한국지능정보시스템학회논문지 제11권 제3호 2005년 12월(pp.69~81)

Support Vector Machine을 이용한 고객구매예측모형

안현철
한국과학기술원 테크노경영대학원
(hcahn@kaist.ac.kr)

김경재
동국대학교 경영대학 정보관리학과
(kjkim@dongguk.edu)

한인구
한국과학기술원 테크노경영대학원
(jghan@kgsm.kaist.ac.kr)

고객관계관리는 치열한 경쟁환경에서 각 기업이 생존하기 위해 반드시 필요한 하나의 기업전략이 되었다. 고객관계관리의 방법은 다양하지만 가장 기본적인 방법은 특정 고객이 어떤 상품 혹은 상품군을 구매할 것인지를 정확히 예측하는 것이다. 이미 국내외 실무현장에서 전통적인 데이터마이닝 기법을 활용한 고객구매예측모형이 널리 적용되고 있다. 하지만 전통적인 기법의 경우, 정확도가 상대적으로 떨어지거나 혹은 모형의 구축 및 유지관리가 어렵다는 문제가 종종 제기되어 왔다. 이에 본 연구에서는 기존 모형의 문제점을 개선하기 위한 대안으로, 매우 높은 예측력을 나타내면서 동시에 일반화 능력이 우수한 것으로 알려진 Support Vector Machine(SVM)을 이용하여 고객구매예측모형을 구축하고자 한다. 본 연구에서는 고객구매예측의 도구로써 SVM의 적합성을 판단하기 위하여 전통적인 기법인 로지스틱 회귀분석, 인공신경망과 그 성과를 비교하였다. 그 결과, SVM이 다른 기법들에 비해 상대적으로 우수한 성과를 나타냄을 확인할 수 있었다.

논문접수일 : 2005년 5월 게재확정일 : 2005년 12월 교신저자 : 김경재

1. 서론

최근 시공간의 제약을 줄여주고, 상대적으로 고객으로부터 풍부한 정보를 확보할 수 있는 인터넷 매체의 확산이 본격화됨에 따라, 이른바 '대량 맞춤(mass customization)'이 기업 경쟁력 강화에 있어 중요한 이슈로 대두되고 있다. 이러한 조류의 하나로 인터넷을 기반으로 한 고객관계관리인 e-CRM에 대한 세간의 관심이 점점 더 높아지고 있다. e-CRM의 여러 분야 중에서도 각 고객이 자사의 특정 상품 혹은 상품군의 구매와 관련해 관심이나 호응을 갖고 있는 고객인지, 아닌지를 분류하는 '고객구매예측모형'은 오늘날 마케팅 분야에서 매우 중요한 이슈 중 하나로 자리매김하고 있다. 고객의 구매행동을 정확하게 파악할 수 있는 능력을 기업이 보유하고 있는 경우, 그 기업은 이를 이용해 다양한 사업기회를 발굴, 육성할 수 있기 때문이다(Chiu, 2002).

전통적인 고객구매예측모형에는 로지스틱 회귀분석, ANN(artificial neural network), CBR(case-based reasoning) 등과 같은 데이터마이닝 기법들이 주로 적용되어 왔다. CBR의 경우, 구현이 용이하고, 실시간으로 항상 새로운 구매 정보가 갱신된다는 측면에서 장점이 있고, 로지스틱 회귀분석의 경우 통계적 기법에 근간한 모형으로서 각 변수의 영향력을 정확하게 설명할 수 있다는 장점이 있으

며, ANN의 경우 정확도가 매우 우수하다는 장점이 있다. 하지만, CBR과 로지스틱 회귀분석은 종종 예측 성과가 높지 않을 때가 발생한다는 단점이, ANN은 모형 구축에 많은 시간이 소요되며, 모형의 설명력이 매우 부족하다는 단점이 있다.

이처럼 기존 전통적인 기법들을 적용할 경우 발생하는 한계점을 최소화하기 위해, 본 연구에서는 최근 새롭게 각광 받고 있는 SVM(support vector machine)을 고객구매 예측에 적용하고자 한다. SVM은 Vapnik에 의해 제안된 학습이론으로 분류문제를 해결하기 위해 최적의 분리 경계면을 제공한다(Vapnik, 1995). SVM이 주목 받는 이유는 첫째, 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내고, 셋째, 적은 학습자료만으로도 신속하게 분류학습을 수행할 수 있기 때문이다. 또한 SVM은 기존의 학습 알고리즘이 경험적 위험 최소화 원칙(empirical risk minimization)을 구현하는 것인데 비해 구조적 위험 최소화 원칙(structural risk minimization)에 기반하므로 인공신경망에서 흔히 발생하는 과대적합현상을 피할 수 있다(박정민 등, 2005).

이런 장점을 바탕으로 경영학 분야에서도 Tay & Cao(2002), Kim(2003), Huang et al.(2005) 등이 SVM을 주가지수 예측에 응용하였으며, Huang et al.(2004), Shin et al.(2005), 박정민 등(2005), 민재형과 이영찬(2005) 등은 기업신용평가를 위한 모형에 응용하였다. 본 연구에서는 경영학 분야에서 성공적으로 활용되고 있는 SVM을 지금까지 적용되지 않았던 새로운 분야인 고객구매예측모형 구축에 적용하여 그 응용가능성을 확인하고자 한다.

본 연구는 총 5장으로 구성하였다. 1장에서는 연구의 내용과 목적을 간단히 소개하고, 2장에서는 기존의 고객구매모형에 적용된 기법들과 SVM에 관해 알아보도록 한다. 이어서 3장에서는 고객구매예측을 위한 SVM모형을 제안하고, 4장에서는 3장의 실험 결과를 비교대상 기법들의 실험결과와 함께 제시한다. 끝으로 5장에서는 본 연구의 시사점과 한계점을 제시한다.

2. 선행연구

고객구매예측에 관해서는 다양한 데이터마이닝 방법들이 적용될 수 있다. 본 장에서는 기존 연구들에 적용된 다양한 데이터마이닝 기법들을 간략하게 살펴보고, 본 연구에서 제안하고자 하는 Support Vector Machine에 대해 알아보도록 한다.

2.1 로지스틱 회귀분석

분류를 하는데 있어서, 가장 흔한 경우는 이분법 (二分法)을 기준으로 분류하는 경우다. 예를 들어, 인터넷 쇼핑몰의 경우에도, 특정 고객이 물건을 살 것인가, 말 것인가 혹은 특정 이벤트에 고객이 반응을 할 것인가, 말 것인가 하는 문제들과 같이 이분법을 필요로 하는 분류 문제들의 해결이 종종 요구되는 것이다. 그런데, 이처럼 이항 확률을 가진 종속변수를 통계적으로 설명하고자 할 때, 일반적인 회귀 분석은 적용하기에 어려움이 있다. 왜냐하면 일반적인 회귀분석에서는 종속변수를 연속변수의 형태를 가정하는데 반해, 이러한 분류의 문제에서는 종속변수가 이항변수이므로 오차의 분포가 정규분포를 한다는 회귀분석의 일반적인 가정이 위배되기 때문이다. 또한 일반적인 회귀분석을 사용할 경우에는 예상값이 0과 1 사이에만 국한되지 않는다는 문제점 역시 발생할 수 있다. 이러한 문제들을 해결하기 위해 사용되는 기법이

바로 로지스틱 회귀분석(Logistic Regression; LOGIT) 기법이다.

로지스틱 회귀분석은 기업부도예측, 주가지수 예측 등 다양한 경영학 분야에 적용되어 왔다 (Ohlson, 1980; Fanning and Cogger, 1994; Barniv et al., 1997; Zhang et al., 1999 참고). 로지스틱 회귀분석을 적용해 고객구매예측모형을 구축한 사례는 많지 않으며, 대표적으로 안현철과 한인구 (2002)의 연구를 들 수 있다.

2.2 인공신경망(ANN)

인공신경망은 생물학적인 뇌의 작동 원리를 그대로 모방하는 방법으로, 그 강력한 예측력과 범용성으로 인해 오늘날 예측, 분류, 군집분석 등 다양한 용도에 적용되고 있는 데이터마이닝 기법이다. 경영학 분야에서도 인공신경망은 다양하게 활용되고 있는데, 초기의 대표적인 연구로는 Odom and Sharda(1990), Tam and Kiang(1992), Fletcher and Goss(1993) 등이 있다.

인공 신경망은 (1) 복잡하고, 비선형적인 자료에서 지식이나 패턴을 추출할 수 있고, (2) 입력-출력 사상 (input-output mapping) 기법이라서 자료에 대한 통계적인 분석 없이 결정을 수행할 수 있으며, (3) 상대적으로 적응력 (adaptability) 이 뛰어나고 견고한 (robust) 모형이라는 점이 장점이다. 하지만 모형이 제시하는 결과에 대해서 왜 그런 결과가 나오는지에 대한 원인을 명쾌하게 설명할 수 없다는 점과 과도하게 학습을 진행할 경우, 전체적인 관점에서의 최적해가 아닌 지역 내 최적해가 선택될 수 있다는 과적합화 (overfitting) 문제는 인공 신경망 기법의 치명적인 단점이라고 할 수 있다 (Berry & Linoff, 1997).

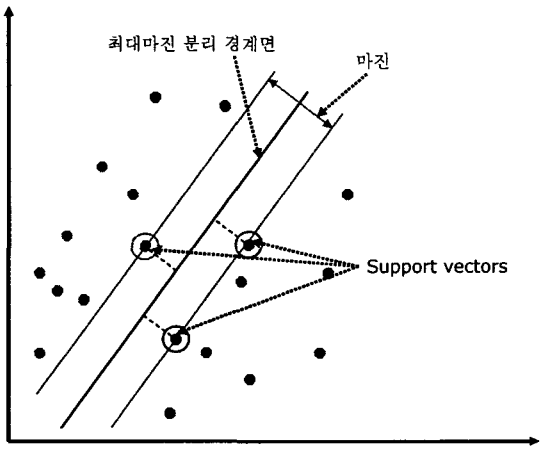
2.3 Support Vector Machine(SVM)

SVM은 통계학자인 Vapnik(Vapnik, 1995)에 의해 개발된 학습기법으로, 입력공간과 관련된 비선형문제를 고차원의 특징공간의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 수월하다 (Hearst et al., 1998). 또한, SVM은 조정해야 할 파라미터의 수가 많지 않아 비교적 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다. 그리고 구조적위험을 최소화함으로써 과대적합문제에서 벗어날 수 있으며, 볼록함수를 최소화하는 학습을 진행하기 때문에 전역 최적해를 구할 수 있다는 점에서 인공신경망보다 성능이 좋은 기계학습기법으로 주목 받고 있다.

이러한 장점을 기반으로 최근 몇 년간 SVM을 사용한 다양한 연구가 진행되었다. 그 예로서 SVM은 문서분류, 영상인식, 문자인식 등에서 뛰어난 일반화 성능을 보여주었다(Osuna et al., 1997; Joachims, 1998). 경영학 분야에서는 주로 재무예측을 위해 SVM을 활용한 연구가 진행되어 왔는데 대표적인 연구로는 주가지수 예측에 응용한 Tay & Cao(2002), Kim(2003), Huang et al.(2005)의 연구가 있으며, 기업신용평가에 적용한 연구로 Huang et al.(2004), Shin et al.(2004), 박정민 등(2005), 민재형과 이영찬(2005) 등의 연구가 있다.

SVM은 훈련데이터들을 서로 다른 두 개의 Class로 분류할 때 분류의 기준이 되는 분리 경계면(hyperplane)을 학습 알고리즘을 이용하여 찾는 원리로 이루어진다(이수용과 이일용, 2002). 따라서, SVM은 입력벡터 x를 고차원의 특징공간 (high-dimensional feature space)으로 사상 시킨 후 두 분류집단 사이의 여백 (margin)을 최대화시키는 분리 경계면를 찾는 것을 목적으로 한다. 이

러한 최대마진 분리 경계면(maximum margin hyperplane)은 두 분류 사이를 최대 거리로 분리한다. 이때 최대마진 결정함수에 가장 근접한 훈련 데이터를 support vector라고 부른다. 최대마진 분리 경계면과 support vector의 예는 [그림 1]과 같다.



[그림 1] Support vector의 예시

전술한 SVM의 개념을 수식으로 간단히 설명하면 다음과 같다. 먼저, 선형분리문제에서 속성변수가 3개인 경우 분리 결정함수는 식(1)과 같다:

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \qquad (1)$$

여기서 y는 출력값이고, x_i는 애트리뷰트값, 그리고 4개의 w_i는 학습 알고리즘에 의해 학습된 가중치이다. 상기 식에서 가중치 w_i는 결정함수를 결정하는 파라미터이다.

최대마진 결정함수는 support vector를 사용해서 식(2)와 같이 나타낼 수 있다.

$$y = b + \sum\alpha_i y_i \mathbf{x}(i) \cdot \mathbf{x} \qquad (2)$$

여기서, y_i는 훈련데이터 x(i)의 분류값이고, ·는 내적(dot product)이다. 벡터 x는 검증데이터를 나타내고, 벡터 x(i)는 support vector를 나타낸다. 이 식에서, b와 α_i는 결정함수를 결정하는 파라미터이다. support vector를 찾아내고, 파라미터 b와 α_i를 결정하는 것은 선형적으로 제약된 이차계획문제(linearly constrained quadratic programming)를 푸는 것과 같다.

앞에서 언급한 바와 같이, SVM은 입력변수를 고차원의 특징 공간으로 이동시킴으로써 비선형 분류문제를 선형모델로 구현한다. 비선형 분류문제에서 식(3)과 같은 고차원 버전은 다음과 같이 간단하게 나타낼 수 있다.

$$y = b + \sum\alpha_i y_i K(\mathbf{x}(i),\mathbf{x}) \qquad (3)$$

식(3)에서 함수 $K(\mathbf{x}(i),\mathbf{x})$는 커널함수(kernel function)라고 정의된다. 커널함수는 원래 데이터를 고차원 공간으로 사상시킴으로써 특징공간 내에 선형으로 분리가능한 입력 데이터셋을 만든다. 커널함수를 선택하는 것은 문제에 따라 다르며, SVM을 적용하는 데 있어서 가장 중요한 요소이다. 일반적인 커널함수의 예로는 다항식 커널(polynomial kernel)과 가우시안 RBF(Gaussian radial basis function)를 들 수 있다:

$$\text{가우시안 RBF} : K(x,y)=\exp(-1/\delta^2(x-y)^2) \qquad (4)$$

$$\text{다항식커널함수}: K(x,y)=(xy+1)^d \qquad (5)$$

여기서 d는 다항식 커널의 차수이고, δ^2은 가우시안 RBF 커널의 대역폭이다. 분리가능한 문제에 있어서 상기 식의 계수 α_i의 하한은 0이다. 분리가불가능한 문제에서 SVM은 계수 α_i의 하한 이외에

상한 C를 추가함으로써 일반화된 결과를 얻을 수 있다(Kim, 2003; 박정민 등, 2005).

3. 실증연구: 자료수집과 변수선정

고객구매예측모형을 SVM을 이용해 구축하고, 그 성과를 기존의 다른 기법과 비교해 보기 위해, 본 연구에서는 실제 데이터에 모형을 적용, 그 결과를 도출한다. 데이터는 G 인터넷 쇼핑몰의 구매 데이터로서, 이 쇼핑몰은 국내 최대 규모의 다이어트 전문 인터넷 쇼핑몰이다. 오프라인에서 이미 다이어트 분야에 확고한 입지를 가지고 있는 P사가 운영하고 있는 G 사이트는 방대한 컨텐츠와 우수한 서비스, 그리고 운영사의 신뢰도 높은 브랜드 이미지로 인해 국내 다이어트 인터넷 쇼핑몰 분야에서 단연 선두를 달리고 있는 전문 인터넷 쇼핑몰이다.

G 사이트로부터 본 연구의 모델링을 위해 확보한 데이터는 2001년 5월부터 8월 사이에 구매한 총 3298명의 회원의 4353건에 대한 구매 내역 데이터이다. 가용한 변수는 전처리 이후를 기준으로 해서 총 46개인데, 이 중 41개의 변수가 명목형(nominal) 변수이고, 나머지 5개가 비율(ratio)척도로 된 변수이다. 데이터는 상품과 관련된 정보를 비롯해, 구매자의 연령, 체중, 키, 기타 건강 정보 등 고객의 개인적인 특성과 관련된 정보가 대부분을 이루고 있다. 예측의 대상이 되는 변수는 고객이 각 상품군 i를 구매한 이력이 있는지, 없는지를 토대로 생성하였으며, 이 때, 상품군의 구분은 현재 온라인 G 다이어트 쇼핑몰에서 대분류로 사용하고 있는 4가지 상품군 분류 체계를 그대로 도입해 사용하였다. 각 상품군의 명칭과 특성은 <표 1>과 같다.

<표 1> 상품군 분류 체계

번호	상품군에 대한 설명
1	**다이어트큐리닉상품군** : 제품과 상담 서비스가 하나의 프로그램 형태로 제공되는 다이어트 종합 솔루션 제공 상품으로서, 고가 제품이 주류임
2	**다이어트식품상품군** : 식사 대신 먹을 수 있는 다이어트 식품이나 다이어트에 도움을 주는 기타 보조 식품들이 여기에 해당됨
3	**다이어트운동상품군** : 다이어트를 위한 각종 운동 기구들, 측정 도구들이 해당되는 상품군
4	**다이어트용품상품군** : 기타 다이어트에 활용될 수 있는 액세서리, 도서 및 음반, 패션 용품 등 나머지 부류의 다이어트 보조 기구들이 이 상품군에 해당됨

SVM의 성과 비교를 위한 로지스틱 회귀분석의 경우, 모형구축을 위한 데이터는 전체 데이터의 80%를 사용하며, 검증용 데이터는 나머지 20%를 사용한다. 한편 과대적합(overfitting) 현상을 방지하기 위해 테스트용 데이터를 별도로 필요로 하는 인공 신경망 기법의 경우에는 학습, 테스트, 검증용 데이터가 각각 6:2:2의 비중이 되도록 설계하여 실험한다.

또한 로지스틱 회귀분석의 경우, 모형에 적합한 변수를 탐색하기 위해 단계별 로지스틱 회귀분석(stepwise logistic regression) 모형의 전진선택방법 (forward-conditional) 을 활용하며 단계별 선택의 확률은 진입 0.05, 제거 0.10이다.

인공신경망 모형은 일반적으로 가장 많이 이용하는 3층 구조의 은닉층이 1개인 네트워크 모형을 기준 모형으로 활용하는데, 은닉층 노드의 개수는 입력변수의 개수를 n이라 할 때, $\frac{1}{2}n$, n, $\frac{3}{2}n$, $2n$의 총 4가지 경우를 대상으로 실험한다. 기타 인공

신경망 모형의 설정으로는 학습률 0.1, 관성률 0.1이며, 학습중지조건은 최소평균오차를 기록한 이후 400,000회로 한다.

SVM의 경우, 본 연구에서는 SVM의 커널함수로서 가장 널리 사용되는 다항식 커널과 가우시안 RBF를 사용한다. 그런데, SVM의 성능에 있어서 커널함수의 상한 C와 커널 파라미터 δ^2, d가 중요한 역할을 한다고 보고하고 있다(Tay & Cao, 2002; Kim, 2003). 본 연구에서도 SVM의 파라미터에 대해 제시된 일반적인 가이드를 따라 보고된 범위 내에서 다양한 값을 대입하여 모형을 변경시킨다. SVM 실험은 LIBSVM(Chang & Lin, 2001)을 사용한다.

4. 실증연구 결과

본 장에서는 SVM의 실험결과를 각 커널함수와 파라미터에 따라 정리해보고, 추가적으로 비교대상 기법인 LOGIT, ANN의 실험결과와 비교해 보도록 한다.

전술한 바와 같이 SVM 모형은 크게 선형 SVM과 비선형 SVM으로 나누어 볼 수 있다. 선형 SVM의 장점 중 하나는 조정해야 할 파라미터가 상수 C 이외에는 존재하지 않는다는 점이다. 그러나 선형 SVM으로 분리되지 않는 훈련 데이터인 경우에는 계수 α_i의 상한인 C가 예측력에 영향을 미친다. 비선형 SVM인 경우에는 커널 파라미터도 조정해야 한다.

본 연구에서는 선형 SVM과 비선형 SVM의 모든 경우를 실험한다. 비선형 SVM의 경우, 가우시안 RBF 그리고 다항식 함수를 커널 함수로 적용해 본다. 특히 이 경우, 상한 C와 커널 파라미터를 변경하면서 실험을 진행한다. 가우시안 RBF에서는 C 이외에 δ^2을 고려해야 한다. Tay and Cao(2002)에 따르면, 적절한 δ^2의 범위는 1에서 100사이이고, C의 값으로 적합한 범위는 10에서 100사이라고 한다. 이를 참고하여 본 연구에서도 C와 파라미터의 값을 세분화하여 실험한다. C의 경우 1, 10, 33, 75, 78, 100의 총 6가지로 구분해 실험하였으며, δ^2의 경우에는 1, 25, 50, 75, 100의 5가지 경우로 나누어 실험한다. 또한 다항식 커널함수에 대해서는 1차식~5차식의 5가지 경우를 모두 고려해 실험을 진행하며, ε은 0.001로 고정한다. <표 2>~<표 4>는 SVM의 다양한 커널함수 및 그의 파라미터에 대한 SVM의 예측력을 나타내고 있다.

<표 2> 선형 SVM 적용시 실험 결과

C	상품군 1		상품군 2		상품군 3		상품군 4	
	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터
1	72.3404%	66.9492%	64.7551%	60.9524%	57.7091%	59.0190%	56.4426%	55.6452%
10	75.1064%	70.3390%	64.7551%	61.4286%	58.0262%	59.0190%	56.4426%	55.6452%
33	75.5319%	69.4915%	64.5161%	61.4286%	58.3036%	58.8608%	56.4426%	55.6452%
55	75.5319%	68.6441%	64.5161%	61.4286%	58.2243%	58.8608%	56.4426%	55.6452%
78	75.1064%	69.4915%	64.6356%	61.4286%	58.1847%	58.7025%	56.4426%	55.6452%
100	74.8936%	68.6441%	64.6356%	61.4286%	58.1451%	60.1266%	56.4426%	55.6452%

<표 3> 비선형 SVM / RBF 커널 적용시 실험 결과

C	상품군 1		상품군 2		상품군 3		상품군 4	
	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터
(a) δ^2=1								
1	75.7447%	69.4915%	66.0693%	54.7619%	57.8280%	60.6013%	59.8787%	55.2419%
10	80.2128%	73.7288%	68.6977%	56.1905%	60.9195%	61.7089%	60.8388%	56.0484%
33	81.4894%	72.0339%	69.1756%	55.7143%	62.3464%	61.8671%	61.4957%	55.4435%
55	82.5532%	69.4915%	69.6535%	55.7143%	62.3860%	61.7089%	61.5968%	55.4435%
78	82.1277%	69.4915%	70.0119%	57.6190%	62.5842%	62.6582%	61.9505%	56.2500%
100	82.7660%	71.1864%	70.0119%	57.6190%	62.9409%	62.1835%	62.0010%	56.6532%
(b) δ^2=25								
1	69.3617%	66.9492%	63.9188%	61.4286%	57.3524%	59.6519%	56.4426%	55.6452%
10	72.1277%	68.6441%	64.0382%	59.5238%	56.9560%	59.3354%	57.6554%	55.8468%
33	74.2553%	69.4915%	64.0382%	60.0000%	57.1938%	58.8608%	58.1102%	56.6532%
55	75.5319%	69.4915%	64.2772%	60.0000%	57.1542%	59.3354%	58.2617%	56.4516%
78	75.7447%	70.339%	63.5603%	60.0000%	57.2335%	59.3354%	58.2617%	56.4516%
100	76.5957%	69.4915%	63.5603%	60.0000%	57.4316%	59.3354%	58.2617%	56.4516%
(c) δ^2=50								
1	67.8723%	68.6441%	63.9188%	60.0000%	57.5902%	58.0696%	56.4426%	55.6452%
10	70.8511%	71.1864%	64.1577%	60.9524%	57.2731%	59.8101%	56.4426%	55.6452%
33	72.7660%	67.7966%	64.0382%	60.0000%	56.9560%	59.1772%	57.2006%	55.8468%
55	73.8298%	69.4915%	63.2019%	58.0952%	56.7975%	58.8608%	58.3123%	56.0484%
78	73.1915%	70.3390%	63.7993%	60.0000%	56.7578%	58.8608%	57.6554%	56.4516%
100	73.8298%	71.1864%	64.0382%	59.5238%	56.8371%	59.1772%	57.9586%	56.4516%
(d) δ^2=75								
1	67.0213%	66.9492%	62.9630%	60.0000%	57.0749%	58.2278%	56.4426%	55.6452%
10	70.4255%	71.1864%	64.1577%	60.9524%	57.2731%	59.8101%	56.4426%	55.6452%
33	72.5532%	68.6441%	63.6798%	60.0000%	57.3127%	59.8101%	56.4426%	55.6452%
55	73.1915%	67.7966%	63.5603%	60.9524%	57.1145%	59.3354%	56.4426%	55.6452%
78	73.4043%	68.6441%	64.1577%	60.0000%	56.7578%	58.7025%	57.5038%	56.0484%
100	72.9787%	70.339%	63.9188%	59.5238%	56.8371%	58.8608%	58.3628%	56.0484%
(e) δ^2=100								
1	67.0213%	66.1017%	63.4409%	61.4286%	56.5993%	58.3861%	56.4426%	55.6452%
10	70.6383%	70.3390%	64.5161%	60.9524%	57.2335%	59.6519%	56.4426%	55.6452%
33	71.4894%	66.9492%	63.9188%	60.9524%	57.2731%	59.8101%	56.4426%	55.6452%
55	72.7660%	67.7966%	63.6798%	60.0000%	57.3127%	59.8101%	56.4426%	55.6452%
78	72.7660%	68.6441%	63.6798%	60.0000%	57.1542%	59.6519%	56.4426%	55.6452%
100	73.4043%	68.6441%	63.4409%	60.9524%	57.1145%	59.3354%	56.4426%	55.6452%

<표 4> 비선형 SVM / 다항식 커널 적용시 실험 결과

C	상품군 1		상품군 2		상품군 3		상품군 4	
	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터	훈련데이터	검증데이터
(a) 1차식								
1	70.2128%	70.3390%	64.3967%	61.4286%	57.6298%	58.7025%	56.4426%	55.6452%
10	71.9149%	68.6441%	64.7551%	60.9524%	58.4621%	58.3861%	56.4426%	55.6452%
33	74.0426%	69.4915%	64.7551%	61.4286%	58.3432%	59.0190%	56.4426%	55.6452%
55	75.1064%	69.4915%	64.7551%	61.4286%	58.6207%	59.1772%	56.4426%	55.6452%
78	75.1064%	70.3390%	64.7551%	61.4286%	58.6603%	58.8608%	56.4426%	55.6452%
100	74.6809%	68.6441%	64.6356%	61.4286%	58.4621%	59.3354%	56.4426%	55.6452%
(b) 2차식								
1	70.0000%	71.1864%	63.5603%	60.0000%	57.0353%	58.8608%	56.5942%	55.6452%
10	74.0426%	66.9492%	63.3214%	59.5238%	56.6786%	58.7025%	58.3123%	56.4516%
33	76.8085%	68.6441%	63.6798%	60.0000%	56.6786%	58.7025%	58.2617%	56.4516%
55	77.6596%	67.7966%	63.4409%	59.5238%	56.6786%	58.7025%	58.2617%	56.4516%
78	77.6596%	68.6441%	63.5603%	59.5238%	56.6786%	58.7025%	58.2617%	56.4516%
100	77.0213%	68.6441%	62.9630%	59.0476%	56.6786%	58.7025%	58.2617%	56.4516%
(c) 3차식								
1	69.7872%	68.6441%	63.3214%	60.0000%	55.3310%	54.4304%	56.6448%	55.6452%
10	74.0426%	68.6441%	64.6356%	58.5714%	57.8280%	60.2848%	58.6155%	54.4355%
33	76.1702%	67.7966%	64.5161%	56.6667%	57.8676%	60.2848%	59.4240%	54.4355%
55	76.3830%	68.6441%	64.7551%	56.6667%	57.8676%	60.7595%	59.7777%	55.4435%
78	76.8085%	70.3390%	65.5914%	55.7143%	57.7487%	60.1266%	59.6261%	55.2419%
100	76.8085%	71.1864%	64.9940%	54.7619%	57.8280%	60.2848%	59.9293%	55.4435%
(d) 4차식								
1	68.9362%	70.3309%	62.9630%	59.0476%	52.7547%	52.0570%	55.1794%	52.8226%
10	73.8298%	68.6441%	65.1135%	56.1905%	56.7578%	57.1203%	57.9080%	55.4435%
33	67.7966%	66.9492%	65.4719%	56.1905%	57.3127%	59.3354%	58.7671%	54.8387%
55	76.5957%	66.9492%	65.1135%	54.2857%	57.5505%	59.0190%	59.0197%	55.2419%
78	76.3830%	67.7966%	65.4719%	54.7619%	57.9073%	59.3354%	59.1208%	55.2419%
100	78.0851%	70.3390%	66.4277%	56.1905%	58.1847%	59.9684%	59.1713%	55.2419%
(e) 5차식								
1	68.9362%	68.6441%	63.9188%	58.0952%	52.1205%	50.3165%	50.0758%	50.0000%
10	73.1915%	68.6441%	65.5914%	55.2381%	55.3706%	53.7975%	55.5331%	53.2258%
33	75.1064%	66.9492%	65.8303%	54.2857%	56.8767%	55.5380%	57.5543%	54.2339%
55	77.0213%	68.6441%	65.2330%	54.7619%	56.9560%	57.1203%	58.1102%	54.8387%
78	77.6596%	68.6441%	66.1888%	55.2381%	57.1145%	57.2785%	58.4133%	55.8468%
100	77.8723%	72.0339%	67.0251%	55.7143%	57.3524%	56.9620%	58.7671%	55.0403%

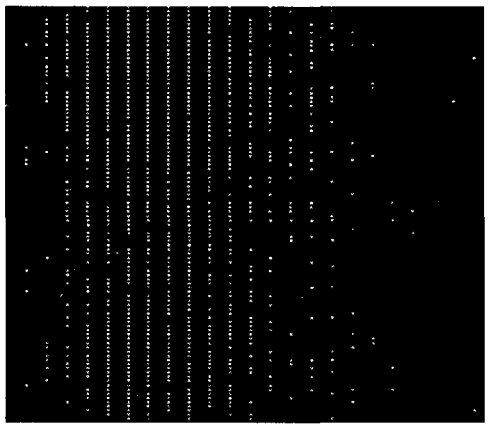
결과에서 볼 수 있듯이, 상품군 2를 제외하고는 선형 SVM보다 비선형 SVM이, 그 중에서도 특히 커널함수를 가우시안 RBF로 사용한 경우에 가장 우수한 성과를 보이고 있음을 알 수 있다.

<표 5>는 로지스틱 회귀분석, ANN 등 비교 기법들의 성과와 SVM의 성과 중 가장 우수한 결과를 함께 비교하고 있다. 이 결과를 통해 알 수 있듯이 모든 상품군에서 SVM이 가장 우수한 예측 성과를 보임을 알 수 있다.

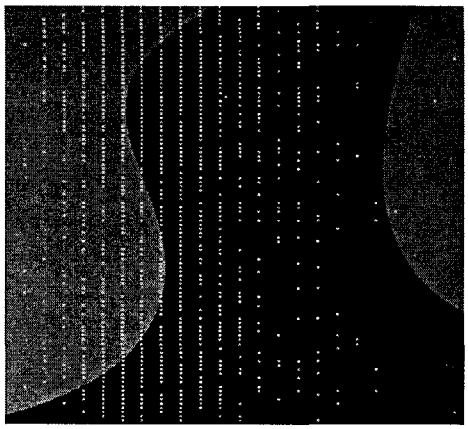
<표 5> 비교 기법과의 성과 비교

구분	로지스틱 회귀분석	ANN	SVM	SVM 설정
(a) 상품군 1				
훈련 데이터	75.32%	77.90%	80.21%	가우시안 RBF / δ^2=1 / C=10
검증 데이터	67.80%	72.88%	73.73%	
(b) 상품군 2				
훈련 데이터	64.99%	65.92%	64.76%	선형 SVM / C=10
검증 데이터	61.43%	60.95%	61.43%	
(c) 상품군 3				
훈련 데이터	58.94%	58.21%	62.58%	가우시안 RBF / δ^2=1 / C=78
검증 데이터	59.65%	60.44%	62.66%	
(d) 상품군 4				
훈련 데이터	57.55%	57.37%	62.00%	가우시안 RBF / δ^2=1 / C=100
검증 데이터	56.25%	56.45%	56.65%	

[그림 2]와 [그림 3]은 상품군 3에 대해 SVM 모형의 분류 전과 후의 분류결과를 그림으로 나타낸 예시이다. 그림에서 점은 각 사례의 출력값을 나타내는데 두 가지 색상으로 표현된 것은 구매와 비구매 두 개의 클래스를 구분하기 위한 것이다. 또한 분류 후 그림에서 두 개의 색상영역을 구분하는 선이 분리경계면이 된다. 따라서 두 가지 색상으로 구분된 영역에 포함된 각각의 점들 중 각 영역의 색상과 동일한 색상의 점들은 바르게 예측이 된 것이고, 일치하지 않는 색상의 점들은 오분류된 사례를 나타내고 있는 것이다.

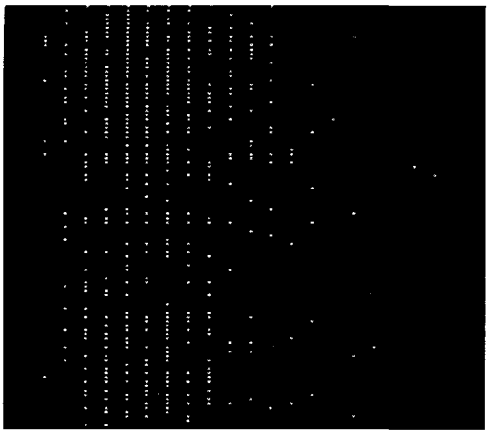


(a) 분류 전

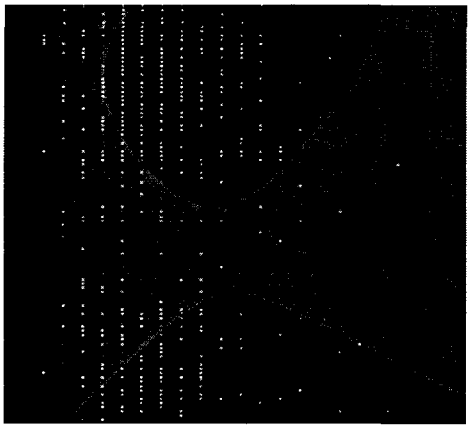


(b) 분류 후

[그림 2] 훈련데이터에 대한 SVM 분류 전과 후의 분류결과



(a) 분류 전



(b) 분류 후

[그림 3] 검증데이터에 대한 SVM 분류 전과 후의 분류결과

5. 결론

본 연구에서는 주가지수예측과 기업신용평가 등의 분야에서 활발하게 응용되고 있는 SVM을 고객구매예측에 최초로 적용하여 그 가능성을 확인해 보았다. 실험 결과, SVM은 기존에 적용되던 로지스틱 회귀분석이나 ANN과 같은 기법들보다 우수한 예측력을 보임을 알 수 있었다. 특히 성과는 우수하나, 모형 구축에 많은 연산과 시간이 소요되는 ANN에 비해, 성과도 우수하고 모형 구축이 훨씬 단순하다는 측면에서 SVM이 높은 적용 가능성을 지닌 기법이라는 점을 확인할 수 있었다.

본 연구는 SVM을 이용한 최초의 고객구매예측 모형을 제안하였다는 의의를 가지고 있으나 몇 가지 한계점도 가지고 있다. 먼저, SVM의 경우, 파라미터를 어떻게 설정하느냐에 따라 성과가 변동될 수 있는데, 아직까지 SVM의 최적 파라미터를 결정할 수 있는 일반적인 방안이 있지 않아 한정된 범위 내에서만 파라미터 변동의 효과를 확인할 수 밖에 없어서 SVM의 우수한 성과를 제대로 제시하지 못하였을 가능성이 있다. 한편, 결과의 일반화와 관련해서 교차검정 등을 시행하여 충분한 표본을 확보하여 다양한 실험조건에서의 강건성 등을 확인해야 할 것이다. 따라서 향후에는 한계점으로 제시한 SVM의 파라미터 최적화와 성과의 일반화를 위한 추가 연구가 필요하다.

참고문헌

[1] 박정민, 김경재, 한인구. "Support Vector Machine을 이용한 기업부도예측", *경영정보학연구*, 제15권, 제2호, pp. 51-63, 2005.

[2] 안현철, 한인구. "데이터 마이닝을 활용한 인터넷 쇼핑몰의 상품 추천 시스템 개발", *경영정보학회 춘계학술대회 논문집*, pp. 739-748, 2002.

[3] 이수용, 이일병. "Fuzzy 이론과 SVM을 이용한 KOSPI 200 지수 패턴분류기", *한국증권학회 제4차 정기학술발표회 논문집*, pp. 787-809, 2002.

[4] 민재형, 이영찬. "Support Vector Machine을

이용한 부도예측모형의 개발", *한국경영과학회지*, 제30권, 제1호, pp. 55-74, 2005.

[5] Barniv, R., Agarwal, A., and Leach, R. "Predicting the outcome following bankruptcy filing: a three-state classification using neural networks", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 6, No.3, pp. 177-194, 1997.

[6] Berry, M.J.A., and Linoff, G.S. *Data Mining Techniques: For Marketing, Sales and Customer Support*, Wiley Computer Publishing, 1997.

[7] Chang, C.-C., and Lin, C.-J. *LIBSVM: a library for support vector machines*, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Available at http://www.csie.edue.tw/~chlin/papers/libsvm.pdf, 2001.

[8] Chiu, C. "A case-based customer classification approach for direct marketing", *Expert Systems with Applications*, Vol. 22, pp. 163-168, 2002.

[9] Fanning, K., and Cogger, K. "A comparative analysis of artificial neural networks using financial distress prediction", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 3, No. 3, pp. 241-252, 1994.

[10] Fletcher, D., and Goss, E. "Forecasting with neural networks: An application using bankruptcy data", *Information and Management*, Vol. 24, pp. 159-167, 1993.

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

[12] Huang, W., Nakamori, Y., and Wang, S.-Y. "Forecasting stock market movement direction with support vector machine", *Computers & Operations Research*, Vol. 32, No. 10, pp. 2513-2522, 2005.

[13] Huang, Z., Chen, H., Hsu, C-J., Chen, W-H., and Wu, S. "Credit rating analysis with support vector machines and neural networks: a market comparative study", *Decision Support Systems*, Vol. 37, pp. 543-558, 2004.

[14] Joachims, T. "Text categorization with support vector machines", *Proceedings of the European Conference on Machine Learning(ECML)*, pp. 137-142, 1998.

[15] Kim, K. "Financial time series forecasting using support vector machines", *Neurocomputing*, Vol. 55, No. 1-2, pp. 307-319, 2003.

[16] Odom, M., and Sharda, R. "A neural network model for bankruptcy prediction", *Proceedings of the International Joint Conference on Neural networks*, pp. 163-168, 1990.

[17] Ohlson, J.A. "Financial ratios and probabilistic prediction of bankruptcy", *Journal of Accounting Research*, Vol. 18, pp. 109-131, 1980.

[18] Osuna, E., Freund, R., and Girosi, F. "Training support vector machines: An application to face detection", *Proceedings of Computer Vision and Pattern Recognition*, pp. 130-136, 1997

[19] Shin, K.-S., Lee, T.S., and Kim, H.-j. "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Applications*, Vol. 28, No. 1, pp. 127-135, 2005.

[20] Tam, K., and Kiang, M. "Managerial

applications of neural networks: the case of bank failure prediction", *Management Science*, Vol. 38, No. 7, pp. 926-947, 1992.

[21] Tay, F.E.J., and Cao, L.J. "Modified support vector machines in financial time series forecasting", *Neurocomputing*, Vol. 48, pp. 847-861, 2002.

[22] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[23] Zhang, G., Hu, M.Y., Patuwo, B.E., and Indro, D.C. "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", *European Journal of Operational Research*, Vol. 116, pp. 16-32, 1999.

Abstract

Purchase Prediction Model using the Support Vector Machine

Hyunchul Ahn* · Kyoung-jae Kim** · Ingoo Han*

As the competition in business becomes severe, companies are focusing their capacity on customer relationship management (CRM) for survival. One of the important issues in CRM is to build a purchase prediction model, which classifies customers into either purchasing or non-purchasing groups. Until now, various techniques for building purchase prediction models have been proposed. However, they have been criticized because their performances are generally low, or it requires much effort to build and maintain them. Thus, in this study, we propose the support vector machine (SVM) as a tool for building a purchase prediction model. The SVM is known as the technique that not only produces accurate prediction results but also enables training with the small sample size. To validate the usefulness of SVM, we apply it and some of other comparative techniques to a real-world purchase prediction case. Experimental results show that SVM outperforms all the comparative models including logistic regression and artificial neural networks.

Key words : Purchase prediction model Support vector machines Logistic regression Artificial neural networks Customer relationship management

* Department of Management Engineering, KAIST Graduate School of Management
** Department of Information Systems, Dongguk University