순차자료를 이용한 고객 이탈 예측 데이터마이닝 방법론에 관한 연구

Integrated Data Mining Methodology for Predicting Customer Defection using Sequential Data

최준연 김성희
한국과학기술원 테크노경영대학원
{zoon, seekim}@kgsm.kaist.ac.kr

Abstract

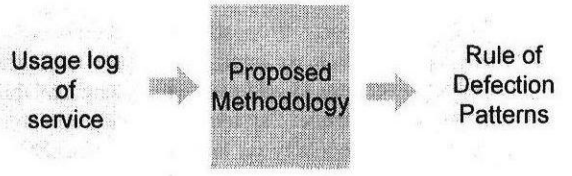
Finding the patterns of defection for predicting customers' behavior is an important issue since holding loyal customers is directly related to the company's profit. This paper proposes an integrated 3-phased methodology for discovering rules of defection process using customers' time-varying usage log. Using moving average method, sequential events from daily usage data were abstracted, then we performed frequent sequence search just before the occurrence of defection. The methodology can be applied to interactive internet service which requires user authentication such as in chatting, community, on-line game etc.

1. 서론

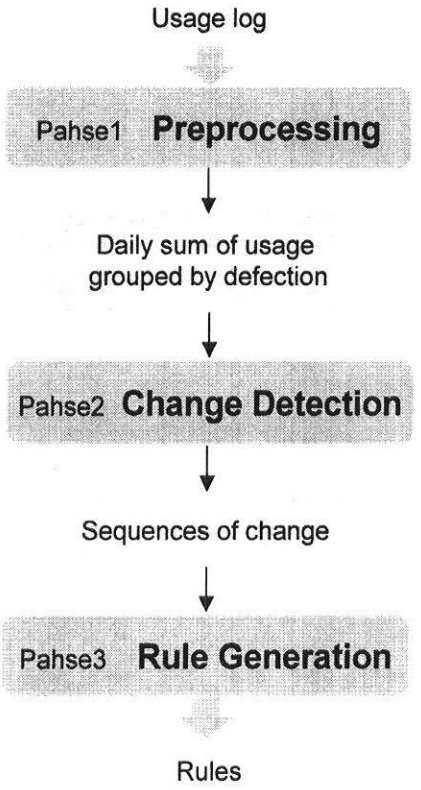
현재의 고객을 유지하는 비용보다 신규 고객을 유치하는 비용이 훨씬 많이 소요되므로, 기업들은 현재의 고객들을 충성고객으로 만들어 이탈하지 않도록 해야만 수익을 극대화할 수 있다[2]. 과거 인터넷이 등장하기 전까지는 고객에 관한 정보를 수집하기가 무척 어려웠으나, 인터넷이 보편화되고, 인터넷을 이용한 새로운 서비스가 등장하면서 고객정보의 획득이 용이해졌고, 기업이 보유한 고객정보가 방대한 양에 이르기까지 했다. 상호작용성 디지털 상품이라고 할 수 있는 온라인게임, 이메일, 증권거래 서비스 등에서는 사용자의 모든 서비스 이용 기록을 저장할 수 있고, 이를 통해 고객의 사용 패턴을 분석하면 향후 이탈 가능성을 예측할 수 있다. 본 논문에서는 이탈고객과 체류고객의 서비스 이용 시간을 순차자료로 변환하여, 현재 고객의 이탈 가능성 및 이탈 고객들의 이탈 규칙을 알아내는 방법을 제안하고자 한다.

2. 문제 정의 및 방법론

인터넷 기반의 서비스들 중 사용자 인증을 필요로 하는 경우, 각 사용자들의 로그인시각, 로그아웃시각, 접속위치 등에 관한 정보를 서버에 기록할 수 있다. 본 연구에서는 이러한 로그정보를 입력 데이터로 사용하여, 최종 결과물로 이탈 고객의 이탈 패턴에 관한 규칙을 추출해내는 것을 목표로 한다.



[그림1] Overall Procedure



[그림2] 3-phased Methodology

2.1 Phase1

Phase1은 전처리 과정으로 매 접속마다 생성된 로그화일을 사용자별/일별 단위로 접속시간을 정리한다. 이 작업 후 이탈자와 체류자를 분류하게 되는데, 이탈자의 정의는 각 서비스의 특성에 따라 달라진다. 유무선 통신서비스와 달리 인터넷기반의 서비스에서는 이탈비용,경쟁사로의 전환비용, 재진입 비용이 매우 낮기 때문에, 이탈자와 체류자를 구분하기가 상대적으로 어렵다. x일 이상 사용 후, y일 이상 서비스에 접속하지 않은 사용자를 이탈자로 정의하며, 사용일수가 x일 이하인 사용자는 이탈자도 체류자도 아닌 진입자로 분류하여 분석에서 제외된다.

If access_days < x
Then 진입자
Else if non_access_days after last access >y
Then 이탈자
Else if
Then 체류자

2.2 Phase2

Phase2는 Phase1에서 걸러진 이탈자들의 일별 사용량을 입력 데이터로 받아, 의미 있는 변화시점으로 이루어진 순차 자료 (Sequences of change)를 생성해 내는 단계이다. 본 연구에서는 상대적 사용시간의 증가 또는 감소와 같은 변화 시점을 추출해 내기 위해 이동평균선의 교차점을 사용하였으며[3], 이에 필요한 인수로서 이동평균선의 개수와 평균선 길이를 추가적인 입력자료로 사용한다.

사용할 이동평균선의 개수와 발생하는 이벤트의 종류의 개수 사이에는 다음과 같은 관계가 성립한다.

Number of Moving Average =n
Number of Cross = ${}_nP_2$

이동평균선의 평균선 길이는 최적값을 산출하기 어려우며, 사용량의 변동성에 반비례하도록 설정하고, Phase1에서 진입자와 체류자를 구분하는 기준으로 사용했던 x값을 기준으로 이용하여, a*x day (0<a<=1)를 계산하는 것이 실험적으로 좋은 결과를 보인다.

't' 시점에서의 p-day 이동평균값은 다음과 같다.

$$M_t^p = \frac{1}{p}\sum_{i=t-p+1}^{t} x_i$$

p-day 이동평균선과 q-day 이동평균선의 교차점은 다음과 같이 구해진다.

$$(M_{t-1}^p - M_{t-1}^q)(M_t^p - M_1^q) \le 0$$

이와 같이 두개의 이동평균선을 사용하는 경우 발

생할 수 있는 이벤트의 종류는 p-day 이동평균선이 q-day 이동평균선을 아래에서 위로 관통하는 경우와, 위에서 아래로 관통하는 경우 두가지이다.

Event Type "a"

$$(M_{t-1}^{p} - M_{t-1}^{q}) < 0 \_and\_ (M_{t}^{p} - M_{1}^{q}) > 0$$

Event Type "b"

$$(M_{t-1}^{p} - M_{t-1}^{q}) > 0 \_and\_ (M_{t}^{p} - M_{1}^{q}) < 0$$

3개 이상의 이동평균선을 사용하는 경우에도 두 개씩 짝을 이룬 상태에서 위의 식을 적용하면 각각의 교차점을 찾아낼 수 있다.

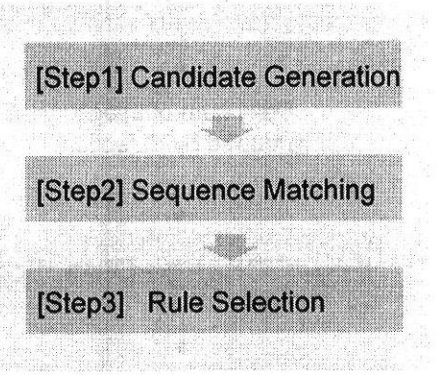
이러한 과정을 거치게 되면, 아래와 같은 시퀀스 테이블이 나온다. 한 사용자당 하나의 레코드를 형성하며, 시퀀스의 길이는 모두 다를 수 있다.

사용자	시퀀스길이	시퀀스
1	L1	$S_1(e_1^1, e_1^2, \ldots, e_1^{L1})$
2	L2	$S_2(e_2^1, e_2^2, \ldots, e_2^{L2})$
3	L3	$S_3(e_3^1, e_3^2, \ldots, e_3^{L3})$
...	...	...
n	Ln	$S_n(e_n^1, e_n^2, \ldots, e_n^{Ln})$

[표1] 시퀀스 테이블

2.3 Phase3

마지막 단계인 Phase3에서는 최종적인 규칙을 추출해 내는 단계이며 그림00과 같이 3개의 스텝으로 구성되어 있다. Step1에서는 후보 시퀀스를 찾아내는 과정으로 이탈시점 직전의 이벤트로부터 과거방향으로 이벤트를 하나씩 추가하며 최소지지도(minimum support) 요건을 충족하는 후보를 찾아낸다. 이 과정에서 후보 시퀀스의 최대 길이를 입력인수로 받아, 이에 도달했을 때 탐색 작업을 멈춘다. 여기서 계산되는 지지도는 Agrawal[1]이 사용했던 전체 인원수 중 해당 시퀀스를 갖고 있는 사용자의 비율로 산출한다.



[그림3] 3 steps of Phase 3

Step2에서는 후보 시퀀스들 중 어떤 시퀀스가 가장 의미있는가에 관한 강도(strength)를 측정하기 위해 지지도와 신뢰도를 계산한다. 기존의 순차마이닝 연구에서는 사람수의 비율만을 갖고 지지도를 계산하였는데, 위와 같은 고객 이탈 문제에서는 다음과 같은 이유로 지지도의 새로운 정의가 필요하게 되었다.

1. 동일 사용자가 같은 시퀀스를 여러 번 갖을 수 있으며, 이 횟수가 지지도에 영향을 미친다.
2. 동일한 시퀀스가 이탈자와 체류자 중 어디에 등장하느냐에 따라 지지도에 영향을 미친다.

기존의 방법인 사람의 비율 개념으로 지지도를 정의하게되면 다음과 같다.

$$Support_u(C_k) = \frac{NU(C_k)}{n_c + n_s}$$

C_k : 후보 시퀀스

$NU(C_k)$: Ck를 1회 이상 갖고 있는 사용자 수

n_c : 이탈자 수

n_s : 체류자 수

사람수의 비율만으로 지지도를 계산할 경우, 한

사용자에게서 복수로 발생하는 시퀀스의 지지도는

$$Support_c(C_k) = \frac{NC(C_k)}{n_c + n_s}$$

파악하기 어렵다는 한계를 갖게 된다 이를 위해 아래와 같은 추가적인 지지도의 정의를 도입한다

$NC(C_k)$ C_k의 총 등장 횟수

$Support_c$는 사용자 1인이 평균적으로 갖고 있는 C_k의 개수를 의미하게 된다

순차마이닝에서의 신뢰도에 대해서는 일관된 정의가 없었는데[4], 본 연구에서는 이탈이라는 명확한 결과변수를 갖고 있으므로,

$$Confidence(C_k) = \frac{count\_defection\_C_k}{NC(C_k)}$$

count_defection_C_k Ck후 바로 이탈한 사용자 수

Step3에서는 앞에서 구한 척도를 이용해 각각의 시퀀스의 규칙 강도의 서열을 매긴다 규칙 강도가 높을수록 해당 시퀀스를 보이는 사용자가 나타날 경우 이탈하지 않도록 예방조치가 필요함을 의미한다

3. 결론

본 연구에서는 인터넷 서비스 사용자의 사용시간을 이용하여 고객 이탈 패턴을 찾아내는 방법론을 제시하였다. 본 연구 결과는 인증 과정을 거치고 접속 빈도가 높은 서비스에 주로 적용 가능하며, 서비스의 성격상 필수재보다는 선택재에 가까울수록 예측력이 우수 할 것으로 보인다. 향후 연구로서 사용시간과 같은 단일 시계열 변수 이외에 정적인 인구통계학적 정보를 동시에 사용하거나 복수 시계열 변수를 같이 입력 받아 고객이탈을 예측하는 방법론에 관한 연구가 필요할 것으로 보인다

참고문헌

[1] Agrawal, R & Srikant R. (1995) Mining Sequential Patterns 11th International Conference on Data Engineering

[2] Datta, P., Masand, B, Mani, D. & Li, B. (2001). Automated Cellular Modeling and Prediction on a Large Scale *Artificial Intelligence Review*, (Vol 14) (pp 485-502)

[3] Lebaron, B (1992) Do Moving Average Trading Rule Results Imply Nonlinearities in Foreign Exchange? Technical Report, University of Wisconsin – Madison, Medison, Wisconsin

[4] Lu, H, Feng, L., Han, J (2000) Beyond Intratransaction Association Analysis Mining Multidimensional Intertransaction Association Rules *ACM Transaction on Information Systems*, 18(4), 423-454

[5] Mannila, H, Toivonen, H, Verkamo, I (1997) Discovery of Frequent Episodes in Event Sequences Data Mining and Knowledge Discovery, 1, 259-289

[6] Mozer, M, Wolniewicz, R, Grimes, D & Kaushansky, H (2000) Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry *IEEE Transactions on Neural Networks*, 11(3), 690-696