Support Vector Machine을 이용한 고객이탈 예측모형에 관한 연구

- A Study on Customer Segmentation Prediction Model using Support Vector Machine -

서 광 규 *
Seo Kwang Kyu

Abastrct

Customer segmentation prediction has attracted a lot of research interests in previous literature, and recent studies have shown that artificial neural networks (ANN) method achieved better performance than traditional statistical ones. However, ANN approaches have suffered from difficulties with generalization, producing models that can overfit the data. This paper employs a relatively new machine learning technique, support vector machines (SVM), to the customer segmentation prediction problem in an attempt to provide a model with better explanatory power. To evaluate the prediction accuracy of SVM, we compare its performance with logistic regression analysis and ANN. The experiment results with real data of insurance company show that SVM superiors to them.

Keyword : Support Vector Machine, Artificial Neural Networks, Customer Segmentation Prediction

* 상명대학교 산업정보시스템공학과

1. 서 론

급변하는 새로운 기업 환경에서 국내 기업들이 살아남기 위한 노력의 일환으로 IT시장 및 기업 마케팅의 중심 개념으로서 고객관계관리 (CRM) 바람이 거세게 불고 있다. 기업의 마케팅 개념이 공급자 중심에서 고객중심으로 전환되면서 기업의 수익을 극대화하기 위한 핵심 전략으로서 무엇보다도 고객과의 지속적인 관계유지를 통해 평생고객으로 확보하는 전략의 수립 및 추진이 요구되고 있다. 그 결과 기업 내에 축적되어 있는 정형, 비정형의 데이터를 별도의 데이터베이스로 구축하여 데이터분석 및 시뮬레이션을 실행하고 전략적 의사결정을 도출하는 시스템 구축의 절대적 필요를 가져왔고 이러한 정보 분석 기술로 데이터마이닝이 활용되고 있다.

과거 금융업계는 고유한 업무영역을 고수하면서 고객을 공유하였다. 즉 은행, 보험사, 투신사 등 각 회사들은 같은 업종 내에서만 경쟁을 하며 발전하여 왔으나, 시장개방화가 촉진되고 업무영역에 대한 규제가 완화되면서 이들 업종간의 고객 쟁탈전은 더욱 심화되고 있으며, 보험업계는 신규고객의 유치보다는 원가 효율이 높은 기존고객의 이탈을 방지하는 측면에 한층 전략적 무게를 두고 접근하고 있다.

보험회사가 시장점유율 혹은 매출을 유지, 증가시키기 위해서는 기존 고객의 이탈을 방지해야 한다. 그러나 보험업의 특성인 인적판매방식의 계속되는 생산성 저하는 고객의 이탈을 부추기고 있다. 우리나라 보험회사의 특징은 과도할 정도로 보험 설계사에 의존하고, 설계사의 양적 팽창에 보험사들이 심혈을 기울여 왔기 때문이다. 무리한 설계사의 양적인 팽창은 필연적으로 설계사의 대량유입 및 대량이탈로 이어지고 설계사에 대한 교육비용의 지출이 증가하는 한편 생산성은 점차 하락하는 현상이 발생하고 있다. 이러한 악순환 속에서 기업이 택한 방법은 기존의 고객들을 유지시키기 위한 노력을 강화하는 것이다. 이를 위해서는 먼저 고객에 대한 정보가 있어야 한다. 특히 기존고객의 성향을 정확히 분석하고, 이탈하는 고객을 분류해내는 작업이 필요하다.

전통적으로 데이터마이닝 기법으로는 로지스틱 회귀분석과 인공신경망 기법이 주로 이용되었다. 그러나 로지스틱 회귀분석의 경우에는 통계적 기법에 근간한 모형으로서 각 변수의 영향력을 정확하게 설명할 수 있다는 있으나, 정확도가 떨어지고, 인공신경망 모델은 정확도가 매우 우수하다는 장점이 있으나, 모형구축에 많은 시간이 소요되고, 모형의 설명력이 부족하다는 단점이 있다.

본 연구에서는 기존의 전통적인 기법들을 적용할 경우 발생하는 한계점을 최소화하기 위해, 최근들에 다양한 예측 문제 영역에 도입되어 성과가 우수하다고 알려진 SVM (support vector machine)을 보험회사의 고객이탈예측모형에 적용하고, 로지스틱 회귀분석과 인공신경망 기법을 적용하여 예측한 결과들과 비교하고자 한다.

2. 관련 이론 고찰

고객이탈예측모형에는 데이터마이닝 방법들이 적용될 수 있다. 본 장에서는 기존의 예측연구들에 적용된 다양한 데이터마이닝 기법들을 간략하게 살펴보고, 본 연구에서 제안하고자 하는 SVM에 대하여 알아보기로 한다.

2.1 로지스틱 회귀분석(Logistic Regression Analysis)

회귀분석은 목표변수가 입력변수들에 의해서 어떻게 설명 또는 예측되는지를 알아보기 위해 자료를 적절한 함수식으로 표현하여 분석하는 통계적 방법을 말한다. 이때 입력변수가 하나인 경우를 단순회귀(simple regression) 분석이라 하고, 입력변수가 여러 개인 경우를 다중회귀(multiple regression) 분석이라 한다 [1].

예를 들어 목표변수 Y에 대해 n개의 입력변수 $x_1, x_2, \ldots, x_n$이 있다면, 다중선형회귀모형은 다음과 같이 표현된다.

$$Y=\alpha+\beta_1 x_1+\beta_2 x_2,\ldots,\beta_n x_n+\xi_1 \quad (1)$$

여기서, $\alpha, \beta_1, \beta_2, \ldots, \beta_n$은 추정되어야 할 (n+1) 개의 회귀계수이며, ε은 기대값 0, 분산 σ^2을 갖는 오차항이다.

그러나 만약에 목표변수 Y의 관측값이 이항형일 때는 이러한 선형회귀모형은 문제점을 가지게 된다. 목표변수 Y의 관측값은 이항형이지만 예측값의 유형은 이항형이 아니라는 것이다. 또 다른 문제점은 목표변수 Y에 대한 확률분포가 선형회귀 모형에서 가정되는 확률분포와 맞지 않는다는 것이다. 즉 Y가 이항형이기 때문에 베르누이(Bernoulli) 분포와 같이 이진변수(binary variable)를 가지는 분포에 의해서 모형화하는 것이 타당한데, 선형회귀모형에서는 Y를 연속형인 것으로 간주하기 때문에 흔히 정규분포를 가정하고 모형을 만들게 된다.

로지스틱 회귀분석은 목표변수가 이항형일 경우 선형회귀모형의 이러한 단점을 극복하기 위해 확률에 대한 로짓변환 (logit transformation)을 고려하여 분석하는 것이다. 즉, 식 (2)와 같이 모형화하여, 모형식의 좌변과 우변이 모두 실수상의 값을 가지도록 하는 것이다.

$$\log\left[\frac{P(Y=1|x_1, x_2, \ldots, x_n)}{1-P(Y=1|x_1, x_2, \ldots, x_n)}\right]=\alpha+\beta_1 x_1+\beta_2 x_2,\ldots,\beta_n x_n \quad (2)$$

여기서 log는 자연로그를 의미한다. 로지스틱 회귀분석의 목적은 추정된 로짓모형을 이용하여 자료를 분류하기 위한 것이기 때문에, 일반적인 판별분석과 비교하여 로지스틱 판별분석(logistic discrimination)이라고 불린다. 위의 모형으로부터 추정된 회귀계수 α, β_1, β_2,..., β_n을 이용하여 다음과 같이 사후확률(posterior probability)에 대한 추정식을 얻을 수 있다.

$$\hat{P}(Y=1|x_1, x_2, \ldots, x_n)=\frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2, \ldots, \beta_n x_n)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2, \ldots, \beta_n x_n)} \quad (3)$$

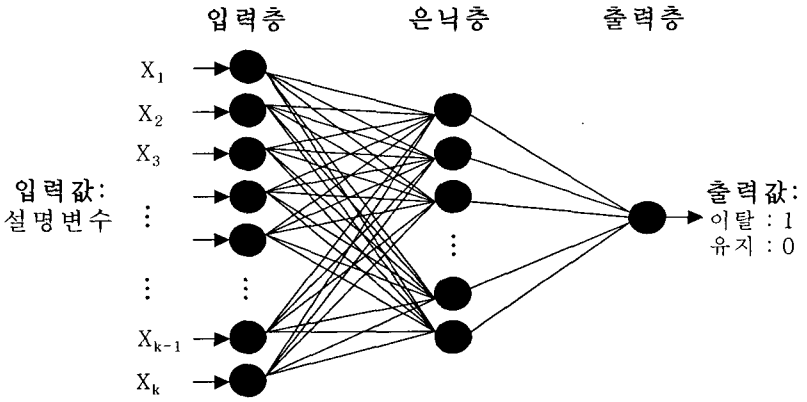
이렇게 얻어진 각 개체에 대한 사후확률은 그 개체를 분류하기 위해 사용될 수 있다($[\hat{P}(Y=1|x_1, x_2, \ldots, x_n)=1-\hat{P}(Y=1|x_1, x_2, \ldots, x_n)]$). 다시 말해 추정된 사후확률은 0과 1사이의 값을 가지게 되므로 적절한 절단값(cutoff value)을 정하여 이 값을 기준을 각 개체를 분류하는 것이다.

2.2 인공신경망

인공신경망 분석은 인간의 신경망을 모방하여 실제 자신이 가진 데이터로부터의 반복적인 학습과정을 거쳐 데이터에 숨어 있는 패턴을 찾아내는 모델링 기법이다. 인공신경망 (artificial neural networks)에 관한 연구는 뇌 신경생리학(neurophysiology)으로부터 영감을 얻어 시작되었다. 자료분석 분야에서 신경망은 복잡한 구조를 가진 자료에서의 예측문제를 해결하기 위해서 사용되는 유연한 비선형모형(nonlinear models)의 하나로 분류될 수 있다. 그러나 신경생리학과의 유사성 때문에 일반적으로 다른 통계적 예측모형에 비해 보다 흥미롭게 받아들여지고 있다. 신경망은 은닉마디(hidden units)라고 불리는 독특한 구성요소에 의해서 일반적인 통계모형과 구별되어진다. 은닉마디는 인간의 신경세포를 모형화한 것으로써, 각 은닉마디는 입력변수들의 결합(combination)을 수신하여 목표변수에 전달한다. 이 때 결합에 사용되는 계수(coefficient)들의 연결강도(synaptic weights)라고 부르며, 활성함수는 입력값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하게 된다 [1, 8, 10].

신경망의 활성함수는 대개 S자형 곡선형태를 갖는 시그모이드 함수(sigmoid function)가 사용된다. 가장 보편적으로 사용되는 활성함수는 쌍곡선탄젠트와 로지스틱 함수이고, 연결함수로는 선형함수가 주로 이용된다. 출력계층으로 연결되는 활성함수는 목표변수의 속성에 따라 정해질 수 있는데 일반적으로 두 개의 값을 갖는 이분형의 목표변수에 대해서는 로지스틱 함수를 주로 이용한다.

< 그림 1 >은 인공신경망의 구조이다. 인공신경망에서 학습 알고리즘의 기본원리는 입력층의 각 유니트에 입력자극을 주면, 이 신호는 각 유니트에서 변환되어 은닉층에 전달되고 최후에 출력층에서 결과를 출력하게 된다. 또한 관리학습(supervised learning)에서는 입력 및 원하는 출력패턴이 네트워크에 제시된다. 네트워크 입력층에 주어진 입력자극이 출력층에 전파되면서 변한 출력패턴을 목표패턴과 비교한다. 네트워크에서 출력된 패턴이 목표패턴과 일치하는 경우에는 학습이 일어나지 않으며 그렇지 않은 경우는 얻어진 출력패턴과 목표패턴의 차이를 감소시키는 방향으로 네트워크의 연결 강도를 조절하여 학습을 한다.



< 그림 1 > 인경신경망의 구조

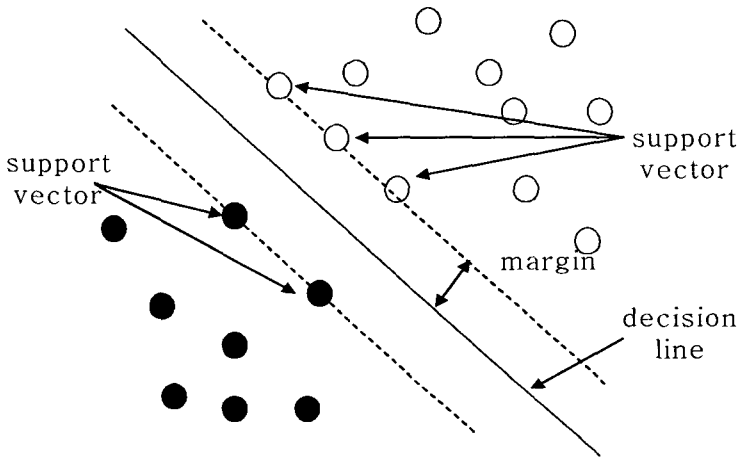
2.3 Support Vector Machine (SVM)

SVM은 1998년 Vapnik [9]에 의해 개발된 학습기법으로, 데이터를 고차원 공간으로 투사시켜 분리경계가 단순한 문제로 변환시키기 때문에 수학적으로 분석하는 것이 수월하다. 또한 SVM은 조정해야 할 모수(parameter)의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다. 그리고 구조적 위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 전역적 최적해를 구할 수 있다는 점에서 인공신경망보다 성능이 우수한 기계학습 기법으로 주목 받고 있다 [3, 4, 5, 7].

최근 인공지능 분야에서 SVM을 이용한 연구가 매우 활발히 진행되고 있으며, 우수한 성과를 나타내고 있는 것으로 알려져 있고, 또한 SVM은 인공신경망이나 판별분석, 로지스틱 회귀분석, 사례기반추론 등과 같은 다른 분류기법과 비교하여 비슷하거나 더

우수한 성능을 나타낸 것으로 보고되고 있다. 본 연구에서는 이러한 연구배경을 토대로 SVM을 고객이탈예측에 적용하는 연구를 수행하였다.

< 그림 2 >에서 보는 바와 같이 SVM 알고리즘은 Support Vector라고 불리는 training set point들에 의해 결정되는 결정 경계(decision boundary)를 결정함으로써 2개의 classification 문제에 주로 많이 사용되고 있다. 이때, 결정 경계는 두 클래스간의 최대 거리를 유지하도록 결정된다. SVM의 기본 아이디어는 구조적 리스크 최소화를 통해 벡터공간에서의 최적의 결정경계영역을 찾아내는 것으로 이진분류문제를 푸는 방법으로 이용되고 있다.



< 그림 2 > 2차원 공간에서의 SVM의 결정경계영역

SVM은 두 집단으로 구분된 입력벡터를 가지는 훈련용 자료에 대해 집단을 분류할 때 기준이 되는 분리 초평면(separating hyperplane)을 특수한 학습 알고리즘을 이용하여 찾게 되는데[2, 6], 이를 구체적으로 설명하면 다음과 같다.

두 개의 집단 $t_i \in \{-1,+1\}$으로 분리된 입력벡터 $x_i = (x_i^{(1)}, \ldots, x_i^{(n)})^T \in R^n$을 가지는 훈련 데이터셋 $D=\{x_i, t_i\}_{i=1}^{N}$이 있다고 하자. Vapnik이 최초 제안한 공식에 따르면 SVM은 다음의 조건을 만족한다[9].

$$w^T\phi(x_i) + b \geq +1, \text{ if } t_1=+1 \;\&\; w^T\phi(x_i) + b \leq +1, \text{ if } t_1 = -1 \qquad (4)$$

또는

$$t_1 = [w^T\phi(x_i)+b] \geq 1,\ I = 1, \ldots, N \quad (5)$$

여기서, w는 가중치 벡터를 나타내며, b는 편차(bias)를 나타낸다. 비선형 함수 $\phi(\cdot): R^n \rightarrow R^{nh}$은 입력벡터를 고차원의 특징공간(feature space)으로 이동시키는(mapping) 역할을 수행한다. 식 (5)는 특징공간에서 분리 초평면 $w^T\phi(x)+b=0$을 사이에 두고 반대쪽에 두 개의 평행한 경계 초평면(bounding hyperplane)을 만들게 되는데, 그 마진(margin)의 폭은 $\frac{2}{w}$가 된다. 이러한 조건을 만족하는 가중치 공간에서 분류의 결과는 식 (6)과 같은 식으로 도출된다.

$$y^t(x) = \mathrm{sgn}(w^T\phi(x)+b) \quad (6)$$

그러나 선형분리가 불가능한 문제가 대부분이므로 식 (7) 및 (8)과 같이 오분류(misclassification)를 허용할 수 있도록 오차항(ξ_i)을 도입한 후 가중치 벡터를 찾는 것이 일반적이라고 할 수 있다.

$$\min_{w,b,\xi} J(w, \xi) = \frac{1}{2}w^Tw + \sum_{i=1}^{N}\xi_i \quad (7)$$

$$t_i(w^T\phi(x_i)+b) \geq 1-\xi_i, i = 1, \ldots, N \ \&\ \xi_i \geq 0,\ i = 1, \ldots, N \quad (8)$$

여기서, w는 제약식의 부등식에서 오분류를 허용하는데 필요한 여유변수(slack variable)이며, 목적함수에 있는 $c \in R^+$는 마진폭에 대응하는 분류오차의 중요도 가중치이다. 식 (7)과 (8)로 구성된 최소화 문제는 선형 제약식을 가진 이차계획(quadratic programming: QP) 모형으로, 최적해는 라그랑지 승수 α_i를 이용하여 구하게 된다[9]. 승수 α_i는 훈련용 데이터 각각에 곱해지는데, 만약 비음인 α_i가 존재한다면 이 승수에 대응하는 데이터를 support vector라고 한다.

그러나 고차원 문제를 주로 다루는 SVM에서 ω나 $\phi(x)$를 실제로 계산할 수 없기 때문에 매핑함수인 $\phi(x)$를 식 (9)와 같은 커널함수 $K(\cdot,\cdot)$로 연결시켜 주는 Mercer의 조건을 이용하여 문제를 풀게 된다.

$$K(x_i, x_j) = \phi(x_i)^T\phi(x_j) \quad (9)$$

이와 같이 커널함수는 이론적으로는 가능하나 실제로는 풀기 힘든 매핑함수를 대신하여 원자료를 고차원 공간으로 사상시켜 특징공간 내에 선형으로 분리가능한 입력자료 집합을 만들어 주는 역할을 수행한다. 어떤 커널함수를 선택하는 것이 바람직한가는 문제의 종류에 따라 다르며, SVM을 적용하는데 가장 중요한 요소이다. 대표적인 커널함수로는 $K(x_i, x_j)=(x_i^T x_j + 1)^d$ (수준 d를 가지는 커널)과 $K(x_i, x_j)=\exp(\frac{-\|x_i-x_j\|^2}{2\sigma^2})$인 가우시안 RBF 커널이 있으며 $d\in N$와 $\in R^+$는 상수이다. 따라서 식 (6)의 분류기는 커널함수를 이용하여 식 (10)의 SVM 분류기로 최종 구성된다.

$$y^t(x) = \mathrm{sgn}(\sum_i^n \alpha_i t_i K(x, x_i) + b) \quad (10)$$

3. 보험회사 고객이탈예측 모형

최근 보험업계는 고객요구의 다양화와 더불어 규제완화에 따른 가격자유화, 종합 금융화, 신채널의 등장 등으로 치열한 경쟁을 예고하고 있다. 이러한 시장의 급속한 변화에 따라 보험회사들은 보험업계에서 경쟁우위를 확보하기 위해서 고객관계관리에 관심을 갖기 시작하였다. 본 연구에서는 A 보험회사의 2002년 초부터 2003년 말까지 12개월에 걸쳐 보유고객에 관한 데이터베이스를 확보하여, 이 데이터를 바탕으로 고객이탈분석을 수행하였다.

3.1 분석자료 및 분석방법

본 분석에 사용된 자료는 고객 데이터베이스를 이용하여 54,820개의 고객 데이터 표본을 획득하였다. 이중 보험계약의 유지고객은 34,284명(62.54%)이고, 이탈고객은 20,536명(37.46%)이다. 목표변수는 이탈고객은 1의 값을 주고 유지 고객은 0의 값을 갖는 변수로 설정하였다. 설명변수로는 인구통계학적인 변수들과 보험 가입시 작성한 가입신청서를 기초로 한 변수들과 거래를 통해 얻어진 변수들로 구성되었다. 설명변수는 총 50개로 연속형 변수와 명목형 변수가 함께 사용되었다 (< 표 1 > 참조). 분석 데이터를 train data와 test data로 분할할 경우에는 train data와 test data의 유지고객과 이탈고객의 데이터들의 구성비를 유사하게 하였다.

< 표 1 > 분석에 사용된 설명변수

1. 계약자 성별	26. 약관대출유무
2. 계약자 연령	27. 최종약관대출액
3. 결혼 유무	28. 약관대출횟수
4. 거래시기와의 거리	29. 최근 1년간 약관대출횟수
5. 상품구분	30. 직전 6개월간 약대유무
6. 납입방법	31. 직전 12개월간 약대유무
7. 수금방법	32. 약대이자 연체유무
8. 납입기간	33. 약대이자 연체일수
9. 보험기간	34. 직전 6개월간 약대이자연체유무
10. 특약유무	35. 직전 12개월간 약대이자연체유무
11. 특약갯수	36. 최종연체보험료
12. 주보험금	37. 최종연체이자
13. 진단코드	38. 직전 6개월간 연체유무
14. 만기환급형	39. 직전 12개월간 연체유무
15. 주피보험자 위험등급	40. 모집설계사와의 거래기간
16. 주피보험자 운전여부	41. 수금사원변경횟수
17. 납입횟수_총횟수	42. 수금사항변경횟수
18. 납입횟수_할인횟수	43. 총납입 보험료
19. 모집자_수금자동일여부	44. 최근 1년동안의 총납입보험료
20. 지급	45. 합계보험료
21. 변경	46. 직전 6개월간 보험료 연체횟수
22. 계약변경여부	47. 직전 12개월간 보험료연체횟수
23. 부활계약여부	48. 대출금총액
24. 부활 횟수	49. 대출금상환액
25. 계약자부활경험의 유무	50. 대출금잔액

SVM의 성과비교를 위한 로지스틱 회귀분석의 경우, 학습을 위한 데이터는 전체데이터의 80%를 사용하였으며, 검증용 데이터는 20%를 사용하였다. 한편 과적합 현상을 막기 위해 테스트용 데이터의 별도로 필요로 하는 인공신경망 기법의 경우에는 학습: 테스트: 검증용 데이터가 60%:20%:20%의 비율이 되도록 설계하여 실험하였다.

또한, 로지스틱 회귀분석의 경우 단계별 로지스틱 분석 모형의 전진선택방법을 활용하였으며 단계별 선택의 확률은 진입 0.1, 제거 0.05로 하였다.

인공신경망 모형은 은닉층이 한개인 3층 구조인 네트워크 모형을 선정하였고, 은닉층의 노드 수는 입력변수의 개수를 n개라고 할 때, $\frac{1}{2}n, n, \frac{3}{2}n, 2n$의 총 4가지로 실험하였다. 기타 인공신경망 모형의 설정으로는 학습률 0.1, 관성률 0.05이었으며, 학습 중단 조건은 최소평균오차를 기록한 이후 50,000회로 하였다.

SVM의 경우, 본 연구에서는 SVM의 커널 함수로서 널리 사용되는 가우시안 RBF를 사용하였다. SVM의 성능에 있어서 α_i의 상한 C와 σ^2이 중요한 역할은 한다. 따라서 본 연구에서는 SVM의 파라미터에 대해 다양한 값을 대입하여 모형을 변형하여 실험하였다.

3.2 분석결과

본 절에서는 SVM의 실험결과를 각 커널변수와 파라미터에 따라 실험한 결과와, 추가적으로 비교기법인 로지스틱 회귀분석과 인공신경망의 실험결과와 비교하였다.

본 연구에서 적용한 가우시안 RBF에서는 SVM의 성능에 영향을 미치는 α_i의 상한 C와 σ^2을 동시에 고려하여야 하는데, C의 적합한 값이 범위는 10에서 100사이로 알려져 있고, σ^2의 범위값으로 1에서 100사이로 알려져 있다. 이를 기반으로 본 연구에서는 C와 파라미터 σ^2의 값을 세분화하여 실험하였다. C의 경우에는 1, 10, 30, 50, 70, 100의 6가지로 구분하여 실험하였고, σ^2의 값은 1, 25, 50, 70, 100의 5가지 경우를 실험하였다.

본 연구에서는 SVM의 다양한 실험중 예측력이 가장 우수한 경우의 SVM의 실험값을 로지스틱 회귀분석과 인공신경망의 실험결과값과 비교하여 < 표 2 >에 나타내었다.

< 표 2 > 적용 기법들간의 예측력 비교

구분	로지스틱 회귀분석	인공신경망	SVM	SVM 설정
훈련데이터	75.43%	78.93%	82.35%	σ^2=1, C=10 설정시 예측력 가장 우수
검증데이터	64.54%	75.37%	77.25%	

이 결과에서 알 수 있듯이 SVM이 가장 우수한 성과를 보임을 확인할 수 있었다. 그러나, < 표 2 >에서 보는 바와 같이 인공신경망에 비해, SVM은 훈련데이터를 이용한 학습능력에서는 매우 우수하지만, 검증데이터를 이용한 경우의 예측정확성은 훈련데이터에 비해 상대적으로 떨어지는 것을 확인할 수 있었다.

4. 결 론

본 연구에서는 최근 패턴인식 및 분류문제와 관련하여 활발하게 연구되고 있는 SVM을 보험회사의 고객이탈예측모형에 적용하여 보았다. SVM은 통계적 이론을 기반으로 하여 설명력이 우수하고, 구조적 위험 최소화 접근에 따라 과대적합문제에서 벗어날 수 있으며, 볼록 집합을 실행가능 영역으로 하는 최적화 기법을 사용하기 때문에 유일한 최

적해를 구할 수 있다는 점에서 관심을 받고 있다. 특히 기본의 데이터마이닝 기법에서 많이 사용하는 로지스틱 회귀분석과 인공신경망과 비교하여 볼 때, 그 성과가 우수함을 확인할 수 있었다. 특히 인공신경망에서 과대적합을 방지하기 위해 수많은 시행착오를 거쳐 아키텍처를 설계해야 하는 번거로움 없이도 커널함수의 선택과 기본적인 모수값 설정만으로도 인공신경망의 성과보다 우수한 예측정확성을 보였다.

그러나, SVM은 이러한 장점에도 불구하고 본 연구결과에서 확인한 바와 같이 우수한 학습능력에 비해 새로운 테이터 판단에 대한 예측력이 상대적으로 덜 우수함을 확인할 수 있었다. 따라서 이에 대한 향후 추가적인 연구가 필요하다고 판단되며, 다른 알고리즘과의 통합 등으로 그 성능을 개선하는 연구가 필요하다.

5. 참 고 문 헌

[1] Bell, T., G. Ribar, and J. Verchio, "Neural nets vs. logistic regression: a comparison of each model's ability to predict commercial bank failures," Proceedings of the 1990 Deloitte & Touche/University of Kansas Symposium on Auditing Problems, 1990, pp. 29-58.

[2] Cristianini, N., and J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge, England: Cambridge University Press, 2000.

[3] Drucker, H., D. Wu, and V.N. Vapnik, "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, Vol. 10, No. 5, 1999, pp. 1048-1054.

[4] Fan, A., and M. Palaniswami, "Selecting bankruptcy predictors using a support vector machine approach," Proceedings of the International Joint Conference on Neural Networks, 2000.

[5] Gunn, S.R., Support Vector Machines for Classification and Regression, Technical Report, University of Southampton, 1998.

[6] Hearst, M.A., S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intelligent System, Vol. 13, No. 4, 1998, pp. 18-28.

[7] Hsu, C.-W., C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," Technical Report, Department of Computer Science and Information Engineering, National Taiwan University.

[8] Smith, M., Neural Networks for Statistical Modeling, NY: Van Nostrand Reinhold, 1993.

[9] Vapnik, V., Statistical Learning Theory, Springer, New York, 1998.

[10] Zhang, G.P., "Neural Networks for Classification: A Survey," IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 30, No. 4, 2000, pp. 451-462.

저 자 소 개

서 광 규 : 고려대학교 산업정보시스템공학과에서 박사학위취득,
한국과학기술연구원(KIST) 시스템연구부 연구원으로 재직,
현재 상명대학교 산업정보시스템공학과 교수로 재직중.
관심분야는 데이터마이닝과 CRM, 정보시스템, 생산시스템, e-business 등이다.