Journal of The Korea Society of Computer and Information
Vol. 19, No. 9, September 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.9.125

준지도 지지 벡터 회귀 모델을 이용한 반응 모델링

김 동 일*

Response Modeling with Semi-Supervised Support Vector Regression

Dong-Il Kim*

요 약

본 논문에서는 준지도 지지 벡터 회귀 모델(semi-supervised support vector regression)을 이용한 반응 모델링(response modeling)을 제안한다. 반응 모델링의 성능 및 수익성을 높이기 위해, 고객 데이터 셋의 대부분을 차지하는 레이블이 존재하지 않는 데이터를 기존 레이블이 존재하는 데이터와 함께 학습에 이용한다. 제안하는 알고리즘은 학습 복잡도를 낮은 수준으로 유지하기 위해 일괄 학습(batch learning) 방식을 사용한다. 레이블 없는 데이터의 레이블 추정에서 불확실성(uncertainty)을 고려하기 위해, 분포추정(distribution estimation)을 하여 레이블이 존재할 수 있는 영역을 정의한다. 그리고 추정된 레이블 영역으로부터 오버샘플링(oversampling)을 통해 각 레이블이 없는 데이터에 대한 레이블을 복수 개 추출하여 학습 데이터 셋을 구성한다. 이 때, 불확실성의 정도에 따라 샘플링 비율을 다르게 함으로써, 불확실한 영역에 대해 더 많은 정보를 발생시킨다. 마지막으로 지능적 학습 데이터 선택 기법을 적용하여 학습 복잡도를 최종적으로 감소시킨다. 제안된 반응 모델링의 성능 평가를 위해, 실제 마케팅 데이터 셋에 대해 다양한 레이블 데이터 비율로 실험을 진행하였다. 실험 결과 제안된 준지도 지지 벡터 회귀 모델을 이용한 반응 모델이 기존 모델에 비해 더 높은 정확도 및 수익을 가질 수 있다는 점을 확인하였다.

▸ Keywords : 반응 모델링; 준지도 학습; 지지 벡터 회귀 모델; 고객관계관리; 데이터마이닝

Abstract

In this paper, I propose a response modeling with a Semi-Supervised Support Vector Regression (SS-SVR) algorithm. In order to increase the accuracy and profit of response modeling, unlabeled data in the customer dataset are used with the labeled data during training. The proposed SS-SVR algorithm is designed to be a batch learning to reduce the training complexity. The label distributions of unlabeled data are estimated in order to consider the uncertainty of labeling. Then,

•제1저자 : 김동일 •교신저자 : 김동일
•투고일 : 2014. 7. 29, 심사일 : 2014. 8. 11, 게재확정일 : 2014. 8. 19.
* 삼성전자 시스템기술팀(System Engineering Team, Samsung Electronics, Co. Ltd.)

multiple training data are generated from the unlabeled data and their estimated label distributions with oversampling to construct the training dataset with the labeled data. Finally, a data selection algorithm, Expected Margin based Pattern Selection (EMPS), is employed to reduce the training complexity. The experimental results conducted on a real-world marketing dataset showed that the proposed response modeling method trained efficiently, and improved the accuracy and the expected profit.

▸ Keywords : Response Modeling, Semi-Supervised Learning, Support Vector Regression; Customer Relationship Management; Data Mining

I. 서 론

최근 빅데이터 분석의 적용 분야는 마케팅, 제조, 의료, IT 등 매우 다양해지고 있다. 특히 마케팅에서 빅데이터 분석을 통한 수익 창출에 대한 연구가 많았는데, 반응 모델링(response modeling)은 그 대표적인 사례이다. 어떤 제품이나 서비스에 대한 구매를 유도하는 마케팅 캠페인을 할 때, 일반적으로 마케팅 캠페인을 받고 구매로 이어지는, 즉 마케팅에 반응하는 고객은 10% 미만으로 알려져 있다. 하지만 어떤 고객이 마케팅에 반응할지 모르기 때문에, 다수의 고객을 무작위로 골라 마케팅 캠페인이 진행되고, 이는 높은 마케팅 비용 지출로 이어진다. 반면, 만약 마케팅 캠페인 전에 반응 고객과 구매 금액을 미리 예측할 수 있다면, 이들만을 목표로 삼아 보다 낮은 비용으로 높은 수익을 기대할 수 있다. 이렇게 마케팅 캠페인에 반응하여 구매할 확률 및 금액이 높은 고객을 예측하는 방법론을 반응 모델링이라고 한다. 잘 구축된 반응 모델은 마케팅 수익을 높이지만, 그렇지 못한 반응 모델은 마케팅 수익뿐 아닌 고객과의 관계 또한 악화시킨다고 알려져 있다[1-3].

일반적인 반응 모델링은 분류(classification) 기법을 이용하여 고객의 잠재적 반응 확률을 예측하는 모델을 뜻한다. 그 후 반응 확률이 높은 순서대로 선별하여 마케팅을 진행하게 된다. 하지만 KDD98 Cup에서 증명되었듯이, 고객의 반응 확률과 기대 반응 금액은 서로 상관관계가 낮을 수 있다[4-6]. 따라서 기존의 분류 기반 반응 모델은 반응 확률을 최대화 시키는 모델링을 하였지만, 이와 별개로 마케팅 수익을 최대화 하는 반응 모델의 필요성이 대두되었다. 2단계 반응 모델링(two-stage response modeling)은 고객의 반응 확률뿐 아닌 반응 금액 또한 예측하여 마케팅의 기대 수익을 최대화하려는 목적에서 제안되었다[5,6]. 2단계 반응 모델의 첫 단계에서는 기존 반응 모델링에서와 마찬가지로 분류 모델을 통해 반응할 확률이 높은 고객을 찾는다. 그리고 두 번째

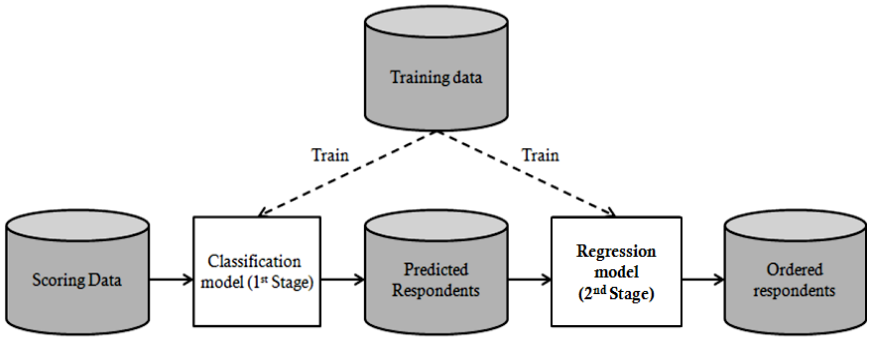


그림 1. 2단계 반응 모델링[6]
Fig. 1. Two-Stage Response Modeling[6]

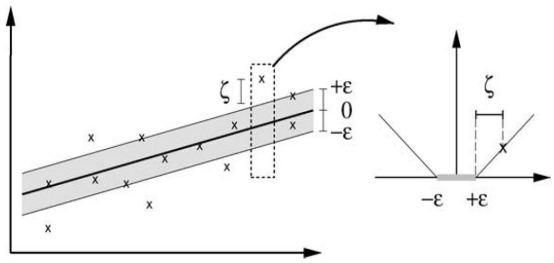


그림 2. ε-튜브를 이용한 지지 벡터 회귀 모델[7]
Fig. 2. ε-tube based Support Vector Regression[7]

단계에서는 회귀(regression) 모델을 이용하여 각 고객의 기대 반응 금액을 예측하여 마케팅 대상을 선정한다. 2단계 반응 모델링의 과정은 그림 1과 같다. 2단계 반응 모델은 기존 분류 모델 기반 1단계 반응 모델이나, 회귀 모델 기반 1단계 반응 모델, 그리고 두 개의 결과를 조합한 하이브리드(hybrid) 모델에 비해 더 높은 수익을 낼 수 있다[6].

2단계 반응 모델에서 가장 널리 사용되고 있는 회귀 모델은 지지 벡터 회귀(support vector regression) 모델이다(그림 2 참고). 지지 벡터 회귀 모델은 지지 벡터 머신과 마찬가지로 커널(kernel) 함수를 이용하여 비선형 문제를 해결할 수 있어 높은 학습 성능이 보장되며, 구조적 위험 최소화(structural risk minimization) 방식의 학습 기법이기 때문에 높은 일반화 성능을 보여주고 있다[7-11]. 하지만 지지 벡터 회귀 모델은 지도 학습(supervised learning)에 기반한 학습 방법으로 레이블(label)이 존재하는 데이터만을 학습할 수 있다는 한계가 있다. 즉, 반응 모델링의 경우 과거 마케팅 캠페인의 반응 여부를 알고 있는 고객만 모델링에 사용될 수 있다. 일반적으로, 마케팅 캠페인은 시간과 비용의 제약 때문에 전체 고객 중 일부의 고객에게만 진행된다. 따라서 다수의 고객들은 마케팅 캠페인의 대상이 되지 못했으므로, 이들의 데이터는 학습에 사용할 없다. 이러한 경우, 반응 모델링이 전체 고객의 성향을 학습하지 못하고 일부의 편향(biased)된 부분 집합만을 학습하여 수익성이 감소한다는 한계가 있다.

본 논문에서는 이러한 단점을 극복하고 반응 모델링의 정확도와 수익성을 높이기 위해, 레이블이 없는 데이터를 이용하여 학습 성능을 높이는 준지도 지지 벡터 회귀 모델(semi-supervised support vector regression)을 이용한 반응 모델링을 제안한다. 준지도 학습(semi-supervised learning)은 레이블이 없는 대량의 데이터를 레이블이 있는 데이터와 함께 학습하여 더 좋은 성능의 모델을 구축하는 방법론을 뜻한다[12]. 일반적으로 마케팅은 레이블이 없는 데이터가 대량으로 존재하는 환경이기 때문에 준지도 학습을 통해 얻을 수 있는 기대 효과가 높은 영역이라고 볼 수 있다.

본 논문에서 제안하는 알고리즘에서는 학습 복잡도를 낮은 수준으로 유지하여 효율적인 학습을 진행하기 위해 반복적 학습법을 택하지 않는다. 대신 레이블 없는 데이터의 레이블 추정에서 불확실성(uncertainty)을 고려하기 위해, 점추정(point estimation)을 하지 않고 분포추정(distribution estimation)을 하여 레이블이 존재할 수 있는 영역을 정의한다. 레이블 분포는, 지역성(locality)을 고려한 분포와 전역성(globality)을 고려한 분포를 각각 추정하여 이들의 조합(conjugation)을 통해 추정하게 된다. 그리고 불확실성을 지지 벡터 회귀 분석의 마진에 충분히 반영하기 위해 오버샘플링(oversampling)을 통해 레이블이 없는 데이터에 대한 레이블을 추출한다. 이 때, 불확실성의 정도에 따라 샘플링 비율을 다르게 함으로써, 불확실한 영역에 대해 더 많은 정보를 학습하게 한다. 마지막으로 오버샘플링에 따른 데이터 수 증가를 완화하기 위해 기존에 제안되었던 지지 벡터 회귀 모델에서의 학습 데이터 선택 기법인 Expected Margin Based Pattern Selection(이하 EMPS, [6])을 사용하여 최종적인 학습 데이터 셋을 구축한다. 본 연구의 실험에서는 반응 모델링의 실제 데이터로 많이 사용되는 DMEF4 데이터 셋이 이용되었다. 샘플링을 통해 DMEF4 데이터 셋에서 10개의 학습 상황을 만든 후, 레이블 데이터의 비율을 바꿔가면서 실험을 진행하였다. 실험 결과에서는 학습 정확도와 학습 속도뿐 아닌, 반응 모델링에서의 마케팅 활동 수준에 따른 기대 수익을 함께 제시한다.

II. 관련 연구

1. 반응 모델링

반응 모델링은 고객들의 인구통계학 및 과거 구매 이력 데이터 등을 입력 변수로 사용하여, 다가올 마케팅 캠페인에서의 각 고객의 구매 확률 및 기대 수익을 예측하는 모델링을 의미한다.

반응 모델링에 기본적으로 사용되는 기법들로는 주로 능형 회귀 모델(ridge regression)과 같은 통계적 기법[13]이나 의사 결정 나무와 같은 기계 학습 알고리즘이다[14,15]. 또한 마케팅 분야에서 전통적으로 널리 사용되어 왔던 로지스틱 회귀분석과 인공 신경망 알고리즘을 이용한 반응 모델링도 제안되었으며[16,17], 특히 인공 신경망의 경우 좋은 학습 성

능을 보장하므로 앙상블 러닝[18], 변수 선택[19]을 통해서 반응 모델링의 성능을 높이는 방향으로 보다 깊은 연구가 진행되었다. 최근 지지 벡터 알고리즘이 높은 일반화 성능으로 주목을 받은 후, 이를 이용하여 반응 모델링의 성능을 높이는 연구들이 주로 진행되었다. 우선 지지 벡터 분류 모델(support vector classifier)을 이용한 반응 모델링이 제안되었으며[2], 지지 벡터 회귀 모델을 이용해 기대 수익을 높이는 2단계 반응 모델링도 제안되었다[6]. 또한 비대칭(imbalanced) 문제를 해결하기 위해 이상치 탐지 지지 벡터 모델(1-class support vector classifier)을 이용하여 반응하지 않은 고객만 학습하는 방식 또한 제안되었다[20].

최근에는 위와 같이 지도 학습에 기반한 각 알고리즘을 적용하는 데에 그치지 않고, 새로운 학습 방법론을 이용한 반응 모델링 연구가 진행되고 있다. 그 예로, 비지도 학습과 지도 학습을 조합하는 방법이 제안되었고[21], 데이터 전처리와 조합하는 방법[22], 군집화, 언더샘플링, 앙상블을 모두 이용하는 방법[23] 준지도 학습을 이용하는 방법[24] 등이 제안되었다. 그러나 위 연구들은 1단계 반응 모델인 분류 모델에만 그 연구를 집중했다는 한계점이 있다. 한편, 기존의 마케팅 영역에서 벗어나 소셜 미디어(social media), 하이테크(high-tech) 마케팅 등 새롭게 떠오르는 마케팅 분야에서도 반응 모델링을 적용하려는 연구가 진행되고 있다[25,26].

2. 준지도 학습

일반적으로 레이블이 있는 데이터는 그 개수가 적고 수집 비용이 많이 드는 반면, 레이블이 없는 데이터는 그 양이 많으며 수집 비용이 적은 것으로 알려져 있다. 따라서 이러한 대량의 레이블이 없는 데이터를 이용해 학습 성능을 높이는 것이 준지도 학습이다.

준지도 학습의 기본적인 아이디어는 레이블이 있는 소량의 데이터를 이용하여 레이블이 없는 대량의 데이터의 레이블을 추정한 후, 이렇게 레이블이 부여된 데이터를 모델링에 추가적으로 사용한다는 데에 있다. 준지도 학습의 기본적인 아이디어는 자기 학습(self-training) 방법에 적용되었다. 자기 학습은 레이블 데이터로 학습 모델을 구축하여 레이블이 없는 데이터의 레이블을 추정한 후, 학습 성능을 높여주는 데이터를 학습 데이터에 추가하는 반복적(iterative) 방식이다. 자기 학습은 간단하고 준지도 학습의 아이디어를 쉽게 이해시킬 수 있으나, 지역 최적해(local optimum)에 빠져 잘못된 해로 수렴할 위험이 높다.

이러한 자기 학습의 단점을 극복하기 위해 제안된 방법이 Co-Training이다[27,28]. Co-Training은 자기 학습 방법에서 발전해서, 서로 다른 관점을 가지는 두 학습 모델을 만들고 각각을 이용해 레이블 없는 데이터의 레이블을 추정한 후 성능을 높여주는 데이터를 선택하지만, 해당 데이터는 자기 자신이 아닌 상대편 모델의 학습 데이터에만 추가된다. 즉, 지역 최적해에 빠지지 않기 위해 높은 신뢰성을 갖는 데이터를 선별하되, 자신이 사용하는 것이 아니라 다른 관점을 가진 상대편 학습 모델에게 넘겨주게 된다. Co-Training에서는 지역 최적해에 빠지는 경우는 드물게 발생하나, 많은 반복을 하는 동안 점점 더 많아지는 학습 데이터에 대해 두 개의 모델을 계속 학습시켜나감으로써 학습 복잡도(training complexity)가 매우 높다는 단점이 있다.

회귀 모델에서의 준지도 학습법 역시 분류 모델과 비슷한 방향으로 진행이 되어왔다. 그러나 회귀 모델의 경우 레이블이 없는 데이터의 추정해야 할 레이블이 이진(binary) 레이블이 아닌 실수(real number)이므로, 분류 모델에서의 준지도 학습보다 더 까다롭다고 볼 수 있다. 따라서 회귀 모델을 위한 준지도 학습법은 학습 복잡도가 높아도 성능 향상에 유리한 Co-Training 기반 방법이 가장 일반적으로 제안되었다. COREG[29] 은 가장 대표적인 알고리즘으로, 두 개의 k-인접 이웃(k-nearest neighbors) 회귀 모델을 사용하여 레이블이 없는 데이터의 레이블을 추정해 나간다. 이와 유사한 알고리즘으로 제안된 Co-SVR[30]은 마찬가지로 Co-Training 방식을 사용하나 k-인접 이웃이 아닌 지지 벡터 회귀 모델을 기본 모델로 사용한다는 차이점이 있다. 이러한 방식들은 높은 성능을 보여주지만, 두 개의 모델을 대량에 데이터를 추가해가며 반복적으로 학습한다는 데에서 높은 학습 복잡도를 가진다는 단점이 있다. 또한 마진 최대화(maximum margin) 형태로 문제를 풀어나가는 지지 벡터 회귀 모델에 적합하지 않다는 단점 또한 있다.

III. 준지도 지지 벡터 회귀 모델

$$D = L, U \qquad (1)$$
$$where\ L = L_x, L_y, L_x \in R^d, L_y \in R$$
$$U = U_x, U_x \in R^d$$

준지도 학습법의 아이디어는 레이블이 없는 데이터의 레이블을 추정하여, 이를 기존 레이블이 있는 데이터와 함께 학습함으로써 학습 모델의 성능을 높이는 데에 있다. 이 과정에서 학습 모델의 성능에 가장 큰 영향을 미치는 것은 레이블의 추정 부분이며, 대부분의 준지도 학습법은 사실상 이 부분을 명

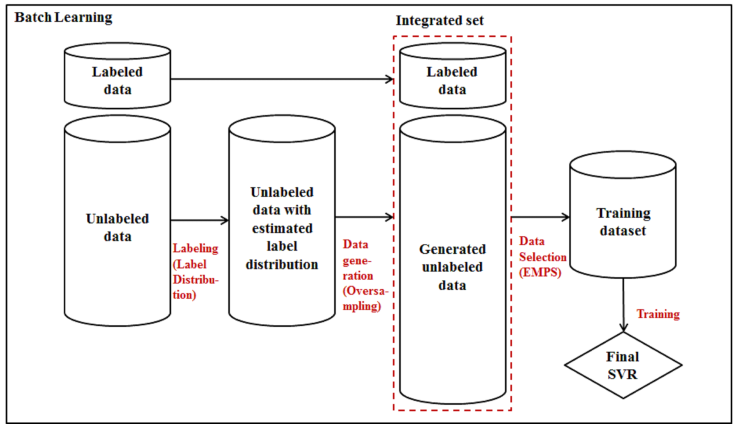


그림 3. 준지도 지지 벡터 회귀 모델 학습 방법론[31]
Fig. 3. Semi-Supervised Support Vector Regression (SS-SVR)[31]

명한다고 봐도 무방하다. 즉, 식 (1)에서 전체 데이터 셋 D는 레이블이 있는 L과 레이블이 없는 U로 나뉘며, 준지도 학습은 L을 이용하여 U_x의 레이블인 $U_{\hat{y}}$을 추정하여 학습에 사용하도록 하는 것이다. 그러나 레이블 없는 데이터의 레이블을 추정하여 학습에 사용한다는 것은, 이미 레이블을 갖고 있는 데이터가 만든 모델의 평가 값을 이용하는 것이므로, 이로 인해 모델이 새로운 정보를 학습하여 모델 성능을 향상시킨다는 것은 매우 어려운 일이다. 따라서 많은 준지도 학습 방법론들은 실제 학습 성능을 향상시키는 데이터를 순차적으로 추가하여 학습 데이터 셋의 크기를 점점 키워나가는 반복적 방법을 사용하며, 자기 자신이 아닌 상대 모델에게 추가함으로써 새로운 정보를 창출하는 효과를 노린다. 하지만 이는 매우 학습 복잡도가 높은 방식이며 레이블 과정의 불확실성을 고려하지 않는다는 한계점이 있다.

본 논문에서 제안하는 학습 방법론의 개요는 그림 3과 같다. 우선 레이블이 있는 데이터와의 관계를 바탕으로 레이블이 없는 데이터의 레이블을 추정한다. 단 이 때 레이블 자체를 추정하는 것 이 아니라 레이블 분포를 추정한다. 그리고 오버샘플링을 이용하여 하나의 분포에서 여러 레이블을 추출하여 레이블이 없는 데이터에 붙여 학습 데이터로 만들어준다. 이 때, 불확실성이 적은 데이터는 중복을 피하기 위해 더 적은 숫자의 데이터를, 불확실성이 높은 데이터는 해당 영역에 고루 존재할 가능성이 높으므로 더 많은 숫자의 데이터를 오버샘플링한다. 이렇게 만들어진 학습 데이터는 기존 레이블이 있는 데이터와 합쳐지는데, 오버샘플링 과정에서 늘어난 데이터는 EMPS 알고리즘을 통해 학습 복잡도를 낮추는 과정을 거쳐 정제하여 최종 학습 데이터 셋이 된다.

1. 레이블이 없는 데이터의 레이블 분포 추정

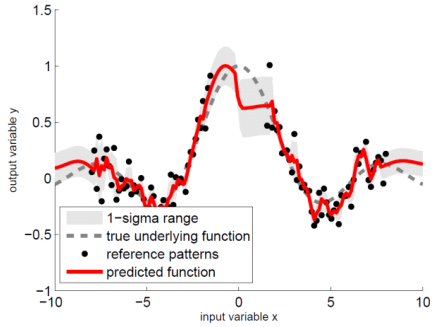
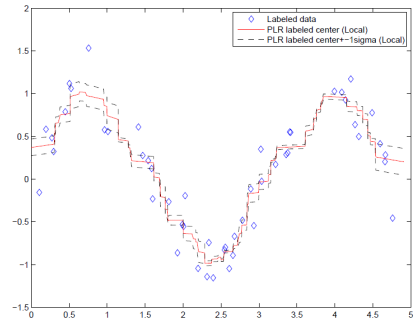


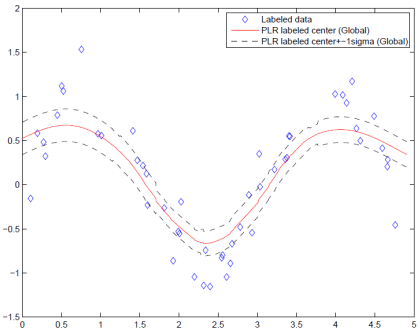
그림 4. PLR의 결과와 레이블 추정 신뢰 구간[32]
Fig. 4. Result of PLR and its estimated confidence interval[32]

레이블이 없는 데이터의 레이블 분포를 추정하기 위해, 확률적 지역 재구축(Probabilistic Local Reconstruction, 이하 PLR, [32]) 알고리즘을 사용했다. PLR 알고리즘은 사례기반 학습(instance based learning)의 한 분야인 지역적 선형 사상(locally linear embedding, [33]) 알고리즘을 수정하여 회귀 모델 형태로 만든 알고리즘이다. PLR에서는 $p(y_i) = N(\overline{y_i}, \sigma_i^2)$와 같이 y_i가 존재하는 확률을 정규분포로 표현한다. 그리고 선형 회귀 모형은 $y_i = w^T \bullet x_i + N(0, \sigma_i^2)$과 같이 가중치 벡터인 w와 정규성을 가진 노이즈인 $N(0, \sigma_i^2)$로 표현한다. 이후 PLR은 우도 확률 분포(likelihood distribution) 추정 및 사전 확률 분포(prior probability distribution) 추정을 통해 최종적인 사후 확률 분포(posterior probability

distribution)을 추정하고, 이는 베이지안 선형 회귀 모델(Bayesian linear regression)과 같이 출력 값이 정규 분포의 형태로 도출된다. 다만 이 과정에서 테스트 데이터가 학습 데이터로 잘 설명이 되는, 즉 유사한 형태의 데이터라면 출력 값인 정규 분포의 분산의 값은 작아진다. 반면, 테스트 데이터가 학습 데이터와 거리가 멀어서 해당 테스트 데이터를 잘 설명할 수 없는 경우, 출력 값이 정규 분포의 분산은 커지게 된다. 본 논문에서 준지도 학습을 위해 PLR 알고리즘을 사용한 이유는 다음과 같다. 우선 출력 값이 정규 분포의 형태로 나오므로, 추후 샘플링을 통해 데이터 생성 과정을 거치기 유리하다. 또한 무엇보다 출력 값의 신뢰도를 데이터 간의 거리에 기반하여 정규 분포의 분산 형태로 반영하기 때문에, 레이블 추정의 불확실성을 정량적으로 계산해준다(그림 4 참고). 마지막으로 사례기반 학습 방법론이기 때문에, 함수기반 학습(function based learning)에 비해 출력 값이 존재할 수 있는 범위가 넓어서, 마진 기반 학습인 지지 벡터 회귀 모델의 학습에 도움을 줄 수 있다.



(a)



(b)

그림 5 PLR의 결과, 작은 k를 썼을 경우(a)와 큰 k를 썼을 경우(b)[31]
Fig. 5. The result of PLR with a small k-value(a), and a big k-value(b)[31]

PLR 알고리즘에는 인접 이웃의 개수를 결정하는 파라미터 k가 존재한다. 기본적으로 사례기반 학습은 이러한 파라미터 k에 매우 민감하다. 본 논문에서는 그런 단점을 극복하기 위해 두 개의 서로 다른 k를 가지는 PLR을 학습하고 이들의 결과를 조합함으로써 최종적인 추정된 레이블 분포를 도출하도록 하였다. 즉, 작은 k를 가지는 모델(PLRlocal)은 보다 이웃 데이터의 특성을 잘 반영하는 방향으로 학습되지만 노이즈에 민감해진다. 반면 큰 k를 가지는 모델(PLRglobal)은 전체 분포를 고루 잘 반영하지만 이웃 특성을 잘 반영하지는 못한다(그림 5 참고). 이러한 두 개의 PLR을 통 해 추정된 정규 분포를 조합하여 최종적인 정규 분포의 모습으로 만들어 주기 위해, 베이지안 기법(Bayesian method)에서 주로 사용되는 사전 확률과 우도 확률을 조합(conjugation)하는 방식을 사용하였다. 즉, PLRglobal은 보다 전체적인 잠재 함수를 모델링하므로 사전 확률의 개념으로 사용될 수 있고, PLRlocal은 보다 현 데이터에 대한 지역적 일치성을 모델링하므로 우도 확률의 개념으로 사용될 수 있다. 두 개의 확률을 조합한 최종 정규 분포 형태는 다음 식 (2)와 같다.

$$\hat{y} = N(\bar{y}, \sigma^2), \tag{2}$$

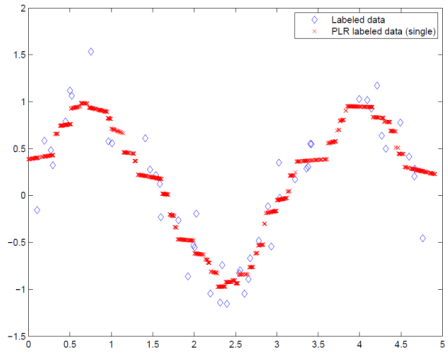
$$where\ \bar{y} = \frac{\frac{\bar{y}_{global}}{\sigma^2_{global}} + \frac{n \times \bar{y}_{local}}{\sigma^2_{local}}}{\frac{1}{\sigma^2_{global}} + \frac{n}{\sigma^2_{local}}},$$

$$\sigma^2 = \frac{1}{\frac{1}{\sigma^2_{global}} + \frac{n}{\sigma^2_{local}}}$$

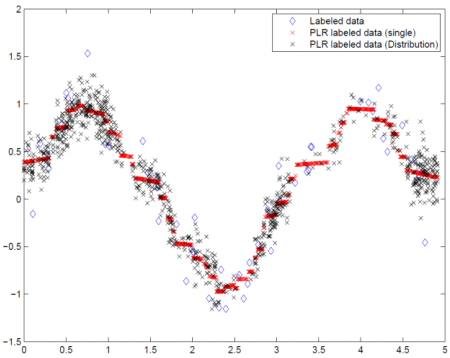
2. 추정된 레이블 분포에서의 오버샘플링을 통한 학습 데이터 생성

앞 단계에서 레이블이 없는 데이터 각각에 대해 정규 분포 형태로 추정된 레이블 분포를 구했다. 그 이유는 레이블이 없는 데이터 모두를 같은 방식으로 학습 데이터로 사용하는 것이 아닌, 이들 레이블 추정의 불확실성에 따라 다르게 사용하기 위함이다. 지지 벡터 회귀 모델은 마진 최대화 과정을 통해 학습이 이루어진다. 즉, 미리 설정된 파라미터인 ε의 영역만큼 데이터가 고루 분포되어 있어 마진이 일정 수준으로 유지가 될 때 좋은 학습 성능을 유지할 수 있다.

그림 6을 보면, 점추정을 통한 데이터 추출(a)은 기존에 갖고 있던 레이블이 있는 데이터에 비해 작은 마진을 가진 데이터만 생성된다는 것을 알 수 있다. 이런 경우 레이블이 없는 데이터를 새로운 정보로 이용해서 학습 성능을 높인다기보다, 기존 레이블이 있는 데이터가 주는 정보의 부분 집합에



(a)



(b)

그림 6. 점추정을 통한 데이터 추출(a)과 분포추정을 통해 오버샘플링 된 데이터 추출(b)[31]
Fig. 6. Data Generation from Point Estimation(a), and Data Generation from the Oversampling(b)[31]

해당된다고 볼 수 있다. 그리고 이런 경우는 지지 벡터 회귀 분석에서 지지 벡터로 선택이 되지 않을 가능성이 높기 때문에, 레이블이 없는 데이터는 최종 회귀 모델에서 무의미할 수 있다. 반면, 분포 추정을 통해 해당 분포 안에서 오버샘플링 된 데이터(b)는 마진의 영역에 고루 존재하며 데이터 셋을 풍성하게 만들어주는 것을 확인할 수 있다. 즉, 레이블이 없는 데이터가 지지 벡터 회귀 모델의 최종적인 학습에 새로운 정보를 부여함으로써 학습 성능 향상을 추구할 수 있게 된다.

본 논문에서는 이러한 오버샘플링 과정에서 레이블 불확실성을 고려하기 위해 추정된 레이블 분포의 분산을 이용하였다. 추정된 레이블 분포에서 분산이 작은 데이터들은 샘플링 비율이 높아도 비슷한 영역에 중복된 데이터를 만들게 된다. 즉, 이러한 데이터들은 학습에 추가적인 정보를 부여하지 않고, 다만 학습 복잡도를 높일 뿐이다. 반면, 추정된 레이블 분포의 분산이 큰 데이터는 해당 레이블이 존재할 영역이 넓고 불확실성이 높다는 의미이므로, 오버샘플링의 샘플링 비율을 높임으로써 데이터가 고루 분포할 수 있도록 도와준다. 최종적으로 학습하는 지지 벡터 회귀 모델은 이러한 영역에서 마진을 최대화할 수 있는 방향으로 학습을 하며, 이는 기존 레이블이 있는 데이터만 학습했을 때 보다 더 데이터의 영역을 잘 학습할 수 있게 된다. 이렇게 분산에 따라 달라지는 오버샘플링 비율, 즉 σ_i^2 에 기반한 x_i의 추출 확률은 최대 오버샘플링 개수인 t에 대해, 각 오버샘플링 시행마다 식 (3)에서 부여되는 확률 값에 따라 샘플링 여부가 결정된다.

$$p_i = \frac{\sigma^2 - \min(\sigma^2)}{\max(\sigma^2) - \min(\sigma^2)} \tag{3}$$

3. 학습 데이터 선택 및 최종 회귀 모델 학습

준지도 학습 방법은 기본적으로 학습 성능 향상을 위해 높은 학습 복잡도를 가지는 알고리즘을 사용한다. 본 논문에서 제안하는 방법은 반복적 학습을 사용하지 않음으로써 낮은 학습 복잡도를 가져갈 수 있지만, 오버샘플링을 통해 학습 데이터의 양이 늘어났으므로 최종 회귀 모델의 학습 복잡도는 다소 높아질 수 있다. 이러한 단점을 극복하기 위해 전체 학습 데이터 중 학습에 꼭 필요한 부분 집합을 선택해서 학습 복잡도를 낮추는 알고리즘인 EMPS를 적용하였다. 기본적으로 지지 벡터 회귀 모델에서는 마진이 미리 설정된 파라미터인 ε보다 큰 데이터만 지지 벡터로 뽑혀서 학습 모델에 영향을 미친다. 따라서 EMPS는 학습 이전에 지지 벡터가 될 가능성이 높은 데이터만 선택하는 알고리즘이다. EMPS의 작동 원리는 그림 7과 같다. 우선 전체 데이터 셋(a)에서 랜덤 샘플링을 통해 데이터를 추출하여 학습한 후(b), 학습된 모델을 이용해 전체 데이터 셋의 마진을 계산한다(c). 이 과정을 여러 번 반복함으로써 모든 데이터에 대해 마진을 추정한 후, 마진이 ε보다 큰 데이터만 학습 데이터로 선택된다(d). 최종적인 지지 벡터 회귀 모델을 이용한 반응 모델은 이렇게 선택된 학습 데이터를 이용해 구축된다. 본 논문의 준지도 지지 벡터 회귀 모델의 알고리즘은 다음 그림 8과 같다.

IV. 준지도 지지 벡터 회귀 모델을 이용한 반응 모델링

1. 실험 설정

우선 최종 모델인 지지 벡터 회귀 모델의 파라미터인 Cost, ε은 각각 {0.1, 0.5, 1, 3, 5, 7, 10, 20, 50, 100}

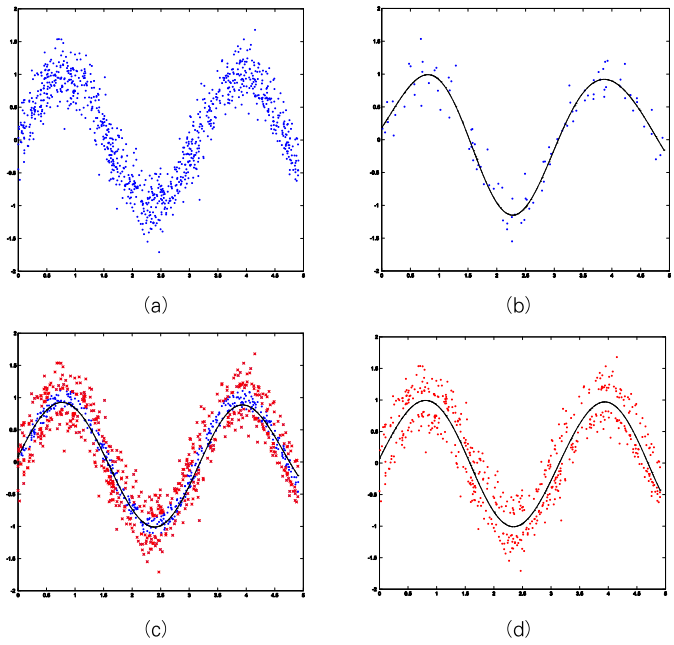


그림 7. EMPS 알고리즘의 작동 예시[6]
Fig. 7. An Example of EMPS[6]

1. 초기화
K_{local}: PLR_{local}을 위한 인접 이웃 수
K_{global}: PLR_{global}을 위한 인접 이웃 수
t: 오버샘플링 최대 개수
ADD_{Ux}: 레이블이 없는 데이터의 입력 변수를 추가해나갈 빈 셋
ADD_{Uy}: 레이블이 없는 데이터의 레이블을 추정하여 추가해나갈 빈 셋

2. 레이블이 없는 데이터의 레이블 추정
$N_{local} \leftarrow PLR_{local}(Lx,\ Ly,\ Ux,\ k_{local})$
$N_{global} \leftarrow PLR_{global}(Lx,\ Ly,\ Ux,\ k_{global})$
$N_{conjugate} \leftarrow Conjugate(N_{local},\ N_{global})$

3. 오버샘플링을 통한 데이터 생성
FOR Ux
FOR 1 to t
  $r \leftarrow Uniform(0,1)$
  IF $p_i > r$
  $ADD_{Ux} \leftarrow ADD_{Ux} \cup Ux_i$
    $ADD_{Uy} \leftarrow ADD_{Uy} \cup y_i$($y_i$는 $N_{conjugate}$에서 랜덤 생성)
  END IF
END FOR
END FOR

4. 데이터 선택
$\{TRNx,\ TRNy\} \leftarrow EMPS(\{Lx,\ ADD_{Ux}\}, \{Ly, ADD_{Uy}\})$

5. 최종 반응 모델링 학습
$Model \leftarrow SVR(TRNx,\ TRNy)$

그림 8. 준지도 지지 벡터 회귀 모델 알고리즘
Fig. 8. The Algorithm of Semi-Supervised Support Vector Regression

과 {0.01, 0.05, 0.07, 0.1, 0.15, 0.3, 0.5, 0.7, 0.9, 1}의 범위에서 교차검증(cross-validation)을 통해 구하였다. 그리고 커널 파라미터는 정규화된 데이터에 대해서 모두 1로 통일되어 설정되었다. 비교 알고리즘으로는 앞에서 언급한 COREG과 Co-SVR을 사용하였다. COREG에서는 최종 반응 모델을 구축하는 알고리즘에 따라 $COREG_{kNN}$과 $COREG_{SVR}$을 모두 구현하여 실험하였다. COREG은 두 개의 모델을 학습하므로 서로 다른 두 개의 인접 이웃 파라미터 k가 필요한데, 본 실험에서는3과 5로 설정하였다. Co-SVR의 지지 벡터 회귀 모델 파라미터는 최종 모델의 파라미터와 동일하게 설정하였다. COREG과 Co-SVR에서 작동 셋(working set)의 크기는 각각 100, 40으로 설정하였다. 제안하는 기법의 파라미터인 k_{local}과 k_{global}은 각각 5와 10으로 고정하였으며, 오버샘플링의 최대 허용치는 3배수로 설정하였다. EMPS의 파라미터는 모두 10으로 통일되었다. 각 실험에서 레이블이 있는 데이터의 비율은 20%, 10%, 5%, 그리고 1%로 설정하여 다양한 환경에서의 일반화 성능을 비교하였다. 레이블 데이터는 전체 데이터에서 랜덤으로 설정하였으며, 따라서 각 레이블 비율에 대해 10번씩 반복 실험을 하여서 그 평균을 실험 결과로 제시한다.

2. 알고리즘 성능 실험 결과

표 1. 실험 데이터 셋
Table 1. Experiment Datasets

번호	이름	학습 데이터 수	테스트 데이터 수	입력 변수 수	출처
1	Santa Fe D	2,000	2,000	10	Santa Fe Competition
2	Melbourne Temperature	2,000	1,000	10	Tim Series Data Library
3	Abalone	2,000	2,000	10	Delve
4	Bank 8NH	2,000	2,000	8	Delve
5	Pumadyn 8NH	2,000	2,000	8	Delve

반응 모델링 실험 결과를 얻기 전에, 알고리즘의 성능 자체를 평가할 수 있는 실험을 진행하였다. 실험은 총 5개의 데이터 셋을 사용하여 진행하였다. 실험에 사용된 데이터 셋은 표 1에 정리되어 있다. Santa Fe D 데이터 셋과 Melbourne Temperature 데이터 셋은 시계열 데이터 셋이고, 나머지는 일반 회귀 분석 데이터 셋이다. Bank와 Pumadyn 데이터 셋에서의 8NH는 8개의 변수, 비선형(nonlinear), 그리고 높은 노이즈 수준(high noise)을 의미한다. 각 데이터 셋 마다 레이블 데이터 비율을 20%, 10%, 5%, 그리고 1%로 설정하여 실험하였고, 랜덤성이 있는 실험이기 때문에 총 10회 반복 실험을 한 후 그 평균을 성능으로 기록하였다. 성능은 학습 효과를 측정할 수 있는 정확도와 학습 효율을 측정할 수 있는 학습 시간을 기록하였다. 정확도는 Root Mean Squared Error(이하 RMSE)를 기반으로 레이블이 있는 데이터만 학습했을 때를 100으로 했을 때, 상대적인 비율을 제시한다. 학습 시간은 준지도 학습 및 학습 데이터 선택 등 학습에 관여한 모든 과정을 실제 CPU 시간으로 기록한 초 단위의 시간을 의미한다.

표 2와 표 3은 각각 정확도와 학습 시간에 대한 실험 결과를 나타낸 것이다. 각 표는 레이블이 있는 데이터의 비율을 20%부터 1%로 변화시켜가며 실험한 결과를 담고 있으며, 레이블 데이터의 비율을 기준으로 평균적인 성능을 제시하고 있다. (*)표시가 되어있는 것은 평균적으로 해당 레이블 데이

표 2. 알고리즘 성능 실험 결과(RMSE 비율)
Table 2. Experimental Result in RMSE Ratio

데이터 셋	$COREG_{kNN}$	$COREG_{SVR}$	Co-SVR	제안기법
L = 20%				
1	110.22	102.11	113.00	98.17
2	100.53	99.52	102.49	98.19
3	102.88	99.88	103.14	100.30
4	96.55	99.73	102.01	97.84
5	98.08	99.63	106.81	97.71
평균	**101.65**	**100.17**	**105.49**	**98.44(*)**
L = 10%				
1	104.03	100.4	120.55	93.12
2	98.25	98.86	103.42	96.42
3	102.14	99.82	104.06	100.83
4	95.58	99.62	102.12	97.48
5	95.56	99.30	106.02	95.57
평균	**99.11**	**99.60**	**107.23**	**96.68(*)**
L = 5%				
1	90.64	99.45	118.61	83.96
2	94.00	97.58	104.30	93.66
3	99.86	100.33	102.18	99.65
4	95.63	99.53	101.69	96.81
5	93.40	98.93	104.06	94.29
평균	**94.70**	**99.16**	**106.16**	**93.67(*)**
L = 1%				
1	87.03	95.16	108.79	89.72
2	90.95	94.46	103.96	90.76
3	96.16	100.09	103.30	98.35
4	97.83	98.47	99.95	97.82
5	95.30	99.18	94.46	94.70
평균	**93.45(*)**	**97.47**	**102.09**	**94.27**

터의 비율 기준으로 가장 좋은 성능을 보여준 알고리즘을 의미한다. 표 2를 보면 정확도를 RMSE 비율로 표현하였다. 레이블 데이터의 비율이 1%일 때를 제외하고 모든 경우에서 제안 기법의 성능이 평균적으로 가장 좋았다. 그리고 레이블 데이터의 비율이 1%일 때 역시 가장 성능이 좋았던 $COREG_{kNN}$와 비교했을 때 크게 뒤쳐지지 않는 정확도를 보여준다. 표 3은 학습 시간을 초 단위로 기록한 것으로, 빅데이터 환경에서 가장 중요하게 여겨지는 것 중 하나인 학습 알고리즘의 실제 문제 풀이 효율성을 측정한 것이다. 제안 기법은 반복적 학습을 오버샘플링으로 대체하였으므로 Co-Training에 기반한 비교 알고리즘들에 비해 2~10배 가량 빠른 학습 속도를 기록했다. Co-SVR은 학습 시간의 편차가 커서 레이블 데이터의 비율이 1%였을 때 제안 기법과 비슷했으나, RMSE 비율에서는 좋지 않은 성능을 기록하였다. 알고리즘 성능을 공개 데이터 셋에 실험해본 결과, 제안 기법이 준지도 회귀 모델 중에서 가장 효율적인 학습이 가능함과 동시에, 가장 높은 수준의 학습 성능 또한 기대할 수 있다는 결론을 얻을 수 있었다.

표 3. 알고리즘 성능 실험 결과(학습 시간)
Table 3. Experimental Result in Training Time

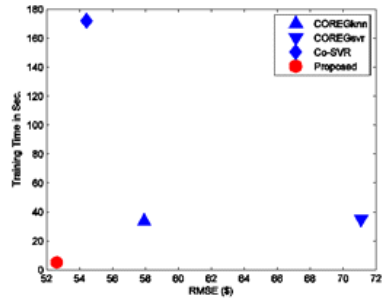
데이터 셋	$COREG_{kNN}$	$COREG_{SVR}$	Co-SVR	제안기법
L = 20%				
1	37.17	47.87	225.99	12.54
2	38.11	47.53	59.13	10.64
3	37.67	46.64	70.47	12.53
4	39.25	42.95	86.41	10.03
5	38.11	42.58	82.46	11.00
평균	**38.06**	**45.51**	**104.89**	**11.34(*)**
L = 10%				
1	28.07	35.19	45.81	7.24
2	28.43	33.39	11.87	6.12
3	29.15	33.37	21.27	6.82
4	28.11	31.28	26.73	6.44
5	28.33	29.80	25.13	6.50
평균	**28.41**	**32.60**	**26.16**	**6.62(*)**
L = 5%				
1	24.39	25.75	12.74	4.55
2	24.22	25.76	8.83	3.88
3	24.45	25.58	8.72	3.97
4	24.09	25.39	9.69	4.18
5	24.12	25.43	9.84	4.07
평균	**24.25**	**25.58**	**9.96**	**4.13(*)**
L = 1%				
1	22.01	22.58	3.08	2.56
2	21.79	23.07	2.99	2.23
3	21.99	23.04	2.97	2.14
4	21.88	22.87	5.22	2.11
5	21.95	22.95	2.99	2.35
평균	**21.92**	**22.90**	**3.45**	**2.27(*)**

3. 반응 모델링 실험 결과

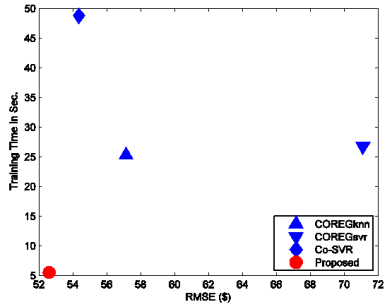
준지도 지지 벡터 회귀 모델을 이용한 반응 모델링의 성능 평가를 위해 실제 반응 모델링 데이터 셋인 DMEF4 데이터 셋을 사용하였다. DMEF4 데이터 셋은 101,532 고객들을 대상으로 메일을 통해 캠페인을 진행한 것으로, 고객에 대한 총 91개의 변수가 존재하며 전체 반응률은 9.4%이다. 본 논문에서는 91개의 입력변수 중, 표 4와 같이 과거 연구들을 통해 널리 사용되고 있는 15개의 변수만을 사용하였다[2,6,18,19,20,34]. 성능 평가를 위해 데이터 셋을 랜덤샘플링을 통해 학습 데이터 셋과 테스트 데이터 셋으로 나누었다[20]. 즉 데이터 셋은 전체 데이터 셋의 반인 50,766 고객들을 학습으로, 나머지 고객들을 테스트로 사용하여서 성능 평가용으로 재구축되었으며, 다양한 상황에 대한 평가를 이루기 위해 같은 방식으로 총 10개의 서로 다른 데이터 셋을 구축하였다. 각 알고리즘의 성능은 이 10개의 데이터 셋에 대한 평균으로 도출하였으며 모든 데이터 셋은 정규화 되었다. 2단계 반응 모델링에서 첫 단계인 분류 모델을 통해 반응 확률을 예측하는 것은 이 논문에서 다루고 있는 범위가 아니므로, 두 번째 단계인 구매 금액을 예측하는 회귀 분석 문제만을 대상으로 실험을 진행하였다.

그림 9는 반응 모델링 실험 결과를 정확도와 학습 속도의 측면에서 그림으로 표현한 것이다. 그림에서 x축은 RMSE를 원 타겟 변수 스케일인 달러($)로 표현되었으며, y축은 학습 시간을 초 단위로 측정한 것을 표현하였다. 파란 색으로 표현된 것은 삼각형, 역삼각형, 다이아몬드 모양이 각각 비교 알고리즘들인 $COREG_{kNN}$, $COREG_{SVR}$,그리고 Co-SVR을 의미하며, 빨간색 동그라미로 표현된 것은 제안 기법을 의미한다. RMSE와 학습 시간 모두 작을수록 좋으므로, 각 그래프에서 좌하단에 존재할수록 더 효율적이면서 효과적인 알고리즘임을 의미한다. 그림 9에서 확인된 것처럼, 제안 기법은 가장 빠른 시간에 준지도 학습을 할 수 있는 알고리즘이다. 레이블 데이터의 비율이 1%일 때(그림 9 (d)) Co-SVR에 비해 느린 학습 시간을 보여주었지만, 레이블 데이터의 비율이 5%~20%일 때에는 다른 알고리즘 대비 2~10배 가량 더 빠른 속도를 보여주고 있다. 학습 정확도 측면에서는 모든 경우에서 제안 기법이 가장 정확한 모델을 구축하고 있다는 것을 확인할 수 있다.

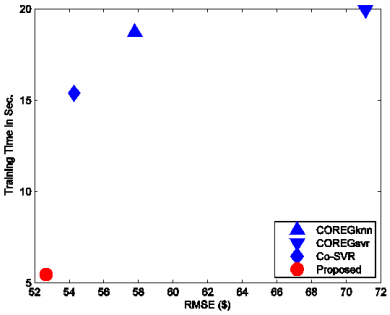
그림 10은 반응 모델링 실험 결과의 수익 분석을 나타낸 그림이다. 각 그림은 마케팅 캠페인을 진행하는 범위를 고객의 1%~10%로 진행할 때(x축), 그에 따른 수익(y축, 달러)



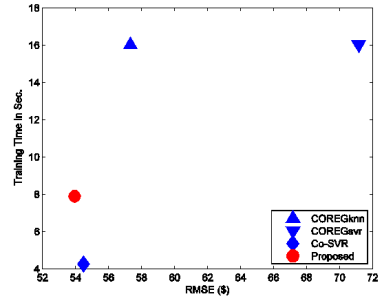
(a)



(b)



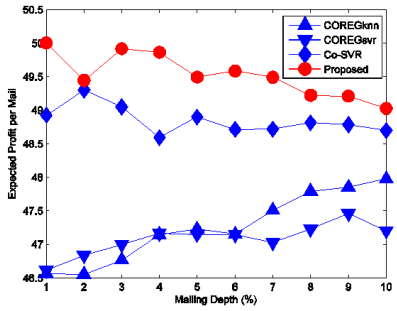
(c)



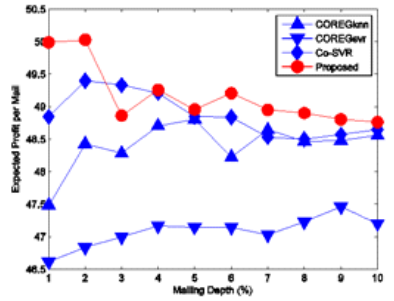
(d)

그림 9. 반응 모델링 실험 결과(레이블 비율(a-d): 20%, 10%, 5%, 1%)

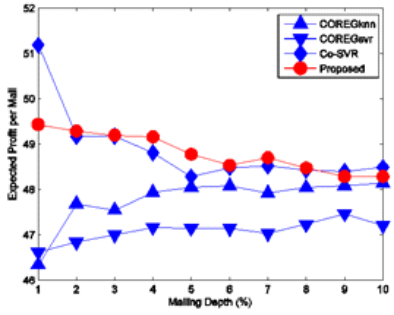
Fig. 9. Experimental Results of Response Modeling(Label Ratio (a-d): 20%, 10%, 5%, and 1%)



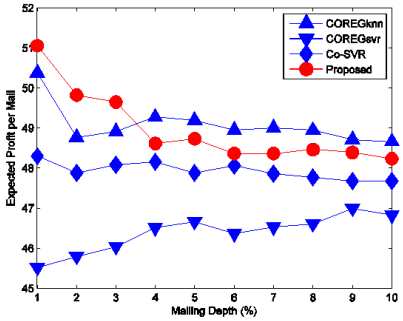
(a)



(b)



(c)



(d)

그림 10. 반응 모델링 수익성 분석 결과(레이블 비율(a-d): 20%, 10%, 5%, 1%)

Fig. 10. Profit Analysis Results of Response Modeling(Label Ratio(a-d): 20%, 10%, 5%, and 1%)

표 4. DMEF4 데이터 셋에서 사용된 변수들
Table 4. Input Variables Used for DMEF4 Dataset

이름	변환 식	설명
원 변수		
Purchase	-	구매한 시즌 수
Falord	-	LTD 가을 구매 수
Ordtyr	-	올 해 구매 수
Puryear	-	구매한 연 수
Sprord	-	LTD 봄 구매 수
변환 변수		
Recency	-	1992년 10월부터 구매 수
Tran53	I(180 < recency < 270)	
Tran54	I(270 < recency < 366)	
Tran55	I(366 < recency < 730)	
Tran38	1/recency	
Comb2	sum(ProdGrp)	올해 구매한 제품 그룹 수
Tran46	srqt(comb2)	
Tran42	log(1 + ordtyr * falord)	구매 수의 상호작용
Tran44	sqrt(ordhist * sprord)	LTD 구매와 LTD 봄 구매의 상호작용
Tran25	1/(1+lorditm)	지난 시즌 제품의 역

을 의미한다. 일반적으로 마케팅 캠페인에서는 제품이 고급화되고 마케팅이 고도화될수록 더 적은 수의 고객만을 타겟으로 선정한다. 따라서 반응 모델링의 수익 분석을 할 때에는 마케팅 캠페인의 진행 범위를 소수에서 다수로 펼쳐나갈 때에 따른 기대 수익의 변화를 살펴본다. 그림 10을 보면 전반적인 상황에서 제안 기법이 가장 많은 수익을 기대할 수 있다는 것을 알 수 있다. 레이블 데이터의 비율이 20%와 10%일 때는 한 경우만을 제외하고 제안 기법에 따른 수익이 다른 비교 알고리즘보다 우수했다. 레이블 데이터의 비율이 5%와 1%에서 역시 비슷했다. 다만, 레이블 데이터의 비율이 5%일 때 마케팅 캠페인 진행 범위가 고객의 1%일 때에는 Co-SVR의 수익성이 더 좋았다. 고객의 1%를 선정할 때에는 뽑히는 고객의 수가 때문에 결과의 편차가 크게 나타날 수 있다. 반대로 레이블 데이터의 비율이 1%일 때에는 마케팅 캠페인 진행 범위가 고객의 4% 이상일 때 COREGkNN의 수익성이 더 좋았다. 이 경우 RMSE와는 반대되게 COREGkNN이 조금 더 좋은 결과를 냈다고 볼 수 있다. 그러나 마케팅 캠페인의 진행 범위가 고객의 1~3%일 때에는 제안 기법이 가장 많은 수익을 내었다. 제품 및 마케팅의 고급화 전략을 사용할 때 제안 기법의 수익성이 더 높을 것으로 기대할 수 있다.

V. 결론

본 논문에서는 준지도 지지 벡터 회귀 모델을 이용한 반응 모델링을 제안했다. 최근 반응 모델링은 반응 확률뿐 아닌 구매 금액까지 예측해서 기대 수익을 최대화하는 2단계 반응 모델링으로 확장되어 연구되고 있다. 이러한 2단계 반응 모델링은 분류 모델과 회귀 모델이 함께 사용되는데 회귀 모델로는 지지 벡터 회귀 모델이 높은 일반화 성능을 바탕으로 널리 사용된다. 단, 지도 학습 기반 반응 모델링은, 전체 고객 데이터 중 마케팅 캠페인의 대상이 되었던 고객 데이터만을 사용해야 한다는 한계점이 있다. 이러한 한계를 극복하고 더 많은 고객 데이터를 사용해서 보다 높은 정확도와 수익을 가지는 반응 모델을 얻기 위해, 본 논문에서는 준지도 지지 벡터 회귀 모델을 이용하여 반응 모델을 구축하고, 실제 데이터 셋에 대한 실험을 통해 정확도 및 수익 분석을 하였다.

본 논문에서 제안한 준지도 지지 벡터 회귀 모델은, 레이블 불확실성을 고려하면서 학습 복잡도를 최소화하기 위해 반복적 학습법 대신 일괄(batch) 학습법을 사용하였다. 레이블이 없는 데이터에 대해 레이블을 점추정을 통해 얻어내지 않고, 두 개의 PLR을 조합하여 신뢰성 있는 레이블 분포를 추정했다. 그리고 불확실성을 고려하며 지지 벡터 회귀 모델의 마진을 보장해주기 위해 레이블이 없는 데이터와 그의 추정된 레이블 분포를 바탕으로 오버샘플링을 통해 학습 데이터를 생성했다. 이 과정에서 불확실성이 높은 지역에 대해서 더 많은 데이터를 생성함으로써 전 영역을 고루 학습할 수 있도록 하였다. 마지막으로 오버샘플링이 된 데이터의 수가 너무 많으면 학습 복잡도가 증가할 수 있으므로, 기존에 제안된 학습 데이터 선택 알고리즘인 EMPS를 이용하여 학습 복잡도를 낮은 수준으로 유지하였다.

알고리즘의 성능을 검증하기 위해 공개 데이터 셋에 실험한 결과, 비교 알고리즘들에 비해 제안 기법이 더 높은 정확도와 더 빠른 속도를 동시에 보여주었다. 마지막으로 실제 마케팅 데이터 셋에 반응 모델링을 구축해서 제안 기법의 실제 수익성을 분석하였다. 제안 기법은 정확도와 속도 측면에서 비교 알고리즘들 보다 더 우수하였다. 수익 분석의 경우 레이블 데이터의 비율에 따라 다소 다른 결과가 나왔지만, 전반적으로 제안 기법이 가장 우수하였으며, 특히 제품이나 마케팅 고도화에 따라 마케팅 캠페인 진행 범위가 좁은 경우 가장 뛰어난 성능을 발휘할 수 있다는 결론이 도출되었다.

본 논문의 한계점 및 추후 연구 방향은 다음과 같다. 우선, 학습 복잡도를 낮추기 위해 일괄 학습 방식을 채택한 결과,

학습 시간에서는 만족할만한 성과를 거두었지만 정확도와 수익성에서는 개선의 여지가 존재했다. 이를 위해 몇 단계의 반복을 거치며 정확도를 개선하는 방향으로 학습을 진행하는 점진적 학습(incremental learning)을 조합하는 연구를 진행하려 한다. 또한, 현재 알고리즘은 레이블의 분포 추정 및 오버샘플링 과정에서 최종적인 반응 모델링을 고려하지 않고 데이터를 복원하는 기능만 수행한다. 하지만 이 과정에서도 반응 모델링의 성능에 도움이 되는 데이터를 생성해내는 지능적 샘플링(intelligent sampling)의 기능이 존재한다면 더 좋은 성능을 기대할 수 있다. 마지막으로 가장 널리 사용되어 온 DMEF4 데이터 셋에 대한 성능 평가를 하였지만, 추후 다른 마케팅 데이터 셋 및 다른 분야로 확장하는 연구를 진행하려고 한다. 특히 공정 품질 데이터의 경우 품질 계측이 일부 샘플만 이루어져서 반응 모델링과 마찬가지로 준지도 학습으로 얻을 수 있는 기대 효과가 크다. 이와 같이 준지도 학습을 적용했을 때 기대 효과가 큰 도메인을 대상으로 하여 제안한 알고리즘의 활용도를 다양하게 실험해보는 연구를 진행하려 한다.

참고문헌

〔1〕 F.F. Gönül, B.D. Kim, and M. Shi, "Mailing Smarter to Catalog Customer," Journal of Interactive Marketing, Vol. 14, No. 2, pp.2-6, Apr. 2000.

〔2〕 H. Shin, and S. Cho, "Response Modeling with Support Vector Machines," Expert Systems with Applications, Vol. 30, No. 4, pp.746-760, May. 2006.

〔3〕 R.C. Blatberg, B.D. Kim, and S.A. Neslin, "*Database Marketing: Analyzing and Managing Customers,*" Springer, pp.245-287, 2008.

〔4〕 K. Wang, 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, pp.57-79, Jul. 2005.

〔5〕 D. Kim, H.J. Lee, and S. Cho, "Response Modeling with Support Vector Regression," Expert Systems with Applications, Vol. 34, No. 2, pp.1102-1108, Feb. 2008.

〔6〕 D. Kim, and S. Cho, "Pattern Selection for Support Vector Regression based Response Modeling," Expert Systems with Applications, Vol. 39, No. 10, pp.8975-8985, Aug. 2012.

〔7〕 A. Smola, and B. Schölkopf, "*A Tutorial on Support Vector Regression,*" NeuroCOLT Technical Report NC-TR-98-030, University of London, 2002.

〔8〕 V. Vapnik, "*The Natural of Statistical Learning Theory,*" Springer, pp.549-557, 1995.

〔9〕 H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," Advances in Neural Information Processing System, Vol. 9, pp.155-161, May. 1997.

〔10〕 B. Choi, and K. Cho, "Comparison of HMM and SVM Schemes in Detecting Mobile Botnet," Journal of the Korea Society of Computer and Information, vol.19, no.4, pp.81-90, 2014 Apr.

〔11〕 K. Huh, and S. Kim, "Context-Aware Fusion with Support Vector Machine," Journal of the Korea Society of Computer and Information, vol.19, no.6, pp.19-26, 2014 Jun.

〔12〕 X. Zhu, "*Semi-Supervised Learning Literature Survey,*" Technical Report 1350, University of Wisconsin at Madison, 2006.

〔13〕 E.C. Malthouse, "Ridge Regression and Direct Marketing Scoring Models," Journal of Interactive Marketing, Vol. 19, No. 4, pp.10-23, Nov. 1999.

〔14〕 D. Haughton, and S. Oulabi, "Direct Marketing Modeling with CART and CHAID," Journal of Direct Marketing, Vol. 11, No. 4, pp.42-52, Nov. 1997.

〔15〕 D.L. Olson, and B. C, "Direct Marketing Decision Support Through Predictive Customer Response Modeling," Decision Support Systems, Vol. 51, No. 1, pp.443-451, Dec. 2012.

〔16〕 E.H. Suh, K.C. Noh, and C.K. Suh, "Customer List Segmentation using the Combined Response Model," Expert Systems with Applications, Vol. 17, No. 2, pp.89-97, Aug.

1999.

[17] Y. Bentz, and D. Merunka, "Neural Networks and the Multinomial Logit for Brand Choice Moldeing: a Hybrid Approach," Journal of Forecasting, Vol. 19, No. 3, pp.177-200, Apr. 2000.

[18] K. Ha, S. Cho, and D. MacLachlan, "Response Models based on Bagging Neural Networks," Journal of Interactive Marketing, Vol. 19, No. 1, pp.17-30, Feb. 2005.

[19] E. Yu, and S. Cho, "Constructing Response Model using Ensemble based on Feature Subset Selection," Expert Systems with Applications, Vol. 30, No. 2, pp.352-360, Feb. 2006.

[20] H. Lee, and S. Cho, "Focusing on Non-Respondents: Response Modeling with Novelty Detectors," Expert Systems with Applications 33(2), pp.522-530, Feb. 2007.

[21] M. Daneshmandi, and M. Ahmadzadeh, "A Hybrid Data Mining Model to Improve Customer Response Modeling in Direct Marketing," Indian Journal of Computer Science and Engineering, Vol. 3, No. 6, pp.844-855, Dec. 2012.

[22] A.N. Aliabadi, "Hybrid Model of Customer Response Modeling Through Combination of Neural Networks and Data Preprocessing," In Proceedings of the 2013 IEEE International Conference on Fuzzy Systems (FUZZ), pp.1-4, Hyderabad, India, 2013 Jul.

[23] P. Kang, S. Cho, and D. MacLachlan, "Improved Response Modeling based on Clustering, Under-Sampling, and Ensemble," Expert Systems with Applications, Vol. 39, No. 8, pp.6738-6753, Jun. 2012.

[24] H. Lee, H. Shin, S. Hwang, S. Cho, and D. MacLachlan, "Semi-Supervised Response Modeling," Journal of Interactive Marketing, Vol. 24, No. 1, pp.42-54, Feb. 2010.

[25] M. Sun, Z.Y. Chen, and Z.P. Fan, "A Multi-task Multi-Kernel Transfer Learning Method for Customer Response Modeling in Social Media," Procedia Computer Science, Vol. 31, pp.221-230, Jun. 2014.

[26] H. Risselada, P.C. Verhoef, and T.H.A. Bijmolt, "Dynamic Effects of Social Influence and Direct Marketing on the Adoption of High-Technology Products," Journal of Marketing, Vol. 78, No. 2, pp.99-118, Mar. 2014.

[27] A. Blum, and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," In Proceedings of the Workshop on Computational Learning Theory, pp.92-100, New York, NY, USA, 1998 Jul.

[28] T. Mitchell, "The Role of Unlabeled Data in Supervised Learning," In Proceedings of the 6th International Colloquium on Cognitive Science, San Sebastian, Spain, 1999 May.

[29] Z.H. Zhou, and M. Li, "Semisupervised Regression with Cotraining-Style Algorithms," IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 11, pp.1479-1493, Nov. 2007.

[30] X. Wang, L. Fu, and L. Ma, "Semi-Supervised Support Vector Regression Model for Remote Sensing Water Quality Retrieving," Chinese Geographical Science, Vol. 21, No. 1, pp.57-64, Feb. 2011.

[31] D. Kim, P. Kang, and S. Cho, "Semi-Supervised Support Vector Regression Considering Labeling Uncertainty with Label Distribution Estimation and Oversampling," Neurocomputing, Submitted, Jun. 2014.

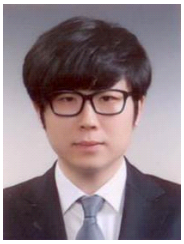
[32] S.K. Lee, P. Kang, and S. Cho, "Probabilistic Local Reconstruction in k-NN Regression and Its Applications to Virtual Metrology in Semiconductor Manufacturing," Neurocomputing, Vol. 131, pp.427-439, May. 2014.

[33] D. de Ridder, and R. Duin, "*Locally Linear Embedding for Classification*," Technical Report PH-2002-01, Delft University of Technology, 2002.

[34] E.C. Malthouse, "Performance-based Variable Selection for Scoring Models," Journal of

Interactive Marketing, Vol. 16, No. 4, pp.37-50, Nov. 2002.

저 자 소 개



김 동 일

2005: 서울대학교
산업공학과 공학사.
2013: 서울대학교
산업공학과 공학박사
(Data Mining)
현 재: 삼성전자 책임연구원
관심분야: 데이터마이닝, 빅데이터분석
Email : dikim01@snu.ac.kr