Journal of the Korean Institute of Industrial Engineers
Vol. 41, No. 1, pp. 10-16, February 2015.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2015.41.1.010

<Original Research Paper>

# Continuous Conditional Random Field에 의한 인터넷 쇼핑몰 신규 고객등급 예측

안길승 · 허 선†

한양대학교 산업경영공학과

# Prediction of New Customer's Degree of Loyalty of Internet Shopping Mall Using Continuous Conditional Random Field

Gil Seung Ahn · Sun Hur

Department of Industrial and Management Engineering, Hanyang University

In this study, we suggest a method to predict probability distribution of a new customer's degree of loyalty using C-CRF that reflects the RFM score and similarity to the neighbors of the customer. An RFM score prediction model is introduced to construct the first feature function of C-CRF. Integrating demographical similarity, purchasing characteristic similarity and purchase history similarity, we make a unified similarity variable to configure the second feature function of C-CRF. Then parameters of each feature function are estimated and we train our C-CRF model by training data set and suggest a probabilistic distribution to estimate a new customer's degree of loyalty. An example is provided to illustrate our model.

***Keywords:*** Customer Loyalty, Continuous Conditional Random Field(C-CRF), RFM Score, Similarity

## 1. 서 론

인터넷을 이용한 전자상거래 이용규모는 급속도로 증가하고 있는데, 통계청 자료에 의하면 2013년 연간 전자상거래의 총 거래액은 1,200조 원 규모로, 2004년부터 2013년까지 연평균 18% 성장해 왔다(Shin, 2014). 인터넷 전자상거래에서는 지리적인 상권이 존재하지 않아 사이버 상에서의 고객 확보가 매우 중요하다. 신규 고객을 확보하는데 소요되는 비용은 기존 고객을 유지하는데 소요되는 비용의 5배에 이르며, 고객 유지율을 5% 향상하면 기업의 이윤을 85%까지 증가시킬 수 있는 것으로 조사되고 있어, 인터넷 쇼핑몰에서의 고객 관계관리(Customer Relationship Management, CRM)가 매우 중요하다(Kalacota and Robinson, 1999).

CRM은 기업에 가치를 제공할 수 있는 고객과의 관계를 구축하는 것이다. 특히 고객 분류는 CRM의 한 부분으로, 기업에 이익을 주는 고객의 속성과 구매패턴을 분석함으로써 목표 고객을 설정하는 것을 의미한다(Kumer and Reinartz, 2012).

기존의 CRM에 관련된 연구는 주로 CRM의 평가 방법론 개발과 CRM 활동을 통해 얻을 수 있는 실질적인 관계편익을 분석하는데 초점을 두고 있다. Kim과 Jeong(2012)은 국제적으로 널리 알려진 CRM scorecard 프레임워크를 활용하여 기업의 CRM을 진단할 수 있는 방법론을 개발하였다. 또한, Sim *et al.* (2011)은 최근 기업들이 진행하고 있는 CRM 활동이 기업의 성과에 긍정적인 영향을 미치는가에 대한 분석과 CRM의 대상이 되는 고객들이 기업수익에 공헌하는 충성도에 따라 각각 어떻게 인지하고 있는가에 대한 연구를 진행하였다.

고객 분류는 여러 데이터마이닝 기법을 통하여 이루어져 왔는데, RFM 방법을 사용한 교차분석(Cross tabulation)이 가장

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2012R1A2A2A01005219).
† 연락저자 : 허 선 교수, 426-791 경기도 안산시 상록구 한양대학로 55, Tel : 031-400-5265, Fax : 031-400-5265,
E-mail : hursun@hanyang.ac.kr
2014년 7월 20일 접수; 2014년 10월 8일 수정본 접수; 2014년 10월 21일 게재 확정.

많이 사용되는 방법론이다(Lim *et al*., 2003). RFM은 Recency, Frequency, Monetary의 약자로, Recency는 최근 구매일, Frequency는 구매 빈도, Monetary는 구매 금액을 나타낸다(Lee and Choi, 2002). RFM은 구매 가능성이 높은 고객을 선정하기 위한 데이터 분석방법으로, 이를 통해 풍부한 고객데이터로부터 의미있는 정보를 추출한다.

e-CRM은 상품과 서비스, 컨텐츠 등과 관련된 그리고 수시로 저장되는 고객 관련 데이터를 통합, 가공, 재정리, 분류, 그리고 분석함으로써 고객만족도를 향상시켜 궁극적으로는 수익구조를 개선하는 경영관리활동을 말한다(Song, 2013). 기존의 e-CRM, 특히 인터넷 쇼핑몰 고객에 관련된 연구는 고객 세분화를 통한 CRM에 초점을 두고 있다. Park *et al*.(2002)은 인터넷 쇼핑몰의 고객을 온라인 구매빈도와 쇼핑몰 애호도에 의해 네 개의 집단으로 분류한 뒤 이들의 특성을 비교하고, 이들 특성에 적합한 고객관리 방안을 제시하였다. 다른 연구에서는 온라인 서점을 이용하는 고객들의 효과적인 CRM을 위하여 소비자들을 비슷한 특성을 가진 집단으로 분류하고 각 집단의 특성에 따라 고객 유형을 정의하고 세분하였다(Jeon *et al*., 2007).

또한, Joe와 Kim(2006)은 e-CRM에서의 최적화 모형을 이용한 고객 세분화 방법을, Lee와 Lee는(2004)은 온라인 소매상점에서 클릭 페이지 수, 세션시간, 게시판 가입 여부, 마케팅 이벤트 참여여부를 속성으로 하는 의사결정나무기법을 적용하여 효과적으로 고객을 분류하는 방법을 각각 소개하였다.

그러나 인터넷 쇼핑몰에서의 고객들은 매우 다양한 형태의 특성을 가지기 때문에 단지 몇 가지 그룹으로 확정적(deterministic) 분류를 하는 것은 사실상 불가능하며 설사 분류한다고 해도 고객 행태의 불확실성으로 인해 그 예측의 유효성이 낮을 수밖에 없다.

기존 인터넷 쇼핑몰의 고객에 관한 연구에서 초점을 두고 있는 다른 분야는 충성고객과 이탈고객 예측이다. Kim *et al*.(2009)은 충성고객 네트워크와 이탈고객 네트워크를 구축하여 사회 네트워크 분석을 통해 충성고객 그룹과 이탈고객 그룹의 연결관계와 연결구조의 차이를 파악하여, 충성고객과 이탈고객을 예측하는 방법이 제시하였다. 한편, Jeon과 Kim(2012)은 고객의 나이, 거주지역, 성별, 할부구매여부 등의 고객의 특성을 속성으로 하는 의사결정나무를 이용하여 충성고객과 비충성고객을 구분하였다.

최근 인터넷 쇼핑몰의 경쟁이 심화되면서 고객은 여러 인터넷 쇼핑몰로부터 구매 유혹을 받게 되고 이로써 고객의 구매행동은 동적으로 변화하고 있다(Jeong and Yeon, 2013). 이는 고객의 기업에 대한 로열티 지수, 즉 고객의 등급 점수를 정적인 결과로 제시한다는 것이 불가능하다는 것을 의미한다. 따라서 고객의 등급에 대한 예측을 특정한 값으로 나타내는 것보다는 확률적인 값으로 표현하는 것이 적절하다.

본 연구에서는 고객 등급의 확률분포를 유도하기 위해 고객의 등급 점수를 확률적으로 표현할 수 있는 Continuous Conditional Random Field(C-CRF) 함수를 도입한다. 우선 고객의 인구 통계학적 변수와 구매 특성 변수를 이용한 RFM 점수 예측 모형을 만들고 이로써 C-CRF의 첫 번째 특징함수를 구성한다. 그리고 인구 통계 유사도, 구매 특성 유사도, 구매 이력 유사도 등을 통합하여 통합 유사도 변수를 생성한 후, 이를 이용하여 C-CRF의 두 번째 특징함수를 구성한다. 훈련데이터를 가지고 각 특징함수의 모수를 추정하여 C-CRF 모형을 완성하면 이를 이용하여 신규 고객의 등급 점수의 확률 분포를 제시한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 신규 고객의 등급예측을 위해 도입한 C-CRF를 설명한다. 제 3장에서는 신규 고객의 등급 점수를 예측하기 위한 프로세스를 단계적으로 설명한다. 제 4장에서는 가상의 고객 구매데이터를 가지고 이 연구의 방법을 예시하고 제 5장에서 결론을 정리한다.

## 2. C-CRF

Conditional Random Fields(CRF)는 은닉 마코프모형(Hidden Markov Model, HMM)에서 독립성 가정을 완화하고 순차적 데이터를 구분하고 분류하기 위한 확률모형이다. 조건부 확률인 $P(y|x)$를 직접 모델링하여, 종속변수 간의 복잡한 의존 구조를 이용하기 위한 프레임워크를 제공한다. CRF는 순차 데이터의 분류를 위해 고안되었으므로 여기서 사용되는 종속변수 y는 순차 데이터이다. 이 종속변수를 연속 데이터로 발전시킨 것이 C-CRF이다(Qin *et al*., 2009).

투입변수 $\boldsymbol{x} = (x_1, \cdots, x_T)^T$와 종속변수 $\boldsymbol{y} = (y_1, \cdots, y_T)^T$에 의해 정의되는 $K_1$개의 특징 함수의 집합 $\{f_k(y_i, \boldsymbol{x})\}_{k=1}^{K_1}$과, 투입변수 $\boldsymbol{x}$와 두 개의 관측치에 의해 정의되는 $K_2$개의 특징 함수의 집합 $\{g_k(y_i, y_j, \boldsymbol{x})\}_{k=1}^{K_2}$을 고려하자. C-CRF은 다음과 같은 밀도함수를 가지는 조건부 확률분포이다(Kosta *et al*., 2013).

$$\mathrm{p(y \mid x)} = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} exp\{\sum_{i=1}^{T}\sum_{k=1}^{K_1}\alpha_k f_k(y_i, \boldsymbol{x}) + \sum_{i=1}^{T}\sum_{j=1}^{T}\sum_{k=1}^{K_2}\beta_k g_k(y_i, y_j, \boldsymbol{x})\} \quad (1)$$

여기서 $\boldsymbol{f}(y_i, \boldsymbol{x})$는 투입변수인 $\boldsymbol{x}$와 종속변수 $y_i$와의 의존성을 나타내는 함수이고, $\boldsymbol{g}(y_i, y_j, \boldsymbol{x})$는 $i$번째 데이터와 $j$번째 데이터의 종속변수간의 상호관계를 나타내는 함수이다. 이 때, $\boldsymbol{\alpha}$는 $K_1$차원의 모수이고 $\boldsymbol{\beta}$는 $K_2$차원의 모수이다. 또한, $Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$는 다음을 만족하는 정규화 함수이다.

$$Z(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \exp\{\sum_{i=1}^{T}\sum_{k=1}^{K_1}\alpha_k f_k(y_i, \boldsymbol{x}) + \sum_{i=1}^{T}\sum_{j}^{T}\sum_{k=1}^{K_2}\beta_k g_k(y_i, y_j, \boldsymbol{x})\} dy \quad (2)$$

본 연구에서 특징함수 **f**(·)는 RFM 점수 예측에 의한 오차를 등급 점수에 반영하는 부분이고 특징함수 **g**(·)는 이웃 고객과의 유사도를 등급 점수에 반영하는 부분이다.

C-CRF의 주요 장점은 임의적이고 비독립적인 투입변수를 다룰 수 있는 유연성이다. 이러한 장점 덕분에, 다양한 속성을 가지고 있는 고객의 특성을 나타내는 변수들을 바탕으로 쉽게 모형화할 수 있다. 또한, C-CRF는 종속변수 간의 의존성이 있는 경우에 사용하기 적합한 모형화 방법이다.

## 3. 신규 고객 등급 점수 예측 프로세스

제 3장에서는 고객의 등급을 예측하는 구체적인 절차를 제시한다. <Figure 1>은 본 논문에서 제안하는 고객등급 예측에 관한 프로세스이다.

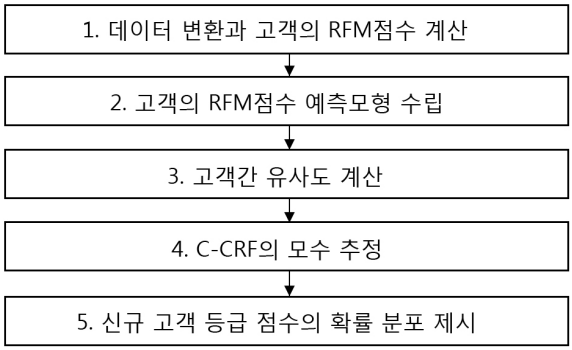


**Figure 1.** Prediction process of the model

### 3.1 단계 1 : 데이터 변환과 고객의 RFM 점수 계산

고객의 등급 분류와 예측에 앞서, 연구 목적에 맞게 데이터의 범위를 일치시키거나 분포를 유사하게 만들어주는 데이터 변환 작업이 필요하다. 본 연구에서 사용한 변수와 각 변수에 적용된 변환 방법은 <Table 1>과 같다.

**Table 1.** Variable list and conversion method

| | 변수명 | 변환방법 | 비고 |
|---|---|---|---|
| 인구 통계 | 나이 | 범주화 | 7개 구간 |
| | 성별 | 이진 변수화 | |
| 구매 특성 | 반품횟수 | 정규화 | 변심에 의한 반품만 고려 |
| | 구매간격 | 정규화 | |
| RFM 변수 | 구매빈도 | 정규화 | |
| | 총 구매액 | 정규화 | |
| | 최근 구매일 | 정규화 | 기준일(예 : 쇼핑몰이 영업을 개시한 날) 이후 경과한 일수 |
| 구매 이력 | 상품군별 | 이진 변수화 | |

본 연구에서 나이는 19세 미만, 20~24세, 25~29세, 30~34세, 35~39세, 40~49세, 50세 이상 등 7개 구간으로 나누어 범주를 생성한다. 또한 각 변수 값들은 다음과 같이 표준화한다.

$$\bar{x}_l \equiv \frac{x_i - \mu_d}{\sigma_d} \tag{3}$$

단, $x_i$는 변수 $x$의 $i$번째 데이터를 나타내며, $\mu_d$와 $\sigma_d$는 각각 변수 $x$의 평균과 표준편차를 나타낸다. 구매 이력은 상품들을 군으로 묶은 후 해당 상품군의 상품을 구매했으면 1, 아니면 0으로 나타내는 이진 변수로 변환한다. 만약 한 고객이 같은 상품을 여러 개 구입하더라도 구매는 한 번 일어난 것으로 간주하여 1로 표현한다(Park *et al*., 2009). 본 연구에서는 통계청 온라인 쇼핑 동향조사에서 사용하는 온라인 쇼핑 상품분류에 의거, 아래 <Table 2>와 같이 상품군을 분류하였다.

**Table 2.** Product classification

| 변수명 | 해당 상품군 |
|---|---|
| P1 | 컴퓨터 및 주변기기 |
| P2 | 소프트웨어 |
| P3 | 가전 · 전자 · 통신기기 |
| P4 | 서적 |
| P5 | 음반 · 비디오 · 악기 |
| P6 | 여행 및 예약서비스 |
| P7 | 아동 · 유아용품 |
| P8 | 음 · 식료품 |
| P9 | 꽃 |
| P10 | 스포츠 · 레저용품 |
| P11 | 생활 · 자동차 용품 |
| P12 | 의류 · 패션 및 관련상품 |
| P13 | 화장품 |
| P14 | 사무 · 문구 |
| P15 | 농수산물 |
| P16 | 각종 서비스 |
| P17 | 기타 |

데이터를 변환한 후 고객의 RFM 점수를 다음과 같이 계산한다.

$$\text{RFM} = W_R \times R + W_F \times F + W_M \times M \tag{4}$$

$R$, $F$, $M$은 각각 최근 구매일, 구매 빈도, 구매 금액을 나타내고, $W_R$, $W_F$, $W_M$은 각각에 부여된 가중치이다(Lim *et al*., 2003).

### 3.2 단계 2 : 고객의 RFM 점수 예측모형

고객의 인구 통계변수와 구매 특성변수들을 독립변수로 하고 고객의 RFM 점수를 종속변수로 하는 다중회귀모형을 다음과 같이 도입한다.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (5)$$

여기서 $X_1$은 나이, $X_2$는 성별, $X_3$은 반품횟수를, 그리고 $X_4$는 구매간격을 나타낸다. 식 (4)에서 계산한 RFM 점수와, 식 (5)에 의하여 예측된 RFM 점수간의 차이가 RFM 점수 예측에 의한 오차이며 이는 C-CRF 모형에서 특징함수 $\mathbf{f}(\cdot)$에 반영된다.

### 3.3 단계 3 : 고객 간 유사도 계산

고객 간의 유사도는 다음 네 가지로 구분한다. 첫째 유사도는 성별과 나이를 기준으로 한 인구 통계학적 유사도(Demographic Similarity, DS)이다. 성별을 기준으로 한 유사도와 나이를 기준으로 한 유사도를 각각 구하여, 그 평균을 인구 통계학적 유사도로 사용한다. 서열형 변수를 기준으로 한 다음의 유사도를 사용한다(Tan *et al*., 2006).

$$\mathrm{s} = 1 - \frac{|x_i - x_j|}{n-1} \quad (6)$$

이 때, $x_i$와 $x_j$는 각각 $i$번째와 $j$째 데이터의 값이고 $n$은 해당 변수가 가질 수 있는 값의 개수이다. 즉, 성별의 경우에는 $n=2$이고, 나이의 경우에는(본 연구에서는 7개 구간으로 구분하였으므로) $n=7$이다. 각각의 유사도를 구하여 평균을 계산하면 식 (7)과 같다.

$$DS_{ij} = \frac{(1-|Gen_i - Gen_j|) + (1-|Age_i - Age_j|/6)}{2} \quad (7)$$

단, $Gen_i$와 $Age_i$는 각각 $i$번째 고객의 성별과 나이를 나타낸다.

두 번째 유사도는 반품횟수와 평균 구매간격을 기준으로 한 고객 간의 유사도인 구매 특성 유사도(Purchasing Characteristic Similarity, PS)이다. 연속형 변수를 기준으로 한 유사도는 다음과 같이 유클리드 거리를 변형하여 계산한다(Tan *et al*., 2006).

$$\mathrm{s} = \frac{1}{1+d_{i,j}} \quad (8)$$

이 때, $d_{i,j}$는 $i$번째 데이터와 $j$번째 데이터간의 유클리드 거리이다. 따라서 구매 특성 유사도는 다음과 같이 계산한다.

$$PS_{ij} = \frac{1}{1+\sqrt{(Return_i - Return_j)^2 + (PI_i - PI_j)^2}} \quad (9)$$

단, $Return_i$와 $PI_i$는 각각 고객 $i$의 반품횟수와 평균 구매간격을 나타낸다.

세 번째 유사도는 구매 이력 유사도(Product Purchase Similarity, PPS)로, 이는 총 17개 상품군의 구매 내역을 기준으로 한 고객 간의 유사도를 뜻한다. 구매 이력 유사도는 이진 변수 간의 거리를 구할 때 널리 사용되는 자카드(Jaccard) 계수를 이용하며 식 (8)과 같다(Tan *et al*., 2006).

$$\mathrm{J(A,\,B)} = \frac{|A \cap B|}{|A \cup B|} \quad (10)$$

따라서 고객 $i$와 고객 $j$간의 구매 이력 유사도는 다음과 같이 계산된다.

$$PPS_{i,j} = \frac{\text{고객 } i\text{와 고객 } j\text{가 모두 구매한 상품군의 개수}}{\text{고객 } i \text{ 혹은 고객 } j\text{가 구매한 상품군의 개수}} \quad (11)$$

네 번째 유사도는 통합 유사도(Unified Similarity, US)이다. 고객 $i$와 고객 $j$간의 통합유사도는 위의 세 유사도의 가중합으로 계산한다.

$$US_{i,j} = w_1 \times DS_{ij} + w_2 \times PS_{ij} + w_3 \times PPS_{ij} \quad (12)$$

이제 두 고객 간 통합유사도를 바탕으로 이웃을 결정한다. 한 고객의 이웃은 그 고객과의 통합유사도가 임계치 이상인 고객들의 집합이라 정의한다. 모든 고객의 이웃을 파악하기 위해 고객 네트워크를 사용한다. 고객 네트워크의 행과 열 모두 고객을 나타내며, 행렬의 각 셀은 고객과의 관계를 나타낸다. 고객 $i$와 고객 $j$의 관계를 나타내는 $i$행 $j$열의 값 $C_{ij}$는 다음과 같다.

$$C_{ij} = \begin{cases} 1, \text{고객 } i\text{와 고객 } j\text{의 유사도가 } \kappa \text{ 이상} \\ 0, \text{고객 } i\text{와 고객 } j\text{의 유사도가 } \kappa \text{ 미만} \end{cases} \quad (13)$$

단, $\kappa$는 고객 간의 관계가 있는지를 판단하는 유사도의 임계치이다.

### 3.4 단계 4 : C-CRF의 모수 추정

이 단계에서는 C-CRF 모델에 필요한 특징함수를 구성하고 C-CRF 모형을 완성한다. 본 연구에서는 식 (1)에 나타난 $f(y_i, \boldsymbol{x})$를 단계 1에서 계산한 고객의 실제 RFM 점수와 단계 2에서 다중회귀분석을 통해 계산된 고객의 RFM 점수 간의 차이를 나타내는 특징함수로 구성한다.

또한, $g(y_i, y_j, \boldsymbol{x})$는 이웃 고객과의 통합유사도와 고객 간의 등급 점수차이에 관련된 특징함수이다. 즉, 단계 1에서 계산한 고객의 RFM 점수와 단계 3에서 계산한 통합유사도를 바탕으로 함수를 구성한다.

특징함수를 구성한 후, 여기에 포함되는 모수 $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$는 식 (14)의 로그가능도(log likelihood)를 최대로 하는 값으로 추정한다.

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \log P(y_i | x_i; \boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (14)$$

추정된 모수를 바탕으로 식 (2)의 정규화 함수를 계산하여 모델을 확정한다.

### 3.5 단계 5 : 신규 고객의 고객 등급 점수 확률 분포 제시

신규 고객이 유입되면 다음과 같은 단계를 거쳐 신규 고객의 고객 등급 점수의 확률 분포를 제시한다.

단계 (1) : 신규 고객의 데이터를 정규화한다.
단계 (2) : 신규 고객의 RFM 점수를 계산한다.
단계 (3) : 신규 고객의 RFM 예측점수를 추정한다.
단계 (4) : 신규 고객에 대해 기존 고객들과의 통합 유사도를 계산하여 이웃 고객집단을 확정한다.
단계 (5) : 단계 (3)과 단계 (4)에서 각각 계산된 RFM 점수의 오차 및 이웃 고객과의 통합 유사도를 C-CRF 모형에 대입하여 신규 고객의 고객 등급 점수의 확률 분포를 얻는다.

## 4. 신규 고객 등급 점수 예측 예제

이 장에서는 제 3장에서 설명한 신규 고객의 등급 점수 확률분포 유도과정을 임의로 만든 가상의 고객 데이터를 통해 예시한다. 가상의 인터넷 쇼핑몰에서는 <Table 2>에서 제시된 상품군에 속하는 모든 상품을 판매하고 있으며, 상품을 구매하였던 기존 고객은 총 156명이고 이들의 구매 데이터를 바탕으로 신규 고객의 등급 점수를 예측하고자 한다.

### 4.1 단계 1 : 데이터 변환과 정규화

단계 1에서는 <Table 1>에서 제시된 바와 같이 데이터를 정규화하며 식 (4)에 의해 각 고객의 RFM 점수들을 계산한다. 여기서 가중치는 RFM 기법에서 일반적으로 널리 설정되는 가중치 $W_R = 0.2$, $W_F = 0.3$, $W_M = 0.5$를 그대로 사용한다(Kang *et al*., 2004). 그 결과 고객 등급 점수의 분포는 <Table 3>과 같다.

**Table 3.** Distribution of customer RFM score

| 고객 등급 점수 | 빈도 |
|---|---|
| -0.5 미만 | 19 |
| -0.5 이상 0미만 | 83 |
| 0 이상 0.5미만 | 33 |
| 0.5이상 1미만 | 13 |
| 1 이상 | 9 |

### 4.2 단계 2 : 다중회귀분석을 통한 고객의 예상 등급 변수 생성

단계 2에서는 나이, 성별, 반품 횟수, 평균구매간격을 독립변수로 하고 고객의 RFM 점수를 종속변수로 하는 다중회귀분석을 시행했다. 다중회귀분석의 결과는 <Table 4>와 같다.

**Table 4.** Result of multiple regression

| 모형 | 비표준화계수 | T 값 | 유의확률 |
|---|---|---|---|
| (상수) | -0.115 | -0.947 | 0.345 |
| 나이 | 0.062 | 1.827 | 0.070 |
| 성별 | -0.075 | -0.742 | 0.459 |
| 반품횟수 | 0.177 | 3.583 | 0.000 |
| 평균구매간격 | 0.259 | 5.239 | 0.000 |

<Table 4>에서 제시된 회귀계수를 바탕으로 식 (15)와 같은 회귀식을 구한다.

$$\hat{Y} = -0.115 + 0.062X_1 - 0.075X_2 + 0.177X_3 + 0.259X_4 \quad (15)$$

이 때, $X_1$, $X_2$, $X_3$, $X_4$는 각각 정규화 된 나이, 성별, 반품횟수, 평균구매간격을 나타낸다.

### 4.3 단계 3 : 고객간 유사도 계산

식 (7), 식 (9), 식 (10)에서 제시된 방법을 따라 각각 인구 통계학적 유사도, 구매 특성 유사도, 그리고 구매 이력 유사도를 계산하였다. 이때, 식 (12)에서 통합 유사도에 사용되는 가중치는 $w_1 = 0.2$, $w_2 = 0.4$, $w_3 = 0.4$로 임의로 설정하였다.

이렇게 계산된 고객 간 통합 유사도가 0.5 이상인 경우를 이웃이라고 설정하였다. 그 결과, 평균적으로 고객 당 25명의 다른 고객과 이웃 관계에 있는 것을 확인하였다.

### 4.4 단계 4 : C-CRF 특징 함수 구성과 모수 추정

Qin 등에서는 C-CRF를 이용하여 사용자의 질의(query)를 포함하는 문서들을 추출하고 이들의 순위를 매김에 있어서 문서의 내용정보뿐만 아니라 문서간의 관계를 이용하는 특징함수를 도입하였는데, 유사한 문서들은 랭킹점수도 유사할 것이라는 기본가정을 바탕으로 하고 있다(Qin *et al*., 2009). 본 연구의 목적이 신규 고객의 등급을 예측하기 위한 것이므로 문서 랭킹 방법과 유사하다고 보아 Qin 등에서 도입한 특징함수를 활용한다. 첫 번째 특징함수는 식 (16)와 같다.

$$f(y_i, x) = \Sigma_i - \alpha(PC_i - AC_i)^2 \quad (16)$$

이 때, $PC_i$와 $AC_i$는 각각 고객 $i$의 예측 등급 점수와 실제 등급 점수이다. 두 번째 특징함수는 식 (17)과 같다.

$$g(y_i, y_j, x) = \Sigma_{i,j} - \beta \times US_{i,j}(C_i - C_j)^2 \quad (17)$$

이 때, $C_i$와 $C_j$는 각각 i번째 고객과 j번째 고객의 RFM 점수이다.

고객 등급 점수에 관한 확률분포의 일반식은 (18)과 같다.

$$\Pr(y|x) = \frac{1}{z(x;\alpha,\beta)} exp(\Sigma_i - \alpha(PC_i - AC_i)^2 + \Sigma_{i,j} - \beta \times US_{i,j}(C_i - C_j)^2) \quad (18)$$

고객 등급 점수에 관한 확률분포 식 (18)에 포함된 모수 $\alpha$와 $\beta$를 추정하기에 앞서, $Z(x_i;\alpha,\beta)$에 대한 계산을 할 필요가 있다. 식 (2)에서 정의된 고객 $i$의 정규화 함수를 본 연구에서 도입한 특징함수에 맞게 정의하면 식 (19)과 같다.

$$Z(x_i;\alpha,\beta)=\int_{y=0}^{\infty}\exp(-\alpha(y-AC_i)^2-\beta\delta)dy \qquad (19)$$
$$=\exp(-\beta\delta)\int_{0}^{\infty}\exp(-\alpha(y-AC_i)^2dy$$

이 때, $\delta$는 $\Sigma_{i,j}US_{i,j}(C_i-C_j)^2$를 계산한 값이다. 식 (19)에서 적분 부분을 계산하기 위해 $y-AC_i=t$로 치환하면 $y=t+AC_i$, $dy=dt$를 얻으며 적분 범위는 $t=(-AC_i,\infty)$이 된다. 다시 $s=t\sqrt{2\alpha}$로 치환하면 $dt=\frac{1}{\sqrt{2\alpha}}ds$이고 적분 범위는 $s=(-\sqrt{2\alpha}AC_i,\infty)$이다. 결과적으로 식 (19)는 다음과 같다.

$$Z(x_i;\alpha,\beta)=\exp(-\beta\delta)\int_{-\sqrt{2\alpha}AC_i}^{\infty}\exp(-\frac{s^2}{2})ds \qquad (20)$$
$$=\exp(-\beta\delta)\sqrt{\frac{\pi}{\alpha}}(1-\phi(-\sqrt{2\alpha}AC_i))$$

여기서 $\phi(\cdot)$는 표준정규분포의 누적확률분포이다. 이제 식 (14)를 이용하여 모수 $\alpha$와 $\beta$를 추정한다. 식 (20)에서 구한 $Z(x_i;\alpha,\beta)$를 식 (18)에 대입하면 $\beta$가 약분되므로 모수 $\alpha$값을 고정시키고 $\beta$값이 변화하여도 최대 로그가능도값이 동일하다. 따라서 $\beta$를 1로 고정 후 $\alpha$를 바꿔가면서 추정을 하였다. 추정 결과, $\alpha=1.65$일 때 로그 가능도 값이 최대가 되었다.

### 4.5 단계 5 : 신규 고객의 등급 점수 확률 분포 제시

이제 신규 고객의 고객등급 점수의 확률분포를 구해 본다. 신규 고객의 예시 데이터는 <Table 5>와 같다.

**Table 5.** New customer's sample data

| 범주명 | 변수명 | 값 |
|---|---|---|
| 인구 통계 | 나이 | 24세 |
| | 성별 | 남성 |
| 구매특성 | 반품횟수 | 1회 |
| | 구매간격 | 32.0일 |
| RFM 변수 | 구매빈도 | 5회 |
| | 총 구매액 | 740,000원 |
| | 최근구매일 | 40일전 |
| 상품구매이력 | 상품군별 | P6, P11 |

신규 고객의 데이터를 정규화 후, RFM 점수를 계산한다. 신규 고객의 RFM 점수는 0.3207이었다. 그리고 유사도가 0.5 이상인 고객들을 이웃으로 설정하였다. 이 때, 신규 고객의 이웃들의 특성은 <Table 6>과 같다.

**Table 6.** Neighbor's data of new customer

| 범주명 | 변수명 | 값 |
|---|---|---|
| 인구 통계 | 나이 | 평균 28세 |
| | 성별 | 남성(6명) |
| 구매특성 | 반품횟수 | 평균 0.33회 |
| | 구매간격 | 평균 42.9일 |
| RFM 변수 | 구매빈도 | 평균 4회 |
| | 총 구매액 | 평균 128,416원 |
| | 최근구매일 | 평균 19.9일 전 |
| 상품구매이력 | 상품군별 | P1(3명), P3(2명), P4(2명), P6(1명), P8(1명), P10(2명), P11(3명) |

계산한 $\delta$값은 0.31이었으며 정규화 상수 $Z(x_i;1.65,1)$을 계산한 결과 0.7285였다. 이를 바탕으로 계산한 신규 고객의 등급 점수의 확률밀도함수는 식 (21)과 같다.

$$\Pr(Y|X)=\frac{1}{0.7285}\exp(-1.65\times(\mathrm{Y}-0.3207)^2-0.31) \qquad (21)$$

<Figure 2>는 식 (21)의 확률밀도함수를 그래프로 나타낸 것이다.

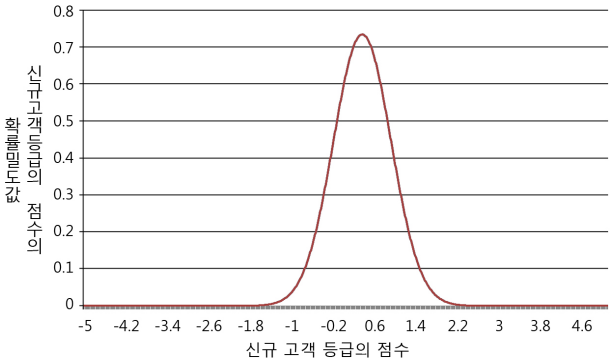


**Figure 2.** Predicted loyalty distribution of new customer

위의 확률밀도함수는 신규 고객 등급의 점수가 a점 이상 b점 이하일 확률이 얼마인지를 계산하는데 사용될 수 있다. 즉, 이를 이용하면 신규 고객이 충성고객이 될 것인지, 된다면 충성도는 얼마인지를 예측할 수 있다.

## 5. 결 론

본 연구에서는 신규 고객의 고객등급 점수를 예측하기 위해 C-CRF의 개념을 적용하는 방법을 제시하였다. 고객 등급은 RFM 기법을 적용하여 계산하였고, 예상 고객 등급은 다중 회귀 분석을 적용하여 계산하였다. 고객간 유사도는 고객의 인구 통계 특성, 구매 특성, 구매 이력을 기준으로 계산하였다. 또한 C-CRF를 이용하여 고객간 유사도를 반영한 고객 등급

점수의 확률분포를 제시하였다.

본 연구에서는 기존연구에서 유사문서의 순위를 정하는 모형에서의 C-CRF의 특징함수를 도입하였으며 이로써 신규 고객등급을 예측할 때 단순히 신규 고객의 특징만 사용하여 예측하는 것이 아니라 유사고객과의 유사성을 반영할 수 있다는 장점이 있다.

추후 연구과제로는 C-CRF를 구성하는 특징함수의 적절성을 어떻게 평가할 것인가가 연구되어야 한다. 그리고 RFM 기법에 적용된 가중치와 통합 유사도에 적용된 가중치를 설정하는 객관적인 방법도 개발될 필요가 있다.

## 참고문헌


Jeon, H-.C., Park, S.-S., Shin, Y.-G., and Jang, D.-S. (2007), A Study on Purchase Intention of Consumers in the Online Bookstore, *Proceedings of KIIE Spring Conference*, 720-725.

Jeon, H.-R. and Lee, D.-W. (2012), Prediction Loyal Customer and Conversion Method to Loyal Customer Using Purchasing Characteristic, *Proceedings of KIIE Spring Conference,* 2583-2597.

Jeong, Y.-P. and Yeon, C.-S. (2013), Customer Relationship Management of the Internet Shopping Mall Using Customer Segmentation, *Journal of Korean Institute of Information Technology*, **11**(12), 159-167.

Joe, Y.-B. and Kim, C.-B. (2006), An Effective Classifying Methodology for On-Line Retail Customers : Application to Decision Trees, *Journal of the Korean Operations Research and Management Science Society*, **19**(6), 2117-2134.

Kalacota, R. and Robinson, M. (1999), *e-Business : Roadmap for success,* Addision-Wesley.

Kang, C.-W., Lee, S.-W., Choi, S.-B., and Kim, K.-K. (2004), A Study of Construction Optimum RFM Model for Customer Segmentation, *Journal of the Korean Data Analysis Society*, **6**(6), 1829-1840.

Kim, H.-S. and Jeong, H.-G. (2012), A Diagnosis and Assessment Methodology for Enterprise CRM Strategy, *Journal of the Korean Operations Research and Management Science Society*, **37**(3), 23-37.

Kim, J.-K., Choi, I.-Y., Kim, H.-K., and Kim, N.-H. (2009), Social Network Analysis to Analyze the Purchase Behavior of Churning Customers and Loyal Customer, *International Journal of Management Science*, **26**(1), 183-196.

Kosta, R., Vladan, R., Slobodan, V., and Zoran, O. (2013), Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs, 27th *AAAI Conference on Artificial intelligence*, 840-846.

Kumer, V. and Reinartz, W. (2012), *Customer Relationship Management : Concept, Strategy and Tools*, Springer.

Lee, J.-H. and Lee, S.-J. (2004), Customer Classification System using Optimized Form in eCRM, *Proceedings of KFIS Autumn Conference*, **14**(2), 149-152.

Lee, Y.-K. and Choi, H.-K. (2002), Comparison Analysis of RFM Using a Statistic Technique, *Journal of Mathematics and Statistics*, **9**(1), 5-6.

Lim, S.-J., Seo, E.-H., and Jeong, T.-S. (2003), A Study of Dynamic Customer Segmentation at Internet Shopping Mall, *Proceedings of KIIE Spring Conference*, 587-591.

Park, C. and Jeon, J.-G. (2002), A CRM Strategy of Internet Shopping Mall : Focused on a Classification of Online Consumer Group by Buying Frequency and Mall Loyalty, *Journal of Information Technology Management and Application*, **19**(4), 127-149.

Park, J.-H., Joe, Y.-H., and Kim, J.-K. (2009), Social Network : A Novel Approach to New Customer Recommendations, *Journal of Intelligence and Information Systems*, **15**(1), 123-140.

Qin, T., Liu, T.-Y., Zhang, X.-D., Wang, D.-S., and Li, H. (2009), Global Ranking Using Continuous Conditional Random Fields, *Advances in Neural Information Processing Systems*, 1281-1288.

Radosavljevic, V., Vucetic, S., and Obradovic, Z. (2010), Continuous Conditional Random Fields for Regression in Remote Sensing, *Proceedings of the 2010 Conference on ECAI 2010 : 19th European Conference*, 809-814.

Shim, B.-H., Kim, M.-C., Ko, J.-Y., and Kim, S.-Y. (2011), Does Customer Relationship Management for VIP Customers Affect Repurchase and Positive WOM in Premium Hotels? *Journal of the Korean Operations Research and Management Science Society*, **37**(3), 185-206.

Shin, G.-H. (2014), *Annual and fourth quarter 2013 e-commerce and Cyber Shopping Trends Press*, Statistics Korea.

Song, H.-S. (2013), *e-CRM Implementation and Management Strategies,* Saerowoon Jaen.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006), *Introduction to Data Mining*, Addison Wesley.