인터넷전화(VoIP)의 신규고객 유치를 지원하는 데이터마이닝 모델

하 성 호* · 양 정 원** · 송 영 미***

A Data-Mining Model to Support new Customer Acquisition for Internet Telephony(VoIP)

Sung Ho Ha* · Jeong Won Yang** · Young Mi Song***

Abstract

Recently, Internet Telephony has become increasingly popular in telecommunication industry. However, previous research on Internet Telephony has focused on analyzing specific Internet Telephonysolutions, identifyingthe Internet Telephony movement itself. The research on prediction models about Internet Telephony adoption has been minimal. The main propose of this study is to develop models for predicting transition intention from using traditional telephones to using Internet Telephony. To do so, this study uses data mining methods to analyze demands in the IT communications market and to provide management strategies for Internet telephony providers. Especially this study uses discriminant analysis, logistic regression, classification tree, and neural nets to develop those prediction models toward Internet Telephony adoption. The models are compared with each other and a superior model is chosen.

Keywords : Internet Telephony, Data Mining, Adoption Prediction

논문접수일 : 2010년 04월 22일 논문게재확정일 : 2010년 06월 11일

* 경북대학교 경영학부 교수, e-mail : hsh@mail.knu.ac.kr
** 경북대학교 일반대학원 경영학부 박사과정
*** 경북대학교 일반대학원 경영학부 박사과정

1. 서 론

정보통신기술이 빠른 속도로 발전함에 따라 통신 서비스도 나날이 진보하고 있다. 차별화되고 진보된 통신서비스가 기존 서비스를 대체하는 형태로 늘어나고 있는 것이다. 그 중 인터넷전화는 기존의 유선전화를 대체할 수 있는 가장 대표적인 서비스로 향후 발전과 이용 확산이 예견된다.

유선전화(Plain Old Telecommunication Service : POTS)란 가입자와 교환국 사이의 유선선로 (가입자망)를 통해 시내전화(통화권내), 시외전화(통화권간), 국제전화, 공중전화가 가능한 통신서비스이다. 인터넷전화(Voice over Internet Protocol : VoIP)는 기존의 유선전화 기능을 제공하면서도 요금이 저렴하여 유선전화를 대체할 수 있는 잠재력을 보유한 서비스이다[권오상 외, 2002].

특히 최근 사업자들의 품질개선 노력과 인터넷전화로의 번호이동이 간편해지면서 인터넷전화에 대한 인지도가 향상되고 유선전화에 대한 대체가능성이 높아지고 있다. 또한, 인터넷전화는 물리적 네트워크에 종속되지않고, 통화품질도 우수할 뿐 아니라 초고속 인터넷의 높은 보급률에 힘입어 전망이 아주 밝다고 할 수 있다[함창용 외, 2007].

KISDI 조사에 따르면, 국내 유무선 및 인터넷전화 서비스 가입현황은 유선전화 72%, 이동전화 97%, 인터넷전화가 21%이며, 인터넷전화 서비스 가입자 중에서 유선전화를 해지한 고객은 81%에 이르는 것으로 조사되었다[KISDI, 2009]. 유선전화 해지 사유로는 '인터넷전화의 통화요금이 저렴해서(75%)' 또는 '결합요금제 때문에(18%)'라는 응답비중이 높아, 유선전화에서 인터넷전화로의 전환에 비용이 중요한 역할을 하는 것으로 분석되었다.

또한, 인터넷전화와 유선전화에 동시 가입한 사용자 중, 70%가 유선전화를 해지하겠다고 응답한 반면, 인터넷전화를 해지하겠다는 응답은 27%에 불과한 것으로 나타났다. 이것은 인터넷전화가 기존 유선전화를 대체할 수 있는 서비스로서의 가능성을 보여주며, 기존 유선전화에서 인터넷전화 서비스로 전환하도록 고객을 유인할 수 있는 서비스 개발이 필요함을 시사한다.

인터넷전화는 저렴한 요금 뿐만 아니라 다양한 부가서비스와 영상통화를 통하여 새로운 시장을 창출할 것으로 기대되고 있다. 특히 와이브로를 비롯하여 무선 인터넷이 활성화되면, 유무선 융합의 비즈니스 모델을 통해 유비쿼터스 환경에서 다양한 인터넷전화 서비스를 제공할 수 있을 것으로 기대된다[김문구 외, 2008]. 그러므로 유선 인터넷전화의 효율적인 확산을 위해서는고객의 특성을 파악하고 적절한 비즈니스전략을 수립하는 것이 필요하다.

인터넷전화에 대한 수요가 급증하고, 기존 유선전화를 대체할 수 있는 서비스로서 가능성이 커지고 있으나 현재 인터넷전화에 대한 문헌연구나 실증적 분석은 미미한 실정이다. 인터넷전화에 대한 대부분의 기존 연구들은 기술동향[강태규 외, 2004], 제도 및 정책연구[김도환, 2009; 권오상 외, 2002], 관련 기술에 대한 개선 방안[송관호, 2007] 등이다. 이에 본 연구에서는 기존의 유선전화에서 인터넷전화로의 전환을 유도할 수 있는 요인들을 도출하는 인터넷전화 사용의향 예측모델을 개발하고자 한다.

또한, 정보통신 기술의 급속한 발달에 따라 통신산업의 비용예측과 합리적인 통신기술 수요예측의 필요성이 제기되고 있다. 본 연구에서는 인터넷전화 도입에 관한 예측모형을 개발하기위해 판별분석(Discriminant analysis), 로지스틱 회귀분석(Logistic regression), 분류나무(Classification tree), 신경망(Neural network) 알

고리즘을 비교 분석하고 가장 우수한 예측모델을 제시한다.

2. 문헌 연구

2.1 인터넷전화

인터넷전화의 핵심 기술인 VoIP는 PSTN(Public Switched Telephone Network)을 통해 이루어졌던 음성전달 서비스를 IP(Internet Protocol)를 사용하여 제공하는 것으로 음성은 물론 팩스, 웹콜, 통합 메시지, 화상회의 등의 서비스가 가능한 기술이다[박대우 외, 2006]. 또한, 인터넷전화는 음성신호를 주고받기 위해 IP와 관련된 표준과 포맷을 사용하는 시스템이라고 정의되기도 한다[Leddy, 2005]. 국제전기통신연합(International Telecommunication Union)은 공중망을 이용하여 음성을 송수신하는 모든 경우를 포괄하여 인터넷전화라고 정의하였다.

해외 인터넷전화 시장 현황을 살펴보면, 인터넷전화 서비스 제공사업자 중 가장 규모가 큰 Skype는 2008년 말 기준, 4억 5백만이 넘는 가입자와 551백만 달러의 매출을 기록하였다[이주영, 2009]. 또한, Lin[2003]의 연구에서는 2008년에 인터넷전화 점유율이 55%를 상회할 것으로 예상한 바 있다. 북미의 경우, 인터넷전화 판매율이 2002년에 4억 명을 넘어섰고 인터넷전화에 필요한 장치 구매가 2006년에 2001년의 6배에 이를 것이라고 예측한 바 있다. 또, World IT Report[2003]는 호주, 홍콩, 한국, 싱가포르에서 인터넷전화가 증가할 것이며 인도가 차세대 시장이 될 것으로 예상하였다.

Ovum[2007]은 VoIP가 확산됨에 따라 유선전화의 수익이 감소할 것이며 2010년에서 2011년 사이에 호주의 음성전화 요금을 낮추는 요인이 될 것으로 전망하였다. 또한, 2011년까지 VoIP의 수익은 업무용이 18%, 가정용이 21%까지 증가하여 가정용 시장에서 더 많은 비중을 차지할 것으로 예상하였다. ACMA[2007]에 따르면, 호주의 VoIP 시장은 2007년 4월에서 9월 사이 사업자 수가 27개에서 269개로 늘어 매우 빠르게 성장하였으며, 다양한 비즈니스 모델과 일반전화보다 저렴한 가격을 적용하여 시장에서 경쟁력을 가지고 성장하고 있는 것으로 나타났다.

2.2 통신서비스의 데이터마이닝 적용

Ezawa and Norton[1995]은 통신 서비스 고객의 연체 및 부정사용 유형 분류에 베이지안을 근거로 네트워크모델을 개발하였다. 또한, 이홍재 외[2000]는 국내 통신 서비스 시장에 적합한 신규고객 예측방법론을 개발하였다. 이 연구에서 이동전화 시장에서의 선택적 고객이전(selective migration)에 관한 모형을 제시하여 신규 통신서비스의 시장 잠재력에 대한 예측방법을 제시하였다. 또한, 신규 통신서비스의 최적도입시기에 관한 예측분석에 유사 서비스의 경험이나 경제이론에 기초한 베이지안확산 모형을 적용하였으며, 특정 시기별로 목표 소비자그룹을 찾아내고 각 시기에 환경변수(요금, 사업자수 등)가 평균적인 소비자의 가입확률에 미치는 영향을 보여주었다. Lingfen and Ifeachor[2006]는 비선형 회귀모델을 개발하여 VoIP의 음성 품질을 측정하였으며 실제 인터넷 VoIP 추적 데이터를 활용하여 개발된 모델이 높은 예측 정확도를 나타냄을 보였다.

이동통신 회사의 고객관리에 데이터마이닝 기법이 적용된 경우를 많이 찾아볼 수 있다. 이극노 외[2003]의 연구에서는 의사결정나무(C4.5)와 신경망기법을 이용하여 이동통신 회사의 고객을 분류하고 패턴을 분석함으로써 우수 고객과이탈 가능 고객을 차별 관리하고 회사의 이익을 극대화할 수 있을 것으로 밝히고 있고, 이병엽 외[2006]의 연구에서는 이동통신업체의 데이터를

이용하여 고객을 세분화한 후 각 고객군에 맞도록 차별 서비스를 제공함으로써 기존 고객의 충성도를 높이고 신규 고객의평생 고객화를 촉진할 수 있다고 말하고 있다.

2.3 인터넷전화의 CRM 적용

통신고객의 서비스에 대한 요구와 기대가 점점 다양해지고 세분화됨으로써 전화서비스 제공업자의 기존 유선전화 고객유지와 신규 인터넷전화 가입자 유치에 부단한 노력과 고도의 전략이 요구되고 있다[Pine et al., 1995; Rinde, 1999].

Kohli et al.[2001]의 연구에서 고객관계관리(Customer relationship management : CRM) 란 고객과의 장기적 관계를 유지하기 위해 제품 또는 서비스제공자가 고객의 요구에 맞춰 지속적인 변화를 추구하고 고객 기대에 부응하는 프로세스라고 정의하였다. Massey et al.[2001]의 연구에서는 CRM이 성공적인 고객관계의 개발과 유지를 포함하는 것으로 정의하였다. Peppers et al.[1999]은 고객에 대한 정보를 기반으로 개별 고객의 기대에 맞춰 기업의 행동을 변화시키려는 의지와 능력이라고 정의하였다. VoIP 서비스제공업자들은 전통적인 전화이용 고객을 적극적으로 유치하기 위해 CRM을 도입/개발/활용한다. 이는 미래에 VoIP가 기존의 유선전화 시스템을 완전히 대체할 수 있을 것이라고 기대하기 때문이다[Zhang et al., 2005].

VoIP와 CRM을 결합함으로써 운영비용절감과 고객 서비스 강화라는 두 가지 목표를 성취할 수 있다. VoIP 서비스제공자들은 고객 경험을 향상시키고, 기업과 고객간에 더 좋은 관계를 형성하기 위해 CRM을 고객 서비스에 포함시킨다[Kabiraj, 2003]. CRM이 VoIP 서비스제공자들에게 가져다 주는 이점은 고객만족을 증대시킴으로써 고객 유지와 충성도를 향상시키는 것이다[Zhang et al., 2005; Oracle, 2005]. 이를 통해 기업은 장기적인 수익을 안정적으로 얻고 산업 내에서의 경쟁력을 강화시킬 수 있게 된다.

Buyut et al.[2008]의 연구에서는 VoIP 서비스에서의 CRM 도전과제를 설명하고 CRM의 이점과 영향력을 강조하였다. 또한, Tseng[2005]의 연구에서는 고객 선호도 분석을 통해 제품, 가격 전략을 도출하였다. 인터넷전화의 주요 속성(음성전송품질, 통화료, 통화시간, 통화 중(line busy) 문제, 장치 속성)을 도출하였고 '음성전송품질'과 '통화 중 문제'를 우선 해결해야 한다는 제품전략을 제시하였다.

인터넷전화의 전략적 관점을 제시한 연구도 있었는데, 장범진 외[2006]의 연구에서는 수직적 차별화 및 기술경쟁차원에서 인터넷전화시장 발전을 위한 로드맵을 제시하고 융합환경에 적합한 규제체계 개선방향에 대하여 고찰하였으며, 인터넷전화와 기존유선전화 간의 차별화된 번호정책을 제시하였다. Jung[2004]은 신기술을 비용절감 기술과 혁신기술의 두 가지 유형으로 분류하고 인터넷전화가 단기적으로는 비용절감기술의 특성을 보이지만 장기적으로는 혁신기술의 특성을 가진다고 주장하였다. 기존 인터넷 망에서 저렴한 음성서비스를 단순 제공하는 것이 초기 인터넷전화 서비스의 전략이지만, 장기적으로는 인터넷 망의 고도화, 다양한 부가서비스와의 결합 등을 통하여 인터넷전화의 서비스 품질을 향상시키고 차별화하기 위한 전략 수립이 필요하다는 지적이다.

또한, 박종현 외[2008]의 연구에서는 국내 유선 VoIP 이용자를 현재이용자, 잠재이용자, 비이용자로 세분화하여 비교하였고 인터넷전화를 이용하는 이유와 이용하지 않는 이유, 이용조건, 지불의사 수준을 파악하여유선 VoIP가 활성화되기 위한 전략을 제시하였다.

본 연구에서는 기존 유선전화 서비스를 해지할 가능성이 높은 고객을 탐지하여 데이터마이닝 기법을 적용하고 이탈고객 또는 전환고객을 예측하여 VoIP 서비스 제공자들의 인터넷전화 서비스로 새로운 고객을 유치할 수 있는 적절한 고객유치전략을 도출한다. 인터넷전화 사용의향을 예측하는 모델 개발은 경쟁이 치열하고 고객의 해지가 빈번한 통신서비스 분야에서 고객 관리에 중요한 시사점을 제공한다. 본 연구에서는 인터넷전화 서비스를 제공하는 기업의 실제 데이터를 이용하여 인터넷전화 사용의향 예측모델을 개발하고 유선전화에서 인터넷전화로의 전환고객에 대한 CRM 전략을 도출할 수 있는 기반을 제공한다.

2.4 예측용 데이터마이닝 기법

본 연구는 설문 자료를 기반으로 인터넷전화 사용의향 예측모델 개발을 목적으로 한다. 설문 자료의 특성상, 모든 자료는 숫자로 구성된 범주형이고 종속 변수는 인터넷전화 사용 의향이 있고 없음을 나타내는 이진형 범주로 나타낼 수 있다. 따라서 범주형 설명 변수와 이진형(Binary) 범주의 종속 변수를 가진 자료를 분석하는 알고리즘은 예측 목적용으로 로지스틱 회귀분석, 분류나무, 신경망 등이 있다. 본 연구에서 사용하게 될 또 하나의 알고리즘인 판별분석의 경우 연속형 설명 변수에 대한 분석 방법이지만 숫자로 구성된 범주형 설명 변수의 경우 적용이 가능하다[Shmueli et al., 2007].

로지스틱 회귀분석은 선형 모형의 단점을 극복하기 위해 종속 변수가 이진형 범주일 때 사용하는 특수 형태의 회귀 분석 알고리즘으로 회귀식을 통해 종속변수의 귀속 확률을 계산하여 예측을 하거나 통계적 추론을 하는 알고리즘이다[Bradley et al., 1998]. 판별분석 알고리즘은 관찰된 데이터를 근거로 판별식을 산출하여주어진 상황에서 판별식의 산출 값에 따라 응답자들의 행동을 두 개 이상의 그룹에 각각 속하도록 예측하는 것으로[정충영 외, 1998] 집단간 공분산의 동일성 등의 조건을 전제로 하지만, 로지스틱 회귀분석과는 달리 두 개 이상의 그룹에 대해 한 개의 판별식으로 예측이 가능하기 때문에 신용 평가나 고객 예측 등의 많은 분야에서 이용되어 온 알고리즘이다.

분류나무 알고리즘은 수집된 데이터를 분석하여데이터 사이에 존재하는 패턴의 분류 규칙을 나무의 형태로 만드는 것으로 신용 예측이나 고객 평가에서 널리 사용되는 기법이다[하성호 외, 2009]. 고객 분류나 신용 예측 규칙을 명시적인 규칙 셋(Rule Set)으로 정의하고 있기 때문에 의사결정에 대한 근거를 설명하기 쉽다는 장점이 있다. 신경망 알고리즘의 경우는 통계적 가설 없이 비선형 회귀 모델을 설명하기에 적당해서 고객 신용 예측에서 뛰어난 결과를 보여주고 있다. 이 밖에 기존의 통계적인 알고리즘들과 데이터마이닝 알고리즘들을 혼합하여 구현한 연구들이 늘어나고 있는데 대표적으로 Lee et al.[2002]의 연구에서는 판별분석에서 선택된 변수를 신경망의 입력변수로 사용하는 혼합 모형을 구현할 경우 신경망만으로 구성된 단일 모형보다 예측률이 더 좋았다고 밝히고 있다. 따라서 본 연구에서는 판별분석, 로지스틱 회귀분석, 분류나무, 신경망(단일, 혼합) 모델을 이용하여 예측모델을 구현하고 성능을 비교 평가한다.

3. 연구모델 개발과 결과 해석

3.1 데이터 분석

본 연구에서 분석한 자료는 국내 모 통신사가

서울과 광역시의 시민들을 대상으로 2008년 8월 실시한 기초조사(집전화를 기반으로 한 추가 서비스 개발) 에서 수집된 것으로 설문에 응한 고객의 인구 통계학적 정보와 고객들이현재 이용하고 있는 통신 기술(유선전화, 초고속인터넷, 인터넷전화, 이동전화)에 대한 정보, 향후 이용할 계획이 있는 통신 기술에 대한 정보를 담고 있다. 설문에 응한 고객의 수는 총 705명이며 모든 설문은 범주형 응답으로구성되어있다.

설문 자료를 기반으로 예측모델을 개발하기 위해 인터넷전화 향후 사용 의향을 나타내는종속변수를 생성하였다. 향후 인터넷전화를 이용할 의향이 있는 고객이란 현재 유선전화를 사용하지 않고 인터넷전화를 사용하고 있는 고객이거나 향후 이용할 계획을 가지고 있는 고객, 현재 유선전화를 사용하고 있지만 추가로 인터넷전화를 사용할 계획을 가지고 있는 고객으로 정의하였다.

현재 유선전화를 사용하지 않고 인터넷전화를 사용하고 있는 고객이거나 향후 이용할 계획을 가지고 있는 고객 수는 36건, 그리고 현재 유선전화를 사용하고있지만 추가로 인터넷전화를 사용할 계획을 가지고 있는 고객은 21건으로 조회되어 총 705건 중 57건이 인터넷전화 사용에 긍정적인 것으로 조회되었다. 인터넷전화 사용의도를 나타내기 위해 생성된 변수를 'D7'으로 명명하고 인터넷전화 사용 의도를 가지면 '1'의 값을, 사용할 의도를 가지지 않으면 '0'의 값을 부여하였다.

<표 1>은 'D7' 변수의 분포를 나타내는데 첫 열의 '있음'과 '없음'은 각각 '사용의도가 있음'과 '사용의도가 없음'을 나타낸다. 자료 분석에는 SPSS 14.0 한글 버전이 이용되었다. <표 1>에서 보듯이 연구에 사용된 자료에서 인터넷전화 사용 의도를 가진 고객(성공 케이스)의 수가 전체 자료의 8.1% 정도를 차지하고 있음을 알 수 있다.

〈표 1〉 종속변수(D7)의 분포

	빈도	퍼센트	유효 퍼센트	누적 퍼센트
없음	648	91.9	91.9	91.9
있음	57	8.1	8.1	100.0
합계	705	100.0	100.0	100.0

본 연구에 사용된 자료에는 응답자의 인구통계학적 정보가 포함되어 있는데 이것은 고객을 이해하기 위한 기본 정보로서 효과적인고객관계관리를 연구한 기존의 많은 문헌에서 입력 변수로 이용하고 있다[정수미 외, 2005]. 따라서 설문에서 얻어진 다양한 인구통계학적 정보를 인터넷전화 사용 의도를 예측하기 위한 모델을 구축하는데 이용한다. 획득된 인구통계학적 정보는 성별, 나이, 거주 형태(단독주택, 아파트 등), 가구 형태(1인 가족, 4인 가족), 직업, 월 가구 소득, 맞벌이 여부 등이다. 모든 응답은 범 주형 유형으로 구성되어 있다. 인구통계학적 변수와 인터넷전화 사용 의향 변수 간 관계를 교차표(Cross-tabulation)를 통해 나타내보면 <표 2>~<표 7>과 같다.

<표 2>는 성별과 인터넷전화 가입 의향과의 관계를 보여주고있는데 응답자가 여자인 경우 인터넷전화 가입 의향을 높게 보여주고 있으나 응답자의 다수가 여자이기 때문에 성별이 인터넷전화 가입 의향에 유의하게 영향을 미친다고는 할 수 없다.

연령별 인터넷전화 가입 의향에 있어서는 30

〈표 2〉 성별에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
A12	남자	188(92.2%)	16(7.8%)	204
	여자	460(91.8%)	41(8.2%)	501
전체		648	57	705

〈표 3〉 연령에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
A13	20대	98(93.3%)	7(6.7%)	105
	30대	216(88.5%)	28(11.5%)	244
	40대	169(92.5%)	14(7.5%)	183
	50대 이상	165(95.4%)	8(4.6%)	173
전체		648	57	705

〈표 4〉 거주형태에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
A15	단독주택	244(93.1%)	18(6.9%)	262
	다세대/연립	149(87.1%)	22(12.9%)	171
	아파트	255(93.8%)	17(6.2%)	272
전체		648	57	705

대와 40대가 다른 연령대의 응답자들에 비해 비교적 높은 사용 의향을 지니고 있음을 알 수 있다<표 3>. 이는 30대와 40대의 경우 대부분 독립된 가정을 이루고 있으며 더 높은 연령대에 비해 새로운 통신 기술에 대한 관심과 이해가 높은 집단이라는 점에서 연령과 인터넷전화 가입 의향 간 상관성을 보여주는 결과라고 할 수 있다.

<표 4>의 거주 형태와 인터넷전화 가입 의향 간 관계에 있어서 다세대 주택이나 연립에 살고 있는 응답자가 비교적 높은 가입 의도를 표시하고 있으며, 가구 형태에 있어서는 <표 5>에서 나타난 것처럼 1인 가구와 노년 가구를 제외하고 비슷한 분포를 보이고 있다.

직업과 인터넷전화 가입 의향간의 관계에 있어서는 가입 의향이 있음을 밝힌 응답자가 직업별로 큰 차이가 있다고 보여지진 않으며 <표 6>, 월 가구 소득에 따라서도 인터넷전화 가입 의향이 크게 차이가 나지 않는다고 보여진다 <표 7>.

〈표 5〉 가구형태에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
A17	1인 가구	97(96.0%)	4(4.0%)	101
	신혼가구	113(90.4%)	12(9.6%)	125
	유아자녀 가구	110(87.3%)	16(12.7%)	126
	초/중/고등자녀	117(91.4%)	11(8.6%)	128
	대학생/성인자녀	114(91.2%)	11(8.8%)	125
	노년 가구	97(97.0%)	3(3.0%)	100
전체		648	57	705

〈표 6〉 직업에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
C14	자영업	119(92.2%)	10(7.8%)	129
	블루칼라	126(90.0%)	14(10.0%)	140
	화이트칼라	134(92.4%)	11(7.6%)	145
	가정주부	226(91.1%)	22(8.9%)	248
	학생/무직	43(100%)	0(0%)	43
전체		648	57	705

〈표 7〉 월 소득에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
C28	300만원 미만	288(92.3%)	24(7.7%)	312
	300~500만원 미만	251(91.3%)	24(8.7%)	275
	500만원 이상	109(92.4%)	9(7.6%)	118
전체		648	57	705

또한 본 연구를 위해 사용되는 자료는 현재 고객들이 사용하고 있는 통신 서비스에 대한 정보도 포함하고 있다. 구체적으로 유선전화, 초고속인터넷, 인터넷전화, 이동전화 서비스 회사에 대한 정보, 유선전화 및 인터넷전화 요금, 현재 사용 중인 유선전화에 대한 해지 의향, 유선전화를 사용하지 않는 이유, 새로운 통신 서비

〈표 8〉 새로 생성된 변수와 정의

변수 이름	정의	설문항목	범주
D3	이용하는 통신 결합 상품의 종류	A22, A23, A24, A25	0-이용 안 함 1-유선전화-초고속인터넷 결합 6-인터넷전화-이동전화 결합
D4	사용하는 통신 결합 상품-통신사별	A22, A23, A24, A25	0(없음) 1(A사)/2(B사)/3(C사)
D6	통신 결합 상품의 이용 여부	A22, A23, A24, A25	0(이용 안 함)/1(이용함)
D9	유선/초고속인터넷 결합상품-통신사별	A22, A23, A24, A25	0(없음)/1(A사)/2(B사)/3(C사)
D10	유선/인터넷전화 결합상품-통신사별	A22, A23, A24, A25	0(없음)/1(A사)/2(B사)/3(C사)
D11	유선/이동전화 결합상품-통신사별	A22, A23, A24, A25	0(없음)/1(A사)/2(B사)/3(C사)
D12	초고속/이동전화 결합상품-통신사별	A22, A23, A24, A25	0(없음)/1(A사)/2(B사)/3(C사)
D8	유선 전화 불만족 여부	A31, A37	0(없음)/1(있음)
D14	통신 요금 (유선전화/인터넷전화)	A26, A43	유선전화-1(2만원 미만)/2(2~3만원)/3(3만원 이상) 인터넷전화-1(1만원 미만)/2(1~2만원)/3(2만원 이상)

스 가입의향 등의 정보를 포함하고 있다.

현재 이용 중인 유선전화 서비스, 초고속인터넷 서비스, 인터넷전화 서비스, 이동전화 서비스 중 같은 회사에 가입하여 이용하는 서비스가 무엇인지를 파악하기 위해 결합상품을 구분하는 변수 'D3'를 생성하였다. 이 변수는 결합통신서비스를 이용하는 경우를 묻는 변수로 인터넷전화 가입에 대한 고객의 의도를 파악하기 위해 설문 응답의 조합으로 만들어진 변수이다.

통신상품 결합서비스란 2개 이상의 방송 또는 통신서비스 상품에 함께 가입해 이용하는 경우 방송 통신 요금을 할인해 주는 상품이다.[1] 방송 통신 사업자들은 자사의 방송 또는 통신 서비스를 묶어서 제공하고 있으며 자사와 타사의 방송 또는 통신 서비스를 묶어서 제공하는 경우도 있다. 본 연구에서는 타사의 상품을 묶어서 결합상품으로 서비스 하는 경우를 확인할 수 없어 같은 회사의 통신 서비스를 이용하는 경우만을 통신 결합 상품 서비스로 보았다.

변수 'D6'는 결합 통신서비스를 이용하는 경우와 그렇지 않는 경우를 나타내기 위해 생성된 변수이고, 변수 'D9~D12'는 결합 서비스 조합별로 이용하는 통신사를 표현하기 위해 생성된 변수이다. 변수 'D8'의 경우 유선전화에 대한 불만족 여부를 나타내기 위해 생성된 변수로 현재 유선전화를 사용하고 있는 응답자 중 유선전화를 해지할 의향이 있는 고객과 현재 유선전화를 사용하지 않고 있는 고객 중 사용하지 않는 이유를 비용이 비싸다거나 이용이 불편해서라고 응답한 경우를 유선전화에 대해 불만이 있는 경우로 해석하였다. 변수 'D14'는 통신 요금을 나타내는 것으로 현재 사용 중인 유선전화 요금과 인터넷전화 요금을 이용하여 생성된 변수이다.

새로 생성된 변수들은 원래 자료와 마찬가지

1) http://blog.daum.net/kcc1335/1531 방송통신위원회 공식 블로그.

〈표 9〉 유선전화 불만족 여부에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
D8	없음	619(92.8%)	48(7.2%)	667
	있음	29(76.3%)	9(23.7%)	38
전체		648	57	705

로 모두 범주형으로 구성되었다. 새로 생성된 변수에 대한 정의와 생성에 이용한 설문 항목, 그리고 각 변수들이 담을 수 있는 범주형 자료에 대한 설명은 <표 8>에 기술되어 있다.

새로 생성된 변수들 중 유선전화 서비스에 대한 불만족 여부 (D8)와 종속변수인 인터넷전화 가입의향(D7) 간 관계를 교차표를 통해 살펴보았다<표 9>. 유선전화 서비스에 불만이있는 고객이 그렇지 않은 고객에 비해서 더 높은 인터넷전화 가입의향을 보이는 것을 알 수 있다(23.68%).

3.2 인터넷전화 사용의향 예측모델

인터넷전화 사용의향 예측을 위한 모델을 개발하기 위해 범주형 독립변수를 이용하여 범주형 종속변수를 예측할 수 있는 데이터마이닝 알고리즘을 이용하였다. 본 연구에서 채택한 알고리즘은 판별분석, 로지스틱 회귀분석, 분류나무, 신경망이다. 데이터마이닝 도구는 SPSS의 클레멘타인 12.0이다.

<표 1>에서 보여지듯 인터넷전화 사용 의도를 가진 고객을 의미하는 성공 케이스가 전체 케이스의 8.1% 정도이기 때문에 성공 케이스와 실패 케이스의 균형을 맞추는 일이 필요하다[허명회 외, 2003]. 개수가 작은 성공 케이스를 부풀려(Oversampling) 성공 케이스와 실패 케이스의 균형을 맞추고 학습(Training) 자료 셋을 구성하며, 검증(Validation) 자료 셋은 원래 자료 분포에 가깝게 성공 케이스를 작게 구성한다. 이와 같이 작은 수의 성공 케이스와 많은 수의

〈표 10〉 모델 구축에 사용된 설명변수와 종속변수

이름	정의	이름	정의
A12	성별	D3	사용하는 통신 결합 상품 유형
A13	나이	D4	사용하는 통신 결합 상품 통신사
A15	주택 형태	D6	통신 결합 상품 이용 여부
A17	가구 유형	D7	인터넷전화 사용의향 (종속변수)
A22	유선전화 서비스 회사	D8	유선전화 불만 여부
A23	초고속인터넷 서비스 회사	D9	통신사별 결합 유선-초고속-이동
A25	이동전화 서비스 회사	D10	통신사별 결합 상품-유선/초고속인터넷
C14	가구주 직업	D11	통신사별 결합 상품-유선/이동전화
C28	한달 평균 수입	D12	통신사별결합-초고속/이동전화
C32	맞벌이 여부	D14	통신 요금

실패 케이스의 균형을 맞추는 일은 성공 케이스에 대한 데이터마이닝 알고리즘의 학습 능력을 극대화하기 위한 방법으로 알려져 있다[허명회 외, 2003; Shmueli et al., 2007].

종속변수의 값으로 자료를 정렬하고 클레멘타인의 샘플링(Sampling) 노드에서 '1-in-n'옵션을 사용하여 학습 자료 셋과 검증 자료 셋이 겹치지 않도록 자료를 분리한다. 선택된 학습 자료는 성공 케이스에 대한 학습을 극대화하기 위해 밸런싱(Balancing) 노드를 사용하여 성공 케이스에 11.172의 균형 가중치를 주고 성공 케이스와 실패 케이스를 동수로 구성한다. 밸런싱 이전에 29개이던 성공 케이스는 밸런싱 이후 실패 케이스와 동수인 324건으로 부풀려졌다. 검증 자료 셋은 원래 자료의 분포와 유사하게 구성한다.

샘플링과 밸런싱 과정을 통해서 구성된 학습 자료 셋은 성공 케이스 324건, 실패 케이스 324

건으로 구성되었고 검증 자료 셋은 성공 케이스 28건, 실패 케이스 324건으로 구성되었다. 검증 자료 셋에서 성공 케이스의 비율은 7.95%로 원 데이터 셋의 구성(8.1%)과 비슷하다. 자료 셋의 분할에 앞서 일부 널값을 가진 독립 변수의 경우 미리 지정된 값으로 치환하는 과정을 밟았다. 널 값을 가진 설명 변수들은 주로 고객이 사용하고 있는 통신 서비스의 종류를 묻는 문항들이었다. 각 알고리즘에 사용된 설명변수와 종속변수들은 <표 10>에 기술되어 있다.

3.3 판별분석(Discriminant Analysis)

판별 분석의 목적은 여러 개의 연속성 값을 가지는 예측 변수에 대해서 집단 멤버십을 예측하고자 하는 것이다[Wiginton, 1980; Grablowsky and Talley, 1981; Tabachnick and Fidell, 2007]. 판별 분석은 각 집단에 속하는 자료의 정규성(Multivariate normality)과 집단 간 예측변수들의 분산-공분산의 일치성(Homogeneity of variance-covariance) 등의 조건이 충족됨을 전제로 예측 성과가 매우 좋은 통계 기법 중의 하나이다.

판별분석 알고리즘을 이용한 모델 구축은 'Enter' 옵션과 'Stepwise' 방식을 각각 수행하여 검증 자료 셋에서 결과를 비교한다. 판별분석 알고리즘에서 집단 간 예측 변수들의 분산 - 공분산의 일치성은 Box-M 테스트를 통해 검증되는데 두 가지 옵션을 사용한 판별분석 검정 결과 유의확률 0.000으로 유의수준 0.05에서 모집단 공분산 행렬이 동일하다는 귀무가설이 기각되었다. 이것은 판별분석의 기본 가정에 위배되는 사항이지만 공분산 행렬의 동일성이 극단적으로 위배되지 않는다면 판별식을 이용할 수 있다고 알려져 왔으므로[정충영 외, 1998], 본 연구에서는 자료를 그대로 이용하였다.

판별분석 실행 옵션은 다음과 같이 설정되었다. 'Stepwise/Prior Probability(All group equal)/Use Covariance Matrix(Within-group covariance)/Method(Wilks' Lambda or mahalanobim distance)/Criteria(F value-Entry(3.84)-Removal(2.71)'

모델 실행 결과, 고유값(Eigenvalue) 0.745, Wilks 람다 값 0.573(유의확률 0.000), 정준 상관계수 0.653을 보여 주었다. 중요변수로 A22, D4, D10, D14, D3, D11, D6, A15, D12가 선택되었다. 판별분석 알고리즘에 의해 선택된 예측 변수의 상대적인 공헌 정도를 나타내는 표준화된 정준 판별 함수의 값(Standardized canonical coefficient)이 비교적 큰 변수는 A22(0.768), D14(0.603), D3(0.597), D12(-0.823)로, 인터넷전화 사용의향은 유선전화 서비스회사, 통신요금, 사용하는 통신 결합 상품 유형, 통신사별 결합(초고속-이동전화) 변수와 밀접한 관계가 있음을 보여주었다.

'Enter' 방식을 사용한 경우 분류 오류율(Classification error rate)이 검증 자료 셋에서 14.2%를 나타내었고, 'Stepwise' 방식을 사용했을 때 분류 오류율은 검증 자료 셋에서 11.36%로 향상되었다. 따라서 'Stepwise' 옵션에 따라 실행한 판별분석의 학습 자료 셋과 검증 자료 셋에 대한 예측 오류율을 각각 <표 11>, <표 12>에 나타내었다.

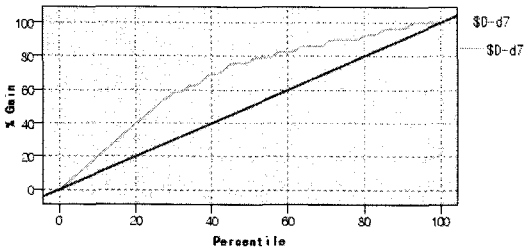
학습 자료 셋에서 예측 오류율은 20.31%이고, 검증 자료 셋에서는 11.36%로 향상되었다. 성공 케이스에 대한 예측실패율을 비교해 보았을 때

〈표 11〉 학습 분류오류(Stepwise)

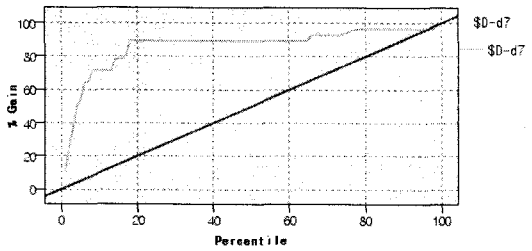
Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	271	53
1	78	246

〈표 12〉 검증 분류오류(Stepwise)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	290	34
1	6	22



〈그림 1〉 Gains chart(학습)



〈그림 2〉 Gains Chart(검증)

학습 자료 셋에서는 324건 중 78개의 오류로 24.07%의 분류 오류가 발생했는데, 검증 자료 셋에서는 28건 중 6개의 오류로 21.43%의 분류 오류를 보여주고 있다. <그림 1>과 <그림 2>는 학습 자료 셋과 검증 자료 셋에서의 판별분석 결과에 대한 게인즈 차트(Gains chart)이다.

3.4 로지스틱 회귀 분석(Logistic regression)

로지스틱 회귀 분석은 다변량 판별분석과 더불어 전통적으로 예측 모형에 이용되던 통계 기법 중의 하나로[Garablowsky and Talley, 1981 Wiginton, 1980; Lee et al., 2002], 종속변수가 이진 형태일 때 사용되는 회귀분석이다[Bradley et el., 1998].

판별분석이 집단 간 공분산의 일치성과 정규성 등의 조건이 충족되어야 함을 전제로 한다면 로지스틱 회귀분석의 경우 이러한 전제 조건에서 좀 더 자유롭다고 알려져 왔다[Tabachnick and Fidell, 2007]. 본 연구는 클레멘타인의 로지스틱 회귀분석 알고리즘을 사용하여 'Enter'방식으로 전체 변수를 대상으로 했을 때와 'Stepwise' 방식으로 변수를 일부 선택했을 때를 비교하여 검증 자료 셋에서의 분류 오류율을 구하였다.

'Enter'방식으로 전체 변수를 대상으로 하였을 때 검증 자료 셋에서 예측 오류율은 11.93%로 나타났고, 'Stepwise(FIN(0.05), FOUT(0.01))' 방식으로 로지스틱 회귀분석을 실행했을 때 예측 오류율은 11.65%로 향상되었다. 이때 선택된 변수들은 A22, D14, A15, A17, D3, A25, C14, A12, D4이다. 판별분석과 비교하였을 때 A22, D14, A15, D3, D4는 일치하는 것으로 나타났다. 예측 변수들의 유형이 모두 셋(Set)으로 설정되어 로지스틱 회귀분석에서는 각 셋에 대한 분석이 이루어졌는데 0.05유의수준에서 통계적 유의성을 가지는 변수들은 A12(1), A15(1, 2), A25(1), D3(3, 5), D14(1, 2)였다. 성공 케이스를 예측하기 위한 로지스틱 회귀식은 아래와 같다.

4.022 × [a12 = 1] + -2.097 × [a15 = 1] + 5.403 × [a15 = 2] + -20.33 × [a17 = 1] + -0.7303 × [a17 = 2] + -2.166 × [a17 = 3] + -0.8774 × [a17 = 4] + 1.564 × [a17 = 5] + -42.49 × [a22 = 1] + -60.34 × [a22 = 2] + -25.1 × [a22 = 3] + 11.07 × [a25 = 1] + -4.703 × [a25 = 2] + 15.19 × [a25 = 3] + 20.44 × [c14 = 1] + 21.09 × [c14 = 2] + 24.26 × [c14 = 3] + 26.39 × [c14 = 4] + 11.58 × [d3 = 0] + -18.58 × [d3 = 1] + 5.548 × [d3 = 2] + 2.106 ×

[d3 = 3] + 6.192 × [d3 = 4] + -3.624 × [d3 = 5] + -5.174 × [d3 = 6] + -0.7326 × [d4 = 0] + 32.62 × [d4 = 1] + 16.8 × [d4 = 2] + -41.72 × [d14 = 0] + -2.434 × [d14 = 1] + -6.4 × [d14 = 2] + -10.28

〈표 13〉 학습 자료 셋 분류오류

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	309	15
1	0	324

〈표 14〉 검증 자료 셋 분류오류

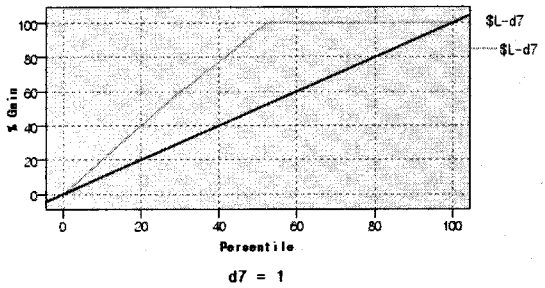
Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	293	31
1	10	18

'Stepwise' 로지스틱 회귀분석결과를 각각 <표 13>, <표 14>에 나타내었다. 학습 자료 셋에서의 분류오류는 2.3%, 검증 자료 셋에서는 11.65%를 보여주고 있다. 성공 케이스에 대한 예측 실패율을 비교해 보았을 때 학습 자료 셋에서는 324건 중 0건 실패로 0%의 예측 실패율을 보여주었고, 검증 자료 셋의 경우 28건 중 10건의 실패로 35.71%의 예측 실패율을 보여주고 있다.

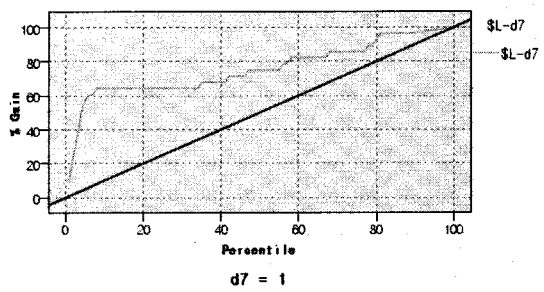
<그림 3>, <그림 4>는 학습 자료 셋과 검증 자료 셋에서의 로지스틱 회귀분석에 대한 게인즈 차트이다.

3.5 분류나무모델(Classification tree model)

분류나무모델 혹은 의사결정나무는 분류 기법의 하나로, 이해하기 쉬운 분류 법칙을 명시적으로 제공한다는 점에서 분류 기법 중 선호도가 높은 알고리즘 중 하나이다[Breiman et al., 1984; Shmueli et al., 2007]. 분류나무모델은 자라난 가지에 최대한 순수한 종속 변수의 값을 가진 자료만을 가질 수 있도록 예측 변수들의반복적인 분할(Partitioning)을 해 나가는 알고리즘이며 이 과정에서 검증 자료 셋을 이용하여 가지치기(Pruning)를 하며 학습 자료 셋에 모델이 과학습(Overfitting)되는 부작용을 방지한다.



〈그림 3〉 Gains Chart(학습)



〈그림 4〉 Gains Chart(검증)

본 연구는 클레멘타인의 C5.0 모델을 사용하여분류나무 알고리즘을 적용하였다. C5.0 실행 옵션은 다음과 같다. 'Simple/Accuracy/Expected Noise(0)'이며 알고리즘 수행 결과 생성된 나무의 깊이(Depth)는 6개였다. C5.0 알고리즘에 의해 선택된 중요 변수는 A22, D14, C14, D3, A12, A15, C28, C32, D11, A17, A13순서였고 성공 케이스(D 7 = 1)를 분류하기 위해 형성된 규칙 셋(Rule set)은 총 10개로 아래와 같았다.

Rule 1 for 1

if a22 in[1]

```
        and a15 = 1
        and d14 in[ 3 ]
        and a13 in[ 4 ]
        then 1
Rule 2 for 1
        if a22 in[ 1 ]
        and a15 = 2
        and c14 in[ 1 ]
        and a12 = 1
        and a17 in[ 2 5 ]
        then 1
Rule 3 for 1
        if a22 in[ 1 ]
        and a15 = 2
        and c14 in[ 2 ]
        and c28 in[ 1 ]
        then 1
Rule 4 for 1
        if a22 in[ 1 ]
        and a15 = 2
        and c14 in[ 4 ]
        and d14 = 1
        and d11 in[ 0 ]
        then 1
Rule 5 for 1
        if a22 in[ 1 ]
        and a15 = 2
        and c14 in[ 4 ]
        and d14 = 3
        then 1
Rule 6 for 1
        if a22 in[ 1 ]
        and a15 = 3
        and c28 in[ 2 ]
        and d3 in[ 3 4 ]
        then 1
Rule 7 for 1
        if a22 in[ 1 ]
        and a15 = 3
        and c28 in[ 2 ]
        and d3 in[ 7 ]
        and c32 = 2
        and a13 in[ 3 ]
        then 1
Rule 8 for 1
        if a22 in[ 9 ]
        and d14 in[ 0 ]
        and c32 = 1
        and c14 in[ 3 ]
        and a15 = 2
        then 1
Rule 9 for 1
        if a22 in[ 9 ]
        and d14 in[ 0 ]
        and c32 = 1
        and c14 in[ 3 ]
        and a15 = 3
        and a12 = 1
        then 1
Rule 10 for 1
        if a22 in[ 9 ]
        and d14 in[ 1 2 3 ]
        then 1
Default : 1
```

첫 번째 규칙에 따르면 고객이 1번 유선전화 서비스 회사를 사용하고, 단독주택에 거주하고 있으며 통신요금이 4만원 이상이고 50대의 연령대라면 인터넷전화를 사용할 의향이 있다고 판단된다. 학습 자료 셋과 검증 자료 셋에 대한 C5.0의 수행 결과는 <표 15>, <표 16>에 제시하였다.

〈표 15〉 학습 자료 셋 분류오류

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	320	4
1	0	324

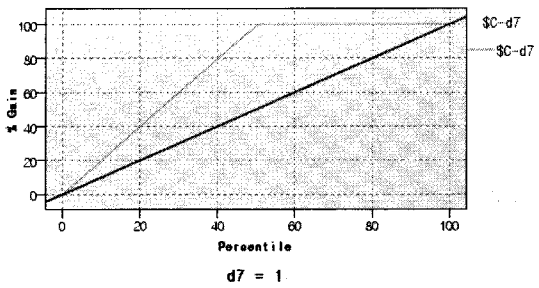
〈표 16〉 검증 자료 셋 분류오류

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	292	32
1	9	19

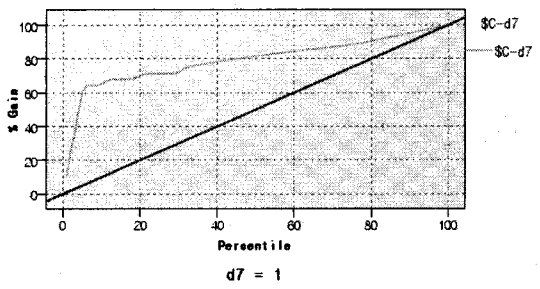
학습 자료 셋에서의 분류 오류는 0.62%를 보여주고 있지만, 검증 자료 셋에서는 11.65%를 보여주고 있다. 그러나 성공 케이스에 대한 예측 실패율을 비교해 보았을 때 검증 자료 셋의 경우 28건 중 9건의 실패로 32.14%의 예측 실패율을 보여주고 있다. <그림 5>, <그림 6>은 학습 자료 셋과 검증 자료 셋에 대한 C5.0 알고리즘의 성능을 게인즈 차트로 나타낸 것이다.

3.6 신경망 모델(Neural Network)

신경망 모델은 통계적 가설이 필요하지 않으면서 비선형 모형을 설명하기에 적당하기 때문에 신용 예측이나 고객 관리에서 뛰어난 성과를 보여주는 알고리즘으로 알려져 왔다[Altman et al., 1994; Cheng and Titterington, 1994; Desai et al., 1997; Desai et al., 1996; 하성호 외, 2009]. 최근의 연구 동향들은 통계적 기법과 데이터마이닝 모형을 비교하여 예측률이 높은 모형을 제안하거나 이들의 혼합 모형을 제안하고 있는데 [김갑식, 2005], Lee et al.[2002]의 경우 판별분석에서 선택된 변수를 신경망 모델에서 입력 변수로 사용할 경우 신경망의 성과가 더 높아졌다



〈그림 5〉 Gains Chart(학습)



〈그림 6〉 Gains Chart(검증)

고 기술하고 있다. 본 연구도 Lee et al.[2002]의 방법을 따라서 1) 판별분석에서 선정된 변수를 입력 변수로 사용한 혼합모형, 2) 로지스틱 회귀분석에서 선정된 변수를 입력 변수로 사용한 혼합모형, 3) 판별분석과 로지스틱 회귀분석에서 공통적으로 나타난 변수를 입력 변수로 사용하는 혼합모형, 4) 전체 변수를 모두 사용한 모형을 상호 비교하여 예측률이 높은 모형으로 신경망 모델의 성능으로 정하기로 한다. 신경망 모델 수행의 파라미터는 <표 17>에 나타나 있다.

<표 17>의 파라미터 설정 값으로 네 개의 후보 모델을 실행한 결과 검증 자료 셋에서의예측 오류율이 전체 변수를 사용한 모형 (4)이 6.82%, 로지스틱 회귀분석의 변수를 사용한 모형 (2)이 10.23%, 판별분석의 변수를 사용한 모형 (1)이 13.92%, 판별분석 모델과 로지스틱회귀분석 모델의 공통 변수를 사용한 모형 (3)의 예측 오차율이 15.06%였다. 전체 변수를 사용한 모형이 가장 좋은 성능을 보였고 이때 D14, D4, A17,

〈표 17〉 신경망 모델의 파라미터 설정

Parameters/Options	
# Hidden layers	1
# Nodes in HiddenLayer-1	20
Persistence	200
Alpha	.9
Initial Eta	.3
High Eta	.1
Eta Decay	30
Low Eta	.01

D10, D9, A25, D6, A12, D11, D8순으로 변수의 중요도가 매겨졌다. 이 모형의 학습 자료 셋에 대한 분류 오류는 0.92%이었다. 신경망 구조는 입력계층에 86개의 뉴론, 은닉계층에 20개의 뉴론, 출력계층에 1개의 뉴론이 생성되었다. <표 18>, <표 19>은 전체 변수를 사용할 경우에 신경망 알고리즘의 성능을 보여준다.

〈표 18〉 학습 자료 셋 분류오류(전체 변수)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	318	6
1	0	324

〈표 19〉 검증 자료 셋 분류오류(전체 변수)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	311	13
1	11	17

그러나 성공 케이스에 대한 예측률에 있어서는 전체 변수를 사용한 경우가 검증 자료 셋에 대해 39.29%의 예측 오류율을 보였고, 판별분석의 변수를 사용한 경우가 28.57%, 두 분석 모델의 공통 변수를 사용한 경우에 예측 오류율 28.57%를, 로지스틱 회귀분석의 변수를 사용한 경우가 32.14%를 보여 주었다. 성공 케이스에 대해서는 판별분석 변수를 사용하거나 공통 변수를 사용하는경우가 가장 성능이 좋은 것으로 나타났다. <표 20>, <표 21>은 성공 케이스 예측에서 성과가 우수했던(판별분석의 변수를 이용한) 혼합모형의 성능을 보여준다. 생성된 신경망 구조는 입력계층에 39개의 뉴론, 은닉계층에 20개의 뉴론, 출력계층에 1개의 뉴론을 사용하였다.

〈표 20〉 학습 자료 셋 분류오류(혼합모형)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	305	19
1	33	291

〈표 21〉 검증 자료 셋 분류오류(혼합모형)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	0	1
0	283	41
1	8	20

4. 예측모델의 비교 검증

4.1 모델의 비교

구축된 다섯 가지 모델의 성과를 비교하기 위해, 검증 자료 셋에 대한 분류 오류율을 구하였다. 모든 모델에 대해서 확률 0.5를 기준으로성공 케이스를 예측하였으며 각 모델의 예측 오류률은 <표 22>~<표 26>에 요약하였다.

〈표 22〉 판별분석 분류 오류

Error Report			
Class	# Cases	# Errors	% Error
0	324	34	10.49
1	28	6	21.43
Overall	352	40	11.36

〈표 23〉 로지스틱 회귀 분류오류

Error Report			
Class	# Cases	# Errors	% Error
0	324	31	9.26
1	28	10	35.71
Overall	352	41	11.65

〈표 24〉 분류나무 분류 오류

Error Report			
Class	# Cases	# Errors	% Error
0	324	32	9.88
1	28	9	32.14
Overall	352	41	11.65

〈표 25〉 신경망 분류 오류(전체변수)

Error Report			
Class	# Cases	# Errors	% Error
0	324	13	4.01
1	28	11	39.29
Overall	352	24	6.82

〈표 26〉 신경망 분류 오류(혼합모형)

Error Report			
Class	# Cases	# Errors	% Error
0	324	41	12.65
1	28	8	28.57
Overall	352	49	13.92

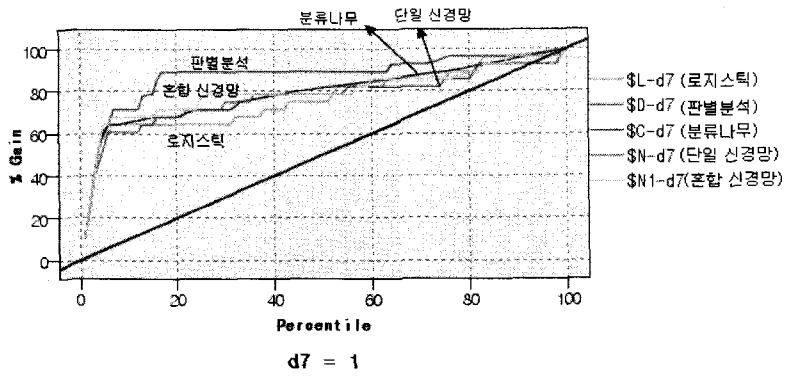
<표 27>은 각 모델에 대한 검증 자료 셋에서의 예측 오류율과 Type I 오류율, Type II 오류율을 나타낸다. 전체 예측 오류율을 비교해 보았을 때 신경망(전체변수)을 사용했을 경우가 가장 성과가 좋은 것으로 나타났다. 그러나 Type I 오류율을 비교해 보았을 때는 판별분석이 더 좋은 성과를 보이는 것으로 나타났다. Type I 오류는 실제로 인터넷전화를 사용할 의향이 있는 고객인데 그렇지 않은 고객으로 예측하여 이익을 낼 수 있는 고객을 잃어버리는 경우에 해당한다.

〈표 27〉 모델 별 예측 오류률 비교

모 델	전체 예측 오류률	Type I 오류률	Type II 오류률
판별분석	11.36%	21.43%	10.49%
로지스틱 회귀	11.65%	35.71%	9.26%
분류나무	11.65%	32.14%	9.88%
단일 신경망(전체변수)	6.82%	39.29%	4.01%
혼합 신경망(혼합모형)	13.92%	28.57%	12.65%

반면에 Type II 오류란 실제로는 인터넷전화를 사용할 의향이 없는 고객임에도 의향이 있는 것으로 예측하여 불필요한 마케팅 비용을 지불해야 하는 경우를 말한다. Type I 오류에 대한 비용과 Type II 오류에 대한 비용의 차이가 크지 않다면 인터넷전화 사용 의향을 예측하기 위한 최적의 모델은 전체 예측 오류율에서 최고의 성과를 보인 신경망(전체변수 사용) 모델이 된다.

그러나 Type I 오류에 대한 비용이 Type II 오류에 비해 월등하게 크다면, 즉 고객을 잃는 비용이 마케팅 비용을 불필요하게 지불하는 것보다 더 높다면 인터넷전화 사용 예측을 위한 최적의 모델은 Type I 오류가 가장 작은 판별분석 모델이 된다. <그림 7>은 다섯 가지 모델의 게인즈 차트를 나타낸 것으로 가장 위쪽에 나타난 판별분석 모델이 가장 성과가 좋은 모델이다. 그러나 전체 예측 오류율의 차이에서 볼 수 있듯이 모델 별 오류율의 차이가 크기 않기 때문에 어떤 알고리즘이 절대적으로 월등하다



〈그림 7〉 전체 모델에 대한 게인즈 차트

고 단정지을 수는 없다.

4.2 예측모델로 분석한 경영정보

인터넷전화를 사용할 의향을 가진 고객을 그렇지 않은 고객으로 잘못 예측하는 경우의 손해가 더 크다고 전제한다면(Type I 오류가 더 크다고 할 경우), 본 연구에서 최고의 성과를 보인 모델은 판별분석 모델이 된다. 판별분석 모델에서 선정된 변수를 중요도에 따라 열거하면 A22(유선전화 서비스회사), D4(사용하는 통신 결합 상품/통신사별), D10(통신사별 결합상품-유선/초고속인터넷), D14(통신요금), D3(사용하는 통신 결합 상품 유형), D11(통신사별 결합상품-유선/이동전화), D6(통신 결합 상품 이용여부), A15(주택형태), D12(통신사별 결합-초고속/이동전화)이다.

이들 변수 중 D4(사용하는 통신 결합 상품-통신사별)와 인터넷전화 사용의향을 가진 고객과의 관계를 교차표를 통해 살펴보면 <표 28>에서 'C 통신 회사'의 결합상품을 쓰고 있는 고객들의 인터넷전화 가입의향이 다른 통신 회사의 고객보다 높은 것을 알 수 있다. 따라서 C 통신사의 결합상품을 사용하고 있는 고객을 표적 마케팅(Target marketing)할 경우 성공적으로 인터넷전화에 대한 사용을 늘릴 수 있을 것이다.

〈표 28〉 결합상품 통신사에 따른 인터넷전화 가입의향 비교

		인터넷전화 가입의향		전체
		없음	있음	
D4	없음	237(97.5%)	6(2.5%)	243
	A 통신사	280(94.3%)	17(5.7%)	297
	B 통신사	110(94%)	7(6%)	117
	C 통신사	21(43.8%)	27(56.2%)	48
전체		648	57	705

<표 29>는 D14(통신요금)과 인터넷전화 사용의향을 비교하고 있는데 통신요금이 비싼 범주에(범주 3) 들어가는 고객이 비록 수는 작지만 인터넷전화 사용의향을 매우 많이 가지고 있음을 알 수 있다. 따라서 현재 유선전화나 인터넷전화 요금을 포함한 통신 요금이 비교적 높은 고객을 인터넷전화 마케팅의 대상으로 하는것이 더 효과적인 방법임을 암시한다고 볼 수 있다.

〈표 29〉 통신요금에 따른 인터넷전화 가입의향 비교

		인터넷전화 가입의향		전체
		없음	있음	
D14	모름	134(96.4%)	5(3.6%)	139
	1	226(92.2%)	19(7.8%)	245
	2	238(93.7%)	16(6.3%)	254
	3	50(74.6%)	17(25.4%)	67
전체		648	57	705

〈표 30〉 통신상품 결합형태에 따른 인터넷전화 가입의향

		인터넷전화 가입의향		전체
		없음	있음	
D3	사용 안함	237(92.6%)	19(7.4%)	256
	유선전화-초고속인터넷 결합	174(98.3%)	3(1.7%)	177
	유선전화-인터넷전화 결합	0(0%)	1(100%)	1
	유선전화-이동전화 결합	74(93.7%)	5(6.3%)	79
	초고속인터넷-인터넷전화 결합	0(0%)	1(100%)	1
	초고속인터넷-이동전화 결합	76(90.5%)	8(9.5%)	84
	인터넷전화-이동전화 결합	0(0%)	3(100%)	3
	세 개 이상 결합	87(83.7%)	17(16.3%)	104
전체		648	57	705

<표 30>은 D3(통신 결합 상품의 이용 여부)와 인터넷전화 사용의향 간의 관계를 나타낸 것이다. '유선전화, 초고속인터넷, 인터넷전화, 이동전화' 중 세 개 이상의 서비스를 같은 통신사의 상품으로 이용하는 경우가 가장 높은 인터넷전화 가입의향을 보여주고 있어서 통신 결합 상품을 이용하는 고객의 경우 인터넷전화에 대한 잠재적 고객이 될 가능성이 높음을 암시한다.

이 중 변수 D3와 D14는 모든 모델에서 공통적으로 선택된 변수로 통신상품 결합형태와 통신비용이 인터넷전화 가입의향에 영향을 미치고 있음을 보여준다. 이것은 결합통신제와 전화요금을 유선전화 해지의 주요한 요인으로 지적하였던 KISDI 연구[KISDI, 2009]와 비슷한 결과이다.

5. 결 론

5.1 연구의 요약 및 의의

본 연구는 인터넷전화 사용의향에 대한 예측 모델을 개발하기 위한 것으로 기존에 유선전화를 사용하고 있는 고객을 대상으로 실시한 설문에서 자료를 수집, 분석하였다. 응답자의 인구통계학적 정보와 통신서비스 이용 정보를 분석하여 인터넷전화 이용의향에 관한 예측을 하였다.

이를 위해 새로운 변수를 추가 생성하였고, 예측 모델 개발을 위해 판별분석, 로지스틱 회귀분석, 분류나무, 신경망을 활용하여 각 모델의 성과를 비교하였다. 단일 신경망이 전체 예측 오류율에서 6.82%로 최고의 성과를 보여주었고 판별분석(11.36%), 로지스틱 회귀분석(11.65%), 분류나무(11.65%), 혼합 신경망(13.92%) 순으로 성과가 나타났다.

신경망 모델에서는 판별분석과 로지스틱 회귀모델에서 공통으로 선정된 변수들을 입력 변수로 사용할 경우(혼합모형) 성공 케이스에 대한 예측 성과가 향상되는 것을 확인하였다(28.57%). 그러나 Type I 오류에 대한 비용이 클 경우 판별분석모델이 최적의 모델로 평가되었고, Type II 오류에 대한 비용이 클 경우 단일 신경망 모델이 최적의 모델로 평가되었다. 판별분석과 로지스틱회귀, 분류나무, 단일 신경망 모델에서 공통적으로 나타난 중요 변수는 D14와 D3로 통신요금과 통신결합 상품유형을 나타내는변수들이었다.

5.2 향후 연구 과제

첫째, 본 연구에서 이용한 자료는 설문의 특성상 응답하지 않은 문항에 대해서는 자료 값이 없었기 때문에 알고리즘의 실행을 위해서 기존 자료값과 겹치지 않는 임의의 값을 입력하였다. 둘째, 성공케이스의 수가 작아 오버샘플링 기법을 이용했고, 모델의 성과에 유의한 차이가 존재하는지 통계적으로 검증할 수 없었다. 셋째, 모든 알고리즘에서 Type I 오류가 Type II 오

류보다월등하게 높게 나타나고 있다. 향후 연구에서는 Type I 오류를 줄일 수 있는 방법이 논의되어야 할 것이다. 특히 성공 케이스가 작을 경우 예측률을 높일 수 있는 방법들에 대한 논의가 필요하다.

참 고 문 헌

[1] 강태규, 김도영, 김영선, "BcN 인터넷전화(VoIP) 기술 동향", *전자통신동향분석*, 제19권 제6호, 2004년 12월, pp. 66-73.

[2] 권오상, 안재홍, "인터넷전화-시장, 요금, 규제", *정보통신정책 ISSUE*, 제14권 제1호, 통권 131호, 2002년 9월, pp. 1-107.

[3] 김갑식, "신용평가를 위한 데이터 마이닝 분류 모형의 통합 모형에 관한 연구", *정보처리학회지D*, 제12권 제2호, 2005년 4월, pp. 211-218.

[4] 김도환, "인터넷 전화의 품질보장제도 및 번호이동제도의 게임이론적 효과분석", *경영학연구*, 제38권 제1호, 2009년 1월, pp. 35-49.

[5] 박대우, 윤석현, "VoIP 서비스의 도청공격과 보안에 관한 연구", *한국컴퓨터정보학회지*, 제11권 제4호, 2006년 9월, pp. 155-164.

[6] 김문구, 권수천, 박종현, "모바일 2.0 촉진을 위한 핵심 성공요인과 모바일 브로드밴드 전개를 수용 특성에 관한 연구", *전자통신동향분석*, 제23권 제6호, 2008년 12월, pp. 112-123.

[7] 박종현, 박희진, 백종현, "국내 유선VoIP 이용특성과 수용 영향요인 분석", *전자통신동향분석*, 제23권 제3호, 2008년 6월, pp. 163-174.

[8] 송관호, "인터넷전화 상호접속 개선을 위한 ENUM 도입방안과 전망-사업자 ENUM 시범사업 결과를 중심으로", *한국통신학회지*, 제24권 제1호, 2007년 2월, pp. 55-64.

[9] 이극노, 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구", *한국지능정보시스템학회논문지*, 제9권 제1호, 2003년 6월, pp. 139-155.

[10] 이병엽, 조규하, 송석일, 유재수, "통신산업의 고객분류를 위한 예측모델 설계", *한국콘텐츠학회논문지*, 제6권 제1호, 2006년 1월, pp. 180-190.

[11] 이주영, "해외의 모바일 VoIP 서비스 제공 현황", *KISDI*, 제21권 제9호, 2009년 5월, pp. 56-64.

[12] 이홍재, 김용규, 유제국, *통신서비스 수요예측 방법론*, 정보통신정책연구원, 2000.

[13] 장범진, 나성현, 이은곤, *인터넷 전략 시장에서의 상품차별화 전략연구*, 정보통신정책연구원, 2006.

[14] 정수미, 이건호, "제품별 구매고객 예측을 위한 인공신경망, 귀납규칙 및 IRANN 모형", *한국경영과학회지*, 제30권 제4호, 2005년 12월, pp. 117-129.

[15] 정충영, 최이규, *SPSSWIN을 이용한 통계분석*, 무역경영사, 1998.

[16] 하성호, 양정원, 민지홍, "코호넨 네트워크와 생존분석을 활용한 신용 예측", *한국경영과학회지*, 제24권 제2호, 2009년 6월, pp. 35-54.

[17] 함창용, 광정호, 맹승찬, 나상우, 천병준, *VoIP 시장의 국내외 현황 및 시사점*, KISDI Issue Report, 정보통신정책연구원, 2007.

[18] 허명회, 이용구, *데이터마이닝 모델링과 사례*, SPSS 아카데미, 2003.

[19] KISDI, *2009년 상반기 통신시장 경쟁상황평가(소비자) 보고서*, 정보통신정책연구원, 한

국리서치, 2009.

[20] ACMA(Australian Communications and Media Authority), *The Australian VoIP Market*, December, 2007.

[21] Altman, E. I., Marco, G., and Varetton, F., "Corporate Distress Diagnosis : Comparisons Using Linear Discriminant Analysis and Neural Networks", *Journal of Baking and Finance*, Vol. 18, May 1994, pp. 505-520.

[22] Bradley, P. S., Fayyard, U. M., and Mangasarian, O. L., "Data Mining : Overview and Optimization Opportunities", *Journal of Computing*, Special issue on Data Mining, January 1998, pp. 17-22.

[23] Breiman, L., Friedman, J., Olshen, R., and Stone, C., *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

[24] Buyut, V. C., Siadat, S. H., and Abidin, W. Z., "Electronic Customer Relationship Management for VoIP Service", *ICCEE : International Conference on Computer and Electrical Engineering*, December 2008, pp. 419-423.

[25] Cheng, B. and Titterington, D. M., "Neural Networks : A review from a Statistical Perspective", *Statistical Science*, Vol. 9, No. 1, 1994, pp. 2-30.

[26] Desai, C. S., Convay, D. F., Crook, J. N., and Overstreet, G. A., "Credit Scoring Models in the Credit Union Environment Using Neural Networks and Genetic Algorithms", *IMA Journal of Mathematics Applied in Business Industry*, Vol. 8, No. 4, 1997, pp. 323-346.

[27] Desai, C. S., Crook, J. N., and Overstreet, G. A., "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment", *European Journal of Operational Research*, Vol. 95, No. 1, November 1996, pp. 24-37.

[28] Grablowsky, B. J. and Talley, W. K., "Probit and Discriminant Functions for Classifying Credit Applicants : A Comparison", *Journal of Economics and Business*, Vol. 33, 1981, pp. 254-261.

[29] Jung. R., *Korea IP Telephony Service Forecast And Analysis, 2004~2008 : 2003*, Year End Review, IDC, 2004.

[30] Kabiraj, S., "Electronic Customer Relationship Management : Origin and Opportunities", *IEMC : Engineering Management Conference*, November 2003, pp. 484-488.

[31] Kohli, R., Piontek, F., Ellington, T., VanOsdon, T., Shepard, M., and Brazel, G., "Managing Customer Relationships through e-Business Decision Support Application : A Case of Hospital-Physician Collaboration", *Decision Support Systems*, Vol. 32, No. 2, December 2001. pp. 171-197.

[32] Ezawa, K. and Norton, S., "Knowledge Discovery in Telecommunication Services Data Using Bayesian Network Models", *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining(KDD-95)*, 1995, pp. 101-105.

[33] Leddy, C., *Cable's 15 most critical VoIP questions*, CableWorld, June 2005.

[34] Lee, T. S., Chiu, C. C., Lu, C. C., and Chen, I. F., "Credit Scoring Using the Hybrid Neural Discriminant Technique", *Expert Systems with Applications*, Vol. 23, No. 3, October 2002, pp. 245-254.

[35] Lin, S. H., VoIP Business broke up, Industry and Business News, Taiwan, 2003.

[36] Massey, A. P., Montoya-Weiss, M. M. and Holcom, K., "Re-engineering the Customer Relationship : Leveraging Knowledge Assets at IBM", *Decision Support Systems*, Vol. 32, No. 2, December 2001, pp. 155-170.

[37] Lingfen, S. and Ifeachor, E. C., "Voice Quality Prediction Models and Their Application in VoIP Networks", *IEEE Transactions on Multimedia*, Vol. 8, No. 4, 2006, pp. 809-820.

[38] Oracle Corporation, *The Implications of VoIP for Service Providers*, 2005.

[39] Ovum, *Fixed Voice Services : Market Development Scenario*, Wireline Strategy, 2007.

[40] Peppers, D., Rogers, M. and Dorf, B., *The One to One Fieldbook*, Garden City, NY, Doubleday, 1999.

[41] Pine II, B. J., Peppers, D., and Rogers, M., "Do you want to keep your customer forever?", *Harvard Business Review*, Vol. 73, No. 2, March 1995, pp. 103-114.

[42] Rinde, J., "Telephony in the year 2005", *Computer Networks*, Vol. 31, No. 3, February 1999, pp. 157-168.

[43] Shmueli, G., Patel, R. N., and Bruce, C. P., *Data Mining for Business Intelligence-Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, A John Wiley and Sons, 2007.

[44] Tabachnick, C. B. and Fidell, S. L., *Using Multivariate Statistics 5th*, Pearson Educations, 2007.

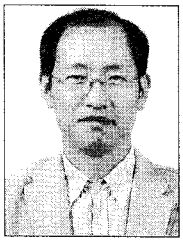
[45] Tseng, F. M., "Forecasting the Taiwan Customer Market for Internet Telephony", *Journal of the Chinese Institute of Industrial Engineers*, Vol. 22, No. 2, 2005, pp. 93-105.

[46] Wiginton, J. C., "A note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior", *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3, September 1980, pp. 757-770.

[47] World IT Report, *Demand for VOIP services across Asia-Pacific to boom*, 2003.

[48] Zhang, G., Hillenbrand, M., and Müller, P., "Facilitating the Interoperability among Different VoIP Protocols with VoIP Web Services", *First International Conference on Distributed Frameworks for Multimedia Applications*, February 2005, pp. 39-44.

■ 저자소개

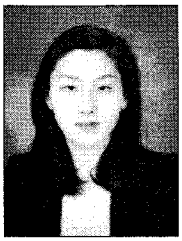


하 성 호

연세대학교 경영학과를 졸업하고, 한국과학기술원에서 정보시스템으로 공학석사, 공학박사를 받았다. 현재 경북대학교 경영학부 부교수로 재직중이다. 다수 국내외 저널의 편집위원으로 활동하고 있으며 관심분야는 지능형정보시스템, 데이터마이닝, e-비즈니스, 지식서비스 등이 있다.

양 정 원

현재 경북대학교 경영학과 박사과정에 재학 중이며 서울대학교 문학사, Texas A&M University Mays Business School에서 MS/MIS 석사 학위를 취득하였다. 주요 연구분야는 온라인 지식 공유 서비스와 데이터 마이닝 등이다.



송 영 미

경북대학교 경영학부에서 석사 학위를 취득하였으며, 현재 동대학원에서 박사과정 중에 있다. 주요 관심분야는 Web 2.0, OSS, 모바일 서비스 등이 있다.

◈ 이 논문은 2010년 04월 22일 접수하여 2010년 06월 11일 게재확정되었습니다.