한국지능정보시스템학회논문지 제10권 제3호 2004년 12월(pp.115~132)

이동통신 사용패턴을 이용한 고객의 직업판정

이재식
아주대학교 경영대학 e-비즈니스학부 교수
(leejsk@ajou.ac.kr)

조유정
아주대학교 대학원 경영학과 박사과정
(mispower@empal.com)

최근 기업들이 고객관계관리의 중요성을 인식함에 따라 고객에 대한 이해의 필요성이 증대되고 있다. 고객의 직업은 고객을 이해하는데 있어서 매우 중요한 정보이다. 하지만 대부분의 고객들이 자신의 직업을 노출하는 것을 꺼리기 때문에 기업에게 그들의 직업을 알려주지 않는 것이 다반사이고, 심지어는 잘못된 직업을 알려주기도 한다. 본 연구의 대상은 이동통신서비스 업체이다. 본 연구에서 우리는 통화상세이력 데이터를 이용하여 고객의 직업을 판정하는 모델을 구축하였다. 인공신경망을 이용해서 우리는 두 단계로 이루어진 직업판정 모델을 구축하였다. 첫 번째 단계에서는 먼저 4개의 직업군을 판정하였고, 두 번째 단계에서 이 4개의 직업군을 세분하여 총 7개의 직업을 판정하였다. 이러한 방식으로 7개의 직업을 판정한 모델의 최종적중률은 71.9%이었다.

Key Words : 고객관계관리, 고객직업판정, 자기조직화지도, 인공신경망, 이동통신 산업.

논문접수일 : 2004년 11월 게재확정일 : 2004년 12월 교신저자 : 조유정

1. 서론

치열한 경쟁상황하의 이동통신서비스 산업에서 '고객'의 중요성은 계속 증가하고 있다. 각 이동통신서비스 업체들은 기존의 가치 있는 고객들의 이탈을 막고 새로운 고객을 유치하기 위한 전략을 구사하는데 총력을 기울이고 있다. 이를 위해서는 무엇보다도 정확한 고객정보를 바탕으로 한 올바른 고객 이해가 필수적이다. 하지만 요즘 고객들은 정보노출에 따른 부작용들을 염려하여 자신에 대한 데이터 공개를 꺼리고 있다. 그들은 서비스에 가입할 때에, 가입에 필요한 최소한의 데이터만을 기록하고 기업의 입장에서 고객을 이해하는데 중요한 데이터인 직업이나 소득 등은 숨기는 것이 다반사이다.

또한, 직업 데이터는 성별이나 생년월일 데이터처럼 한번 입력된 후 변하지 않는 것이 아니라 시간에 따라 변할 가능성을 가지고 있다. 그러므로 올바른 고객 이해를 위해서는 주기적으로 직업 데이터를 갱신 및 관리할 필요가 있다. 하지만 수많은 고객을 일일이 접촉하여 직업을 파악하는 작업은 거의 불가능할 뿐만 아니라, 일부 샘플에 대해서만 직업 파악 작업을 수행한다고 할지라도 막대한 비용이 소요되는 작업이다. 그러므로 대부분의 이동통신서비스 업체들이 직업 데이터의 중요성을 인식하면서도 이를 획득 및 관리하는 데는 소극적일 수밖에 없는 실정이다.

이에 본 연구에서는 이동통신서비스 사용자들

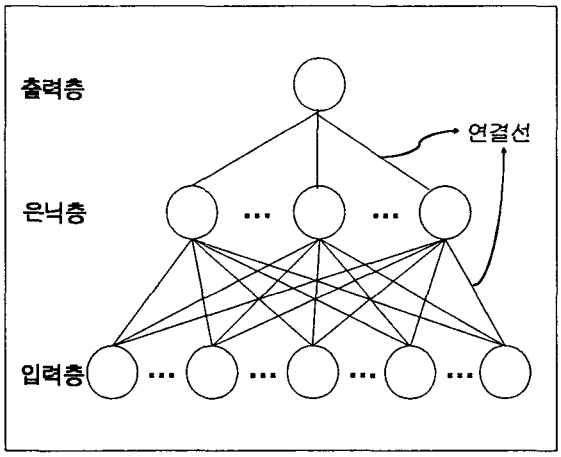
이 자신을 암시적으로 알리는 통화상세이력 데이터(Call Detail Record)를 분석하여 고객의 직업을 판정하는 시스템을 구축하고자 한다. 직업판정 시스템의 구축으로 기업은 직업 정보를 정기적으로 갱신 및 관리할 수 있게 된다. 더불어, 이동통신서비스 업체에서 단순히 사용요금을 청구하기 위한 목적으로만 수집하고 활용하였던 대용량의 통화상세이력 데이터를 고객관계관리를 위해 적극적으로 활용한다는데 본 연구의 의의가 있다.

본 논문은 다음과 같이 구성되어 있다. 제 2장에서는 통신서비스 산업에서의 고객관계관리에 대한 기존연구를 간략하게 살펴본다. 제 3장에서는 본 연구의 모델구축기법으로 사용한 인공신경망에 대해 간략하게 소개한다. 제 4장에서는 본 연구에서 사용한 데이터를 설명하고, 제 5장에서는 직업판정시스템의 구축과정에 대해서 기술한다. 제 6장에서는 직업판정시스템을 평가하고, 본 연구의 결론과 한계점을 기술한다.

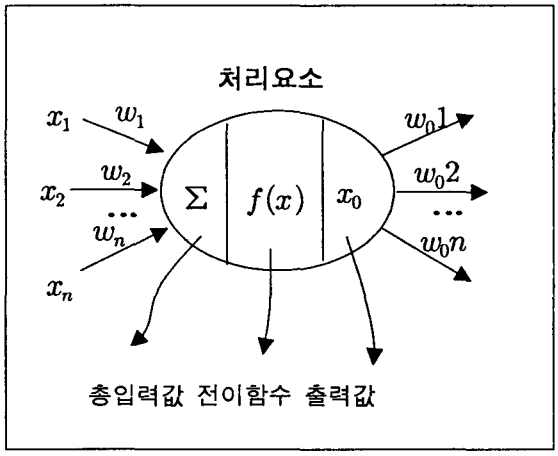
2. 통신서비스산업에서의 고객관계관리에 대한 기존연구

유선통신서비스 산업도 이동통신서비스 산업과 마찬가지로 고객관계관리가 매우 중요하다. Hahm *et al.*은 유선통신서비스 산업에서 고객서비스 만족도를 측정 및 관리하는 방법에 대한 연구를 수행하였다[3]. 먼저 고객을 의미 있고 동질성이 있는 그룹으로 세분화한 후 하나의 그룹을 선정하여 고객서비스 만족도를 측정하고 관리할 수 있는 의사결정지원시스템을 구축하였다. 그룹의 세분화는 서비스 종류와 고객 종류를 중심으로 수행하였는데, 서비스는 일반전화, 공중전화, Fax, ISDN, Hi-NetP 그리고 Hi-Tel 등 6개의 종류를 고려였고, 고객은 개인고객과 법인고객의 두 종류로 나누었다. 이러한 서비스/고객 매트릭스를 통해서 12개의 그룹이 만들어지는데, 그 중 해당사항이 없는 3개 그룹을 제외시키고 나머지 9개의 그룹을 연구대상으로 선정하였다. 그들은 9개의 그룹 가운데 일반전화/개인고객 그룹의 고객서비스 만족도 측정을 위한 설문지를 개발하였고, 설문 결과로 얻은 고객만족도 측정 데이터를 분석 및 관리하기 위한 의사결정지원시스템을 개발하였다. 개발한 의사결정지원시스템은 입력모듈(Input Module)과 분석모듈(Analysis Module)로 구성되었다. 입력모듈은 입력, 편집 그리고 삭제 기능을 포함하고 있고, 분석모듈은 데이터추적(Data Tracking)과 통계분석 기능을 포함하고 있다. 데이터추적 기능은 과거 5년 동안 고객서비스 만족도의 추세를 보여주고, 고객서비스 만족도가 가장 향상 또는 하락한 특정 서비스 영역을 식별하는 것을 지원한다. 통계분석 기능에서는 전체 고객서비스 만족도 점수를 결정하는 요인들을 판단하기 위해 상관관계분석 등을 수행한다.

본 연구의 대상과 동일한 이동통신서비스 산업에 대한 연구로서, Mozer *et al.*은 이동통신사용자가 어떤 불만족요인이 생겼을 경우에 이탈하는지를 예측하는 연구를 수행하였다[7]. 이탈예측을 위해 그들은 로지스틱회귀분석, 의사결정나무 그리고 인공신경망 등의 기법을 이용하였다. 가입자의 사용이력, 요금청구내역, 신용관련사항 그리고 불만사항에 대한 기록 등을 이탈예측의 주요변수로 사용하였는데, 궁극적인 이탈요인은 통화요금에 대한 불만족과 서비스품질의 저하 등인 것으로 나타났다. 또한, 리프트 곡선(Lift Curve)을 기준으로 다양한 기법들의 성능을 비교한 결과, 가장 우수한 성능을 보인 것은 인공신경망을 이용



[그림 3-1] 인공신경망의 구조



[그림 3-2] 처리요소에서의 값 처리과정

한 모델이었다.

Hwang *et al.*은 이동통신사용자의 고객평생가치(LTV: Life Time Value)를 계산하는 새로운 모델을 제안하고, 고객이탈 가능성과 교차판매(Cross Selling) 가능성을 고려해서 고객을 세분화하는 연구를 하였다[4]. 기존의 고객평생가치 모델들은 고객이탈 가능성을 고려하지 않았다. 고객이탈은 서비스 기간의 길이와 미래수익창출(Future Profit Generation)에 영향을 미치기 때문에 고객평생가치 모델을 구축하는데 있어서 중요하다. 이에 그들은 과거수익에 대한 공헌도, 잠재적인 이익 그리고 이탈가능성을 고려하여 고객평생가치 모델을 구축하였다. 또한 고객가치를 분석해서 그것에 따라 고객을 세분화하였는데 고객가치는 현재가치(Current Value), 잠재가치(Potential Value) 그리고 고객충성도(Customer Loyalty) 세 개의 범주로 구분하였다. 잠재가치는 교차판매의 가능성을 반영한 것이고, 고객충성도는 고객유지(Customer Retention)에 대한 척도가 된다.

3. 인공신경망

3.1 인공신경망의 개요

일반적으로 인공신경망은 입력층, 은닉층 그리고 출력층으로 구성되며, 각각의 층은 처리요소라고 불리는 노드들의 집합으로 구성된다. 각 층 사이에 이러한 처리요소들은 [그림 3-1]과 같이 연결강도(Connection Weight)를 갖는 연결선으로 연결된다[8].

[그림 3-2]는 처리요소에 의해 입력값이 어떻게 처리되어 출력값이 생성되는지를 보여준다. 은닉층과 출력층을 구성하는 노드들은 전단계의 출력값에 연결강도를 곱한 값을 입력값으로 받아서 이것들을 합산한 후에 특정 전이함수(Transfer Function)에 의해 출력값을 생성하게 된다. 이 출력값은 다시 다음 단계의 입력값으로 사용된다. 인공신경망 학습의 종류에는 감독학습(Supervised Learning)과 무감독학습(Unsupervised Learning)이 있다. 감독학습에서는 출력값과 목표값(Target Value)이 비교된 후, 출력값과 목표값 사이의 오차를 줄이기 위해 연결강도(Connection Weights)

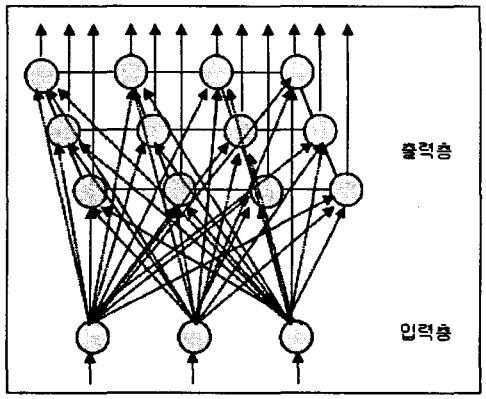
를 조절한다. 이러한 연결강도의 조절을 인공신경망의 학습이라 하고, 이것은 출력값과 목표값이 일치하지 않는 경우에만 이루어진다. 무감독학습에서는 목표값에 대한 정보없이 입력값만을 가지고 주어진 학습규칙에 따른 연결강도 조절로 인공신경망을 자기조직화(Self-Organizing) 한다.

3.2 Kohonen의 Self-Organizing Map(SOM)

Kohonen Feature Map이라고도 불리는 Self-Organizing Map(SOM)은 인공신경망의 한 종류로서 Tuevo Kohonen에 의해 개발되었다[5]. SOM은 주로 군집화(Clustering)와 차원축소(Dimension Reduction)를 목적으로 사용되지만 다른 인공신경망처럼 분류(Classification), 예측(Forecasting) 그리고 최적화(Optimization)의 목적으로도 사용된다. 무감독학습을 하는 SOM은 입력벡터를 받아서 이 입력벡터를 유사성에 따라 군집화한다[6].

SOM은 [그림 3-3]에서 보는 것과 같이 입력층과 2차원 격자모양의 출력층으로 구성된 네트워크 구조를 이루고 있다. 이 네트워크는 완전연결이기 때문에 모든 입력노드는 모든 출력노드와 연결되어 있다.

SOM은 연결강도가 임의의 값으로 초기화된 후에 훈련이 시작되는데, 연결강도와 입력벡터들은 0에서 1사이로 정규화(normalized)되어야 한다. SOM의 학습에서는 입력노드와 출력노드 간의 거리 d_j 를 계산한다. 입력벡터가 N개의 값들로 구성되어 있을 때에, t번째 훈련에서 i 번째 입력노드의 입력값을 $X_i(t)$라 하고, i 번째 입력노드와 j번째 출력노드 사이의 연결강도를 $W_{ij}(t)$라 할 때, 거리 d_j 는 식 (3-1)과 같이 계산한다[9].



[그림 3-3] SOM의 구조

$$d_j = \sum_{i=1}^{N} (X_i(t) - W_{ij}(t))^2$$

식 (3-1)

SOM의 학습철학이 승자독점(Winner takes all)이기 때문에 승자노드만이 출력신호를 낼 수 있고, 승자노드와 그것의 이웃 노드들만이 제시된 입력벡터에 대한 학습을 할 수 있다. 그러므로 각 노드들은 학습할 수 있는 특권을 부여받기 위해서 서로 경쟁하는데, 거리 d_j 가 가장 가까운 노드가 승자노드로 선정된다. 승자노드 j^*가 결정되면, SOM은 입력벡터와 더 가까워지기 위해서 학습규칙에 따라 연결강도를 조절한다. 승자노드와 인접한 출력노드들을 이웃들(neighborhood)이라고 하는데, 이 이웃들도 입력벡터와 더 가까워지기 위해서 그들의 연결강도를 조절한다. 승자노드 j^*와 이웃노드 j는 식 (3-2)에 따라 연결강도 벡터를 조절한다. α는 0과 1사이의 값을 가지는 학습계수이다.

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(X_i(t) - W_{ij}(t))$$

식 (3-2)

이 과정이 끝나면 또 다른 입력벡터가 들어오게 되고 위의 과정을 반복한다. 즉 새로운 승자노드가 선택되고 선택된 승자노드는 출력신호를 내며 승자노드와 그 이웃노드들의 연결강도 벡터는 입력벡터와 가까워지게 된다. 이 과정은 모든 훈련이 끝날 때까지 계속 반복된다.

4. 연구에 사용한 데이터

4.1 이동통신 데이터

본 연구에 사용한 데이터는 한 이동통신회사로부터 사용허가를 받고 수집한 개별고객의 통화상세이력 데이터, 단말기에 따른 고객데이터 그리고 고객에 대한 인구통계학적인 데이터 등이다. 분석시점을 기준으로 이동통신서비스를 사용 중에 있고, 최소 3개월 이전부터 현재까지 사용이력데이터가 존재하는 고객들 가운데 총 2,170명의 고객을 선정하였다.

이동통신서비스 가입고객 중에는 명의자와 실제 사용자가 다른 경우가 많다. 본인명의의 이동통신단말기를 어떠한 이유에서건 가족들이나 타인이 사용하는 것이다. 본 연구를 위해 선정한 2,170명의 고객데이터내의 직업 기입율은 7%에 불과했다. 또한 가입당시시점과 분석시점사이에 시간차가 존재하기 때문에 기입된 직업들도 변동했을 가능성이 존재한다. 그러므로 우리는 선정된 고객들에게 직접 개별적인 연락을 취해 분석시점에서의 실제직업데이터를 획득하였다. 직업데이터가 정확하게 파악된 2,170명 고객의 데이터 중 70%에 해당하는 1,519명의 데이터는 Training용으로 사용하였고, 20%에 해당하는 434명의 데이터는 Test용으로 그리고 나머지 10%인 217명의 데이터는 Validation용으로 사용하였다[2]. 데이터는 다음과 같이 총 4개의 테이블로 구성되어 있다.

(1) 고객에 대한 인구통계학적인 데이터테이블.
(2) 각 고객이 소유한 단말기에 대한 데이터테이블.
(3) 각 단말기의 발신에 대한 발신통화상세이력 데이터테이블.
(4) 각 단말기의 착신에 대한 착신통화상세이력 데이터테이블.

위의 각각의 테이블들을 연결하는 키(Key)는 모두 단말기번호(Handset Telephone Number)가 된다. 개별고객에 대해서, 발신통화상세이력 데이터테이블에는 월평균 약 150건의 사용이력이 존재하고 착신통화상세이력 데이터테이블에는 월평균 약 100건의 사용이력이 존재한다. 발신통화상세이력 데이터테이블은 이동통신회사의 주 수입원인 사용요금을 청구하는데 직접적으로 관련되기 때문에 매우 상세하게 기록된다. 실제 원데이터(Raw data)에는 정상적으로 통화완료가 이루어지지 않은 기록도 약 50%에 달한다. 또한 상대방의 번호입력시에 발생할 수 있는 실수들도 전산상에 기록으로 남아있다. 이렇게 불필요할 정도로 상세한 통화기록은 데이터의 크기를 방대하게 할 뿐만 아니라 데이터의 품질을 저하시킨다. 따라서 통화상세이력 데이터를 유용한 정보로 변환하기 위해서는 전문가의 지식과 많은 시간을 요구하는 정교한 전처리 작업이 선행되어야만 한다.

4.2 직업의 분류

노동부에서는 매 5년마다 한국의 직업을 새롭게 분류하고 있는데, 현재 사용되고 있는 2000년 1월에 마련된 한국표준직업분류에 의하면 우리나라 직업의 총 개수는 세세분류시 1,567개이다[1]. 하루가 다르게 기존의 직업이 사라지고, 새로운 직업이 생겨나고 있다. 2004년 현재 한국의 표준직업의 개수는 2000년 1월의 통계치보다는 다소 증가되어 있으리라고 생각된다.

본 연구의 목적은 통계청의 분류처럼 직업을 세세하게 구분하고자 하는 것이 아니다. 직업에 대한 정보를 활용하여 고객에게 마케팅 활동을 펼칠 때에는 고객의 특성을 고려한 직업분류가 필요하다. 그러므로 본 연구에서는 기업이 고객을 이해하고 고객에게 마케팅 활동을 펼치는데 유용하다고 판단되는 7개의 직업군(이하 7개의 직업으로 통칭)을 새롭게 정의하였다. 궁극적으로 본 연구를 통해 판정하고자 하는 직업명과 전체 데이터에서 각 직업별 분포는 <표 4-1>과 같다.

<표 4-1>에 제시된 것처럼 전체 데이터에서 가장 높은 비율을 차지하고 있는 직업은 내근직으로서 43.9%를 차지하고 있고, 그 다음이 외근직으로서 약 17.0%를 차지하고 있다. <표 4-2>는 모델의 적중률 평가에 사용되는 Validation 데이터의 각 직업별 분포를 보여주고 있다.

<표 4-1> 전체 데이터의 직업별 분포

직업명	빈도	비율(%)
내근직	953	43.9
외근직	369	17.0
주부	248	11.4
대학생	206	9.5
정주형자영업	206	9.5
무직	99	4.6
중고생	89	4.1
계	2,170	100.0

<표 4-2> Validation 데이터의 직업별 분포

직업명	빈도	비율(%)
내근직	95	43.8
외근직	39	18.0
주부	24	11.1
대학생	20	9.2
정주형자영업	20	9.2
중고생	10	4.6
무직	9	4.1
계	217	100.0

4.3 요약변수와 파생변수의 생성

통화상세이력 데이터에 기록된 값들을 직접 모델에 이용하는 것은 불가능하다. 수많은 결측치(Null)가 존재하고 사용패턴에 대한 정보가 문자형으로 함축되어 있는 경우도 있기 때문에 해당 데이터를 활용하기 위해서는 데이터의 전처리 작업을 하여야 한다. 전처리 작업을 통해 요약변수와 파생변수를 생성한다. <표 4-3>과 <표 4-4>는 발신과 착신통화상세이력 데이터테이블에 저장되어 있는 변수들을 보여주고 있다.

발신과 착신통화상세이력 데이터테이블에 저장된 변수들을 이용해 요약변수를 생성한다. 요약변수 생성을 위해 통화호수, 통화량, 통화자의 크기, 기지국수, 통화시간대 그리고 요일을 고려하였다.

① 통화호수 : 통화호수는 몇 번의 발신 또는 착신이 이루어졌느냐에 대한 것이다. 사용자가 주로 발신을 많이 하는 사람인지 아니면 착신을 많이 하는 사람인지 등의 행동유형을 파악할 수 있는 단서가 된다. 발신통화상세이력 데이터(또는 착신통화상세이력 데이터)에서 사

<표 4-3> 발신통화상세이력 데이터테이블의 일부 변수들

변수	변수설명	변수	변수설명
1	통화상세이력버전	15	통화시간
2	통화상세이력종류	16	통화시작네트워크
3	부가서비스현황	17	통화시작스위치
4	호종료상태	18	통화시작지역
5	통화상세이력네트워크	19	통화시작기지국
7	요금청구네트워크	20	통화종료스위치
8	요금청구스위치	21	상대기지국
9	요금청구번호	22	부가서비스
10	발신번호	23	통화종료지역
11	착신번호	24	통화상대네트워크
12	호시작시간	25	통화상대스위치
13	등록상태	26	통화상대지역
14	호종료시간		. . .

<표 4-4> 착신통화상세이력 데이터테이블의 일부 변수들

변수	변수설명	변수	변수설명
1	통화상세이력버전	7	통화상세이력생성
2	통화상세이력종류	8	호시작시간
3	호서비스종류	9	호종료시간
4	호종료상태	10	통화시간
5	발신번호	11	착신과금여부
6	착신번호		. . .

용자별로 해당 통화호수를 세어서 해당 데이터를 얻는다.

② 통화량 : 통화량은 사용자가 얼마나 오랫동안 통화를 지속하느냐에 대한 정보를 준다. 각 통화에 대해 '통화시간(Duration)' 변수로 주어진다. 사용자가 짧은 통화를 자주하는지 아니면 긴 통화를 자주하는지 등의 사용패턴을 파악할 수 있다.

③ 통화자크기(Network Size) : 통화자크기는 사용자가 얼마나 많은 사람들과 통화를 하는지를 나타낸다. 한사람과 집중적인 통화를 하는 사람은 통화자크기가 작을 것이고, 여러 명과 다양한 통화를 하는 사람은 통화자크기가 클 것이다. 통화자크기는 통화상세이력 데이터에서 산출할 수 있다. 발신통화상세이력 데이터에서 유일한 착신자의 수를 세어보고, 착신통화상세이력 데이터에서 유일한 발신자의 수를 세어봄으로써 데이터를 획득할 수 있다.

④ 기지국수 : 기지국수를 통해서는 사용자의 공간이동성을 알 수 있다. 공간이동이 많은 직업에 종사하는지 여부를 판단하는데 유용하다. 점심시간대에는 불규칙한 공간의 이동이 있을 수 있기 때문에 제외시켰다. 공간이동성은 발신통화상세이력 데이터에서 추출이 가능하다.

⑤ 요일 : 요일은 주중과 휴일로 구분을 하였다. 주중에는 월, 화, 수, 목, 금이 해당이 되고 휴일에는 일요일과 공휴일이 포함된다. 토요일은 직업과 상관없이 개별회사의 사정에 따라 다양성을 나타나기 때문에 휴일이나 주중구분에서는 제외시켰다. 하지만, 전체요일을 산정할 때는 주중과 휴일 그리고 토요일을 모두 포함한다.

⑥ 시간대 : 시간대는 모든 요일에 대해 크게 10가지로 구분하였다. 시간대에 대한 명칭은 요일에 상관없이 일반적인 직장인의 시간대를 기준으로 하였다. 출근시간대, 오전시간대, 점심시간대, 오후시간대, 순근무시간대, 퇴근시간전시간대, 퇴근시간대, 저녁시간대, 여가시간대 그리고 심야시간대 등이 10가지의 시간대에 해당된다. 여가시간대란 퇴근 후 저녁식사를 마친 다음 자유롭게 쉬는 시간을 말한다. 순근무시간대는 출근시각부터 퇴근시각까지의 총근무시간대에서 점심시간대를 제외한 것을 말한다. 시간대 분할은 연구자가 임의로 일반적인 직장인의 표준시간대를 기준으로 분할한 것이다.

⑦ 통화상대구분 : 통화상대구분은 통화상대가 유선전화사용자인지 무선전화사용자인지를 구분하는 것이다.

생성된 요약변수는 총 125개이다. 요약변수 중 일부는 모델에 그대로 사용하였고, 일부는 새로운 파생변수를 생성하는데 사용하였다. 이들 요약변수와 파생변수를 생성할 때 주중의 요일수가 5일이고, 휴일의 요일수가 1일이거나 공휴일이 포함된 경우 2일이상이기 때문에 주중과 휴일의 불규칙한 요일수로 인해 통화패턴분석에 Bias가 발생한다. 이 Bias를 해소하기 위해 일일평균값을 사용하였다. 125개의 요약변수 중 모델 구축을 위해서 사용한 변수 중 일부가 <표 4-5>에 제시되어 있다. 요약변수들은 직업별로 개인차가 크게 발생하기 때문에 개인차를 줄이기 위해 변수들의 '비

<표 4-5> 요약변수 중 일부

변수	변수설명	변수	변수설명	변수	변수설명
1	전체 무선/무선발신호수	15	휴일 타무선통화자크기	29	전체 무선/유선통화자크기
2	주중 무선/무선발신호수	16	전체 타무선착신호수	30	주중 여가시간대 발신호수
3	휴일 무선/무선발신호수	17	주중 타무선착신호수	31	주중 출근시간대 발신호수
4	전체 무선/무선통화자크기	18	휴일 타무선착신호수	32	주중 오전시간대 발신호수
5	휴일 무선/무선통화자크기	19	전체 타무선착신통화자크기	33	주중 점심시간대 발신호수
6	전체 타무선발신호수	20	주중 타무선착신통화자크기	34	주중 퇴근시간대 발신호수
7	주중 타무선발신호수	21	휴일 타무선착신통화자크기	35	주중 근무시간대 발신호수
8	휴일 타무선발신호수	22	전체 무선/유선발신호수	36	주중 퇴근시간이후 착신호수
9	전체 타무선통화자크기	23	주중 무선/유선발신호수	37	주중 퇴근시간이후 발신호수
10	주중 타무선통화자크기	24	휴일 무선/유선발신호수	38	휴일 무선/유선착신호수
11	주중 무선/유선통화자크기	25	휴일 자무선발신호수	39	주중 타무선통화자크기
12	전체 자무선발신호수	26	전체 타무선통화자크기	40	휴일 타무선통화자크기
13	휴일 무선/유선통화자크기	27	주중 무선/유선착신호수	41	전체 무선/유선착신호수
14	주중 자무선발신호수	28	휴일 점심시간대 발신호수		· · ·

<표 4-6> 파생변수 중 일부

변수	변수설명	변수	변수설명	변수	변수설명
1	주중/휴일 발신호수차이	6	주중 근무시간대 착신호수비율	11	무선통화자크기비율
2	무선발신호수 비율	7	주중 퇴근시간대 통화자크기 비율	12	근무시간대 주중/휴일 발신호수 차이
3	휴일 농촌기지국이용율	8	무선/유선 통화자크기차이		
4	유선 발신/착신비율	9	주중 퇴근시간대 통화자크기비율	13	주중 농촌기지국이용율
5	주중 최빈기지국의존도	10	타이동통신 착신호수비율		...

율'과 '차이' 등을 고려한 파생변수를 추가로 생성하였다. <표 4-6>에 파생변수 일부가 제시되어 있다.

5. 직업판정시스템 구축

본 절에서는 구체적으로 직업판정시스템이 구축되는 과정을 제시한다. 본 연구에 사용된 데이터에는 성별, 연령대 그리고 요금제도 등 소수의 범주형 변수(Categorical Variable)가 포함되어 있다. 하지만, 대부분이 수치형의 요약변수와 파생변수이기 때문에 수치형 변수에 대해 뛰어난 성능을 보이는 인공신경망기법을 사용하여 모델을 구축하였는데, 감독학습을 하는 인공신경망(이하 ANN(Artificial Neural Network)으로 칭함)과 무감독학습을 하는 인공신경망인 SOM을 모두 사용하였다.

본 연구에서는 모델의 적중률을 향상시키기 위하여 다양한 시도를 하였는데 <표 5-1>과 같이 4가지 유형의 모델을 구축하였다.

5.1 Model_1의 구축

<표 5-1>에서 보는 것처럼 Model_1은 SOM을 통해 구축하였다. Model_1 구축에 사용된 변수들은 <표 5-2>와 같다. 모델에 사용된 수치형 변수들은 식 (5-1)에 의해 모두 0.1과 0.9 사이의 값으로 정규화하였다.

<표 5-1> 구축된 4가지 모델

Model_1	기법	SOM	
	기능	7개의 직업을 판정	
Model_2	기법	SOM ⇨ ANN	
	기능	4개의 직업군을 판정	7개의 직업을 판정
Model_3	기법	ANN ⇨ ANN	
	기능	4개의 직업군을 판정	7개의 직업을 판정
Model_4	기법	ANN ⇨ ANN	
	기능	사전에 정의된 순서에 의해 4개의 직업군을 Stepwise로 판정	7개의 직업을 판정

<표 5-2> Model_1에 사용된 변수들

변수	변수설명	변수	변수설명	변수	변수설명
1	주중 근무시간대 통화자크기	8	주중 유선전화발착신비율	15	마일리지합계
2	오전시간대 평균발신호수	9	주중 여가시간대 착신호수	16	요금제도
3	주중 통화자크기비율	10	주중 무선통화자크기비율	17	월평균요금
4	주중 퇴근시간대 통화자크기	11	오전/오후 발신비율차이	18	연령대
5	주중 퇴근시간대 착신율	12	주중 무선/유선통화자크기비율	19	성별
6	주중 점심시간대 착신호수비율	13	주중 타무선발착신비율		· · ·
7	주중 근무시간대 평균발신호수	14	주중 출근시간대 발신비율		

$$Y^* = 0.1 + \frac{Y - \min}{\max - \min} \times 0.8 \qquad \text{식 (5-1)}$$

(Y^* = 정규화한 값, Y = 실제값, max = 실제값 중 최대값, min = 실제값 중 최소값)

범주형 변수들에 대해서는 각 범주별로 Dummy 값을 부여하는 방식으로 전처리 작업을 수행하였다. 그 다음, SOM을 이용하여 7개의 직업을 한번에 판정하였는데, 그 결과 63.1%의 적중률을 보였다. Validation용 데이터에 대한 결과가 <표 5-3>에 나타나있다. 각 셀에 적힌 숫자는 판정개수이고, 괄호 안에 적힌 백분율(%)은 각 직업별 적중률이다.

<표 5-3>에서 보듯이, 내근직과 외근직의 적중률은 70.5%와 79.5%이다. 하지만 나머지 직업들의 적중률은 55.0%이하로서 불균형한 분포를 보인다. 적중률의 분포를 보다 균형적으로 만들기 위해서 대학생과 중고생을 하나의 군(群)으로 묶고 주부, 정주형자영업 그리고 무직을 또 하나의 군으로 묶은 다음에 Model_1의 결과를 재구성하였다. 재구성한 결과는 <표 5-4>에 제시되어 있다.

<표 5-3> Model_1의 각 직업별 판정개수 및 적중률

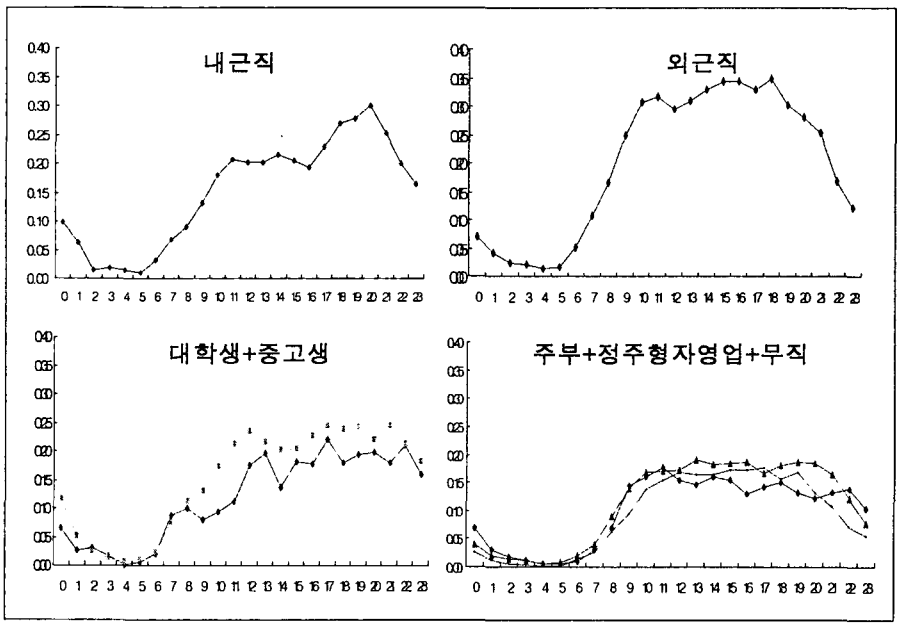
예측 \ 실제	내근직	대학생	중고생	주부	정주형 자영업	무직	외근직	총계
내근직	67 (70.5%)	3		3	3	1		77
대학생	7	11 (55.0%)	3		1	1	2	25
중고생	3	4	5 (50.0%)			1		13
주부	2		1	11 (45.8%)	6	2	3	25
정주형자영업	12			4	9 (45.0%)	1	1	27
무직	2		1	5	1	3 (33.3%)	2	14
외근직	2	2		1			31 (79.5%)	36
총계	95	20	10	24	20	9	39	

<표 5-4> Model_1에서 직업을 군으로 재구성한 후의 적중률

실제 / 예측	내근직	대학생 + 중고생	주부 + 정주형자영업 + 무직	외근직	총계
내근직	67 (70.1%)				77
대학생 + 중고생		23 (76.7%)			38
주부 + 정주형자영업 + 무직			42 (79.2%)		66
외근직				31 (79.5%)	36
총계	95	30	53	39	

<표 5-4>를 보면, 각 직업군을 판정한 적중률이 모두 70%를 넘고, <표 5-3>보다 균형적인 분포를 보이고 있다. 이와 같이 직업군을 형성하는 작업에 대한 타당성을 [그림 5-1]을 통해서 볼 수도 있다. [그림 5-1]은 각 직업군별 시간대별 평균통화건수에 대한 그래프로서 가로는 하루의 시간대를 나타내고 세로는 평균통화건수를 나타낸다.

[그림 5-1]을 보면, 각 직업군간에는 통화패턴이 확연히 구분되지만, 각 직업군 내의 직업들은 비교적 유사한 통화패턴을 가진다는 것을 알 수 있다. 앞으로 본 연구에서는, <표 5-5>에서 보는



[그림 5-1] 직업군별 시간대별 평균통화건수

바와 같이 먼저 4개의 직업군으로 분류하고, 자유직군은 다시 3개의 세부직업으로, 그리고 학생직군은 다시 2개의 세부직업으로 더 분류함으로써 직업 판정의 적중률을 향상시키고자 한다. 이러한 사항을 적용한 첫 번째 모델이 Model_2이다.

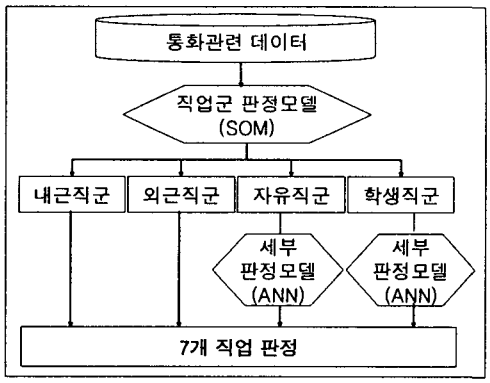
<표 5-5> 직업군의 명칭과 포함된 직업

직업군	포함된 직업
내근직군	내근직
외근직군	외근직
자유직군	주부, 정주형자영업, 무직
학생직군	중고생, 대학생

5.2 Model_2의 구축

[그림 5-2]는 Model_2의 결합모델에 대한 개념도이다. [그림 5-2]에 제시된 것처럼 Model_2는 먼저 직업군을 판정한 후, 각 직업군에 속한 세부직업들을 판정한다. Model_2에서는 직업군 판정모델에서 무감독학습기법인 SOM을 사용하고, 세부직업판정시에는 감독학습기법인 인공신경망을 사용한다.

직업군판정에 사용된 변수들은 Model_1을 구축할 때 사용했던 변수들과 동일하다. Model_2가 4개의 직업군을 판정한 적중률은 70.9%이다.



[그림 5-2] Model_2의 결합모델 개념도

무직, 정주형자영업 그리고 주부가 포함되어 있는 자유직군의 세부직업판정모델에 사용된 변수들은 <표 5-6>과 같다.

자유직군에는 주부가 포함되어 있기 때문에 성별을 입력변수로 선택하였다. 연령대는 20대, 30대 그리고 60대 이상으로 구분하여 Dummy 방식의 변환을 하였다. 무직을 고려해서 전체발신대비 타이동통신발신율과 타이동통신과의 발신 및 착신 통화자크기도 세부직업 판정을 위한 변수에 포함하였다. 휴일저녁시간대 평균발신호수는 자녀를 둔 주부의 경우 유용한 판정자 역할을 할 것으로 기대되는 변수이다. 성별과 연령대를 제외한 나머지 수치형 변수들은 식 (5-1)을 통해 0.1과 0.9 사이의 값으로 정규화를 하였다.

<표 5-6> 자유직군 세부직업판정을 위한 변수들

변수	변수설명	변수	변수설명	변수	변수설명
1	성별	6	퇴근시간대 발신호수비율	11	근무시간대 발신대비평균통화량
2	연령대	7	휴일저녁시간대 평균발신호수	12	오전시간대 통화자크기비율
3	주중 무선통화자크기	8	주중/휴일 저녁시간대 발신호수차	13	타이동통신 착신통화자크기
4	주중 무선발신호수	9	점심시간대 발신호수비율	14	타이동통신발신율
5	휴일 무선발신호수비율	10	타이동통신 발신통화자크기		. . .

<표 5-7> 학생직군 세부직업판정을 위한 변수들

변수	변수설명	변수	변수설명
1	요금제도	5	오후시간대 발신호수비율
2	연령대	6	근무시간대 통화자크기
3	휴일 저녁시간대 유선전화착신호수	7	출근시간대 발신호수
4	근무시간대 평균발신호수		...

대학생과 중고생이 포함되어 있는 학생직군을 세분하기 위해 사용한 변수들이 <표 5-7>에 제시되어 있다.

학생직군 세부직업 판정모델의 입력변수 생성 과정은 다음과 같다. 요금제도의 경우는 발신량에 제한이 있는 요금제인지 아닌지에 따라 Dummy 방식의 변환을 하였다. 연령대는 대학생과 중고생이 구분될 수 있는 좋은 지표이기 때문에 연령대가 10대인지 여부에 따라 Dummy 방식의 변환을 하였고, 나머지 수치형 변수는 식 (5-1)에 따라 정규화를 하였다. 출력노드는 하나로서 직업이 대학생인지 아닌지를 나타낸다. Model_2의 결과는 <표 5-8>과 같다.

<표 5-8> Model_2의 적중률

단위(%)

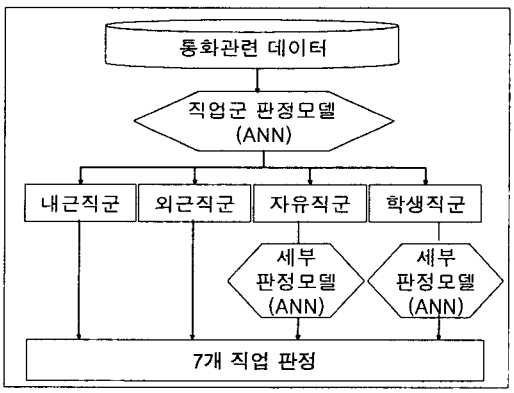
		직업군 판정모델	직업군별 세부직업판정 모델	최종적중률
외근직군		76.9	·	76.9
내근직군		74.7	·	74.7
학생직군	대학생	73.3	85.7	60.0
	중고생		75.0	60.0
자유직군	주부	58.5	83.3	41.7
	정주형 자영업		71.4	50.0
	무직		60.0	33.3
최종적중률		70.9	77.4	**65.4**

<표 5-8>에서 보는 것처럼 Model_2의 최종적중률은 65.4%로서, Model_1의 63.1%에 비해 다소 향상된 적중률을 보이고 있다.

직업판정모델의 적중률을 향상시키기 위해서, 추가적으로 고려해 볼 수 있는 방법은 우리가 Model_2에서 SOM으로 분류해낸 직업군을 이미 알고 있는 사실로 받아들이고 감독학습기법을 적용하는 것이다. 이러한 사항을 반영한 것이 Model_3이다.

5.3 Model_3의 구축

[그림 5-3]은 Model_3의 결합모델에 대한 개념도이다. Model_3에서는 Model_2와 달리 직업



[그림 5-3] Model_3의 결합모델 개념도

군을 판정하는 모델에 감독학습을 하는 인공신경망을 사용하였다. 세부직업 판정모델은 Model_2와 동일한 기법과 구조를 사용하였다.

직업군판정에 사용된 입력변수는 앞서 제시한 <표 5-2>와 동일하다. Mode_3이 4개의 직업군을 판정한 적중률은 74.7%이다.

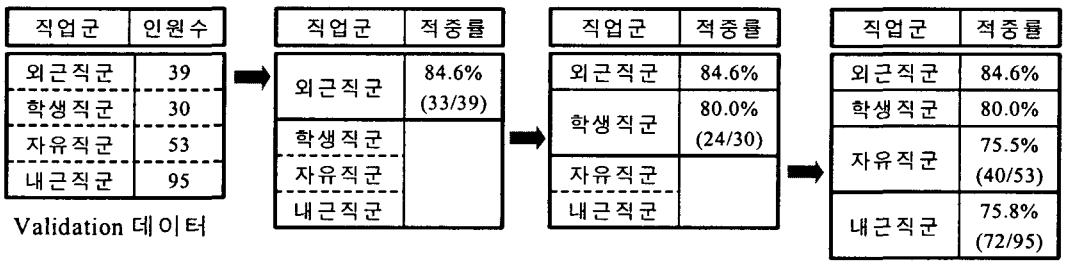
직업군 판정모델로 판정한 네 개의 직업군 가운데 자유직군과 학생직군에 대해서 세부직업판정모델을 구축한다. 자유직군과 학생직군의 세부직업 판정모델을 구축하기 위해 사용한 입력변수와 인공신경망의 구조는 Model_2에 사용한 것과 동일하다.

Model_3의 최종적중률은 <표 5-9>에 제시된 것처럼 69.1%이다. Model_3의 적중률은 Model_1보다 6.0% 포인트 향상되었고, Model_2보다도 3.7% 포인트 향상되었다.

<표 5-9> Model_3의 최종적중률

단위(%)

		직업군 판정모델	직업군별 세부직업판정 모델	최종적중률
외근직군		82.1	·	82.1
내근직군		71.6	·	71.6
학생직군	대학생	80.0	86.7	65.0
	중고생		88.9	80.0
자유직군	주부	77.5	82.3	58.3
	정주형 자영업		80.0	60.0
	무직		50.5	33.3
최종적중률		74.7	80.1	**69.1**



[그림 5-4] Model_4의 직업군 판정과정

<표 5-10> Model_4의 최종적중률

		직업군 판정순서	직업군판정모델 적중률 (%)	직업군별 세부직업 판정모델 적중률 (%)	최종적중률 (%)
외근직군		1	84.6	·	84.6
학생직군	대학생	2	85.7	73.7	70.0
	중고생			77.8	70.0
자유직군	주부	3	72.7	53.8	58.3
	정주형자영업			63.2	60.0
	무직			40.0	44.4
내근직군		4	75.8	·	75.8
최종적중률			77.8	79.7	**71.9**

5.4 Model_4의 구축

Model_4는 직업군을 판정할 때, Model_3에서 직업군을 판정해낸 결과를 이용한다. <표 5-9>에 따르면, 적중률이 높은 직업군의 순서는 외근직군(82.1%), 학생직군(80.0%), 자유직군(77.5%) 그리고 내근직군(71.6%)이다. Model_4는 이 순서에 따라 감독학습을 하는 인공신경망을 사용하여 직업군을 단계적(Stepwise)으로 판정한다. 즉, 제 1 단계에서는 전체 고객을 외근직군과 나머지(편의상 나머지_1군으로 부르기로 함)로 판정하고, 제 2 단계에서는 나머지_1군을 학생직군과 나머지_2군으로 판정하고, 마지막 제 3 단계에서는 나머지_2군을 자유직군과 나머지_3군으로 판정하는데, 이 나머지_3군이 바로 내근직군이 되는 것이다. 직업군이 판정되면, Model_3과 마찬가지로 각 직업군에 속한 세부직업을 판정한다.

직업군 및 세부직업판정에 사용한 입력변수들은 Model_3을 구축하는데 사용한 것과 동일하다. 직업군 판정모델의 결과도출과정을 [그림 5-4]가 보여주고 있으며, Model_4의 최종적중률은 <표 5-10>과 같다.

Model_4의 최종적중률은 71.9%이다. 이 적중률은 Model_2 보다는 6.5% 포인트, Model_3 보다는 2.8% 포인트 향상된 것이다.

5.5 구축된 모델들의 종합 평가

본 연구에서 구축된 각 모델의 직업별 적중률 및 최종적중률은 <표 5-11>과 같다.

<표 5-11>에서 보듯이, Stepwise기법을 적용한 Model_4는 Model_1보다 8.8% 포인트, Model_2보다 6.5% 포인트, 그리고 Model_3보다 2.8% 포인트 향상된 적중률을 보였다. 단순히 수치적인 의미에서는 근소한 차이지만 현실에서는 전체 고객의 규모 등에 따라 매우 큰 차이일 수 있다.

또한 본 연구에서 구축한 모델들은 단일모델과 결합모델의 비교 관점에서도 평가할 수 있는데, Model_1은 단일모델로서 적중률이 63.1%이지만 결합모델인 Model_4의 적중률은 71.9%로서, 앞서 언급한 바와 같이 8.8% 포인트 향상된 적중률을 보이고 있다.

6. 결론

이 논문에서 우리는 이동통신서비스 산업에서의 고객관계관리 구현에 데이터마이닝을 활용하는 연구를 수행하였다. 고객관계관리를 구현하기 위해서는 무엇보다도 고객에 대한 정확한 이해가 필수적이다. 고객의 직업은 그 고객을 이해하는데 필요한 중요한 속성이다. 하지만 본 연구에 사용된 이동통신 데이터에 포함된 직업 기입율은 7%에 불과하다. 고객을 이해하는데 활용하기에는 불가능할 정도로 직업 기입율이 낮은 대표적 이유

<표 5-11> 각 모델의 직업별 적중률 및 최종적중률

단위(%)

	내근직	대학생	무직	외근직	정주형자영업	주부	중고생	최종적중률
Model_1	70.5	55.0	33.3	79.5	45.0	45.8	50.0	63.1
Model_2	74.7	60.0	33.3	76.9	50.0	41.7	60.0	65.4
Model_3	71.6	65.0	33.3	82.1	60.0	58.3	80.0	69.1
Model_4	75.8	70.0	44.4	84.6	60.0	58.3	70.0	71.9

는 대부분의 고객이 자신의 직업이 알려지는 것을 꺼린다는 것이다. 또한, 직업데이터는 영구적인 것이 아니라 고객의 상황에 따라서 비정기적으로 변할 수 있다. 그러므로 이동통신서비스 업체는 직업데이터를 주기적으로 갱신하여야만 한다. 하지만 고객과 직접 접촉하여 정확한 직업을 확인한 후에 갱신하는 것은 거의 불가능할 뿐만 아니라, 가능하다해도 매우 많은 비용을 수반하게 된다.

고객이 자신에 대한 데이터를 직접 공개하지 않을 경우에, 우리는 고객이 남긴 통화상세이력 데이터를 통하여 고객을 파악할 수 있다. 즉, 본 연구에서는 통화상세이력 데이터로부터 통화패턴에 대한 정보들을 요약변수와 파생변수의 형태로 생성한 후에, 다각도로 인공신경망기법을 적용하여 고객의 직업을 판정하는 모델을 구축하였다. 여러 가지 다양한 모델들을 구축해본 결과, 고객 직업판정의 적중률을 71.9%까지 얻을 수 있었다. 고객의 직업이 판정되면, 각 직군에 따라 상이한 마케팅을 전개함으로써 불필요한 마케팅 비용을 절감할 수 있고, 보다 정확한 타게팅이 가능해지므로 신상품개발에도 유용하게 활용될 수 있을 것이다.

본 연구의 한계점으로는 적용기법의 단순성과 변수선정의 미비를 들 수 있다. 데이터마이닝시스템을 개발하는 데는 다양한 기법들이 사용될 수 있다. 본 연구에서는 단일기법, 즉 인공신경망만을 사용하였으나, 여러 가지 다양한 기법들을 사용한 결합모델을 개발함으로써 좀더 높은 적중률 향상을 꾀할 수가 있을 것이다. 또한, 어느 모델에서건 변수선정의 문제는 중요한 것인데 본 연구에서는 변수선정 부분을 다루지 않았다. 통계학적인 또는 다른 기계학습기법들을 통한 변수선정 단계를 거침으로써 시스템의 효율성을 꾀할 수도 있을 것이다.

참고문헌

[1] 노동통계정보망 홈페이지, http://laborstat.molab.go.kr, 2004년 11월.

[2] Berry M. J. A. and G. S. Linoff, *Mastering Data Mining*, John Wiley & Sons, Inc., 2000.

[3] Hahm J., W. Chu. and J. W. Yoon, "A Strategic Approach to Customer Satisfaction in the Telecommunication Service Market," *Computers & Industrial Engineering*, Vol. 33, 1997.

[4] Hwang H., T. Jung and E. Suh, "An LTV Model and Customer Segmentation based on Customer Value: A Case Study on the Wireless Telecommunication Industry," *Expert Systems with Applications*, Vol. 26, 2004.

[5] Kohonen T., "The Self-Organizing Map," *Proceedings of the IEEE*, 1990.

[6] McMullen P. R., "A Kohonen Self-Organizing Map Approach to Addressing a Multiple Objective, Mixed-model JIT Sequencing Problem," *International Journal of Production Economics*, Vol. 72, 2001.

[7] Mozer C. M., W. Richard and D. B. Grimes, "Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry," *IEEE*

Transaction on Neural Networks, Vol. 11, 2000.

[8] Nelson M. and W. T. Illingworth, *A Practical Guide to Neural Nets*, Addison-Wesley Publishing Company, Inc., 1991.

[9] Yang C., H. Chen and K. Hong, "Visualization of Large Category Map for Internet Browsing," *Decision Support Systems*, Vol. 35, 2003.

Abstract

Customer's Job Identification using the Usage Patterns of Mobile Telecommunication

Jae Sik Lee* · You Jung Cho*

Recently, as most companies recognize the importance of the customer relationship management, they strongly believe that they must know who their customers are. The job of a customer is very important information for us to understand the customer. However, since most customers are reluctant to reveal themselves, they do not let us know their jobs, and even provide false information about their jobs. The target domain of our research is mobile telecommunication. In this research, we developed a system that identifies the customer's job by utilizing the Call Detail Record. Using artificial neural networks, we developed a two-step Job Identification System. In the first step, it identifies the four job classes, then in the second step, it subdivides these four job classes into seven jobs. The accuracy of identifying the seven jobs was 71.9%.

Key Words : Customer Relationship Management, Customer's Job Identification, Self-Organizing Map, Artificial Neural Networks, Mobile Telecommunication Industry.

* Department of Business Administration, Ajou University