데이터마이닝 기법을 이용한 일반전화고객의 인터넷전화로의 전환의향에 대한 연구

하성호*, 권은경**, 박현선**, 임광혁***

요약

국내 유·무선 시장이 포화상태를 보이면서 최근 몇 년간 국내 주요 통신사업자들은 인터넷전화 서비스에 나서기 시작했다. 이에 따라 인터넷전화 시장은 지속적인 성장세를 보이기 시작했으며 유선전화를 대체하는 서비스로 부각되면서 사용자가 꾸준히 증가하고 고객들의 통신 이용 패러다임도 변화하고 있다. 따라서 본 연구에서는 유선전화 사용자들의 인터넷전화 전환 의도를 예측해보고자 한다. 이를 위해 데이터마이닝 군집화 기법(TwoStep, K-means) 및 예측기법(C5.0, Neural Net, Support Vector Machine)을 사용한다. 본 연구의 결과는 통신사업자들이 유선전화 사용자들을 인터넷전화로 유인하기 위한 전략을 수립하는데 시사점을 제공할 것으로 판단된다.

A Study for Customer Switching Intentions to Internet Telephony by Using Data Mining Mechanism

Sung Ho Ha*, Eun-Kyung Kwon**, Hyun-Sun Park**, Kwang Huk Im***

ABSTRACT

Internet telephony market is growing continuously and is being emphasized as a substitution for PSTN market. Telecommunication companies and Vendors make an effort to get into internet telephony market, which is regarded as a blue ocean to boost the telecommunication market in a long-term stagnation. In this regard, we need to understand traditional telephone customer churn and transition intention to internet telephony. To fulfill this purpose, this study builds a customer churn analysis model based on a method of data mining. This study conduct clustering(TwoStep, K-means) and uses c5.0, neural nets, and support vector machine to develop prediction models. The results of this study will provide useful insights to manage customer churn and retention.

Key Words : Internet Telephony, Transition intention, Data mining, Clustering Prediction

* 경북대학교 경영학부(✉hsh@mail.knu.ac.kr)
** 경북대학교 대학원 경영학부
*** 배재대학교 전자상거래학과
· 제1저자(First Author) : 하성호 · 교신저자(Correspondent Author) : 권은경
· 접수일(2012년 5월 2일), 수정일(1차 : 2012년 6월 1일), 게재확정일(2012년 6월 5일)

Ⅰ. 서 론

4~5년 전에 틈새시장에 불과했던 인터넷전화가 최근에는 통화 품질의 개선과 번호이동성 제도 등의 시행으로 집전화 시장에서 유선전화를 대체하는 수단으로 주목을 받으며 빠르게 확산되고 있다. 정보통신정책연구원(KISDI)에 따르면, 국내전화 서비스 가입자 중 인터넷전화는 이동전화(97%), 유선전화(72%)에 이어서 가입자의 21%를 차지하고 있는 것으로 나타났다. 인터넷전화 가입자 중에서 기존의 유선전화를 해지한 비율은 81%로 높았으며, 해지 이유로는 통화요금이 저렴하기 때문(75%), 결합요금제 때문(18%)의 순으로 나타났다[1]. 이는 인터넷전화가 유선전화를 대체하는 서비스로서 부각되고 이동이 촉진되고 있음을 의미한다. 그리고 이 두 가지 서비스 모두를 이용하고 있는 가입자들의 향후 이용계획에 대한 설문에서는 70%가 유선전화를 해지할 것이라고 응답했으며, 27%는 인터넷전화를 해지하겠다고 응답해 유선전화에서 인터넷전화로의 한 방향 대체가 진행 중인 것으로 분석할 수 있다[1]. 이처럼 인터넷전화는 이제 유선전화의 대체 서비스로서 확실하게 자리매김하고 있고 통신사업자들 역시 인터넷전화로의 전환의향을 가진 고객들을 위한 새로운 전략을 수립하거나 그들의 통신 이용 패러다임 변화에 주목하고 있다. 특히, 고객이탈을 관리하는 비용이 신규 가입자 유치에 드는 비용보다 저렴하기 때문에 사업자들은 이탈고객 관리에 대해 더 많은 관심을 가지고 있으므로 고객 유지에 대한 분석 및 연구가 필요하다 할 수 있다. 하지만 인터넷전화 사용자가 급증하고 이탈고객 관리에 대한 관심이 증가하고 있는 반면에 인터넷전화에 대한 연구는 인터넷전화의 제도 현황이나 인터넷전화의 확산, 기술동향 등 대부분의 연구들이 주로 기술적인 측면 및 제도에 대해서만 강조하고 있다[2,3]. 또한, 데이터마이닝 기법을 이용한 이동통신시장의 고객 분석은 연구자들이 꾸준히 관심을 가지고 있는 분야이지만[4,5] 인터넷전화와 서비스전환 예측에 관한 분석은 미미한 실정이다. 이러한 점에서 본 연구는 고객 이탈을 관리하고 유선전화 사용자들을 인터넷전화로 유인하기 위한 전략 수립에 필요한 정보를 얻기 위해 데이터마이닝 기법을 적용시켜 어떤 요인이 이탈에 영향을 미치는지를 알아보고자 한다. 본 연구를 통해 고객 이탈에 영향을 미치는 원인이나 전환의도를 예측할 수 있다면 효과적인 고객 유지 및 이탈 관리를 위한 시사점을 제공할 수 있을 것이라 판단된다. 따라서 본 연구에서는 데이터마이닝 기법을 적용하여 유선전화를 사용 중인 고객들의 인터넷전화 사용의향을 예측하고자 한다.

Ⅱ. 이론적 배경

2.1 인터넷전화

인터넷 전화는 전기통신사업법 제7조에 의해 기간통신역무로 규정된다. 인터넷전화는 인터넷 망을 이용해 통화권 구분 없이 기존의 음성통신 서비스를 실시간으로 제공하는 서비스로 흔히 Voice over Internet Protocol(VoIP), Internet Telephony 등으로 불린다. 유선전화(Plain Old Telecommunication Service: POTS)의 경우 가입자의 단말기로부터 교환기까지의 연결이 단일 선로를 통해 이루어졌지만, 인터넷전화는 음성 데이터를 패킷망(전용망, Internet)을 통해 디지털화시켜 교환기로 전달하는 방식을 이용한다.

인터넷전화는 1995년 3월, 이스라엘 VocalTec사가 VoIP 게이트웨이를 상용화하면서 최초로 시작되었다. 국내에서는 2000년 1월 새롬 기술에서 무료로 PC to Phone 형태의 서비스를 제공하면서 인터넷전화가 도입되었으나 열악한 통화 품질과 소비자의 부정적 인식, 빈약한 수익모델 등의 문제로 성공을 거두지 못했다. 그러나 인터넷전화의 상용화와 번호이동 제도의 도입, 통신비 절감에 대한 소비자들의 관심이 맞물리면서 인

터넷전화 보급률은 급속도로 증가하고 있는 추세이다. 특히, 추가 설비 없이도 기존의 인터넷 망을 통해 전 세계 사람들과 통화할 수 있다는 개방성과 저렴한 통화료는 인터넷전화에 대한 관심을 증대시키고 있다. 다음의 표 1은 유선전화와 인터넷전화의 특징을 비교해 놓은 것이다.

표 1. 유선전화와 인터넷전화 비교[1]
Table 1. Comparison of wired telephone and Internet telephone[1]

구분	유선전화	인터넷전화
이동성 비교	실내(고정)	인터넷 접속방식에 따라 다름
지역 구분	통화지역 구분	통화지역 비구분
주요 서비스	음성	음성과 인터넷서비스 모두 이용
요금	저렴함	매우 저렴함

2.2 고객이탈과 데이터마이닝

최근에는 신규 고객을 유치하는 것보다 기존 고객을 유지하는 비용이 더 저렴하다는 점에서 고객 관리와 이를 위한 전략수립에 대해 기업들의 관심이 증가하고 있다[6,7]. 기존 고객 유지와 이탈 고객 관리는 동일한 관점으로 볼 수 있으며 고객이탈은 자발적 또는 비자발적으로 고객이 현재의 통신서비스를 중단하는 것을 의미한다[8]. 이탈고객에 대한 연구는 인터넷 쇼핑몰이나 백화점, 통신서비스, 금융, 보험 등의 다양한 산업분야에서 수행되고 있으며, 데이터마이닝 기법을 적용한 연구도 많이 다루어지고 있는 있다. 특히, 고객관계관리(Customer Relationship Management, CRM)의 관점에서 기존 고객을 유지하고 고객 수익성을 증대시키려는 접근은 고객의 행동과 특성을 이해하고 분석하려는 데이터마이닝 기법을 적용시킴으로서 경쟁적인 CRM 전략의 개발에 토대를 마련한다는 점에서 중요한 의미를 가진다고 할 수 있다. 또한, 데이터마이닝의 분석결과를 통해 잠재 고객을 유지하고 고객을 세분화하여 고객 가치를 최대화하는 이점을 얻을 수 있다.

기존 고객유지 및 고객이탈을 예측하기 위해 데이터마이닝 기법을 적용한 연구를 살펴보면 통신, 금융, 쇼핑 등 다양한 분야에서 데이터마이닝이 활용되고 있음을 알 수 있으며 CRM 전략 개발에 도움을 줄 수 있는 시사점을 제공하고 있음을 확인할 수 있다. 먼저, 이지영과 김종우의 연구에서는 신용카드사의 실제 데이터를 활용하여 고객 이탈을 연구하였으며, 의사결정나무, 로지스틱 회귀분석, 신경망을 이용하여 분류모델을 생성하였다. 그리고 고객들의 상태변화를 변수로 선정하여 상태변화와 고객이탈 여부간의 연관성 규칙을 생성하여 최종적으로 고객이탈을 예측하였다[9]. 임세헌과 허연은 온라인 자동차 보험에서 고객이탈을 예측하기 위해 의사결정나무, 다변량판별분석, 로짓분석을 적용하여 비교하였다[10]. 조대현 등은 화장품 회사의 구매자료로부터 고객의 구매 행태를 분석하기 위한 연구를 수행하였으며, 의사결정나무, 로지스틱 회귀분석 및 신경망 기법을 사용하였다[11]. 또한, 모형비교 및 평가를 통해 해석과 활용이 쉬운 의사결정나무를 고객 충성도를 예측하기 위한 모형으로 선택하여 제시하였다.

이처럼 다양한 분야에서 다루어지고 있는 데이터마이닝 기법을 본 연구에서는 통신서비스 분야에 접목시켜보고자 한다. 통신서비스의 경우 시장이 포화상태로 성장이 점차 둔화되고 있고 최근에는 번호이동 및 결합상품과 같은 서비스로 인해 고객이탈이 매우 활발하게 이루어지고 있는 상황이다. 또한, 통신서비스 분야가 고객 유지비용에 비해 신규고객 창출비용이 매우 높다는 점에서 많은 연구자들은 해지 가능성이 높은 고객을 사전에 발견하고, 해지가 발생하는 원인을 알 수 있다면 고객이탈에 적절히 대응할 수 있고 둔화되어 가는 시장에서 경쟁력을 가질 수 있음을 설명하였다[4,7].

통신서비스 시장의 고객 이탈에 관한 연구를 살펴보면, Wei et al.의 연구에서는 무선통신 서비스 고객의

해지 원인을 파악하기 위해 대만의 이동통신회사의 데이터를 사용했으며, 계약관련 변수와 통화관련 변수로 변수를 구분하여 중요도를 측정하였다[12]. Daskalaki et al.은 대형 통신회사의 고객자료를 바탕으로 신경망, 의사결정나무, 판별분석을 이용하여 파산고객에 대한 예측을 비교하였으며, 이병엽 등의 연구에서는 인공신경망과 로지스틱 회귀분석을 이용하여 국내 이동통신 산업을 대상으로 고객 분류 예측을 위한 모형을 제시하였다[13,5]. 이들은 국내 이동통신 시장이 성숙화 및 포화로 경쟁방식이 기존 고객의 유지 및 서비스의 차별화를 통해 이루어져야 한다는 점에 주목하고 있다. 홍태호와 전성용은 SVM을 이동통신사들의 이탈고객 관리를 위한 예측에 적용하였으며, 의사결정나무, 로짓모형, 신경망과 성과를 비교하여 제시하였으며, Hung et al.의 연구는 이탈 관리가 통신서비스 분야에서 중요한 요인임을 설명하고 이를 위해 이탈 고객을 예측해 볼 수 있다고 주장하였다[4,14]. 그리고 신경망과 의사결정나무를 이용하여 무선 통신서비스를 이용하는 고객의 이탈 관리를 위한 모형을 개발하였다. 또한, 고객들의 인구통계학적 정보, 지불 정보, 계약이나 서비스 상태, 서비스 변화 등이 예측에 영향을 미치는 요인임을 설명하였다.

Chu et al.의 연구에서는 보유 고객의 이탈을 예측하는 것이 CRM의 주요 이슈라 주장하며 고객 이탈을 최소화하는 것이 기업의 수익을 극대화한다고 하였다[15]. 그리고 C5.0을 이용하여 통신회사 가입자의 이탈을 예측하였다. Ahn et al.의 연구는 이동통신 운영자가 다양해지고 있고 전통적인 전화서비스에서 벗어나 서비스 종류가 많아지고 있음을 설명하였다[16]. 그리고 수익성을 확대하기 위해서 고객을 세분화하여 분류하고 새로운 서비스의 구매가능성을 예측해야 할 필요성을 주장하였다. 이를 위해 인공신경망, 의사결정나무, 로지스틱 회귀분석을 사용하였으며 한국의 통신 기업의 실제 데이터를 적용하여 높은 품질의 정보를 생산하기 위한 모형을 구축하고 기업의 성과 향상에 영향을 줄 수 있는 시사점을 도출하고자 하였다. Farvaresh and Sepehri는 통신 서비스를 제공하는 기업들이 고객 이탈이나 사기와 같은 행동에 의해 잠재적인 손해를 입을 수 있음을 설명하였다[17]. 그리고 고객의 사기성을 예측하기 위해 데이터마이닝을 접목시켰으며 신경망, 의사결정나무, 자기조직화맵(Self-Organizing Map, SOM) 기법을 사용하여 고객을 세분화하였으며 테헤란의 통신회사 데이터를 사용하여 효율적인 방법을 제안하고자 하였다. 이들 선행연구에서 고객의 이탈 및 서비스 해지확률을 예측하기 위해 활용하는 변수는 이용하는 서비스 유형과 통화시간, 나이, 성별, 소득 등의 인구통계학적 기준, 단말기 및 서비스 요금조건, 수용지역 등 이라고 설명하고 있으며, 고객들의 이탈과 유지를 위한 예측은 경쟁력을 강화시키고 수익성을 향상시키기 위해서 중요하게 수행되어야함을 강조하였다[6,7]. 다음의 표 2는 이들 선행연구를 정리하여 나타낸 것이다.

이처럼 통신서비스 분야는 신규고객 창출비용이 높고 고객의 이탈이 빈번하게 이루어지기 때문에 데이터마이닝 기법과 고객에 대한 방대한 데이터를 활용한다면 이탈 고객을 예측할 수 있고 적절한 전략 수립에 이를 활용할 수 있을 것이다. 본 연구에서는 데이터마이닝 기법을 이용해 기존 유선전화를 사용 중인 고객들을 군집화하고 유사한 특성을 가진 고객들의 인터넷전화 사용 의향을 예측하여 인터넷전화 서비스를 제공하는 통신사업자들이 수립할 수 있는 적절한 전략을 제시한다. 그리고 유선전화 사용자들의 인터넷전화 전환의향을 예측할 수 있는 모델의 개발을 통해 이탈 고객 관리에 대한 시사점을 제공하며, 이탈 및 해지고객을 위한 전략적 시사점을 제공한다.

III. 분석의 절차

본 연구의 첫 번째 단계에서는 TwoStep, K-means

를 이용하여 군집화를 수행하는 단계로서 고객의 특성 및 성향을 반영한 고객군집화는 고객의 행동 패턴 정보의 전략 수립에 기반이 될 수 있다. 두 번째 단계는 예측 모형을 개발하는 단계로 유사한 특징을 가진 군집으로

표 2. 고객이탈방지관리와 관련한 선행연구
Table 2. Previous research regarding customer churn management

연구자	연구분야	연구내용
이지영과 김종우 (2007)	신용카드	의사결정나무, 로지스틱 회귀분석, 인공신경망 등을 사용하여 신용카드사의 이탈 고객을 예측하는 방법을 제시함
이병엽 등(2006)	통신	신경망, 회귀분석을 사용하여 이동통신업체의 기존고객을 세분화하기 위한 예측모형을 설계
임세헌과 허연 (2006)	자동차보험	C5.0을 이용하여 온라인 자동차보험 고객이탈을 예측하고 다변량판별분석과 로짓분석을 이용하여 예측 결과를 비교함
홍태호와 전성용 (2006)	통신	SVM을 이용하여 이탈고객 관리를 위한 예측모형을 설계하였으며, 의사결정나무, 로짓모형, 신경망과의 성과를 비교함
Ahn et al.(2011)	통신	통신서비스회사가 수익성을 확대시키고 고객을 유지하며 표적 마케팅을 위해 데이터마이닝의 분류기법을 사용하였음
Chu et al.(2007)	통신	C5.0을 이용하여 통신회사 가입자의 이탈을 예측 및 분석함
Daskalaki et al. (2003)	통신	의사결정나무, 신경망, 판별분석을 이용하여 파산고객에 대한 예측 모형을 개발하고 이들 성과를 비교함
Farvaresh and Sepehri(2011)	통신	신경망, 의사결정나무, SOM 기법을 사용하여 고객의 사기성을 예측하고자 하였음
Hung et al.(2006)	통신	의사결정나무와 신경망을 이용하여 통신서비스를 이용하는 고객의 이탈을 예측 및 관리하기 위한 모형을 개발하고 예측에 영향을 미치는 요인들을 제시함
Wei et al.(2000)	통신	데이터마이닝 기법을 적용하여 고객의 서비스 해지 원인을 분석하고 중요 변수를 도출함

군집화 된 고객들의 인터넷전화 서비스 사용 의향을 예측하기 위하여 의사결정나무(Decision Tree)와 신경망(Neural Net), SVM(Support Vector Machine)을 사용한다.

3.1 고객군집화를 위한 군집분석

군집화(Clustering)란 군집 특성에 대한 사전 정보가 주어지지 않은 상태에서 주어진 데이터들을 유사한 특성을 지닌 군집으로 분할하여 형성하는 기법을 말한다[18]. 데이터의 분포를 탐색하고 각 군집의 특성을 관찰하기 위해 사용될 수도 있으며, 분류와 같은 다른 알고리즘을 위한 전처리 단계로써 사용될 수 있다. 본 연구에서는 설문지를 통해 얻어진 인구 통계학적 정보와 통신서비스 관련 정보를 군집분석에 이용하며, 적절한 군집 수를 선정하기 위해 먼저 TwoStep 군집화를 수행한다.

3.1.1 TwoStep 군집화

TwoStep 군집화는 2단계로 계층적 군집화를 수행함으로써 적절한 수의 군집을 결정할 수 있는 기법으로 자료를 한번만 읽기 때문에 연속형 변수와 범주형 변수 모두를 이용하여 군집화를 할 수 있다. 특히, 계층적 군집분석에서는 군집의 수가 1, 2, 3,.., K와 같이 하나의 군집에서 전체 개체의 수만큼 설정될 수 있으므로, 각각의 군집 수에서 Schwarz Bayesian Criterion 통계량을 구할 수 있다. 이 통계량을 이용하여 군집화 단계에서 차이가 가장 작은 단계의 군집화를 선택한 후에 그 군집화를 기준으로 바로 다음 단계 군집화의 최소 '군집내 거리'를 계산하여 최적의 군집 수를 결정한다.

3.1.2 K-means 군집화

K-means 군집화는 사전에 군집 수 K를 정해 가까운 거리에 있는 관찰치를 K개의 군집으로 나누어 군집 내

유사성은 높이고 군집 사이 유사성은 낮게 하는 방법이다. 중심점으로부터 거리를 계산하고 군집을 구하여, 주어진 데이터 집합을 분류한다. N개의 속성으로 구성되는 각각의 레코드를 벡터로 표시하여 N차원의 데이터 공간에 나타낼 때, 유사한 특성을 갖는 레코드들은 서로 근접하여 위치한다는 가정 에 근거하고 있다. K-means 알고리즘에서 이용되는 변수들은 원칙적으로 범주형이어야 하고 알고리즘 적용에 앞서 표준화되어야 한다.

3.2 인터넷전화로의 전환의향 모형 개발을 위한 예측기법

예측기법은 데이터의 특성을 기술하는 모형을 구축하거나 미래의 경향을 예측하는데 사용될 수 있는 데이터 분석방법이며, 많은 분류 및 예측 기법은 기계학습, 전문가 시스템, 통계학, 그리고 신경 생물학 분야의 연구에서 제안되었다[18].

3.2.1 의사결정나무(Decision tree)

의사결정나무는 의사결정규칙을 나무 구조로 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 하는데 효과적으로 사용되는 분석기법이다. 데이터 레코드들을 분석하여 이들 사이에 존재하는 분류 별 특성을 속성의 조합으로 나타낸다. 그리고 만들어진 분류 모형은 새로운 레코드를 분류하고 해당 분류의 값을 예측하는데 사용된다. 의사결정나무는 생성된 모형이 신경망, 판별분석 등의 다른 방법들에 비해 단순하여 연구자가 분석과정을 쉽게 이해할 수 있도록 도와주고, 예측 정확도가 다른 분류모형보다 높거나 동등하다. 또한 나무를 만드는데 있어서 입력매개변수를 요구하지 않기 때문에 많이 사용되는 방법 중 하나이다[19].

3.2.2 신경망(Neural Net)

신경망은 인간 두뇌의 신경세포를 모방한 개념으로 뉴런(neuron) 또는 노드(node)와 고리(link)로 서로 복잡하게 구성된 망구조를 모델링하고, 의사결정나무와 마찬가지로 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다[20]. 분류, 군집분석 발견과 같은 작업에 널리 사용되며 고객세분화, 수요 및 판매 예측 등의 목적으로도 사용되고 있다. 은닉노드(hidden unit)라 불리는 구성요소에 의해서 일반적인 통계모형과 구별되며, 출력노드가 은닉노드로부터 신호를 전달받아 결합하여 최종 반응을 내보내는 방식이다[21].

3.2.3 SVM(Support Vector Machine)

SVM은 데이터로부터 분류와 규칙을 학습하기 위해 학습 집단에서 마진을 최대화하는 결정면을 찾아내는 알고리즘이다. 기본원리는 학습 데이터들을 서로 다른 두 개의 클래스로 분류할 때 기준이 되는 분리 경계면을 SVM 알고리즘을 이용해 찾는 것이다. 그리고 선형으로 분리할 수 있는 학습 집단에 대해서 최대 마진을 가지고 두 클래스를 구분하는 결정면을 찾아내는 SVM과 선형으로 분리할 수 없는 경우 커널 함수에 의해 만들어진 비선형 결정함수를 이용하여 최적의 초평면을 구축하는 비선형 SVM으로 분류된다. 입력된 값을 간단하고 명료한 알고리즘을 통해 학습한 뒤 얻는 분류 결과는 직관적인 해석이 가능하다는 특징을 가지고 있다.

Ⅳ. 분석 결과

4.1 데이터 분석

본 연구는 설문 자료를 이용하여 유선전화 고객을 군집화를 통해 분리한 후 집단 별로 인터넷전화로의 전환의향을 예측하기 위한 모형을 개발하는 것을 주요 목적으로 한다. 설문데이터는 국내 A 통신사가

2008년 8월에 실시한 것으로 원래는 집전화 기반 신규 통신서비스를 추가 실시하기 전에 소비자 조사를 하기 위함이었다. 설문은 고객들의 인구통계학적 정보, 현재 이용하고 있는 통신 서비스, 향후 새로운 서비스 이용에 대한 정보를 포함하고 있다. 총 705명의 고객이 설문에 응답하였으며, 순서형 응답으로 구성된 세 문항(나이, 월가구소득, 통신요금)을 제외한 모든 설문은 범주형 응답으로 구성되어 있다.

4.1.1 데이터 전처리

데이터마이닝은 일반적으로 통계적인 추론 및 검정보다는 예측모형의 일반화에 초점을 두고 있으며, 숨겨진 지식을 도출하는 방법론이다. 불완전하고 일관성이 부족한 현실 데이터를 활용하여 예측모형을 만들 때, 최소한의 독립 변수로 모형을 만드는 것이 차후에 구축된 모형의 설명이 용이하기 때문에 데이터 전처리 과정은 반드시 필요하다. 본 연구에서는 기존의 설문 데이터에서 새로운 지식을 발견하기위해 데이터를 정제 및 통합하였고 필요한 데이터를 선별 및 변환하는 전처리 과정을 거쳤다.

데이터 전처리 과정으로 설문을 통해 얻어진 데이터에서 예측모형을 개발하기 전에 KISDI의 결과를 바탕으로 인터넷 전화로의 전환에 영향을 주며 중요하게 여겨지는 변수를 중심으로 새로운 변수를 추출 및 생성하였다[1]. 이는 설문 문항 중에 본 연구에 직접 이용 가능한 항목들도 있지만 설문 항목의 결합을 통해서만 얻을 수 있는 변수들이 있었기 때문이다.

따라서 인터넷전화 전환의향을 나타내는 종속변수 'E07'을 생성하였으며, 전환의향을 가진 경우는 '1'의 값을, 전환의향을 가지고 있지 않은 경우는 '0'의 값을 부여하였다. A18, A19, A20은 각각 유선전화, 초고속 인터넷 이용여부, 인터넷 이용여부에 대한 응답으로 이들 문항을 결합하여 E04를 생성하였다. 변수 E05는 유선전화를 사용하는 고객들의 불만족여부에 대한 문항 A31, A37을 결합하여 생성하였으며, 변수 E07은 인터넷전화 사용계획이나 의향에 대한 문항인 A32, A41와 A46을 바탕으로 생성하였다. 표 3은 본 연구를 위해 새로 생성한 변수를 나타낸 것이다.

표 3. 새로 생성된 변수와 변수 설명
Table 3. New variables and their descriptions

변수	정의	이용 변수	변수설명
E04	통신 결합상품 이용 여부	A18, A19, A20	이용하지 않음=0 이용하고 있음=1
E05	유선전화 불만족 여부	A31, A37	없음=0, 보통=1 있음=2
E06	통신요금	A26, A43	유선전화 2만원 미만=0 2~3만원=1 3만원 이상=2 인터넷전화 1만원 미만=0 1~2만원=1 2만원 이상=2
E07	인터넷전화 사용 의향	A32, A41, A46	없음=0, 있음=1
E09	현재 이용 중인 통신서비스 중 유선전화 이용여부	A18	이용하지 않음=0 이용하고 있음=1
E10	현재 이용 중인 통신서비스 중 초고속인터넷 이용여부	A19	이용하지 않음=0 이용하고 있음=1

4.1.2 데이터 정규화

표 4는 본 연구에서 사용된 정규화한 변수를 제시한 것이며 본 연구에서는 이 변수 모두를 고객 군집화와 예측 모형 개발에 사용하였다. 단, E07은 종속변수로서 예측 모형 개발에 사용되었다. 설문 응답을 통해 얻어진 인구통계학적 정보에 포함된 데이터는 거주 지역,

성별, 나이, 주택형태, 가족유형, 직업, 월가구소득, 맞벌이 여부이며 범주형 데이터로 구성되어 있다. 예측을 위해 사용되는 신경망 알고리즘은 입력정보와 출력정보가 0과 1사이의 값을 취할 때 최적의 성능을 제공하므로 변수 값을 0과 1사이의 값으로 변환시키는 것이 바람직하다. 본 연구에서는 분석에 앞서 모든 변수의 값을 0과 1사이로 변환시키기 위한 정규화를 시행하였다. 정규화를 수행하기 위해서 IBM 모델러의 Derive 노드*를 사용하였으며, 0과 1사이의 값으로 변환하였다. 먼저 변수가 0, 1로 2개의 값을 가진 경우에는 그대로 사용하였으며, 3개의 값을 가진 경우에는 0, 0.5, 1로 변환시켜주었다. 그리고 4개의 값을 가진 경우에는 0, 0.33, 0.66, 1의 값으로 변환시켜주었으며, 5개의 값을 가진 경우에는 0, 0.25, 0.5, 0.75, 1의 값으로 변환시켜주었다. 마지막으로 6개의 값을 가진 경우에는 0, 0.2, 0.4, 0.6, 0.8, 1의 값으로 변환시켜주었다. 그리고 IBM 모델러의 Merge 노드**를 사용하여 정규화 된 모든 변수를 합친 SPSS파일을 생성하였다. 이 파일을 사용하여 다음에 나오는 군집화 및 인터넷전화 사용의향 예측모형 개발을 시행하였다.

표 4. 정규화 변수
Table 4. Normalized variables

변수명	변수 설명
A05 (거주지역)	서울=0, 인천=0.2, 부산=0.4, 대구=0.6, 광주=0.8, 대전=1
A12 (성별)	남자=0, 여자=1
A13 (나이)	20대=0, 30대=0.33 40대=0.66, 50대이상=1
A15 (주택형태)	단독주택=0, 다세대/연립=0.5 아파트=1
A17 (가족유형)	1인가구=0 신혼가구=0.2 유아자녀가구=0.4 초/중/고등자녀=0.6 대학생/성인자녀=0.8 노년가구=1
C14 (응답자직업)	자영업=0 블루칼라=0.25 화이트칼라=0.5 가정주부=0.75 학생/무직=1
C16 (가구주직업)	자영업=0 블루칼라=0.33 화이트칼라=0.66 주부/학생/무직=1
C28 (월가구소득)	300만원 미만=0 300~500만원 미만=0.5 500만원 이상=1
C32 (맞벌이여부)	맞벌이부부임=0 맞벌이부부아님=1
E04 (통신결합상품이용여부)	이용하지 않음=0 이용하고 있음=1
E05 (유선전화불만족여부)	없음=0, 보통=0.5, 있음=1
E06 (통신요금)	유선전화 2만원 미만=0 2~3만원=0.5 3만원 이상=1 인터넷전화 1만원 미만=0 1~2만원=0.5 2만원 이상=1
E07 (인터넷전화 사용의향)	없음=0 있음=1
E09 (현재유선전화이용여부)	이용하지 않음=0 이용하고 있음=1
E10 (현재초고속 인터넷이용여부)	이용하지 않음=0 이용하고 있음=1

* 필드옵션 중 특정변수의 변수 값을 연구자의 필요에 따라 변환하는데 사용하는 노드이다.

** 레코드옵션 중 하나로서 같은 규칙에 따라 배열되어 있는 2개 이상의 데이터 집합을 하나의 데이터 집합으로 종합하는 노드이다.

4.2 고객군집화

일반적으로 군집화는 예측을 수행하기 전에 동질의 집단을 분리하기 위해 수행하는 경우가 많다. 그리고 다수의 선행연구에서 군집화를 통해 얻은 동질적인 집

단에 대해 예측을 수행하였을 경우 더 나은 예측결과가 도출되었다[5,22]. 군집화는 명확한 기준에 의해 이루어져야 하며, 여기서 기준은 고객들의 특성을 나타내는 변수들로 구성된다. 고객군집화를 위해 설문 응답 데이터를 이용하여 유선전화 고객집단을 군집화 한다. 먼저 TwoStep 알고리즘을 사용하여 적절한 군집 수를 구하기로 하며 표 4에 제시된 변수를 사용하였다. 본 연구에서는 분석대상들을 상호관련성에 의해 서로 동질적인 집단으로 묶는 고객군집화를 위해 IBM 모델러를 사용하여 군집화 분석을 하였다.

4.2.1 TwoStep 군집화

군집화에는 예측 모형의 개발에서 사용할 종속변수인 E07을 제외한 14개의 범주형 변수를 이용하였다. 군집분석을 할 때 파라미터 설정 값 중 군집 수의 범위는 최소 2개에서 최대 20개로 정의하여 자동적으로 예비-군집(sub-cluster)을 생성하였다. TwoStep 군집화의 군집분석 결과, 2개의 군집이 생성되었으며, 각 군집을 보면, 먼저 군집-1은 478개 레코드, 군집-2는 227개 레코드로 구성되었다. 이와 같은 TwoStep 군집화 결과에 따라 본 연구의 K-means군집화 시, 군집 수를 2개로 정의하여 사용하기로 한다.

4.2.2 K-means 군집화

사전 군집단계인 TwoStep 군집화에서 구한 군집 수 2를 사용하여 K-means 군집화를 수행하였다. K-means 군집 시, 유클리드 거리를 사용하고 최대 20번 반복수행하며 허용오차를 0.0으로 정의하여 군집화하였다. K-means 군집화의 파라미터들은 표 5에 나타나 있다. K-means 군집화 결과, 군집-1이 558개 레코드, 군집-2가 147개 레코드를 가진다.

또한 각 군집의 특성을 이해하기 위해서 각 군집화 변수별-군집별 중심을 보여주고 있다. 두 군집 간의 거리는 1.581이었다.

4.2.3 TwoStep과 K-means 비교 분석

IBM 모델러의 이분형 로지스틱 분석을 통해서 TwoStep의 R-square 값은 0.680으로 K-means의 R-square 값은 0.641로, 군집화에 대한 R-square값을 보여주고 있다. 본 연구에서는 R-square 값이 더 높게 나타나는 TwoStep의 군집을 선택하였다.

표 5. K-means 파라미터 설정 값
Table 5. Parameter values for K-means

Parameters	
Number of clusters	3
Optimize	Memory
Mode	Expert
Stop on	CUSTOMER
Maximum iterations	20
Change tolerance	0.0
Encoding value for sets	0.707

4.2.4 TwoStep 군집분석

표 6은 TwoStep 군집결과와 요인 간에 교차분석을 실시한 결과이며, 그 결과 TwoStep 군집화를 이용해 군집화한 각 군집의 특성을 구체적으로 살펴보면 다음과 같다.

군집-1의 경우, 연령대는 30대가 가장 높은 빈도를 보였으며, 주택형태는 아파트에 거주하고 초/중/고등자녀를 둔 가족유형이 가장 높게 나타났으며 다음으로 유아자녀를 둔 가족유형 대학생/성인자녀를 둔 가족유형으로 나타났다. 월가구소득이 300~500만원 미만이 가장 높게 나타났으며, 다음으로 300만원 미만으로 나타났다. 대부분 통신결합상품을 사용하는 것으로 나타났다. 군집-2의 경우, 연령대는 50대가 가장 높은 빈도를 보이며 주택형태는 단독주택이 가장 높게 나타났다.

표 6. TwoStep 군집결과와 요인 교차분석(n (%))
Table 6. TwoStep cluster analysis and cross-factor analysis(n (%))

요인명		군집분류					
		군집-1(N=478)		군집-2(N=227)		Overall(N=705)	
A05(거주지역)	0	177	(37.03)	87	(38.33)	264	(37.45)
	0.2	57	(11.92)	23	(10.13)	80	(11.35)
	0.4	83	(17.36)	37	(16.30)	120	(17.02)
	0.6	56	(11.72)	25	(11.01)	81	(11.49)
	0.8	49	(10.25)	31	(13.66)	80	(11.35)
	1	56	(11.72)	24	(10.57)	80	(11.35)
A12(성별)	0	117	(24.48)	87	(38.33)	204	(28.94)
	1	361	(75.52)	140	(61.67)	501	(71.06)
A13(나이)	0	53	(11.09)	52	(22.91)	105	(14.89)
	0.33	204	(42.68)	40	(17.62)	244	(34.61)
	0.66	150	(31.38)	33	(14.54)	183	(25.96)
	1	71	(14.85)	102	(44.93)	173	(24.54)
A15(주택형태)	0	157	(32.85)	105	(46.26)	262	(37.16)
	0.5	106	(22.18)	65	(28.63)	171	(24.26)
	1	215	(44.98)	57	(25.11)	272	(38.58)
A17(가족유형)	0	37	(7.74)	64	(28.19)	101	(14.33)
	0.2	87	(18.20)	38	(16.74)	125	(17.73)
	0.4	115	(24.06)	11	(4.85)	126	(17.87)
	0.6	124	(25.94)	4	(1.76)	128	(18.16)
	0.8	95	(19.87)	30	(13.22)	125	(17.73)
	1	20	(4.18)	80	(35.24)	100	(14.18)
C14(응답자직업)	0	86	(17.99)	43	(18.94)	129	(18.30)
	0.25	90	(18.83)	50	(22.03)	140	(19.86)
	0.5	102	(21.34)	43	(18.94)	145	(20.57)
	0.75	196	(41.00)	52	(22.91)	248	(35.18)
	1	4	(0.84)	39	(17.18)	43	(6.10)
C16(가구주직업)	0	131	(27.41)	42	(18.50)	173	(24.54)
	0.33	96	(20.08)	62	(27.31)	158	(22.41)
	0.66	228	(47.70)	65	(28.63)	293	(41.56)
	1	23	(4.81)	58	(25.55)	81	(11.49)
C28(월가구소득)	0	163	(34.10)	149	(65.64)	312	(44.26)
	0.5	220	(46.03)	55	(24.23)	275	(39.01)
	1	95	(19.87)	23	(10.13)	118	(16.74)
C32(맞벌이여부)	0	164	(34.31)	50	(22.03)	214	(30.35)
	1	314	(65.69)	177	(77.97)	491	(69.65)
E04(통신결합상품이용여부)	0	5	(1.05)	213	(93.83)	218	(30.92)
	1	473	(98.95)	14	(6.17)	487	(69.08)
E05(유선전화불만족여부)	0	448	(93.72)	219	(96.48)	667	(94.61)
	1	30	(6.28)	8	(3.52)	38	(5.39)
E06(통신요금)	0	418	(87.45)	85	(37.44)	503	(71.35)
	1	60	(12.55)	142	(62.56)	202	(28.65)
E07(인터넷전화사용의향)	0	438	(91.63)	218	(96.04)	656	(93.05)
	1	40	(8.37)	9	(3.96)	49	(6.95)
E09(현재유선전화이용여부)	0	28	(5.86)	136	(59.91)	164	(23.26)
	1	450	(94.14)	91	(40.09)	541	(76.74)
E10(현재초고속인터넷이용여부)	0	4	(0.84)	122	(53.74)	126	(17.87)
	1	474	(99.16)	105	(46.26)	579	(82.13)

표 7. TwoStep 군집의 변수 ANOVA 검증 결과
Table 7. ANOVA verification on TwoStep clustering

		제곱합	df	평균제곱	F	유의확률
A05(거주지역)	집단-간	.002	1	.002	.017	.896
	집단-내	89.363	703	.127		
	합계	89.365	704			
A12(성별)	집단-간	2.952	1	2.952	14.612	.000
	집단-내	142.018	703	.202		
	합계	144.970	704			
A13(나이)	집단-간	1.760	1	1.760	15.705	.000
	집단-내	78.802	703	.112		
	합계	80.562	704			
A15(주택형태)	집단-간	4.261	1	4.261	23.186	.000
	집단-내	129.203	703	.184		
	합계	133.465	704			
A17(가족유형)	집단-간	.162	1	.162	1.518	.218
	집단-내	75.128	703	.107		
	합계	75.290	704			
C14(응답자직업)	집단-간	.087	1	.087	.912	.340
	집단-내	66.800	703	.095		
	합계	66.887	704			
C16(가구주직업)	집단-간	1.710	1	1.710	16.495	.000
	집단-내	72.899	703	.104		
	합계	74.610	704			
C28(월가구소득)	집단-간	6.557	1	6.557	52.622	.000
	집단-내	87.597	703	.125		
	합계	94.154	704			
C32(맞벌이여부)	집단-간	2.322	1	2.322	11.126	.001
	집단-내	146.719	703	.209		
	합계	149.041	704			
E04(통신결합상품이용여부)	집단-간	132.506	1	132.506	5150.975	.000
	집단-내	18.084	703	.026		
	합계	150.590	704			
E05(유선전화불만족여부)	집단-간	.117	1	.117	2.287	.131
	집단-내	35.835	703	.051		
	합계	35.952	704			
E06(통신요금)	집단-간	38.482	1	38.482	256.081	.000
	집단-내	105.640	703	.150		
	합계	144.122	704			
E09(현재유선전화이용여부)	집단-간	44.970	1	44.970	390.876	.000
	집단-내	80.880	703	.115		
	합계	125.850	704			
E10(현재초고속인터넷이용여부)	집단-간	43.083	1	43.083	501.456	.000
	집단-내	60.398	703	.086		
	합계	103.481	704			

표 8. ANOVA 검증 결과 평균 비교
Table 8. Comparison of means for ANOVA verification

		A05	A12	A13	A15	A17	C14	C16	C28	C32	E04	E05	E06	E09	E10
1.00 (N=478)	평균	0.36	0.76	0.50	0.56	0.49	0.47	0.43	0.43	0.66	0.99	0.06	0.13	0.94	0.99
	표준편차	0.36	0.43	0.29	0.44	0.26	0.29	0.31	0.36	0.48	0.10	0.24	0.33	0.24	0.09
2.00 (N=227)	평균	0.37	0.62	0.60	0.39	0.52	0.49	0.53	0.22	0.78	0.06	0.04	0.63	0.40	0.46
	표준편차	0.36	0.49	0.41	0.41	0.43	0.34	0.35	0.34	0.42	0.24	0.18	0.49	0.49	0.50
합계 (N=705)	평균	0.36	0.71	0.53	0.51	0.50	0.48	0.46	0.36	0.70	0.69	0.05	0.29	0.77	0.82
	표준편차	0.36	0.45	0.34	0.44	0.33	0.31	0.33	0.37	0.46	0.46	0.23	0.45	0.42	0.38

노년가구가 가장 높게 나타났으며, 대부분 1인 가족으로 나타났다. 300만원 미만이 가장 높게 나타났으며 대부분 통신결합상품을 사용하지 않는 것으로 나타났다. 그리고 두 군집 모두 유선전화 불만족도가 낮은 것으로 나타났다.

4.2.5 TwoStep 군집화에 대한 ANOVA 검증 결과

표 7은 TwoStep 군집화의 결과 생성된 2개의 군집에 대해 집단간, 집단내 One-way ANOVA 분석 결과를 보여주고 있다. 분석의 결과, E04(통신결합상품이용여부)를 제외한 모든 변수들이 집단내보다 집단간이 더 높은 제곱값을 가지는 것으로 나타났다. 그리고 A05(거주지역), A17(가족유형), C14(응답자직업), E05(유선전화불만족여부)를 제외한 모든 변수들이 유의한 값을 나타내어 군집 간 차이가 있는 것으로 파악되었다. 표 8은 ANOVA 분석에 이용된 각 변수의 평균과 표준편차를 보여주고 있다.

군집분석은 고객세분화에 가장 보편적으로 사용하는 분석방법으로 본 연구의 목적을 달성하는데 가장 적합한 분석방법이다. 인터넷전화 사용 의향 예측 모형의 개발을 위해서 TwoStep 군집화로 얻어진 2개의 군집 중 유선전화에 대해 만족하면서 통신결합상품을 사용하는 30~40대로 구성된 군집-1을 사용하였다.

4.3 인터넷전화 사용 의향 예측 모형의 개발

군집화를 통해 세분화된 그룹별로 예측기법을 통하여 비즈니스 활동 결과를 예측하고 마케팅 전략을 도출하였다. 인터넷전화 사용 의향 예측을 위한 모형 개발을 위해서 사용되는 독립변수들은 고객군집화에 사용된 변수와 같으며 종속변수는 E07(인터넷 사용의향)을 사용하였다. 그리고 IBM 모델러의 예측기법 중 C5.0과 신경망 그리고 SVM을 사용하여 분석을 수행하였다.

본 연구는 이전 권은경 등의 연구에서 제기된 데이터 불균형 문제점을 해결하고자 예측 분석 전에 Oversampling을 통하여 대조집단과 실험집단의 비율을 맞추었다[23]. 표본 데이터에서 인터넷전화로 전환의향이 있는 집단과 그렇지 않는 집단 간에 수적 차이가 존재하는 것은 데이터 불균형의 사례인데 이러한 데이터 불균형 문제는 기계학습 알고리즘의 성능을 저하시키는 요인으로 이를 해결하기 위해 Shin and Cho는 오분류 패턴에 대해서 서로 다른 패널티를 부과하는 방법을 제시하였다[24]. 이 방법은 소수집단에 속한 데이터가 다른 집단으로 분류되는 것에 대해 패널티를 부과하는 것으로 Oversampling은 패널티 부과 방법 중의 하나인데 이를 통해 대조집단과 실험집단의 비율을 맞출 수 있다 [22]. IBM 모델러의 Balance 노드*를 사용하여 종속변수인 E07의 인터넷 전화를 사용하지 않겠다는 실패케이스(438개)와 사용하겠다는 성공케이스(40개) 중 성공케이스인 경우 가중치를 10.95로 주어서

* 레코드 옵션 중 하나로서 작은 범주의 레코드를 일정한 비율로 증가시키거나 큰 범주의 레코드를 일정한 비율로 감소시켜서 종속변수의 분포를 균등하게 처리하는 전처리 노드이다.

성공과 실패 케이스의 비율을 맞추는 Oversampling을 하였다. Oversampling 결과, 총 케이스 수는 877개가 되었고, Partition 노드*를 적용하여 학습 데이터 셋(603개)과 검증 데이터 셋(274개)을 7:3으로 분리하여 예측모형의 개발에 사용하였다.

4.3.1 C5.0

IBM 모델러에서 제공하는 의사결정나무 모형에는 CART, C5.0, QUEST, CHAID가 있으며, 이 모형은 종속변수가 범주형인 경우에 모두 사용할 수 있다. 그러므로 본 연구에서는 일반적으로 좋은 결과를 보여주는 C5.0 기법을 사용하여 예측 모형을 개발하였다. 그 결과, 선택된 중요 변수 중, 중요도가 0.1이상인 변수는 E09(0.138)로 나타났다. 성공케이스(E07=1)를 위해 형성된 규칙 셋은 총 9개이고, 첫 번째 규칙 셋에 따르면 현재 유선전화를 이용하고 있지 않는 고객이 인터넷전화를 사용할 의향이 있다는 규칙이 형성되었다. 이 규칙은 382개 케이스 중 46%인 174개의 지지를 받았다.

표 9과 표 10는 C5.0 알고리즘 수행 결과인 학습 데이터 셋과 검증 데이터 셋의 분류오류이다. 학습 데이터 셋의 총 예측정확도는 96.2%이며, 검증 데이터 셋의 총 예측 정확도는 91.97%를 보여주고 있다. 그리고 검증 데이터 셋에서 총 145건의 실제 성공 케이스(E07=1) 중에서 132건이 성공 케이스로 예측됨으로써 예측 정확도는 91.97%를 보여주고 있다. 그림 1은 C5.0의 결과에 대한 게인즈 차트(Gains Chart)이다.

표 9. 학습 데이터 셋 분류오류(C5.0)
Table 9. Classification errors on training data set(C5.0)

Actual Class \ predicted Class	0	1
0	294(48.79%)	11(1.82%)
1	13(2.16%)	285(47.26%)

* 필드 옵션 중 하나로서 데이터 셋을 학습 데이터 셋과 검증 데이터 셋으로 분리하는데 사용하는 노드이다.

표 10. 검증 데이터 셋 분류오류(C5.0)
Table 10. Classification errors on validation dataset(C5.0)

Actual Class \ Predicted Class	0	1
0	120(43.80%)	13(4.74%)
1	9(3.28%)	132(48.18%)

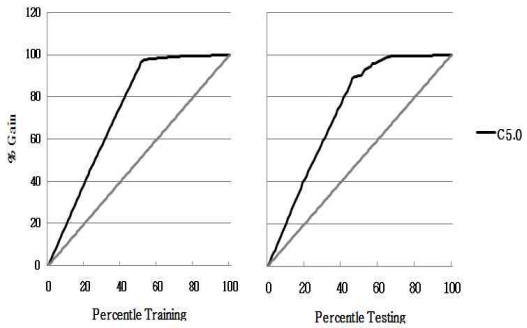

그림 1. C5.0 Gains Chart
Fig 1. C5.0 Gains Chart

4.3.2 신경망(Neural Net)

IBM 모델러의 신경망 노드의 다양한 신경망 모형 구조를 탐색해본 결과, 은닉층이 20뉴런으로 구성된 Prune 알고리즘의 경우가 가장 좋은 예측 효율을 보이고 있으므로 본 연구에서는 Prune을 채택하였다. Prune 알고리즘은 다층퍼셉트론(Multi-layer perceptron)에서 역전파(Backprogagation) 알고리즘으로 학습한 후 필요성이 떨어지는 뉴런을 제거하여 최적 모형을 찾도록 한 알고리즘으로, 역전파 알고리즘의 개선이라고 할 수 있다.

표 11은 신경망 모형의 파라미터 설정 값을 나타낸다. 신경망 알고리즘의 결과, 선택된 중요 변수 중 중요도가 0.1이상인 변수는 E09(0.28), E05(0.121), A13(0.103)이며, 표 12과 표 13는 학습 데이터 셋과 검증 데이터 셋의 분류오류이다. 학습 데이터 셋의 총 예측정확도는 98.01%이며, 검증 데이터 셋의 총 예측 정확도는 97.45%를 보여주고 있다.

표 11. 신경망 모형의 파라미터 설정 값
Table 11. Parameter values for neural network models

Parameters	
Method	Prune
Random seed	12345
Optimize	Memory
Hidden layers	1
Nodes in Hidden Layer-1	20
Hidden rate	0.15
Input rate	0.15
Persistence	100
Alpha	0.9
Initial Eta	0.3
High Eta	0.1
Eta Decay	30
Low Eta	0.01

표 12. 학습 데이터 셋의 분류오류(신경망)
Table 12. Classification errors on training data set(neural nets)

Actual Class \ Predicted Class	0	1
0	293(48.59%)	12(1.99%)
1	0(0.00%)	298(49.42%)

표 13. 검증 데이터 셋의 분류오류 (신경망)
Table 13. Classification errors on validation data set(neural nets)

Actual Class \ Predicted Class	0	1
0	126(45.99%)	7(2.55%)
1	0(0.00%)	141(51.46%)

그림 2는 신경망 알고리즘의 결과에 대한 게인즈 차트(Gains Chart)이다.

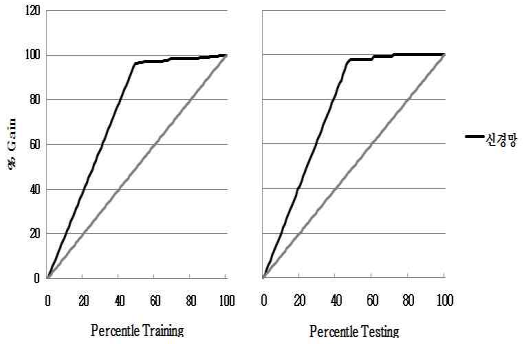

그림 2. 신경망 Gains Chart
Fig 2. Neural Nets Gains Chart

4.3.3 SVM(Support Vector Machine)

IBM 모델러의 SVM의 커널 유형에는 RBF, Polynomial, Sigmoid, Linear 4가지 방식이 있다. 이 중 RBF 커널함수는 종속변수가 이분변수(0/1)일 때 국지적 (local) 학습 시 종속변수 분류에 적합하다[25]. 본 연구에서는 종속변수가 이분변수이기 때문에 RBF 커널함수를 채택하였다. 표 14은 SVM 모형의 파라미터 설정 값을 나타낸다. SVM 알고리즘의 결과, 선택된 중요변수 중 중요도가 0.1이상인 변수는 E09(0.572), E05(0.148)이며, 표 15와 표 16는 학습 데이터 셋과 검증 데이터 셋의 분류오류이다. 학습 데이터 셋의 총 예측 정확도는 99.67%이며, 검증 데이터 셋의 총 예측 정확도는 98.54%를 보여주고 있다. 그림 3는 SVM 알고리즘 결과에 대한 게인즈 차트(Gains Chart)이다.

표 14. SVM 모형의 파라미터 설정 값
Table 14. Parameter values for SVM models

Parameters	
Stopping criteria	1.0E-3
Regularization parameter(C)	10
Regression precision(epsilon)	0.1
Kernel type	RBF
RBF gamma	0.1

표 15. 학습 데이터 셋의 분류오류(SVM)
Table 15. Classification errors on training dataset(SVM)

Actual Class \ Predicted Class	0	1
0	303(47.08%)	2(0.33%)
1	0(0.00%)	298(51.46%)

표 16. 검증 데이터 셋의 분류오류(SVM)
Table 16. Classification errors on Validation data set(SVM)

Actual Class \ Predicted Class	0	1
0	129(47.08%)	4(1.46%)
1	0(0.00%)	141(51.46%)

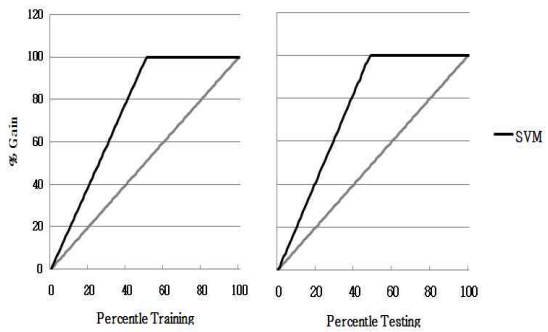

그림 3 SVM Gains Chart
Fig 3. SVM Gains Chart

4.3.4 모형성과 비교

인터넷전화로의 전환 의향 예측을 위해 구축된 세 가지 알고리즘의 20회 반복 수행 결과 중에서 예측 오류가 가장 낮게 나온 연구결과를 제시하였고 알고리즘 간의 성과를 비교하였다. 각 알고리즘의 예측오류율은 아래 표 17, 표 18, 표 19로 제시되어 있다. 전체 예측 오류율을 비교해 보았을 때, C5.0은 8.02%, 신경망은 2.55%, SVM은 1.46%로서 SVM 모형을 사용한 경우가 가장 성과가 좋은 것으로 나타났다. 그리고 표 20의 1종 오류(Type I)의 오류율을 비교하면 SVM 모형과 신경망이 가장 좋은 성과를 나타내고 있다. 1종 오류는 실제로는 인터넷전화를 사용할 의향을 가지고 있으나 그렇지 않은 고객으로 예측하는 경우를 의미한다.

2종 오류(Type II)를 비교하면 SVM 모형이 3.01%로서 가장 성과가 좋은 것으로 나타났다. 2종 오류는 실제로는 인터넷전화를 사용할 의향을 가지고 있지 않은 고객인데도 의향이 있는 고객으로 예측하여 불필요한 마케팅비용을 지불해야 하는 경우를 의미한다. 일반적으로 최종 모형을 선택할 때 1종 오류와 2종 오류, 그리고 전체 예측 오류율을 비교하고 비용적인 측면을 감안한다. 본 연구에서는 3가지 예측 오류율의 비교에서 SVM이 가장 좋은 성과를 나타내고 있기 때문에 SVM을 통해 전략적인 시사점을 도출하고자 한다. 다음의 그림 4와 그림 5는 세 가지 알고리즘에 대한 게인즈 차트(Gains Chart)를 비교해 놓은 것이다.

표 17. C5.0 분류오류
Table 17. C5.0 classification errors

Class	#Class	#Error	%Error
0	120	13	9.77%
1	132	9	6.38%
Overall	252	22	8.02%

표 18. 신경망 분류오류
Table 18. Neural network classification errors

Class	#Class	#Error	%Error
0	126	7	5.26%
1	141	0	0.00%
Overall	267	7	2.55%

표 19. SVM 분류오류
Table 19. SVM classification errors

Class	#Class	#Error	%Error
0	129	4	3.01%
1	141	0	0.00%
Overall	270	4	1.46%

표 20. 모형 성과 비교
Table 20. Comparisons of model performance

구분	C5.0	신경망	SVM
1종 오류	6.38%	0.00%	0.00%
2종 오류	9.77%	5.26%	3.01%
전체오류율	8.02%	2.55%	1.46%

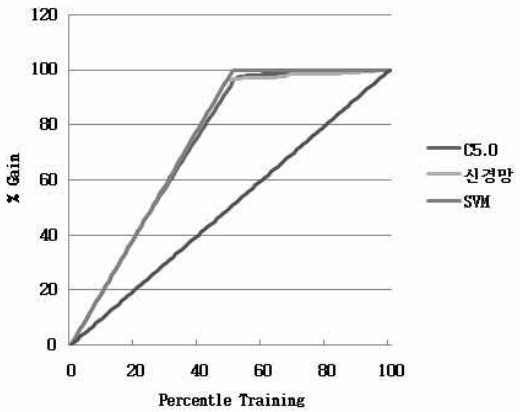

그림 4. 세 알고리즘의 학습 데이터 셋에 대한 Gains Chart
Fig 4. Gains Charts for three algorithms (training dataset)

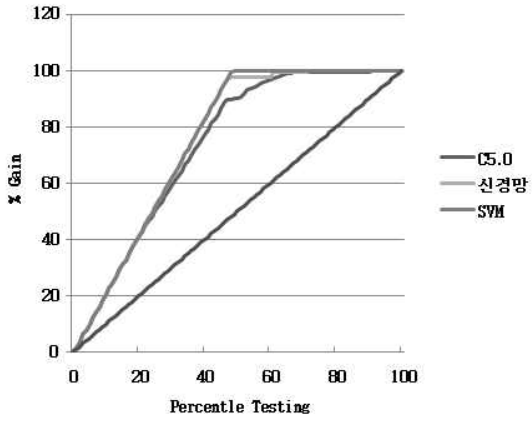

그림 5. 세 알고리즘의 검증 데이터 셋에 대한 Gains Chart
Fig 5. Gains Charts for three algorithms(validation dataset)

4.4 고객군집화와 예측모형으로 분석한 전략적 시사점

본 연구에서는 기존 유선전화 사용자들을 군집화 한 결과 2개의 군집이 형성되었으며 그 중에서 통신결합상품을 사용하고 자녀를 둔 30~40대의 고객들이 분포되어 있는 군집-1을 예측 분석에 사용하였다. 분석 결과, 최고의 성과를 보인 모형은 SVM이었으며 SVM 모형에서 선정된 변수들 중 E05(유선전화불만족여부), E09(현재유선전화이용여부)가 가장 중요도가 높은 변수로 나타났다.

먼저 표 21의 E05(유선전화불만족여부)와 E07(인터넷전화사용의향)의 관계를 살펴보면 유선전화에 불만족을 가지고 있는 고객들 중에서 인터넷전화 사용의향이 있는 경우가 그렇지 않은 경우보다 높게 나타남을 알 수 있다. 이는 유선전화에 불만족을 가진 사람들이 대체서비스로서 인터넷전화로의 전환을 고려하고 있음을 예상할 수 있다. 그러므로 이들 고객들을 대상으로 기존의 유선전화에 대해 불만족을 가지는 이유를 살펴보고 인터넷전화의 장점을 부각시켜 마케팅을 할 필요성이 있다. 그러한 요인들을 결합시켜 마케팅을 할 경우 통신사업자는 유선전화 이탈을 막고 자사의 인터넷전화 사용으로 이들 고객들을 유인할 수 있을 것이다. 본 연구에 사용된 설문에 응답한 고객들은 유선전화에 대해 불만족을 가지는 이유로 '통신요금이 비싸기 때문에', '유선전화가 불편하기 때문에' 라고 응답하였다. 최근에 인터넷전화 보급률이 증가하는 이유 중 하나가 저렴한 통신요금으로 알려져 있다. 따라서 유선전화에 불만족을 가지는 고객들에게 경제적인 측면을 부각시킨 마케팅 활동도 인터넷전화로 이들을 전환시키는데 효과적일 것이다.

다음으로 표 22의 E09(현재유선전화이용여부)와 E07(인터넷전화사용의향)의 관계를 살펴보면, 현재 유선전화를 사용하고 있지 않은 고객들은 유선전화를 사용하고 있는 고객들과 비교하여 인터넷전화 사용 의향

이 높게 나타났다. 이들은 현재 통신서비스를 사용하고 있는 고객들보다 새로운 통신서비스를 선택하고 받아들이는 것이 더 쉽다. 그러므로 이들을 대상으로 표적마케팅(target marketing)을 한다면 인터넷전화 신규고객을 획득하기위한 효과적인 방법이 될 수 있을 것이다.

표 21. 유선전화불만족여부에 따른 인터넷전화 사용의향
Table 21. Use intention on Internet telephone depending on wired telephone dissatisfaction

인터넷전화 사용의향(E07) / 유선전화불만족 여부(E05)	없음	있음
없음	420(57.8%)	307(42.2%)
있음	18(12.1%)	131(87.9%)

표 22. 현재유선전화이용여부에 따른 인터넷전화 사용 의향
Table 22. Use intention on Internet telephone depending on current use of wired telephone

인터넷전화 사용의향(E07) / 현재유선전화 이용여부(E09)	없음	있음
이용하지않음	4(1.5%)	264(98.5%)
이용하고있음	434(72.2%)	167(27.8%)

V. 결론

5.1 연구의 요약

본 연구에서는 설문에서 얻어진 인구통계학적 정보 및 통신서비스 정보에 대한 데이터와 데이터마이닝 기법을 사용하여 유선전화를 사용 중인 고객들을 군집화하고 특정 군집 사용자들의 인터넷전화 사용 의향을 예측하기 위한 모형을 개발하고자 하였다.

이를 위해 먼저 수집된 설문 데이터를 바탕으로 새로운 변수를 생성하였으며, 데이터 전처리를 수행하였다. 본 연구에서의 고객군집화를 위해서는 TwoStep과 K-means가 사용되었고, 예측을 위해서는 의사결정나무(C5.0), 신경망(Neural Net), SVM(Support Vector Machine) 기법이 사용되었다. 고객군집화에서는 TwoStep 군집화를 이용하여 군집의 수가 2개로 정해졌으며, TwoStep과 K-means 군집화에 대한 R-square 값을 통하여 두 가지 군집 알고리즘을 비교하였다. 그 결과, TwoStep 군집화가 더 높은 R-square 값을 나타내었고 군집화의 결과가 강건한지를 판단하기 위해 군집간, 군집내 ANOVA 검증을 수행하였다. 그리고 TwoStep 군집화 결과 얻어진 2개의 군집 중 군집-1에 대해 C5.0, 신경망, SVM 기법을 사용하여 전환의향을 예측하였다. SVM 모형의 예측 오류율이 0.01%로 최고의 성과를 보여주어 SVM 모형을 최적의 모형으로 평가하였다. SVM 모형에서 선정한 변수들 중 E05(유선전화불만족여부), E09(현재유선전화이용여부)가 가장 중요한 변수로 나타났다. 이 두 변수와 E07(인터넷전화 사용의향)간의 관계를 살펴보면서 유용한 마케팅 전략을 도출할 수 있었다. 유선전화에 불만족을 가지는 고객들을 인터넷 전화로 전환시키기 위해서는 비용적인 측면을 부각시킨 마케팅 활동을 하는 것이 효과적이다. 그리고 인터넷전화 신규 고객을 획득하기 위해서는 유선전화를 사용하고 있지 않는 고객들을 대상으로 표적마케팅을 하는 것이 효과적인 방법이 될 수 있다는 것이다.

5.2 연구의 한계점 및 향후 계획

본 연구의 한계점은 다음과 같다 첫째, 본 연구에서 인터넷전화 사용 예측을 위한 인구 통계적 변수와 새로 생성한 변수 외에 인터넷전화 사용 의향을 나타내는 다양한 변수의 개발이 부족하였다. 둘째, 신경망에서 은

닉층을 1개로 하고 은닉노드를 20개로 사용하였는데, 은닉층의 개수와 은닉노드의 수를 변화시키는 방법을 통해 좀 더 우수한 성능을 보이는 모형을 선택해야 하나, 이를 체계적으로 수행하지 못했다. 셋째, 본 연구에서 사용한 설문 데이터의 인구통계학적 변수들은 각 군집에 속한 고객들의 특성을 나타내는데 한계가 있는 것으로 확인되었다.

이러한 한계점을 바탕으로 향후 연구에서는 TwoStep 군집화 결과로 얻어진 모든 군집에 대상으로 예측을 수행하고 그 결과를 비교하는 것에 대한 논의가 필요하겠다. 또한 고객군집화를 위한 TwoStep과 K-means 군집화 결과에서 비슷한 특성을 나타내는 군집 간에 인터넷전화 사용 의향에 관한 예측 결과를 비교해보는 것에 대한 논의도 필요하겠다. 신경망에서 은닉층과 은닉노드의 수를 변화시켜보는 추가적인 연구가 필요할 것으로 생각되며 각 군집에 속한 고객들의 특성을 대표할 수 있는 변수와 유선전화 만족 및 불만족 의사 및 인터넷전화 사용 의향을 대표할 수 있는 변수의 개발과 관련성을 비교할 수 있는 추가 연구가 필요할 것으로 생각한다. 그리고 본 연구의 전략적 시사점이 미치는 영향에 대해 추가적인 모델을 개발하고 제시할 필요가 있다.

참고문헌

[1] KISDI, *Competitive circumstance assessment report (Consumers) for telecommunications market(2009)* , Korea Information Society Development Institute, 2009.

[2] Y. I. Gong, “Internet telephony of diffusion and implications for communications market,” *Broadcasting and Communication Policy*, Vol. 21, No. 5, pp.39-58, 2009.

[3] T. G. Kang, D. Y. Kim, and Y. S. Kim, “The Trend of Internet Telephony(VoIP) Technology for BcN,” *Electronics and Telecommunication Trends*, Vol. 19, No. 6, pp.66-73, 2004.

[4] T. H. Hong, and S. Y. Jeon, “Using Data Mining customer retention rating based on customer segmentation,” *Korea Intelligent Information System Society*, Vol. 2, No. 1, pp.189-198, 2006.

[5] B. Y. Lee, K. H. Joh, S. I. Song, and J. S. Yoo,“Design of a Forecasting Model for Customer Classification in the Telecommunication Industries,” *The Korea Contents Society*, Vol. 6, No. 1, pp.180-190, 2006.

[6] C. Wei, and I. Chiu, “Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach,” *Expert Systems with Applications,* Vol. 23, No. 2, pp.103-112, 2002.

[7] J. K. Kim, “Mobile phone operators of churn strategics,” *Information & communications policy*, Vol. 10, No. 7, pp.19-36, 1998.

[8] R. Siber, “Combating the Churn Phenomenon,” *Telecommunications,* Vol. 31, No. 10, pp.77-80, 1997.

[9] J. Y. Lee, and J. W. Kim, “실시간 CRM을 위한 분류 기법과 연관성 규칙의 통합적 활용 신용카드 고객 이탈 예측에 활용,” *한국경영정보학회학술대회*, pp.135-140, 2007.

[10] S. H. Lim, and Y. Hur, “Customer Churning Forecasting and Strategic Implication in Online Auto Insurance using Decision Tree Algorithms,” *Information systems Review*, Vol. 8, No. 3, pp.125-134, 2006.

[11] D. H. Cho, B. S. Kim, K. H. Seok, J. U. Lee, J. S. Kim, and S. H. Kim, “A study on the behavior of cosmetic customers,” *Journal of the Korea Data & Information Science Society*, Vol. 20, No. 4, pp.615-627, 2009.

[12] C. Wei, H. T. Chang, and Y. H. Lee, “Telecommunications Data Mining for Target Marketing,” *Journal of Computers*, Vol. 12, No. 4, pp.60-74, 2000.

[13] S. Daskalaki, I. Kopanas, M. Goudara, and N. Avouris, “Data Mining for Decision Support on Customer Insolvency in Telecommunications Business,” *European Journal of Operational Research*, Vol. 145, No. 2, pp.239-255, 2003.

[14] S. Y. Hung, D. C. Yen, and H. Y. Wang, “Applying Data Mining to Telecom Churn Management,” *Expert Systems with Application*, Vol. 31, No. 3, pp.512-524, 2006.

[15] B. H. Chu, M. S. Tsai, and C. S. Ho, “Toward a Hybrid Data Mining Model for Customer Retention,” *Knowledge-Based Systems*, Vol. 8, No. 8, pp.703-718, 2007.

[16] H. C. Ahn, J. J. Ahn, K. J. Oh, and D. H. Kim, “Facilitating

Cross-Selling in a Mobile Telecom Market to Develop Customer Classification Model based on Hybrid Data Mining Techniques," *Expert Systems with Applications*, Vol. 38, No. 5, pp.5005-5012, 2011.

[17] H. Farvaresh, and M. M. Sepehri, "A Data Mining Framework for Detecting Subscription Fraud in Telecommunication," *Engineering Application of Artificial Intelligence*, Vol. 24, No. 1, pp.182-194, 2011.

[18] J. Han, and M. Kamber, *Data Mining Concept and Techniques*, Morgan Kaufmann Publishers, 2000.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Belmont, Wadsworth, 1984.

[20] Z. B. Jonathan, "Market Segmentation a Neural Network Application," *Annuals of Tourism Research*, Vol. 32, No. 1, pp.93-111, 2005.

[21] R. J. Eyden, *The Application of Neural Networks in the Forecasting of Share Price*, Finance and Technology Publishers, 1996.

[22] G. Shmueli, N. R. Patel, and P. C. Bruce, *Data Mining for Business Intelligence*, Wiley, 2010.

[23] E. K. Kwon, S. H. Ha, and H. Y. Park, "고객선호 분석을 통한 인터넷전화로의 전환 의향 예측", 2010년 대한산업공학회/한국경영과학회 춘계공동학술대회, pp.150-157, 2010.

[24] H. J. Shin, and S. Z. Cho, "Response Modeling with Support Vector Machines," *Expert Systems with Applications*, Vol. 30, No. 4, pp.746-760, 2006.

[25] M. J. D. Powell, "The Theory of Radial Basis Function Approximation in 1990," *In Advances in Numerical Analysis*, Vol. 2: Wavelets, *Subdivision Algorithms and Radial Functions*, pp.105-210, Oxford University Press, Oxford, UK, 1990.

저자소개

하성호 (Sung Ho Ha)

2001년 한국과학기술원 박사

현재 경북대학교 경영학부 교수로 재직

※ 관심분야 : 데이터마이닝, e비즈니스, 지식경영, 의료경영, 지능의사결정

권은경 (Eun-Kyoung Kwon)

2009 계명대학교 컴퓨터공학과

2011년 경북대학교 대학원 경영학부 (경영학석사)

2011년~현재 경북대학교 대학원 경영학부 박사과정

※ 관심분야 : 데이터 마이닝, SNS, 마이크로 블로그 등

박현선 (Hyun-Sun Park)

2007 영남대학교 불어불문학과

2011년 경북대학교 대학원 경영학부 (경영학석사)

2011년~현재 경북대학교 대학원 경영학부 박사과정

※ 관심분야 : 모바일 서비스, 클라우드 컴퓨팅, IT adaption, 데이터 마이닝 등

임광혁 (Kwang Huk Im)

2006 한국과학기술원 산업공학(공학박사)

2006~2008 삼성전자(주) 반도체연구소 책임연구원

현재 배재대학교 전자상거래학과 교수

※ 관심분야 : 경영정보시스템, 데이터마이닝, 전자상거래, 고객관계관리, 공급사슬관리, 지식서비스