Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 109-120

연관분석을 이용한 데이터마이닝 기법에 관한 사례연구

류귀열[1] · 문영수[2] · 최승두[3]

요 약

본 연구에서는 RFM 분석을 통하여 전체 고객들을 점수화(scoring)하고 이를 다시 5개의 그룹(최우수그룹, 우수그룹, 일반그룹, 하위그룹, 최하위그룹)으로 세분화하고, 세분그룹별 유의성을 검정한다. 이렇게 분류된 5개의 세분화그룹들은 연관분석과 의사결정나무 등을 통하여 고객들의 인구학적 변수와 각 그룹별 유의한 변수들의 패턴을 찾아냄으로써 우수 고객들을 유지하기 위해서는 어떻게 해야 하며, 경쟁업체로 떠날 가능성이 높은 고객은 누구이며, 이러한 이유가 무엇인지에 대하여 효과적인 분석을 할 수 있는 기반이 조성된다. 본 연구의 목적은 통하여 연관규칙(association rules)과 의사결정나무(decision tree)를 비교 분석을 함으로써, 이론적으로 설명할 수 없는 복잡한 세분그룹의 특성들에 대해 효과적으로 파악하는 방법을 제시하는 것이다.

주요용어 : 데이터마이닝, RFM, 연관분석, 의사결정나무.

1. 서론

1.1 데이터마이닝

데이터마이닝(data mining)이란 지식 발견(knowledge discovery in database), 정보 발견(information discovery), 정보 수확(information harvesting)등으로 알려진 자동화되고 지능을 갖춘 데이터 분석 기법이다(안광호 등, 2001). 데이터마이닝은 대용량의 데이터들로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화함으로써 유용한 지식을 추출하는 일련의 과정들이다(강현철 등, 2001). Herb Edelstein은 데이터마이닝이란 "기업이 보유하고 있는 거래 데이터, 고객 데이터, 상품데이터, 각종 마케팅 활동자료, 외부 자료 등을 포함한 모든 데이터를 기반으로 그동안 발견하지 못했던 데이터간의 관계, 패턴을 탐색하고 이를 모형화하여 업무에 적용할 수 있는 정보로 변환함으로써 기업의 전략적 마케팅 의사결정에 적용하는 일련의 프로세스"라고 정의할 수 있으며, 간단히 말해서, 데이터마이닝은 데이터에 숨어 있는 패턴과 관계를 발견하여 더 나은 비즈니스 의사 결정을 돕는 데 사용되는 툴이라고 정의하였다.

[1]136-704 서울시 성북구 정릉동 산16-1, 서경대학교 인터넷정보학과 부교수. E-mail : gyryu@skuniv.ac.kr
[2]130-742 서울시 동대문구 청량리동 206-9, 한국과학기술정보연구원 연구원. E-mail : youngsum@kisti.re.kr
[3]614-714 부산시 부산진구 가야동 산24, 동의대학교 재무부동산학과 조교수. E-mail : mm@deu.ac.kr

이러한 데이터마이닝의 기술에는 여러 가지가 있는데 연관규칙, 순차패턴, 분류, 군집화, 이상점 판별 등이 있으며, 연관규칙은 데이터마이닝을 소개할 때 대표적으로 언급되는 기술로 백화점이나 슈퍼마켓에서 한번에 함께 산 물건들에 관한 연관 규칙을 찾아내는 기술이다. 연관 규칙은 물건을 한 번에 살 때 같이 구매한 것들을 이용해 규칙을 찾는 것인 반면, 순차 패턴 발견은 순서대로 일어나 데이터를 분석해 빈도수가 높은 순차 패턴을 찾아내는 기술을 말한다. 분류는 주어진 데이터와 각각의 데이터에 대한 클래스가 주어진 경우, 그것을 이용해 각각의 클래스를 갖는 데이터들은 어떤 특징이 있는지 분류 모델을 만들고, 새로운 데이터가 있을 때 그 데이터가 어느 클래스에 속하는지 예측하는 것을 뜻한다. 군집화란 주어진 n개의 점을 K개의 그룹으로 나누는 것을 말한다. 분류와 다른 점은 각 클래스에 해당되는 정보가 제공되지 않는다는 것이다. 군집화 기술은 전체 데이터의 분포 상태나 패턴 등을 찾아내는 데 유용하게 이용할 수 있다. 대부분의 데이터마이닝 기술은 데이터를 나타내는 패턴에 관심을 갖고 찾아내려 한다. 하지만 이상점 판별 기법은 이와 반대로 대부분의 데이터와 다른 소수 또는 일부를 찾아내는 기술이다.

1.2 연관분석

여기에서는 본 연구에 이용되는 연관분석에 대해 간략히 알아본다. 연관분석의 개념은 데이터 안에 존재하는 품목 간의 연관규칙(association rules)을 발견하는 과정으로 두 가지 항목의 연관성을 통하여 한 항목의 값을 알 경우 다른 항목의 값을 예측해 내는 방법이다. 즉 연관성분석은 거래내역에 대한 자료는 있지만 찾고자 하는 특정한 규칙을 모를 때 분석의 초기 작업으로 이용되는 데이터마이닝 기법 중 하나로써 무엇을 해야 할지 모를 때 유용하게 쓰여 질 수 있는 방법이다.

Agrawal et al.(1993)은 연관규칙을 만들어 내기 위한 알고리즘을 소개하였다. 각 X와 Y는 항목들(items)의 집합이고, 연관규칙은 X→Y로 나타낼 수 있다. 이러한 규칙의 의미는 X를 포함하는 사건들은 Y도 포함하는 경향이 있다는 것이다. 그러나 이러한 규칙들은 실제 데이터를 분석할 때 100% 신뢰도를 가지는 경우가 드물기 때문에 통상적으로 확률이나 도표 등을 이용하여 정량화 한다. 연관규칙 기술을 적용할 때 두 가지 설정값을 입력해야 하는데 이들은 지지도(support)와 신뢰도(confidence)라 불린다. 어떤 규칙의 지지도가 10%라면 그 의미는 전체 트랜잭션이 10%를 차지한다는 것을 의미한다.

<표 1> 슈퍼마켓 구매물품자료

고객번호	품 목
1	오렌지주스, 사이다
2	우유, 오렌지주스, 식기세척기
3	오렌지주스, 세제
4	오렌지주스, 세제, 사이다
5	식기세척기, 사이다

예를 들어, <표 1>에 있는 5번의 거래에서 사이다와 오렌지주스가 함께 거래된 경우는 2번이다.

즉 2번의 거래가 "사이다를 구입하는 고객은 오렌지주스를 산다." 라는 규칙을 지지한다고 할 수 있다. 따라서 이 규칙에 대한 지지도는 5번의 거래에서 2번만이 동시 구매가 이루어졌기 때문에 2/5 혹은 백분율로 표현하면 40%가 될 것이다.

$$\text{지지도(support, S)} = P(A \cap B)$$

그리고 신뢰도는 규칙의 왼쪽에 있는 것을 산 사람들 중에서 오른쪽에 있는 물건들을 모두 산 사람들의 퍼센트를 말한다. <표 1>을 보면 사이다를 포함하는 3번의 거래에서 2번의 거래는 오렌지주스를 포함하고 있다. 이런 경우에 "사이다를 사면 오렌지주스를 산다"는 규칙은 66%의 신뢰도를 갖는다고 말한다. 그런데 "오렌지주스를 사면 사이다를 산다"는 역 규칙은 4번의 오렌지주스 구매에 대하여 단지 2번의 사이다 구매가 동시에 이루어졌기 때문에 신뢰도가 50%로 떨어지게 된다. 즉 신뢰도는 지지도와 달리 상호 대칭적이지 못하다. 신뢰도는 조건의 구매가 일어난 경우에, 결과의 구매가 일어난 비율(조건부 확률)이며 신뢰도를 수식으로 표현하면 다음 식과 같이 나타낼 수 있다.

$$\text{신뢰도(confidence, C)} = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

그리고 어떤 규칙에 대하여 조건이 없을 때 기대되어지는 결과보다 조건이 추가될 때의 결과가 얼마나 더 향상되었는지를 알려주는 측도로 향상도(lift)가 있으며 수식으로 표현하면 다음 식과 같이 나타낼 수 있다.

$$\text{향상도(lift, L)} = \frac{P(X \cap Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)} = \frac{C}{P(Y)}$$

품목 X와 Y의 구매가 상호 관련이 없다면 P(Y|X)는 P(Y)와 같게 되어 향상도가 1이 된다. 만약 어떤 규칙에 대한 향상도가 1보다 크다면, 이 규칙은 결과를 예측하는데 있어서 우연적 기회(random chance)보다 우수하다는 것을 의미한다.

이러한 연관규칙 알고리즘은 Apriori, AprioriTid, DHP, Partition, Sampling, FUP, DIC 등이 있다. 이중 연관 규칙을 찾아주는 알고리즘 중에서 가장 먼저 개발됐고, 또 가장 많이 쓰이는 알고리즘은 Apriori 알고리즘이다.

2. 실험 분석

2.1 자료의 개요

본 연구에서 사용되는 자료는 K기관에서 운영하고 있는 홈페이지의 고객자료를 바탕으로 고객분석을 실시하고, 분석된 고객들의 인구학적 특성을 통하여 고객 세분화를 함으로써 효율적인 마

케팅 전략을 돕고자 한다. 특히 본 연구는 RFM과 연관분석기법을 함께 적용함으로써 보다 효율적인 모형을 구축함을 목적으로 한다.

고객 분석 및 세분화를 위해 사용된 자료는 2003년에서 2004년까지 두 해 동안 고객으로부터 수집된 자료이며 크게, 고객데이터와 거래 데이터로 구성되어 있다. 먼저 고객 데이터란 회원고객들의 정보로써 고객 ID, 나이, 성별, 직장, SSO 가입사이트 수, 회원구분, 지역, 주 이용 메일호스트, 부서, 학위, 전공, 세부 전공분야 등으로 구성되어 있으며, 거래 데이터란 각 회원 고객들의 거래 내역에 관한 정보로써 구매일자, 구매건수, 구매금액, 신청방법 등으로 구성되어 있으며, 이용된 자료는 4,397건이다. <표 2>는 분석에 사용된 변수들로 고객들의 인구통계학적인 요소 및 구매자료를 요약한 표이다.

<표 2> 분석대상 변수설명

변수명	설 명	내 용
A1	이용자 ID	홈페이지 접속 고유번호
A2	회원구분	0: 국내거주 1:국외거주 2:기업
A3	성별	1: 남자 2:여자
A4	지역	전국을 16개 범주로 구분(1~16)
A5	나이	1: 30대 이하 2: 30대 초반 3: 30대 후반 4: 40대 초반 5: 40대 후반 6: 50대 이상
A6	직업	산업분류에 따른 96개 직업
A7	신청방법	1: 검색의뢰 2: 웹신청
A8	이메일 서비스 도메인	14개 도메인으로 구분
A9	부서	0: 없음 1: 연구/개발 2: 기획 3: 마케팅 4: 생산 5: 영업 6: 관리/행정 7: 전산 8: 전산 9: 기술관리 10: 기타
A10	학위	1: 없음 2: 학사 3: 석사 4: 박사
A11	전공	64개 전공으로 분류
A12	SSO 총 가입수	K기관 가입 사이트 수
B1	최근 구매일자	범위 : 1 ~ 24개월
B2	총 구매건수	단위 : 건
B3	총 구매금액	단위 : 원

2.3 RFM 설정

2.3.1 R값 설정

R값의 정의는 고객들의 구매데이터를 최근성에 근거하여 5개의 세그먼트로 분류하였다. 최근 구매시기가 1개월 이내인 경우를 가장 높은 가중 값 5로 설정하고, 최근 구매시기가 16개월 ~ 24개월까지를 경우를 가장 낮은 가중 값 1로 설정하였다. <표 3>은 최근 구매시기에 대한 R값의 설정을 나타내고 있다.

2.3.2 F값 설정

F값의 정의는 고객들의 총 구매수량에 근거하여 5개의 세그먼트로 분류하였다. 총 구매수량이

37개 이상이면 가장 높은 가중 값 5로 설정하고, 총 구매수량이 1개이면 가장 낮은 가중 값 1로 설정하였다. <표 4>는 총 구매수량에 대한 F값의 설정을 나타내고 있다.

<표 3> R값 설정 및 분포

최근 구매시기	R값	빈도	%
1개월 이내	5	1,188	27.0
2개월 ~ 3개월	4	872	19.8
4개월 ~ 6개월	3	591	13.4
7개월 ~ 15개월	2	881	20.0
16개월 ~ 24개월	1	865	19.7
합 계		15	100

<표 4> F값 설정 및 분포

총 구매수량	R값	빈도	%
37개 이상	5	874	19.9
21개 ~ 36개	4	448	10.2
11개 ~ 20개	3	595	13.5
2개 ~ 10개	2	1,754	39.9
1개	1	726	16.5
합 계		15	4,397

2.3.3 M값 설정

M값의 정의는 고객들의 총 구매금액에 근거하여 5개의 세그먼트로 분류하였다. 총 구매금액이 313,900원 이상이면 가장 높은 가중 값 5로 설정하고, 총 구매금액이 4,000원 이하이면 가장 낮은 가중 값 1로 설정하였다. <표 5>는 총 구매금액에 대한 M값의 설정을 나타내고 있다.

<표 5> F값 설정 및 분포

총 구매금액	R값	빈도	%
313,900원 이상	5	439	10.0
83,000원~313,100원	4	878	20.0
18,100원~82,900원	3	1,312	29.8
4,100원~18,000원	2	1,078	24.5
4,000원 이하	1	690	15.7
합 계		15	4,397

본 연구에서는 각각의 R값, F값, M값을 5개의 세그먼트로 분류하여 전체 고객을 125개의 세그먼트(계층)로 나누려고 하였으나, 빈도가 1 이상인 유효한 세그먼트 계층은 82개가 생성되었으며, 가장 낮은 RFM값은 111이며, 가장 높은 RFM값은 555를 가지고 있다.

2.4 스코어링 함수(scoring function)를 이용한 세분화

점수화(scoring)는 분석결과를 보다 쉽고 직관적으로 알기 쉽게 하기 위하여 고객을 대상으로 점

수를 부여하는 것을 말한다. 본 연구에서는 RFM 분석결과에 K기관의 특수성을 고려하여 각각의 R값, F값, M값에 가중치를 주어 스코어링 함수를 산출하였으며 공식은 아래의 식과 같다.

$$\text{고객의 평균 RFM score} = 0.1 \times R + 0.2 \times F + 0.7 \times M$$

$$\text{RFM score} = (\text{평균 RFM score} \times 100) / 5$$

위의 식에서 각각의 R값, F값, M값에 적용된 가중치 값(0.1, 0.2, 0.7)은 RFM 분석에 있어 가장 중요한 문제로 다양한 방법을 통하여 얻어질 수 있다. 가중치를 구하는 기존의 연구들을 살펴보면, 파레토 법칙의 상위 20%고객의 평균값을 기준으로, 실제 점수화를 실시하였을 때의 평균값과 비교하여 가장 유사한 수준에서 결정하는 방법(이강태, 2002), 통계적 기법중 하나인 다중회귀분석을 통해 추정하는 방법(김승아, 1998), RFM의 가중치를 각각 50%, 35%, 15%로 산정하는 방법(R.S. Hodeson, 1980), RFM의 세 가지 요인에 고객의 구입행동을 평가하는 요인 즉, 고객이 어떠한 종류의 상품을 구입했는가, 고객의 지불방법, 주거지 등을 고려하여 평가하는 방법(B.Stone, 1984) 등이 있으나 ,세 요소에 점수를 부여하는 기준인 가중치가 상품이나 시장 환경에 따라 많은 차이를 보이기 때문에 가장 예측력 있는 현실적인 가중치를 알아내는 것이 RFM 점수를 산출하기 위한 핵심 문제라고 할 수 있으므로, 각 기업에서의 중요도에 따라 가중치를 부여하는 것이 일반적인 현상이다.

본 연구에서도 K기관의 특수성과 고객분석의 결과에 따라 총 구매금액을 가장 높은 0.7의 가중치를 부여하였으며, 다음으로 F값과 R값에 0.2와 0.1의 가중치를 부여하였다. 이러한 가중치는 각 기업에서 시뮬레이션을 통해 이상적인 가중치를 찾을 수도 있을 것이다.

위의 가중치를 통하여 얻어진 점수는 최저 20점에서 최고 100점까지의 점수로 환산되어지는데, 본 연구에서는 5개의 세분화 그룹으로 분류하기 위하여 80점에서 100점 사이의 고객을 최우수그룹으로 분류하고, 60점에서 79점 사이의 고객을 우수그룹으로, 40점에서 59점 사이의 고객을 일반그룹으로, 21점에서 39점 사이의 고객을 하위그룹으로, 그리고 20점인 고객을 최하위그룹으로 분류하였다. 5개 세분 그룹은 <표 6>과 같이 나타나 있다.

<표 6> 세분화 그룹에 대한 R, F, M값의 평균 비교

segment group		최근 구매시기 (R값)	총 구매수량 (F값)	총 구매금액 (M값)
최우수그룹	922명(21%)	1.54	124.16	533,296
우수그룹	837명19%)	3.79	22.17	92,551
일반그룹	1,224명(28%)	6.40	7.17	31,426
하위그룹	1,232명(28%)	13.67	2.83	11,047
최하위그룹	181명(4%)	20.12	1	2,138
전체그룹	4,396명	7.49	33.09	141,407

분산분석을 통해 5개 세분그룹간 차이에 대한 유의수준이 .000임을 알 수 있으며, Tukey나 LSD를 이용한 그룹 차이도 유의한 것으로 나타났다.

2.5 연관분석

연관분석에서 사용하는 고객데이터는 <표 3>의 변수 중에서 범위가 큰 변수들(직업, 전공, 이메일 도메인)을 상위 개념으로 통합하여 변수의 범위를 줄였으며, 최종적으로 사용된 변수는 <표 7>과 같다.

<표 7> 연관분석용 변수

변수명	설 명	내 용
A0	이용자 ID	홈페이지 접속 고유번호
A1	회원구분	0: 국내거주 1:국외거주 2:기업
A2	성별	1: 남자 2:여자
A3	지역	전국을 16개 범주로 구분(1~16)
A4	연령대	1: 30대 이하 2: 30대 초반 3: 30대 후반 4: 40대 초반 5: 40대 후반 6: 50대 이상
A5	직업	1: 공공부문 2: 교육/연구부문 3: 대기업 4: 중소기업 5: 의료/제약 6: 농림/수산/광업 7: 전문직 및 기타
A6	신청방법	1: 검색의뢰 2: 웹 신청
A7	이메일 도메인	1: 직접입력 2: 내부고객 3: 포털 이용
A8	부서	0: 없음 1: 연구/개발 2: 기획 3: 마케팅 4: 생산 5: 영업 6: 관리/행정 7: 전산 8: 전산 9: 기술관리 10: 기타
A9	학위	1: 없음 2: 학사 3: 석사 4: 박사
A10	전공	1: 인문사회 2: 자연과학 3: 공학 4: 의약학 5: 농수해양 6. 기타
A11	SSO 총가입 수	1: 3개이하 2: 4개 3: 5개 4: 6개이상
A12	R 값	범위 : 1 ~ 5
A13	F 값	범위 : 1 ~ 5
A14	M 값	범위 : 1 ~ 5
A15	고객 그룹	총 5개의 고객 그룹

연관규칙의 탐사는 통계적 접근방식에 의해 대용량의 사건 데이터베이스에서 규칙성을 추출하고, 많은 의외의 규칙들이 발견해내고 있다. 이런 점에서 연관규칙은 데이터마이닝의 한 분야로 산업분야전반에서 적용이 가능하다.

본 연구에서 고객별 레코드 한 건은 한 트랜잭션으로 볼 수 있고, 하나의 사건에 포함된 각각의 변수들은 한 트랜잭션 내의 각각의 항목(item)이 된다. 연관규칙을 시행함에 있어 한 트랜잭션에서 중복되는 항목은 존재하지 않는다고 가정하고, 그 항목들을 일정한 순서로 정렬되었다고 가정을 할 때 조사자가 명시한 최소의 지지도 이상을 갖는 트랜잭션의 집합을 빈번한 항목집합이라고 한다.

전체 4,397명의 데이터 중 결측치를 제외한 트랜잭션의 수가 약 500개에 지나지 않아 의미가 있는 연관관계를 가지는 지지도의 수준을 5%로 설정하였다. 목표변수가 2개 이상인 연관규칙을 채택하였으며, 임계치(threshold)는 지지도 5%, 신뢰도 50%로 지정하여 규칙을 도출하였다. 그 결과 총 70,506개의 연관규칙이 도출되었고, 임계치를 만족하는 연관규칙은 4,592개이다. 임계치를 만족하는 연관규칙이 적은 이유는 연관규칙은 설명변수와 목적변수의 구별이 없이 규칙들을 추출하기 때문

이다.

연관분석 결과 전체 5개의 세분화된 그룹들에 대하여 유의한 연관규칙이 도출된 그룹은 최우수그룹, 우수그룹, 일반그룹, 하위그룹 이였으며 최하위그룹에 대한 연관규칙은 도출되지 않았다.

2.5.1 최우수그룹 연관규칙

최우수그룹(ST5)에 대한 연관규칙은 모두 148개가 발생하였으며, 요인으로는 SSO 가입사이트 3개 이하(M1), 이메일 호스트는 직접입력(D1), 개인 이용자(O1), 검색방법에 의한 원문신청(I1), 교육/연구기관 소속(F2), R값은 5(E5), F값은 5(J5), M값은 4~5(G4, G5) 이다.

최우수그룹은 교육/연구기관에서 근무하는 개인 이용자로, 주 이용 사이트는 대표홈페이지를 포함하여 3개 이하를 이용하고 있으며, 정보검색 활용에 능하며, 직장 이메일을 통하여 개인화 서비스를 제공받고 있는 고객들로 이루어져 있다.

<표 8> 최우수그룹에 대한 연관규칙 결과

향상도	지지도	신뢰도	연관규칙				
2.92	11.10	61.23	M1	E5	D1	==>	ST5
4.56	10.48	95.64	O1	J5	I1	==>	ST5
2.58	5.00	54.05	G4	F2		==>	ST5
4.47	18.63	93.71	G5			==>	ST5

2.5.2 우수그룹 연관규칙

우수그룹(GT4)에 대한 연관규칙은 모두 243개가 발생하였으며, 요인으로는 남자(L1), 이메일 호스트는 직접입력(D1), 개인 이용자(O1), SSO 가입사이트는 3개 이하(M1), R값은 5(E5), F값은 3~4(J3, J4), M값은 3(G3) 이다. 우수그룹은 남자인 개인 이용자로, 주 이용 사이트는 대표홈페이지를 포함하여 3개 이하에 가입되어 있으며, 직장 이메일을 통하여 개인화 서비스를 제공받고 있는 고객들로 이루어져 있다.

<표 9> 우수그룹에 대한 연관규칙 결과

향상도	지지도	신뢰도	연관규칙				
5	5.3	95.1	L1	G3	E5	==>	ST4
3.07	5.25	58.48	O1	J3	D1	==>	ST4
3.87	7.51	73.66	M1	J4		==>	ST4

2.5.3 일반그룹 연관규칙

일반그룹(GT3)에 대한 연관규칙은 모두 182개가 발생하였으며, 요인으로는 남자(L1), 개인 이용자(O1), SSO 가입사이트는 3개 이하(M1), 웹을 통한 복사신청(I2), R값은 3(E3), F값은 2(J2), M값은 3(G3) 이다. 일반그룹은 남자인 개인 이용자로, 주 이용 사이트는 대표홈페이지를 포함하여 3개 이하의 사이트에 가입하였으며, 정보검색 활용 능력이 다소 떨어지는 고객들로 이루어져 있다.

<표 10> 일반그룹에 대한 연관규칙 결과

향상도	지지도	신뢰도	연관규칙				
1.74	15.44	55.47	M1	L1	J2	==>	ST3
1.94	5.05	54.01	O1	I2	E3	==>	ST3
1.68	13.99	50.25	G3			==>	ST3

2.5.4 하위그룹 연관규칙

하위그룹(GT2)에 대한 연관규칙은 모두 328개가 발생하였으며, 웹을 통한 복사신청(I2), 개인 이용자(O1), 교육/연구기관 소속(F2), 이메일 호스트는 직접입력(D1), R값은 1~2(E1, E2), F값은 1~2(J1, J2), M값은 1~2(G1, G2) 이다. 하위그룹은 교육/연구기관에 종사하는 개인 이용자로, 주 이용 사이트는 대표홈페이지를 포함하여 3개 이하의 사이트를 이용하고 있으며, 정보검색 활용 능력이 다소 떨어지며, 직장 이메일을 통하여 개인화 서비스를 제공받고 있는 고객들로 이루어져 있다.

<표 11> 하위그룹에 대한 연관규칙 결과

향상도	지지도	신뢰도	연관규칙				
1.6	14.31	51.06	M1	L1	J2	==>	ST2
2.29	5.34	64.03	I2	E2	D1	==>	ST2
2.12	5.89	59.27	O1	F2	E1	==>	ST2
2.03	5.14	56.78	G2	A1		==>	ST2
2.1	6.94	58.88	J1	G1		==>	ST2

2.6 의사결정나무

고객 프로파일링(customer profiling)이란 고객이 갖고 있는 라이프스타일에 따라 기업에 있어 고객과의 커뮤니케이션(communication)을 할 수 있는 토대를 제공해 주는데, 고객 세분화 과정 후에 수행이 된다(Bounsaythip et al., 2001). RFM 분석을 통하여 고객을 세분화한 자료를 기초로 데이터마이닝 기법들을 통하여 추론규칙(induction rule)에 의해서 분류(classification)를 수행한 결과물을 갖고 고객 프로파일링을 실시하고자 한다.

여기에서는 의사결정나무 분석을 통한 고객 프로파일링을 실시할 것이다. 의사결정나무 분석을 위해서는 먼저 분리기준과 정지기준을 설정해 주어야 한다. 목표변수가 범주형인 5개의 고객 세분화 그룹이므로 분석에서는 엔트로피 지수(entropy index)를 이용하였으며, 끝마디에 포함될 관측개체의 최소 개수를 10개로 하고, 임의의 분리기준에 의해 부모마디가 자식마디로 분리되기 위해 요구되는 관측개체의 수를 40개로 지정하였다. 또한 자식마디가 형성될 때 고려될 최대의 분리 개수는 이지분리를 사용하였으며, 나무의 최대 깊이는 6으로 설정하였다. 그리고 4,397개의 전체 변수에 대한 모형을 구축하기 위하여 분석용(Training)자료와 평가용(Validation)자료를 각각 50%로 할당하여 분석을 수행하였다. 그리고 의사결정나무모형과의 비교를 위해 로지스틱 회귀분석도 함께 시행하였다.

먼저 세분화된 고객그룹들에 대한 특성을 파악하기 위하여 유의수준 5%에서 유의하지 않은 결과를 나타낸 변수들을 제외하였다. 이렇게 하여 최종적으로 분석에 사용될 변수의 선정은 성별(A2), 직업(A5), 신청방법(A6), SSO 총 가입 수(A11)의 4개 변수로 결정되었다. 이와 같이 생성된 의사결정나무 모형에 따르는 그룹별 세부특성은 <표 12>과 같다. 각각의 고객 그룹에 있어 특성을 가장 잘 나타낼 수 있는 변수로는 뿌리마디(root node)에 해당되는 변수로 신청방법이 각각의 그룹에 가장 큰 영향을 주는 요소로 평가되었다.

<표 12> 의사결정나무모형을 이용한 고객 그룹별 특성

segment group	그룹의 특성	특성이 그룹과 일치할 확률
최우수그룹	신청방법(검색) & 직업(3, 4 ,5) & 사이트 가입수(3개 이상)	52.6%
우수그룹	신청방법(검색) & 직업(1, 2, 7) & 사이트 가입수(3개 이하) & 성별(남)	21.9%
일반그룹	신청방법(웹) & 사이트 가입수(3개 이하)	60.0%
하위그룹	신청방법(웹) & 사이트 가입수(3개 이상) & 직업(전문직 및 기타)	34.4%
최하위그룹	신청방법(웹) & 사이트 가입수(3개 이하)	12.0%

의사결정나무 분석을 통하여 얻어진 세분화된 5개의 고객그룹에 대한 특성을 살펴보면, 검색방법을 통하여 원문을 신청하고, 직업이 대기업, 중소기업, 의료/제약 분야에 속하며, K기관에서 주로 이용하는 사이트의 수는 3개 이상인 고객들이 최우수그룹으로 분류될 확률은 52.6%로 나타났다.

우수그룹에 대한 분석 결과는 검색방법을 통하여 원문을 신청하고, 직업이 공공부문, 교육/연구기관, 전문직 및 기타에 속하며, K기관에서 주로 이용하는 사이트는 3개 이하인 고객들은 우수그룹으로 분류될 확률은 21.9%로 나타났다. 그리고 일반그룹과 하위그룹, 최하위그룹에 대한 분석 결과는 서로 비슷하였다. 즉, 웹 방법을 통하여 원문을 신청하고, K기관에서 주로 이용하는 사이트는 3개 이하의 고객들인 것으로 분석되었다.

3. 결론

직 · 간접적 데이터마이닝 기법인 연관분석과 의사결정나무를 이용하여, 다섯 개의 세분화된 고객그룹에 대한 분석을 했다. 우선 본 연구에서 실험한 두 종류의 기법에 사용된 변수들을 살펴보면, 연관분석에서는 이용가능한 모든 변수(결측치가 너무 많은 일부 변수 제외)를 활용하여 분석을 시행하였고, 의사결정나무 분석은 평균과 교차분석을 통하여 유의수준 5% 이내에서 유의하다고 판단되는 변수들만을 입력변수로 활용하였다.

연관분석과 의사결정나무를 이용하여 분석된 결과에 대한 요약은 <표 13>에 기술하였다. 내용을 살펴보면, 우선 연관분석에서는 다섯 개의 세분화 그룹에 대하여 최하위그룹에 대한 연관규칙을 발견하지 못하였으며, 의사결정나무 분석에서는 하위그룹들(일반그룹, 하위그룹, 최하위그룹)이 유사한 패턴을 보이고 있어 구분하기 어려웠다.

최우수그룹에서는 K기관의 주이용 사이트는 3개 이하라는 점은 동일하지만, 직업에서 상이하고

연관분석에서는 좀 더 많은 특성을 알게 되었다. 우수그룹에서도 K기관의 주이용 사이트는 3개 이하라는 점은 동일하지만, 직업에서 상이하고 연관분석에서는 좀 더 많은 특성을 알게 되었다. 의사결정나무에서는 일반그룹이나 하위그룹, 최하위그룹에 대한 특성을 구분하지 못하였으나, 연관분석에서는 일반그룹과 하위그룹의 특성을 알 수 있었다. 또한 연관분석에서는 RFM의 대략적인 점수를 구할 수 있었다. 그리고 데이터마이닝 분석을 통하여 밝혀진 중요한 사실은 상위계층(최우수그룹, 우수그룹)과 하위계층(일반그룹, 하위그룹, 최하위그룹)간 뚜렷한 경계를 긋는 중요한 변수를 찾아내었다는 것이다. 이러한 변수는 신청방법으로, 상위계층은 검색을 활용한 원문복사 신청을 많이 하고 있으며, 하위계층은 웹을 통한 원문복사 신청을 많이 하고 있다는 것이다.

따라서 연관분석이 의사결정나무에 비해 많은 세분그룹별 규칙과 특성을 제공하여 보다 직관적으로 이해하기 쉽게 도와주는 도구임을 알 수 있었다. 따라서 폭넓은 이해를 바탕으로 정확하고 효과적인 전략을 구사할 수 있을 것이다.

<표 13> 직·간접 데이터마이닝 기법을 통한 고객 특성 분석 결과

segment group	연관분석		의사결정나무
	특 징	RFM 패턴	
최우수 그룹	교육/연구기관에서 근무하는 개인 이용자로, K기관의 주 이용 사이트는 3개 이하이며, 정보검색에 능하며, 직장 이메일을 통하여 개인화 서비스를 이용하는 경향이 있다.	R=5 F=5 M=4~5	대기업, 중소기업, 의료/제약 분야에 종사하고 있으며, K기관의 주 이용 사이트는 3개 이상인 경향이 있다
우수 그룹	남자인 개인 회원으로, K기관의 주 이용 사이트는 3개 이하이며, 직장 이메일을 통하여 개인화 서비스를 이용하는 경향이 있다.	R=5 F=3~4 M=3	공공부문, 교육/연구기관, 전문직 및 기타에 종사하고 있으며, 정보검색에 능하며, K기관의 주 이용 사이트는 3개 이하인 경향이 있다
일반 그룹	남자인 개인 회원으로, K기관의 주 이용 사이트는 3개 이하이며, 정보검색에 어려움을 겪는 경향이 있다.	R=3 F=2 M=3	정보검색에 어려움을 겪고 있으며, K기관의 주 이용 사이트는 3개 이하인 경향이 있다.
하위 그룹	교육/연구기관에 종하사는 개인 회원으로, 직장 이메일을 통하여 개인화 서비스를 이용하는 경향이 있다.	R=1~2 F=1~2 M=1~2	
최하위 그룹	발견 못함	발견못함	

참고문헌

[1] 강현철 등 (2001). *SAS Enterprise Miner 4.0을 이용한 데이터마이닝 : 기능과 사용법*, 자유아카데미.

[2] 김승아 (1998). *DB 마케팅 분석법*, 경영과 컴퓨터.

[3] 안광호, 김상용, 김주영 (2001). *인터넷마케팅원론*, 법문사.

[4] 이강태 (2002). *eCRM 환경에서 LTV극대화를 위한 고객세분화 기법에 관한 연구*, 석사학위논문, 전주대학교.

[5] Agrawal, R., R., Imielinski, T., and Swami, A. (1993). Database Mining : A Performance Perspective, in *IEEE trans. on knowledge and data engineering,* Vol. 5, No. 6.

[6] B. Stone (1984). *Successful Direct Marketing Methods,* 3rd ed., Crain Books.

[7] Bounsaythip C., E. Rinta-Runsala (2001). *Overiew of datamining for customer behavior modeling,* VTT information Technology.

[8] R. S. Hodgson (1980). *Direct Mail and Mail Order Handbook,* 3rd ed., Dartnell.

ABSTRACT

Huge information has been made due to the current computing environment and could not be acceptable. People want the information which they can understand and accept easily. They may want not only simple information but also knowledge. That is why data mining becomes a center of information. We use RFM analysis in order to create customer score. Customers are classified into five groups(most excellent/excellent/common/lower/lowest) for a various marketing activities.

We can found the significant patterns in each group, and classify customers from loyal customers to leaving customers in the near future by the indirect data mining(e.g. association analysis) and the direct data mining(e.g. decision tree, logistic regression analysis, etc.), which are named in this study.

Our research focuses on the advanced models by applying the association rules in data mining. Our results indicate that the indirect data mining and the direct data mining seem to have same outputs, but the former shows more clear pattern then the latter one.