데이터마이닝을 이용한 이탈확률에 기반한 고객 세분화

홍 태 호[a], 전 성 용[b]

[a] 부산대학교 경영학부
609-735, 부산광역시 금정구 장전동 산 30
Tel: +82-51-510-2531, E-Mail: hongth@pusan.ac.kr

[b] 부산대학교 경영학과
609-735, 부산광역시 금정구 장전동 산 30
Tel: +82-51-510-2531, E-Mail: weblogic@pusan.ac.kr

Abstract

현재의 이동통신시장은 시장의 포화상태로 인해 신규 고객의 확보보다는 기존 고객의 유지에 마케팅 활동을 강화하고 있다. 본 연구에서는 이탈고객관리(churn management)를 위한 방안으로 데이터마이닝 기법에 기반하여 고객을 등급별로 세분화하였다. 이동통신 고객데이터를 활용하여 로짓모형, 인공신경망, SVM 등을 이탈고객 예측모형을 개발하였고, 각 모형별 성과를 통계적으로 비교하였다. 이탈고객 예측모형을 통해 고객의 이탈가능성을 등급화하여 등급별 이탈확률과 점유율, 적중률을 산출하였다. 제안된 고객등급화 방법을 통해 이동통신사들은 고객의 이탈확률에 따른 차별화된 마케팅 전략을 수행할 수 있을 것으로 기대된다.

Keywords: 이탈고객관리, 고객세분화, 데이터마이닝, 인공신경망, Support vector machine

I. 서 론

국내 이동통신시장은 한국이동통신이 1984년 차량용으로 이동전화 서비스를 실시한 이후 급속도로 발전하여 2003년 말 현재 국내인구 중 이동통신 서비스를 사용하는 고객의 수는 약 3,350만명에 이르고 있다. 하지만 최근 이동통신사는 시장의 포화상태로 인해 성장이 둔화되고 있으며 번호이동성 제도의 시행으로 인해서 고객의 이탈 또한 매우 빈번하게 일어나고 있다. 통신시장에서의 고객이탈이란 자발적 또는 비자발적으로 고객이 현재의 통신서비스를 중단하는 것을 의미하며, 2000년까지의 미국의 무선통신에 있어서의 고객이탈(Churn)은 매해 150%의 증가율을 보이고 있다(Swartz, 2001).

각 이동통신사들은 꾸준히 증가하고 있는 고객의 이탈을 방지하기 위해서 다양한 서비스의 개발과 마케팅에 많은 노력을 기울이고 있다.

이탈고객을 관리하고 마케팅 전략을 수립하기 위한 방안으로 데이터마이닝 기법을 활용한 이탈고객관리(Churn management)가 연구되어 왔다. (김충영 등, 2002; 윤충한 등, 2002; Wei & Chiu, 2002). 정보기술의 발달로 데이터의 축적 및 처리능력이 과거와는 비교할 수 없을 정도로 발전하였으며, 이에 따라 데이터마이닝의 활용범위가 더욱 넓어지고 있다. 이탈고객관리에 적용되는 데이터마이닝 기법으로는 로짓모형, 프로빗과 같은 통계적 기법과 인공신경망(Artificial Neural Networks), 의사결정나무 등과 같은 인공지능기법 등이 있다. 이중 인공신경망(ANN)의 성과가 가장 우수한 것으로 알려져 있으나, 최근에 Support Vector Machine(SVM)이 부도예측, 신용등급분석, 시계열예측, 보험사기적발 등의 분야에서 매우 우수한 성과를 내는 것으로 보고되었다. SVM은 명확한 이론적 근거에 기반을 두기 때문에 결과해석이 용이하고 적은 양의 학습자료만으로도 신속하게 학습을 수행할 수 있다는 장점을 갖고 있기 때문에 이동통신사의 고객이탈 예측에 새롭게 적용할 필요가 있다.

또한 기존의 연구들은 이탈고객의 이탈여부만을 예측하는 연구가 주였으나, 이동통신사에서 이탈고객에 대한 세밀한

마케팅을 위해서는 고객세분화를 통한 차별화된 마케팅 전략이 필요하다. 본 연구에서는 이동통신 고객에 대한 이탈고객예측모형을 SVM과 기존의 데이터마이닝 기법인 로짓모형과 인공신경망을 이용하여 개발하고자 한다. 또한 이를 토대로 이탈고객의 수에 따른 점유율을 등급화에 적용시켜 실제 이동통신사들이 각 등급에 맞는 마케팅 전략을 수립하는 데 도움을 주고자 한다.

Ⅱ. 이론적 배경

2.1 Data Mining

데이터마이닝은 데이터로부터 패턴이나 모형을 추출하기 위해 구체적인 알고리즘을 응용하는 과정이다(Fayyad et al., 1996). 다르게 말을 하자면 데이터마이닝은 대량의 데이터로부터 지식을 추출하는 것, 또는 데이터베이스나 데이터웨어하우스 또는 그 밖의 다른 정보 저장소들에 저장되어 있는 대량의 데이터로부터 흥미로운 지식을 발견하는 과정이라고 할 수 있다. 데이터마이닝에 적용되는 기법은 통계적 기법에서 기계학습까지 다양하며, 대용량의 데이터로부터 의미 있는 패턴을 추출하기위한 데이터마이닝 작업은 종속성 분석(Dependency analysis), 분류(Classification), 개념 기술(Concept description), 일탈 탐지(Deviation detection), 데이터 시각화(Data visualization) 등으로 나누어 진다(Shaw et al., 2001). 데이터 시각화는 주로 다른 데이터마이닝 작업을 지원하는데 사용되고, 다른 데이터마이닝 작업은 추출된 지식의 유형에 따라 구분된다.

첫째, 종속성 분석 작업은 주로 최소한도의 신뢰도(Confidence)를 갖는 개체 군간의 연관성을 찾는 것이다. 장바구니 분석이라고도 불리며, 구매한 다른 상품간의 관계를 파악하는 데 사용된다. 둘째, 분류작업은 주로 고객들을 사전에 정의된 여러 등급으로 나누는데 사용된다. 예를 들면, 쇼핑몰에서는 과거의 구매경력 등을 이용하여 고객을 분류할 수 있다. 셋째, 개념기술은 그 분야의 지식과 데이터베이스를 이용하여 고객을 그룹화하는 기법이다. 개념 기술은 마케팅과 고객 지식의 요약, 판별, 비교 등에 사용될 수 있다. 요약을 통해서 마케터는 고객을 직업, 수입, 구매 패턴과 유형 등에 따라 분류하여 고객 프로파일을 생성할 수 있다. 판별은 다른 자료와 구분할 수 있는 특성을 기술하는 것이며, 비교는 다른 자료들과 비교와 분석을 통해서 그룹을 설명하는 것이다. 넷째, 일탈 탐지는 예외적이거나 변화를 발견하는데 유용하다. 예를 들면, 어느 고객이 같은 그룹의 고객들과 비교하여 평균에서 매우 차이가 있는지를 규명하는 것이다. 카드사는 카드사용이 갑자기 증가할 때, 부정카드사용이 있는지를 확인 조사할 수 있다. 마지막으로, 시각화는 복잡한 고객 데이터의 패턴을 볼 수 있게 해준다.

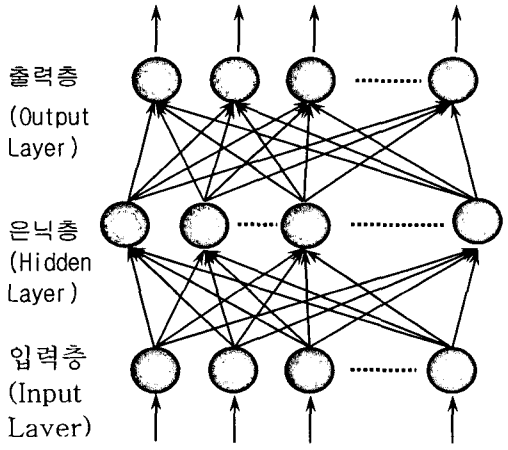
통신시장에서 데이터마이닝을 적용할 수 있는 분야는 여러 가지가 있지만 시장변화를 관리하고 고객의 수요를 파악하기 위한 장바구니 분석과 효과적인 고객유지를 위한 고객행태 분석을 주요 작업으로 들 수 있다. 통신 서비스 분야는 고객 유지비용에 비해 신규고객 창출비용이 매우 높기 때문에 데이터마이닝을 적용하여 해지 가능성이 높은 고객을 사전에 발견하고 어떤 요인에 의해 고객의 해지가 발생하는지를 알 수 있다면 고객의 이탈에 적절히 대응할 수 있을 것이다. 본 논문에서는 고객의 이탈을 알기 위한 일탈탐지 작업을 위한 데이터마이닝 기법으로 통계기법과 인공신경망, SVM을 사용한다. 다음은 인공신경망과 SVM에 대한 내용이다.

(1) 인공신경망

일반적인 인공신경망은 다층퍼셉트론(Multi layer perceptron)이라 불리우며, 다층퍼셉트론은 입력층과 출력층 사이에 하나 이상의 중간층이 존재하는 신경망을 지칭하는 것이다. 이 때, 입력층과 출력층 사이의 중간층을 은닉층(hidden layer)이라 하며 network는 입력층, 은닉층, 출력층으로 연결되어 있다. 다층 퍼셉트론에서의 가중치는 지속적으로 전체 신경망이 만족할 만한 목표에 도달할 때까지 변하게 된다. 즉 인공신경망을 통해 계산된 출력값과 목표출력값(output)을 비교하여 그 차이(오차합수)를 최소화시킬 수 있도록 지속적으로 가중치를 조정하는 것이다.

이러한 신경망의 가중치의 조절은 역전파 알고리즘에 의한 학습과정을 통해 이루어진다. 역전파 학습 알고리즘의 기본 원리는 다음과 같다. 입력층의 각 유니트에 입력패턴을 주면, 이 신호는 각 유니트에서

변환되어 중간층에 전달되고 최후에 출력층에서 신호를 출력하게 된다. 이 출력값과 기대값을 비교하여 차이를 줄여 나가는 방향으로 연결강도를 조절하고, 상위층에서 역전파하여 하위층에서는 이를 근거로 다시 자기층의 연결강도를 조정해 나가게 된다. 인공신경망에 대한 자세한 내용은 Rumelhart & McClelland(1986)을 참고한다.

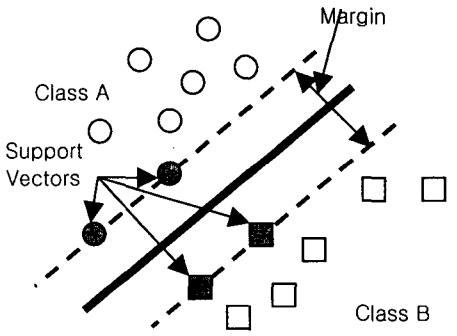


<그림 1> 인공신경망 모형

(2) SVM

SVM은 Vapnik(1995)에 의해 제안된 통계적 학습방법으로 입력공간과 관련된 비선형문제를 고차원의 특징공간(Feature Space)의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 쉽다고 알려져 있다(Hearst et al., 1998). 특히 SVM이 최근 각광을 받는 이유는 대부분의 학습 알고리즘은 경험적 위험최소화 기법인 ERM(Empirical Risk Minimization)에 기반하는 모형인데 비하여, SVM은 구조적으로 오분류율을 최소화시키는 SRM(Structural Risk Minimization)에 기반하기 때문에 일반화하기가 용이하다고 할 수 있다(Tay and Cao, 2001).

SVM은 분류 문제에 있어 훈련 데이터들을 서로 다른 두 개의 클래스로 분류할 때 분류의 기준이 되는 분리 경계면(hyper plane)을 학습알고리즘을 통해 찾으며 클래스를 구분하는 최적의 분리 경계면(maximum margin hyperplane)을 구하기 위해 분리 경계면과 가장 인접한 점(support vector)과의 거리를 최대화한다. 만약 SVM이 2개의 클래스로 나눌 수 없는 경우에는 커널함수를 이용하여 입력 자료를 고차원의 특징공간으로 사상시킨다. 고차원 공간에서는 선형 분리를 가능한 분리경계면을 생성하는 것이 가능하다. 따라서 SVM에서는 커널함수가 매우 중요한 역할을 한다.



<

그림 2> SVM에 의한 두 개의 클래스 분리

SVM 의 알고리즘을 살펴보면 분류문제에 있어서 학습 데이터를 이용하여 함수 $f : R^n \rightarrow \{\pm 1\}$을 추정하도록 한다. 클래스 A 는 x ∈ A, y=1 로, 클래스 B 는 x ∈ B, y=-1 로 표시한다. 즉, $(x_i, y_i) \in R^n * \{\pm 1\}$이 된다.

학습 데이터가 선형으로 분리가능하면 다음과 같이 나타낼 수 있다.

$$w^T x + b \geq +1, \forall x \in A \quad (1)$$

$$w^T x + b \leq -1, \forall x \in B \quad (2)$$

단 w 는 가중치 벡터이고, b 는 바이어스이다. 식 (1)과 (2)를 합치면 아래의 식을 도출할 수 있다.

$$y(w^T x + b) \geq 1, \forall x \in A \cup B \quad (3)$$

최대 마진분류자는 최대마진 경계면을 가지고 데이터를 최적화한다. 즉, $\frac{1}{2\|w\|^2}$을 최소화하고 식 (3)을 제약조건으로 하는 최적화 문제가 된다.

$$\underset{w,b}{Min} \frac{1}{2} w^T w \quad s.t \; y(w^T x + b) \geq 1 \quad (4)$$

위의 문제를 풀기 위해서는 이차계획 문제를 풀면 되고 Karush-Kuhn-Tuker 조건을 사용하면 된다. 이에 대한 자세한 풀이는 Huang et.al(2005)를 참고한다.

최종적으로 최적 의사결정 분리경계면 $f(x, a, b)$는 아래와 같다.

$$f(x,a,b) = \sum y_i \cdot a_i (x_i, x) + b \quad (5)$$

여기서 b와 a_i는 분리경계면을 결정하는 파라미터이며, x 는 학습용 데이터, x_i 는 서포트벡터(최대마진 분리경계면에 가장 근접한 학습 데이터)를 나타낸다.

학습 데이터가 비선형인 경우에는 입력변수를 고차원의 특징공간으로 이동시킴으로서 비선형 문제를 선형문제로 근사시킬 수 있다. 비선형 분류문제에서는 식(5)를 다음과 같이 나타낼 수 있다.

$$f(x,a,b) = \sum y_i \cdot a_i \cdot K(x_i, x) + b \quad (6)$$

식 (6)에서 $K(x_i, x)$ 는 커널함수라고 정의한다. 커널함수는 원래 데이터를 고차원공간으로 사상시킴으로써 특징공간 내에 선형으로 분리가능한 입력 데이터 셋을 만든다. 일반적으로 사용되는 커널함수는 다음과 같다.

- 다항식 커널: $K(x_i, x_j) = (1 + x_i \cdot x_j)^d$
- 단순내적커널: $K(x_i, x_j) = x_i \cdot x_j$
- 가우시안 RBF커널:

$$K(x_i, x_j) = exp\left(-\frac{1}{2\delta^2}(x_i \cdot x_j)^2\right)$$

2.2 이탈고객관리에 관한 선행연구

최근 백화점이나 은행, 보험 등의 판매나 인터넷 서비스제공자, 이동통신 서비스와 연관된 산업 등에서 고객유지에 대한 관심이 증가하고 있다. 경제적 관점에서 신규 고객을 유치하는 것보다 기존의 고객을 지속적으로 유지하는 것이 더 큰 경제적 이익을 가져올 수 있을 것이다(NG & Liu, 2000). 고객유지와 고객이탈은 동일한 관점에서 볼 수 있을 것이다. 고객이탈(Churn)이란 자발적 또는 비자발적으로 고객이 현재의 통신서비스를 중단하는 것을 의미하며, 가입해지비율은 일반적으로 특정 기간의 해지비율로 측정된다(Siber, 1997). 고객이탈의 주요 원인은 사용되는 표본이나 연구자에 따라 조금씩 다르게 나타나고 있다.

김진기(1998)는 <표 1>에서 미국 이동전화 서비스 가입자의 가입해지에 의한 발생비용과 기회비용을 연도별로 정리하였다. 이 연구결과에 따르면 평균가입자 유치비용은 줄어드는데 비해 가입해지자의 수는 크게 늘어나고 있으며, 이로 인해서 가입해지로 인한 총비용의 발생이 매우 크게 늘어날 것으로 예측하고 있다. 또한 이를 기회비용의 측면에서 보면, 가입자당 수입이 줄어들기 때문에 고객을 유지했을 때의 기회비용은 매우 커진다는 것을 보여준다.

<표 1> 미국 이동전화서비스 가입자의 가입해지에 의한 발생비용과 기회비용 (김진기, 1998)

구분		1999	2000	2001
연간가입해지자수 (1,000명)		21,952	29,785	34,315
발생비용	가입자유치비용(달러)	136	119	105
	발생비용 (100만달러)	3,425	3,425	3,603
기회비용	가입자당평균수입(달러)	48.08	45.73	43.64
	가입자당6개월수입(달러)	288.48	274.38	261.84
	기회비용 (100만달러)	6,333	7,898	8,985

현재까지의 고객이탈에 대한 선행연구들은 주로 이동통신사에 관한 선행연구가 주를 이루었으며, 이외에도 보험이나 신용카드, 게임산업에서도 이러한 분석이 진행되었다. 이동통신사의 고객에 대한 이탈예측에 관한 연구는 로짓모형과 프로빗 등의 통계기법을 활용한 연구가 많이 이루어졌는데 최근에는 모형의 예측성과 향상을 위해 인공지능 기법들의 사용이 늘어나고 있다. 이러한 예로 인공신경망(김충영 등, 2002), Decision Tree(Wei & Chiu, 2002; 김충영 등, 2002) 등이 사용되었다.

또한 현재 사용되고 있는 이탈고객관리 시스템인 Coral Systems의 Churn Alert이나 GTE Telecommunication Services의 Churn Manager는 고객서비스나 마케팅담당자에 대한 지원을 하기 위한 도구로 사용되고 있다.

김진기(1998)는 이동통신사의 가입해지의 주요 원인을 서비스 요금조건, 서비스 수용지역, 이용자서비스의 문제, 통화품질, 부적절한 요금청구, 계약 만료 및 단말기 교체 등을 그 원인으로 보았다. 김희수(2000)는 경제적 특성, 사용특성, 만족도 등으로 그 원인을 구분하여 Binominal 프로빗모델을 통해 중요도를 측정하였는데 통화품질과 단말기에 대한 만족도가 가장 큰 중요성을 보이고 있으며 회사 이미지와 이동전화 사용 중 불편경험

여부도 중요한 요소로 작용하고 있는 데 반해, 사업자간 요금차이가 크지 않거나 요금 경쟁이 본격화되지 않고 있어서인지 요금수준, 요금제도, 요금 고지의 정확성 등과 같은 요금 관련된 사항들은 영향력이 별로 없는 것으로 나타나고 있다고 하였다.

김충영 등(2003)은 모 이동통신사의 고객데이터를 사용하여 고객 해지모형을 구축하고 평가하였다. 총 82개의 변수에 대한 고객데이터를 사용하여 로짓모형, Decision Tree, Neural Network 모델을 사용하여 정분류율, 이득율, Kolmogorov-Smirnov 통계치와 예측치 분포에 대한 통합모형을 작성하였는데 기존 이동통신사가 보유하고 있는 모든 변수들에 대한 중요도를 각각의 모델별로 평가하여 채택한 후 사용하였다.

Wei & Chiu(2002)는 고객해지의 원인을 파악하기 위해 Taiwan의 이동통신회사의 데이터를 사용하여 계약관련 변수(서비스 기간, 요금제, 계약형태)와 통화관련 변수(사용시간, 사용빈도, 특정기간 통신횟수)로 구분하여 중요도를 측정하였다.

윤충한 등(2002)은 통화품질, 요금, 부가서비스기능, 단말기교체, 업체전환에 따른 인센티브, 유행 등이 가입해지의 원인이라고 보고 있다

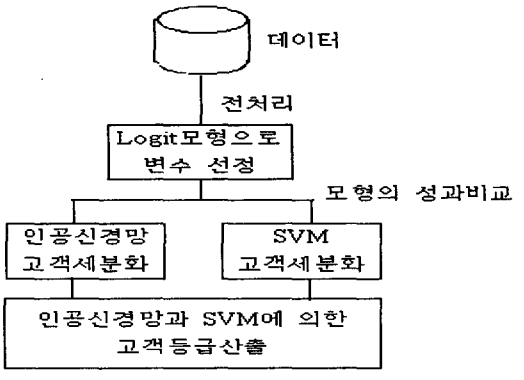
이들을 방법론적인 측면에서 <표 2>로 정리하였다.

<표 2> 선행 연구의 방법론 및 사용변수

데이터마이닝 기법	연구자	사용변수
로짓모형	김충영 등 (2002)	등급, 단말기제품 출시 후 기간, 총 연체, 회선 수, 7월 기본료, 8월 기본료, 6월 기본료, 8월 국내통화, 개통 후 기간,
	윤충한 등(2002)	만족도 변수: 통화만족, 요금만족, 브랜드만족, 단말기만족 사용자 특성변수: 요금사용액, 연령, 소득, 성별 기타 가입변경의사, 가입기간, 과거변경여부, 단말기 보유기간
Binominal프로빗	김희수(2000)	경제적 특성: 소득, 월 평균 사용요금 인구통계학적 변수 : 연령 사용특성: 사용기간, 문제경험, 사용 중인 요금제도 만족도: 통화품질, 월 사용요금, 정확한 요금고지, 부가서비스, 고객 불편 해결, 회사이미지, 단말기
Decision Tree	Wei & Chiu(2002)	계약관련 변수: 서비스 기간, 요금제, 계약형태 통화관련 변수: 사용시간, 사용빈도, 특정기간 통신횟수
	김충영 등 (2002)	등급, 단말기제품 출시 후 기간, 총 연체, 회선 수, 7월 기본료, 8월 기본료, 6월 기본료, 8월 국내통화, 개통 후 기간, 의무사용기간, 정지일수, 마일리지, 교체할부
인공신경망(GA)	김충영 등 (2002)	등급, 단말기제품 출시 후 기간, 총 연체, 회선 수, 7월 기본료, 8월 기본료, 6월 기본료, 8월 국내통화, 개통 후 기간, 의무사용기간, 정지일수, 마일리지, 교체할부

Ⅲ. 연구모형

본 논문에서는 이탈고객의 예측에 많이 이용되고 있는 방법인 로짓모형과 인공신경망, 그리고 최근 들어 문서분류 분야에서 각광을 받고 있는 Vapnik의 SVM을 사용하여 고객의 이탈율을 예측해보고 각 적용기법들의 적중율을 비교하여 보고자 한다. 또한 인공신경망과 SVM의 실험결과로 산출된 이탈율과 적중율로 이탈고객을 등급화하고 실제 현 상황에 맞도록 비율을 조정하여 예상 이탈율을 예측한다.



<그림 3> 인공신경망과 SVM을 이용한 고객등급세분화 연구모형

IV. 실증분석

4.1 자료 수집 및 변수 선정

본 연구는 www.spss.co.kr 제공하는 SK Telecom 데이터를 토대로 Stepwise Method를 수행하여 변수를 채택하였다. 그 결과로 최종 로짓모형의 변수 13개가 의미 있는 변수라고 판명되었으며 이 변수들을 이용하여 인공신경망과 SVM에도 이용하였다. 아래의 <표 3>에 그 결과를 정리하였으며 로짓모형을 수행하기 위하여 요금제, 통화량구분, 납부여부 변수는 더미변수로 처리하였다.

요금제변수는 크게 CAT요금제와 Play요금제로 나뉘어 지는데 CAT요금제는 단계별로 50,100, 200이 있고 Play요금제는 100, 300의 2단계가 있다. 각각 단계별로 올라갈수록 기본요금이 비싸지는 한편 무료통화가 길어지는 특징이 있다. 이 변수와 연관되어서 납부여부 변수는 단계별로 되어 있는 요금제도에서 고객 자신이 속한 요금제도의 수준보다 좀 더 통화량이 많고, 적음을 구분하는 일종의 범주형 필드를 생성하는 것이 목적이다. 이런 필드는 요금제의 잘못된 선택으로 고객이 이탈하는지 안 하는 지의 검정과 비교를 할 수 있고, 동시에 고객이 요금제를 변경하는 경우를 대비한 필드이기도 하다. 이런 고객 자신이 속한 요금제도의 수준이탈 여부를 측정하는 방법으로 자신의 요금 그룹 중 고객 자신의 단계보다 국내통화요금이 바로 위의 단계의 기본요금보다 더 많이 사용한 경우 이를 수준이탈로 간주를 하였다. 각각의 기본요금의 차이에 6(개월 수)을 곱하여 나온 값보다 국내통화요금이 많은 경우는 High로 표현하였다.

통화량구분변수의 경우 국내 통화시간을 기준으로 "무", "저", "중저", "중", "중고", "고" 로 표시하였다(다만 데이터 전처리 과정에서 "무" 의 경우는 존재하지 않았으므로 더미변수 추출 시 제외하였다).

<표 3> 로짓모형 선택변수

변수 명	B[1]	S.E.[2]	Wald[3]	유의확률
연령	0.023	0.001	392.058	0.000
서비스기간	0.008	0.001	64.826	0.000
요금제			434.720	0.000
요금제 더미_1	0.066	0.091	0.525	0.469
요금제 더미_2	-0.030	0.087	0.115	0.735
요금제 더미_3	0.175	0.074	5.589	0.018
요금제 더미_4	1.920	0.123	245.567	0.000
국제통화시간_분	-0.006	0.001	24.180	0.000
평균국내통화시간	-0.017	0.004	21.358	0.000
통화량구분			593.687	0.000
통화량구분_1	-0.010	0.118	0.007	0.934
통화량구분_2	0.011	0.093	0.015	0.904
통화량구분_3	-1.716	0.240	51.053	0.000
통화량구분_4	-1.918	0.147	171.039	0.000
국내통화요금	-0.009	0.004	5.286	0.021
총통화요금	0.011	0.004	8.599	0.003
납부여부			296.506	0.000
납부여부_1	4.057	0.725	31.340	0.000
납부여부_2	0.880	0.756	1.356	0.244
납부여부_3	0.989	0.842	1.378	0.240
평균납부요금	-2.431	0.385	39.943	0.000
주말통화비율	-0.850	0.267	10.094	0.001
국제통화비율	1.760	0.287	37.650	0.000
통화품질불만	-3.643	0.137	711.205	0.000
	-4.074	0.743	30.040	0.000

1) 로짓모형의 계수, 2) 표준오차, 3) Wald 통계량(클수록 선택변수의 유의성이 높음)

4.2 이탈고객 세분화

로짓모형의 성과 분석을 위해서 전체 29390개의 표본을 학습용과 검증용으로 나누어 구성하였으며 그 비율은 8:2(23512:5878)로 구성하였다. 로짓 실험 시 0.5, 0.54, 0.55, 0.56, 0.58, 0.6의 값을 갖는 Cutoff 조정을 통하여 0.56의 Cutoff시 학습용이 66.05%, 검증용이 65.67%로 가장 우수한 예측력을 가졌다고 판명되었다.

<표 4> 로짓모형 성과표(Cutoff 0.56)

	학습용 표본	검증용 표본
유지고객 표본	66.33% (7,798/11,756)	64.75% (1,903/2,939)
이탈고객 표본	65.77% (7,732/11,756)	66.59% (1,957/2,939)
전체	66.05% (15,530/23,512)	65.67% (3,860/5,878)

위의 표에서도 알 수 있듯이 이탈고객의 예측률은 65.93%, 유지고객의 예측률은 66.02%로 두 고객집단간의 예측률은 비슷하며 이탈고객과 유지고객의 비율이 표본 선정 시 50:50으로 선정하였으므로 이들의 예측비율이 가장 비슷한 Cutoff를 선택하는 것이 성과비교에 가장 타당하다고 보았다.

<표 5> 인공신경망 성과표(노드수 20, Cutoff 0.58)

	학습용 표본	검증용 표본
유지고객 표본	71.71% (8,430/11,756)	70.47% (2,071/2,939)
이탈고객 표본	67.75% (7,965/11,756)	67.44% (1,982/2,939)

전체	69.73% (16,395/23,512)	68.95% (4,053/5,878)

본 모형에 적용된 인공신경망의 구조는 역전파 알고리즘 (Back-propagation algorithm)을 이용한 3 layer feedforward 인공신경망이다. 학습률과 모멘텀은 각각 0.1을 사용했으며 은닉층의 노드 수는 1에서 2n중에서 가장 학습이 잘 된 노드수를 선택하였다.

<표 6> RBF 커널함수 사용시 SVM 결과

δ^2	C*	학습용 표본	검증용 표본
1	1	68.65%	67.85%
	25	70.88%	69.82%
	50	69.41%	68.37%
	75	71.46%	70.47%
	100	71.53%	70.50%
25	1	63.97%	63.46%
	25	67.74%	67.20%
	50	68.17%	67.35%
	75	68.27%	67.54%
	100	68.40%	67.86%
50	1	63.97%	63.46%
	25	67.50%	66.67%
	50	67.31%	66.86%
	75	67.52%	66.96%
	100	67.73%	67.22%
75	1	63.97%	63.46%
	25	67.37%	66.67%
	50	67.48%	66.69%
	75	67.28%	66.74%
	100	67.31%	66.83%
100	1	63.97%	63.46%
	25	65.50%	65.19%
	50	67.43%	66.71%
	75	67.47%	66.71%
	100	67.48%	66.66%

* 커널함수의 모수

<표 7> SVM 성과표(Cutoff 0.6)

	학습용 표본	검증용 표본
유지고객 표본	74.29% (8,733/11,756)	71.66% (2,106/2,939)
이탈고객 표본	66.24% (7,787/11,756)	65.12% (1,914/2,939)
전체	70.26% (16,520/23,512)	68.39% (4,020/5,878)

Cut_off는 기본적으로 0.5이지만 훈련용 표본과 검증용 표본의 예측률 차이가 너무 크기 때문에 Cut-off를 각각 0.5, 0.55, 0.6으로 조정하여 실험하였으며, 예측률이 뛰어난 파라미터의 값들 중에서 검증용과 예측용 표본의 예측률 차이가 가장 적은 $\delta^2 = 1$, C=100이며 Cut_off가 0.6을 선택하였다.

아래의 표는 본 연구에서 실험한 기법들의 최고의 성과들을 비교한 표이다. 예측률은 인공신경망이 가장 높았고, SVM, 로짓모형 순이었다. 이러한 예측률의 차이가 통계적으로 유의한가를 검증하기 위해서 Mcnemar 검증을 실시하였으며, 그 결과로서 표에 나타난 바와 같이 SVM과 인공신경망은 유의한 차이가 없었으며, 두 실험결과와 로짓모형과는 차이가 있음을 보여주고 있다.

<표 8> Mcnemar 검정결과표

	로짓모형/인공신경망	로짓모형/SVM	인공신경망/SVM
N	5,876	5,876	5,876
카이제곱	35.756	25.903	1.554
유의확률	0.000*	0.000*	0.213

* 유의수준 1%에서 유의함

본 연구에서는 SVM과 인공신경망의 고객이탈율을 이용하여 고객의 등급을 5단계로 나누어서 등급화를 실시하였으며, 고객의 이탈율이 매우 높은 고객에 있어서는 이동통신사의 입장에서 매우 신경을 써야 하는 등급임을 알 수 있다. 이 연구의 실험결과인 이탈율은 모형에 사용된 전체 고객의 분포가 이탈고객과 유지고객의 비율을 50:50으로 가정하여서 산출한 결과이다. 그러나 현실적인 상황을 고려하여 볼 때, 이탈고객의 비율보다는 유지고객의 비율이 훨씬 많은 것이 사실이므로 이를 현실에 맞게 고객의 비율을 재조정할 필요가 있다. 아래의 표는 SVM과 인공신경망의 실험결과에 실제 이탈고객율을 임의로 산정하여 대입한 대입한 결과이다. <표 11>에서 제시한 바와 같이 SVM의 실험에서 이탈할 확율이 매우 높은 경우에도 실제 예상이탈율 5%를 가정한 결과 9.17%로 환산되어짐을 알 수 있다..

<표 9> SVM을 이용한 고객 등급화

구분(이탈확율)	이탈고객	유지고객	총합계	점유율	이탈율	적중률
1등급(매우 낮음)	365	1,105	1,470	5%	24.83%	75.17%
2등급(낮음)	2,105	5,242	7,347	25%	28.65%	71.35%
3등급(보통)	4,758	6,998	11,756	40%	40.47%	57.21%
4등급(높음)	6,119	1,228	7,347	25%	83.29%	83.29%

5등급(매우 높음)	1,348	122	1,470	5%	91.70%	91.70%
총 합계	14,695	14,695	2,9390	100%		

<표 10> 인공신경망을 이용한 고객 등급화

구분(이탈확율)	이탈고객	유지고객	총합계	점유율	이탈율	적중률
매우 낮음	307	1,163	1,470	5%	20.88%	79.12%
낮음	2,162	5,185	7,347	25%	29.43%	70.57%
보통	4,771	6,985	11,756	40%	40.58%	56.52%
높음	6,035	1,312	7,347	25%	82.14%	82.14%
매우 높음	1,420	50	1,470	5%	96.60%	96.60%
총 합계	14,695	14,695	29,390	100%		

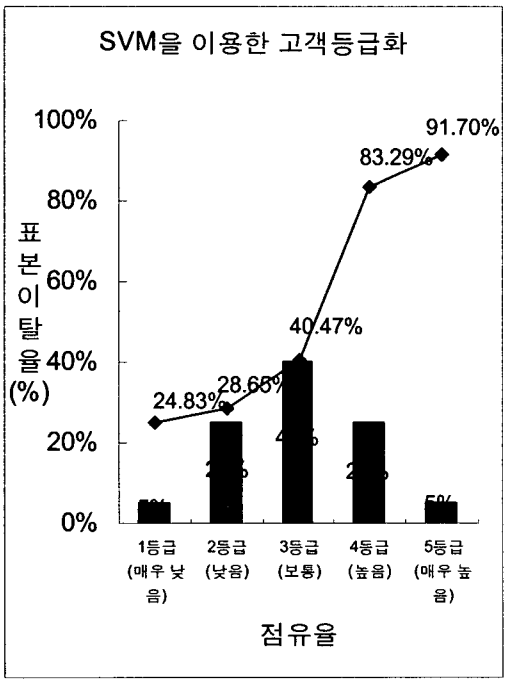


Figure.3 – SVM을 이용한 고객 등급에 따른 표본 이탈율

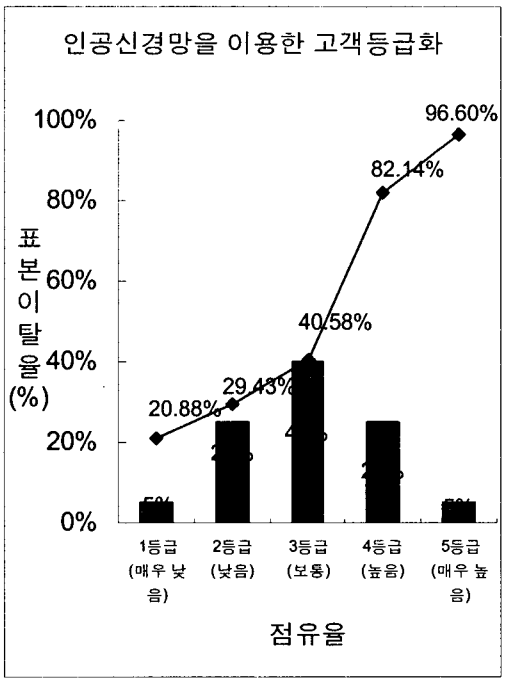


Figure.4 – 인공신경망을 이용한 고객 등급에 따른 표본 이탈율

<표 11> SVM의 고객 등급모형 표본 이탈율과 예상 이탈율(단위: %)

등급(이탈확율)	표본 이탈율 (50%)	예상이탈율 (20% 가정)	예상이탈율 (15% 가정)	예상이탈율 (10% 가정)	예상이탈율 (5% 가정)
1등급(매우 낮음)	24.83	9.93	7.45	4.97	2.48
2등급(낮음)	28.65	11.46	8.60	5.73	2.87
3등급(보통)	40.47	16.19	12.14	8.09	4.05
4등급(높음)	83.29	33.32	24.99	16.66	8.33
5등급(매우 높음)	91.70	36.68	27.51	18.34	9.17
평균	50	20	15	10	5

<표 12> 인공신경망의 고객 등급모형 표본 이탈율과 예상 이탈율(단위: %)

구분(이탈확율)	표본 이탈율 (50%)	예상이탈율 (20% 가정)	예상이탈율 (15% 가정)	예상이탈율 (10% 가정)	예상이탈율 (5% 가정)
1등급(매우 낮음)	20.88%	8.35	6.26	4.18	2.09
2등급(낮음)	29.43%	11.77	8.83	5.89	2.94
3등급(보통)	40.58%	16.23	12.17	8.12	4.06
4등급(높음)	82.14%	32.86	24.64	16.43	8.21
5등급(매우 높음)	96.60%	38.64	28.98	19.32	9.66
평균	50	20	15	10	5

V. 결론

본 연구에서는 최근 패턴 인식 및 분류문제와 관련하여 활발하게 연구되고 있는 SVM을 이탈고객 예측에 적용하여 보았다. SVM은 통계적 이론에 기반하여 설명력이 우수하고, 구조적 위험을 최소화 시키는 모델인 SRM기법을 사용하여 과대적합문제에서 벗어날 수 있는 장점이 존재하므로 이를 이탈고객을 예측하고 고객을 분류해온 기존의 방법들인 로짓모형, 인공신경망과 비교하여 적중율을 예측하였다.

또한 실험결과 성과가 우수했던 인공신경망과 SVM의 적중율에 기반하여 고객의 등급화를 실시하였으며 각 이탈등급별 이탈율과 적중율을 산출하고 실제 현실에 적용 가능하도록 이탈고객의 예상비율을 조정함으로써 이동통신사의 이탈고객 관리에 효과적으로 대응할 수 있을 것으로 기대한다.

본 연구가 이탈고객 예측에 대한 등급화에 SVM을 사용하고 등급화를 산출함으로써 기존의 이탈고객 분석과는 다른 기초적 자료를 제시하였다고 여겨진다. 그러나 실제 분석 결과 인공신경망과 비교하여 유의적인 차이를 보이지는 못했지만 인공신경망의 실험과 비교하여 비슷한 적중율을 보인다는 것을 증명하였다. 이를 토대로 향후 연구에 있어서는 SVM과 인공신경망을 통합한 모형을 제시하여 객관적인 고객 등급을 산출함으로써 이동통신 회사들의 이탈고객 관리에 좀 더 나은 방향을 제시하고자 한다.

참고문헌

[1] 구영규, 강재열(1999), "이동통신 산업에 있어서의 고객이탈관리방안에 관한 연구", 경우논집 Vol.30

[2] 김문구, 정동헌(2002), "이동전화 번호이동성의 고객수요와 시장에 미치는 효과", 통신시장 제 7권 제 8호, pp.66-77

[3] 김재경, 채경희, 송희석(2004), "SOM을 이용한 온라인 게임제공업체의 고객이탈방지 방법론", 경영과학 21권 제3호, pp.85-99

[4] 김진기(1998), "이동전화사업자들의 가입해지(Churn) 방어전략", 정보통신정책 제 10권 제 7호, pp.19-36

[5] 김충영, 장남식, 김준우(2002), "이동통신서비스 해지고객 예측모형의 비교분석에 관한 연구", 경영정보학연구 12권 제1호, pp.139-158

[6] 김희수(2000), "국내 이동전화시장의 가입전환(Churn) 및 고객충성도 결정요인 분석", 정보사회연구, pp.1-18

[7] 박정민, 김경재, 한인구(2003), "Support Vector Machine을 이용한 기업부도예측," 한국경영정보학회 추계학술대회, pp.751-758

[8] 박진기(2003), "데이터마이닝 기법간의 스포츠센터 고객 이탈가능 예측성 평가," 한국체육학회지 제42권 제 5호, pp.369-377

[9] 양희태, 최문기(2003), "번호 이동성 시행 하에서 국내 이동통신 사업자들의 고객 유지 전략", 한국통신학회논문지 Vol.28 No.2B, pp.157-169

[10] 윤성준(2005), "데이터마이닝 기법을 통한 백화점의 고객이탈예측모형 연구,"한국마케팅저널 제 6권 제 4호, pp.45-72

[11] 윤충한, 김희수, 권남훈(2002), "이동전화서비스 이용자의 가입전환(Churning) 및 가입 고착(Lock-in)에 관한 실증분석", 정보통신정책연구 제 9권 제 2호, pp.77-88

[12] 이건창, 권순재, 신경식(2001), "은행고객 세분화를 통한 이탈고객 관리분석 -가계성 예금을 중심으로-," 한국지능정보시스템학회 제7권 제 1호 pp.177-197

[13] 이건창, 정남호, 김재경(2001), "퍼지인식도를 이용한 형식지와 암묵지 결합 메커니즘에 관한 연구 : 신용카드 이탈고객 분석을 중심으로", 경영정보학연구 11권 제4호, pp.113-133

[14] 이건창, 정남호, 신경식(2001), "신용카드 시장에서 데이터마이닝을 이용한 이탈고객 분석", 한국지능정보시스템학회 학술대회 , Vol.1 No.1, pp.421-443

[15] 이수용, 이일병(2002), "Fuzzy 이론과 SVM을 이용한 KOSPI 200 지수 패턴분류기,"한국증권학회 제4차 정기학술발표회, pp.787-809

[16] 이용구, 이현정(2001),

"데이터마이닝을 이용한 보험회사 고객이탈분석에 관한 연구", 수학통계논문집 No.8, pp.37-57
[17] 이학식, 김영(2001), SPSS 10.0 매뉴얼 - 통계분석방법 및 해설 -, 법문사
[18] 허준(2001), 통신시장에서의 데이터마이닝 Telco-CAT, (주) 데이터솔루션
[19] Chih-Ping Wei, I-Tang Chiu(2002), "Turning telecommunications call details to churn prediction : a data mining approach", Expert System With Applications Vol.23 No.2, pp.103-112
[20] Han, Kamber(2003), Data Mining - Concept and Techniques -, 자유아카데미
[21] Hearst, M.A., Dumais, S.T., Osman, E., Platt,j., and Scholkopf, B.(1998), "Support Vector Machine," IEEE Intelligent System, Vol.13, No.4, pp.18-28
[22] KianSing Ng., Huan Liu(2000), "Customer Retention via Data Mining, " Artificial Intelligence Review, Volume 14, Number 6, pp.569-590
[23] Michael J. Shaw., Chandrasekar Subramaniam., Gek Woo Tan., Michael E. Welge(2001), "Knowledge management and data mining for marketing," Decision Support Systems 31, pp.127-137
[24] Rumelhart, D.E., and McClelland, J.L. (1986), *Parallel distributing processing: exploration in the microstructure of cognition*, Vol. 1, Cambridge, MA: MIT Press
[25] Siber, Richard(1997), "Combating the Churn Phenomenon", Telecommunications, Vol.31, No.10, pp.77-80
[26] Swartz, N(2001)., "Churn Alert Competition is getting fierce, churn rates are rising, and carriers are fighting to keep their customers," Wireless review v.18 no.8, pp.44-51
[27] Vladimir N. Vapnik(1995), The Nature of Statistical Learning Theory Second Edition, New York Springer
[28] Wei Huang., Yoshiteru Nakamori., Shou-Yang Wang(2005), "Forecasting stock market movement direction with support vector machine," Computers & Operations Research 32, pp.2513-2522

부록 A- 사용변수

변수명	변수설명
고객ID	고객 ID
이탈여부	0:이탈(14695) 1:유지(14695)
Set1	로짓모형을 위한 셋 구분
Set2	인공신경망과 SVM을 위한 셋 구분
성별	1:(M:남) 0:(F:여)
연령	고객 나이
서비스기간	서비스 개시로부터의 월 사용기간
단선횟수	최근 6개월간의 통화 단선횟수
지불방법	선불/후불 - 제외
요금제	요금제 유형
요금제 더미	CAT50(0000) / CAT100(1000) / CAT200(0100) / PLAY100(0010) / PLAY300(0001)
핸드셋	이동통신의 단말기 - 제외
주간통화횟수	해당 월의 주간통화횟수
주간통화시간_분	해당 월의 주간통화시간
야간통화횟수	해당 월의 야간통화횟수
야간통화시간_분	해당 월의 야간통화시간
주말통화횟수	해당 월의 주말통화횟수
주말통화시간_분	해당 월의 주말통화시간
국제통화시간_분	해당 월의 국제통화시간
국내통화요금_분	해당 월의 국내통화요금
평균주간통화시간	주간통화시간_분/주간통화횟수
평균야간통화시간	야간통화시간_분/야간통화횟수
평균주말통화시간	주말통화시간_분/주말통화횟수
국내통화횟수	국제전화 사용을 제외한 모든 통화횟수의 합 주간통화횟수+ 야간통화횟수+ 주말통화횟수
국내통화시간_분	상동의 의미를 지님
평균국내통화시간	주야주말통화시간의합/국내통화횟수
총통화시간_분	국내통화시간+ 국제통화시간
통화량구분	국내통화량 구분
통화량 구분 더미	저(0000) / 중저(1000) / 중(0100) / 중고(0010) / 고(0001)
요금부과시간	(전체국내통화시간_분- 무료통화시간_분) * 6
분당통화요금	(주간통화요금+ 야간통화요금+ 주말통화요금)/국내통화시간
국내통화요금	금액으로 환산된 고객의 통화자료. (분당통화요금*요금부과시간)/100
총통화요금	국내통화요금+ 국제통화요금
부과요금	실질 부과요금. 통화요금+ 기본요금
납부여부	기본요금 대비 납부 요금액 여부
납부여부 더미	OK(000:A) / High CAT 50(100:B) / High CAT 100(010:C) / High Play 100(001:D)
평균납부요금	부과요금/총통화시간_분
주간통화비율	주간통화 시간/전체 국내통화시간
야간통화비율	야간통화 시간/전체 국내통화시간
주말통화비율	주말통화 시간/전체 국내통화시간
국제통화비율	국제통화시간_분/국내통화시간_분
통화품질불만	0:만족(F,27315) 1:불만(T,2111) / 단선횟수11.517 기준