SOM을 이용한 고객의 이탈 가능성 분석 및 이탈 방지 방법론

채경희[a], 김재경[b], 송희석[c]

[a,b] 서울특별시 동대문구 회기동 1번지 경희대학교 경영대학 경영학과, 130-701
Tel: 02-961-9355, Fax: 02-967-0788, E-mail: jaek@khu.ac.kr

[c] 대전광역시 대덕구 오정동 133번지 한남대학교 경상대학 경영정보학과, 306-791
Tel: 042-629-8344, E-mail: hssong@hannam.ac.kr

요 약

최근 빠르게 성숙되고 있는 시장과 경쟁적 환경으로 인해 고객 유지에 대한 중요성이 증대되고 있다. 이는 기존 고객을 유지하는 것이 비용 면에서 저렴할 뿐 아니라, 고객 충성도나 구전효과와 같은 기타 부수적인 이득을 획득할 수 있다는 측면에서 유리하기 때문이다. 본 논문은 고객의 이탈 가능성을 미리 예측하고 이를 사전에 방지할 수 있는 고객 유지 절차를 제시하고 있다.

이탈고객의 탐지 및 방지를 위해서는 기존의 인구통계학적 자료 외에도 웹로그, 구매 Database 등의 대용량의 고객 행위 데이터에 대한 분석이 요구되기 때문에 데이터 마이닝 기법의 활용이 필수적이다. 그러나 대부분의 데이터 마이닝 연구는 예측 및 분류의 정확성이 높은 모델을 개발하는데 초점이 맞추어져 있으며, 고객의 행위를 이해하고 바람직한 방향으로 유도하고자 하는 연구는 지극히 부족한 상황이다. 그러므로 본 논문은 다양한 데이터 마이닝 기법을 통합하여 잠재 이탈고객을 탐지하고, 기존 연구에서 간과하고 있던 비용적 측면을 고려한 이탈 방지 절차를 제시하고자 한다.

키워드 : 군집수, 데이터 마이닝, 의사결정나무분석, 이탈률, 행위 유도, SOM

1. 서론

1.1 문제제기

최근 치열한 경쟁 환경으로 인해 고객 만족에 대한 중요성이 증대되고 있으며, 신규 고객 유치보다는 기존 고객을 어떻게 유지할 것인가에 대한 관심이 증대되고 있다(Ng and Liu 2000).

지금까지 고객 유지를 위한 많은 연구들이 진행되어 왔으며, 고객 유지 방법에 주로 활용되고 있는 기술 중 하나가 데이터마이닝이다. 데이터마이닝은 기업이 방대한 자료 분석을 통해 고객의 패턴과 성향을 알 수 있는 매우 유용한 방법이며, 많은 연구를 통해 데이터마이닝의 활용과 효과가 입증되고 있다(Yeo et al. 2001).

본 논문에서는 이러한 데이터마이닝 기술을 통해 이탈가능성이 높은 고객을 사전에 발견하여, 이탈 방지 절차를 수립하는 것에 그 목적이 있다. 이 때, 기존의 많은 연구들에서 간과하고 있던 비용적 측면을 고려함으로써 보다 효율적인 방법으로 캠페인 전략을 수립할 수 있도록 한다.

전략 수립에 앞서 먼저 2장에서 분석을 위한 이론적 배경과 사용될 알고리즘들에 대해 간단히 살펴보기로 한다. 다음으로 3장에서 세부 실험 내용들과 절차에 대해 설명하고, 실제 적용한 결과에 대해 분석하고자 한다. 3장의 세부 내용을 살펴보면 먼저 전체적인 연구 절차에 대해 설명한다. 그리고 분석에 사용될 데이터를 수집하고 정제하는 과정을 살펴본다. 다음으로 분석에 적절한 군집수를 결정하고, 각 군집에 대한 이름을 명명하고 군집에 대한 특정을 정의, 설명한다. 군집이 정의되면, 각 군집별 이탈 가능성을 분석하고, 행위 유도 시 발생 가능한 마케팅비용을 산정한다. 이 두 가지를 통해 수익률을 산정함으로써 효율적인 행위 유도 방향을 결정하고, 미리 정의해 놓은 군집의 특성에 따라 행위 유도 전략을 수립한다. 그리고 마지막으로 4장에서 분석 내용을 요약, 정리하고, 추후 연구 방향을 제시하고자 한다.

2. 연구 배경

2.1 Self-Organizing Map(SOM)

SOM은 본 논문의 가장 핵심이 되는 분석 방법으로써, Kohonen (Kohonen 1990; Kohonen 1995; Kohonen et al. 1996)에 의해 제시, 개발되어 Kohonen Maps이라고도 알려져 있다. SOM은 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런(neuron)으로 정렬하여 지도의 형태로 형상화한다. 이러한 형상화는 입력 변수의 위치 관계 그대로 보존한다는 특징이 있다. 다시 말해 실제 공간의 입력 변수가 가까이 놓여있으면, 지도(map) 상에서도 가까운 위치에 놓인다. 이러한 SOM의 특징으로 인해, 입력 변수의 정보와 그들의 관계가 지도 상에 그대로 나타나게 되는 것이다.

SOM은 전문검색 마이닝(full-text mining)과 금융 데이터 분석 등을 통해 가치 있는 마이닝 도구임이 증명되었다. 또한 여러 기술 분야에서의 패턴 발견, 이미지 분석, 프로세스 모니터링, 오류 진단 방법

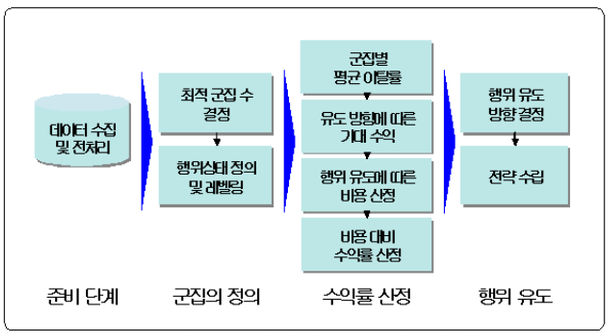
등으로 활용한 결과가 성공적임에 따라 그 적용 분야도 다양하다(Vesanto 1999).

2.1 의사결정나무 분석

의사결정나무분석은 SOM에 의해 나타난 군집들을 레벨링(labeling)할 때 사용되며, 예측과 분류를 위한 보편적이고 강력한 데이터마이닝 도구이다(Berson, Smith and Thearing 2000). 인공신경망 분석과 함께 다양한 분야에서 많이 활용되고 있는데(Chung, Pottenger and Schatz 1998), 신경 분석과는 달리 나무구조로 규칙을 표현하기 때문에 이해하기가 쉽게 나타난다.

3. SOM을 이용한 이탈 방지 연구 모형

3.1 분석 절차



<그림 1> 전체적인 분석 과정

전체적인 절차는 크게 4단계로 이루어지며, 첫 번째 단계에서는 데이터베이스에 저장되어있는 웹로그(web log) 데이터를 수집하여 분석에 적절한 월간 단위로 정리한다.

두 번째 단계에서는 SOM을 이용하여 고객을 군집 별로 나누고 의사결정나무 분석을 이용하여 정의하게 된다. 세 번째 단계에서는 각 군집의 평균 이탈률을 통해 비이탈 고객의 이탈 가능성을 예상해 본다. 마지막 단계에서는 세부적인 전략을 수립하게 된다. 자세한 내용은 다음 절에서 세부적으로 살펴보도록 한다.

3.2 데이터 수집 및 전처리

분석에 사용될 변수는 영역 전문가에 의해 선택된다. 또한 데이터의 값들은 오랜 시간에 걸쳐 데이터베이스에 저장된 웹로그 파일로부터 수집되며, 방대한 양의 웹로그 파일로부터 데이터를 확보하기 위해서는 많은 시간과 노력이 요구된다. 웹로그 데이터의 전처리 과정에 대한 자세한 내용은 Cooley et al. (1999)을 통해 살펴볼 수 있다.

SOM의 학습을 위해 선택된 변수와 데이터들은 정규화 과정을 거친다. 상세한 정규화 과정은 Chakraborty et al. (2000)에 설명되어 있으며, 변수 선택 및 데이터 전처리 과정은 Hall and Smith (1998); Kohavi and John (1997); Almuallim and Dietterich (1994); Kira and Rendell (1992) 등에 잘 나타나 있다.

3.3 군집의 정의

3.3.1 최적 군집수 결정

SOM을 사용하여 군집을 나누기에 앞서, 먼저 적절한 군집의 수를 정해야 하는데, 이는 오분류표를 이용하여 결정한다. '오류 영역 1'이 95% 신뢰수준을 만족시키는 범위, 즉 오류가 5%를 넘지 않는 범위 내에서 '오류 영역 2'가 가장 적게 나타나는 군집의 수를 최적 군집수로 선택하게 된다.

3.3.2 행위상태 정의 및 레벨링(Labeling)

최적 군집수를 결정하고 SOM을 학습시켰으면, 다음으로 각 군집에 어떤 특성을 가진 고객들이 분포되어 있는지, 어떤 기준에 의해 군집이 나뉘어졌는지 살펴보아야 한다. 이는 의사결정나무분석을 통해 나타난 규칙을 살펴봄으로써 알 수 있다.

3.4 수익률 산정

SOM을 통해 고객을 군집화 하고나면, 고객의 행위를 어느 방향으로 유도할지 결정하기 위해 고객의 행위를 유도할 경우 발생되는 수익과 비용을 고려하여 수익률을 산정하고, 가장 수익률이 높은 방향으로 고객의 행위를 유도하게 된다. 세부과정은 다음의 네 단계로 이루어져 있다.

3.4.1 군집별 평균 이탈률 계산

군집 X의 이탈률 P_X를 수식으로 정리하면 다음과 같다. N_X는 군집 X 에 속한 고객의 수, D_X는 군집 X 내의 이탈 고객의 수이고, 이 때 각각의 군집은 X, Y, Z ... 등으로 나뉘며, 각 군집에 속한 개인은 i로 정의한다. 즉, X 그룹에 속한 고객은 x_i, 각각의 고객은 x_1, x_2, x_3 등으로 표시된다.

$$P_X = \frac{D_X}{N_X} \times 100$$

3.4.2 수익(gain) 산정

각 군집의 평균 이탈률을 토대로 목표 방향과 그 방향으로 유도했을 경우의 수익($G_{X,Y}$)을 계산한다. 수익은 이탈률이 높은 곳에서 낮은 곳으로 고객을 유도하였을 경우 발생되는 이탈률의 감소량이다. 이를 식으로 나타내면 다음과 같다.

$$G_{X,Y} = P_X - P_Y$$

3.4.3 비용 산정

앞서 말했듯이 고객의 행위 유도는 수익과 비용을 모두 고려하여 이탈률은 높고 비용은 적은 방향, 다시 말해 비용 대비 수익률이 가장 높은 방향으로 이루어지게 된다. 군집 X의 고객 x_i를 군집 Y로 유도하고자 할 때, 예상되는 비용을 $C_{x_i,Y}$라 하면, $C_{x_i,Y}$는 다음과 같이 정의된다. 이 때, $D(x_i,y_j)$는 x_i와 y_i와의 거리를 의미한다.

$$C_{x_i,Y} = \frac{\sum_{j=1}^{N_Y} D(x_i, y_j)}{N_Y}$$

$C_{x_i,Y}$가 계산되면, 이 값을 이용하여 군집간 거리($C_{X,Y}$), 즉 군집간 유도 비용을 계산할 수 있는데, 이는 다음과 같이 정의 된다.

$$C_{X,Y} = \frac{\sum_{i=1}^{N_X} C_{x_i,Y}}{N_X}$$

3.4.4 비용 대비 수익률 산정

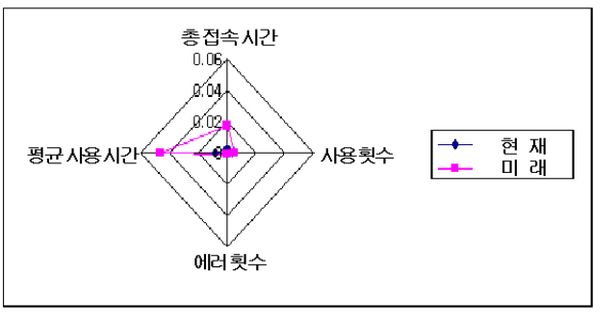
고객의 행위 유도에 따라 발생될 수 있는 수익을

계산하고, 그에 따른 비용을 계산하였으면, 마지막 단계로 비용 대비 수익률을 산정하여 고객의 행위 유도 방향을 결정하게 된다. 군집 X에서 군집 Y로 이동시 발생 가능한 비용 대비 수익률 $E_{X,Y}$의 정의는 다음과 같다.

$$E_{X,Y} = \frac{G_{X,Y}}{C_{X,Y}}$$

3.5 고객의 행위 유도

목표 군집이 결정되면, 앞서 의사결정나무분석을 통해 정의해 놓은 고객 특성을 이용하여 행위 유도 전략을 수립할 수 있다. <그림 2>는 마케팅 캠페인 설계를 위한 사례로써, 현재 고객의 행위 상태(현재)와 유도하고자 하는 행위 상태(미래) 즉, 목표 군집을 비교하여 보여준다. <그림 2>의 그림은 현재 고객의 사용 시간이 크게 부족함을 알 수 있다.



<그림 2> 마케팅 캠페인 설계의 예

4. 사례 연구

4.1 데이터

Korean online game company의 데이터베이스에 저장되어 있는 400명 고객에 대한 114,736건의 Web-log 거래 데이터를 수집하였으며, data set의 이탈자와 비이탈자 비율은 40%와 60%로 구성하였다. 수집된 데이터는 전처리 과정을 통해 한달 동안의 총 사용 시간, 총 사용 횟수, 에러 횟수 및 한달 평균 사용시간을 정리하였다. 기간은 한 달 혹은 한 주가 될 수도 있으나, 고객의 행위 정보를 충분히 파악하고, 변화를 관찰하기 위해 한 달이 적절하다고 판단되었다. 또한 인공신경망 분석과 본 논문에서 제시되고 있는 모델의 성능을 비교하기 위해 데이터를 학습용 데이터 60%와 검증용 데이터 40%로 나누었다.

4.2 최적 군집 수

Korean online game company의 고객을 가장 적절하게 분류하는 최적 군집 수를 결정하기 위하여 6×6, 5×7, 7×5, 6×8, 8×6 등의 군집수를 임으로 정한 후 실험해 보았다. 앞의 과정에서 전처리 된 데이터를 SOM을 이용하여 6×6, 5×7, 7×5, 6×8, 8×6 등으로 군집을 나눈 후 각 경우마다 오분류표를 작성해 보았으며, 결과는 <표 1>과 같다.

<표 1> 각 군집수에 따른 오분류표

실제값 / 예측값	6×6		5×7		7×5		8×6	
	비이탈	이탈	비이탈	이탈	비이탈	이탈	비이탈	이탈
비이탈	-	11.67	-	17.62	-	15.51	-	13.33
이탈	11.90	-	4.56	-	6.90	-	8.57	-

앞의 분석과정 설명에서 언급했듯이, 최적 군집 수는 오류영역 1이 5%를 넘지 않는 범위 내에서 오류영역 2가 가장 적은 것을 선택하기로 하였다. 이와 같은 전제를 토대로 Korean online game company의 경우에 있어서는 오류영역 1이 4.56%인, 5×7이 가장 적절한 군집 분류 방법이다.

4.3 고객 행위 유도 전략 수립

각 유효 군집의 이탈률과 전처리 과정을 통해 생성된 Korean online game company의 고객의 네가지 변수를 사용해 생성된 비용을 기반으로 Visual Basic 6.0 프로그램을 활용하여 현재 군집과 목표 군집을 생성하였다. 또한 의사결정나무 분석을 통해 생성된 규칙을 통해 현재 군집과 목표 군집간의 차이를 살펴보았는데, 그 결과는 <표 2>과 같이 나타났다.

<표 2> 현재 군집과 목표 군집의 비교

현재 군집	목표 군집	현 군집과 목표 군집과의 차이점
0	1	접속횟수의 증가
1	2	
2	9	평균 사용시간의 증가
3	9	
8	9	총 사용시간의 증가
9	17	
10	11	
11	24	접속 횟수 및 총 사용시간의 증가
12	11	총 사용시간의 증가
15	23	평균 사용시간의 증가
16	17	
17	24	
18	17	접속 횟수 및 총 사용시간의 증가
20	24	
21	24	총 사용시간의 증가
22	23	
23	24	총 사용시간 증가 및 사용 오류의 감소
26	24	사용자 오류의 감소
29	23	
32	24	

<표 2>을 통해 각 군집의 목표 군집 및 목표 군집과의 차이를 살펴보았다면 이러한 차이를 기반으로 <표 3>에서 고객 행위의 변화를 유도하기 위한 몇 가지 방법을 제시해 보았다. <표 2>에서 군집 '0'은 수익률 분석 결과 군집 '1'로 유도하는 것이 가장 적절하다는 결론이 나왔다. 이러한 사실을 기반으로 군집 '0'과 군집 '1'의 속성을 비교해본 결과 두 군집간의 차이는 접속횟수이며, '0'에서 '1'로 군집을 유도하기 위해서는 접속횟수를 증가시켜야 한다는 사실을 알 수 있었다. <표 3>는 각 경우에 따라 수행 가능한 행위 유도 방안을 보여주고 있다.

<표 3> 행위 유도 방안

군집간 차이점	행위 유도를 위한 캠페인
접속횟수의 증가	- 사이트의 수시 업데이트 및 그에 따른 소식 발송 - 새로운 게임 정보 제공 및 무료 사용 쿠폰 제공
평균 사용시간의 증가	- 사용시간에 따른 누적 점수 도입 - 누적 점수에 따라 상품 및 무료 쿠폰 발송
총 사용시간의 증가	- 저렴한 월 정액권 판매 및 할인권 증정 - 사용 누적점수에 따른 상품 및 무료 쿠폰 발송
접속횟수 및 총 사용시간의 증가	- 새로운 게임 정보 및 무료 사용 쿠폰 제공
사용 오류의 감소	- 서버 오류인지, 사용자 오류인지 분석 - 사용자가 많은 시간대의 접속에 따른 오류 발생의 경우에는 사용자가 적은 시간 대에 사용할 수 있는 쿠폰 발송

4.4 성과 측정

이번 절에서는 SOM에 의해 구축된 모델과 예측력이 뛰어난 것으로 알려져 있는 인공신경망과 비교하여 그 정확도에 대해 평가해 보도록 하겠다.

평가방법으로는 일반적으로 사용되고 있는 예측률(precision)과 재현률(recall)을 사용하였다.

예측률(precision) :

$$P = \frac{|\text{모델에 의해 예측된 이탈자} \cap \text{실제 이탈자}|}{|\text{모델에 의해 예측된 이탈자}|}$$

재현률(recall) :

$$P = \frac{|\text{모델에 의해 예측된 이탈자} \cap \text{실제 이탈자}|}{|\text{실제 이탈자}|}$$

<표 4> 예측 정확성 측정 비교 결과

분석기법	예측률	재현률	적중률
4×5 SOM	65.5	76	78.7
5×7 SOM	73.1	76	82.7
6×8 SOM	70.3	76	81.3
NN(MLP)	76.2	64	81.3

5. 결론

5.1 요약 및 시사점

신규 고객을 획득하는 것보다 기존 고객을 유지하는데 소모되는 마케팅 비용이 더 저렴하다는 것 외에, 고객 충성도 및 캠페인의 효과성 등 부가적인 효과로 인해 기존 고객의 유지에 대한 중요성이 증대되고 있다. 따라서 본 논문에서는 고객 이탈을 방지하기 위한 새로운 방법론을 제시하였으며, 실제 데이터에 적용해 보았다. 이로써 고객의 이탈이 발생하기 전에 조치를 취할 수 있도록 하였다. 이 때, 기존 연구에서는 간과하고 있던 비용적 측면을 고려함으로써 효율적인 전략을 수립할 수 있도록 하였다.

5.2 연구의 한계 및 향후 연구방안

본 연구에서는 한 달 동안의 실제 데이터를 활용하여 새로운 고객 이탈 방지 절차를 제시함으로써, 효율적인 마케팅 캠페인을 수행하기 위한 근간을 제시하였다. 그러나 마케팅 수행 결과를 분석하고 그 효과를 실험에 반영하지 못한 한계점이 있다.

따라서 앞으로의 연구에서는 본 논문에서 제시된 이탈 방지 절차를 자동하고, 실제 캠페인을 계획하고 실시하여, 그 효과성을 검증해 보고자 한다. 또한 효과성 검증을 위한 기준을 연구하여, 고객 행위 유도에 대해 효과가 높은 캠페인에 대해서는 정규화된 매뉴얼로 가시화 하고, 효과가 없는 캠페인에 대해서는 인터뷰나 설문조사를 통하여 고객의 요구사항을 반영함으로써 보다 실제적인 캠페인 전략을 제시하고자 한다. 본 연구에서는 온라인 게임 사이트의 사례를 통해 적용해 보았으나, 이외에도 통신 산업, 인터넷 서비스 및 컨텐츠 산업에도 적용해 볼 수 있을 것이다.

또한 본 논문에서는 고객의 개인 정보를 침해하지 않는 한도에서 세가지 변수를 추출하는데 그쳤으나, 향후 분석에서는 총 사용시간, 사용 횟수, 에러 횟수 외에 보다 많은 변수를 추출함으로써, 분석의 정확성을 높이고 마케팅 캠페인 수립 시 자세한 기준 및 전략을 제시하고자 한다.

<참고문헌>

Almuallim, H. and Dietterich, T. (1994). Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence 69: 279-306.

Berson, A., Smith, S. and Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill. 277-298.

Chakraborty, G. and Chakraborty, B. (2000). A novel normalization technique for unsupervised learning in ANN. IEEE transactions on neural networks 11(1): 253-257.

Chung, Y., Pottenger, W. and Schatz, B. (1998). Automatic Subject Indexing Using an Associative Neural Network. In Proceedings of the ACM Conference on Digital Libraries (DL-99), 59-68

Cooley, R., Mobasher, B. and Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and information systems 1(1): 5-32.

Hall, M. and Smith, L. (1998). Practical feature subset selection for machine learning. Proceedings of the 21st Australian Computer Science Conference, 181-191.

Kira, K. and Rendell, L. A. (1992). The feature selection problem : Traditional methods and a new algorithm. Proceedings AAAI-92, 129-134.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.

Kohonen, T. (1990). The Self-Organizing Map. Proceedings of the IEEE 78(9): 1464-1480.

Kohonen, T. (1995). Self-Organizing and Associative Memory. Berlin: Springer-Verlag.

Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. (1996). Engineering applications of the Self-Organizing Map. Proceedings of the IEEE 84(10): 1358-1384.

Ng, K. and Liu, H. (2000). Customer retention via data mining. Artificial Intelligence Review 14(6): 569-590.

Vesanto, J. (1999). SOM-based data visualization methods. Intelligent data analysis 3: 111-126.

Yeo, A., Smith, K., Willis, R. and Brooks, M. (2001). Modelling the effect of premium changes on motor insurance customer retention rates using neural networks. Lecture notes in computer science 2074: 390-399. Springer-Verlag.