지능정보연구 제17권 제4호 2011년 12월(pp.95~108)

고객별 구매빈도에 동적으로 적응하는 개인화 시스템 : 음료수 구매 예측에의 적용

박윤주
서울과학기술대학교 국제융합학부 글로벌테크노경영학과
(yjpark@seoultech.ac.kr)

인터넷 비즈니스의 활성화에 따라서 기업은 고객의 인물정보 및 거래정보를 활용하여 보다 맞춤화된 개인화 서비스를 제공하고 있다. 기존의 고객군별 예측기법은 유사한 고객들을 군집화하여 고객군별로 예측모델을 수립하는 것으로, 구매가 많고 충성도가 높은 핵심고객에게 요구되는 일대일 서비스를 제공하는 데는 한계가 있다. 반면 일대일 고객별 예측기법은 각 고객에게 고도로 맞춤화된 서비스를 제공하지만, 과거 구매이력이 많지 않은 고객 이나 신규 고객에게는 정확한 개인화 서비스를 제공하지 못한다. 본 연구는 고객의 구매빈도에 따라서 유사 고객들과의 군집화 수준을 동적으로 조정하는 새로운 지능형 개인화 시스템을 제안한다. 제안된 시스템은 과거 구매가 많은 고객들에 대해서는 일대일 예측모델을 수립하지만, 구매 빈도가 낮은 고객의 경우 다른 고객들과의 최적화된 군집화를 통해 예측모델을 수립한다. 본 기법을 Neilsen의 음료수 구매 데이터셋에 적용하여 고객의 일회 구매금액 및 구매품목을 예측한 결과, 기존 두 예측기법들에 비하여 적정한 계산비용(computational cost)으로 더욱 정확한 개안화 서비스를 제공할 수 있음을 확인하였다.

논문접수일 : 2011년 07월 05일 게재확정일 : 2011년 08월 09일
투고유형 : 학술대회우수논문 교신저자 : 박윤주

1. 서론

인터넷 비즈니스의 활성화에 따라서 기업은 고객의 인물정보, 구매 내역, 웹페이지 체류시간, 클릭한 상품들 목록에 이르기까지 다양한 활동정보를 수집하고 활용할 수 있게 되었다. 이에 따라서, 많은 e-biz 기업들이 유사한 고객군을 세분화하여 고객별 특징에 적합한 맞춤형 서비스를 제공하고 있으며, 최근에는 이를 넘어서 각 고객별로 일대일 개인화된 서비스를 제공하려는 노력이 증가하고 있다. 이러한 맞춤형 서비스를 제공하기 위해서는 고객이 선호하는 제품, 구매 시간, 소비금액 등의 구매 패턴을 정확하게 예측하는 것이 중요하며, 이러한 다양한 정보를 반영한 분석을 수행하기 위하여 데이터마이닝 기법이 널리 활용되고 있다.

기존의 고객군별 예측기법(Customer Segmentation Method)은 나이, 성별, 결혼유무 등 유사한 특징을 가지는 고객들을 군집화한 후, 동일 고객군별로 하나의 데이터마이닝 예측모델을 수립한다(Wedel and Kamakura, 2000; Smith, 1956). 따라서 같은 고객군에 속한 고객들의 소비성향을 예측할 때에는 동일한 예측모델이 적용된다. 이러한 고객군별 예측기법은 예측모델의 수를 적정수로 유지할 수 있고, 또한 구매가 적은 고객들이 다른 유

사고객들의 구매 데이타를 활용할 수 있다. 그러나 본 기법은 군집화로 인하여 고객별로 맞춤화 정도가 낮기 때문에, 구매가 많은 VIP 고객 또는 장기간 해당 기업을 이용하는 충성고객 등에 대한 일대일 개인화 서비스를 제공하는 데는 한계가 있다.

다른 기법으로 일대일 예측기법(1-to-1 Method)은 각 고객을 위한 독립적인 예측모델을 수립하는 것으로, 개별 고객의 데이터만을 학습하여 예측모델을 수립하므로 개인별 맞춤화 정도가 높고, 학습 데이터 수가 적어 고객군별 예측기법에 비해 예측에 소요되는 시간이 짧다(Peppers and Rogers, 1993; Adomavicius and Tuzhilin, 2005). 그러나 신규고객이나 구매빈도가 낮은 고객 등 정확한 예측모델을 수립할 만큼 충분한 데이터를 보유하고 있지 못한 고객에게는 정확한 개인화 서비스를 제공하기 어렵고, 고객수의 증가에 비례하여 예측 모델 수가 증가하므로 모델관리의 확장성(scalability) 측면에서도 한계를 갖는다.

본 논문은 기존 개인화 기법의 한계를 극복하고, 고객의 구매 빈도에 따라서 고객군의 군집화 정도를 동적으로 조절하는 새로운 *지능형 고객세분화 기법(Intelligent Customer Segmentation Method)*을 제안한다. 제안된 기법은 구매빈도가 많은 핵심 고객에게는 고도로 맞춤화된 일대일 서비스를 제공하지만, 구매빈도가 부족한 고객에게는 데이터 희소성을 해결하기 위한 군집화를 수행하는 것을 목적으로 한다. 이를 위하여, 본 논문은 구매빈도가 특정 임계치 보다 높은 고객에 대해서는 일대일 추천을 수행하고, 해당 임계치 보다 적을 경우에는 유사한 다른 고객들과의 군집에 대한 예측모델을 수립하도록 하였으며, 이러한 구매빈도 임계치를 "*슬라이딩 윈도우 상관분석*"이라는 기법을 통하여 산출할 것을 제안하였다.

본 연구는 Neilsen의 음료수 구매 데이터셋에 적용하여 고객의 일회 구매금액 및 구매품목을 예측하였으며, 그 결과 제안된 *지능형 고객세분화기법(Intelligent Customer-Segmentation Method)*을 기존의 일대일 예측기법(*1-to-1 Method*)에 비하여 많은 경우 우수한 예측성능을 도출하였다. 또한, 기존의 *고객군별 예측기법(Customer Segmentation Method)*에 비해서 훨씬 적은 수의 데이터를 사용하면서도 동일한 수준의 예측을 수행할 수 있어 계산비용 측면에서도 장점이 있음을 확인하였다.

본 논문의 나머지 부분은 다음의 다섯 장으로 구성된다. 우선, 제 2장에서는 데이터 희소성으로 인한 예측모델의 성능저하에 대한 관련 연구를 조사하였다. 제 3장에서는 *일대일 기법*과 *고객군별* 기법에 대한 사전 분석을 수행하였으며, 이러한 분석을 바탕으로 새로운 *지능형 고객세분화 기법*을 제안하였다. 다음으로 제 4장에서는 본 연구의 분석환경 및 결과를 제시하였다. 마지막으로 제 6장에서는 본 연구의 결론 및 향후 연구방안을 기술하였다.

2. 관련 연구

데이터마이닝의 다양한 기법들이 고객의 정보를 분석하고, 행동패턴을 예측하여 고객관계관리를 강화하는데 널리 활용되고 있다. 특히, 데이터마이닝 기법은 최근 추천 시스템에서 고객의 상품 선호도 예측에 많이 활용되는 협업필터링 기법과는 달리, 고객의 나이, 성별, 결혼여부 등의 인물정보를 예측에 반영할 수 있고, 단순한 상품 선호도 이외에도 고객이 언제 물건을 구매할지, 얼마나 구매할지, 어떠한 카테고리에서 구매할지 등의 다양한 예측이 가능하기 때문에 보다 정교한 분석을 수행할 수 있다는 장점이 있다. 따라서, 이러한 데

이터마이닝 기법을 장점을 활용하여 고객의 구매 성향 예측 및 추천에 활용한 다수의 연구가 존재한다. Park and Tuzhilin(2008)은 회귀분석, Case-Based Reasoning(CBR), Support Vector Machine (SVM) 등 9개의 데이터마이닝 기법을 사용하여 고객의 선호도를 예측하는 연구를 수행하였으며, 김경재, 김병국(2005)도 데이터마이닝을 이용한 인터넷 쇼핑몰의 상품추천 시스템을 제안하였다. Pazzani and Billsus(1997)는 클러스터링, 의사결정나무, 신경망, 베이시안등 다양한 데이터마이닝 기법을 사용하여 사용자가 선호하는 웹페이지를 분석하고 추천하는 시스템을 제안하였으며, Adomavicius and Tuzhilin(2005)도 다양한 데이터마이닝 기법으로 고객의 프로필을 생성하여 개인화된 추천을 수행하는데 활용하였다. 그 외에도 Lawrence et al.(2001) 은 고객이 슈퍼마켓에서 물품을 구매하는데 개인화 서비스를 제공하기 위해 데이터마이닝 기법을 사용할 것을 제안하였다.

그러나 데이터마이닝 기법들은 학습에 사용하는 데이터수가 증가할수록 모형 도출에 오랜 시간이 소요되고, 도출된 모형은 복합해져 계산비용(computational cost)가 급격히 증가하는 경향이 있다. 이와는 반대로, 학습에 사용할 데이터 수가 부족할 경우에는 데이터 희소성으로 인하여 도출된 모형의 예측 성능이 저하된다. 즉, 신규로 가입하였거나 구매빈도가 부족한 기존 고객들의 경우 데이터 희소성 문제로 정확한 추천이나 맞춤형 서비스를 제공하는데 한계가 있다. 이러한 문제를 해결하기 위한 기존 연구로는 Schein et al.(2002)이 content-based 기법을 사용하여 부족한 거래정보를 보완하는 방안을 제시하였으며, Jiang and Tuzhilin(2009)은 몇몇 유사고객을 거래에 기반하여 군집화하는 micro-segmentation 기법으로 데이터 희소성을 해결하는 방안을 연구하였다.

3. 고객별 구매빈도에 따른 동적 개인화 추천시스템

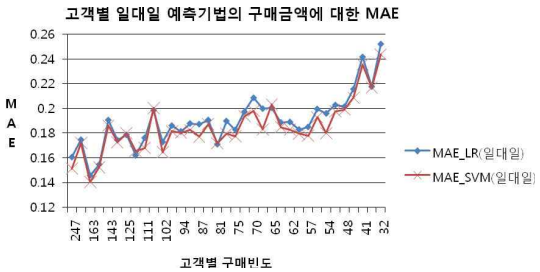
본 장에서는 기존의 *고객군별 예측기법*과 *일대일 예측기법*을 비교분석하고, 고객의 구매빈도에 따라 동적으로 적응하는 새로운 *지능형 고객세분화 기법*을 제안한다. 우선, 제 3.1절에서는 각 개인별로 일대일 맞춤화된 예측 모델을 수립하는 "*일대일 예측기법*"을 기술하고, 이를 음료수 구매 예측에 적용한 사전분석 결과를 제시한다. 다음으로 제 3.2절에서는 유사한 고객군을 군집화한 후, 군집된 고객군별 예측모델을 수립하는 "*고객군별 예측기법*" 및 실험 결과를 제시한다. 제 3.3절에서는 본 논문에서 제안하는 고객의 구매빈도에 따라서 동적으로 적응하는 "*지능형 고객세분화기법*"의 알고리즘을 제시한다. 다음 각 절에서는 이를 보다 세부적으로 기술하고 있다.

3.1 일대일 예측기법(1-to-1 method)

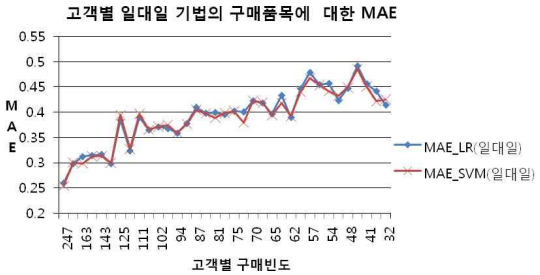
고객은 나이, 성별, 직업과 같은 인물정보를 가지며, 상품은 가격, 카테고리, 출시일 등의 상품 특성에 대한 정보를 갖는다. 고객이 상품을 구매하면, 구매한 "요일", "소비금액", "상품 카테고리" 등의 거래정보가 발생된다. 일대일 추천모델은 각 고객별로 거래정보를 취합하고, 이를 학습 데이터와 테스트 데이터로 구분한 후, 학습 데이타를 활용하여 예측모델을 수립한다. 예를 들어, 고객 A가 과거에 다섯 번 상품을 구매한 이력이 있다면, 이 다섯 번에 대한 거래 정보만을 활용하여 예측모델이 수립되는 것이다. 일대일 추천모델의 경우, 모델 생성에 사용되는 고객의 인물정보는 모두 동일하기 때문에 예측모델에 영향을 미치지 않으며, 따라서 거래 데이터만을 학습하여 예측모델이 수립한다.

본 사전분석은 제 4장에서 기술한 바와 같이, Neilsen의 음료수 구매데이타에 적용하였다. 실험에는 Support Vector Machine과 회귀분석 두 가지의 데이터마이닝 기법을 사용하여, 두 개의 종속변수 "소비금액"과 "상품 카테고리"를 예측하였다. 학습 및 테스트에는 10-fold cross validation이 사용되었으며, 예측성능은 Mean Absolute Error(MAE)와 Root Mean Square Error(RMSE) 예측오차를 통하여 산출하였다.

다음으로는, 고객의 구매빈도와 구매성향에 대한 예측오차의 관계를 파악하기 위하여 고객의 구매빈도를 X축에 표기하고, 예측오차를 Y축에 표기하여 <그림 1>, <그림 2>와 같이 나타내었다.



<그림 1> 일대일 기법의 구매금액에 대한 예측오차



<그림 2> 일대일 기법의 구매품목에 대한 예측오차

그 결과, 일대일 추천모델은 구매빈도가 높은 그림의 좌측의 고객들에 대해서는 낮은 예측오차를 도출하지만, 구매빈도가 낮은 우측의 고객일수록 점차 예측오차가 증가하는 경향이 있음을 볼 수 있었다. 이러한 현상을 통계적으로 검증하기 위하여, 본 연구는 선호도 평가 횟수 x와 예측오차 y 간에 상관분석을 수행하였으며, 그 결과를 <표 1>에 제시하였다. <표 1>에서 보는 바와 같이, 구매빈도와 구매금액에 대한 예측오차 MAE 사이에는 회귀분석을 적용했을 때 -0.72의 상관계수가 도출되었으며, 구매품목에 대한 MAE 사이에는 상관계수 -0.9가 도출되어 두 변수 사이에 강한 음의 상관관계가 있음을 알 수 있다. 마찬가지로, SVM 데이터마이닝으로 예측을 수행한 경우에도 상관계수가 구매금액의 경우 -0.71, 구매품목의 경우 -0.91가 도출되어 강한 음의 상관관계 존재하였다.

<표 1> 각 기법 별 구매빈도와 예측오차간 피어슨 상관계수

Pearson's Corr. Coeff.	구매빈도 vs. 구매금액	구매빈도 vs. 구매품목
MAE_회귀분석	-0.7204	-0.9006
MAE_ SVM	-0.7061	-0.9055
MAE_회귀분석	-0.7957	-0.9015
MAE_ SVM	-0.762	-0.906

이러한 사전분석을 통하여, 기존 일대일 예측기법은 고객의 구매빈도가 높은 경우에는 고객의 구매성향을 정확한 예측을 수행할 수 있지만, 고객의 구매빈도가 낮을수록 점차 예측 성능이 선형적으로 감소한다는 것을 통계적으로 확인하였다.

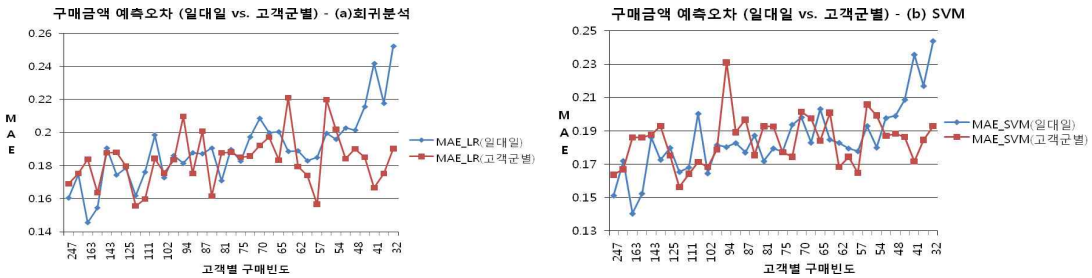
3.2 고객군별 예측기법(Customer Segmentation Method)

고객군별 추천모델은 고객의 나이, 성별, 직업 등과 같은 인물정보가 유사한 고객들을 하나의 그

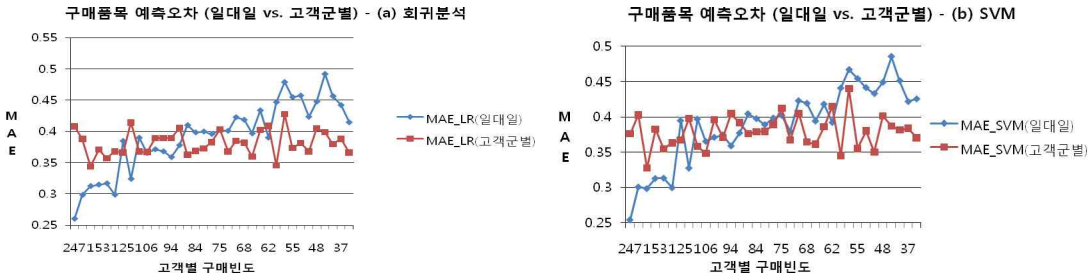
룹으로 군집화하고, 같은 군집으로 분류된 고객들의 인물 데이타 및 거래 데이타를 모두 사용하여 예측모델을 수립한다. 예를 들어, 고객 A에 대한 예측모델 수립 시, A와 유사한 성향을 갖는 고객 B, C, D를 군집화한 후, 이들의 인물 및 거래정보를 취합하여 학습에 활용하는 것이다. 고객군별 예측모델은 수립 시에는 이질적인 인물정보를 갖는 고객들이 하나로 군집화되었기 때문에, 예측모델 수립 시에 고객의 거래정보 뿐만 아니라 인물정보도 독립변수로 활용하여 학습을 수행한다. *고객군별 예측기법*은 *일대일 예측방법*에 비해 고객군의 "클러스터링 수"라는 하나의 추가적인 요소를 갖게 된다. 본 실험에서는 k-means 클러스터링(Lloyd, 1982) 기법을 사용하여 클러스터링을 수행하였으며, 전체의 고객들을 100개의 군집으로 세분화하여 사전분석을 수행하였다. 일대일 예측기법과 마찬가지로 10-fold cross validation을 사용하여 실험을 수행하였고, MAE와 RMSE 예측오차로 성능을 확인하였다.

다음에서는 *고객군별 예측기법*의 성능을 제 3.1절에서 소개한 일대일 예측기법과 비교하여 분하였다. 이러한 비교는 두 개의 데이터마이닝 기법(회귀분석, SVM), 두 개의 종속변수(구매금액, 구매품목), 두 개의 예측오차(MAE, RMSE)에 대하여 각각 수행되었으므로, 총 8번의 비교 실험을 수행하였다. <그림 3>과 <그림 4>는 이 중 4개의 비교실험 결과를 제시하고 있다.

이러한 비교분석을 수행한 결과, *고객군별 예측기법*은 *일대일 기법*에 비하여 구매빈도가 적은 고객에 대한 예측오차가 상대적으로 낮아 더욱 정확한 예측을 수행할 수 있음을 알 수 있었다. 그러나 고객의 구매 빈도가 많은 고객들에게는 <그림 3>의 좌측에서 볼 수 있듯이, *고객군별 예측기법*의 성능이 *일대일 기법*보다 우수하지 않으며, <그림



<그림 3> 구매금액에 대한 일대일 vs. 고객군별 예측기법의 MAE 예측오차



<그림 4> 구매품목에 대한 일대일 vs. 고객군별예측기법의 MAE 예측오차

4>의 경우 오히려 예측오차가 더 높게 도출되어 일대일 기법보다 예측성능이 낮음을 볼 수 있다. 즉, 클러스터링을 통한 고객 군집화는 구매빈도가 부족한 고객들에게는 예측성능을 향상시킬 수 있지만, 구매빈도가 많은 고객들에게는 예측성능 향상의 효과가 낮으며, 오히려 예측모델의 고객별 맞춤화 수준을 떨어뜨려 예측성능 저하의 원인으로 작용할 수 있음을 보여주고 있다.

3.3 지능형 고객세분화 기법(Intelligent Customer Segmentation Method)

일대일 기법은 각 고객별로 개인화 정도가 높지만, 구매빈도가 적은 고객에게 정확한 예측을 수행하지 못한다. 반면 고객군별 기법은 구매빈도가 적은 고객들에게 일대일 기법보다 향상된 예측성능을 보이지만, 구매빈도가 높은 핵심고객에게는 오히려 예측성능이 저하될 뿐만 아니라 불필요한 학습 데이타 증가로 인하여 계산비용 증가를 초래한다.

본 절에서는 고객의 구매빈도에 따라서 고객군의 군집화 정도를 동적으로 조절하는 새로운 *지능형 고객세분화 기법*을 제안한다. 제안된 기법은 고객의 구매빈도 α가 기준점 $\alpha_{boundary}$보다 클 경우에는 개별 고객에 대해 일대일 예측모델을 수립하지만, 고객의 구매빈도 α가 이보다 적을 경우에는 유사고객들과 군집화한 후 고객군별로 예측 모델을 수립한다. *일대일 예측기법*과 *고객군별 예측기법* 간의 경계가 되는 구매빈도 기준점 $\alpha_{boundary}$는 구매빈도 감소에 따라서 예측오차가 급격하게 증가하는 지점으로 설정하였다.

본 연구는 구매빈도 기준점 $\alpha_{boundary}$를 산출하기 위하여, <그림 5>와 같은 "*슬라이딩 윈도우 상관분석*"이라는 알고리즘을 제안한다. 제안된 기법은 전체 고객들을 구매빈도 내림차순으로 정렬한 후, 구매빈도가 높은 고객들부터 윈도우 사이즈 w만큼의 고객 구간을 설정한다. 설정된 구간에 속한 고객들은 일대일 예측기법을 적용하여 구매성향을 예측하게 되며, 실제 값과의 예측 값의 차이를 예측오차를 통하여 산출한다. 다음으로는, 해당 구간 내에서 예측오차가 급격히 증가하였는지 확인하기 위하여, 예측오차와 구매빈도간 상관분석을 수행하여 상관계수 β를 산출한다. 이렇게 산출한

Step 1 : 고객들을 구매 빈도에 따라서 내림차순으로 정렬한다. 초기 $\alpha_{boundary}$는 가장 많이 구매한 고객의 구매빈도 α_{max}로 설정한다. : $\alpha_{boundary} = \alpha_{max}$

Step 2 : 슬라이딩 윈도우 사이즈 w만큼의 고객들에 대한 구매성향을 일대일 기법을 적용하여 예측한 후, 각 고객별 예측오차를 산출한다.

Step 3 : 슬라이딩 윈도우 구간 w안에서 구매빈도와 예측오차간의 상관계수 β를 산출한다.

Step 4 : 산출된 상관계수 β와 음의 임계치 $\beta_{criterion}$을 비교한다.

1) $\beta < \beta_{criterion}$인 경우
해당 슬라이딩 윈도우 구간의 마지막 번째 구매 빈도인 α_{w_last}를 구매빈도 기준점인 $\alpha_{boundary}$로 설정한다. : $\alpha_{boundary} = \alpha_{w\_last}$

2) $\beta >= \beta_{criterion}$인 경우
슬라이딩 윈도우를 우측으로 한 칸 이동시켜, 구매빈도가 한 단계 낮은 고객들을 대상으로 슬라이딩 윈도우 w를 재설정한 후, Step 2 과정부터 다시 수행한다.

<그림 5> 슬라이딩 윈도우 상관분석 알고리즘

상관계수 β는 음의 상관관계가 있다고 판단되는 임계 값 $\beta_{criterion}$와 비교하는데, 만약 β 값이 $\beta_{criterion}$ 보다 더욱 강한 음의 상관관계를 가질 경우에는 해당 윈도우 구간의 마지막 구매빈도인 α_{w_last}를 $\alpha_{boundary}$로 설정한다. 그러나 상관계수 β가 $\beta_{criterion}$ 보다 약한 상관관계를 가진다면, 해당 구간에서는 구매빈도 감소로 인한 예측성능 저하가 발생하지 않았다고 판단하고, 윈도우 구간을 <그림 6>과 같이 오른쪽으로 한 칸 이동하여 구매빈도가 한 단계 적은 고객들을 대상으로 새로운 구간을 설정한다. 이렇게 새로 설정된 윈도우 구간의 고객들을 대상으로, 위 과정을 반복한다. 본 연구에서는 강한 음의 상관관계가 있다고 판단하는 임계치 상관계수 $\beta_{criterion}$을 -0.7로 설정하였다.

제안된 알고리즘은 고객이 구매빈도 임계값 $\alpha_{boundary}$보다 적은 구매빈도를 가지면 다른 유사고객들과 군집화를 수행하게 되는데, 이때 군집당 평균적으로 $\alpha_{boundary}$개 이상의 데이터를 가질 수 있도록 클러스터링 수 C를 식 (1)을 사용하여 설정하였다.

$$C = \Sigma(\#\text{of transaction for } U_i)/\alpha_{boundary} \quad (1)$$

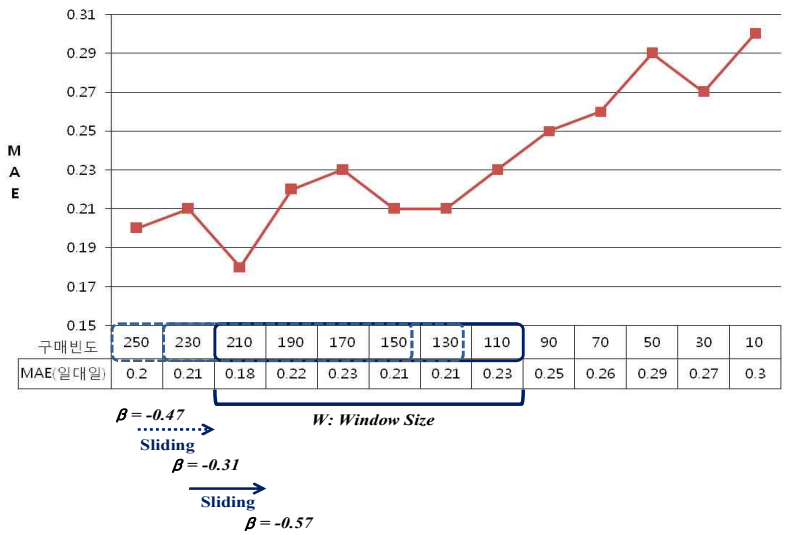
(C : *클러스터링 수*, $\alpha_{boundary}$: 구매빈도 임계값, U_i : $\alpha_{boundary}$보다 적은 구매빈도를 가진 고객)

4. 분석

다음에서는 본 연구의 분석환경과 논문에서 제안한 *지능형 고객세분화 기법*을 실제 사례인 Neilsen 음료수 구매 데이타에 적용한 결과를 제시한다.

4.1 분석환경

본 분석에는 Neilsen 패널리스트 데이터가 사용되었다. 이 데이터는 1992년부터 1993년 동안 1508개 가구에 대한 음료수 구매정보를 수집한 것으로 구매자 성별, 나이, 구매요일, 구매 금액, 수량, 구매 품목, 쿠폰사용여부, 인종, 자녀유무, 가족구성원 규모, 수입, 보유 TV수 등의 정보로 구성되어 있다. 가구별 최대 구매 구매빈도는 568번이며, 최



<그림 6> 슬라이딩 윈도우 상관분석 사례

소 구매빈도는 30번이다. 본 실험에는 구매금액과 구매품목 두 개의 종속변수가 사용되었으며, 각 종속변수를 예측하기 위한 모델 수립 시에는 데이터마이닝 기법들 중 회귀분석과 Support Vector Machine(SVM)이 사용되었다. 수립된 예측모델의 성능 측정에는 예측오차인 Mean Absolute Error (MAE)와 Root Mean Square Error(RMSE)를 사용하였다. 결과적으로, 본 실험은 *일대일 예측*, *고객군별 예측*, *지능형 고객세분화* 기법 각각에 대하여 8차례 실험이 수행되었다. 이 때, *고객군별 예측기법*과 *지능형 고객세분화 기법*에서 유사 고객들을 군집화하기 위해 클러스터링 기법 중 *k-means clustering*(Lloyd, 1982)을 활용되었다. 본 실험은 10-fold cross validation을 사용하여 수행하였으며, 본 실험에 사용한 모든 데이터마이닝 기법들은 java 기반 패키지인 weka(Witten and Frank, 1996)를 사용하여 구현하였다.

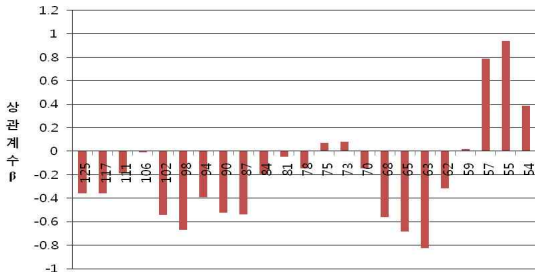
$$MAE = \frac{\sum_{i=1}^{n} |r_t - \hat{r}_t|}{n} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (r_t - \hat{r}_t)^2}{n}} \quad (2)$$

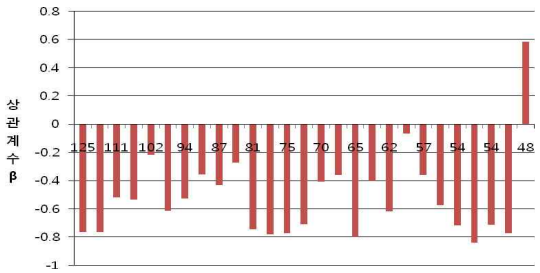
(r_t : 실제값, $\hat{r}_t$: 예측값, n : 테스트 데이타 수)

4.2 분석결과

본 절에서는 제 3장에서 제안한 *지능형 고객세분화기법*을 실제 마케팅 데이터인 Neilsen에 적용한 분석 결과를 제시한다. 우선, 제 3장에서 제안한 '슬라이딩 윈도우 상관분석' 알고리즘으로 *일대일 기법*과 *고객군별 예측기법* 간의 경계점이 되는 구매빈도 $\alpha_{boundary}$를 산출하였다. 다음의 <그림 7>과 <그림 8>은 윈도우 구간을 이동시키면서 '슬라이딩윈도우 상관분석'를 수행했을 때의 결과를 나타내고 있다. 본 연구에서는 상관계수의 임계치 $\beta_{criterion}$ 를 -0.7로 설정하였으며, 따라서 윈도우구간의 상관계수 β가 -0.7이하인 구간의 구매빈도가 $\alpha_{boundary}$로 설정된다. 예를 들어, 구매금액을 종속변수로 사용한 <그림 7>의 결과에서는 $\alpha_{boundary}$가 63으로 설정되며, 구매품목을 종속변수로 사용한 <그림 8>에서는 $\alpha_{boundary}$는 125가 설정된다.



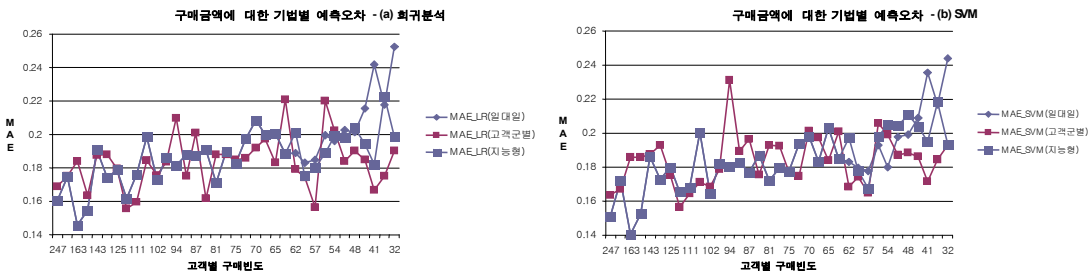
<그림 7> 슬라이딩 윈도우 구간별 구매금액 MAE_회귀분석 vs. 구매빈도간 상관계수



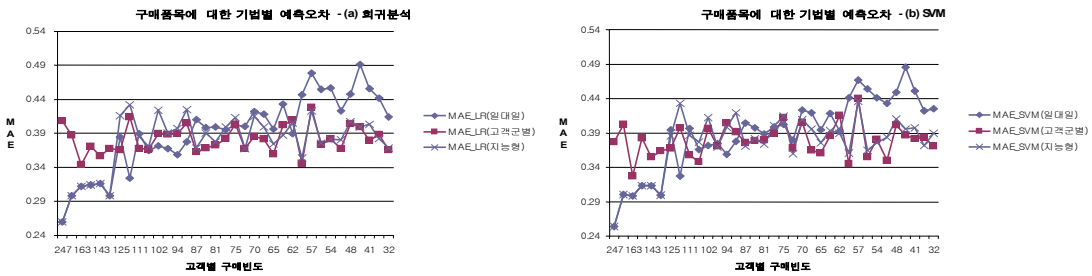
<그림 8> 슬라이딩 윈도우 구간별 구매품목 MAE_회귀분석 vs. 구매빈도간 상관계수

다음으로, 이렇게 찾은 경계점 구매빈도 $\alpha_{boundary}$를 사용하여 *지능형 고객세분화 기법*을 수행한 결과를 <그림 9>과 <그림 10>에 나타내었다. <그림 9>는 데이터마이닝 기법 중 회귀분석 및 SVM을 사용하여, 종속변수인 구매빈도를 예측한 결과를 나타내고 있다. 마찬가지로, <그림 10>은 구매품목을 종속변수로 하였을 때의 결과를 제시한다. 구

매빈도 변화에 따른 예측오차 증감 추이를 파악하기 위하여, x축에는 구매빈도를 내림차순으로 정렬하여 나타내었고, y축에는 MAE 예측오차를 나타내었다. 그 결과에서 볼 수 있듯이, 제안된 *지능형 고객세분화 기법*은 구매빈도가 높은 고객들에게는 일대일 기법과 동일한 예측방식을 적용하기 때문에, <그림 9>와 <그림 10>의 왼편에서는 두 기법의 예측오차에 차이가 없다. 그러나 구매빈도가 $\alpha_{boundary}$보다 적은 오른편에 위치한 고객들에 대해서는 *지능형 고객세분화 기법*의 예측오차가 *일대일 기법*보다 낮게 도출되는 경향이 있다. 이러한 성능 차이를 보다 체계적으로 비교하기 위하여, *일대일 기법* vs. *고객군별 기법* vs. *지능형 고객세분화 기법*의 평균 예측오차를 비교하여 순위를 부여하는 순위 테스트(rank test)를 수행하여, <표 2>와 <표 3>에 결과를 나타내었다. 구매금액을 예측하였을 때 각 기법 별 성능은, <표 2>에서 제시한 바와 같이, *지능형 고객세분화 기법이* 기존의 다른 두 기법에 비해 총 4회 중 2회에서 가장 우수한 예측성능을 도출함을 볼 수 있었다. 또한, 구매품목에 대한 예측을 수행하였을 때에는 <표 3>과 같이, *지능형 고객세분화 기법*의 예측성능이 4회 모두 가장 우수한 결과를 도출하였다. 이러한 성능차이가 통계적으로 유의한지 확인하기 위하여 *지능형고객세분화 기법*과 *일대일 기법*의 예측성능을 paired t-test로 비교한 결과를 <표 4>에 제시하였다. 그 결과, 총 8회 중 4회에서 *지능형 고객세분화 기법*이 *일대일 기법*보다 95% 유의수준에서 통계적으로 낮은 예측성능을 도출하였으며, 나머지 4회는 두 기법상 성능차이가 없는 것으로 나타났다. 마찬가지로, <표 5>에서는 *지능형 고객세분화 기법*을 기존의 *고객군별 예측기법*과 paired t-test를 통하여 비교하였으나, 두 기법의 예측성능은 통계적으로 유의한 차이가 없는 것으로 나타났다. 그



<그림 9> 구매금액에 대한 일대일 vs. 고객군별 vs. 지능형기법 예측오차



<그림 10> 구매금액에 대한 일대일 vs. 고객군별 vs. 지능형기법 예측오차

러나 *지능형 고객세분화 기법*은 계산비용(computational cost)의 측면에서 *고객군별 예측기법*에 비해 훨씬 적은 비용으로 동일한 수준의 성능을 도출한다. 본 실험에서 *지능형 고객세분화 기법*은 구매금액 예측 시 평균 130개, 구매품목 예측 시 평균 424개 데이터를 모델 수립에 사용하였으나, *고객군별 예측기법*은 평균적으로 1859개의 데이터를 사용하였다.

<표 2> 구매금액 예측성능에 대한 순위 테스트 결과

순위	1	2	3
Aver.MAE_LR	고객군별 0.1833	지능형 0.1858	일대일 0.1898
Aver.MAE_SVM	지능형 0.1836	고객군별 0.1838	일대일 0.1848
Aver.RMSE_LR	고객군별 0.4149	지능형 0.4180	일대일 0.4239
Aver.RMSE_SVM	지능형 0.4152	고객군별 0.4159	일대일 0.4173

<표 3> 구매품목에 대한 예측성능에 대한 순위 테스트 결과

순위	1	2	3
Aver.MAE_LR	지능형 0.3776	고객군별 0.3815	일대일 0.3920
Aver.MAE_SVM	지능형 0.3746	고객군별 0.3792	일대일 0.3895
Aver.RMSE_LR	지능형 0.6682	고객군별 0.6731	일대일 0.6897
Aver.RMSE_SVM	지능형 0.6623	고객군별 0.6661	일대일 0.6820

<표 4> 쌍체 t-검정결과(지능형고객세분화 vs. 일대일)

종속변수	Err.*지능형* < Err.*일대일*	Err.*지능형* > Err.*일대일*	Err.*지능형* = Err.*일대일*
구매 금액	1	0	3
구매 품목	3	0	1

<표 5> 쌍체 t-검정결과(지능형고객세분화 vs. 고객군별)

종속변수	Err.*지능형* < Err.*고객군별*	Err.*지능형* > Err.*고객군별*	Err.*지능형* = Err.*고객군별*
구매 금액	0	0	4
구매 품목	0	0	4

5. 결론

본 연구는 기존의 *일대일 예측기법*이 구매빈도가 부족한 고객의 구매성향을 정확히 예측하지 못하는 문제와, 기존의 *고객군별 예측기법*이 구매빈도가 높은 핵심고객에게 불필요한 계산비용(computational cost) 증가를 초래하면서도 맞춤화 정도는 낮은 서비스를 제공한다는 한계를 제시하였다. 이러한 문제를 해결하기 위하여 본 논문에서는 구매가 많은 고객들에게는 일대일 예측모델을 수립하지만, 구매 빈도가 낮은 고객들은 유사 고객들과 군집화하여 예측모델을 수립하는 새로운 *지능형 고객세분화 기법*을 제안하였다.

본 연구는 다음의 세 가지 측면에서 의의를 갖는다. 첫째, 기존의 고객별 *일대일 예측기법*이 구매빈도가 적은 고객의 구매성향을 정확히 예측하지 못하는 문제를 실제 사례에 적용하여 통계적으로 검증하였다는 점에서 의의가 있다. 둘째, 구매빈도에 따라서 동적으로 적응하는 *지능형 고객세분화 기법*을 제안하여, 구매빈도가 낮은 고객들에게는 데이터 희소성 문제를 해소하고, 구매빈도가 높은 고객들에게는 보다 개인화된 서비스를 제공하였다는 점에서 의의가 있다. 마지막으로, 기존의 고객군별 예측기법이 불필요한 계산비용의 증가를 초래하는 한계를 극복하고 필요시에만 최적화된 군집화를 수행하여, 훨씬 적은 수의 데이터를 사용하면서도 *고객군별 예측기법*과 동일한 수준의 예측성능을 도출하였다는데 의의가 있다.

본 연구의 한계점은 다음과 같다. 첫째, 각 예측 기법에 소요되는 계산비용(computational cost)을 측정하는 요소로 학습모델 수립에 사용된 데이터의 수만을 고려하였다는 점이다. 본 연구에서는 모델 수립에 사용한 데이터 수를 기준으로 각 기법별 예측에 소요되는 비용을 비교하였으나, 그 외에도 예측모델의 수, 모델 수립 시 소요된 시간, 소요된 컴퓨팅 자원 등 비용과 관련된 다양한 요소들을 고려할 수 있다. 또한, 본 연구를 Neilsen의 음료수 구매 데이터에만 적용하여 실험한 점도 결과의 보편성 측면에서 한계로 지적될 수 있다. 향후 이를 보다 다양한 문제 해결에 적용하여 보편적인 결과를 도출할 필요가 있다.

참고문헌

김경재, 김병국, "데이터마이닝을 이용한 인터넷 쇼핑몰 상품추천 시스템 데이터마이닝을 이용한 인터넷 쇼핑몰 상품추천 시스템", *한국 지능정보시스템학회논문지*, 11권 1호(2005), 191~205.

Adomavicius, G. and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6(2005), 734~749.

Lawrence, R. D., G. S. V. Almasi, M. S. Kotlyar, Viveros, and S. S. Duri, "Personalization of supermarket product recommendations", *Data Mining and Knowledge Discovery*, Vol.5, No.1/2(2001), 11~32.

Lloyd, S. P., "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, Vol.28(1982), 129~136.

Park, Y.-J. and A. Tuzhilin, "The Long Tail of Recommender Systems and How to Leverage It", *ACM Conference On Recommender Systems*, Vol.3(2008), 11~18.

Peppers, D. and M. Rogers, The One-to-One Future. New York : Doubleday, 1993.

Schein, A., A. Popescul, L. Ungar, and D. Pennock, "Methods and Metrics for Cold-Start Recommendations", Proceeding of the 25th ACM SIGIR Conference, 2002.

Wendell, S. R., "Product Differentiation and Market Segmentation as Alternative Marketing Strategies", *Product Differentiation and Market Segmentation as Alternative Marketing Strategies Journal of marketing*, Vol.21, No.1(1956), 3~8.

Truong, K. Q., F. Ishikawa, and S. Honiden, "Improving Accuracy of Recommender System by Item Clustering", IEICE TRANSACTIONS on Information and Systems, E90-D-I(9), 2007.

Jiang, T. and A. Tuzhilin, "IEEE transactions on knowledge and data engineering", Vol.21 No.3(2009), 305~320.

Wedel, M. and W. Kamakura, "Market Segmentation : Conceptual and Methodological Foundations", 2nd ed. Kluwer Publishers, 2000.

Witten, I. H. and E. Frank, "Data Mining : Practical machine learning tools and techniques with Java implementations", Morgan Kaufmann, 2005.

Abstract

The Adaptive Personalization Method According to Users Purchasing Index : Application to Beverage Purchasing Predictions

Yoon-Joo Park*

TThis is a study of the personalization method that intelligently adapts the level of clustering considering purchasing index of a customer. In the e-biz era, many companies gather customers' demographic and transactional information such as age, gender, purchasing date and product category. They use this information to predict customer's preferences or purchasing patterns so that they can provide more customized services to their customers.

The previous *Customer-Segmentation* method provides customized services for each customer group. This method clusters a whole customer set into different groups based on their similarity and builds predictive models for the resulting groups. Thus, it can manage the number of predictive models and also provide more data for the customers who do not have enough data to build a good predictive model by using the data of other similar customers. However, this method often fails to provide highly personalized services to each customer, which is especially important to VIP customers. Furthermore, it clusters the customers who already have a considerable amount of data as well as the customers who only have small amount of data, which causes to increase computational cost unnecessarily without significant performance improvement.

The other conventional method called *1-to-1* method provides more customized services than the *Customer-Segmentation* method for each individual customer since the predictive model are built using only the data for the individual customer. This method not only provides highly personalized services but also builds a relatively simple and less costly model that satisfies with each customer. However, the *1-to-1* method has a limitation that it does not produce a good predictive model when a customer has only a few numbers of data. In other words, if a customer has insufficient number of transactional data then the performance rate of this method deteriorate.

In order to overcome the limitations of these two conventional methods, we suggested the new method called *Intelligent Customer Segmentation* method that provides adaptive personalized services according to the customer's purchasing index. The suggested method clusters customers according to their purchasing index, so that the prediction for the less purchasing customers are based on the data in more intensively clustered groups, and for the VIP customers, who already have a considerable

* Seoul National University of Science and Technology

amount of data, clustered to a much lesser extent or not clustered at all. The main idea of this method is that applying clustering technique when the number of transactional data of the target customer is less than the predefined criterion data size. In order to find this criterion number, we suggest the algorithm called *sliding window correlation analysis* in this study. The algorithm purposes to find the transactional data size that the performance of the *1-to-1* method is radically decreased due to the data sparity. After finding this criterion data size, we apply the conventional *1-to-1* method for the customers who have more data than the criterion and apply clustering technique who have less than this amount until they can use at least the predefined criterion amount of data for model building processes.

We apply the two conventional methods and the newly suggested method to Neilsen's beverage purchasing data to predict the purchasing amounts of the customers and the purchasing categories. We use two data mining techniques (Support Vector Machine and Linear Regression) and two types of performance measures (MAE and RMSE) in order to predict two dependent variables as aforementioned. The results show that the suggested *Intelligent Customer Segmentation* method can outperform the conventional *1-to-1* method in many cases and produces the same level of performances compare with the Customer-Segmentation method spending much less computational cost.

Key Words : Personalization, Customer Segmentation, Data Sparsity, Clustering, Intelligent Recommendation

저 자 소 개

박윤주

고려대학교 컴퓨터학과에서 학부 및 석사학위를 취득하였으며, 2006년 한국과학기술원에서 경영공학 박사학위를 취득하였다. 이 후, New York University의 Stern Business School에서 초빙연구원으로 근무하였고, 삼성생명 정보기획부서에서 과장으로 근무하였다. 2010년부터 현재까지 서울과학기술대학교 국제융합학부에서 조교수로 재직 중이다. 기존 연구는 Artificial Intelligence in Medicine, Expert Systems with Applications 등의 논문지에 게재되었으며, 2008 ACM conference on Recommender Systems에서 논문을 발표하였다. 이러한 기존 연구들은 한국 경영정보학회 및 한국지능정보시스템 학술대회에서 최우수논문상 및 우수논문상을 네 차례 수상한 바 있다. 주요 연구분야는 데이터마이닝을 이용한 질병 예측, 개인화 시스템, 그리고 온라인 매칭시스템 등이다.