신경망과 SVM을 이용한 주가지수예측의 비교

Comparison of Stock Price Index Prediction Performance Using Neural Networks and Support Vector Machine

저자
(Authors)

김유일, 신은경, 홍태호
Yuil Kim, Eunkyoung Shin, Taeho Hong

출처
(Source)

인터넷전자상거래연구 4(3), 2004.12, 221-243 (23 pages)
The Journal of Internet Electronic Commerce Resarch 4(3), 2004.12, 221-243 (23 pages)

발행처
(Publisher)

한국인터넷전자상거래학회
Korea Internet Electronic Commerce Association

URL

http://www.dbpia.co.kr/Article/NODE00583576

APA Style

김유일, 신은경, 홍태호 (2004). 신경망과 SVM을 이용한 주가지수예측의 비교. 인터넷전자상거래연구, 4(3), 221-243.

ISSN 1598-1983
한국인터넷전자상거래학회
인터넷전자상거래연구 제4권 제3호 2004년

신경망과 SVM을 이용한 주가지수예측의 비교

Comparison of Stock Price Index Prediction Performance Using Neural Networks and Support Vector Machine

김유일* · 신은경** · 홍태호***
Yuil Kim · Eunkyoung Shin · Taeho Hong

Abstract

This study presents the comparative analysis of data mining performance for the prediction of stock price index using neural networks and support vector machine. The prediction of stock price index was performed on the basis of technical analysis using technical indicators which is able to find the change of the present and future prices in the market. Neural networks have a few problems such as the lack of explanation and over-fitting although their outstanding performance in the financial prediction area. On the other side, SVM is capable of generalizing the model because it can be explained mathematically and minimize the structured risk.

In this study, we predicted the stock price index using neural networks and SVM and compared their performance with the prediction performance of discriminant analysis and logit, called statistical techniques. For the comparison of performance among each

* 부산대학교 경영학부 교수
** 부산대학교 대학원 경영학과
*** 부산대학교 경영학부 조교수

models, KOSPI 200 and S&P 500 index are utilized to experiments and the results are tested statistically. In addition, we analyzed the experimental results considering the characteristics of data in Korea and US.

Keywords : data mining, neural networks, SVM, stock price index prediction, technical analysis

Ⅰ. 서 론

데이터마이닝(data mining)은 많은 양의 데이터에서 의미 있는 규칙과 패턴을 탐색하는 기법으로 마케팅, 웹분석 등 여러 영역에서 활용되고 있다(Changchien & Lu, 2001; Park, Piramuthu & Shaw, 2001). 데이터마이닝은 초기에 통계기법으로 시작하여 최근에 기계학습기법으로 발전 되고 있다. 데이터마이닝에서 사용되는 통계기법에는 판별분석, 회귀분석 등이 있으며, 기계학습기법으로는 신경망(neural networks), 귀납적 학습법(inductive learning), 사례기반 추론(case-based reasoning), 유전자 알고리즘 (genetic algorithms), support vector machine(SVM) 등이 있다.

신경망과 통계적 기법을 비교하면, 높은 비선형성과 동적인 성질을 가진 문제에는 신경망이 통계적 기법보다 예측과 분류를 더 정확하게 수행한다. 또한 신경망은 변화하는 환경에 쉽게 적응이 가능하며, 샘플의 크기, 변수의 수와 데이터의 분포에 대해서 덜 민감하여 잡음을 포함하고 있는 데이터도 이용이 가능하다(Sun et al, 1997). SVM은 Vapnik에 의해서 개발된 학습 기법으로 실험 설계시 해당하는 함수의 파라미터를 조정하여 비교적 간단하게 학습할 수 있다. 대부분의 학습 알고리즘은 경험적 위험 최소화 원칙을 구현하는 것에 비해, SVM은 구조적 위험의 최소화 원칙에 기반을 두기 때문에 모델을 일반화하기가 용이하다고 할 수 있다(Tay and Cao, 2001).

주가예측은 예측분야 중 가장 어려운 분야로 여러 가지 예측기법들이 개발되어 있다. 그러나 어떤 기법이든 예측율이 크게 높지 못하다. 이 논문에서는 주가지수 예측율을 높이기 위하여 데이터마이닝 기법 중 신경망기법과 SVM기법을 응용하고자 한다. 신경망기법과 SVM기법의 성능을 평가하기 위해서 통계기법인 판별분석과 로짓의 결과와 비교한다. 그리고 SVM이 주가지수 예측에서의 활용 가능성이 있는지를 이 연구에서 확인하고자 한다. 또한 주가지수 예측에 사용할 데이터

마이닝 기법의 적용가능성을 일반화하기 위해서 한국과 미국의 증권시장을 대표하는 KOSPI 200 지수와 S&P 500 지수를 대상으로 하고, 양 지수를 주간예측, 일간예측으로 예측기간을 나누어 분석한다. 이러한 기법간의 예측율 차이가 통계적으로 있는가를 밝히기 위해서 McNemar Test를 통해 검정한다.

Ⅱ. SVM (Support Vector Machine)

Support Vector Machine(SVM)은 VC(Vapnik Chernonenkis)이론을 근거로 하여 개발되었다(Vapnik, 1995). SVM은 예측에 우수하며, 벌칙항(penalty)을 이용하여 과대적합을 피하는데 성공적이고, 함수근사의 문제에서 이상치에 둔감하다는 장점을 갖고 있다는 것을 여러 연구의 결과에서 보여 준다. VC이론은 통계적 학습이론의 하나로 N차원을 구성하고 있는 x_i와 각 클래스에 대한 인덱스를 가지고 있는 y_i를 대상으로 학습을 통해 얻어진 함수 f를 추정하여 x를 {+1, -1} 중 하나로 분류하는 방법론이다. SVM의 함수 근사는 다음과 같은 절차로 이루어진다.

훈련자료 $\{(x_i, y_i), i = 1, ..., n\} \subset X \times R$ 이 주어졌다고 가정한다.
여기서 X는 d차원 입력벡터 공간 R^d를 나타낸다.

SVM은 모든 훈련자료에 대해서 실제 목표값 y_i로부터 최고 ϵ만큼의 편차이내에 있으며 가능한 작은 크기의 w값을 갖는 함수로 다음과 같다.

$$f(x) = w^t x + b \, with \, x \in X, b \in X$$

이를 해결하기 위한 한 가지 방법은 제곱 $\|w\|^2$을 최소화 시키는 것이다. 이러한 문제는 공식적으로 다음과 같은 볼록 최적화 문제로 간주 된다.

$$minimize \frac{1}{2}\|w\|^2,$$
$$subject to \left| y_i - w^t x_i - b \right| \leq \epsilon$$

여기서 기본 가정은 볼록 최적화 문제의 해결이 가능하다. 이를 해결하기 위해 새로운 슬랙변수 ξ, ξ^*를 도입하여 Vapnik(1995)이 제안한 최적화문제는 다음과 같이 전개된다.

$$minimize \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^2 + \xi_i^{*2}),$$

$$svbjcet \ to \ \ y_i - w^t x_i - b \le \epsilon + \xi_i, \ i = 1, ..., n$$

$$w^t x_i + b - y_i \le \epsilon + \xi_i^*, \ i = 1, ..., n$$

$$\xi_i, \xi_i^* \le 0$$

여기에서 C는 함수의 평평한 정도와 추정오차의 크기를 조절하는 모수이다. 이 문제를 라그랑제 함수로 표현하면 다음과 같이 된다.

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i^2 + \xi_i^2) - \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i - y_i + w^t x_i + b)$$

$$- \sum_{i=1}^{n}\alpha_i^*(\epsilon + \xi_i^* + y_i - w^t x_i - b) - \sum_{i=1}^{n}(\eta_i \xi_i \ + \eta_i^* \xi_i^*)$$

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$$

위의 라그랑제 함수를 최소화하는 w와 b를 구하여 위 식에 대입하고 커널을 이용한 비선형함수의 추정으로 확장을 하면 다음의 최대화문제가 된다.

$$maximaize \quad -c\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\{K(x_i, x_j + \delta_{ij}/C)\}$$

$$-\epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i - \alpha_i^*),$$

$$svbject \ to \quad \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \ \text{and} \ \alpha_i, \alpha_i^* \ge 0$$

이때 많이 사용되고 있는 커널함수는 일반적으로 다음과 같다

$$Polynomid \qquad K(x_i, x_j) = ((x_i, x_j) + 1)^d$$

$$Gaussian\ RBF\ K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

여기서 d와 σ는 커널함수의 모수로 SVM에서 중요한 역할을 한다.

Ⅲ. 주가지수 예측에 관한 문헌 연구

1. 주가지수 예측의 특징과 분석법

주식시장 참여자들이 미래의 주식시장 가격을 예측하기 위해 사용되는 분석방법으로는 크게 기본적 분석(fundamental analysis)과 기술적 분석(technical analysis)이 있다. 기본적 분석은 시장에서 거래되고 있는 어떤 상품의 본질적 가치(intrinsic value)를 연구하여 현재 시장 가격과의 괴리를 밝힘으로써 향후의 가격을 예측하고자 하는 기법이다. 즉 언젠가는 시장 가치가 본질적 가치를 향해 움직이리라고 가정하는 것이다. 반면에 기술적 분석이란 상품의 본질적 가치를 분석하기보다 과거 및 현재의 시장가격 변동을 연구하여 어떤 특징이나 패턴을 찾아내고 이를 통해 미래의 가격변화를 예측하고자 하는 기법이라고 할 수 있다.

주가지수 예측에 초점이 맞추어져 있는 많은 연구가 수행되었다. 그러나 주식시장 데이터에 많은 양의 노이즈와 비고정성 특징 때문에 뛰어난 예측 정확성을 나타내는 연구는 많지 못하다. 특히 주식 시장의 주가지수 데이터는 비선형인 특징을 가지기 때문에 선형모델로 주식 시장을 분석하기에는 여러 가지 제약이 따른다. 따라서 단기적인 주식 시장 분석과 예측에는 기술적 분석을 이용한 인공지능 기법에 의한 비선형 패턴 분석이 많이 활용된다.

Tasih et al. (1998)는 규칙 기반 시스템과 신경망과의 통합 모형을 만들어 주가지수를 예측하였는데, 이 통합모델은 실제 데이터 실험에서 약 58%정도의 예측 정확성을 보였다. 이 연구에서 주가 지수 예측에 이용한 기술적 지표는 SP, SN, LU, LD, UD, AUD, 와 기존에 널리 쓰이는 상대강도지수(relative strength index : RSI)를 수준 별로 나눈 RSI1, RSI2, RSI2, RSI3, RSI4이었다. 여기서 SP와 SN은 14일간의 회귀선 기울기를 음수와 양수로 구분하여 가격의 경향을 알아보기 위해 사용한 지표이다. LU와 LD는 14일후 예측한 값과 실제 값의 차이로 계산되는 예측 진동자(Oscillator)로 쓰였다. UD와 AUD는 오늘의 종가와 전날 종가를 비교해서 상승과 하락을 구분하여 상승시에는 UD를 1

로 하락시에는 AUD를 1로 둔다. 위와 같은 10가지 변수 중에서 규칙 기반 시스템에서 LU, LD, RSI1, RSI2를 트리거로 중요하게 다루었다. 이들 4가지 변수의 신호에 따라서 나머지 6개의 변수들은 최종적으로 현재 시장의 경향을 얻기 위해서 사용되었다.

Shen & Loh (2004)의 연구에서는 러프셋을 이용하여 S&P 500을 대상으로 주가지수에 대한 예측을 하였다. 실제 실험 데이터에서는 약 58%의 예측 정확성을 보였다. 이 연구의 기술적 지표는 이동평균의 수렴/발산(moving average convergence & divergence: MACD), 가격변화율(price rate of change : ROC), 확률적 진동자(stochastic oscillator: RSI), 방향변동(directional movement: DI), 선형회귀선(linear regression lines), 가중누적개편계열(weighted accumulated reconstruction series: WARS)이다. 이 중 중요하게 언급된 지표는 WARS이다. 기술적 지표에는 이동평균법을 많이 사용하게 되는데 WARS는 데이터에서 정보를 더 많이 가지고 있고, 가격의 변화 추이를 잘 반영하는 특징을 가지고 있다고 언급했다.

인공지능 분야에서 SVM을 이용한 연구가 활발히 진행되고 있으며, 우수한 성과를 나타내고 있는 것으로 알려져 있으며(Joachims, 1998; Osuna et al., 2003), SVM은 문자인식, 얼굴인식 등 많은 응용분야에서 성공사례를 보여주고 있다(Burges, 1998; Gunn, 1998; Smola, 1998). 최근 SVM을 재무분야에 적용한 연구로는 주로 시계열 예측 및 분류에 관한 것이 있다(박정민 등 2003; Huang et al., 2003; Kim, 2003 ; Tay et al., 2001). Kim(2003)의 연구에서는 SVM을 주가 상승과 하락을 예측하는데 사용하였다. 그는 SVM을 사례기반추론(case-based reasoning)및 신경망(neural networks)과 성능을 비교 하였다. 입력 변수로 그는 기술적 지표를 사용하였으며 출력 변수로는 일간 주가의 상승, 하락으로 하였다. 이 연구에서 사용한 기술적 지표는 %K, %D, Slow %D, Momentum, ROC, Williams' %R, A/D Oscillator, Disparity5, Disparity10, OSCP, CCI, RSI이다. 검증용 데이터의 적용결과를 보면 사례기반 추론은 51%, 역전파 신경망 알고리즘은 54%, SVM은 57%의 적중률을 나타내었다. 따라서 이 실험에서 SVM이 주가 시장을 예측하는데 하나의 대안이 될 수 있음을 증명하였다.

2. 기존연구의 기술적 지표

Tsaih et al.(1998)은 기술적 지표를 가격경향, 예측 오실레이터, 현재 시장

경향, 경향이 없는 시장의 가격정보를 추출해 내는 지표 등으로 기술적 지표를 4가지로 분류하였다. 가격경향을 알아보기 위한 지표는 SP, SN을 사용하였으며, 예측 오실레이터 지표는 LU, LD를 사용하였다. 현재 시장 경향을 알아보기 위한 지표는 UD, AUD를 사용하였으며, 경향이 없는 시장에서 가격 정보를 알아보기 위한 지표는 RSI를 수준별로 4가지로 나눈 RSI1, RSI2, RSI3, RSI4를 사용하였다. 그는 유전자 알고리즘과 신경망 모형의 통합 모형을 제시하였는데 유전자 알고리즘에서 LU, LD, RSI1, RSI2를 트리거로 사용하여 나머지 지표보다 더 중요하게 다루었다.

Shen & Loh (2004)의 연구에서는 기술적 지표를 7개를 사용하였는데, 그 중에서 WARS를 기존에 쓰던 평균이동법보다 주가의 상태를 더 잘 반영한다고 강조하였다. 나머지 지표는 MACD, ROC, Stochastic Oscillator, RSI, DI과 Linear regression lines를 사용하였다. 그들의 연구에서는 이러한 지표들이 기술적 분석에서 잘 구성된 지표이기 때문이라고 변수 선정 근거로 밝혔다. 6개의 지표들을 살펴보면 시장의 상승·하락의 어느 단계에 있는 가를 보여주는 추세지표로 MACD, ROC, Linear regression lines를 사용하였고, 주가의 전환점을 알려주는 오실레이터 지표로 Stochastic Oscillator, RSI, DI를 사용하였다.

Kim(2003)의 연구에서는 %K, %D, Slow %D, Momentum, ROC, Williams' %R, A/D Oscillator, Disparity5, Disparity10, OSCP, CCI, RSI를 기술적 지표를 사용하였다. 변수 선정은 이전 연구와 전문 분야의 분석으로 기술적 지표의 선정 근거로 밝혔다. 이 연구에서 사용한 지표의 특성을 살펴보면 기술적 지표는 오실레이터 지표와 주식가격 추세지표로 분류할 수 있다. 오실레이터 지표로는 %K, %D, Slow %D, Momentum, Williams' %R, A/D Oscillator, CCI, RSI를 사용하였으며, 주식가격의 추세지표로는Disparity5, Disparity10, OSCP, ROC를 사용하고 있다.

Ⅳ. 연구모형

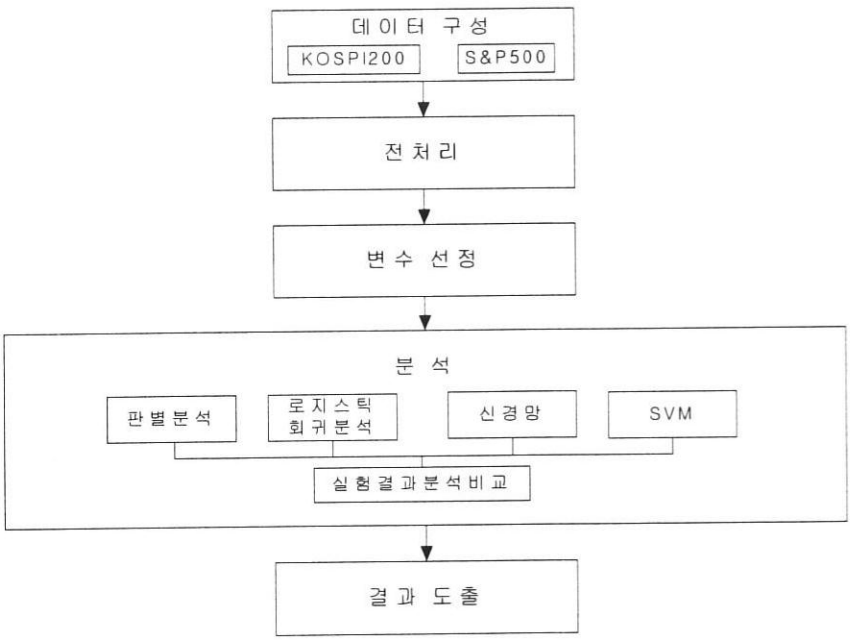
1. 연구절차

이 논문의 연구절차는 데이터마이닝을 적용하는 일반적인 과정에 따라 〈그림 1〉과 같이 진행한다. 연구 대상의 데이터는 한국의 주가 지수를 대표하는 KOSPI 200 지수와 미국의 주가지수를 대표하는 S&P 500 지수를 선정하였으며 기간은

1995년 1월부터 2004년 9월까지 하였다.

분석에 들어가기 전의 처리과정으로 5-fold cross validation을 수행할 수 있도록 데이터 셋을 나누었다. 5-fold cross validation을 통해서 본 연구의 실증결과의 일반성을 확보할 수 있으며, 샘플링에 따른 오류를 줄일 수 있다. 또한 데이터를 날짜별로 저가, 고가, 종가에 따라 정리하는 전처리를 하고 난 뒤 선행 연구에서 살펴보았던 중요한 기술적 지표로 각 기법에 들어갈 입력 변수를 선정한다.

분석 방법으로는 신경망, SVM 분석 방법을 사용하여 결과를 도출하고 전통적인 통계 기법인 판별분석과 로짓의 결과와 비교한다. 위와 같은 4가지 데이터마이닝 기법으로 주가지수 예측에 대한 분석을 한 후 각 기법간의 차이가 통계적으로 유의한지 검증하기 위해서 McNemar Test로 검정한다. 최종적으로 이러한 실험결과를 평가하여 이 연구의 결론을 도출한다.



〈그림 1〉 연구절차

2. 변수선정

입력변수의 선정은 모형의 정확도에 커다란 영향을 미치며, 입력변수가 잘못 선정된 경우 예측 정확도는 현저히 낮아진다. 그러나 최적의 입력 변수군을 선정하

는 문제는 매우 어려운 과제 중 하나로, 선행 연구들에게서는 주로 전문가의 의견을 반영하거나, 문헌을 통해 도출, 혹은 통계적 기법을 활용하여 입력 변수를 선정하는 것이 일반적이다. (홍승현, 신경식; 2003)

이 연구에서는 앞에서 살펴본 선행 연구를 기초로 하여 기술적 지표를 가격변화, 오실레이터, 장기간의 추세경향, 이동평균 등 4가지로 분류하였다. 가격변화를 나타내는 지표는 Momentum, ROC를 사용한다. 주가의 전환점을 알려주는 오실레이터 지표로는 %K, %D, Slow %D, CCI, ROC를 사용하며, 이 지표들은 비교적 단기간의 경향을 알아보는 것으로 기간 n을 5일로 두었다. 장기간의 추세경향을 알아보는 지표는 기간 n을 14일로 한 선형회귀선을 응용한 Linear Slope와 LU/LD를 사용하였다. 여기서 LU, LD 지표는 선행 연구에서는 14일 후 예측한 값과 실제 값의 차이로 계산하여 상승과 하락을 구분하여 2개의 지표로 설정하였으나 본 논문에서는 이 값들을 그대로 반영하여 하나의 지표로 나타내었다. 이동평균은 기복이 심한 변수의 움직임을 완만하게 표현해 내는 지표로 MACD, WARS를 사용하였다. 선택된 변수의 산식은 〈표 1〉과 같으며, 변수 WARS의 산식은 다음과 같다.

〈표 1〉 선택된 변수와 관련 설명

변수명	산식	설명	연구자
Momentum	$C_t - C_{t-n}$	주어진 기간 동안의 주식 가격의 변화량 측정	Kim (2003)
ROC	$\frac{C_t}{C_{t-n}} \times 100$	n일 전과 현재 가격의 비율	Kim (2003), Shen & Loh (2004)
%K	$\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$	주어진 기간 동안의 가격의 범위에 대해서 종가의 상대적인 비교	Kim (2003)
%D	$\frac{\sum_{i=0}^{n-1} \%K_{t-i}}{n}$	%K의 이동 평균	Kim (2003)
Slow %D	$\frac{\sum_{i=0}^{n-1} \%D_{t-1}}{n}$	%D의 이동 평균	Kim (2003)

변수명	산식	설명	연구자
CCI	$\frac{(M_t - SM_t)}{(0.015D_t)}$ $M_t = (H_t + L_t + C_t)/3$ $SM_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$, and $D_t = \frac{\sum_{i=1}^{n} \lvert M_{t-1+1} - SM_t \rvert}{n}$	통계적 평균으로부터 주식 가격의 변동성을 측정	Kim (2003)
RSI	$100 - \frac{100}{1_n + RS}$ $RS = \sum_{t=1}^{n}(C(t) - C(t-1))$ $\sum_{t=1}^{n}(C(t) - C(t-1))^{-}$	0에서 100범위에서 오실레이터를 따라가는 가격	Tsaih et al. (1998), Kim (2003), Shen & Loh (2004)
Linear Slope	$C = \alpha + \beta \times t + \epsilon_t$, $t = 1, 2, \cdots, 14$	14일간의 추세의 상승과 하락	Achelis (1995)
LU/LD	$\frac{FO_t + FO_{t-1} + FO_{t-2}}{3}$ $FO = 100 \times \frac{C_t - C_f}{C_t}$ C_f : 회귀식으로부터 14일전에 예측한 값	주식 가격의 잠재적인 경향 변화값	Tsaih et al. (1998)
WARS	이론 배경 참고	지수의 변동을 반영하는 지표	Shen & Loh (2004)
MACD	12일 지수식 이동평균 -26일 지수식 이동평균	장, 단기 이동평균선의 상호관계 중 중요한 특성인 수렴과 확장을 반복하는 특성과 최근 주가에 더 큰 가중치를 둠	Shen & Loh (2004)

Ct : t일의 종가, H_t: t일의 고가, L_t: t일의 저가, HH_t: t기간 중 최고가, LL_t: t기간 중 최저가

WARS는 가중누적개편계열으로 시리즈의 모든 값을 {-1, 1}사이의 값으로 정규화하고 시점간의 간격을 유지한 다음, 가중치누적을 통하여 값을 얻는다. 계산하는 과정은 다음과 같다;

Step 1: -1과 1사이의 이 시리즈의 값을 정규화 한다.

$$x_i = x_i / max(\lvert x_i \rvert); \quad (i=1,n)$$

Step 2: 한 시리즈의 처음 값을 뺀다. $x_i = x_i - x_1;$ $(i=1,n)$

Step 3: 전체 시리즈로부터 평균값을 뺀다. $x_i = x_i - Mean\_x;\quad (i=1,n)$

$$where: Mean\_x = \frac{1}{n}\sum_i x_i$$

Step 4: 실제 값을 가중치 누적을 통해서 WARS를 재구성한다.

$$y_n = \frac{1}{1+2+\cdots+n}x_1 + \frac{1+2}{1+2+\cdots+n}x_2 + \cdots + \frac{1+2+\cdots+n}{1+2+\cdots+n}x_n$$

이러한 과정에서 최근의 포인트가 WARS에 더 많은 가중치를 주게 된다.

Step 5: 이 간격의 area를 계산한다.

$$Area = \left| y_1 + y_2 + \cdots + y_n \right|$$

V. 연구모형의 실증 분석

1. 데이터 구성

이 연구의 연구대상 데이터는 주가지수의 상승 및 하락을 예측하기 위해 한국의 주가 지수를 대표하는 KOSPI 200 지수과 미국의 주가지수를 대표하는 S&P 500 지수를 선정하였으며 기간은 1995년 1월부터 2004년 9월까지 하였다. 단 지 대상기간 중에 우리나라 IMF기간은 분석 대상에 포함하기가 적절하지 않다고 판단하여 제외시켰으며, 분석 대상 기간을 동등하게 하기 위해서 S&P 500 지수도 이 기간을 제외 하였으며, 데이터 대상과 기간은 〈표 2〉와 같다. 일간모형은 1일 후의 거래일의 종가의 상승/하락을 예측하는 모형이며, 주간모형은 일주일 후의 거래일의 종가의 상승/하락을 예측하는 모형으로 구성된다.

〈표 2〉 데이터 대상과 기간

데이터 대상	기간	데이터 갯수
KOSPI 200 지수	1995. 01. 20 ~ 1997. 08. 25	1,489개
	2001. 09. 24 ~ 2004. 09. 9	
S&P 500 지수	1995. 01. 24 ~ 1997. 08. 22	1,389개
	2001. 10. 01 ~ 2004. 08. 27	

2. 데이터 전처리

KOSPI 200 지수와 S&P 500 지수의 예측대상은 일주일 후의 종가와 다음 날의 종가이며, 5-fold cross validation을 수행했다. 데이터 셋 구분은 사용하는 각 기법에 따라서 달리 구분하였다. 통계적 기법을 사용할 때 데이터는 학습용과 검증용으로 나누었으며, 신경망 기법을 사용할 때 데이터는 과대적합(over-fitting)을 방지하기 위해서 학습용, 평가용, 검증용으로 나누었다. 기계학습은 학습용, 평가용 데이터에서 행해지며 이 두 데이터에서 구해진 가중치로 검증용 데이터에 대입하여 예측율을 나타낸다. 그래서 전체 결과를 비교 할 때에는 학습용의 예측율은 평가용 데이터을 포함한 예측율을 나타내기로 한다.

SVM 또한 기계학습 기법을 사용하므로 신경망과 동일한 분석을 하기 위해서 학습용, 평가용, 검증용으로 구분을 두었다. 또한 이러한 데이터 구분은 검증용 예측율 가지고 성능을 평가하기 보다는 평가용에서 미리 그 예측율을 판단하여 모형을 재구성하는 것이 더 합리적이라고 판단되기 때문이다. 그래서 전체 결과를 비교 할 때에는 평가용 예측율은 실질적인 기계 학습에 참가하지 않았으므로 그 결과는 나타내지 않는다.

〈표 2〉 데이터 구분

	KOSPI 200 지수		S&P 500 지수	
	판별분석 로짓 모형	신경망 SVM 모형	판별분석 로짓 모형	신경망 SVM 모형
학습용	1190개	893개	1111개	833개
평가용		298개		278개
검증용	299개	298개	278개	278개

3. 분석 결과

이 연구에 사용된 인공신경망은 3 계층 구조를 다층퍼셉트론(multi layer perceptron)을 사용하였으며 학습 알고리즘은 역전파 학습방법(back-propagation algorithm)을 사용한다. 실험 데이터의 입력벡터와 그에 대응하는 출력값을 함께 신경회로망에 입력시킨 후 학습시키는 방법으로써, 입력이 주어짐에 따라 원하는 출력값이 활성화되도록 가중치를 조정한다. 출력 결과가 의도한 결과와 일치하는지를 확인하는 것이다.

신경망과 SVM은 조절할 수 있는 파라미터들을 비교해가며 실험을 하였다. 신경망은 과대적합을 피하기 위해서 평가용에서의 성과가 가장 좋은 것을 선택하였다. SVM도 기계학습 기법이며 평가용에서 성과가 좋은 것을 택해서 검증용 예측율을 비교하는 것이 더 합리적이라 판단하여 신경망과 동일하게 모형을 재구성하였다. 실험의 결과는 상승과 하락의 예측에 대한 정확성을 예측율로 하여 각 기법간의 성과를 비교하였다.

신경망은 노드수를 6, 10, 14, 18, 22로 하고 학습율 0.1, 모멘텀 0.1로 실험을 수행하였다. SVM에는 비선형 함수로 널리 사용되는 가우시안 RBF 함수를 사용하였다. SVM은 가우시안 RBF 함수의 모수 C, δ^2로 그 성과가 달리 나타나게 된다. C값과 δ^2는 1, 25, 50, 75, 100으로 조절하면서 실험을 수행하였다.

1) KOSPI 200지수의 주간 예측

KOSPI 200 지수 주간 예측을 위해서 사용된 신경망과, SVM의 결과는 평가용에서 가장 높은 예측율을 보인 것으로 정리한 모형의 결과는 아래와 같다.

먼저 신경망 기법으로 학습시킨 KOSPI 200지수의 주간 예측을 보면 〈표 4〉와 같이 노드 수가 14일 때 평가율이 집중적으로 높게 나타났으며, 노드 수가 18개와 22개일 때는 예측율이 높게 나타나지 않았다. 평가용의 예측율이 높은 5개 셋의 검증용 데이터의 예측율은 낮은 것이 51.3%, 높은 것은 56.4%로 나타났다.

〈표 4〉 KOSPI 200 지수 주간 예측 - 신경망모형 결과(%)

데이터구분	SET1	SET2	SET3	SET4	SET5
히든 노드수	10	14	14	6	14
학습용	55.5	62.0	60.7	57.0	57.3
평가용	56.0	62.8	56.0	55.4	60.4
검증용	55.0	56.4	56.4	53.0	51.3

SVM 기법으로 학습시킨 KOSPI 200지수의 주간 예측의 결과는 〈표 5〉와 같이 δ^2이 25일 때 평가용 데이터 셋에서 높은 예측율을 가장 많이 보였으며, C값은 비교적 골고루 분포되어 있었다. 이 때 검증용 데이터의 예측율은 54.0%에서 59.47%로 높게 나타났다.

〈표 5〉 KOSPI 200 지수 주간 예측 - SVM모형 결과(%)

데이터구분	SET1	SET2	SET3	SET4	SET5
δ^2/C	25/1	25/75	25/50	1/75	1/25
학습용	53.4	57.4	58.7	56.7	78.0
평가용	54.7	62.4	53.0	56.0	59.1
검증용	56.7	54.0	59.4	56.7	56.7

KOSPI 200지수의 주간 예측에 대한 기계학습의 신경망모형과 SVM모형 결과와 통계기법인 판별분석과 로짓모형의 결과를 요약하면 〈표 6〉과 같다. KOSPI 200 지수의 주간예측 결과의 평균 예측율은 판별분석모형 53.8%와 로짓모형 53.4%에 비해서 신경망모형이 54.4%로 아주 근소하게 높은 성과를 나타내었다. 그러나 SVM모형은 56.7%로 판별분석, 로짓, 신경망 모형보다 상당히 높은 성과를 나타내었음을 알 수 있다.

〈표 6〉 KOSPI 200 지수 주간 예측 전체 결과 (%)

	판별분석모형		로짓모형		신경망모형		SVM모형	
	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용
SET1	55.4	55.7	56.6	56.4	55.5	55.0	53.4	56.7
SET2	56.3	53.4	56.8	54.7	62.0	56.4	57.4	54.0
SET3	55.2	57.7	55.6	53.7	60.7	56.4	58.7	59.4
SET4	57.2	48.7	57.4	49.7	57.0	53.0	56.7	56.7
SET5	57.3	53.4	58.7	52.7	57.3	51.3	78.0	56.7
평균	56.3	53.8	57.0	53.4	58.5	54.4	60.8	56.7

2) KOSPI 200지수의 일간예측

KOSPI 200 지수 일간예측에 사용된 신경망 기법과 SVM 기법의 학습에서 평가용의 예측율이 가장 높은 실험결과를 정리한 것은 아래와 같다.

먼저 신경망 기법의 학습결과를 보면 〈표 7〉에 나타난 것과 같이 평가용에서 예측율이 높은 노드 수는 골고루 나타나고 있다. 검증용 데이터의 평균 예측율은 55.4%에서 59.1% 사이로 상당히 좋은 성과를 보였다.

〈표 7〉 KOSPI 200 지수 일간 예측 - 신경망모형 결과 (%)

데이터구분	SET1	SET2	SET3	SET4	SET5
히든 노드수	14	22	6	18	14
학습용	65.8	59.2	56.6	62.2	58.9
평가용	66.8	61.4	57.3	61.1	62.1
검증용	59.1	55.4	57.7	58.1	59.1

SVM기법의 학습결과는 〈표 8〉에서 보는 바와 같이 δ^2가 25일 때 평가용에서 높은 예측율을 가장 많이 보였으며, C값은 대체적으로 50이상의 값에 분포되어 있었다. 검증용 데이터의 예측율은 55.4%와 58.4% 사이에 나타났으나, 신경망보다는 다소 성능이 떨어지는 것으로 나타났다.

〈표 8〉 KOSPI 200 지수 일간예측 - SVM모형 결과(%)

데이터구분	SET1	SET2	SET3	SET4	SET5
δ^2/C	25/75	75/50	25/75	25/50	1/25
학습용	61.0	59.5	61.6	60.9	75.8
평가용	57.5	56.7	55.0	57.7	57.0
검증용	56.7	55.7	58.4	55.0	56.0

KOSPI 200 지수의 일간예측에 대한 각 기법의 예측율을 종합한 것은 〈표 8〉과 같다. 평균 예측율을 보면 판별분석모형 56.1%, 로짓모형 55.4%, 신경망모형 57.7% 그리고 SVM모형 56.4%로 나타났다. KOSPI 200 일간예측은 신경망모형이 다른 기법에 비해서 다소 우수한 예측율을 보여주었다. 그리고 주간예측보다 일간예측이 평균적으로 더 높은 예측율을 보여 주었다.

〈표 9〉 KOSPI 200 일간 예측 전체 결과 (%)

	판별분석모형		로짓모형		신경망모형		SVM모형	
	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용
SET1	57.9	56.7	57.9	57.2	65.8	59.1	61.0	56.7
SET2	59.5	56.4	58.1	53.2	59.2	55.4	59.5	55.7
SET3	59.9	58.1	57.1	56.3	56.6	57.7	61.6	58.4
SET4	61.0	53.4	57.2	56.1	62.2	58.1	60.9	55.0
SET5	57.9	55.7	58.9	54.3	58.9	59.1	75.8	56.0
평균	59.2	56.1	57.8	55.4	60.5	57.9	63.8	56.4

3) S&P 500 지수 주간 예측

S&P 500 지수 주간예측의 신경망모형과 SVM모형의 결과를 정리한 것은 아래와 같다.

먼저 신경망모형의 결과를 보면 〈표 10〉에서 보는 바와 같이 노드 수가 6, 10, 18개일 때 평가용의 예측율이 높게 나왔다. 검증용 데이터의 평균 예측율을 보면 58.6%와 59.7%사이로 상당히 우수하게 나왔다.

〈표 10〉 S&P 500 지수 주간 예측 - 신경망모형 결과 (%)

데이터구분	SET1	SET2	SET3	SET4	SET5
노드수	10	18	10	6	18
학습용	54.6	63.5	61.3	58.2	62.4
평가용	59.7	62.6	62.8	60.4	61.2
검증용	59.4	59.7	58.6	59.7	59.4

SVM모형의 결과를 보면 〈표 11〉에서 보는 바와 같이 δ^2는 1과 25에서만 평가용에서 높은 예측율을 보였으며 C값은 대체적으로 1과 100으로 극단 값에 분포되어 있었다. 검증용 데이터의 예측율은 57.2%와 61.2% 사이에 우수한 결과로 나타났다.

〈표 11〉 S&P 500 지수 주간 예측 - SVM모형 결과(%)

데이터구분	SET1	SET2	SET3	SET4	SET5
δ^2/C	25/50	25/100	1/1	1/1	25/100
학습용	60.1	61.9	63.7	61.9	60.0
평가용	61.1	59.4	58.3	59.7	59.0
검증용	58.3	60.8	57.2	59.7	61.2

S&P 500 지수의 주간예측의 결과를 정리하면 〈표 12〉와 같다. S&P 500 지수 주간 예측은 데이터마이닝 기법 간에 가장 뚜렷한 성과 차이가 있었다. 평균 예측율의 성과는 SVM모형 59.4%, 신경망모형 58.8%, 로짓모형 58.5%, 판별분석모형 52.8%의 순으로 나타났다.

〈표 12〉 S&P 500 주간 예측 전체 결과 (%)

	판별분석모형		로짓모형		신경망모형		SVM모형	
	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용
SET1	56.8	52.9	58.2	58.6	54.6	59.4	60.1	58.3
SET2	56.6	51.8	57.4	56.1	63.5	59.7	61.9	60.8
SET3	55.9	50.0	59.5	55.8	61.3	58.6	63.7	57.2
SET4	55.0	51.4	59.5	54.3	58.2	56.8	61.9	59.7
SET5	54.0	57.8	58.0	58.1	62.4	59.4	60.0	61.2
평균	55.7	52.8	58.5	56.6	60.0	58.8	61.5	59.4

S&P 500 지수 주간 예측의 SVM모형 결과를 살펴보면 모든 데이터 셋에서 거의 비슷한 예측율을 가지는 것으로 나타났다. 이는 KOSPI 200 지수는 상승과 하락의 비율이 동일하였고, S&P 500 지수는 데이터의 분포가 상승에 치우쳐 있어서 상승과 하락을 분리하게 되는 초평면이 지나치게 단순화 모형을 가지게 된 것을 알 수 있다. 따라서 SVM기법은 주가 예측할 때 상승과 하락의 비율의 차이가 있으면 예측율에 영향을 미치는 것을 알 수 있다.

4) S&P 500지수 일간예측

S&P 500 지수 일간예측을 위해서 사용된 신경망모형의 학습 결과는 〈표 13〉에서 보는 바와 같이 노드수 10을 제외하고 비교적 다양한 노드 수에서 평가용 예측율이 높게 나왔다. 검정용의 예측율은 52.5%에서 59.0% 사이에 나타났다.

〈표 13〉 S&P 500 지수 일간 예측 - 신경망모형 결과 (%)

데이터구분	SET1	SET2	SET3	SET4	SET5
히든 노드수	22	18	6	6	18
학습용	56.0	63.0	56.4	55.9	58.9
평가용	55.8	63.7	59.7	55.2	57.9
검증용	57.6	56.5	55.8	59.0	52.5

S&P 500 지수 일간예측을 위해서 사용된 SVM모형의 학습 결과는 〈표 14〉에서 보는 바와 같이, δ^2가 50이상의 값에서 높은 예측율이 나왔다. C값은 대체적으로 25, 75, 100에서 각 평가용 예측율이 높게 나왔다. 검증용 데이터의 예측

율은 54.0%와 61.5% 사이로 나타났다. 주간예측 학습에서는 대체적으로 δ^2가 25에서 평가용에서 높은 예측율을 보였던 것과 대조적으로 일간예측에서는 δ^2가 50이상의 값에서 예측율이 높게 나타났다.

S&P 500 지수 일간예측에 대한 기계학습을 하는 신경망모형과 SVM모형의 결과와 통계기법인 판별분석과 로짓모형의 결과는 〈표 15〉와 같이 요약된다. S&P 500 지수 일간예측은 주간예측보다 예측기간이 단기간인데도 불구하고 성과가 더 낮았다. 그 이유는 주간예측에 있어서 상승과 하락의 비율이 상승에 치우쳐져 있어서 상승에 예측을 더 많이 하게 되는 이유에서 비롯된 것이라 본다. 여기서 기법간의 성과 차이는 KOSPI 200 지수 일간 예측과 마찬가지로 주간 예측에 비해서 차이가 뚜렷이 나타나지 않았다.

〈표 15〉 S&P 500 일간 예측 전체 결과 (%)

	판별분석모형		로짓모형		신경망모형		SVM모형	
	학습용	검증용	학습용	검증용	학습용	검증용	학습용	검증용
SET1	56.9	54.7	57.9	57.2	56.0	57.6	61.5	56.8
SET2	58.2	50.0	58.1	53.1	63.0	56.5	61.0	54.0
SET3	57.1	59.6	57.1	56.3	56.4	55.8	56.3	57.9
SET4	56.8	59.4	57.2	56.1	55.9	59.0	56.1	61.5
SET5	57.2	51.8	58.9	54.3	58.9	52.5	55.8	54.0
평균	57.2	55.1	57.8	55.4	58.1	56.3	58.1	56.8

4. 실험 결과 비교 분석

위에서 사용한 데이터마이닝 기법들간의 성과 비교를 위해 McNemar Test로 검정해 보기로 한다. 먼저 KOSPI 200 지수에 대한 각 기법의 예측율 간의 차이에 대한 검정통계량을 보면 〈표 16〉에서 보는 바와 같이 KOSPI 200 지수의 주간예측은 SVM모형이 판별분석모형과 비교해서 유의수준 10%에서 유의한 예측성과의 차이가 있었고, 로짓모형과 SVM모형은 유의수준 5%에서 모형간 성과 차이가 있다고 나타났다.

일간 예측에서는 로짓모형과 신경망모형이 유의수준 5%에서 예측 성과간 차이가 있다. 이러한 결과를 볼 때 통계모형보다 신경망모형과 SVM모형이 예측율 성과가 다소 우수하다고 볼 수 있다.

〈표 16〉 KOSPI 200 지수 McNemar Test 결과

	주간예측			일간예측		
	로짓모형	신경망모형	SVM모형	로짓모형	신경망모형	SVM모형
판별분석모형	0.223	0.308	3.549*	0.402	1.990	0.070
로짓모형		0.773	4.673**		4.088**	0.637
신경망모형			1.184			0.718

* : 10% 수준에서 유의, ** : 5% 수준에서 유의, ***: 1% 수준에서 유의

S&P 500 지수의 주간 예측과 일간 예측을 Mcnemar Test 결과를 보면 주간 예측에서는 데이터마이닝 기법간의 차이가 유의수준 1%에서 신경망모형과 SVM 모형이 판별분석이나 로짓모형과 같은 통계적 기법보다 우수하다고 나타났다. 그러나 일간 예측에서는 통계적으로 유의한 예측율 차이가 없음을 알 수 있다.

〈표 17〉 S&P 500 지수 McNemar Test 결과

	주간예측			일간예측		
	로짓모형	신경망모형	SVM모형	로짓모형	신경망모형	SVM모형
판별분석모형	5.112**	12.291***	14.278***	0.012	0.648	1.006
로짓모형		2.461	8.843***		0.716	1.096
신경망모형			0.595			0.093

* : 10% 수준에서 유의, ** : 5% 수준에서 유의, *** : 1% 수준에서 유의

V. 결 론

실험 결과 SVM모형과 신경망모형이 통계모형보다 높은 성과를 보여주었으며 통계적으로는 주간예측에서 일간예측보다 유의한 차이가 있는 것으로 나타났다. 그러나 SVM모형과 신경망모형의 예측율을 비교해보면 SVM모형이 신경망모형보다 조금 높은 수준으로 나타났으나 McNemar Test에서는 통계적으로 유의한 차이를 보여주지 않아 어느 모형이 우월하다고 단정 지을 수는 없다.

SVM모형의 결과는 모수를 변경시킬 때마다 일정한 변화가 나타났다. C가 일정할 때 δ^2가 증가하면 과소 적합되는 양상을 나타났다. 그리고 δ^2이 일정 할 때 C가 증가할수록 과대 적합되는 경향을 보였다. 그리고 신경망과 SVM을 비교하면 신경망은 실험 설계시 모멘텀, 학습율, 학습 횟수, 노드수 변경과 같이 조절해야

할 요인들이 많아서 연구자의 경험에 따른 기교적인 요소가 많이 요구되었다. 그러나 SVM은 모형 구축에 있어서 해당하는 함수의 모수만 변화시키면 되는 용이함이 있다는 이점이 있었다.

이 연구는 다음과 같이 두 가지 측면에서 의의를 찾을 수 있다.

첫째, 최근 패턴인식과 이진 의사결정에 활발하게 연구가 진행되고 있는 SVM 기법을 주가지수 상승과 하락 예측에 적용해 보았다는 점이다. 주가지수를 예측할 때 사용되는 데이터마이닝 기법은 판별분석, 로짓, 인공신경망, 유전자 알고리즘이 기존 연구에서 쓰였으나, 새로운 기계학습 방법인 SVM기법을 이용하여 주가지수의 상승과 하락예측을 다루는 연구는 아직 활발하지 못한 실정이다.

많은 연구에서 신경망은 전통적인 통계 기법보다 우수한 성능을 보여주었다. 그러나 우수한 신경망의 성능에도 불구하고 과대적합과 부족한 설명력과 같은 문제점들이 있다. 그러나 이러한 한계점을 극복하기 위해서 본 논문에서는 새로운 기계학습 기법인 SVM을 사용하였다. SVM은 수학적으로 분석이 되는 간단한 기법이면서 결과 면에서는 우수한 성능을 가져다준다.

두 번째 의의는 주가지수의 상승과 하락 예측을 기존 연구에서 보여주지 않았던 우리나라의 KOSPI 200 지수와 미국의 S&P 500 지수를 실험에 사용하여 SVM의 적용가능성을 일반화를 시도한 점이다. 이 논문에서는 여러 자료를 SVM과 신경망을 적용하여 그 결과를 분석하였다. 분석결과 대체적으로 SVM이 신경망보다 나은 예측율을 보였으나 통계적으로 유의한 차이를 나타내지 않았다. 그러므로 성능면에서 SVM이 신경망보다 우수하다고 단정지을 수는 없으나 주가지수의 상승과 하락 예측을 할 때, 많은 분석 대상을 통하여 SVM이 또 다른 대체방안으로 사용될 수 있음을 입증하였다.

이 연구의 한계점으로는 첫째, 입력 변수들이 각 데이터마이닝 기법에 적절한 변수인지 심층 연구없이, 모든 데이터마이닝 기법에 동일하게 적용하여 데이터마이닝 기법간의 차이를 알아보았다. 그러나 예측율을 향상시키기 위해서는 신경망과 SVM에 맞는 변수를 찾기 위해서 유전자 알고리즘을 사용하거나 통합 모형을 제시하는 것이 바람직하다고 할 수 있겠다.

둘째, 본 연구 결과에서 주간예측이 일간예측보다 상대적으로 기법간의 차이가 유의하게 나타났는데 그 이유는 입력변수를 계산할 때 주로 단기를 나타내는 지표 계산기간 n을 5일로 하였으며, 장기를 나타내는 지표인 선형회귀를 응용하는 변수에는 계산기간 n을 14일로 했다는 점에서 찾을 수 있다. 그러므로 단기간 예측인 일간 예측을 할 때에는 기간 n을 좀 더 짧은 기간으로 패턴을 감지하는 것이 바람직하다고 할 수 있다. 그리고 KOSPI 200 지수와 S&P 500 지수의 상승과 하락 비율에 따라서 그 결과가 다름을 확인 할 수 있었는데 이러한 비율의 차이가

예측력에 영향을 미치는지 검증할 필요성이 있다.

마지막으로 이 연구 결과를 기초로 거래비용을 고려한 수익률 분석을 통하여 실제 매매 시스템에서의 적용을 해 보는 것이 바람직하나 시도하지 못했다.

참 고 문 헌

박정민, 김경재, 한인구, "Support Vector Machine을 이용한 기업부도 예측", 한국경영정보학회 추계학술대회 발표논문집, 2003, pp 751-758.

석경하, "서포터벡터학습의 효율적 알고리즘", Journal of the Korea Data & Information Science Society vol 12, 2001, pp 95-102

여운승, 사회과학과 마케팅을 위한 다변량 행동 조사, 민영사, 서울 , 2000.

이학식, 김영, SPSS 10.0 매뉴얼 서울, 법문사, 2001.

Allen, l.e. "Mining gold from database", Mortagage Banking, 56(8), 1996, pp 99-100

Berry, M. J. A., and G. Linoff, Data mining techniques for marketing, sales and customer support, New York, Wiley, 1997

Burges, C. "A Tutorial on Support Vector Machines for Pattern Recognition", In Data Mining and Knowledge Discovery 2, Kluwer Academic Publishers, Boston, 1998.

Chang, C-C., and C.-J. Lin, 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

Changchien, S. W., and T. Lu, "Mining association rules procedure to support on-line recommendation by customer and products fragmentation", Expert Systems with ApplicationPs, vol 20, 2001, pp 325-335.

Fayyads, U. M., G. Piatetsky-Shapiro, and P. Smith, " The KDD processes for extracting useful knowledge from volumes of data, Communications on the ACM 39, 1996, pp 27-34

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

Joachims, T, "Text categorization with support vector machines", Proceeding of the European Conference on Machine Learning, 1998, pp. 137-142.

Kyoung-jae Kim, "Financial time series forecasting using support vector machines", Neurocomputing 55, 2003, pp 307-319.

Lixiang Shen and Han Tong Loh, "Appling rough sets to market

timing decisions", Decision Support System 37, 2004, pp. 583-597.

Osuna, E., R. Freund and F. Girosi, "Training support vector machines: an application to face detection," Proceedings of Computer Vision and Pattern Recognition, 1996, pp. 13-136.

Park, S. C., S. Piramuthu, and M. J. Shaw, "Dynamic rule refinement in knowledge-based data mining systems", Decision Support Systems, vol 31, 2001, pp 205-222.

Ray Tasih, Yenshan Hsu, and Charles C. Lai, "Forecasting S&P 500 stock index futures with a hybrid AI system", Decision Support Systems 23, 1998, pp 161-174.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, "Learning Internal Representation by Error Rropagaion," in D. E. Rumelhart and J. L. McClelland eds. Parallel Distributed Processing, Vol. 1, MIT Press, 1986.

Schalkoff, R., Pattern Recognition : Statistical, Structural and Neural Approaches, John Wiley & Son, New York, 1992.

Smola, A. J. and B. Scholkropf, "A Tutorial in Support Vector Regression", NeuroCOLT2, Technical Report, NeuroCOLT, 1998.

Steven B. Achelis, " Technical Analysis from A to Z", Probus, Chicago, 1995.

Taeho Hong, Ingoo Han, "Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks", Expert Systems with Applications 23, 2002, pp1-8.

Tay, F. E. H., and L. Cao, "Application of support vector machines in financial time series forecasting", Omega, Vol. 29, 2001, pp 309-317.

Vapnik, V., "The Nature of Statistical Learning Theory", Springer, 1995.

Vapnik. V, S. Golowich , and A. Smola "Support vector method for function approximation, regression estimation, and signal processing," In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, Cambridge, MA, MIT Press, 1997, pp. 281-287.