Journal of Information Technology and Architecture
Vol. 13. No. 2, June 2016, Pages 331-340

기술적 지표와 기계학습을 이용한 KOSPI 주가지수 예측

박재연[1] · 유재필[2] · 신현준[2*]

[1]상명대 기술융복합학과
[2]상명대 경영공학과
jaeyeon.park@kispricing.com; jaepilryu@kispricing.com; hjshin@smu.ac.kr

(2016년 6월 6일 접수; 2016년 6월 28일 수정; 2016년 6월 30일 채택)

요약: 최근 다양한 분야에서 기계 학습에 대한 관심이 높아지고 있는 가운데 금융 분야에서도 로보어드바이저(robo-advisor) 등 기계 학습을 현업에 접목시키려는 시도가 많아지고 있다. 특히 이러한 기계적이고 정량적인 의사결정은 수수료와 같은 비용 절감은 물론 효과적인 의사결정을 가능하게 한다는 점에서 큰 장점이 있다. 본 연구에서는 SVM(support vector machines)와 라쏘 회귀분석(lasso regression) 그리고 인공신경망(Artificial Neural Network; ANN) 등의 기계학습 기법들을 이용하여 KOSPI 지수를 예측하는 모형을 개발하고 제시한 기계학습 기법들의 예측력을 비교, 분석한다. 실험은 학습 데이터(in sample)와 실험 데이터(out of sample)로 나눠서 진행하였는데, 전자는 2000년 1월 1일부터 2009년 12월 31일까지이며 후자는 2010년 1월 1일부터 2015년 9월 15일까지이다. 실험을 진행한 결과 학습 데이터에서는 SVM이 ANN에 비해서 더 높은 예측력을 보인 반면에 실험 데이터에서는 ANN의 예측력이 더 우수했다. 한편 라쏘 회귀분석의 경우에는 학습 데이터와 실험 데이터 모두에서 예측력이 우수하지 않았다.

키워드: SVM, 인공신경망, 라쏘 회귀분석, KOSPI, 기계학습, 주가지수 예측

Predicting KOSPI Stock Index using Machine Learning Algorithms with Technical Indicators

Jae Yeon Park[1], Ryu Jaepil[2], and Shin Hyun Joon[2*]

[1]Department of Convergence, Sangmyung University
[2]Department of Management Engineering, Sangmyung University

(Received June 6, 2016; Revised June 28, 2016; Accepted June 30, 2016)

Abstract: Recently, there have been many attempts to employ the machine learning methodologies such as Robo-Advisor in the financial sector with growing interest in machine learning in various sectors. Especially, these mechanical and quantitative decision making have some big advantages not only reduces the costs such as fees but also enable us to make effective decision. In this research, we developed a machine learning model to forecast the KOSPI index and analyze the accuracy of the prediction. We use three machine learning model : SVM(support vector machines), Lasso regression, and ANN(Artificial Neural Network). We divided our data into two parts : 'in sample' data and 'out of sample' data. The 'in sample' data is from January 1st, 2000 to December 31st, 2010. And the 'out of sample' data is from January 1st, 2011 to September 15th, 2015. The result of the experiment, the 'in sample' data, the SVM showed higher accuracy compare to the ANN. On the other hand, in the 'out of sample' data, ANN was superior than SVM. For the Lasso

교신저자
본 연구는 2016년도 상명대학교 교내연구비를 지원받아 수행하였음.

Regression analysis, it showed worst predictive accuracy in both 'in sample' and 'out of sample' data.

Keywords: SVM, ANN, Lasso Regression, KOSPI, Machine Learning

1. 서 론

한국은 IMF(International Monetary Fund) 금융 위기를 계기로 금융 시장을 개방하면서 1997년 당시 약 1,640억 달러 규모였던 외국인 자본이 2016년에는 약 6,000억 달러까지 급증하였다. 이는 동일 기간 동안 평균 물가상승률을 고려해도 약 260%나 급등한 수치이다. 이처럼 주식시장에 유입된 외국인 자본으로 인해서 KOSPI 지수의 변동성은 커졌으며 외국인 자본의 의존도는 크게 높아졌다(이석준, 2011).

또한 금융 시장에 선물(futures)과 옵션(option)을 비롯한 다양한 파생상품(derivatives)이 크게 증가하면서 개인 및 기관 투자자 등은 파생상품 거래에 대한 헤지(hedge) 수단을 필요로 하고 있으며, 다양한 헤지 전략을 통한 매매차익을 기대하고 있다. 또한 레버리지(leverage) 기능이 큰 금융상품에서의 위험(risk)은 파생상품의 투자자들로 하여금 크나큰 부담을 줄 수밖에 없는 현실이다.

이러한 주식 시장의 환경 속에서 경제적 소득과 직접적인 영향을 미치는 주가에 관한 예측은 투자자들로 하여금 많은 관심을 받는다. 때문에 금융 분야와 함께 다양한 연구 분야에서 주가 예측에 관한 방법론들을 연구하였고 지금까지도 주가 예측에 대한 연구는 관심도가 매우 높다.

일반적으로 주가 예측은 크게 기본적 분석(fundamental analysis)과 기술적 분석(technical analysis) 그리고 과학 기술적 방법(technological methods)이 있다(김선웅, 안현철, 2010). 기본적 분석은 기업의 재무 정보와 같은 기업의 과거 성과를 평가하여 주가를 예측하는 방법으로 펀드 매니저나 주식 투자자 등이 주로 사용하는 방법이다. 이러한 방법은 기업의 실적과 사업의 성과 등을 자세하게 알 수 없기 때문에 미래의 주가를 예측하는 것은 한계가 있다는 단점이 있다. 기술적 분석은 과거 주식의 가격의 동향을 분석하고 패턴을 이해하여 미래 주가의 방향을 예측하는 방법으로 EMA(exponential moving average) 등과 같은 통계적 기법을 활용한다. 과학 기술적 방법은 컴퓨터 기술의 발달로 인해서 수많은 연산을 짧은 시간에 처리할 수 있게 되면서 인공신경망(artificial neural network)과 유전자알고리즘(genetic algorithm)과 같은 기계학습(machine learning) 방법론을 이용하여 미래의 주가를 예측하는 것을 말한다. 그러나 주가 또는 지수는 다양한 직·간접적인 변수들에 의해서 불규칙적으로 변동하기 때문에 정확한 예측을 하는 것은 한계가 있다. 더불어 주식시장은 비정상성(non- stationary)과 잡음(noise) 그리고 비선형성(non-linearity) 등으로 인해서 미래의 주가 움직임을 예측하는 것은 매우 어렵고 복잡하다.

주가 및 지수를 예측하는 방법은 위에서 기술했듯이 기본적 분석과 기술적 분석 등을 거쳐서 최근에는 기계학습 방법으로 진화하고 있다. 빅데이터(big data) 또한 우리나라가 세계에서 두 번째로 관심도가 높은 분야이면서 2020년에는 약 40제타바이트의 데이터 시대가 올 것이라고 예상한다(김정래, 정찬기, 2013). 이러한 현황에 맞춰 최근 알파고(AlphaGo)로 인해서 금융시장에서는 로보어드바이저(robo-advisor)에 대한 기대심이 높아지고 있다. 자산관리 시스템인 로보어드바이저는 높은 고객 접근성과 상대적으로 낮은 수수료 등의 장점이 있다. 현재 로보어드바이저에 의해서 운용되는 자금은 전 세계적으로 약 24조원이고 2020년에는 약 450조원이 될 전망이다. 특히 우리나라의 경우에는 금융위원회가 이르면 2016년 10월부터 검증된 로보어드바이저에 한해서 온라인 자문과 일임 업무를 허용할 계획이다.

본 연구에서는 과학 기술적 방법을 통한 주가 예측의 관심이 높아지면서 본 연구에서는 SVM(support vector machines)과 라쏘 회귀분석(lasso regression) 그리고 인공신경망 학습 기법을 통해서 한국 주가의 대표 지수인 KOSPI 지수의 예측력을 분석하고자 한다. 실험기간은 2000년 1월 1일부터 2015년 9월 15일까지 일별(daily) 지수 종가 값을 기초로 한다. 이 데이터는 학습 데이터(training data)와 실험 데이터(test data)로 나눠 실험을 진행하는데, 전자는 2000년 1월 1일부터 2009년 12월 31일까지이며 후자는 2010년 1월 1일부터 2015년 9월 15일까지이다.

Figure 1은 기본적인 기계학습의 이론을 설명해주는

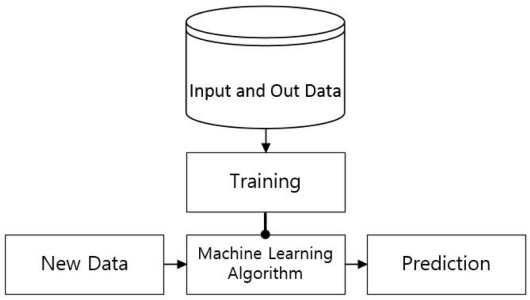

Figure 1. 기계학습의 기본적인 모형

그림이다(Jae Yeon Park, 2016). 일반적으로 학습, 즉 training을 위해서 input data와 output data를 수집한다. 여기서 전자는 독립변수(independent variable) 그리고 후자는 종속변수(dependent variable)가 된다. 과거 시계열의 독립 및 종속변수를 바탕으로 기계학습을 하고 학습된 모형에 새로운 input data를 적용하면 예측된 output data를 추출한다.

본 논문의 2장에서는 주가 예측에 관한 연구와 기계학습에 관한 선행 연구를 기술하고, 3장에서는 실험을 위한 데이터 설명과 함께 본 연구에서 지수 예측을 위해서 사용하는 학습 방법인 SVM과 라쏘 회귀분석 그리고 인공신경망 학습 기법에 대해서 설명한다. 4장에서는 본 연구의 실험 계획 및 실험 결과 그리고 본 연구에서 사용한 3가지 학습 방법론 간의 예측성과 분석에 대해서 기술한다. 마지막으로 5장에서는 결론으로 마무리 한다.

2. 관련문헌 연구

주가 및 지수 예측에 대한 연구는 2000년 이전에는 기본·기술적 분석이 일반적 이였으나 1987년 Lapedes와 Farber이 기계학습 모델을 이용하여 과거 10일의 주가 자료를 바탕으로 미래 10주의 주가를 예측하는 연구가 발표되었다. 2000년 이후에는 기계학습을 통해서 주가를 예측하는 연구가 진행되고 있는데, 최근 컴퓨터 기술 발달하고 기업과 기관 등 여러 분야에서 기계학습이 중요시되면서 금융 분야에서도 기계학습을 통한 주가 예측 연구가 활발하게 진행되고 있다.

Jeantheau(2004)는 약 5년의 시계열 자료를 갖고 ARCH 모델을 사용하여 주가를 예측한 결과 기업에서 공시한 과거의 재무 자료를 바탕으로 주가의 움직임을 예측하는 것보다 ARCH 모델이 우수한 예측력을 보인다는 것을 입증하였다. 이와 유사한 연구 방법론을 바탕으로 Amilon(2003)는 중국 주식시장에서 존재하는 유통 거래량 정보와 GARCH 모델을 사용하여 주가를 예측하는 방법을 제시하였다. 반면 Tay and Cao (2001)는 금융시계열 데이터를 기계학습 기법 중에 하나인 SVM에 도입하는 방법을 제안하였는데 특정한 시계열들 간에 선행적 상관관계가 있음을 입증하였다. 최근 지수나 주가 예측에 있어서 전통적인 예측 기법과는 다르게 빅데이터를 예측 기법에 적용하는 연구가 증가하고 있는데, 유재필과 신현준(2016)은 포탈 트렌드(trend)를 이용하여 약 5년 동안의 시험 자료를 갖고 ETF(exchange traded fund) 상품 가격의 방향성을 예측하였다. 그 결과 트렌드를 활용하여 정량적으로 매매 포지션을 선정하는 것이 벤치마크 전략의 수익률에 비해서 높은 성과를 보인 것으로 나타났다. 또한 유재필과 신현준(2014)은 KOSPI200 지수와 선행적 역상관관계를 갖는 시계열을 바탕으로 관리도 모형을 적용한 KOSPI200 지수 예측 모형을 제안하였는데, 이는 관리도 모형을 적용한 정량적 매매 전략이 랜덤 매매 전략(random trading strategy) 등에 비해서 월등하게 우수한 성과를 보였다.

일반적으로 기계학습을 통한 주가 예측에 관한 연구는 인공신경망 학습 모델이 가장 대표적이다. Kanas (2003)은 인공신경망 학습 모델을 통해서 S&P500 지수를 예측하였는데, 본 연구에서는 지수의 변동성이 낮을 때 보다 더 높은 예측력을 보였다. Yang et al. (2001)은 인공신경망 학습 모델을 이용해 은행의 대출 리스크(Commercial Bank Loan Risk) 관리를 위한 조기경보시스템을 제안하였으며, Bekiros and Georgoutsos(2008)는 미국의 상장 기업들과 관련된 보도 자료 구분하기 위해서 불명확한 뉴스가 NASDAQ 지수에 어떠한 영향을 주는지 인공신경망 학습 모델을 이용하여 분석하였다. Yoon and Swales(1991)는 비선형적인 기법인 인공신경망 학습 모델과 선형적 기법인 다변량 판별분석(Multivariate Discriminant Analysis)을 주가 예측능력의 향상여부를 파악하기 위하여 비교하였다. 그 결과 다변량 판별분석은 학습 기간에는 약 74%이며 예측 기간에서는 약 65%의 예측력을 보였고, 인공신경망 학습 모델은 학습 기간에서는 약 91%이며 예측 기간에는 약 78%의 예측력을 나타냈다. Wong(2000)은 Fuzzy System과 인공신경망 학습 모델을 결합시킨 주식 예측 시스템을 개발하였는데 입력 자료(input data)를 전문가의 지식으로 전환시킬

수 있는 규칙을 활용하여 퍼지 시스템(fuzzy system)으로 가공한 후에 인공신경망 학습 모델에 규칙을 입력하는 방식을 적용하여 예측하였다. George and Matthew(1995)는 인공신경망 학습 모델을 이용하여 미래의 주가 방향을 예측하였는데 이는 예측의 정확도가 매우 높다는 것을 입증하였다. 또한 효율적 시장에 대한 가설을 검증하기 위해서 귀무가설(null hypothesis) 검정을 실험한 결과 인공신경망 학습 모델을 이용한 경우의 예측력이 귀무가설의 5%를 상회하는 결과를 보였다. Hamid and Zahid(2002)의 연구에서는 옵션모형, 선형모형, 인공신경망 학습 모형을 이용하여 S&P 500 지수 선물의 변동성을 예측한 결과, 인공신경망 학습 모델의 예측력이 우수하다는 것을 입증하였다. 허준영(2015)은 재무 정보를 기반으로 SVM을 이용하여 주가를 예측함으로써 회사의 내재 가치를 나타내는 재무정보가 주가의 예측에 얼마나 효과적인지를 검증하였다. 그 결과 전문가 점수와 기계학습 방법인 인공신경망, 결정트리, 적응형 부스팅보다 SVM을 이용하여 주가를 예측하는 것이 보다 더 우수한 결과를 보였다.

3. 자료와 모형

본 연구에서는 KOSPI 지수를 예측하는 기계학습 모형을 개발하고 성능을 실험하기 위해서 크게 학습 데이터와 실험 데이터로 나눠진다. 본 장에서는 앞서 기술한 학습 및 실험 데이터의 수집에 대해서 설명하고 본 연구에서 기계학습 모형으로 사용하는 SVM, 라쏘 회귀분석, 인공신경망 학습 모델에 대해서 설명하고자 한다.

3.1 자료 선정

본 연구에서는 학습 및 실험 데이터를 구분하며 학습 데이터는 2000년 1월 1일부터 2009년 12월 31일까지이며 실험 데이터는 2010년 1월 1일부터 2015년 9월 15일까지이다. 학습 데이터 기간은 기계학습 모형을 구성하기 위한 기간이고 실험 데이터 기간은 학습된 모형에 입력변수를 투입시킴으로써 추출되는 값과 실제 값과의 비교를 통해서 학습 모형을 검증하기 위함의 기간이다.

데이터는 한국거래소에서 공시하는 KOSPI 지수를 통계 소프트웨어인 R을 통해서 수집한다. 종속변수는 영업일 기준으로 10일 후의 KOSPI 지수의 종가로 정의한다. 예컨대 오늘 입력변수를 통해서 상승 신호(signal)가 나왔다면 10일 후 KOSPI 지수의 종가는 상승 신호가 나온 날의 종가 대비 상승하고, 반대로 하락 신호가 나오면 10일 후 KOSPI 지수의 종가는 하락 신호가 나온 날의 종가 대비 하락하는 것을 말한다.

본 연구에서 기계학습을 위한 즉, 예측을 하는데 이용하는 변수는 총 6개의 종류인데 이는 모두 KOSPI 지수를 가공하여 산출된 변수이다. 산출된 변수는 다음과 같다.

- ATR(average true range)
- DMI(directional movement indicators)
- 60 Volatility
- MACD(moving average convergence divergence)
- SAR(stop and reversal)
- Mean

ATR은 시간에 따른 가격 변화의 정도인 변동성을 나타내는 지표이다. 지수가 급락하여 낮은 수준에 머물 경우 ATR은 높게 나타나고 반대로 주가가 급속히 변동하기 직전까지 일정한 수준으로 지속되어 온 경우에는 ATR은 낮게 나타난다. DMI는 시장의 방향성을 나타내는 지표로써 전일 대비 해당일의 고가와 저가 그리고 종가의 최고 값을 이용하여 현재 추세와 매수 및 매도 시점을 판단해주는 지표이다. Volatility은 KOSPI 지수의 과거 60일 변동성을 의미한다. MACD는 26일간의 지수 평균과 12일간의 지수 평균 간의 차이를 갖고 산출하며, 이 두 지수 평균의 차이를 다시 9일간의 지수 평균으로 산출한다. SAR는 시장 가격의 추세가 더 이상 지속하지 못하면 추세의 전환이 발생한다는 점을 착안하여 산출하는 지표로써 추세 전환 시 기존 포지션을 청산(stop)하고 반대의 신규 포지션을 확보(reversal)한다. 시장 가격 이 상승 추세 시 SAR 값은 시장 가격 아래쪽에서 상승을 지속하고, 시장 가격 하락 추세 시 SAR 값은 시장 가격의 위쪽에서 하락을 지속한다. SAR과 시장 가격이 만나는 시점이 추세가 반전되었음을 확인하는 지점이다. 이처럼 위에서 설명한 입력변수들은 모두 투자 시에 매매 시점을 참고하기 위한 지표로써 수집된 KOSPI 지수를 기반으로 공개 소프트웨어인 R을 이용해서 6 가지 변수를 산출한다. 본 연구에서는 6개의 변수로 이뤄진 모델을 이용하는데 이러한 6개의 변수는 예측에 있어서 중요한 변수이다. 예컨대 위 변수들이 얼마만큼 중요한지 실험을 한 결과,

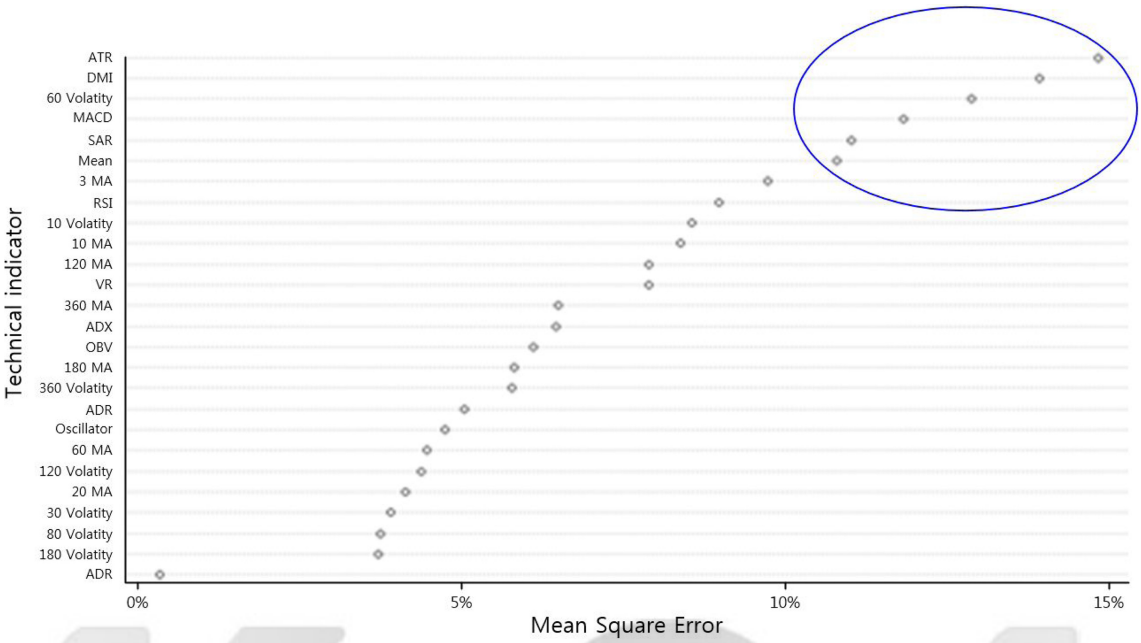

Figure 2. 특정 기술적 지표가 제외됐을 경우의 MSE (단위 %)

위 변수들을 제외했을 때, 학습 모델의 MSE(Mean Square Error)의 증가율은 크게 증가하는 것을 알 수 있었다. Figure 2는 학습 모델에서 각각의 변수를 제거했을 때 증가하는 MSE를 보여주는 그림이다. 세로축은 총 26개의 입력변수들로 나타나 있고 가로축은 학습 모델에서 해당 변수가 제외되었을 때 산출된 MSE이다. 그림에서 원 안에 있는 6개의 변수가 본 연구에서 선정한 6개의 변수이다. 특히 6개의 변수 중에서 MSE가 약 15%정도로 Mean 변수가 가장 높았고 SAR이 약 11%로 가장 낮았다. MA(moving average)는 이동평균 값을 의미하며 앞에 숫자는 과거 일 수를 의미한다. 예컨대 3 MA는 3일 이동평균 값을 의미하고, 180 MA는 과거 180일 이동평균 값을 의미한다. 이와 유사하게 변동성을 의미하는 Volatility 또한 앞에 명시된 숫자가 과거 일 수를 의미한다. 본 연구에서 사용하는 6개의 입력 변수들은 서로 상관관계의 유무가 중요한 사항은 아니다. 따라서 6개의 입력 변수들의 인과관계 분석은 본 연구에서 제외한다. 그 외 ADX(average directional movement indicator), OBV(on balance volume) 등 다른 technical indicator는 본 연구에서 의미 있는 자료가 아니기 때문에 추가적인 설명은 생략한다.

3.2 기계학습 모형

본 연구에서는 기계학습을 통한 KOSPI 지수 예측 능력을 실험하기 위해서 대표적인 3 종류의 기계학습 모형을 이용한다. 다음 절에서는 SVM과 라쏘 회귀분석 그리고 인공신경망 학습 기법에 대해서 설명하고자 한다.

3.2.1 SVM(support vector machines)

SVM은 1995년 Vapnik에 의해서 고안된 알고리즘이며 입력공간과 관련된 비선형문제를 고차원적인 특정 공간으로 변화시켜 선형문제로 적용한다. 이 공간에서 훈련 샘플(training sample)이 분리경계면(hyperplane)까지 거리가 최대가 되도록 두 개의 클래스(class)로 나누는 최대 마진 분리경계면(maximum-margin hyperplane)을 찾는다. 이진 클래스 분류 문제에서 훈련 자료 N개의 패턴 $(x_i, d_i), i = 1, \ldots, N$ 이 주어지면 x는 두 클래스 중에서 하나에 포함되고, $d \in \{-1, 1\}$ 는 해당 클래스를 표현하는 레이블이 된다. SVM은 각 각의 클래스를 최대치로 구분하는 최적의 분리 경계면을 산출하기 위해서 분리 경계면과 가장 가까운 구간(support vector)까지의 거리를 최대화 한다. 최적의 선형 분리 경계면을 $f(x) = w^t x + b$ 로 정의하면

support vector와 $f(x)$의 거리 $1/\|w\|$을 최대치로 할 수 있도록 $\|w\|^2$을 최소화 하는 문제로 변화된다. 결국 다음과 같은 블록 최적화(convex optimization) 문제가 된다.

$$min\ \frac{1}{2}w^t w$$
$$s.t.\ d_t(w^t x_i + b) \geq 1\ \text{ for }\ i = 1, \cdots, N$$

선형 분리가 되지 않는 데이터들을 처리하기 위해서는 유화 변수(slack variable) $\zeta_i(\zeta_i \geq 0)$을 오분류 척도로 정의하면 결정 함수는 다음과 같이 수정된다. 여기서 C는 penalty parameter이다.

$$min\ \frac{1}{2}w^t w + C\sum_{i=1}^{N}\zeta_i$$
$$s.t.\ d_t(w^t x_i + b) \geq 1 - \zeta_i\ \text{ for }\ i = 1, \cdots, N$$

라그랑지 승수(lagrange multiplier) $a = (a_1, \ldots, a_N)$을 도입하고, 비선형 패턴을 분리하기 위해서 입력 공간을 선형 패턴의 특징적 공간으로 변환해야 한다. 여기서 커널(kernel) 함수 $K(x_i, x_j) = \varnothing(x_i)\varnothing(x_j)$를 정의하면 비선형 패턴을 분리하기 위한 최적의 결정 함수는 다음과 같다.

$$maximize\ \sum_{i=0}^{N}a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j d_i d_j K(x_i, x_j)$$
$$s.t.\ \sum_{i=1}^{N}a_i d_i = 0,\ 0 \leq a_i \leq C$$

위 모델에서 라그랑지 승수인 a_i를 산출하면 특징 공간에서 가장 평평한 함수를 구할 수 있다.

3.2.2 라쏘 회귀분석(Lasso Regression)

통상적으로 회귀분석(regression analysis)에서 회귀계수(regression coefficient)의 추정량을 구하기 위해서는 잔차(residual)의 제곱 합을 최소로 하는 최소제곱법(least squared method)을 사용한다. 그러나 설명변수(explanatory variables)의 개수가 증가하면 설명변수들 사이의 강한 상관관계로 인한 다중공선성(multicollinearity)이 존재할 수 있기 때문에 최소제곱 회귀계수 추정량의 분산이 커져 추정회귀식의 예측 정확도가 떨어지는 문제점이 발생할 수 있다. 또한 설명변수의 개수가 증가하면 변수에 대한 해석력이 떨어진다. 즉, 다양한 설명 변수 중에서 어떠한 변수가 중요한 역할을 하는지에 대한 판단이 힘들어진다. 라쏘 회귀분석은 능형 회귀(ridge regression)의 장점인 회귀계수 축소를 통해 예측 정확도(prediction accuracy)를 향상시키며 동시에 영향력이 적은 회귀계수 값을 쉽게 0으로 만드는 변수 선택(variable selection)의 기능이 있어서 해석력(interpretability)을 높여준다. 따라서 라쏘 회귀분석은 능형회귀의 예측 정확도와 변수 선택의 해석력을 모두 갖출 수 있는 분석방법으로 알려져 있다. 라쏘 회귀분석의 추정량은 다음의 식과 같이 구한다.

$$arg\ min_{\beta_1, \cdots, \beta_p}\left\{\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

여기서 회귀계수 $\beta_1, \ldots, \beta_p$의 값은 설명변수 x_{ij}의 척도(scale)에 의존하기에 회귀계수 값의 크기가 그 변수의 영향력을 반영할 수 없기 때문에 x_{ij}에 표준화된 값을 사용한다. 또한 β_0에 대한 추정은 관심 대상에서 삭제되고 $\beta_1, \ldots, \beta_p$에 대한 최소화 문제로 정의된다. 위 식을 다음과 같은 제약조건이 주어진 최소화 문제로 나타낼 수 있다.

$$arg\ min_{\beta_1, \cdots, \beta_p}\left\{\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\}$$
$$subject\ to\ \sum_{j=1}^{p}|\beta_j| \leq t$$

위 식의 제약조건인 $t(\geq 0)$는 회귀계수 값에 대하여 축소 정도를 조절하는 조절모수(tuning parameter)이다. 이 조절모수 t값이 줄어들면 중요치 않는 변수의 회귀계수 값은 축소되면서 순서대로 0으로 만들어지면서 변수 선택이 되는 효과가 발생한다. 조절모수 t값이 충분히 커지면 회귀계수 값에 대한 제약이 없어지므로 최소제곱부분만 남아 로쏘 회귀 추정량이 최소제곱 추정량이 된다.

3.2.3 ANN(Artificial Neural Network)

데이터들의 패턴을 인식하고 분류 문제(classification problem)를 해결하기 위해서 인공지능(artificial intelligence) 분야로부터 도입된 비선형 분석 모델(nonlinear analytical model)인 인공신경망은 계산하는 능력의 단점을 제외하고는 인식이나 의사결정에 있어서 컴퓨터보다 우수한 인간 두뇌(human brain)를 모방하려는 시도 속에서 연구되었다. 특히 인공신경망 모형은 인간의 우뇌 부분의 기능 가운데 학습(learning), 병렬처리(parallel processing), 패턴인식

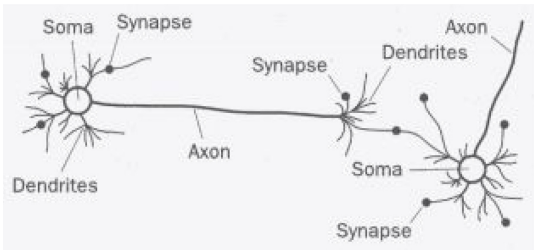

Figure 3. 인간의 신경 세포의 구조

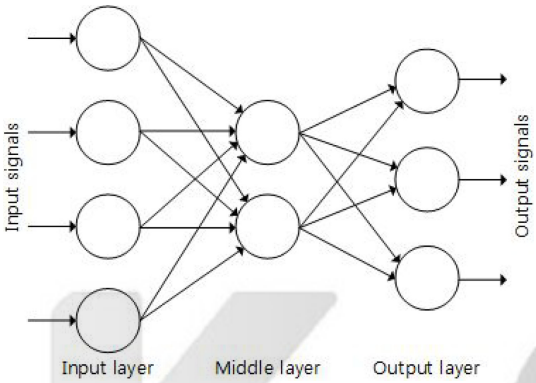

Figure 4. 인공신경망의 기본적인 구조

(pattern recognition), 오차용인(fault tolerance)과 같은 생물학적 프로세스를 공학적으로 컴퓨터에 적용한 기법으로 인간 두뇌의 기본 단위인 뉴런(neuron)의 구조에 착안하여 개발되었다. Figure 3는 뉴런을 설명하는 그림인데 시냅스(synapse), 축색돌기(axon), 가지돌기(dendrites), 신경세포체(cell body) 등으로 이뤄져 있으며 하나의 뉴런은 시냅스에 의해서 서로 다른 뉴런에 연결되어 있다.

인공신경망 학습 모델은 뉴런의 구조와 유사하게 Figure 4과 같이 입력층(input layer), 은닉층(hidden layer), 그리고 출력층(output layer)으로 구성되어 있으며, 입력층에서 보내지는 값을 가중치에 따라 은닉층이 합산하고 이를 활성함수(activation function)에서 변환하여 출력층으로 보내는 구조를 가지고 있다.

인공신경망이 회귀분석과 차별화되는 이유는 입력층과 출력층 사이에 은닉마디(hidden units)로 이뤄진 은닉층이 존재하기 때문이다. 은닉층이 다수로 존재 가능하기 때문에 다층퍼셉트론(multi-layer perceptron)으로 정의하기도 한다. 이처럼 인공신경망 모형은 은닉층의 활성함수로 인해서 인공신경망이 비선형자료의 유형분류와 인식에서 우수한 성과를 보이게 된다.

일반적으로 활성함수로는 주로 S자형 곡선형태를 갖는 함수가 사용되며 가장 보편적인 활성함수로는 로지스틱 함수(logistic function)와 쌍곡탄젠트 함수(hyperbolic tangent function)가 사용된다. 인공신경망 학습 모형은 퍼셉트론, 다층퍼셉트론, 코호넨의 자기조직화 네트워크 등 다양한 알고리즘이 있는데 본 연구에서는 다층퍼셉트론 알고리즘을 사용하였다.

4. 실험결과 및 분석

4.1 실험계획

본 연구에서는 ATR, DMI, Volatility, MACD, SAR, Mean의 변수를 이용하여 기계 학습 모형을 개발하고 약 15년의 KOSPI 지수를 갖고 10일 이후의 지수의 상승 또는 하락에 대한 예측 성능을 실험한다. 기계 학습 방법론은 SVM과 라쏘 회귀분석 그리고 인공신경망 학습 기법을 이용하며, KOSPI 지수 일별 자료의 수집과 기계 학습 모형 개발 그리고 예측력에 대한 분석은 통계 소프트웨어인 R을 이용한다. 본 연구의 실험 계획을 정리하면 Table 3과 같다.

본 연구에서 SVM R package에서 kernlab을 이용하고, 인공신경망 학습은 nnet를 그리고 라쏘 회귀분석은 glmnet를 사용한다.

4.2 실험결과 및 분석

본 절에서는 앞서 설명은 3가지 기계 학습 방법을 통해서 학습 모형을 개발하고 각각의 방법론들이 어느 정도의 예측력을 보이는지 설명하고자 한다. 예측력을 분석하기 위해서 두 가지로 나눠서 실험을 진행한다. 하나는 T를 독립변수로 두고, 회귀(regression) 분석을 통해서 T를 추측하고, 그 이후에 지수가 움직이는 방향성에 대한 신호를 추론하는 방법이다. 이는 본 연구에

Table 3. 실험 계획

실험 요인		설 명
목표 값		KOSPI 지수
학습 모형		SVM, Lasso Regression, ANN
입력 변수		ATR, DMI, Volatility, MACD, SAR, Mean
실험 기간	In sample	2000.01.01~2009.12.31
	Out of sample	2010.01.01~2015.09.15

서 RP(regression problem)라고 정의한다. 다음은 신호를 직접 구하는 방법이다. 이는 -1, 0 그리고 +1로 클래스(class)를 나눠서 sell, hold 그리고 buy 로 분류(classification)을 하였다. 이는 본 연구에서 CP(classification problem)라 정의한다. 즉 In sample 기간과 Out of sample 기간을 각각 RP와 CP를 이용하여 신호 값을 추출한다.

우선 In sample에서 RP 기반으로 실험을 한 경우, SVM는 precision이 0.75, recall이 0.61로 인공신경망 학습 기법과 라쏘 회귀분석에 비해서 우수했다. 여기서 precision은 예측력을 나타내며 recall은 재현율을 의미한다. 예컨대 precision는 1이라고 예측한 것 중에서 실제로 1인 것의 비율이고, recall은 실제로 1인 것 중에서 정확한 예측을 얼마나 했는지를 나타내는 값 이다. 즉 전체 1 중에서 몇 개를 예측했는지를 의미한다. precision과 recall을 산출하는 식은 다음과 같다.

$$precision = \frac{TP}{TP+FP}$$

$$recall = \frac{TP}{TP+FN}$$

TP	1이라고 예측, 실제 값 1
FP	1이라고 예측, 실제 값 0
FN	0이라고 예측, 실제 값 1

CP 기반으로 실험한 결과도 precision과 recall 모두 약 0.79로 ANN에 비해서 SVM의 성과가 좋았다. 라쏘 회귀분석의 경우에는 모형 특성 상, CP 기반으로 실험이 불가능하다. Out of sample에서 RP 기반으로 실험을 한 경우, ANN이 precision이 0.57로 가장 예측력이 높았으며 recall 또한 ANN이 0.55로 가장 우수한 결과를 보였다. CP 기반으로 실험을 한 경우에는 SVM이 precision과 recall 모두 각각 0.54, 0.52로 가장 우수한 성과를 보였다. 성과가 좋았던 학습 방법론의 경우에 precision이 recall에 비해서 산출된 값이 다소 높은 것을 알 수 있다. 일반적으로 학습 기법에 대한 성과를 측정하기 위해서 precision과 recall의 수준을 함께 봐야하기 때문에 평균을 구한다. 그러나 평균을 산출하면 학습 모형의 학습이 잘 되었는지 판단하는 것에 한계가 있다. 예컨대 In sample의 RP를 기반으로 실험을 했을 때, 라쏘 회귀분석의 precision과 recall의 평균은 0.29이다. 하지만 이는 precision이 0.48로 recall에 비해서 월등하게 높은 것을 알 수 있다. 이처럼 한쪽으로 치우친 것은 학습 모형이 잘 구성되지 않는 것을 의미한다. 따라서 본 실험에서는 이를 보완하기 위해서 F Score를 산출하며 산출 식은 다음과 같다.

$$F = 2\frac{PR}{P+R}$$

P	precision
R	recall

F Score를 산출했을 경우에 In sample의 경우에 RP 기반은 SVM이 0.67로 가장 우수하며, CP 기반도 SVM이 0.78로 가장 높았다. Out of sample의 경우에 RP 기반은 ANN이 0.54로 가장 높았으며, CP 기반은 SVM이 0.54로 가장 우수했다. 일반적으로 In sample에서 F Score가 0.8 이상이면 학습률이 매우 높은 모형이며 Out of sample에서는 F Score이 0.5 이상이면

Table 4. 실험 결과

		Model	Precision	Recall	Average	F_1 Score
In sample	regression (=RP)	SVM	0.75	0.61	0.68	0.67
		ANN	0.63	0.47	0.55	0.54
		Lasso	0.48	0.10	0.29	0.17
	classification (=CP)	SVM	0.79	0.78	0.79	0.78
		ANN	0.18	0.20	0.19	0.19
		Lasso	N/A	N/A	N/A	N/A
Out of sample	regression (=RP)	SVM	0.20	0.30	0.25	0.24
		ANN	0.52	0.57	0.55	0.54
		Lasso	0.42	0.04	0.23	0.07
	classification (=CP)	SVM	0.54	0.55	0.55	0.54
		ANN	0.41	0.44	0.43	0.42
		Lasso	N/A	N/A	N/A	N/A

예측력이 높다고 한다. 또한 In sample data와 Out of sample data의 F Score가 차이가 매우 큰 것을 알 수 있다. 일반화 오류(generalization error)가 상대적으로 큰 이유는 KOSPI 지수가 기술적 지표들의 과거 데이터이외에도 수많은 시장 상황에 영향을 받기 때문이라고 판단된다.

5. 결 론

본 연구는 최근 관심이 높아진 기계 학습을 이용하여 한국의 대표적인 주가 지수인 KOSPI 지수를 예측하고 그 성능을 분석하였다. 특히 기계 학습에 대한 관심이 높아지면서 금융 분야 또한 로보어드바이저라는 기계 학습을 이용한 매매 시스템에 대한 관심이 더욱 높아지고 있다. 이는 거래 비용과 이상적인 의사결정을 할 수 있다는 장점이 있는 반면에 다양한 금융 시장에 예측하기 어려운 시장 변수로 인해서 정성적인 판단에는 미흡하다는 우려도 함께 존재한다.

본 연구에서는 전통적인 학습 기법인 SVM과 라쏘 회귀분석 그리고 인공신경망 학습 모델을 갖고 KOSPI 지수의 예측력을 연구함으로써 현재 트렌드(trend)가 실효성이 있는지 분석하였다. 약 15년의 실험 기간을 학습 기간과 실험 기간으로 나눠서 실험해본 결과 라쏘 회귀분석의 경우에는 예측력이 매우 저조한 반면에 SVM과 인공신경망 학습 모델의 경우에는 예측력이 학습 데이터와 실험 데이터 모두 0.5보다 높았다. 이처럼 기계 학습을 이용하면 지수의 방향성을 예측하면 인간의 심리적인 부분을 통제하고 실제 과거 데이터에 의존하여 정량적인 의사결정을 할 수 있다는 장점이 있다. 또한 시간이 지나면서 축적되는 데이터를 바탕으로 의사 결정을 위한 모델을 강화할 수 있기 때문에 시간이 지나면서 지능형 운용 시스템을 구축해 갈 수 있다. 이렇게 구축된 시스템은 거래 수수료 등 운용에 대한 비용을 절감할 수 있다.

본 연구를 통해서 아무리 우수한 예측 변수(predictor)를 가지고 있다고 하더라도 미래 주가의 움직임을 예측하는 것은 쉽지 않다는 것을 알 수 있었다. 또한 교차 타당화(cross-validation)와 같은 기법을 이용해야만 최적화된 결과를 얻을 수 있을 것으로 사료된다.

References

[1] 김선웅, 안현철,“선물시장의 시스템트레이딩에서 동적시간와핑 알고리즘을 이용한 최적매매빈도의 탐색 및 거래전략의 개발”, 한국데이터정보과학회지, 제22권, 제2호, pp. 255-267, 2011.

[2] 김정래, 정찬기, “전화통화 빅데이터 분석에 관한 연구”, 한국정보기술아키텍처논문지, 제10권, 제3호, pp. 387-397, 2013.

[3] 이석준, 오경주, “Support Vector Machines와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발”, 한국데이터정보과학회지, 제16권, 제1호, pp. 71-92, 2010.

[4] 유재필, 신현준, “시계열의 역상관관계를 이용한 KOSPI200 지수선물 매매 전략”, 한국파생상품학회, 제22권, 제4호, pp. 7231-746, 2014.

[5] 유재필, 한창훈, 신현준, “빅데이터 트렌드를 이용한 섹터 투자 전략”, 한국정보기술아키텍처논문지, 제13권, 제1호, pp. 111-122, 2016.

[6] 허준영, 양진용, “SVM 기반의 재무 정보를 이용한 주가 예측”, 한국정보과학회, 제21권, 제3호, pp. 167-172, 2015.

[7] Amilon, H., “GARCH estimation and discrete stock prices : an application to low-priced Australian stocks”, Economics Letters, Vol. 81, No. 2, pp. 215-222, 2003.

[8] Bekiros, S. and Georgoutsos, D., “Direction-of-Change Forecasting using a Volatility-Based Recurrent Neural Network”, Journal of Forecasting, Vol. 27, pp .407-417, 2008.

[9] Jeantheau, T., “A link between complete models with stochastic volatility and ARCH models”, Finance Stochast, Vol. 8, pp. 111-131, 2004.

[10] Kanas A., “Non-linear forecasts of stock returns”, Journal of Forecasting, Vol. 22, No. 4, pp. 299-315, 2003.

[11] Jae Yeon Park, “Predicing Stock Returns using Machine Learning Algorithms”, New York University, 2015.

[12] Tay, F. E. H. and Cao, L., “Application of support vector machines in financial time series forecasting”, Omega, Vol. 29, No. 4, pp. 309-317, 2001.

[13] Yang, B., Li, L. X., and Xu, J., “An early warning system for loan risk assessment using artificial neural networks”, Knowledge-Based Systems, Vol. 14, pp. 303-306, 2001.

[14] Yoon and Swales, “Predicting stock price performance: a neural network approach”, System Sciences, Vol. 4, No. 5, pp .156-162, 1991.

[15] Wong, L. K., Leung, F. H. F. and Tam P. K. S, “An Improved Lyapunov Function Based Stability Analysis Method for Fuzzy Logic Control Systems”, Electronics Letters, Vol. 36, pp. 1085-1086, 2000.

박재연(Jae Yeon Park)
뉴욕대학교 석사
현재 : 상명대학교 박사과정, KIS채권평가 금융공학연구소 선임연구원
관심분야 : 금융공학
이메일 : jaeyeonpark@kispricing.com

유재필(Jae Pil Ryu)
상명대학교 학사, 석사
현재 : 상명대학교 박사과정, KIS채권평가 금융공학연구소 선임연구원
관심분야 : 금융공학, 데이터마이닝
이메일 : jaepilryu@kispricing.com

신현준(Hyun Joon Shin)
고려대학교 학사, 석사, 박사
미국 Texas A&M 대학교 연구원
(주)삼성전자 책임연구원
현재 : 상명대학교 경영공학과 교수
관심분야 : 금융공학, 조합최적화 응용, 데이터마이닝
이메일 : hjshin@smu.ac.kr