지능정보연구 제16권 제4호 2010년 12월(pp.213~225)

Support Vector Regression에서 분리학습을 이용한 고객의 구매액 예측모형

홍태호
부산대학교 경영대학
(hongth@pusan.ac.kr)

김은미
부산대학교 일반대학원 박사과정
(keunmi100@pusan.ac.kr)

본 연구에서는 기업의 마케팅 프로모션에 따른 반응고객의 구매액 예측을 위한 방법을 제시하고 SVR의 효과적인 학습방법을 제시하였다. 프로모션에 의한 고객의 구매액을 기반으로 고객을 5등급으로 등급화하고 각 등급 내에서 SVR을 적용하여 고객의 구매액을 예측하였다. 본 연구에서 제안하는 예측된 고객의 등급 내에서 고객 구매액을 예측하는 분리데이터 학습법이 프로모션에 반응한 모든 고객을 대상으로 구매액을 예측하는 전체데이터 학습법보다 높은 예측성과를 보여주었다. 일반적으로 세분화된 고객집단을 하나의 집단으로 보고 동일한 마케팅 전략을 제시하나 본 연구를 통해 구매액에 따라 등급화 된 고객의 등급 내에서 다시 고객의 거래 구매액을 예측하여 동일한 집단 내에서도 차별화된 마케팅 전략을 제시할 수 있는 기반을 제시하였다. 즉 동일한 등급에서도 고객 구매액에 따라 고객의 우선순위를 정할 수 있으며, 이는 마케팅 담당자가 프로모션을 제시할 고객을 선정할 때 유용한 정보로 활용될 수 있다.

논문접수일 : 2010년 11월 24일 게재확정일 : 2010년 12월 08일 교신저자 : 김은미

1. 서 론

기업은 새로운 고객을 획득하고 기존고객을 유지하기 위해 다양한 마케팅 프로모션을 제공하고 있다. 정보기술의 발달로 고객 데이터베이스에 대한 접근이 용이해 짐에 따라 기업은 고객정보, 거래정보 등의 고객데이터를 활용하여 마케팅 프로모션을 제공할 목표고객을 선정하고자 한다. 모든 고객에게 무차별적으로 제공되는 프로모션은 불필요한 비용의 지출은 물론이고 고객과의 관계도 악화시킬 수 있기 때문에 프로모션을 제공할 목표고객의 선정은 중요하다(Cönül et al., 2000). 특히 다이렉트 마케팅에서 목표고객의 선정 및 예측에 대한 중요성이 강조되고 있으며 이를 위해 고객반응 예측모형을 활용하고 있다(Kim and Street, 2004; Suh et al., 1999). 다이렉트 마케팅의 고객 데이터베이스를 통해 잠재고객을 구별하고 정확한 목표고객의 선정은 프로모션을 통한 판매증진을 기대할 수 있다. 고객반응 예측모형을 통해 프로모션에 대한 고객의 반응여부에 따라 고객을 반응고객과 비반응고객으로 구별하여 프로모션을 제공할 고객집단과 프로모션을 제공하지 않을 고객집단을 구별한다. 프로모션에 반응할 고객집단으로 분류된 고객들은 일반적으로 하나의 반응고

* 본 연구는 2010학년도 부산대학교 특성화분야 육성사업의 연구비를 지원 받아 연구되었음.

객 집단으로 동일한 프로모션을 제공받게 되나 반응고객 집단 내에서도 고객의 특성에 따라 고객은 다양하게 분류될 수 있다. 또한 반응고객으로 분류된 모든 고객들에게 프로모션을 제공할 수 없을 경우 반응고객 집단 내에서도 프로모션을 제공할 목표고객의 선정이 이루어져야 한다.

이를 위해 반응고객 집단으로 분류된 고객의 거래 구매액을 적용할 수 있다. 고객의 구매액은 소액부터 고액까지 다양하게 분포하며 구매액이 고액인 고객이 소액인 고객보다 수익성 있는 고객으로 분류되어 고액의 구매액을 나타내는 고객에게 보다 적극적인 프로모션을 제공할 수 있다. 고객의 구매액 예측은 고객의 반응여부와 함께 프로모션을 제공할 목표고객의 선정을 위해 중요한 요소이나 반응고객의 구매액 예측은 구매액에 대한 범위가 매우 넓게 나타나기 때문에 이에 대한 예측오류의 범위 또한 넓게 나타나는 문제점을 안고 있어 구매액 예측에 관한 연구는 많지 않았다. 고객의 구매액은 일반적으로 두 단계를 통해 예측할 수 있다. 먼저 기업이 제공하는 프로모션에 대한 고객의 반응여부를 예측한 다음, 반응한다고 예측된 고객을 대상으로 구매액 예측이 이루어진다(Kim et al., 2008). Kim et al.(2008)은 프로모션에 대한 반응고객을 대상으로 SVR(Support Vector Regression)을 적용하여 고객의 구매액을 예측하였으며 Wang et al.(2005)는 비영리조직의 고객데이터를 이용하여 고객의 기부액을 예측하고자 하였다. 이는 프로모션에 대한 고객의 반응여부를 기반으로 한 이진분류를 통해 반응한 고객을 대상으로 구매액 예측이 이루어졌다.

본 연구에서는 프로모션에 대한 고객의 반응여부를 기반으로 이루어지는 고객의 구매액 예측을 이진분류가 아닌 다분류를 통한 고객등급별 구매액 예측방법으로 분리데이터 학습법(LMS; Learning Method using Separated data for classified purchasing customers)제안한다. 분리데이터 학습법은 고객을 구매 거래액에 따라 5개의 등급으로 등급화 하여 고객의 구매액 범위를 좁히고 각 등급별로 구매액 예측모형을 개발하였다. 프로모션에 반응하는 모든 고객을 대상으로 구매액을 예측하는 것보다 고객의 등급을 적용한 분리데이터 학습법은 구매액 예측오류의 범위를 좁힐 수 있으며 동일한 등급 내에서 구매액에 따라 고객의 우선순위를 정할 수 있기 때문에 마케팅 담당자가 프로모션을 제시할 고객을 선정할 때 등급별 고객의 구매액 예측을 통해 유용한 정보를 제공받을 수 있다.

본 연구는 다음과 같이 구성된다. 제 2장에서는 고객반응 예측모형과 SVR에 대한 기존연구를 제시하였으며 제 3장에서는 연구 프레임워크를 제시하였다. 제 4장에서는 등급별 구매액 예측을 위한 실험결과 및 분석을 제시하였고 제 5장에서는 결론 및 향후연구방향을 제시하였다.

2. 이론적 배경

2.1 고객반응 예측

기업들은 고객의 구매를 촉진하기 위해 다양한 프로모션을 펼치고 있다. 프로모션은 고객들의 제품구매를 유도하거나 일정기간 동안 반복구매가 더 많이 일어나도록 하기 위한 모든 방법으로 기업의 이윤을 증대시키며 고객과의 관계를 보다 강화시킬 수 있다(고용식, 2005). 기업의 프로모션에 대한 고객의 반응은 고객이 기업의 마케팅 활동에 노출되면서 일어나는 고객의 선호도, 기대, 허락, 태도 등이 있을 수 있으며 고객의 반응이 높을수록 기업은 더 많은 이익을 기대할 수 있을 것이다.

또한 프로모션에 대한 고객의 반응은 구매행위로 이어질 가능성이 매우 크다고 할 수 있기 때문에 프로모션에 대한 고객의 긍정적인 반응이 장기적으로 지속될 수 있도록 해야 한다.

고객반응 예측모형은 다이렉트 마케팅에서 프로모션을 제공할 목표고객의 선정에 중요한 영향을 미친다(Prinzie and Van der Poel, 2005). 고객반응 예측모형을 통해 기업이 제시하는 마케팅 프로모션에 대한 고객의 반응여부를 예측하고 반응할 확률이 높은 고객을 선별하며 얼마나 많은 고객들에게 프로모션을 제공할지 결정하도록 도와준다(Kim et al., 2008). 모든 고객에게 프로모션을 제공하기보다 프로모션에 대한 고객의 반응예측을 통해 목표고객을 선정하면 필요이상으로 소요되는 마케팅 비용을 줄일 수 있어 기업의 비용절감과 이윤의 극대화를 기대할 수 있다(Baesens et al., 2002; Cönül et al., 2000; Suh et al., 1999). 고객의 프로모션에 대한 반응예측모형을 위해 고객정보, 구매정보, 상품정보 등이 사용되어지며(Kim and Street, 2004), 기업은 프로모션에 대한 고객의 반응이 높은 집단에는 프로모션을 통해 고객의 구매를 촉진 시킬 수 있으며 프로모션에 대한 고객의 반응이 낮은 집단에는 굳이 비싼 비용을 들여서 프로모션을 할 필요가 없다.

프로모션에 대한 고객의 반응을 예측하기 위해 전통적으로 로지스틱 회귀모형이나 판별분석 등과 같은 통계적 방법이 적용되었으나 방대하고 비선형적이며 복잡한 데이터의 분석을 위해 연관규칙, 의사결정나무, 인공신경망, SVM 등과 같은 다양한 데이터마이닝 기법들이 적용되고 있다(김진화 등, 2008; Shin and Cho, 2006; Zahavi and Levin, 1997). 김진화 등(2008)은 고객의 구매의도를 예측하기 위해 구매데이터를 사용하여 인공신경망, 로지스틱 회귀분석, 의사결정나무, 베이지안 망, SVM 등 다양한 데이터마이닝 기법을 적용하였으며 SVM 모형이 가장 우수한 성과를 보였다고 보고하였다. 안현철 등(2005)은 인터넷 쇼핑몰의 구매데이터를 사용하여 고객구매예측을 위해 로지스틱 회귀분석, 인공신경망, SVM 등을 적용하였고, Baseens et al.(2002)은 베이지안 신경망을 적용하여 통신판매회사의 고객에 대한 구매를 예측하였다.

2.2 Support Vector Regression

Support Vector Machines(SVM)은 Vapnik(1995)이 제시한 새로운 종류의 신경망 알고리즘으로 통계적 학습이론에 기반하여 주로 분류문제에 적용되어 우수한 예측성과를 보여 왔다. SVM은 구조적 위험 최소화(SRM; Structural Risk Minimization)에 기반하기 때문에 경험적 위험 최소화(ERM; Empirical Risk Minimization)에 기반한 신경망(neural networks)보다 일반화하기가 더 용이하고 우수한 성과를 보여 왔다(Tay and Cao, 2001). 이러한 이유 때문에 SVM은 기업신용평가(Huang et al., 2004), 프로모션 고객의 반응예측(Cho and Shin, 2006), 이탈고객의 예측(Coussement and Van den Poel, 2008) 등 다양한 분야에서 성공적으로 적용되어 왔다. SVM은 분류문제의 예측에 적용된 반면에, SVM의 회귀모형에 ϵ-무감도 손실함수(ϵ-insensitive loss function)을 도입하여 Support Vector Regression(SVR)이 회귀문제의 영역까지 확장되어 왔다(Vapnik et al., 1997). 다음은 SVR에 대한 설명이며 자세한 내용은 Vapnik et al.(1997)을 참고한다.

SVR을 설명하기 위해 일반적인 회귀모형 문제를 예를 들어 설명하도록 한다. 데이터 셋 $G=(x_i, q_i)_i^n$을 살펴보도록 하고, 여기서 x_i는 모형입력을 위한 입력 벡터값이고, q_i는 출력값이다. 일반적

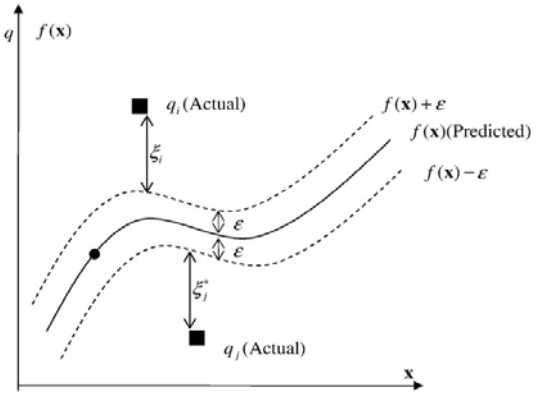
회귀함수는 $q_i = f(x_i) + \delta$과 같이 나타낼 수 있으며, 여기서 δ는 $N(0, \sigma^2)$ 분포를 따르는 랜덤 오류이다. SVR에서는 비선형 회귀문제를 풀기 위해서 고차원(high dimension)의 형상공간(feature space)으로 사상(mapping)시킨다. 즉, 비선형회귀를 이용한 원래의 최적화문제는 형상공간에서 선형함수를 탐색하는 문제로 재정의 된다.

$$f(x) = (v \cdot \Phi(x)) + b \qquad (1)$$

ϵ-무감도 손실함수 L_ϵ는 일반적으로 SVR에 사용되는 비용함수로 다음과 같다.

$$L_\epsilon(f(x), q) = \begin{cases} |f(x) - q| - \epsilon & if\ |f(x) - q| \geq \epsilon \\ 0 & o.w \end{cases} \qquad (2)$$

여기서 ϵ는 회귀함수 $f(x)$의 주변에 위치한 튜브의 반지름을 나타내는 정밀모수(precision parameter)이다(<그림 1> 참조). SVR의 선형 추정함수 $f(x) = (v \cdot \phi(x)) + b$의 가중벡터 (v)와 상수 (b)는 다음의 정규화 된 위험함수에 의해 추정될 수 있다.



<그림 1> ε-무감도 손실함수를 사용하는 SVR의 도식적 표현(Lu et al., 2009)

$$RC = C\frac{1}{n}\sum_{i=1}^{n} L_\epsilon f(x_i), q_i) + \frac{1}{2}|w|^2 \qquad (3)$$

$\frac{1}{2}|w|^2$는 회귀모형의 복잡성과 정확성의 균형을 조정하는 정규화 항이며, C는 경험적 위험과 정규화 항의 균형을 맞추는데 사용되는 정규화된 상수이다. 또한, C와 ϵ는 사용자가 정하여야하는 모수가 된다.

여유변수를 이용하면 식 (3)은 아래의 제약식을 갖는 식 (4)로 변환된다.

$$Min: R_{reg}(f) = \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad (4)$$

s.t.

$$q_i - (w \cdot \Phi(x_i)) - b \leq \epsilon + \xi_i$$
$$(w \cdot \Phi(x_i)) + b - q_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0, \quad for\ i = 1, \cdots, n$$

라그랑지 승수와 Karush-Kuhn-Tucker조건을 식 (4)에 적용하면, 최종적으로 SVR 기반 회귀함수의 일반적 형태는 식 (5)와 같다.

$$f(x, v) = f(x, \alpha, \alpha^*) \qquad (5)$$
$$= \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b$$

여기서 $K(x_i, x_j)$는 커널함수로 RBF(Radial Basis Function)을 주로 사용한다. 즉, $k(x_i, x_j) = \exp\left(\frac{|x_i - x_j|^2}{2}\sigma^2\right)$가 되며, σ는 RBF의 넓이를 나타낸다.

3. 연구프레임워크

본 연구에서는 고객의 구매액에 따른 등급을 적

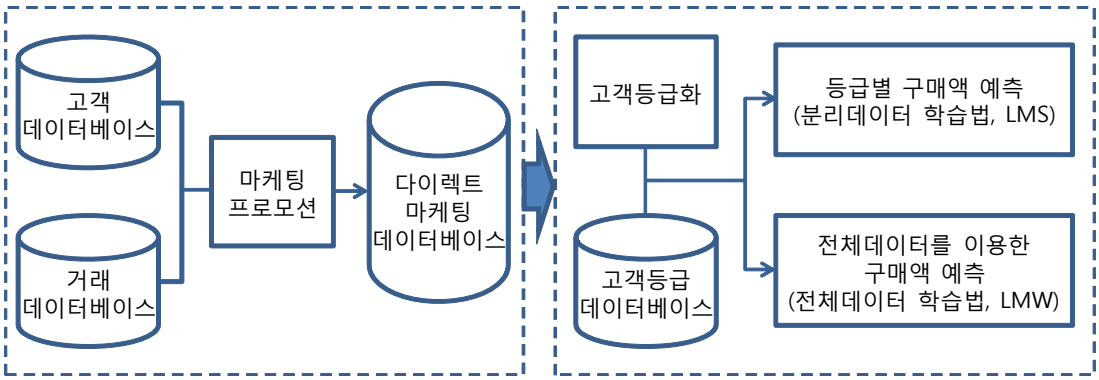
용한 고객등급별 구매액 예측을 위해 <그림 2>와 같이 연구 프레임워크를 제시하였다. 기업은 고객정보와 거래정보를 기반으로 프로모션을 제공하고 프로모션에 대한 고객의 반응정보를 다이렉트 마케팅 데이터베이스에 포함시킨다. 프로모션에 대한 반응정보가 포함된 다이렉트 마케팅 데이터베이스에서 고객을 등급화하기 위해 프로모션 이후 고객의 구매액을 활용한다. 고객의 구매액을 기반으로 고객을 등급화하고 각 등급별로 고객의 구매액 예측모형을 개발하여 동일한 등급 내에서도 프로모션을 제공하기 위한 목표고객의 선정을 효과적으로 할 수 있도록 한다.

고객의 구매액 예측을 위해 Vapnik(1995)이 제안한 SVR을 적용하였으며 전체데이터 학습법(LMW; Learning Method using Whole data for purchasing customers)과 분리데이터 학습법(LMS; Learning Method using Separated data for classified purchasing customers)을 비교분석 한다. 전체데이터 학습법은 프로모션에 대한 고객의 반응여부에 따라 반응한 고객집단 전체를 대상으로 학습하는 이진분류 기반의 학습법으로 반응고객의 구매액을 예측한다. 분리데이터 학습법은 구매액에 대한 고객의 등급을 적용한 다분류 기반의 학습법으로 동일한 등급 내에서의 학습을 통해 고객의 구매액을 예측하여 고객의 구매액 범위를 좁히고 이에 대한 오류의 범위도 좁힐 수 있다. 일반적으로 세분화된 고객집단을 하나의 집단으로 보고 동일한 마케팅 전략을 제시하나 본 연구에서 제안하는 분리데이터 학습법인 LMS를 통해 구매액에 따른 세분화 된 집단 내에서도 차별화된 마케팅 전략을 제시할 수 있는 방안을 제시한다. 세분화된 집단 내에서 고객의 구매액에 따른 우선순위를 정할 수 있으며 이는 마케팅 담당자가 프로모션을 제시할 목표고객을 선정할 때 유용한 정보를 제공할 수 있다.

4. 실험 및 결과분석

4.1 데이터

본 연구에서는 다이렉트 마케팅 데이터로 Direct Marketing Association의 DMEF04 데이터셋(http://www.directworks.org)을 사용하였다. DMEF04 데이터셋은 실제 마케팅 활동을 통해 얻어진 101,532명의 고객들에 대한 데이터로 주문수, 주문금액, 주문항목 등에 대한 91개의 변수로 이루어져



<그림 2> 고객등급별 구매액을 위한 연구 프레임워크

있다. 기업에서 정기적으로 카탈로그를 보낸 후 프로모션에 의한 구매여부가 제시되어 있으며 홍태호와 박지영(2010)에서 프로모션 이후 고객의 구매액을 기준으로 등급화한 고객등급 자료를 사용하였다. 홍태호와 박지영(2010)은 전체 101,532명의 고객 중 프로모션 시점을 기준으로 2년 이내에 거래가 있는 41,924명의 고객을 대상으로 고객을 등급화 하였다. 프로모션이후 고객 구매액을 기반으로 전체 구성비를 반영하여 고객을 구매액이 200달러 이상인 고객 군(1등급), 100달러 이상 200달러 미만인 고객 군(2등급), 30달러 이상 100달러 미만인 고객 군(3등급), 1달러 이상 30달러 미만인 고객 군(4등급), 고객 반응이 나타나지 않은 고객 군(5등급)으로 하여 고객을 5등급으로 등급화 하였으며 고객등급은 <표 1>과 같다.

<표 1> 구매액 기반의 고객 등급화

고객등급	구매액($)	데이터수	분포
1등급	200이상	427	1%
2등급	100~199	745	2%
3등급	30~99	3,037	7%
4등급	1~29	3,159	8%
5등급	0	34,556	82%
합계		41,924	100%

4.2 실험설계

고객의 구매액 예측을 위해 91개의 입력변수 중 Kim et al.(2008)과 Malthouse(2001) 등의 연구에서 사용된 16개의 독립변수를 사용하여 구매액 예측모형을 개발하였다. <표 2>에 제시되어 있는 독립변수를 사용하여 프로모션에 반응한 고객집단으로 분류된 고객을 대상으로 구매액을 예측하였다. 구매액에 의한 5등급의 고객들 중 프로모션에 반응을 보이지 않은 5등급의 고객은 제외하고 나머지 등급의 고객을 대상으로 구매액을 예측하였다. 전체데이터 학습법인 LMW와 구매고객의 등급을 적용하여 동일한 등급 내에서 학습하여 고객의 구매액을 예측한 분리데이터 학습법인 LMS에서 SVR을 적용하였다. SVR을 위해 각 등급별로 데이터를 랜덤하게 4 : 1로 나누어 학습용과 검증용으로 사용하였다.

<표 2> 구매액 예측에 사용된 변수

변수명		설 명	
독립변수	Purseas	구매가 일어난 시즌 수	
	Falord	현재까지 가을 시즌 주문 수	
	Ordtyr	올해 주문 수	
	Puryear	구매가 일어난 해의 수	
	Sprord	현재까지 봄 시즌 주문 수	
	Recency	프로모션 기준일로부터 주문일 수	
	Tran38	1/recency	
	Tran52	90 ≤ recency < 180	1
		그 외	0
	Tran53	180 ≤ recency < 270	1
		그 외	0
	Tran54	270 ≤ recency < 366	1
		그 외	0
	Tran55	366 ≤ recency < 730	1
		그 외	0
	Comb2	$\sum_{i=1}^{14}$ 올해 구매한 제품그룹	
	Tran25	1/(1+최근 시즌 아이템)	
	Tran42	log(1+올해 주문 수 × 현재까지 가을시즌 주문 수)	
	Tran44	√현재까지 주문 수 × 현재까지 봄 시즌 주문 수	
	Tran46	$\sqrt{comb^2}$	

SVR의 커널함수로 RBF 커널을 사용하였으며 모수 C는 0.1, 1, 50로 하였고 σ는 1, 0.1, 0.05로 설정하였으며 ϵ-무감도 손실함수(ϵ-insensive loss function)의 ϵ값은 0.05, 0.1, 0.5로 설정하여 공개소

프트웨어인 LIBSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm)을 사용하였다.

4.3 제안모형의 결과 및 분석

SVR을 이용한 고객 구매액의 예측성과는 예측값과 실제값의 예측오차를 통해 MAE(Mean Absolute Error)와 MAPE (Mean Absolute Percent Error)로 나타내었다. <표 3>은 반응고객 전체를 학습하는 LMW를 이용한 고객등급별 구매액 예측성과로 MAE와 MAPE가 대체적으로 높게 나타난다. LMW에서는 반응한 모든 고객을 대상으로 학습했기 때문에 고객의 거래 구매액 범위도 넓으며 이에 대한 예측오차도 크게 나타난다. LMW의 전체 성과를 보면 MAPE가 0.670으로 나타나고 각 등급별로 예측성과를 살펴보면 다른 등급에 비해 3등급에서의 MAPE가 0.327로 예측성과가 가장 좋게 나타나며 4등급에서의 MAPE가 1.023으로 예측성과가 가장 낮게 나타난다. 고객의 등급을 반영한 LMS의 예측성과는 <표 4>와 같다. 고객의 등급을 적용하여 고객의 구매액에 대한 범위를 좁힌 결과 LMS을 이용한 고객등급별 구매액의 예측성과는 LMW보다 높은 예측성과를 보인다. LMS을 이용한 각 등급의 MAE와 MAPE는 모든 등급에서 전체데이터 학습법인 LMW의 MAE와 MAPE보다 향상된 예측성과를 보인다. LMW에서 가장 높은 예측성과를 보였던 3등급에서의 예측성과가 LMS에서도 가장 높은 예측성과를 보여주며 MAPE가 0.275로 나타나LMW의 0.327에서보다 향상되었다. 또한 LMW에서 예측성과가 가장 낮게 나타났던 4등급의 예측성과는 LWS에서 0.386으로 0.637만큼 향상된 결과를 보여주었다<그림 3>. LMS에서는 구매액에 따른 고객의 각 등급별로 구매액을 예측하여 등급별로 예측성과가 비슷하게 나타났

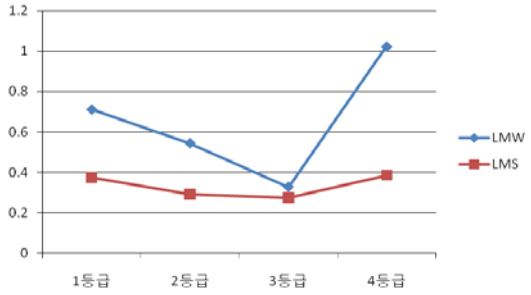
<표 3> LMW를 이용한 고객 등급별 구매액 예측성과

고객 등급	학습용				검증용			
	실제값 (평균)	예측값 (평균)	MAE	MAPE	실제값 (평균)	예측값 (평균)	MAE	MAPE
전체	48.371	34.278	26.624	0.654	48.607	34.462	27.167	0.670
1등급	180.569	47.103	133.478	0.686	188.349	45.275	143.074	0.711
2등급	101.666	40.971	61.238	0.555	99.827	41.150	59.414	0.544
3등급	49.113	35.917	17.560	0.323	49.316	36.282	17.736	0.327
4등급	17.196	29.387	12.716	0.992	17.056	29.683	13.034	1.023

<표 4> LMS를 이용한 고객 등급별 구매액 예측성과

고객 등급	학습용				검증용			
	실제값 (평균)	예측값 (평균)	MAE	MAPE	실제값 (평균)	예측값 (평균)	MAE	MAPE
1등급	180.569	151.407	57.964	0.284	188.349	152.725	72.841	0.374
2등급	101.666	96.348	25.900	0.273	99.827	96.531	26.316	0.291
3등급	49.113	44.857	13.002	0.257	49.316	45.044	13.786	0.275
4등급	17.196	17.087	4.807	0.358	17.056	17.117	4.977	0.386

으나 LMW에서는 각 등급별로 예측성과의 차이가 크게 나타났다. 고객의 각 등급별로 LMW와 LMS의 구매액 예측성과가 통계적으로 유의한지 대응표본 t-test를 통해 살펴 본 결과 <표 5>와 같다. 각 등급에서 LMW와 LMS의 차이를 통계적으로 검증한 결과 모든 등급에서 유의하게 나타났다.



<그림 3> 등급별 LMW와 LMS의 MAPE

<표 5> 대응표본 t-test 결과

대응표본		t-통계량	MAE	MAPE
1등급	LMW & LMS	평균차이 유의확률 N	70.233 0.000** 85	0.338 0.000** 85
2등급	LMW & LMS	평균차이 유의확률 N	33.098 0.000** 149	0.253 0.000** 149
3등급	LMW & LMS	평균차이 유의확률 N	3.951 0.000** 607	0.053 0.000** 607
4등급	LMW & LMS	평균차이 유의확률 N	8.057 0.000** 632	0.637 0.000** 632

주) ** : $p < 0.01$, * : $p < 0.05$.

5. 결론 및 한계점

본 연구에서는 고객의 구매액을 예측하기위해 SVR을 적용하였다. 구매액 예측모형의 성과를 높이기 위해 고객의 등급을 분류하고 본 연구에서 제안된 분리데이터 학습을 적용한 결과 제안한 분리데이터 학습법이 통계적으로 유의하게 우수한 성과를 보였다. 기존의 고객반응예측 연구들은 고객의 구매 여부에 초점을 둔 이진분류 모형의 개발이 주였다. 하지만, 고객의 구매액은 구매여부와 함께 프로모션 마케팅에서 매우 중요한 요소이며 이러한 구매액의 예측과 관련된 연구는 많지 않았다. 그 이유는 구매액의 범위가 너무 넓기 때문에 예측의 오류 또한 넓어지는 문제점을 안고 있기 때문이다. 본 연구에서는 이러한 문제점의 해결방안으로 고객을 5등급으로 분류하는 모형을 통해 고객의 구매액에 대한 구매 범위를 좁혔다. 고객 등급화 모형은 일반적으로 많이 사용되는 RFM 모형에 근거하였으며, 각 등급별로 구매액을 예측하는 4개의 SVR 모형을 개발하였다. SVR과 같은 데이터를 통해서 학습을 하는 데이터마이닝 기법은 학습표본이 동질적인(homogenous) 표본에서의 학습이 용이하다는 가정에 근거하여 분리데이터 학습법을 제시하였다. 실증분석 결과 분리데이터 학습방법은 기존의 모든 샘플에서 하나의 예측모형을 개발할 때보다 매우 효과적임을 통계적으로 밝혀냈다.

본 연구에서는 고객의 구매액을 예측하기 위해 SVR을 적용하였으나 다양한 파라미터를 적용해 보지 못했다. 예측성과를 높이기 위해 다양한 파라미터를 적용하여 예측성과를 더 향상 시킬 수 있도록 해야 할 것이며 고객의 구매거래액이 아닌 다른 특성을 통해서도 동일 집단 내에서 차별화할 수 있는 방안을 제시할 수 있도록 해야 할 것이다.

참고문헌

고용식, “세일즈 프로모션전략으로서의 VMD에 관한 연구”, 한국마케팅과학회 2005 춘계학술대회 발표논문집, (2005), 321~339.

김진화, 남기찬, 이상종, “Support Vector Machine 기법을 이용한 고객의 구매의도 예측”, *Information Systems Review*, 10권 2호(2008), 137~158.

안현철, 김경재, 한인구, “Support Vector Machine을 이용한 고객구매예측모형”, *한국지능정보시스템학회논문지*, 11권 3호(2005), 69~81.

홍태호, 박지영, “RCMDE를 적용한 프로모션에 따른 고객등급예측”, 한국인터넷전자상거래학회, 한국정보시스템학회 2010년 춘계공동 학술대회논문집, (2010), 155~168.

Baesens, B., S. Viaene, D. Van den Poel, J. Vanthienen and G. Dedene, “Bayesian neural network learning for repeat purchase modelling in direct marketing”, *European Journal of Operational Research*, Vol.138, No.1 (2002), 191~211.

Cho, S. and H. Shin, “Response modeling with support vector machines”, *Expert Systems with Applications*, Vol.30, No.4(2006), 746~760.

Cönül, F. F., B. D. Kim and M. Shi, “Mailing smarter to catalog customer”, *Journal of Interactive Marketing*, Vol.14, No.2(2000), 2~16.

Coussement, K. and D. Van den Poel, “Churn prediction in subscription services : An application of support vector machines while comparing two parameter-selection techniques”, *Expert Systems with Applications*, Vol.34, No.1(2008), 313~327.

Huang, Z., S. Chen, C. Hsu, W. Chen and S. Wu, “Credit rating analysis with support vector machines and neural networks : a market comparative study”, *Decision Support Systems*, Vol.37, No.4(2004), 543~558.

Kim, D., H. Lee and S. Cho, “Response Modeling with Support Vector Regression”, *Expert Systems with Applications*, Vol.34, No.2(2008), 1102~1108.

Kim, Y. S. and W. N. Street, “An intelligent system for customer targeting a data mining approach”, *Decision Support Systems*, Vol.37, No.2(2004), 215~228.

Lu, C., T. Lee, C. Chiu, “Financial time series forecasting using independent component analysis and support vector regression”, *Decision Support Systems*, Vol.47, No.2(2009), 15~125.

Malthouse, E. C., “Assessing the performance of direct marketing scoring models”, *Journal of Interactive Marketing*, Vol.15, No.1(2001), 49~62.

Prinzie, A. and D. Van den Poel, “Constrained optimization of data-mining problems to improve model performance : A direct-marketing application”, *Expert Systems with Applications*, Vol.29, No.3(2005), 630~640.

Shin, H. and S. Cho, “Response Modeling with Support Vector Machine”, *Expert Systems with Applications*, Vol. 30, No.4(2006), 746~760.

Suh, E. H., K. C. Noh and C. K. Suh, “Customer list segmentation using the combined response model”, *Expert Systems with Applications*, Vol.17, No.2(1999), 89~97.

Tay, F. E. H. and L. Cao, “Application of support vector machines in financial time series forecasting”, *Omega*, Vol.29, No.4(2001), 497~505.

Vapnik, S. Golowich and A. Smola, “Support vector method for function approximation regression estimation, and signal processing”, In Mozer, M., Jordan, M., Petsche, T. editors, Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, (1999), 281~287.

Vapnik, The Nature of Statistical Learning The-

ory, Springer, N.Y. 1995.

Wang, K., S. Zhou, Q. Yang and J. M. S. Yeung, "Mining customer value : From association rules to direct marketing", *Data Mining and Knowledge Discovery*, Vol.11(2005), 57~79.

Zahavi, J. and N. Levin, "Applying neural computing to target marketing", *Journal of Direct Marketing*, Vol.11, No.4(1997), 76~93.

Abstract

The Prediction of Purchase Amount of Customers Using Support Vector Regression with Separated Learning Method

Taeho Hong* · Eunmi Kim**

Data mining has empowered the managers who are charge of the tasks in their company to present personalized and differentiated marketing programs to their customers with the rapid growth of information technology. Most studies on customers' response have focused on predicting whether they would respond or not for their marketing promotion as marketing managers have been eager to identify who would respond to their marketing promotion. So many studies utilizing data mining have tried to resolve the binary decision problems such as bankruptcy prediction, network intrusion detection, and fraud detection in credit card usages. The prediction of customer's response has been studied with similar methods mentioned above because the prediction of customer's response is a kind of dichotomous decision problem. In addition, a number of competitive data mining techniques such as neural networks, SVM(support vector machine), decision trees, logit, and genetic algorithms have been applied to the prediction of customer's response for marketing promotion. The marketing managers also have tried to classify their customers with quantitative measures such as recency, frequency, and monetary acquired from their transaction database. The measures mean that their customers came to purchase in recent or old days, how frequent in a period, and how much they spent once. Using segmented customers we proposed an approach that could enable to differentiate customers in the same rating among the segmented customers.

Our approach employed support vector regression to forecast the purchase amount of customers for each customer rating. Our study used the sample that included 41,924 customers extracted from DMEF04 Data Set, who purchased at least once in the last two years. We classified customers from first rating to fifth rating based on the purchase amount after giving a marketing promotion. Here, we divided customers into first rating who has a large amount of purchase and fifth rating who are non-respondents for the promotion. Our proposed model forecasted the purchase amount of the customers in the same rating and the marketing managers could make a differentiated and

* Associate Professor, School of Business, Pusan National University
** Ph.D. Candidate. School of Business, Pusan National University

personalized marketing program for each customer even though they were belong to the same rating. In addition, we proposed more efficient learning method by separating the learning samples. We employed two learning methods to compare the performance of proposed learning method with general learning method for SVRs. LMW (Learning Method using Whole data for purchasing customers) is a general learning method for forecasting the purchase amount of customers. And we proposed a method, LMS (Learning Method using Separated data for classification purchasing customers), that makes four different SVR models for each class of customers. To evaluate the performance of models, we calculated MAE (Mean Absolute Error) and MAPE (Mean Absolute Percent Error) for each model to predict the purchase amount of customers. In LMW, the overall performance was 0.670 MAPE and the best performance showed 0.327 MAPE. Generally, the performances of the proposed LMS model were analyzed as more superior compared to the performance of the LMW model. In LMS, we found that the best performance was 0.275 MAPE. The performance of LMS was higher than LMW in each class of customers. After comparing the performance of our proposed method LMS to LMW, our proposed model had more significant performance for forecasting the purchase amount of customers in each class. In addition, our approach will be useful for marketing managers when they need to customers for their promotion. Even if customers were belonging to same class, marketing managers could offer customers a differentiated and personalized marketing promotion.

Key Words : Customer response model; Customer rating; Support Vector Regression; The prediction of purchase amount of customers

저 자 소 개

홍태호

현재 부산대학교 경영학부 부교수로 재직하고 있다. KAIST에서 산업공학사를 취득하였고 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 딜로이트 컨설팅에서 컨설턴트로 재직했으며, 주요 관심분야는 데이터마이닝, CRM, Business Intelligence 그리고 Social Networks 등이다. Expert Systems, Expert Systems with Applications, Asia Pacific Journal of Information Systems, 그리고 정보시스템연구, 지능정보시스템연구, Information Systems Review 등을 비롯한 국내외 학술지에 에 논문을 발표하였다.

김은미

부산대학교 경영학과에서 석사학위를 취득하고, 현재 동대학원에서 박사과정 중에 있다. 주요 연구분야는 데이터마이닝, 고객관계관리, 지식경영 등이며 인터넷전자상거래연구, 정보시스템연구, Information Systems Review 등에 논문을 게재하였다.