지능정보연구 제16권 제1호 2010년 3월(pp.71~92)

Support Vector Machines와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발

김선웅
국민대학교 BIT전문대학원
(swkim@kookmin.ac.kr)

안현철
국민대학교 경영대학 경영정보학부
(hcahn@kookmin.ac.kr)

최근 트레이딩 시스템에 대한 관심이 높아지면서, 인공지능을 이용한 지능형 트레이딩 시스템의 개발과 관련한 연구들이 활발하게 이루어지고 있다. 그러나 현재까지 소개된 트레이딩 시스템 관련 연구들은 트레이딩에 적용될 수 있는 다양한 변수들이 실무에서 활용되고 있음에도 불구하고, 주가지수에서 파생된 기술적 지표에만 과도하게 의존하는 경향이 있었다. 또한, 실제 수익창출에 초점이 맞추어진 트레이딩 시스템의 모형보다는 주가 혹은 주가지수의 등락에 대한 정확한 예측에 초점을 맞춰 모형을 개발하려고 하는 한계도 존재했다. 이에 본 연구에서는 기존 연구에서 주로 활용되어 온 기술적 지표 외에 현업에서 유용하게 활용되는 다양한 비가격 변수들을 시스템에 반영함으로서 예측 성과의 개선을 도모하는 동시에, Support Vector Machines 기반의 등락예측모형의 결과를 트레이딩 시스템의 매수, 매도, 혹은 유지의 신호로 해석할 수 있도록 설계된 새로운 형태의 지능형 트레이딩 시스템을 제안한다. 제안시스템의 유용성을 검증하기 위해, 본 연구에서는 2004년 5월부터 2009년 12월까지의 KOSPI200 주가지수에 제안모형을 적용하여 그 성과를 살펴보았다. 그 결과, 제안시스템이 수익률 관점에서 다른 비교모형들에 비해 더 우수한 성과를 도출함을 확인할 수 있었다.

논문접수일 : 2010년 02월 03일 논문수정일 : 2010년 02월 14일 게재확정일 : 2010년 02월 23일 교신저자 : 안현철

1. 서론

효율적 시장가설(Efficient Market Hypothesis, EMH)은 주가가 다양한 정보를 즉각적으로 반영하여 움직이기 때문에 랜덤워크(random walk)에 따라 움직인다고 가정하고 있다. 이 경우 과거의 주가자료를 이용하여 내일의 주가를 예측하는 것은 불가능하기 때문에, 효율적 시장가설 하에서 기술적 분석이나 주가의 예측모형을 찾아내려는 노력은 무의미해진다. 시장의 효율성에 대한 초기의 연구들은 주로 과거 주가자료의 통계적 분석을 통해 시장이 효율적임을 주장해 왔다. 그러나 실제 주식시장은 거래비용이 발생하고 거래제도 등의 일정한 규제를 받고 있기 때문에 주가 움직임에 부분적으로 비효율성이 존재할 수 있다는 주장도 상당히 많다. 특히 욕망과 공포(greed and fear)에 지배되는 시장참여자들의 투자행태로 인해 주가 움직임에 잡음(noise)이 발생하고, 이는 다시 투자

* 본 논문은 2009년도 국민대학교 교내연구비를 지원받아 수행된 연구임.

심리에 영향을 미쳐 주가는 복잡성을 더하게 되는데, 이처럼 주가가 다양한 정보를 즉각적으로 그리고 정확히 반영하지 못하는 경우가 발생함에 따라 시장에는 비효율성이 부분적으로 나타날 수 있다. Elton and Gruber(1984)는 과거 주가를 결합하는 방법이 무수히 많기 때문에 제한된 특정 방법을 써서 시장이 효율적임을 주장하는 것은 불가능하다고 하였다. 실제로 오래 전부터 시장 참여자들은 다양한 기술적 분석을 실제 투자에 활용해오고 있으며, 이를 통해 좋은 투자성과를 거두는 투자자들이 속속 등장하고 있다(Schwager, 1992).

이러한 상황에서 정보처리능력의 발달로 복잡하고 방대한 양의 주가에 대한 실시간 분석이 가능해지면서 자동화된 트레이딩 시스템(trading system)에 대한 연구와 실무에서의 활용이 증가하고 있다. 특히 2008년 한 해에만 3조 7000억 원 가량의 연봉을 받아 세계 최고 연봉자가 된 제임스 사이먼스(James H. Simons)가 운용하는 르네상스 테크놀로지스社의 성공사례로 인해 최근 트레이딩 시스템에 대한 관심은 폭발적으로 증가하고 있다[1]. 트레이딩 시스템이란 투자위험을 줄이면서 수익을 극대화하기 위해 일정한 규칙에 따라 매매를 체계적으로 수행하기 위한 시스템이다. 구체적으로는 과거의 주가자료를 분석해 최적의 수익을 제공하는 포지션 진입규칙(position entry rule)과 청산규칙(position exit rule)을 말한다. 트레이딩 시스템은 전통적 매매방법에 비해 계량화(quantifiablity), 검증성(verifiability), 객관성(objectivity), 일관성(consistency), 자동화(automation)의 장점을 가지고 있기 때문에 실제 투자에서도 중요도가 점점 커지고 있다.

1) 조선일보, 28억 달러 번 사이먼스 사장 작년 헤지펀드 고소득자 1, 2009, 4. 1.

최근에는 고도의 통계처리 기법이나 인공지능 기법들을 이용하여 주가 움직임이 랜덤워크가 아닌 고도의 복잡한 패턴으로 설명될 수 있다는 연구들이 소개되면서, 인공지능을 이용해 트레이딩 알고리즘을 찾아내려는 연구가 활발히 진행되고 있다(McNelis, 2005; 이형용, 2008; 안현철과 이형용, 2009). 그러나 대부분의 기존 연구들은 공통적인 한계를 갖고 있다. 우선 이들 기존 연구들은 대부분 시장을 예측하기 위한 입력변수로 과거의 주가 자체로부터 계산되어지는 가격 기반의 기술적 지표(technical indicators)를 주로 사용하고 있다. 간혹 거래량과 같은 비가격 기술적 지표를 입력변수로 사용하는 기존 연구도 있기는 하지만, 이들 연구에서도 입력변수의 절대 다수는 가격 기반 기술적 지표로 이루어져 있다(Kim, 2003; Kim and Lee, 2004; 이형용, 2008; 안현철과 이형용, 2009 등). 하지만 이는 트레이딩 시스템을 활용하는 현업의 상황과 상당한 괴리가 있다. 현업에서 고려하고 있는 수많은 시장 상태 변수들 중 가격 기반의 지표들이 차지하는 비중은 그다지 높지 않기 때문이다. 또한 기존 연구들은 실제 투자자들이 실전투자에 활용하여 수익을 내는데 활용할 수 있는, 즉 트레이딩과 관련한 사용자의 의사결정을 직접적으로 지원할 수 있는 모형보다는 주가지수의 등락을 정확하게 예측할 수 있는 간접적인 모형을 개발하는데 주로 초점을 맞추고 있다.

이에 본 연구에서는 기존 연구에서 사용되어 온 가격 기반 기술적 지표 외에 주가에 영향을 미치는 다양한 비가격 변수(non-price variables)를 추가하여 트레이딩 시스템의 투자성과를 높이고자 한다. 가격 기반 기술적 지표는 오래 전부터 시장 참여자들 사이에 많이 알려지고 실제 투자에 활용되면서, 이를 활용한 최근의 투자성과는 많이 줄어들고 있는 추세다(Sullivan et al., 1999). 비가격 변

수에는 투자주체별 매매동향이나 선물시장이나 옵션시장의 관련 정보를 포함하며, 이러한 비가격 변수는 최근에 와서 활용 가능해진 정보들로서 기술적 지표에 비해서 주가 예측력이 높은 것으로 알려지고 있다.

또한 주가예측 모형이 실제 투자에 활용되기 위해서는 주가의 등락예측 확률을 높이는 것 보다는 주가예측을 활용한 실제 투자의 수익률이 높아야 한다. 이를 위해서는 주가예측의 성과기준이 투자수익률이 되어야 하며, 투자에 따른 자금관리에 적합한 누적 투자수익률을 목적함수로 사용해야 한다. 이러한 배경에서 본 연구는 투자수익률을 극대화시킬 수 있는 새로운 지능형 시스템을 제안하는데, 이를 위해 이중 임계치 기법을 도입하고자 한다(이재식 외, 2000; 안현철과 이형용, 2009). 이 기법을 적용하게 되면, 예측 값이 확실하지 않을 경우 보유(hold) 의사결정을 내릴 수 있도록 이중 임계치를 설정하여 강력한 매수(buy) 또는 매도(sell) 신호만 활용하여 매매하는 것이 가능해 진다.

본 연구에서 이중 임계치를 적용하게 될 주가지수 등락의 예측모형으로는 최근 가장 활발하게 적용되고 있는 이분류(二分類, binary classification) 모형인 Support Vector Machines 알고리즘을 적용한다. 또한 최적의 이중 임계치를 결정하기 위한 방법으로는 유전자 알고리즘을 사용한다. 본 연구에서는 이러한 제안 시스템을 실제 KOSPI200 지수에 다른 비교모형들과 함께 적용함으로서, 모형의 유용성을 실증적으로 검증해 보고자 하였다.

본 논문은 다음과 같이 구성된다. 우선 제 2장에서는 본 연구와 관련한 이론적 배경을 설명하고, 그 다음 제 3장에서는 본 연구의 제안 시스템의 구조와 작동원리를 제시한다. 이어 제 4장에서는 제안 시스템의 유용성을 검증하기 위한 실험설계에 관한 설명이 이루어질 것이며, 제 5장에서 실증 분석을 통한 실험결과가 종합적으로 제시될 것이다. 끝으로, 마지막 장에서는 본 연구의 의의 및 한계점을 토의하고자 한다.

2. 이론적 배경

본 장에서는 우선 트레이딩 시스템과 관련한 기존 문헌들을 고찰해 보고, 본 연구에서 사용될 2가지 핵심 기법인 Support Vector Machines와 유전자 알고리즘에 대해 살펴보도록 한다.

2.1 트레이딩 시스템

트레이딩 시스템에 대한 연구는 실무에서의 높은 관심과는 반대로 학계에서는 큰 관심을 끌지 못하였다. 1960년대에 정립되기 시작한 효율적 시장가설이 학계에서 큰 호응을 얻고 있었기 때문이다. Alexander(1961)는 주가가 저점으로부터 일정한 비율만큼 오르면 매수하고 고점으로부터 일정 비율만큼 하락하면 매도하는 필터기법(filter rule)을 적용한 결과 비용을 고려하면 추가수익을 낼 수 없음을 증명하였다. Fama(1965)는 주가의 상관관계분석(serial correlation)과 런검정(runs test)을 통해 주가가 효율적임을 주장하였다. 이어지는 효율적 시장가설에 대한 연구 결과들은 주로 가설을 지지하는 결론을 보여주고 있다.

그러나 Granger(1981)에 의해 주가와 같은 불안정한 시계열자료 분석에 선형회귀분석을 적용하는 것이 타당하지 않음이 밝혀진 이후부터는, 인과관계분석(causality) 기법이 등장하면서 계량경제학적 주가예측모형에 대한 연구가 활발히 이루어지기 시작하였다. Caporale and Pittis(1998)는 공적분 검정(cointegration test)을 이용해 시장에서 부분적으로는 주가를 예측할 수 있음을 증명하였

다. 또한 McMillan(2007)은 영국 등 4개국의 주식 시장의 거래량을 입력변수로, 비선형회귀모형을 적용한 결과 선형모형보다 좋은 투자성과를 얻을 수 있음을 보여주었다.

최근에는 인공지능기법을 주가예측에 적용하려는 연구가 활발히 진행되고 있다. 주가예측을 위한 입력변수로는 기본적 분석(fundamental analysis), 기술적 분석(technical analysis) 등이 주로 이용되고 있다. 전자는 기업의 부채비율, 배당률, PER(Price/Earnings Ratio), 경제성장율, 금리, 환율 등 주가에 영향을 미치는 경제변수를 입력변수로 이용하려는 접근방법이고 기술적 지표는 과거의 주가나 거래량 자료를 이용하여 변환과정을 거쳐 이동평균(moving average), 스토캐스틱(stochastic), MACD(Moving Average Convergence and Divergence), 모멘텀(momentum) 등으로 변환된 지표이다. 기본적 분석은 주가를 결정하는 본질적 정보를 이용하는 분석방법이나 경제변수들은 주가에 장기적 영향을 미치고 있기 때문에 단기적 거래를 지향하는 트레이딩 시스템에서는 활용하기에 어려움이 있다. 기술적 지표는 주가자료에서 쉽게 계산해 낼 수 있고 단기적 주가의 움직임을 포착하는 데 적절한 지표로서, 오래 전부터 시장 참여자들에 의해 실제 투자에 많이 활용되고 있다. 한편, 목표함수는 크게 통계적 기준과 비통계적 기준으로 나누어지는데, 예측값의 정확도를 평가하기 위해 통계적 기준으로는 MAE(mean absolute error), RMSE(root mean square error) 등의 지표들이 사용되며, 비통계적 기준으로는 적중률(hit ratio)이나 수익률(rate of return)이 주로 사용된다.

주가지수 예측과 관련한 연구들을 우선 살펴보면, Yudong and Lenan(2009)은 BP-ANN을 이용하여 S&P 500 주가지수를 예측하고 있다. 입력변수는 S&P 500 주가지수의 기술적 지표 10개를 이용하고 있으며, 예측값의 MSE(Mean-Squared Error)를 목표함수로 이용하고 있다. Schulmeister(2009)는 2,580개의 기술적 지표를 이용하여 S&P 500 주가를 분석하고 거래별 수익률을 단순 덧셈하는 총수익의 관점에서 예측모형을 평가하였다. 그 결과 1990년대에 접어들면서 기술적 지표의 수익성이 하락함을 확인할 수 있었다. Atsalakis and Valavanis(2009b)는 주가의 3일 이동평균값을 입력변수로 하는 Neuro-Fuzzy 모형을 이용하여 내일의 상승, 하락을 예측하고 그 결과를 적중률과 단순수익률 기준으로 기존의 예측방법들과 비교하여 제안 모형의 성과가 우수함을 보여주고 있다.

트레이딩 시스템과 관련한 연구로는 대표적으로 Núñez-Letamendia(2007), Bao and Yang(2008), Chavarnakul and Enke(2009), 그리고 안현철과 이형용(2009) 등이 있다. 우선 Núñez-Letamendia(2007)는 유전자 알고리즘을 이용하여 기술적 트레이딩 시스템을 최적화하였다. 그러나 입력변수로는 기술적 지표 중 가장 잘 알려진 두 이동평균선의 교차만을 사용하고 있고, 적합함수(fitness function)로 비록 누적수익률함수를 사용하고 있지만, 연구의 주목적이 최적의 기술적 규칙을 찾기보다는 통제 파라미터 값에 대한 유전자 알고리즘의 강인성 검정(robustness test)에 초점이 맞추어져 있다는 한계가 있다. 또한, 이 연구에서 제안된 시스템은 매일매일 매수 또는 매도신호를 발생하도록 설계되어 있다. 따라서, 실제 투자에 활용하는데 있어, 다소 무리가 따른다는 한계가 있다.

Bao and Yang(2008)은 고차원 표현(high-level representation)을 확률모형과 결합하는 인공지능 트레이딩 시스템을 개발하였다. 입력변수로는 과거의 주가자료와 4개의 기술적 지표를 사용하고 있어, 비가격 변수를 전혀 고려하지 않았다는 한계가 있다.

Chavarnakul and Enke(2009)는 인공신경망, 퍼지로직과 유전자알고리즘을 이용한 트레이딩시스템을 제시하였다. 이들의 연구에서 입력변수는 과거의 주가와 거래량을 결합하여 사용하고 있고, 매수매도 신호도 매매를 하지 않는 중립지대를 두고 있지만 중립지대를 결정하는 threshold 값을 비교적 큰 값인 +0.5와 -0.5로 주관적으로 결정하여 사용하고 있다.

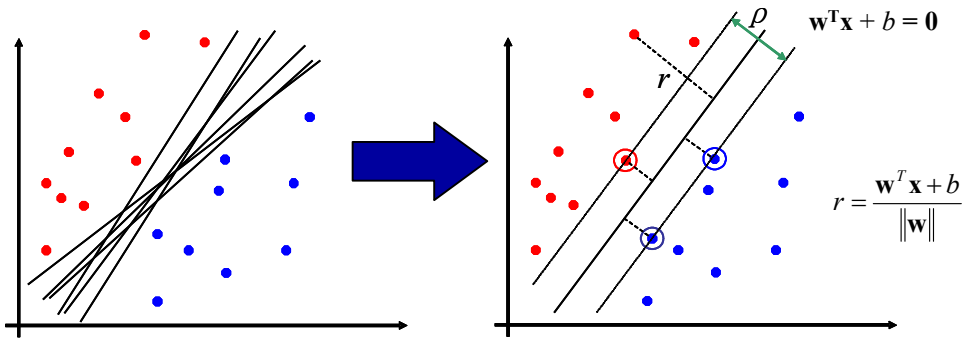
안현철과 이형용(2009)은 이중 임계치 모형을 사용하여 익일의 주가패턴이 확실치 않은 경우 보류할 수 있도록 중립지대를 설정하여 실제 투자에 활용할 수 있도록 고안하였다. 본 연구 역시 입력변수로 12개의 기술적 지표를 사용하고 있는데, Bao and Yang(2008)의 연구와 마찬가지로 비가격 변수를 전혀 고려하지 않았다는 한계가 존재한다.

Atsalakis and Valavanis(2009a)는 다양한 인공지능기법을 주식시장 예측에 사용한 100편 이상의 연구논문을 조사하여 입력변수(input variables), 예측방법(forecasting methodology), 성과측정방법(performance measures) 등의 기준에 따라 분류작업을 시도하였다. 여러 논문에서 분석대상이 된 주식시장은 미국 뿐만 아니라 유럽, 아시아와 남미까지 24개 이상의 주식시장으로 골고루 분포하고 있다. 입력변수의 수는 4~10개 사이가 대부분이지만 2개를 사용한 연구부터 61개를 사용한 연구까지 많은 차이를 보이고 있다. 입력변수로는 30% 이상이 주가나 주가지수의 과거 종가자료를 사용하고 있음을 보였다. 예측방법으로 가장 많이 활용된 기법은 인공신경망이었다. 전체 분석대상의 90% 이상이 인공신경망 혹은 인공신경망의 변형모형을 사용한 것으로 나타났다. 반면 인공신경망보다 더 진보된 방법론으로 알려진 SVM의 경우, 전체 분석대상 중 단 한 편만 주식시장 예측에 적용한 것으로 나타났다.

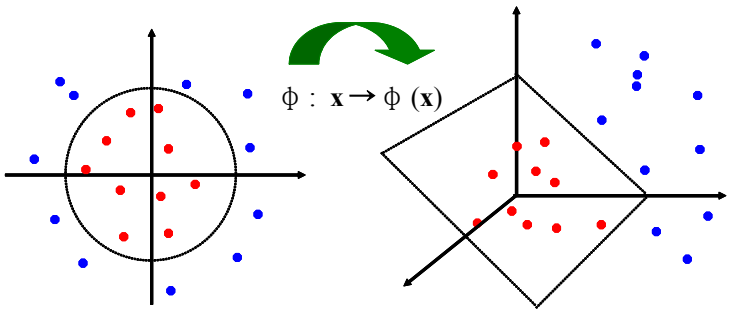
기본적 변수를 이용한 예측모형을 제시한 논문으로는 Atiya et al.(1997), Casas(2001), Olson and Mossman(2003) 등이 있다. 이 중, Olson and Mossman(2003)은 캐나다 기업 2,352개를 상대로 61개의 회계지표를 이용하여 인공신경망기법이 회귀분석보다 우수한 예측성과를 보임을 제시하였다. 그러나 이러한 기본적 변수를 이용한 연구들은 예측기간이 1년 단위로 되어있어 트레이딩 목적에는 적합하지 않다는 한계를 갖는다.

2.2 Support Vector Machines

Support Vector Machines(이하 SVM)은 1995년 러시아의 통계학자인 Vapnik이 처음 제안한 분류기법으로, 인공신경망과 마찬가지로 복잡한 비선형 관계를 갖는 이분류 문제를 해결하는데 적합한 분류 기법이다. 복잡한 분류문제에서 상당히 우수한 예측 정확도를 보인다는 점은 인공신경망과 동일하나, SVM은 상대적으로 여러 측면에서 장점을 갖고 있다. 우선, SVM은 '경험적 위험 최소화(empirical risk management)'를 추구하는 인공신경망과 달리 '구조적 위험(structural risk)'을 최소화하는 방향으로 학습을 수행하도록 설계되어 있다. 때문에, SVM은 상대적으로 과적합화(overfitting) 문제에서 자유로울 수 있다. 또한, 추정해야 할 많은 가중치들로 인해 학습 시 많은 양의 데이터를 요구하는 인공신경망과 달리, SVM은 서포트 벡터(support vector)라 불리는 소수의 데이터만을 최종적으로 학습에 사용하기 때문에, 일반적으로 적은 양의 학습 데이터로도 우수한 예측성과를 나타낸다. 아울러, SVM은 인공신경망에 비해 조정해야 할 파라미터의 수가 많지 않아, 상대적으로 간단하게 학습에 영향을 미치는 요소들을 규명할 수 있다(Kim, 2003; 안현철 외, 2005b; 안현철과



(a) 최대 마진 분류



(b) 고차원 공간으로의 데이터 위치 이동

<그림 1> SVM의 작동원리

이형용, 2009).

SVM의 작동 원리는 크게 2가지로 설명될 수 있다. 첫 번째는 '최대 마진 분류(maximum margin classification)'의 원리이다. <그림 1>의 (a)에서 볼 수 있듯이, SVM은 기본적으로 서포트 벡터라 불리는 두 집단의 경계에 있는 점들로부터 가장 멀리 떨어져 있는 분류기(classifier)를 찾도록 설계되어 있다. 이를 통해, 두 집단을 가장 잘 나눌 수 있는 선형 분류기를 선형적으로 제약된 이차계획(QP) 문제(linearly constrained quadratic programming)로 정형화하여 찾을 수 있도록 설계되어 있다(Vapnik, 1998).

두 번째 SVM의 핵심 작동원리는 '고차원 공간으로의 데이터 위치이동(mapping data into higher dimensional space)'이다. <그림 1>의 (b)가 이러한 두 번째 원리를 도식적으로 잘 설명하고 있다. 이 그림은 X축과 Y축으로 구성된 2차원 공간에서 비선형 분류기(원)로만 구분되던 2개의 집단

이 3차원 공간으로 전이되자 선형 분류기(평면)에 의해 구분이 되는 모습을 나타내고 있다. 이처럼 저차원 공간에서 선형 분류기로 분류되지 않는 문제들도 보다 높은 차원의 공간으로 사상시킬 경우, 선형 분류기로 더 잘 분류될 수 있다는 것이 SVM의 또 다른 핵심 작동원리이다(Vapnik, 1998).

그런데, 이 때 원래 데이터를 고차원 공간으로 사상시킴으로써 특징공간 내에 선형으로 분리가능한 입력 데이터셋을 만드는 전이함수가 필요하게 되는데, SVM에서는 이러한 전이함수를 커널함수(kernel function)라 부른다. 어떠한 커널함수를 선택할 것인가 하는 것은 주어진 문제에 따라 다르며, SVM을 적용하는 데 있어서 가장 중요한 요소이다(Tay and Cao, 2002; Kim, 2003; 안현철 외, 2005b). 보통 커널함수로는 아래 식 (1) - 식 (3)과 같은 선형함수(linear function), 다항함수(polynomial function), 그리고 가우시안 RBF 함수(Gaussian radial basis function)가 가장 많이 사용된다.

$$K(x, y) = xy \tag{1}$$

$$K(x, y) = (xy+1)^d \, (d : \text{다항함수의 차수}) \tag{2}$$

$$K(x, y) = e^{-\frac{(x-y)^2}{\sigma^2}}$$
$$(\sigma^2 : \text{가우시안 RBF커널함수의 대역폭}) \tag{3}$$

앞서 살펴본 바와 같이 SVM은 기본적으로 두 집단을 가장 잘 분류할 수 있는 '선형 분류기'를 찾는데 초점을 맞추고 있는 기법이다. 때문에, SVM 모형을 이용해 분석의 대상이 되는 데이터가 어느 집단에 속하는지에 대해서는 예측할 수 있지만, 각 집단의 소속확률에 대해서는 산출하지 못한다. 이는 다른 이분류 모형인 로지스틱 회귀모형이나 인공신경망과 비교해, SVM이 갖는 상대적인 약점이라고 할 수 있을 것이다. 이를 극복하기 위해 지금까지 여러 연구자들이 추정확률(estimated probability)을 결과로 산출할 수 있는 변형된 SVM 모형을 제안하였다(Platt et al., 1999; Sollich, 2002; 홍태호와 신택수, 2005). 본 연구 역시 예측 정확도가 높은 SVM을 트레이딩 시스템에 적용하고자 하였는데, 이를 위해서는 각 집단의 추정 소속확률을 제공할 수 있는 변형된 SVM의 적용이 요구되었다. 이를 위해 본 연구에서는 Platt et al.(1999)이 제안한 방법에 기초한 변형 SVM 모형을 적용하였다. 기본적으로 이 방법은 분류기로부터 데이터가 얼마나 멀리 떨어져 있는지를 활용하여, 해당 집단에 소속될 사후 조건부 확률(conditional posterior probability)을 추정하는 형태로 이루어져 있다. 본 방법은 SVM의 실험용 소프트웨어인 LIBSVM의 2.6버전 이후에 옵션으로 내장되어 있다.[2)]

2.3 유전자 알고리즘

유전자 알고리즘은 찰스 다윈(Charles Darwin)의 적자생존의 원리와 멘델(Mendel)의 유전법칙을 응용한 최적화 기법으로, 생물의 진화과정을 모사하여 적응적으로 탐색공간을 탐색하여 최적 또는 유사 최적해를 찾아내는 탐색기법이다(홍승현과 신경식, 2003). 유전자 알고리즘은 점에 의한 탐색이 아닌 개체들이 모여 이룬 개체군에 의해 병렬적으로 탐색이 이루어진다는 점에서 기존의 최적화 알고리즘과 차별화된다. 또한 탐색의 방향이나 영역이 초기값에 과도하게 의존하지 않고, 세대에 따라 확률적으로 변화한다는 점에서 전역 최적

2) http://www.csie.ntu.edu.tw/~cjlin/libsvmtools 참고

화가 가능하다는 장점을 가진다(옥중경과 김경재, 2009).

유전자 알고리즘은 다음과 같은 프로세스에 의해 작동된다. 우선 본격적인 진화에 앞서 유전자 알고리즘은 해결하고자 하는 문제에 대한 해(solution)를 적당한 형태의 이진 벡터값으로 표현하게 되는데, 이렇게 표현된 하나의 개체를 염색체(chromosome)라 부른다. 유전자 알고리즘에서는 전체 탐색공간 내에서 n개의 염색체들을 임의로 선택하여 이 값들을 진화시켜 나가게 되는데, 이러한 염색체들의 집합은 모집단(population)으로 호칭된다. 본격적인 유전자 알고리즘의 진화과정이 시작되기 전, 모집단을 구성하는 모든 염색체들의 값은 임의값으로 초기화된다.

1단계의 초기화 작업이 끝나면, 다음 단계에서는 이렇게 생성된 모집단이 문제해결에 얼마나 적합한지를 평가하게 되는데, 이른바 적합도 함수(fitness function)라는 평가기준을 이용해 각 개체(염색체)의 적합도를 평가하게 된다. 이렇게 각 개체의 적합도가 평가되고 나면, 유전자 알고리즘은 평가된 개체집단을 확률적으로 선택(selection)하거나, 교배(crossover), 혹은 돌연변이(mutation) 등의 유전적 조작을 통해 이전 세대와 다른 새로운 개체들로 구성된 새로운 세대를 생성하게 된다.

이렇게 생성된 새로운 세대의 개체집단은 다시 적합도 함수에 의해 평가되며, 그 결과에 의해 다시 유전적 조작이 이루어지게 된다. 이러한 평가와 유전적 조작 작업은 종결조건에 도달할 때까지 즉, 목표로 한 적합도 수준에 도달하거나 사전에 정해진 최대 진화수에 도달할 때까지 반복적으로 이루어지게 된다. 그렇게 하여, 유전자 알고리즘은 전체 생성된 세대 중에서, 가장 최적의 적합 정도를 나타낸 개체를 최종적으로 선택하여, 그 결과를 전역 혹은 유사전역 최적해로 도출하게 된다(안현철 외, 2005a).

이러한 유전자 알고리즘은 여러 장점으로 인해 지금까지 많은 연구에서 다른 인공지능 기법의 파라미터를 최적화하는데 널리 적용되어 왔다(Ahn et al., 2003; Kim and Han, 2003; 안현철 외, 2005a 등). 최근에는 이형용(2008), 안현철과 이형용(2009) 등의 연구에서 임계치를 최적화하는데 사용되고 있는데, 본 연구에서도 임계치의 최적화 도구로 유전자 알고리즘을 적용한다.

3. 제안 시스템

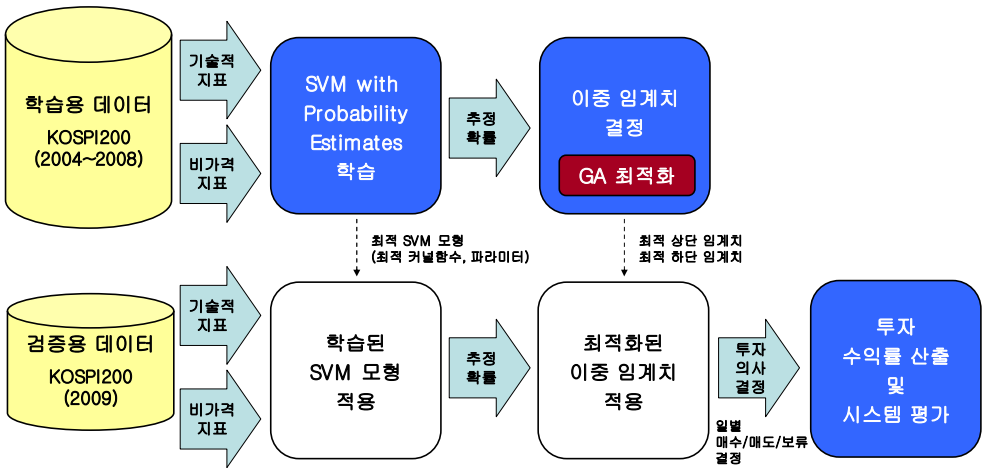
앞서 이론적 배경을 통해 살펴본 바와 같이 기존 트레이딩 시스템과 관련된 연구들은 가격에 기반한 기술적 지표에만 주로 의존하고 있다는 점, 그리고 실제 트레이딩 보다는 주식 시장의 정확한 예측에 초점을 맞추고 있다는 점에서 한계가 있었다. 또한 SVM이 인공신경망과 비교해 여러모로 더 우수한 성능을 보이는 것으로 여러 문헌에서 소개되고 있음에도 불구하고, 지금까지는 인공신경망에 비교해 SVM을 활용한 연구가 소수에 불과했다는 점 역시 개선이 필요한 부분이다.

이에 본 연구에서는 이러한 기존 연구의 한계점을 극복할 수 있는 대안으로 새로운 개념의 지능형 트레이딩 시스템의 모형을 제안하고자 한다. 제안 시스템은 크게 3단계에 의해 구현되도록 설계되었는데, 제안시스템의 전체적인 진행체계는 다음의 <그림 2>와 같다.

작동 원리에 대한 상세한 설명은 아래와 같다.

3.1 추정확률을 산출하는 SVM 모형의 학습

첫 번째 단계로 제안 시스템은 SVM을 활용해



<그림 2> 제안 시스템의 진행체계

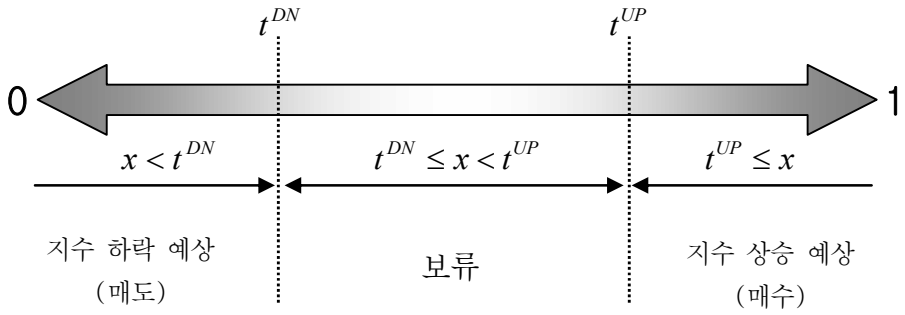
학습용 데이터에 대한 이분류 예측모형을 구축하게 된다. 이 때, 입력변수로는 기존 연구들에서 주로 제시된 기술적 지표 뿐 아니라, 실무에서 활용되는 각종 비가격 지표들도 함께 사용하여, 예측성과 개선을 도모하도록 한다. 그리고 예측모형으로는 SVM, 그 중에서도 예측결과를 0~1사이의 '추정확률' 값으로 산출할 수 있는 변형 SVM 모형을 활용한다. 이를 통해 전통적으로 많이 적용되어 온 인공신경망 모형과 비교해, 보다 안정적이면서도 우수한 예측력을 갖춘 분류모형을 구축한다.

SVM의 경우, 어떤 커널함수를 사용하는지, 그리고 커널함수에 포함되는 파라미터 값들을 어떻게 설정하는가에 따라 성과가 달라질 수 있다. 이에 본 시스템에서는 여러 경우에 대해 모두 실험을 진행해 보고, 가장 우수한 성과를 보이는 커널함수와 파라미터 값을 탐색하여, 최적의 SVM 모형을 결정할 수 있도록 하였다.

3.2 이중 임계치 결정

문헌에서 cut-off value 혹은 threshold로 표현되는 임계치는 보통 0에서 1사이의 확률값으로 제공되는 예측결과를 토대로 하여 그 결과를 어떤 소속집단으로 해석할 지 판단할 때 기준으로 활용되는 값을 의미한다. 예를 들어, 내일의 주가지수가 오를지 내릴지를 판단하는 모형을 구축했는데, 예측결과(확률값)가 x로 나왔다고 하자. 이 때 만약 이 모형이 내일의 주가가 하락하는 경우를 0으로 학습하고, 내일의 주가가 동일하거나 상승하는 경우를 1로 학습했다고 한다면, 기준 임계치를 t라고 할 때, $x \geq t$라면 익일의 주가가 유지 혹은 상승할 것으로 예측하고, $x < t$일 경우, 익일의 주가가 하락할 것으로 예측하게 되는 것이다. 보통 관련 연구에서는 이 값을 0.5로 설정하여 사용한다.

그런데 이렇게 하나의 기준 임계치(single cut-off value)를 사용하는 것은 두 집단(상승/하락)을 판단하는데에는 유용하게 적용될 수 있으나, 투자의사결정처럼 최종 예측 결과가 세 개의 집단(매수/매도/보류)으로 구분되는 문제에서는 효과적으로 적용될 수 없다. 때문에 이를 해결하기 위하여 최근 일부 연구들에서 이중 임계치 방식(two-



<그림 3> 이중 임계치 방식

threshold mechanism)을 도입하고 있다(이재식 외, 2000; 안현철과 이형용, 2009; Ahn et al., 2010). 상단의 <그림 3>은 이중 임계치 방식의 원리를 나타낸다.

단일 임계치는 적용 시 내일의 주가지수가 오를 것인지(매수), 내릴 것인지(매도)에 대해서만 판단을 할 수 있도록 되어 있다. 하지만, 상기 그림에서 볼 수 있듯이, 이중 임계치를 적용하게 되면 내일의 주가지수 방향 패턴이 확실치 않은 경우 보류라는 판단을 내릴 수 있다. 즉, 상단 임계치(upper threshold)를 t^{UP}라하고, 하단 임계치(lower threshold)를 t^{DN}이라고 할 때, 예측결과 x가 $x \geq t^{UP}$를 만족한다면 이는 내일의 주가지수가 상승할 것이라는 의미이므로 오늘 주가지수를 매수하라는 신호로 해석할 수 있다. 반대로 $x < t^{DN}$을 만족하면 이는 익일 주가지수 하락을 의미하기 때문에, 오늘은 주가지수를 매도하는 전략을 취하라는 의미가 된다. 하지만 만약 $t^{UP} > x \geq t^{DN}$인 상황이라면, 이 경우는 주가지수의 움직임을 명확하게 판단하기 어려운 상황이라고 할 수 있다. 이 때는 판단불가(보류)로 판별하여, 특별한 대응을 하지 않게끔 하는 것이다. 이렇게 하면, 정확한 판단을 통해 투자 수익의 관점에서 더 높은 수익을 창출시킬 수 있을 뿐만 아니라, 거래횟수를 줄여 거래비용을 감소시키는 효과도 누릴 수 있다.

그런데 앞서 소개한 단일 임계치와 마찬가지로, 이중 임계치 역시 설계자가 직관에 의해 임의로 값을 설정해야 하는 어려움이 있다. 이러한 문제를 해결하기 위해, 본 연구에서는 유전자 알고리즘을 이용해 이중 임계치 값을 최적화하고자 한다. 아래의 <그림 4>는 본 연구의 제안 시스템에서 활용된 염색체의 구조를 나타내고 있다. 그림에서 볼 수 있듯이, 염색체는 14비트가 할당된 2개의 변수로 구성되는데 앞의 변수는 '하단 임계치'를, 그리고 뒤의 변수는 '상단 임계치와 하단 임계치의 차이'를 나타내도록 하였다. 이렇게 설계함으로서, 항상 상단 임계치는 하단 임계치보다 크거나 같은 값을 갖게끔 하였다.

아울러, 본 연구에서는 유전자 알고리즘을 위한 적합도 함수로서 모형구축용 데이터셋에 대한 수익률을 활용하였다. 이 때 수익률은 다음과 같이 계산된다. 익일 지수 상승을 예측한 날에 당일 주가지수로 매수하고, 보유 후 익일 지수 하락을 예측하는 날에 매도한다고 가정할 때, 주어진 기간 중에 발생한 n번의 매매 활동 중에서 k번째 매매 활동에 대한 수익률 R(k)는 다음의 식 (4)와 같다.

모집단 (1세대)	염색체 구조													
	하단 임계치(T^{DN})							임계치 간 차이(GAP)						
	T^{DN}_1	T^{DN}_2	T^{DN}_3	…	T^{DN}_{12}	T^{DN}_{13}	T^{DN}_{14}	GAP_1	GAP_2	GAP_3	…	GAP_{12}	GAP_{13}	GAP_{14}
염색체 1	1	1	0	…	1	1	1	1	0	0	…	1	0	0
염색체 2	0	1	1	…	0	1	0	0	0	1	…	0	1	0
염색체 3	1	1	1	…	1	0	0	1	0	1	…	1	0	1
⋮					⋮					⋮				
염색체 *m*	1	1	0	…	1	1	1	1	0	1	…	0	1	0

<그림 4> 제안 시스템의 염색체 구조

$$R(k) = \frac{I(t_k^2) - I(t_k^1)}{I(t_k^1)} \quad (4)$$

$I(t)$: t시점의 지수, t_k^1 : k번째 매매활동의 시작일(매수일), t_k^2 : k번째 매매활동의 종료일(매도일)

단, 이 때 마지막 매매(n)의 경우에는 매수 후 자료의 마지막 날 강제로 매도된 것으로 가정하여 수익률을 계산한다. 전체 수익률(TR)은 다음의 식 (5)와 같이 계산된다.

$$TR = \prod_{k=1}^{n}(1 + R(k)) \quad (5)$$

그렇게 하여, 이 단계에서는 상기 전체 수익률을 극대화하는 최적의 상단 및 하단 임계치를 유전자 알고리즘을 이용해 탐색하게 된다.

3.3 일반화 검증

상기 과정을 통해 유전자 알고리즘을 이용한 학습까지 모두 마무리되면, 마지막 3단계에서는 최종적으로 구축된 SVM 모형과 최적의 상/하단 임계치를 검증용 데이터에 적용해 봄으로서, 유전자 알고리즘을 통해 탐색된 값들이 과연 일반화된 결과인지를 마지막으로 검증하게 된다. 간혹 유전자 알고리즘을 통해 최적화된 결과가 학습용 데이터에 대해서는 잘 작동하지만, 새로운 데이터에 대해서는 잘 적용되지 않는 과적합 현상이 나타나기도 한다. 본 과정은 이러한 과적합의 여부를 판단하고, 최종적으로 개발된 시스템의 성능이 유용한지를 검증하기 위한 단계라고 할 수 있다.

4. 실험 설계

4.1 실험 데이터

본 연구에서는 제안 시스템의 우수성을 검증하기 위해, KOSPI200 지수의 일별자료에 적용해 보았다. 본 연구를 위해 수집된 데이터는 2004년 5월 17일부터 2009년 12월 31일까지의 약 6년치 데이터였다.[3] 이 중, 전체 데이터의 약 80%를 차지하

<표 1> 데이터 구성 현황

구 분	연도	익일하락	익일상승	총합계	비중
학습용	2004	69	69	138	79.84%
	2005	100	100	200	
	2006	109	109	218	
	2007	101	101	202	
	2008	122	122	244	
검증용	2009	112	141	253	20.16%
총합계		617	786	1255	100.0%

는 2004년부터 2008년까지의 데이터를 학습용으로 활용하였고, 나머지 2009년의 데이터를 검증용으로 활용하였다. 학습용 데이터의 경우, 학습의 왜곡을 막기 위해 주가지수의 상승사례와 하락사례의 비율이 서로 1 : 1이 되도록 조정되어야 한다. 이에 '무작위 표본 추출'을 통해, 학습용 데이터가 각 년도별로 1 : 1의 비율로 구성되도록 전처리하였다. <표 1>은 최종적으로 시스템 구축 및 검증에 활용된 데이터의 구성 현황을 나타내고 있다.

본 시스템에서 고려된 입력변수들은 총 48개의 변수들이었다. 이 중 15개의 변수는 Kim(2003), 이형용(2008) 등의 연구에서 사용된 기술적 지표들이며, 나머지 33개의 변수는 20년 이상 트레이딩 시스템 분야에 종사한 전문가로부터 추천된 비가격 변수들로 구성되었다. 이러한 48개의 후보 변수들 중에서 KOSPI200 지수의 등락을 가장 잘 설명하는 변수들을 선정하였는데, 독립표본 t-검정 결과 유의수준 70% 이상에서 유의하면서, 전문가들이 타당하다고 추천한 변수만 추려 총 15개의 변수를 SVM의 입력변수로 최종 선정하였다. 다음의 <표 2>는 최종 선택된 15개 변수의 명칭과 의미, 산식 등을 정리하여 제시하고 있다. 표에서 볼 수 있듯이, 최종 선택된 15개의 변수 중 단 4개만이 기존 연구에서 적용되어 온 기술적 지표임을 확인할 수 있다. 이를 통해 학술연구에서 지금까지 많이 활용되지 못했던 비가격 지표들이 시장을 예측하는데 있어 상당히 중요한 역할을 할 수 있음을 미루어 짐작할 수 있다.

3) 본 연구에서 사용된 데이터가 2004년 1월 1일이 아닌 2004년 5월 17일부터 시작한 이유는 연구에서 사용된 일부 비가격 변수들이 측정되기 시작한 시점이 바로 이 때부터이기 때문이다.

4.2 실험 설계 및 시스템 개발

제안 시스템에 포함된 SVM은 선형 그리고 가우시안(Gaussian) RBF의 2가지 커널함수를 적용하여, 학습용 데이터를 기준으로 가장 우수한 성과를 보이는 커널함수를 최종적으로 선정하였다. 또한 Tay and Cao(2002), Kim(2003), 안현철 외(2005b)는 SVM의 성과를 결정짓는데 있어서, 상한계수 C나 σ^2와 같은 커널함수 내 매개변수들의 값에 대한 설정이 중대한 영향을 미칠 수 있음을 지적하였다. 만약 이러한 매개변수 값들이 적절하게 설정되지 않은 경우, SVM은 과적합 되거나 혹은 불충분적합이 될 수 있기 때문이다. 이에, 본 연구에서도 상기 매개변수들의 값을 다양하게 바꾸어가면서 실험하여, 가장 우수한 성과를 보이는

<표 2> 선정된 입력변수 현황

변수코드	변수명	의미(산식)
X5	KP/외	외국인 주식 순매수액
X6	KP/개	개인 주식 순매수액
X14	C/개	개인 콜옵션 순매수액
X16	F 종가	선물가격 종가
X18	F거래대금	선물 거래대금
X21	만기	만기월
X22	F 시가	선물가격 시가
X23	F 고가	선물가격 고가
X24	F 저가	선물가격 저가
X27	OPTION/개	개인의 풋옵션 매수금액에서 개인의 콜옵션 매수금액을 뺀 차이
X28	KPF/개	개인의 주식매수대금과 선물매수대금의 합
Y2	Stochastic %D	$\frac{\sum_{i=0}^{n-1}\%K_{t-i}}{n}$ where $\%K_t = \frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}} \times 100$ C_t : t시점 종가, $HH(LL)_{t-n}$: t시점 전 n일간 최고(최저)가
Y5	Price Rate-of-Change	$\frac{C_t}{C_{t-n}} \times 100$
Y10	Price Oscillator	$\frac{MA_5 - MA_{10}}{MA_5}$ where MA_t : 최근 t일간 이동평균
Y13	Disparity 10	$\frac{C_t}{MA_{10}} \times 100$

매개변수 값들을 최종적으로 선택하였다.

유전자 알고리즘의 각종 통제변수들과 관련해서는 본 연구에서는 탐색해야 할 공간이 그다지 크지 않은 공간임을 감안해, 모집단의 크기를 50개체로 설정하고, 교배율과 돌연변이율은 각각 50%와 10%로 설정하였다. 그리고, 종료조건으로는 총 1000회의 연산(즉, 20세대)을 시도하게끔 설정 하였다.

아울러, 제안 시스템의 성과를 좀 더 깊이 있게 분석하기 위해, 몇 가지 비교모형을 함께 실험해 보았다. 우선 본 시스템은 성과의 향상을 위해 SVM을 도입해 적용하였는데, 이 효과를 정밀하게 검증하기 위해 로지스틱 회귀모형(LOGIT, logistic regression) 및 인공신경망(artificial neural network) 모형을 추가로 적용하여 실험해 보았다. 아울러, 이중 임계치의 효과도 검증해 보기 위하여, 로지스틱 회귀모형, 인공신경망, SVM의 각 개별 모형에 이중 임계치가 아닌 단일 임계치(0.5)를 적용하여, 매일 매일 매수/매도를 예측해 실행했을 경우의 수익률도 함께 구해 보았다. 이렇게 함으로서, (1) LOGIT+단일 임계치 모형, (2) ANN+단일 임계치 모형, (3) SVM+단일 임계치 모형, (4) LOGIT+이중 임계치 모형, (5) ANN+이중 임계치 모형, (6) SVM+이중 임계치 모형(제안 시스템)의 총 6가지 경우에 대한 실험을 모두 수행해 보고 결과를 비교하였다.

실험을 위한 프로토타입 시스템은 MS Excel의 VBA(Visual Basic for Applications)를 이용해 구현되었다. 프로토타입 시스템은 SVM을 통해 예측값이 생성될 경우, 이를 임계치에 의해 투자의사결정 신호로 해석하고, 이 신호에 기반해 예상 투자수익률을 계산해 낼 수 있도록 설계되었다. 추정확률값을 산출할 수 있는 변형 SVM 모형은 LIBSVM 2.91버전을 이용해 구축되었으며(Chang and Lin, 2001), 이중 임계치를 최적화하기 위한 유전자 알고리즘은 상용 소프트웨어인 Evolver 4.08버전을 이용해 구현되었다.

비교모형인 로지스틱 회귀모형은 PASW Statistics 17버전을 이용해 도출되었으며, 인공신경망 모형의 경우 Neuroshell 4.0버전을 이용해 실험하였다. 로지스틱 회귀모형에서 입력변수의 선정방식은 입력(enter), 전진 선택(forward selection), 후진 선택(backward selection)의 3가지 방식을 모두 실험하여 가장 우수한 성과를 보이는 방식을 채택하였다. 인공신경망에 대해서는 입력층과 출력층 사이에 은닉층을 1개 포함하는 3계층 역전파 네트워크(three layer back propagation network)를 적용하였다. 인공신경망의 학습율과 모멘텀율은 모두 0.1로 설정하였으며, 은닉층과 출력층의 노드들은 시그모이드 전이함수(sigmoid transfer function)를 사용하게끔 설계하였다. 은닉층의 노드수와 관련해서는 8, 16, 24, 32 등 4가지 경우를 모두 적용해 보았으며, 그 중에서 가장 우수한 결과를 보이는 설정을 선택하였다. 아울러, 과적합화를 막기 위한 테스트용 데이터는 학습용 데이터 중 2008년 1월~12월까지의 데이터를 사용하였고, 테스트용 데이터의 오류가 최저값에 도달한 뒤 50,000회 반복해도 성과가 더 이상 개선되지 않을 경우 학습을 중지하도록 설정하였다.

5. 실험 결과

전술한 바와 같이, 본 연구의 제안 시스템은 SVM을 이용해 주가지수의 등락예측 모형을 먼저 구축하도록 설계되어 있다. 이에 확보된 지난 6년간의 KOSPI200 지수 예측과 관련한 데이터를 이용하여, SVM 모형을 먼저 학습하였다. 다음의 <표 3>은 다양한 조건 하에서 수행된 SVM 모형의 학습결과를 나타내고 있다.

<표 3>에서 볼 수 있듯이, SVM 모형 중 가장 우수한 예측성과를 보인 설정은 가우시안 RBF 커널을 사용하면서, C = 100과 σ^2 = 100을 적용한 경우였다. 이 경우 학습용 데이터에 대해 61.98%의 예측정확도를 보이면서, 검증용 데이터에 대해서도 56.92%의 높은 예측정확도를 보이는 것으로 나타났다. 이는 KOSPI200 주가지수 등락 예측에 SVM을 적용한 안현철 외(2009)의 연구와 비교해 볼 때, 월등히 예측성과가 높아진 것이다. 이러한 현상의 원인으로는 여러 가지가 제시될 수 있겠으나, 기존 안현철 외(2009)의 연구에서는 단순히 기술적 지표만 입력변수로 사용한 반면 본 연구에서는 기술적 지표 외에 비가격 변수들까지 함께 활용했다는 점이 주요 원인들 중 하나로 추정된다.

또 한가지 흥미로운 결과는 커널함수의 종류에 관계없이 C가 클수록 예측정확도가 대체로 더 높게 나타나는 경향이 보인다는 점이다. 이론적으로 SVM의 C 파라미터는 선형으로 완벽하게 구분되지 않는 문제에 SVM을 적용하기 위한 조정 파라미터(regularization parameter)로서, 두 집단간 구분되는 패턴이 불분명한 상황일 때 C 값이 커지는 특징이 있다. 이러한 점을 고려해 보면, 본 연구의 대상이 되고 있는 'KOSPI200 지수의 등락 구분'은 패턴이 상당히 불안정하고, 복잡성이 높은 이분류 문제에 해당된다는 점을 미루어 짐작할 수

<표 3> SVM 학습결과

구 분			학습용 데이터	검증용 데이터
선형커널	c = 1		50.00%	55.73%
	c = 10		49.30%	56.13%
	c = 33		50.30%	54.15%
	c = 55		51.00%	53.36%
	c = 78		54.19%	51.78%
	c = 100		54.59%	47.43%
가우시안RBF커널	c = 1	σ^2 = 1	3.09%	55.73%
		σ^2 = 25	60.38%	50.99%
		σ^2 = 50	57.58%	54.15%
		σ^2 = 75	55.69%	53.36%
		σ^2 = 100	55.49%	53.36%
	c = 10	σ^2 = 1	50.00%	55.73%
		σ^2 = 25	48.40%	53.36%
		σ^2 = 50	61.38%	52.96%
		σ^2 = 75	61.68%	53.36%
		σ^2 = 100	60.68%	52.57%
	c = 33	σ^2 = 1	50.00%	55.73%
		σ^2 = 25	27.94%	49.80%
		σ^2 = 50	50.30%	56.13%
		σ^2 = 75	59.38%	55.34%
		σ^2 = 100	58.38%	51.78%
	c = 55	σ^2 = 1	50.00%	55.73%
		σ^2 = 25	24.45%	50.20%
		σ^2 = 50	50.00%	55.73%
		σ^2 = 75	60.78%	52.96%
		σ^2 = 100	57.98%	54.55%
	c = 78	σ^2 = 1	50.00%	55.73%
		σ^2 = 25	22.46%	51.38%
		σ^2 = 50	36.03%	52.57%
		σ^2 = 75	60.68%	56.52%
		σ^2 = 100	61.98%	54.55%
	c = 100	σ^2 = 1	50.00%	55.73%
		σ^2 = 25	22.65%	50.99%
		σ^2 = 50	33.33%	51.78%
		σ^2 = 75	59.18%	55.73%
		σ^2 = 100	61.98%	56.92%

<표 4> 단일모형 실험결과

모형구분	학습용	테스트용	검증용	비고
LOGIT	55.59%		49.41%	변수선택법 : 후진선택(총 4개의 입력변수 선정)
ANN	52.90%	49.18%	53.36%	은닉층 노드의 수 : 16
SVM	61.98%		56.92%	가우시안 RBF 커널, C = 100, σ2 = 100

있다.

SVM 실험에 이어, 비교모형인 로지스틱 회귀모형과 인공신경망에 대해서도 동일한 데이터를 기준으로 실험을 진행해 보았다. 그 결과가 다음의 <표 4>에 제시되어 있다. 이 결과를 통해 SVM이 비교모형인 로지스틱 회귀모형이나 인공신경망과 비교해 월등히 높은 예측 정확도를 나타내고 있음을 확인할 수 있다.

예측모형에 임계치까지 적용한 본 연구의 최종 실험결과는 다음의 <표 5>에 제시되어 있다. <표 5>에서 볼 수 있듯이, SVM과 유전자 알고리즘을 결합한 본 연구의 제안 시스템은 다른 비교 모형들에 비해 훨씬 더 우수한 성과를 보이고 있음을 알 수 있다. 본 연구의 제안 시스템을 활용해 2004년 5월부터 2008년 12월까지의 패턴을 학습하고, 이를 2009년 한 해 동안의 실제 매매에 적용할 경우, 약 48.21%의 수익을 거둘 수 있는 것으로 나타났다. SVM에 기반한 모형의 경우, 유전자 알고리즘을 이용한 이중 임계치를 적용하지 않고 단일 임계치로 등락을 판단해 매매를 할 경우에도 약 47%에 가까운 높은 수익을 거둘 수는 있으나, 이 경우에는 상당히 자주(46회) 매매를 해야 한다는 한계가 있다. 그에 비해 제안 시스템은 절반 정도의 횟수(27회)로만 매매를 하고도, 더 높은 수익을 거둘 수 있다는 점에서 전반적으로 가장 우수한 성능의 모형이라는 점을 확인할 수 있다.

6. 연구의 의의 및 한계점

본 연구에서는 최근 금융분야에서 크게 주목받

<표 5> 최종 실험결과

모형 구분	최적 임계치		전체 수익률		연평균 수익률		거래횟수	
	상단임계치	하단임계치	학습기간	검증기간	학습기간	검증기간	학습기간	검증기간
LOGIT + 단일임계치	0.500000	0.500000	123.67%	12.91%	19.01%	12.91%	186	49
ANN + 단일임계치	0.500000	0.500000	30.18%	34.83%	5.87%	34.83%	98	37
SVM + 단일임계치	0.500000	0.500000	547.00%	46.96%	49.74%	46.96%	151	46
LOGIT + 이중임계치	0.497770	0.491353	135.81%	11.48%	20.38%	11.48%	165	50
ANN + 이중임계치	0.468103	0.413088	86.02%	42.39%	14.36%	42.39%	15	2
SVM + 이중임계치	**0.498915**	**0.490540**	**470.26%**	**48.21%**	**45.71%**	**48.21%**	**85**	**27**

고 있는 트레이딩 시스템과 관련하여, SVM과 유전자 알고리즘을 결합한 새로운 형태의 지능형 트레이딩 시스템을 제안하였다. 본 연구의 제안 시스템은 다른 분류 모형들에 비해 효율적이면서도 예측 정확도가 높은 것으로 보고되고 있는 SVM을 기반으로, 유전자 알고리즘에 의해 최적화 된 이중 임계치를 통해 매수, 매도 혹은 보류와 같은 매매 의사결정을 스스로 내릴 수 있도록 설계되었다. 제안 시스템의 유용성을 약 6년 여간 축적된 KOSPI 200 지수에 적용해 본 결과, 제안 시스템이 기존 연구들에서 다루어졌던 다른 비교 모형들과 비교해, 더 높은 투자 수익률을 제공함을 확인할 수 있었다.

본 연구의 의의 혹은 시사점은 크게 다음의 3가지 정도로 요약될 수 있다. 우선 첫째, 본 연구는 복잡다단한 주식시장의 등락을 예측하는데 있어, 다른 이분류 모형에 비해 SVM 모형이 훨씬 우수함을 실증적으로 증명하고 있다. 오랜 시간 동안 국내외를 막론하고 주식시장의 흐름을 예측하기 위한 도구로는 주로 인공신경망 모형이 활용되어 왔는데, 본 연구는 앞으로 SVM에 대해 관련 연구자들이 더 많이 관심을 가질 필요가 있다는 점을 시사한다.

둘째로 본 연구는 '이중 임계치' 방법이 매매 수익률 극대화를 목표로 하는 트레이딩 시스템에 상당히 유용하게 활용될 수 있음을 시사하고 있다. 우리는 본 연구에서 유전자 알고리즘을 통해 이중 임계치를 최적화 할 경우, 로지스틱 회귀모형을 제외한 나머지 두 모형에서 모두 단일 임계치를 매매에 적용할 때에 비해 더 적은 거래횟수로도 더 높은 수익을 창출하는 것이 가능함을 확인할 수 있다. 이런 점으로 미루어 볼 때, 앞으로 이중 임계치 방법은 향후 다른 매매 최적화와 관련된 연구에서도 유용하게 적용될 수 있을 것으로 기대된다.

마지막으로 본 연구는 트레이딩 시스템에 있어 '비가격 변수'들이 시장의 흐름을 예측하는데 상당히 유용한 정보가 될 수 있다는 점을 시사한다. 비슷한 조건에서 이루어진 기존 연구와의 비교를 통해, KOSPI200 지수의 등락을 예측하는데 가격 기반의 기술적 지표만 사용하는 것 보다는 비가격 지표들을 함께 활용해 예측을 하는 것이 보다 더 정확한 예측을 가능케 함을 확인할 수 있었다. 이는 상당히 많은 주식시장 예측과 관련한 연구들이 기술적 지표만을 입력변수로 활용하고 있는 점을 고려할 때, 매우 의미 있는 발견이라고 할 수 있다.

하지만 본 연구는 여러 한계점도 함께 내포하고 있다. 우선 실제 매매에서는 거래비용이 발생함에도 불구하고, 본 연구에서는 거래비용이 없다고 가정하고 실험을 수행하고 있다는 점을 들 수 있다. 물론 거래비용에 대한 추정의 어려움은 있겠지만, 향후 연구에서는 직/간접적인 거래비용까지 반영해서 제안 시스템의 성능을 보다 정밀하게 측정하는 노력이 수반되어야 할 것이다.

둘째로 검증용 데이터가 좀 더 확보될 필요가 있다는 점을 지적할 수 있다. 현재 연구에서는 2009년 한 해 동안의 KOSPI200 지수를 검증용 데이터로 활용하고 있는데, 1년이라는 기간은 시스템의 성능을 검증하기에 충분하지 않으며, 2009년이라는 시점 자체가 2008년 말 갑작스레 터진 글로벌 금융위기가 극복되면서 상당히 빠르게 주식시장이 오른 다소 특별한 시기였다는 점을 감안할 때 검증 기간을 더 확대할 필요가 있다고 할 수 있다. 향후 연구에서 이런 부분이 더 보완되어야 할 것이다.

세 번째 본 연구의 또 다른 한계점으로 학습용 데이터에 적용된 '무작위 추출'을 통한 지수 상승 및 하락 사례의 균형화가 시계열 정보를 왜곡할 가능성이 있다는 점을 지적할 수 있다. 일반적으로

이분류 모형에서는 학습 시 예측값이 한쪽 방향으로 쏠리는 현상을 막기 위해, 종속변수를 구성하는 두 집단의 표본수를 동일하게 맞추는 균형화 작업을 사전에 수행한다. 하지만, 주가 정보와 같은 시계열 정보에 이러한 인위적인 표본 추출 과정을 적용할 경우, 시계열 패턴이 명확하게 반영되지 않을 위험이 존재할 수 있다. 때문에 학습의 왜곡도 방지하면서, 시계열 패턴도 그대로 반영할 수 있는 현명한 학습 알고리즘에 대한 깊이 있는 연구가 향후 수반되어야 할 것으로 보인다.

끝으로 현재 연구에서는 SVM의 단일 모형만 다루고 있는데, 보다 나은 성과의 개선을 위해 향후 연구에서는 모형 간 결합을 시도해 볼 수 있다. 옥중경과 김경재(2009)의 최근 연구에 따르면, 분류모형의 경우 각 기법마다 각각의 장단점이 있어 서로 상호 보완적인 관계에 있기 때문에, 단일 기법을 사용하는 것보다 다수의 기법을 동시에 결합해 사용하는 것이 성과를 개선하는데 도움이 될 수 있다. 이러한 아이디어를 반영한다면, 본 연구에서 제안된 트레이딩 시스템 보다 매매 수익률을 더 높일 수 있는 트레이딩 시스템의 개발도 가능할 것으로 예상된다.

참고문헌

안현철, 김경재, 한인구, "효과적인 고객관계관리를 위한 사례기반추론 동시 최적화 모형", *한국지능정보시스템학회논문지,* 11권 2호(2005a), 175~195.

안현철, 김경재, 한인구, "Support Vector Machine을 이용한 고객구매예측모형", *한국지능정보시스템학회논문지,* 11권 3호(2005b), 69~82.

안현철, 이형용, "투자의사결정 지원을 위한 유전자 알고리즘 기반의 다중인공지능기법 결합모형", *e-비즈니스 연구,* 10권 1호(2009), 267~288.

안현철, 이형용, 김경재, "효과적인 투자의사결정 지원을 위한 인공지능 결합모형 : KOSPI200 지수에의 응용", 2009 한국BI데이터마이닝학회 춘계학술대회, 407~415, 2009.

옥중경, 김경재, "유전자 알고리즘 기반의 기업부실예측 통합모형", *지능정보연구,* 15권 4호(2009), 97~118.

이재식, 송영균, 허성회, "인공신경망 앙상블을 이용한 옵션 투자예측 시스템", 한국지능정보시스템학회 2000년 학술대회 2권, 489~497, 2000.

이형용, "한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형", *Entrue Journal of Information Technology,* 7권 2호(2008), 33~43.

홍승현, 신경식, "유전자알고리즘을 활용한 인공신경망모형 최적입력변수의 선정 : 부도예측모형을 중심", *한국지능정보시스템학회논문지,* 9권 1호(2003), 227~247.

홍태호, 신택수, "Using Estimated Probability from Support Vector Machines for Credit Rating in IT Industry", 한국지능정보시스템학회-웹코리아포럼 2005 공동추계정기학술대회, 509~515, 2005.

Ahn, H., K. j. Kim and I. Han, "Determining the optimal number of cases to combine in an effective case based reasoning system using genetic algorithms", *Proceedings of International Conference of Korea Intelligent Information Systems Society 2003 (ICKI-ISS 2003),* 178~184, 2003.

Ahn, H., C. Song, J. J. Ahn, H. Y. Lee, T. Y. Kim, and K. J. Oh, "Using Hybrid Data Mining Techniques for Facilitating Cross-selling of a Mobile Telecom Market to Develop Customer Classification Model", *The 43th Hawaii International Conference on System Sciences(HICSS-43),* Hawaii, USA, 2010.

Alexander, S. S., "Price movements in speculative markets : trends or random walks", *Industrial Management Review,* Vol.2, No.2 (1961), 7~26.

Atiya, A., N. Talaat and S. Shaheen, "An efficient stock market forecasting model using neural networks", *Proceedings of the IEEE International Conference on Neural Networks,* 2112~2115, 1997.

Atsalakis G. S. and K. P. Valavanis, "Surveying stock market forecasting techniques Part II : Soft computing methods", *Expert Systems with Applications,* Vol.36, No.3(2009a), 5932~5941.

Atsalakis G. S. and K. P. Valavanis, "Forecasting stock market short term trends using a neuro fuzzy based methodology", *Expert Systems with Applications,* Vol.36, No.7 (2009b), 10696~10707.

Bao, D. and Z.Yang, "Intelligent stock trading system by turning point confirming and probabilistic reasoning", *Expert Systems with Applications,* Vol.34, No.1(2008), 620 ~ 627.

Caporale, G. M. and U. N. Pittis, "Cointegration and predictability of asset prices", *Journal of International Money and Finance,* Vol. 17, No.3(1998), 441~453.

Chang, C.-C. and C.-J. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chavarnakul T. and D. Enke, "A hybrid stock trading system for intelligent technical analysis based equivolume charting", *Neurocomputing,* Vol.72, No.16-18(2009), 3517~3528.

Casas C. A., "Tactical asset allocation : An artificial neural network based model", *Proceedings of the International Joint Conference on Neural Networks,* 1811~1816, 2001.

Elton, E. J. and M. J. Gruber, *Modern Portfolio Theory and Investment Analysis,* Wiley, 1984.

Fama, E. F., "The Behavior of Stock Market Prices", *Journal of Business,* Vol. XXXVIII (1965), 34~105.

Granger C., "Some Properties of Time Series Data and Their Use in Econometric Model Specification", *Journal of Econometrics,* Vol. 16, No.1(1981), 121~130.

Kim, K.-j., "Financial time series forecasting using support vector machines", *Neurocomputing,* Vol.55, No.1-2(2003), 307~319.

Kim, K. j. and I. Han, "Application of a hybrid genetic algorithm and neural network approach in activity based costing", *Expert Systems with Applications,* Vol.24, No.1 (2003), 73~77.

Kim, K.-j. and W. B. Lee, "Stock market prediction using artificial neural networks with optimal feature transformation", *Neural Computing and Applications,* Vol.13, No.3 (2004), 255~260.

McMillan, D. G., "Non-linear forecasting of stock returns: Does volume help?", *International Journal of Forecasting,* Vol.23, No.1(2007), 115~126.

McNelis, P. D., *Neural Networks in Finance: Gaining the Predictive Edge in the Market,* Elsevier Academic Press, 2005.

Núñez-Letamendia, L., "Fitting the control parameters of a genetic algorithm : An application to technical trading systems design", *European Journal of Operational Research,* Vol.179, No.3(2007), 847~868.

Olson, D. and C. Mossman, "Neural network forecasts of Canadian stock returns using

accounting ratios", International Journal of Forecasting, Vol.19, No.3(2003), 453~465.

Platt, J., "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods", In A. J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers,* Cambridge, MA, 2000. MIT Press.

Schulmeister, S., "Profitability of technical stock trading : Has it moved from daily to intraday data?", *Review of Financial Economics,* Vol. 18, No.4(2009), 190~201.

Schwager, J. D., *The New Market Wizards : Conversations with America's Top Traders,* Harper Business, 1992.

Sollich, P., "Bayesian Methods for Support Vector Machines : Evidence and Predictive Class Probabilities", *Machine Learning,* Vol. 46, No.1-3(2002), 21~52.

Sullivan R., Timmermann A., and H. White, "Data snooping, technical trading rule performance, and the bootstrap", *The Journal of Finance,* Vol.LIV(1999), 1647~1691.

Tay, F. E. J. and L. J. Cao, "Modified support vector machines in financial time series forecasting", *Neurocomputing,* Vol.48, No. 1-4(2002), 847~861.

Vapnik, V., *Statistical Learning Theory,* Wiley, 1998.

Yudong, Z. and W. Lenan, "Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network", *Expert Systems with Applications,* Vol.36, No.5(2009), 8849~8854.

Abstract

Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms

Sun-Woong Kim* · Hyunchul Ahn**

As the use of trading systems increases recently, many researchers are interested in developing intelligent trading systems using artificial intelligence techniques. However, most prior studies on trading systems have common limitations. First, they just adopted several technical indicators based on stock indices as independent variables although there are a variety of variables that can be used as independent variables for predicting the market. In addition, most of them focus on developing a model that predicts the direction of the stock market indices rather than one that can generate trading signals for maximizing returns.

Thus, in this study, we propose a novel intelligent trading system that mitigates these limitations. It is designed to use both the technical indicators and the other non-price variables on the market. Also, it adopts 'two-threshold mechanism' so that it can transform the outcome of the stock market prediction model based on support vector machines to the trading decision signals like buy, sell or hold. To validate the usefulness of the proposed system, we applied it to the real world data–the KOSPI200 index from May 2004 to December 2009. As a result, we found that the proposed system outperformed other comparative models from the perspective of 'rate of return'.

Key Words : Intelligent Trading Systems, Support Vector Machines, Genetic Algorithms, Two-threshold Mechanism, KOSPI200

* The Graduate School of Business IT, Kookmin University
** School of Management Information Systems, Kookmin University

저 자 소 개

김선웅
현재 국민대학교 BIT전문대학원 초빙교수로 재직 중이다. 서울대학교 경영학과에서 경영학사를 취득하고, KAIST 경영과학과에서 증권투자론을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 투자공학, 트레이딩시스템, 헤지펀드와 자산운용이다.

안현철
현재 국민대학교 경영대학 경영정보학부 전임강사로 재직 중이다. KAIST에서 산업경영학사를 취득하고, KAIST 테크노경영대학원에서 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 주요 관심분야는 금융 및 고객관계관리 분야의 인공지능 응용, 정보시스템 수용과 관련한 행동 모형 등이다.