO > 1 and CARD > 479195 and LOAN > 0 and DEPOSIT > 2229005 and INTERES > 7 and INTERES <= 7.9 then -> 1 Rule #2 for 1:(40, 0.929) if CARD > 479195 and MON_AVER > 1995960 and MON_AVER <= 2827120 and DEPOSIT <= 2700984 and INTERES > 5.8 then -> 1 Rule #3 for 1: (209, 0.919) if BOOK_NO > 3 and TRANS_NO > 0 and CARD > 479195 and MON_AVER > 1995960 and MON_AVER <= 15986000 and LOAN <= 0 and DEPOSIT <= 11200000 and INTERES > 3.5 then -> 1 Rule #4 for 1: (2933, 0.911) if BOOK_NO > 1 and TRANS_NO <= 7 and CARD > 479195 and MON_AVER > 2827120 and INTERES > 5.8 then -> 1	Rule #5 for 1: (8, 0.9) if CARD > 479195 and DEPOSIT <= 1243334 then -> 1 Rule #6 for 1: (923, 0.896) if BOOK_NO > 1 and CARD > 479195 and MON_AVER > 1995960 and LOAN <= 0 and DEPOSIT <= 11200000 and INTERES > 3.5 then -> 1 Rule #7 for 1: (1914, 0.856) if BOOK_NO > 1 and CARD > 479195 and MON_AVER > 12017600 then -> 1 Rule #8 for 1: (10, 0.833) if TERM <= 9.88 and TRANS_NO <= 0 and CARD > 479195 and MON_AVER > 1995960 and LOAN > 0 and DEPOSIT <= 11200000 and INTERES > 3.5 and INTERES <= 5.8 then -> 1	Rule #1 for 0:(2136, 1.0) if CARD <= 479195 then -> 0 Rule #2 for 0:(372, 0.997) if BOOK_NO <= 1 then -> 0 Rule #3 for 0: (129, 0.992) if BOOK_NO <= 2 and TRANS_NO > 0 and MON_AVER <= 10091100 and INTERES <= 7.3 then -> 0 Rule #4 for 0: (442, 0.991) if MON_AVER <= 10091100 and DEPOSIT > 10324100.0 and INTERES <= 7.3 then -> 0 Rule #5 for 0: (96, 0.99) if BOOK_NO <= 3 and TRANS_NO > 0 and MON_AVER <= 15986000 and INTERES <= 4.5 then -> 0	Rule #6 for 0:(95, 0.99) if TERM > 7.95 and TRANS_NO > 0 and MON_AVER <= 10091100 and LOAN > 0 and DEPOSIT > 6200000 and INTERES <= 7.3 then -> 0 Rule #7 for 0: (781, 0.985) if AGE <= 40 and MON_AVER <= 2827120 and DEPOSIT > 2700984 then -> 0 Rule #8 for 0: (52, 0.981) if TERM > 5.99 and MON_AVER > 1995960 and MON_AVER <= 2827120 and DEPOSIT > 2700984 and INTERES > 5.8 then -> 0 Rule #9 for 0: (962, 0.98) if MON_AVER <= 1995960 and DEPOSIT > 1243334 then -> 0 Rule #10 for 0:(172, 0.977) if TRANS_NO <= 0 and MON_AVER <= 15986000 and LOAN <= 0 and INTERES <= 4.5 then -> 0

* 해당 건수와 이때의 신뢰도를 나타낸다.

<표 10> 로짓, 인공신경망, C5.0의 결과비교

방법론 / 데이터종류	로짓			인공신경망			C 5.0		
	예측율	1종오류	2종오류	예측율	1종오류	2종오류	예측율	1종오류	2종오류
학습데이터	80.4%	12%	7.6%	88.7%	5.6%	5.7%	93.9%	3.3%	2.8%
검증데이터	77.5%	12%	10.5%	88.9%	4.5%	6.6%	94.2%	3.3%	2.5%

>C5.0 순으로 분석되었다. 결론적으로 예측력에서는 C5.0이 가장 높은 것으로 분석되었으며, 1종 오류의 경우에도 C5.0이 우수한 것으로 분석되었다.

결론적으로 C5.0의 분석결과 이 모형의 예측율이 매우 높았으므로 다음 장에서는 이때 도출된 규칙을 중심으로 가계성예금 고객의 특성을 분류하고 이들에게 적합한 마케팅 전략을 제시하고자 한다.

6. 데이터마이닝을 통한 고객관계 전략제시

원래 이탈고객 및 유지고객의 유형을 세분화하고 이들의 특징을 분석하는 연구는 통신산업에서 기존의 통신수단을 해지하고 다른 업체의 통신수단으로 이동하는 고객들의 특징을 분석하기 위하여 연구가 시작되었다. 이에 본 연구에서는 이러한 방법론을 활용하여 가계성 예금의 분야에 적용시켜 보았다는데에 그 공헌점이 있다고 하겠다. 특히 IMF이후로 우리나라의 은행들은 현재 큰 구조조정을 맞이하고 있으며 이 속에서 살아남기 위하여 기존의 고객의 유형을 분석하고 이를 마케팅 전략에 활용하는 본 연구는 매우 시의 적절한 연구라고 사료된다. 이에 본 연구에서는 앞서 다양한 데이터마이닝을 이용한 분석한 결과를 바탕으로 적절한 마케팅 전략을 제시하고자 한다.

첫째, 충성 고객 및 이탈고객을 분류하는 고객세분화 전략이다. 본 연구의 분석결과 유지고객과 이탈고객을 분류할 수 있었으며 이들의 특성 또한 도출할 수 있었다. 따라서, 새로운 고객을 선정할 때의 시장세분화 전략과는 달리 기존의 고객에 대해서도 차별화된 유지전략이 필요하다. 이를 위하여 앞서 분석된 결과(가계성예금의 월평균 잔액, 자동이체건수, 정기예금 금액 등)를 중심으로 고객을 몇 개의 군집화한 다음 이들 각 고객집단별로 특성을 분석하고 이들을 대상으로 하는 전략의 수립이 요구된다.

둘째, 새로운 상품권장전략이다. 예를 들어, 본 연구결과 거래하는 가계성 예금의 통장수가 1개 이하인 고객들은 현재 거래하고 있는 은행에 대한 충성도가 낮은 것으로 분석되었다. 따라서 새로운 금융상품이 나왔을 때 통장의 수가 1개인 고객들에게 집중적인 마케팅 전략이 필요하다. 또한 금융상품의 이자율이 일정액이 이상이 되는 고객이 이탈률이 낮은 것으로 분석되었다. 따라서 충성고객에게는 새로운 금융상품이 나왔을 때 이자율 부분을 중심으로 홍보를 실시하여 계속적으로 현재 거래하고 있는 고객들이 새로운 금융

상품에 대한 가입동기를 유발시키는 전략이 필요할 것이다. 정기예금 금액이 높은 고객이 충성도가 높은 것으로 분석되었다. 따라서 정기예금이 만기에 다다른 고객들에 대하여 새로운 금융상품을 소개함으로써 기존의 고객이 가계성 예금을 만기 후에 재 예치할 수 있는 전략이 필요하다. 실제적으로 정기예금이 만기가 된 후에 새롭게 계좌를 개설하거나 만기 후 수령되는 금액을 자신의 다른 가계성 예금에 재 예치하는 고객은 10명중 4명이 불가하다고 한다. 따라서 정기예금 만기일에 해당하는 고객들을 특별히 관리하는 전략이 필요하다.

셋째, 해당 은행의 카드 사용의 권장전략이다. 본 연구의 분석결과 월평균 카드 사용액이 높은 고객이 가계성 예금을 이탈하지 않고 계속적으로 거래를 하는 것으로 분석되었다. 이것은 해당 은행의 카드 사용액이 높을수록 이 은행을 주 거래은행으로 사용하고 있다는 뜻이다. 따라서 고객들이 더 많은 제품을 저렴한 비용에 구매하도록 유인하기 위하여 백화점 등과 전략적 제휴를 통하여 3개월 무이자로 제품을 구매할 수 있게 하는 상품전략이 필요하다. 또한 자동이체건수가 적은 고객들이 쉽게 가계성 예금을 해지하고 다른 은행으로 옮기는 것으로 분석되었다. 따라서 카드사용액에서부터 각종 공과금까지 다양한 자동이체를 권장하도록 하여 거래은행에 대한 지지도를 높일 필요가 있다.

7. 결론 및 향후 연구방향

본 연구에서는 가계성 예금 고객을 대상으로 전통적인 통계적인 분석방법이며 마케팅에서 가장 많이 활용하여 로짓 모형과 전통적인 데이터마이닝 기법으로 알려진 인공신경망과 C5.0 분석방법을 활용하여 가계성 예금 고객의 유형을 분석하고 이들의 특징을 도출하였다. 본 연구결과를 간략히 정리하면 다음과 같다.

첫째, 데이터마이닝 기법을 이용하여 가계성 예금의 고객의 유형을 분류하고 이들의 이용행태에 따른 전략을 제시하였다. 이미 서론에서도 언급한 바와 같이 국내외의 가계성 예금에 대한 연구는 주로 설문방법에 의한 연구들이 많았으며 이것도 주로 가계성 예금을 현재 가입하고 있는 고객을 대상으로만 이루어졌다. 또한 은행들의 구조조정으로 금융업에서는 충성도가 높은 고객을 대상으로 적극적인 마케팅 전략을 제시하는 일이 그 무엇보다도 필요하게 되었다. 본 연구의 경우 데이터마이닝 기법을 이용하여 80% 이상 가계성 예금의 해지여부를 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 통계적 기법인 로짓 모형의 결과를 유기적으로 해석에 이용함으로써 기법이 가지고 있는 특성을 이용하여 보완이 가능하다. 즉, 인공신경망은 견고하고 학습성이 뛰어나지만 해석하기 어려운 점을 로짓 모형이나 C5.0이 보완한다는 개념이다.

셋째, 가계성예금의 고객에 대한 마케팅 전략의 제시이다. 본 연구에서는 가계성예금 고객들의 행위분석을 실제 거래자료를 이용하였으며 이를 통하여 가계성예금 유지자와 이탈자를 규정짓

는 영향변수를 도출하였다. 이러한 변수의 도출은 가계성예금 고객들의 특성을 파악한 기존의 고객을 충성도가 높은 고객과 충성도가 낮은 고객으로 세분화할 수 있으므로 이들에 대한 차별화 된 마케팅 전략을 수립할 수 있는 근거를 제공할 수 있을 것이다.

하지만 본 연구에서는 고객의 유형을 유지 및 이탈자로만 한정하였을 뿐 이를 다양한 집단으로 분류하고 이들의 특성을 반영하지 못했다는 한계점을 가지고 있다. 따라서 향후 표본의 수를 조금 더 늘려서 이들을 몇 개의 집단으로 분류하고 이들의 특성을 분석하는 연구가 필요하다고 하겠다.

참고문헌

안광호, 임영균, 이산적 확률 선택 모형을 이용한 경쟁적 시장구조분석에 관한 연구, *소비자학 연구*, 7권, 1호(1996), 75-90.

안광호, 채서일, Multinomial 로짓 모델을 이용한 점포선택행위에 대한 실증 연구, *경영학 연구*, 22권, 2호(1993), 101-120.

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

김정수, *통합데이터베이스 마케팅 시스템*. 범우사, 1997.

박찬욱, *데이터베이스 마케팅*. 연암사, 1996.

송현수, *통합 DBMS*. 새로운 제안, 1998.

채서일, *마케팅조사론*. 3판, 학현사, 2000.

하영원, "금융시장의 세분화에 관한 연구-가계성예금의 고객세분화를 중심으로," 서강경영논총, Vol. 7(1996), 519-17.

한충민, 이한구, 외국산 수입 제품의 소비자 채택행위에 관한 실증적 연구 - 수입담배를 중심으로, *소비자학 연구*, 9권, 2호(1998), 79-89.

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.

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

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

Bishop, C., *Neural Network for Pattern Recognition*. Oxford Press, 1995.

Fayyad, U., G. Piatetsky-Shapiro, 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.

Haykin, S., *Neural Network*. Prentice Hall, 1994.

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

Jain, B. A. and 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.

Marsh, J., *Managing Financial Services Marketing*. London: Pitman Publishing, 1998.

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

Quinlan, R., "Induction of Decision Trees", *Machine Learning*, Vol. 1(1996), 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

Analysis of Defection Customer Using Customer Segmentation on Bank
-Focusing on Personal Deposit-

Kun-chang, Lee*
Soon-jae, Kwon*
Kyung-shik Shin**

This paper is aimed at proposing a data mining-driven analysis to manage the customer defection rate in the bank. After 1997 IMF crisis, Korean banks were suffering from hard-pressed restructuring. At the heart of such restructuring efforts, there was the need to manage the customer more effectively than ever. So far, many banks in Korea used to a poor management of customers without any highly-skillful techniques. In line with this argument, we propose several data mining techniques to determine more effective technique for managing customer defection. We applied three data mining techniques such as logit model, neural network, and C5.0. Experiment data were collected from personal deposit account data of a specific bank in Korea. After experiments, we found that C5.0 showed more robust performance compared to other two techniques. On the basis of those experiment results, we proposed customer defection management policy.

* School of Business Administration, Sungkyunkwan Univ.
** College of Business Administration, Ewha Womans University