K-Nearest Neighbors(K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구

Predictability Test of K-Nearest Neighbors (K-NN) Algorithm : Application to the KOSPI 200 Futures

저자 (Authors)

김명현, 이세호, 신동훈
Myeong-Hyeon Kim, Seho Lee, Dong-hoon Shin

출처 (Source)

대한경영학회지 28(10), 2015.10, 2613-2633(21 pages)
Korean Jouranl of Business Administration 28(10), 2015.10, 2613-2633(21 pages)

발행처 (Publisher)

대한경영학회
DAEHAN Association of Business Administration, Korea

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06558615

APA Style

김명현, 이세호, 신동훈 (2015). K-Nearest Neighbors(K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구. 대한경영학회지, 28(10), 2613-2633

대한경영학회지
제28권 제10호(통권 132호)
2015년 10월(pp. 2613-2633)

http://dx.doi.org/10.18032/kaaba.2015.28.10.2613
Print ISSN 1226-2234 / Online ISSN 2465-8839

K-Nearest Neighbors(K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구*

김명현(고려대학교 경영대학 박사)
이세호(고려대학교 이과대학 수학과)
신동훈(인하대학교 글로벌금융학과 교수)

요약

본 논문에서 저자들은 머신러닝의 패턴분석기법 중 하나인 K-nearest neighbors(K-NN) 알고리즘을 KOSPI200 선물지수에 적용, 동 알고리즘을 이용한 기술적 분석의 예측력을 검증했다. 기술적 분석의 예측력 검증은 효율적 시장가설과 밀접한 연결고리가 있다. 효율적 시장가설에서 강형의 성립은 사적 내부정보(Private Information Set)를 이용해야만 시장에서 초과수익률 창출이 가능하다는 것으로써, 과거 가격 시계열의 움직임을 고려해서 투자하는 기술적 분석 혹은 차트 분석의 경우 현재 시장가격에 반영이 되어 있기 때문에 초과수익률 창출이 불가능하다는 것을 의미한다. K-NN 알고리즘은 머신 러닝의 대표적인 비모수 및 비선형 알고리즘으로 금융 시계열 데이터를 이용한 기술적 분석에 이 알고리즘을 선택한 이유는 다음과 같다. K-NN 알고리즘은 기계 학습의 방법 중 가장 간단한 방법으로 분류되며 모형 위험(Modeling Risk)을 최소화할 수 있다는 장점이 있다. 또한 정상성(Stationary)의 제약조건을 벗어나 비정상성의 동학을 갖는 가격 레벨에서 분석을 진행할 수 있기 때문에 실제 시장참여자들의 투자패턴을 그대로 적용하는데 용이한 점이 있다. 단변량 분석의 결과 K-NN의 두 가지 방법론 중 절대거리(Absolute) 방법론은 선물지수 하락기에 실현된 값보다 지속적으로 과대 예측하는 경향을 보였고, 반면 지수 횡보기에는 예측에 변동을 보이는 상관계수(Correlation) 방법론보다 안정적인 예측력을 보였다. 미결제약정과 프로그램 순매수 변화를 독립변수로 고려해 분석한 다 변량 분석의 결과 두 독립변수들의 추가적인 예측 기여도는 제한적인 것으로 나타났으며, 미결제약정 변수의 예측력 감소는 기존 논문의 결과와 배치되고 있어 추가 연구가 필요한 것으로 보인다. 또한 기존의 중요 기술적 지표들에 K-NN 알고리즘을 결합할 경우, 기술적 지표 자체를 이용한 투자전략보다 뛰어난 거래결과를 보임을 확인하였다. 본 논문은 약형 효율적 시장가설 관점에서 머신러닝 알고리즘을 적용해 기술적 분석의 유효성을 검증했다는 점에서 의미가 있다. 또한 K-NN 방법론의 KOSPI200 선물 적용은 본 논문에서 최초로 시도하는 것으로 국내 선물시장의 효율성 검증에 새로운 의미를 가져다줄 것으로 기대된다.

주제어: K-Nearest Neighbors, 기술적 분석, 효율적 시장가설, 예측력 검증, KOSPI200 선물

· 접수일(2015. 8. 6), 수정일(2015. 10. 22), 게재확정일(2015. 10. 22), 게재일(2015. 10. 31)
* 이 논문은 2015년도 인하대학교 학술연구비 지원에 의한 것임.
본 논문의 심사과정에서 유익한 조언을 해주신 익명의 두 분의 심사위원님께 감사드립니다.

ⓒ Korean Journal of Business Administration
2015, Vol. 28, No. 10, 2613-2633.
http://dx.doi.org/10.18032/kaaba.2015.28.10.2613
Print ISSN 1226-2234 / Online ISSN 2465-8839

Predictability Test of K-Nearest Neighbors (K-NN) Algorithm: Application to the KOSPI 200 Futures

Myeong-Hyeon Kim(Korea University)
Seho Lee(Korea University)
Dong-hoon Shin(Inha University)

Abstract

In this paper, we apply K-nearest neighbors (K-NN) Algorithm being one of the pattern analysis techniques of machine learning to the KOSPI 200 futures index, and test the forecasting power of the technical analysis of the algorithm. The test of predictive accuracy for technical analysis has close links with the efficient market hypothesis. The strong form of the efficient market hypothesis implies that primitive information must be used to generate excess returns in the market. Therefore, a technical analysis or chart analysis of the investment taking into account the past price movement time series is impossible to generate the excess return because all informations for pricing are reflected in current market prices. As K-NN algorithm is the representative non-parametric and non-linear algorithm on machine learning, the reason to take this algorithm for the technical analysis of financial time-series data is following. First, K-NN algorithm is the simplest method among machine learning methods, so we minimize the modeling risk from the analysis of this algorithm. Second, it allows analyzing on price levels whose dynamics can be non-stationary. Hence, it is easy to apply the actual investment pattern of market participants.

K-NN algorithms for analyzing the univariate time series can be separated by two sub category methods, an absolute distance method and the correlation method, depending on how to measure the neighborhood. As the reason, we compared the predictive powers of the two ways. As the result of the univariate analysis, the absolute method, as the one of two K-NN algorithm methods, tended to consistently over-predicted than the realized value on the downturn of the future index. While the index walked sideways, the absolute method showed more reliable predictive power than the correlation method showing volatile prediction. Since K-NN algorithm analyzing a univariate sime series predict with using only past data set, there is a disadvantage that no additional information set is available. Therefore, it is available to study the effectiveness of aditional analysis containing

more than one information set which help to increase predictive power. Candidates of the independent variables were selected for the two variables closely associated with connecting KOSPI200. As the first independent variable, we considered open interest to see the effect of an increase of the net quantity to prices of futures and options in a new contract. As the second independent variable, we considered the program net buying because the derivative market is freakishly large comparing with the underlying market in Korea. As the results of the multivariate analysis considering open interest and net buying in program trading, the additional contributions of two variables for the forecasting was limited, and the reduction of the prediction power of the open interest is contrary to the results of the existing papers, so it seems to need further studies. Also, a combination of K-NN algorithm of the existing main technical indicators confirmed to show superior trading results than the investment strategies dealing with the main indicators.

This paper is meaningful in that it verifies the validity of technical analysis to apply machine learning algorithms in terms of a weak-form of efficient market. In addition, as the first attempt in domestic market, the application of K-NN algorithm to the KOSPI 200 futures market is expected to bring new meaning to verify an efficiency of the domestic futures market.

Recently, a number of studies of various subjects have coming out such as effects of algorithmic trading on recent markets, contentions against the efficient market hypothesis, and verifying the validity of such trading strategies. Adjusting this trend, this study applying machine learning method to KOSPI200 contributes to have led to new finding that the predictive power using the univariate time series is not worse than multivariate time series.

Keywords: K-nearest Neighbors, EMH, Algorithm Trading, KOSPI200 Futures, Predictability Test

Contents

Ⅰ. 서 론

주식가격의 예측가능성에 대한 연구는 Fama (1965)가 제시한 효율적 시장가설의 검증과 밀접하게 연결되어 있으며, 주가가 과거패턴 혹은 외생 변수를 이용한 예측치 가능한지의 여부는 금융시계열 분야에서 꾸준한 실증분석을 통해서 검증되어 왔다. 이를 위해 경제, 수학, 통계, 컴퓨팅 등 다양한 학문 분야의 학제 간 융합 연구를 통해 주식시장의 분석과 주식가격의 예측을 위한 연구가 이루어져 왔다. 최근 주식시장의 예측력과 관련해서 효율적 시장가설은 준 강형까지 성립한다는 연구결과가 대다수이다. Fama (1965)는 Runs Test를 통해 주가움직임은 확률보행과정을 따른다고 밝혀낸 바 있다. 이에 대해 Lo and Mackinlay(1988)은 시계열 상관관계 분석을 통해 약한 양의 관계를, Fama and French(1998)와 Campbell and Shiller(1988)의 연구는 예측 가능한 패턴이 존재함을 과거 주가움직임에 근거하여 주식을 사거나 파는 전략을 통해 밝혀낸 바 있다. 효율적 시장가설에서 강형의 성립은 사적 내부정보(Private Information Set)를 이용해야만 시장에서 초과수익률 창출이 가능하다는 것으로써, 과거 가격 시계열의 움직임을 고려해서 투자하는 기술적 분석 혹은 차트 분석의 경우 현재 시장가격에 반영이 되어 있기 때문에[1] 초과수익률 창출이 불가능하다는 것을 의미한다. 하지만 Cochrane(2011)가 지적하듯, 대부분의 시장에서 장기 수익률 예측의 효과가 단기 수익률 예측보다 좋다는 점 그리고 주식시장에서 배당수익률, 외환시장에서 국가간 이자율 스프레드 등 예측력 향상에 도움이 되는 기본적인 변수(Fundamental Variables) 등의 역할에 대한 연구들, 경기 변동과 강한 동행성(Procyclicality)을 띄는 공통요인이 예측력 향상에 기여한다는 연구 등도 속속 나오고 있어, 효율적 시장가설에 대한 적용 여부는 아직도 정답이 없는 질문이라 할 수 있다.

금융에서는 Ord, Koehler, and Snyder(1997) 연구를 포함, 시계열분석 방법론에 기반을 둔 연구가 주를 이루고 있다. 이를 위해 시계열이 무조건부 평균 주위에서 변동하는 정상성(Stationary)을 갖는 확률과정을 활용해왔다. 평균방정식(Mean equation)은 ARIMA(Auto Regressive Integrated Moving Average)로, 분산방정식(Variance equation)은 대칭이나 레버리지 효과를 포함한 다양한 GARCH(Generalized Auto-Regressive Conditional Heteroskedasticity) 류의 모형들을 통해 많은 연구가 된 바 있다. 하지만 확률과 통계의 추론 기법을 시계열에 적용하기 위해, 비정상성(Non-stationary)의 성질을 갖는 금융데이터는 보통 로그차분 혹은 추세와 순환으로 분해해서 정상성의 성질을 갖는 데이터로 변환해 예측에 적용해 왔다. 보통 분산모델의 예측력이 변동성 군집현상, 비대칭 반응 등을 고려해 평균모델보다 좋은 것으로 알려져 있다. 뿐만 아니라 주식 수익률의 예측을 위해 고려된 통계적 모형은 주가수익률 모집단생성구조(Data Generating Process)를 선형(Linearity)으로 가정했기 때문에 대부분이 선형 모형으로 경제적인 의미부여가 쉽다는 장점이 있다. 하지만 실제 주가 혹은 주가수익률이 비선형성을 종종 보인다는 점을 감안할 때, 통계적 모형 집합에 강한 제약이 부여됐음을

[1] 현재 효율적 시장가설은 정보반영의 여부를 기준으로, 약형 효율적 시장가설의 검증은 수익률 예측능력연구와 시장이례현상(Anomaly)에 관한 연구로 나뉘며, 준강형의 경우 사건연구로, 강형의 경우 사적정보연구를 통해 검증한다.

알 수 있다. 다양한 형태의 비선형 관계를 특정 모형설정에 의존하지 않고 추정하는 비모수(Non-parametric) 추정의 경우 모형의 후보 군이 무한대라는 모형설정의 문제와 난해한 경제적 의미부여의 문제는 있지만 예측력은 기존의 선형모형보다 월등함을 보여주는 사례(Guidolin, Hyde, McMillan, & Ono, 2009)가 많이 있다. 특히 높은 비선형성과 동적 성질을 갖는 문제에서 신경망(Neural Network) 혹은 SVM(Support Vector Machine) 분석 등 산업공학과 컴퓨터 공학 등에서 주로 개발된 머신 러닝 기법들은 분포가정이 내재된 통계적 기법(판별분석, 로짓, 프로빗 모형 등)보다 예측력이 더 뛰어나다는 연구가 많이 있으며, 이를 주식시장에 적용한 연구 역시 마찬가지 결과를 보이고 있다.

금융경제학적 접근과 달리, 머신 러닝(Machine Learning) 혹은 패턴인식(Pattern Recognition) 분야의 방법론들이 성과가 좋은 이유는, 주식시장의 시장 참여자인 개인 및 기관 투자자가 과거의 패턴으로부터 미래예측이 가능하다는 전제로 기술적 분석 혹은 차트분석이라 불리는 방법론을 믿고 자기실현적 거래를 하는 데 그 이유가 있다. 실제 기술적 분석 문헌은 선물과 외환시장 참여자들의 30~40%(Irwin & Brorsen, 1985; Brorsen & Irwin, 1987; Billingsley & Chance, 1996) 정도가 기술적 분석 혹은 차트 분석이 단기적으로 최대 6개월까지(e.g., Menkhoff, 1997; Cheung & Wong, 2000; Cheung, Chinn, & Marsh, 2000; Cheung & Chinn, 2001) 가격 결정에 중요한 요인으로 생각하고 있음을 밝혀낸 바 있다.

기술적 분석에 기반을 둔 거래유형의 한 예는 다음과 같다. 현재로부터 과거 200일 동안의 데이터를 이용한 이동평균 선을 강하게 믿는 시장참여자가 많다면, 거래상품의 가격이 그 이동평균 선의 값과 같을 때 다음 거래상품의 가격은 그 이동평균 값보다 상승할 것으로 예측해서 매수거래를 다수가 할 것이다. 수없이 많은 참여자들의 거래활동의 합이 시장에 자기실현적으로 나타나고, 큰 구조적 충격이 없다면, 그 이동평균 선에 기반을 둔 거래는 의미를 가질 것이다. 시장참여자들이 사용하는 기술적 분석은 이외에도 지지 및 저항 등의 가격 레벨을 활용하거나, 가격과 거래량을 조합하거나, 주가변동성을 이용하여 가격변동 대역을 탄력적으로 계산하는 등 다양한 가격 변동의 패턴을 찾으려 한다. 이외 엘리어트 파동, 피보나치, 쌍바닥, 헤드앤숄더, 갠각도(Gann), 다우이론 등이 기술적 분석의 일종이다. 합리적 투자자를 가정하는 전통적 금융경제의 가정 하에서는 이런 패턴들의 존재를 우연들이 누적된 동전던지기의 결과 혹은 데이터 스누핑(Data Snooping)으로 취급하기 때문에 전통적인 통계적 기법 외에 다양한 분석 방법론을 시도하지 않은 것이 전통적 금융경제학 접근의 한계로 지적된다. 이와 달리 행동 재무학(Behavioral Finance)에서는 제한된 합리성을 가정하는 등 좀 더 현실적인 시장참여자를 가정, 주식 시장의 가격이 주목할 만한 수준의 일정한 흐름을 보이는 것을 확인한 바 있다.

문헌상 기술적 분석의 정의는 "기술적 분석이란 미래 가격 트렌드를 예측하고자 하는 목적으로, 시장거래행위를 분석하는 연구"라 정의를 내릴 수 있다.[2] 흥미로운 점은 머신 러닝의 데이터 마이닝 정의가 "대용량 데이터를 활용해 의

[2] 헤지펀드인덱스 홈페이지: Technical analysis is the study of market action, for the purpose of forecasting future price trends.

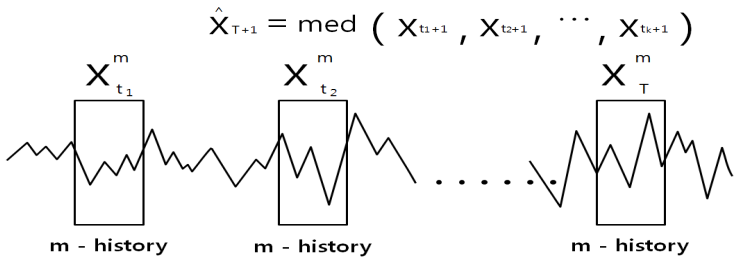


Figure 1. **Nearest Neighbor Prediction Technique:** Using non-overlapping m-period K sample data among all financial time series data, it predicts most present observation as the median value of the K sample data.

미 있는 패턴과 관계를 찾아내는 과정"[3]과 같이 주어지기 때문에, 대용량 데이터를 과거 자산 가격으로만 치환하면 두 정의는 같은 문제를 다루고 있음을 알 수 있다. 따라서 금융 시계열 데이터를 이용한 기술적 분석에 머신 러닝 기법의 적용은 자연스럽다 하겠다. 최근 들어 주식시장의 움직임과 주식가격 예측에 머신 러닝 기법들의 적용이 늘어나고 있다. 머신러닝기법의 금융데이터에 대한 적용은 Farmer and Sidorowich (1987)를 시작으로 Casdagli(1989), Qian, and Rasheed(2007), Tsai and Hsiao(2010), Sugihara and May(1990) 등이 대표적인 논문들이다. Bajo-Rubio, Rivero, and Rodríguez(2002)는 K-NN (K-nearest neighbors)기법을 외환시장에 적용해 K-NN방법론의 예측력을 확인하였다. 본 논문에서는 머신 러닝의 대표적인 비모수 및 비선형 알고리즘인 K-NN 알고리즘을 유동성이 풍부한 KOSPI200 선물시장에 적용하여 그 유용성을 분석한다. K-NN 알고리즘을 선택한 이유는 다음과 같다. K-NN 알고리즘은 기계 학습의 방법 중에 가장 간단한 방법 중 하나로 분류되고 있기 때문에 모형 위험(Modeling Risk)을 최소화할 수 있다는 장점이 있다. 이는 최절약원리(Parsimony)의 적용과도 관련 있으며, 다른 머신러닝방법론의 결론보다 강건한(Robustness) 결과를 생산할 수 있을 것으로 기대된다. 뿐만 아니라 정상성의 제약조건을 벗어나 비정상성의 동학을 갖는 가격 레벨에서 분석을 진행할 수 있어 실제 시장 참여자들의 투자패턴을 그대로 적용할 수 있기 때문이다.

본 논문의 주요한 발견과 공헌은 다음과 같다. K-NN 방법의 단 변량 분석의 결과 일반적으로 시장에서 애용되는 이동평균, RSI(Relative Strength Index), MACD 등의 기술적 분석의 시장보다 거래전략측면에서 RMSE(Root Mean Square Error)기준으로 예측성과가 나음을 보였다. 또한 두 가지 K-NN 방법 중 절대거리(Absolute) 방법론은 선물지수 하락기에 실현된 값보다 지속적으로 과대 예측하고, 지수 횡보기에는 예측에 변동을 보이는 상관계수(Correlation) 방법론보다 안정적인 예측력을 보이고 있음을 밝혀냈다. 미결제약정과 프로그램 순매수 변화를 독립변수로 고려해 분석한 다 변량 분석의 결과 두 독립변수들의 추가적인 예측 기여도는 제한

[3] 브리태니카 백과사전: The process of discovering interesting and useful patterns and relationships in large volumes of data.

적인 것으로 나타났다. K-NN의 단기예측력을 이용, 일별 데이터를 이용한 거래전략(Trading Rules)의 존재는 단기에 효율적 시장가설의 존재를 부정하는 하나의 증거로 고려될 수 있을 것이다. 뿐만 아니라, K-NN 방법론의 국내 KOSPI200 선물 적용은 본 논문에서 처음 시도하는 것으로 국내 선물시장의 효율성 검증에 새로운 의미를 가질 것으로 기대한다.

본 연구의 구성은 다음과 같다. 제 Ⅱ장에서는 K-nearest neighbors(K-NN) 알고리즘의 설명과 구현에 대해 설명한다. 제 Ⅲ장에서는 K-NN 알고리즘을 적용할 데이터에 대해 소개한다. 제 Ⅳ장에서는 K-NN의 적용결과와 기타 기술적 분석방법의 비교결과를 보인다. 마지막으로 제 Ⅴ장에서 연구결과를 요약하고 시사점을 제시한다.

Ⅱ. K-Nearest Neighbors 알고리즘

K-NN 알고리즘은 머신 러닝에서 분류 및 예측(Classification and Prediction)을 위해서 많이 사용되는 방법으로, 간단한 알고리즘 구조와 견고한 예측력을 보여주고 있어, 지금은 머신 러닝의 다양한 분야에 광범위하게 쓰이고 있다. Karlsson and Yakowitz(1987), Lall and Sharma(1996)은 K-NN의 정의를 종속변수의 최적추정치를 얻기 위해 설명변수와 종속변수의 관측 값들 사이에 유사성을 측정하는 방법으로 내린바 있다.[4] 국내에서는 김광용・이경락(2008)이 인공신경망 모형을 이용하여 상장기업의 주가를 예측하였다. 기본 아이디어는 새로운 입력 값이 일정한 기준 하에서 제시된 입력 값과 가깝다면 그 새로운 입력 값을 주어진 기준 하에서 제시된 입력 값이 속한 집단으로 구분하는 것이다. 특히 금융데이터의 적용에 있어서 SVM 및 신경망분석과 더불어 가장 자주 쓰이는 알고리즘으로 알려져 있으며, K-NN 알고리즘은 1개의 nearest neighbors(NN) 알고리즘을 K개로 확장한 버전이다. Kuramochi and Karypis(2005)는 금융데이터의 적용에 있어서 훨씬 정교한 구조를 띄는 SVM보다 K-NN 알고리즘의 성과가 뛰어나다는 결과를 제시하였다. 주가 예측관점에서 우선 NN 알고리즘은 다음과 같은 분류 프로세스를 갖는다.

우선 K-NN 알고리즘의 목적함수는 새로운 금융시계열데이터(주가)가 거래시스템에 도달했을 때, 그 새로운 데이터가 어떤 집단(거래전략에서는 매수/매도)에 속하는가를 최적으로 결정하는 것이다. 이를 위해 주어진 전체 금융시계열 데이터 중 얼마만큼의 데이터(Training Dataset)를 이용해 알고리즘을 훈련시킬 것인가, 유사한 패턴을 정하는 데이터 기간(m)은 얼마 만큼인가, 유사한 패턴을 몇 개(K)를 이용할 것인가에 대한 입력변수의 결정, 그리고 마지막으로 수학적인 의미의 유사하다(Similarity)의 계산 가능한 정의가 필요하다. 트레이닝 기간과 패턴 개수는 알고리즘 사용자의 선택의 문제로 다양한 값을 기반으로 한 시뮬레이션으로 모델의 강건성(Robustness)을 테스트 할 필요가 있다. 이외 유사하다의 수학적 정의는 크게 유클리디안 $D(x,\ x_i)\Sigma_{j=1}^{d}(x_j-x_{i,j})^2$ 혹은 가중평균 유클리디안 메저 $D(x,\ x_i)\Sigma_{j=1}^{d}w_i(x_j-x_{i,j})^2$를 이용한다. 이렇게 주어진 입력변수를 이용, 가장 근접한 K

[4] K-nearest neighbor (K-NN) methods use the similarity (neighborhood) between observations of predictors and similar sets of historical observations (successors) to obtain the best estimate for a dependent variable.

개의 데이터에 대한 다수투표기준(Majority voting), 혹은 가중평균합(Weighted sum)의 방식으로 매수/매도 거래전략(Classification)을 결정하게 된다. 1개의 NN 알고리즘은 1개의 근접 패턴을 보고 새로운 시계열데이터를 일방적으로 같은 분류(Class)에 분류하는 알고리즘으로, 주어진 시계열 데이터에 의미 있는 신호 외에 잡음이 껴 있을 경우 좋지 않은 성능을 나타낸다. 이와 같은 단점을 K-NN 알고리즘은 K개의 근접패턴을 통해 잡음을 Averaging Out 시키는 방식으로 극복하고, 주어진 신호를 통해 새로운 금융 시계열 데이터가 어떤 분류(Classification)에 속하는지 결정하는 것이다. 각 클래스에 대해 에러 비용(Costs of error)이 같다고 한다면, 미지의 샘플이 속하는 것으로 추정되는 클래스는 K개의 가장 가까운 이웃 집단 중에서 가장 흔하게 표현되는 클래스를 선택하는 것이 기본 원리이다.

하지만 만약 K가 너무 크다면, 하나의 근린(얼마나 가까운가)에 포함되는 데이터가 너무 많아져서 식별의 문제가 발생하며, 파라미터가 적용되는 대부분의 모델들이 그렇듯, K-NN 역시 최적의 K를 정하는 것이 쉽지 않다. 일반적으로는 여러 개의 K값을 대입 후 테스트한 뒤 가장 성능이 좋은 K값을 선택하는 과정을 거치며, 노이즈가 심한 데이터일수록 K값이 큰 것으로 알려져 있다. 따라서 다양한 값을 알고리즘에 적용한 뒤 최적의 값을 선택하는 과정이 필수적이다. K개의 근접패턴이 결정이 되면, 다수투표기준(Majority voting)은 다음과 같이 목적함수를 극대화시키는 방식으로 결정이 된다.

$$\text{Majority voting: } y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$$

$x = (x', y')$가 새로 도착하는 테스트 대상으로 $(x_i, y_i) \in D$ 데이터를 통해 트레이닝을 한 뒤, K-NN 알고리즘, 즉 z와 모든 D집합의 순서쌍 집합과의 거리(Distance) 혹은 비슷한 정도 (Similarity)를 계산하여 가장 근접한 K개의 nearest neighbors 리스트를 D_z가 구하는 과정을 거치게 된다. i번째 nearest neighbors y_i가 v 클래스에 해당이 되면 지시함수(Indicator Function)를 통해 카운트가 된다. Appendix I에 상기의 알고리즘이 정리되어 있다.

III. 데이터

KOSPI200 선물은 유가증권시장본부에 상장된 주식 200종목의 시가총액 기준으로 산출된 KOSPI200 지수[산출기준시점 1990.01.03]를 기초자산으로 하는 상품으로 전 세계에서 파생상품 중 거래량 기준 8위에 해당될 만큼 유동성이 풍부한 상품이다. 본 논문은 거래일 기준 2010년 12월 29일부터 2014년 12월 19일까지 총 1452개의 KOSPI200 선물 최근월물 거래가격을 이용해서 분석한다. <표 1>에서 정리된 바와 같이, KOSPI200 선물가격은 약 4년간 262근처에서 움직였으며, 평균미결제 약정은 106321, 거래량은 213574값을 갖는 것을 알 수 있다. 선물가격 변화에 있어서 특이한 점은 2011년 3분기에 급격한 가격 하락을 시작으로 4년간 종종 Fat-tail에 해당하는 가격움직임이 관측된다는 점이다.

2010년 12월 29일부터 2014년 12월 19일 동안 KOSPI200 선물 지수는 평균 회귀의 모습을 보이며, 박스 권에서 머무르고 있다. 대략 230을 기점으로 하방 경직성을 보이고 있음을 알 수 있다. K-NN 알고리즘은 데이터의 정상성 여부와 관련 없이 적용할 수 있는 장점이 있지만, 분석 데이터의 특징을 설명하기 위해 정상성 검증을 위한 ADF(Argumented Dickey-Fuller)

Table 1

Statistics of the Variables for 2011~2014 KOSPI200 Futures

Statistics	High	Low	Sattle	Open Interest
Mean	263.9726	260.8199	262.3724	106321.5
Median	262.3500	259.9000	261.1500	107726.0
Maximum	307.4500	302.8500	307.0500	142127.0
Minimum	232.7500	224.9500	225.6000	0.000000
Std. Dev.	13.86153	14.06340	13.97494	16525.72
Skewness	0.693679	0.548670	0.617554	-2.956381
Kurtosis	3.200907	3.171306	3.203382	20.05812
Jarque-Bera	118.8900	74.62690	94.79484	19719.39
Probability	0.000000	0.000000	0.000000	0.000000
Observations	1452	1452	1452	1452

테스트를 진행했다. <표 2>의 결과를 통해 이 기간 동안 데이터는 가격 레벨에서 정상성을 보임을 확인할 수 있다.

Table 2

ADF test for KOSPI200 Futures

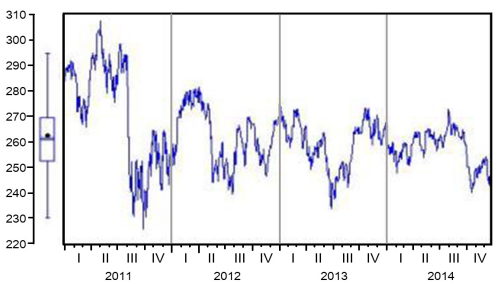
Null Hypothesis: KOSPI200F closed price has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic-based on SIC, maxlag = 23)

		t-Statistic	Prob.*
ADF test statistic		-3.3924	0.0114
Test critical values:	1% level	-3.4346	
	5% level	-2.8633	
	10% level	-2.5678	

* MacKinnon (1996) one-sided p-values.

특히 외부충격이 왔을 때, 상방으로의 반응보다 하방으로의 반응이 격하고 빠른 점이 관측이 되는데, 변동성 문헌의 비대칭성(Asymmetry) 속성도 나타나고 있음을 파악할 수 있다. 가격레벨 차원에서 정상성을 보이는 점은 데이터 선택기간에 따른 우연이며, 이를 제외하고 대부분의 변동성 관련 현상-비대칭성, 군집현상, 레버리지 효과 등을 보이고 있음을 확인했다.

본 논문은 총 1452개의 데이터 중에서 2010년 12월 19일부터 2014년 9월 10일까지의 거래데이터 1352개의 데이터를 훈련데이터(Training Data)로 이용해 K-nearest neighbors(KNN) 알고리즘을 적용하고, 2014년 9월 11일부터 2014년 12월 19일까지 100거래일 동안의 예측력을 테스트 한다.



Fiture 2. **KOSPI200 First Month Futures Time Series**

Ⅳ. Empirical Results

본 논문의 가장 큰 연구주제는 K-NN 알고리즘을 통한 KOSPI200 선물지수 예측이 가능한가

를 검증하는데 있으며 이 예측력의 성과가 기존에 알려진 대표적 기술적 분석-이동평균, RSI, MACD 등 보다 나은지 검증하는데 있다. 예측력 검증은 데이터의 종류에 따라 단일 시계열을 이용한 분석과 복수 시계열을 이용한 분석으로 나뉘는데, 본 논문은 KOSPI200 단일 시계열에 대한 분석만을 진행한다. Kuramochi and Karypis (2005)가 주장한 대로 SVM보다 K-NN 알고리즘의 성과가 뛰어나다는 결과에 따라, 머신 러닝의 다른 방법론인 SVM이나 신경망 등과의 비교 연구는 향후 연구주제로 남겨둔다. 뿐만 아니라 K-NN 알고리즘은 상대적으로 최절약원리(Parsimony)에 기반한 모델링이 가능하기 때문에 본 논문은 K-NN 알고리즘의 예측성과를 직접 측정하는데 중점을 둔다. 본 논문의 예측력 비교는 일반적인 통계적 예측오차를 측정하는 오차제곱 합을 제곱근한 RMSE(Root Mean Squared Error)기준으로 성과비교를 한다.

4.1 단변량 분석

단일 시계열을 분석하는 K-NN 알고리즘은 근린을 측정하는 방식에 따라 절대거리 방법과 상관계수(Correlation) 방법으로 구분이 된다. 이에 하부 연구주제는 두 방법 중 어떤 방법의 예측력이 더 뛰어난가를 탐구한다. 절대거리와 상관계수 방법에 관한 자세한 사항은 Farmer and Sidorowich(1987)와 Bajo-Rubio et al.(2002)을 참고하길 바란다. 2010년 12월 19일부터 2014년 9월 10일까지의 거래데이터 1352개의 데이터를 훈련데이터(Training Data)로 이용해서, 2014년 9월 11일부터 2014년 12월 19일까지 1거래일 예측(1-step ahead forecasting) 값의 100거래일 예측 결과는 다음 <그림 3> 시계열 그래프에 나타나 있다.

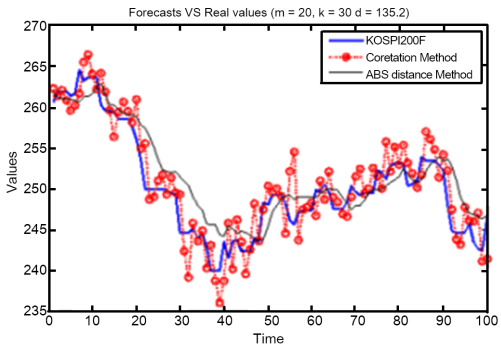
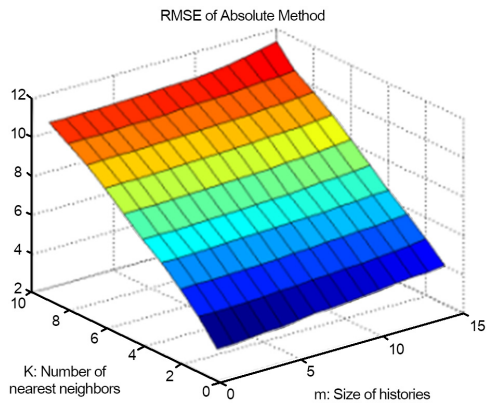


Figure 3. **KOPSI200 Futures Index and the Forecasting Time Series Graphs:** It shows the KOSPI200 connection futures index during 100 trading days from September 11, 2014 to December 19, 2014, and the predictive time series by absolute method and corelation method.

상관계수 방법론은 실현된 KOSPI200 선물지수 값 근처를 중심으로 변동(Variations)을 보이고 있지만 실현된 값을 평균적으로 잘 예측하고 있음을 볼 수 있다. 반면 절대거리 방법론의 경우, 선물지수 하락기에 실현된 값보다 큰 값을 지속적으로 예측하고, 지수 횡보기에는 변동을 보이는 상관계수 방법론보다 안정적인 예측력을 보이고 있다. K-nearest neighbors 알고리즘의 경우 주어진 전체 금융시계열데이터 중 얼마만큼의 테스트 데이터를 이용해 알고리즘을 훈련시킬 것인가, 유사한 패턴을 정하는 데이터 기간은 얼마만큼 인가, 유사한 패턴을 몇 개를 이용할 것인가에 대한 입력변수의 결정에 따라 결과 값이 얼만큼 변하는지에 대한 민감도 분석이 필수적이다. 다음 하단의 <그림 4>는 유사한 패턴을 정하는 데이터 기간, 유사한 패턴을 몇 개를 이용할 것인가에 해당하는 파라미터를 각각 20부터 30일, 30부터 45일까지 변화시켜 각 파라미터 순



Fiture 4. **Panel A RMSE of Absolute Method**

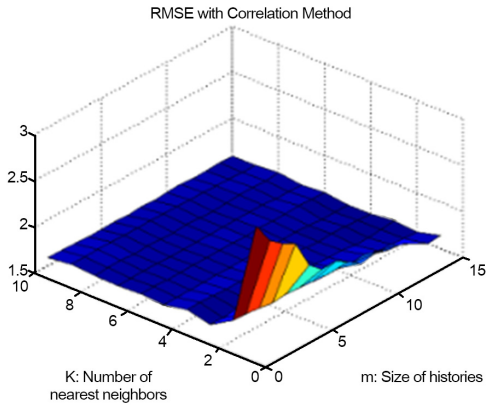


Figure 4. **Panel B RMSE of Correlation method**

Figures are illustrates the RMSE values under the change of variables in the absolute method(Panel A) and in the corelation method(Panel B)

서쌍 별로 RMSE를 보여주고 있다. <그림 4>에서 X축은 유사한 패턴을 정하는 데이터 기간, Y축은 패턴의 개수 그리고 Z축은 RMSE를 나타낸다.[5] 상관계수 방법론의 경우, 유사한 패턴을 정하는 데이터 기간과 유사한 패턴을 몇 개를 사용할 것인가를 정하는 파라미터에 대한 민감도가 초기 변화를 제외하고는 매우 안정적임을 볼 수 있다. 반면 절대거리 방법론의 경우, 유사한 패턴을 정하는 데이터 기간에 대한 민감도는 변화가 크지 않지만 유사한 패턴(K)을 몇 개를 사용할 것인가에 대한 민감도 반응은 그 개수의 증가에 따라 RMSE가 단조증가(Monotonic)함을 알 수 있다. 따라서 절대거리 방법론의 경우는 민감도 분석의 결과 패턴의 개수(K)를 적게 가져가는 것이 필요하겠다.

4.2 다변량 분석

단일 시계열을 분석하는 K-NN 알고리즘은 과거 값들만 이용해서 예측을 하기 때문에 KOSPI 200이 생성하는 정보집합 외에 이용할 수 있는 정보가 없는 단점이 있다. 추가로 관심변수의 예측에 도움이 되는 독립변수를 예측에 활용했을 때 예측성과는 어떻게 될 것인가 하는 것이 확장 연구주제가 된다. 독립변수의 후보군은 연결선물 KOSPI200과 밀접하게 연결된 두 변수를 선정했다. 우선 첫 번째 독립변수는 선물옵션의 신규계약진입-균형에서 순 수량의 증가가 가격에 미치는 효과를 따져보기 위해 미결제약정을 고려했다. 미결제약정의 예측력은 Figlewski(1981)과 Hong and Yogo(2012) 등의 결과로부터 선정했다. 두 번째 독립변수는 파생상품시장이 기초자산보다 기형적으로 큰 한국 파생시장의 독특한 특성 때문에 프로그램 순매수를 독립변수를 선정했다.

특히 선물옵션 거래는 한국 시장에서 독특한 위상을 띄는데, 이는 '왝더독(Wag the Dog)'이라 불리는 현상에서 의미를 찾을 수 있다. 왝더독은 개의 꼬리가 몸통을 흔든다는 뜻으로, 주

[5] RMSE의 계산을 위해 편의상 데이터는 평균과 분산을 조정했다.

식시장에서 선물시장(꼬리)이 현물시장(몸통)에 큰 영향을 미치는 현상을 가리킬 때 보통 사용한다. 주식시장에서 왝더독은 기관투자자와 외국인 등이 선물시장과 현물시장의 가격 차이를 이용하는 '차익거래' 기법을 흔히 사용하기 때문에 주로 발생하며, KOSPI200 주가지수와 KOSPI200 선물지수의 차이를 따져, 비싼 것은 팔고 상대적으로 값이 싼 것을 사들이는 과정에서 프로그램 매매를 통해 대량 매수 또는 매도 주문을 내게 되면 선물시장이 현물시장에까지 영향을 미치는 것이다. 따라서 두 번째 독립변수 후보 군은 프로그램 순매수와 관련된 변수들을 고려한다.

<그림 5> Panel A는 KOSPI200 미결제 약정수량, Panel B는 차익/비차익 프로그램 순매수 그래프를 나타낸다. 차익 비차익 프로그램 순매수는 프로그램 순매수(NETF), 차익 순매수(NETS), 비차익 순매수(NETPRO)를 고려했다.

<그림 5>의 Panel A는 KOSPI200 선물지수의 미결제 약정 시계열로, 각 만기 날 일부는 청산되거나 일부는 다음 만기로 롤오버(Roll-over) 되는 움직임을 보이고 있음을 알 수 있다. 미결제약정이 파생상품의 가격에 미치는 영향은 다음과 같다. 투자자들이 기존의 매매 물량을 청산하지 않는 상태에서 추가로 신규매수나 신규매수 물량을 늘린다면 미결제약정 수량이 증가하게 되며, 전매나 환매로 많이 청산하면 미결제약정 수량은 줄어든다. 따라서 앞으로 강세장이 예상된다면 선물·옵션을 신규 매수한 투자자는 목표수익에 도달할 때까지 청산하지 않고 신규 매수를 계속하여 늘릴 것이기 때문에 선물가격이 상승하면서 미결제약정 수량이 증가한다면 향후 KOSPI200 지수의 가격이 계속 상승할 것이라고 선물시장 참여자들이 기대가 형성되고 있음을 추론할 수 있다. K-NN 알고리즘 관점에서, 패턴인식을 강화시키는 역할을 하는 보조 데이터라 할 수 있다.

<그림 5>의 Panel B는 KOSPI200 현/선물 거래와 관련된 프로그램 순매수 시계열을 나타낸다. 프로그램 매매는 차익거래와 비차익거래로 나뉘며, 차익 거래는 KOSPI200 선물과 현물을 사고팔면서 차익을 올리는 전략이다. 비차익거래란 선물과 상관없이 KOSPI200 구성종목 중 15개 종목 이상으로 바스켓(Basket)를 구성해 일시에 거래하는 프로그램 매매의 일종으로 비차익거래는 지수의 등락에 따라 시가총액 상위 대형주를 일시에 사들이거나 팔아 치운다. 비차익거래에는 외국인과 국내 기관투자가들이 주로 사용하며, 외국인이 지수관련 대형주를 일시에 사들이는 비차익거래 순매수를 통해 KOSPI 현

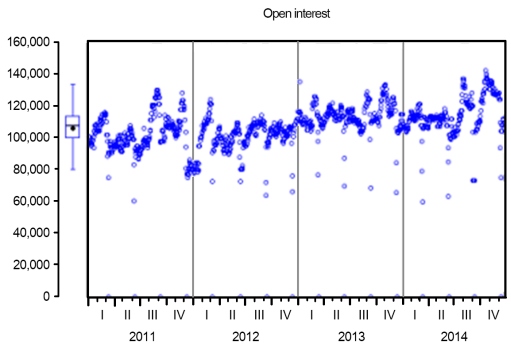


Figure 5. **Panel A Open interest rates in KOPSI200**

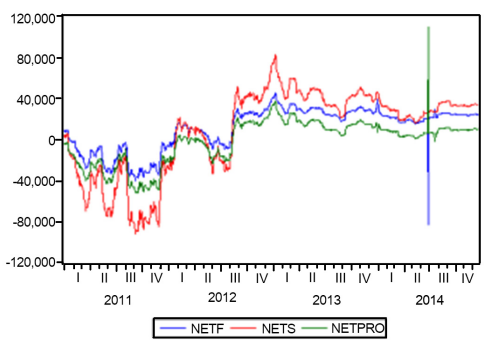


Figure 5. **Panel B Program net buying graph**

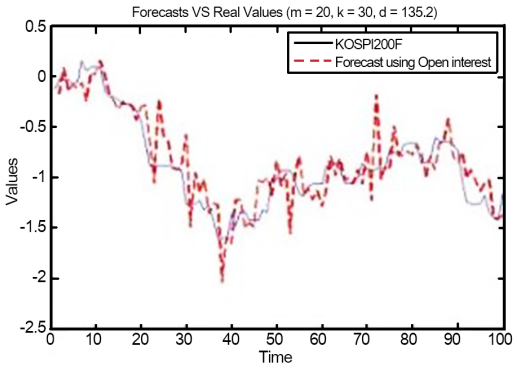


Figure 6. **Panel A Forecasting graph using open interest**

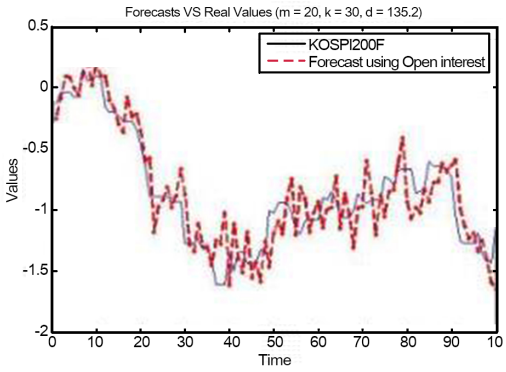


Figure 6. **Panel B Forecasting graph using program net buying**

물지수를 상승시키는 일이 빈번히 일어난다.

<그림 6>의 그래프는 KOSPI200 미결제 약정수량, 차익/비차익 프로그램 순매수를 독립변수로 이용한 예측 그래프를 나타낸다. 단일 시계열을 이용한 분석한 <그림 3>보다 미결제 약정 및 프로그램 순매수를 독립변수로 이용할 경우, 변동 폭의 변화를 볼 수 있다. 특히 미결제 약정을 이용할 경우 단일 시계열만 이용할 때보다 예측 변동이 상당히 감소함을 확인할 수 있다. <표 3>은 좀 더 엄밀한 비교를 위해 단일시계열, 미결제약정, 프로그램 순매수를 이용한 RMSE 값을 담고 있다.

Table 3

RMSE Comparisons of K-NN Algorithm Multivariate Analysis[6]

Method	Single time series		Open Interest	Net purchase by program
	Corelation method	Absolute method		
RMSE	0.1865	0.2094	0.2248	0.1996

[6] 다변량 예측을 위한 데이터는 표준화 함.

미결제 약정의 경우, 예측 시계열상의 변동은 작아지는 걸 확인할 수 있지만 실제 RMSE 값은 단일 시계열 방법론 두 개, 미결제 약정과 프로그램 순매수를 독립변수로 한 방법론 총 네 가지 방법에서 가장 큰 값을 갖기 때문에 미결제 약정의 경우 예측력 향상에 득이 되는 정보를 담고 있지 않거나, 그 효과가 제한적임을 알 수 있다. 프로그램 순매수를 독립변수로 활용할 경우, 단일 시계열의 상관계수 방법론보다 못하며, 절대거리 방법론보다 예측력 향상에 미미하게 도움이 되고 있다. 전체적으로 <표 3>은 시장에서 의미 있다고 여겨지는 미결제 약정, 프로그램 순매수 등의 수량 정보가 KOSPI200 선물 예측에 큰 도움이 되지 않음을 확인시켜주고 있다.

4.3 이동평균지표와의 성과비교

K-NN 알고리즘의 예측값과 같이 시장참여자들이 가장 많이 사용하는 가격레벨의 기술적 지표는 이동평균으로 Brock, Lakonishok, and Lebaron(1992)의 논문 결과가 대표적이다. 이들은 여러 기술적 지표 중, 단순이동평균법과

대역대돌파(Range Break)의 유효성을 검증했으며, 전반적으로 기술적 전략의 사용의 유효성을 지지하고 있다. 여기에 Blume, Easley, and O'hara(1994)은 과거 주가변화의 정보에 관한 질과 정확도에 거래량이 추가적 정보를 부여함을 밝힌바 있다. 국내는 김형도(1995)가 기술적 지표의 활용이 단순 매입보유 전략보다 높은 수익률을 제공함을 보고했으며, 이윤선(1999, 2002)은 단순주가이동평균을 이용한 매매전략의 의미 있는 성과가능성을 기술한 바 있다. 유일성(2011)은 뉴욕증권시장과 KOSPI200 선물시장을 동조화 시킨 기술적 거래전략의 수익실현가능성에 대해 연구했으며, 김상환 · 조태근(2003)은 거래소시장보다 코스닥시장에서 기술적 분석의 투자성과를 신뢰할 수 있음을 White의 Reality Check(RC)방법과 Hansen의 Superior Predictive Ability(SPA)기준으로 검증하고 있다. 본 연구에서 K-NN 알고리즘과 기술적 지표의 대표인 이동평균의 예측력을 RMSE 기준으로 직접 비교해 보고 그 결과는 <표 4>에 보고한다. Panel A 결과는 가격레벨에서의 K-NN 알고리즘과 여러 다양한 기간을 활용한 이동평균의 예측오차를 보이고 있다. 모든 이동평균기간의 예측결과보다 상관계수 및 절대거리 방법론의 예측오차가 작음을 알 수가 있으며, 상관계수방법론의 예측오차가 절대거리 방법론의 예측오차보다 작음을 알 수 있다. 실제 거래전략의 결과를 측정하는 누적수익률 및 샤프비율을 보고한 Panel B를 통해 K-NN 알고리즘의 활용가능성을 알아볼 수 있다. 이동평균 투자기법은 실무적으로 제일 많이 활용이 되는 방법인 단기 5일선이 장기 20일선보다 커지는 시점에서 매수한 뒤 5일선이 20일선보다 작아지는 시점에서 매도하는 전략을 선택했다. K-NN 알고리즘의 경우, 동 알고리즘을 통해 예측한 값을 기준으로 같은 투자기법을 적용한 결과이다. 누적수익률 기준으로 상관계수 방법론은 이동평균기법보다 2.54배, 절대거리 방법론은 4.58배 더 나은 결과를 보이고 있다. 추가로 가격의 상승압력과 하락압력 간의 상대적인 강도를 나타내는 RSI(Relative Strength Index)의 결과도 함께 보고한다. RSI 기법의 결과는 상관계수 방법론의 경우만 K-NN의 성과가 나음을 보이고 있어, 결과적용 및 해석이 절대적이지 않음을 강조한다. 자세한 거래전략의 성과비교는 Appendix Ⅱ를 참조하길 바란다.

Ⅴ. 결 론

본 연구는 머신 러닝의 대표적인 비모수 (Non-parametric) 알고리즘인 K-NN 알고리즘을 유동성이 높은 KOSPI200 선물시장에 적용하여 그 활용도와 예측의 유용성을 가격레벨에서 분석했다. 절대거리와 상관계수 두 방법론의 K-NN 알고리즘을 이용, 과거 가격패턴으로부터 의미있는 예측성과를 검증했다. 분석 방법은 KOSPI 200 데이터의 특성 따라 단일 시계열 방법론을 적용했다. 통계적 예측오차를 측정하는 오차제곱합을 제곱근한 RMSE 기준으로 분석했으며, 상관계수 방법론은 실현된 KOSPI200 선물지수 값 근처를 중심으로 변동(Variations)을 보이고 있지만 실현된 값을 평균적으로 잘 예측함을 밝혀냈으며, 절대거리 방법론의 경우, 선물지수 하락기에 실현된 값보다 큰 값을 지속적으로 예측하고, 지수 횡보기에는 변동을 보이는 상관계수 방법론보다 안정적인 예측력을 보이고 있음을 발견했다. 또한 다변량 시계열 분석에서는 선물옵션의 신규계약진입-균형에서 순 수량의 증가가 가격에 미치는 효과를 따져보기 위해 미결제

약정을 고려했으며, 한국파생시장의 독특한 특성 때문에 나타나는 프로그램 순매수와 관련된 수량변화를 독립변수를 고려해 분석했다. 미결제 약정의 경우 예측력 향상에 득이 되는 정보를 담고 있지 않거나, 그 효과가 제한적이었으며, 프로그램 순매수를 독립변수로 활용할 경우, 단일 시계열의 상관계수 방법론보다 못하며, 절대거리 방법론보다 예측력 향상에 미미한 효과를 미치는 것으로 분석됐다. 전체적으로 시장에서 의미 있다고 여겨지는 미결제 약정, 프로그램 순매수 등의 수량 정보가 KOSPI200 선물 예측에는 큰 도움이 되지 않음을 확인시켜주고 있다. 이는 과거 주가변화의 정보에 관한 질과 정확도에 거래량이 추가적 정보를 부여함을 밝힌 Blume et al.(1994)의 결과와 배치되는 발견으로 향후 추가연구가 요구된다.

최근 알고리즘 트레이딩이 시장에 미치는 효과, 효율적 시장가설과의 경합 여부, 거래 전략의 유효성 검증 등 다채로운 연구가 나오고 있다. 본 논문은 이런 흐름에 맞춰 머신러닝의 기법을 KOSPI200에 적용시켜 단일 시계열을 이용한 예측력이 기타 수량과 관련된 독립변수를 추가한 예측력보다 뒤처지지 않는다는 새로운 발견을 이끌어 냈다는데 의의가 있다.

References

김광용 · 이경락 (2008). 인공지능시스템을 이용한 주가예측에 대한 연구. 대한경영학회지, 21(6), 2421-2449.

김상환 · 조태근 (2003). 기술적 거래전략의 예측력 검정. 재무연구, 16(2), 67-93.

김형도 (1995). 한국주식시장에서의 기술적 분석을 통한 약형 효율시장가설 검증과 기술적 시스템의 수익성 분석. 증권 금융연구, 1, 49-67.

유일성 (2011). 증권시장 동조화와 기술적 거래전략-한국 주가지수 선물시장의 정보효율성. 대한경영학회지, 24(2), 837-857.

이윤선 (1999). 거래량을 이용한 투자전략에 관한 실증연구. 산업경제연구, 12(6), 63-74.

이윤선 (2002). 주가이동 평균선을 이용한 기술적 분석의 효과. 금융공학연구, 1, 1-20.

Bajo-Rubio, O., Rivero, S. S., & Rodríguez, F. F. (2002). Non-Linear Forecasting Methods: Some Applications to the Analysis of Financial Series. *FEDEA Working Paper*, 2002-01.

Billingsley, R. S., & Chance, D. M. (1996). Benefits and Limitations of Diversification Among Commodity Trading Advisors. *The Journal of Portfolio Management, 23*(1), 65-80.

Blume, L., Easley, D., & O'hara, M. (1994). Market Statistics and Technical Analysis: The Role of Volume. *The Journal of Finance, 49*(1), 153-181.

Brock, W., Lakonishok, J., & Lebaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance, 47*(5), pages 1731-1764.

Brorsen, B. W., & Irwin, S. H. (1987). Futures Funds and Price Volatility. *Review of Futures Markets, 6,* 118-138.

Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *The Review of Financial Studies, 1*(3), 195-228.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena, 35*(3), 335-356.

Cheung, Y., & Wong, C. (2000). A Survey of Market Practioners' Views on Exchange Rate Dynamics. *Journal of International Economics, 51,* 379-400.

Cheung, Y., & Chinn, M. D. (2001). Currency traders and exchange rate dynamics: a survey of the US market. *Journal of International Money and Finance, 20*(4), 439-471.

Cheung, Y., Chinn, M. D., & Marsh, I. W. (2000). How Do UK-Based Foreign Exchange Dealers Think their Market Operates? *Working Paper*, 7524, NBER.

Cochrane, J. H. (2012). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047-1180.

Cowles, A. (1960). A Revision of Previous Conclusions Regarding Stock Price Behavior. *Econometrica, 28*(4), 909-915.

Fama, E. F., & French, K. R. (1998). Value versus Growth: The International Evidence. *The Journal of Finance, 53*(6), 1975-1999.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business, 38*(1), 34-105.

Farmer J. D., & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters, 59,* 845-848.

Figlewski, S. (1981). Futures Trading and Volatility in the GNMA Market. *The Journal of Finance, 36*(2), 445-456.

Guidolin, M., Hyde, S., McMillan, D., & Ono, S. (2009). Non-linear predictability in stock and bond returns: When and where is it exploitable? *International Journal of Forecasting, 25*(2), 373-399.

Hong, H., & Yogo. M. (2012). What does futures market interest tell us about the macroeconomy and asset

prices? *Journal of Financial Economics, 105*(3), 473-490

Irwin S. H., & Brorsen, B. W. (1985). Public futures funds. *Journal of Futures Markets, 5*(2), 149-171.

Karisson, M., & Yakowitz, S. (1987). Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research, 23*(7), 1300-1308.

Kelso, J. A. S., Mandell, A. J., & Shlesinger, M. F. (1988). *Dynamic Patterns in Complex Systems*. World Scientific Press.

Kendall, M. G., & Hill, A. B. (1953). The Analysis of Economic Time-Series-Part I: Prices. *Journal of the Royal Statistical Society*. Series A (General), *116*(1), 11-34.

Kuramochi, M., & Karypis, G. (2005). Finding Frequent Patterns in a Large Sparse Graph. *Data Mining and Knowledge Discovery, 11*(3), 243-271.

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research, 32*(3), 679-693.

Lo, A. W., & MacKinlay, A. C. (1988). Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. *The Review of Financial Studies, 1*(1), 41-66.

Menkhoff, L. (1997). Examining the use of technical currency analysis, International Journal of Finance & Economics. *Special Issue: Technical Analysis and Financial Markets, 2*(4), 307-318.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1995). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association, 92*(440), 1620-1629.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence, 26*(1), 25-33.

Sharma, U. L. (1996). A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resources Research, 32*(3), 679-693.

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature, 344,* 734-741.

Tsai, C., & Hsiao, Y. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches, *Decision Support Systems, 50*(1), 258-269.

Weber, M., Camerer, C., & Lowewnstein, G. (1989). The Curse of Knowledge in Economic Settings: An Experimental Analysis. *The Journal of Political Economy, 97*(5), 1232-1254.

Yakowitz, S. & Karlsson, M. (1987). Nearest Neighbor Methods for Time Series, with Application to Rainfall/Runoff Prediction, Advances in the Statistical Sciences: Stochastic Hydrology. *The University of Western Ontario Series in Philosophy of Science, 37,* 149-160.

Appendix I K-nearest neighbors (KNN) 알고리즘과 Pseudocode

Input: D, set of K training objects, and test object $z=(x',\ y')$

Process:

1) Compute $d(x',\ x)$, the distance between z and every object, $(x,\ y)\in D$

2) Select $D_z \subseteq D$, set of K closest training objects to z

Output: $y' = \arg max_v \sum_{(x_i, y_i)\in D_z} I(v = y_i)$

Function KNN

Input: A finite set D of points to be classified

A finite set T of points

A function c: T-> {1, ···, m}

A natural number k

Output: A function r: D-> {1, ···, m}

Begin

For each x in D do

Let U <- {}

For each t in T

Add the pair (d(x, t), c(t)) to U

Sort the pairs in U using the first components

Count the class labels from the first k elements from U

Let r(x) be the class with the highest number of occurrence

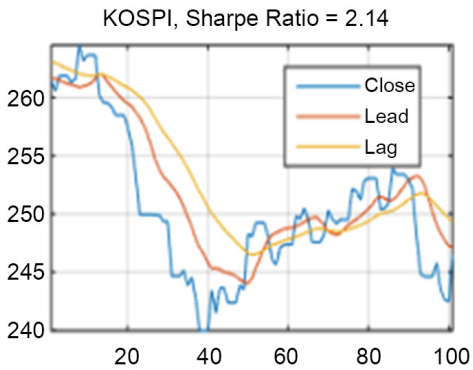
End For each

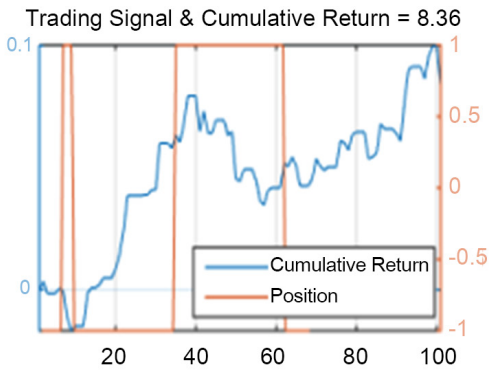
Return r

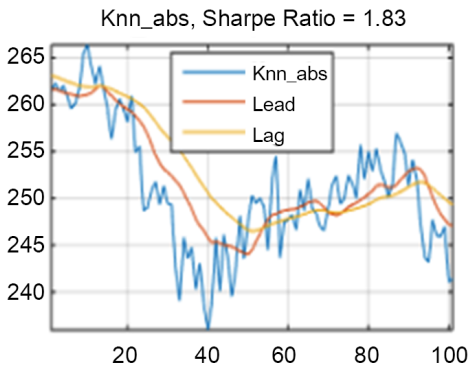
End

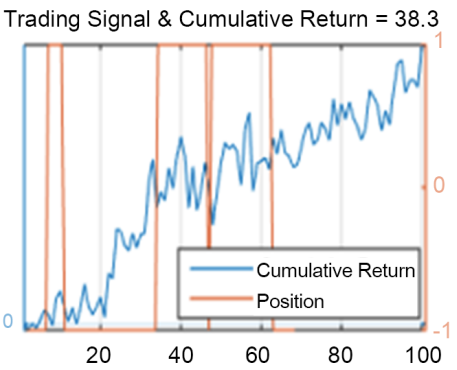
Appendix II K-NN 예측값의 이동평균 및 RSI를 이용한 거래전략 비교

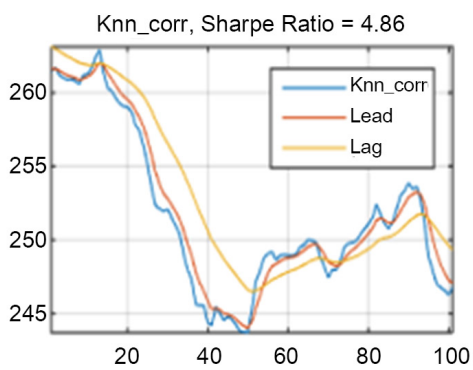
<Panel A> 이동평균 거래전략 성과

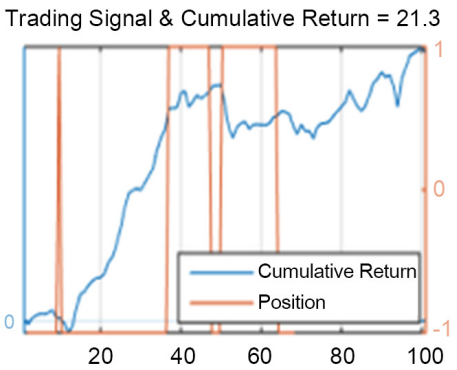




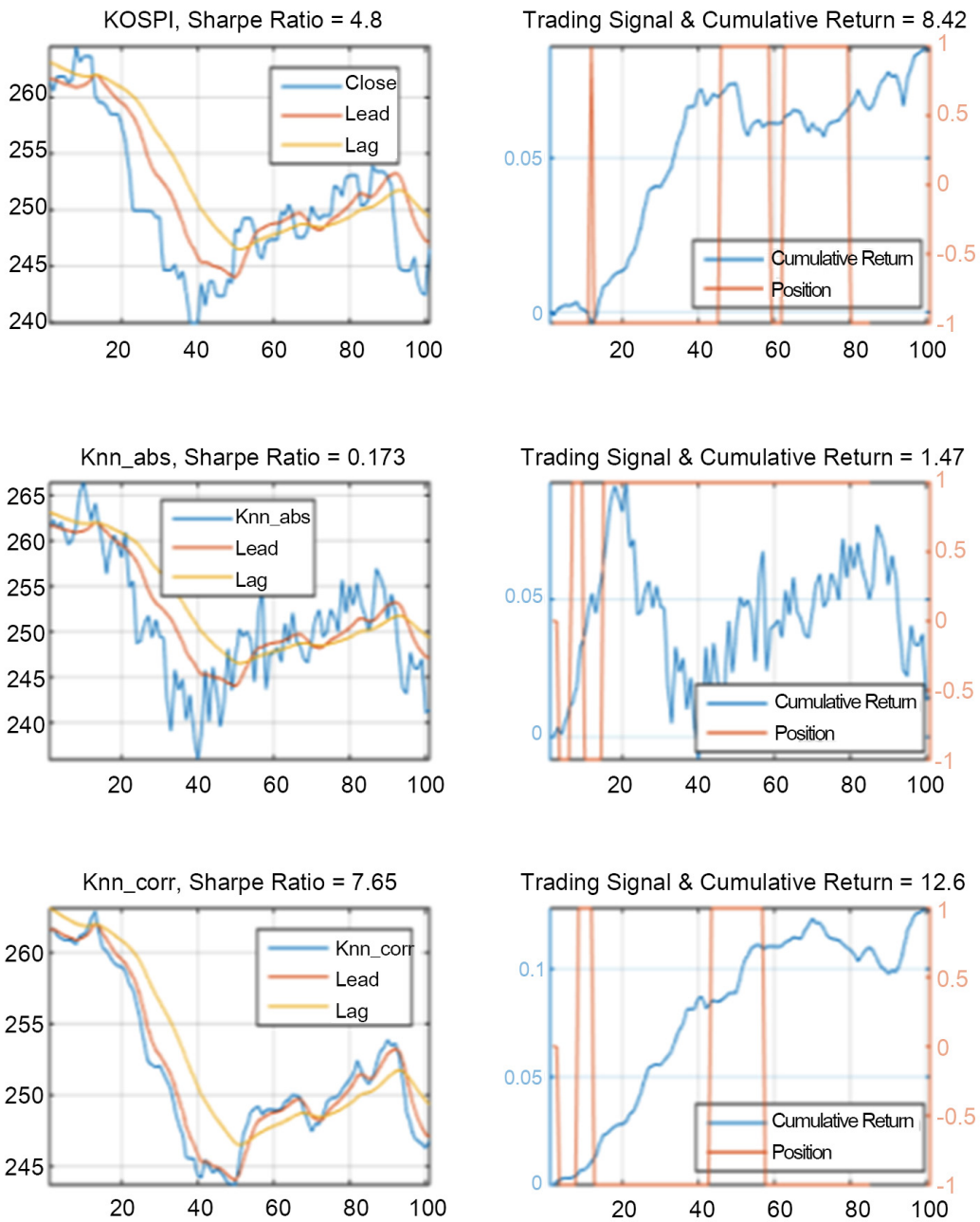








<Panel B> RSI 거래전략 성과



저자사항(author(s) note)

김명현(Myeong-Hyeon Kim)

・제1저자(First author)
・macrovue@korea.ac.kr
・고려대학교 기업지배구조연구소 연구원
・관심분야 : Systemic Risk, Asset Pricing, Financial Derivatives
・Researcher, The Asian Institute of Corporate Governance (AICG) at Korea University, Seoul, Korea

이세호(Seho Lee)

・공동저자(Co-author)
・sztp3266@gmail.com
・런던정경대 응용수학과 석사과정
・관심분야 : Financial Engineering, Asset Pricing, Derivatives
・Student in MSc program in Applicable Mathematics, Department of Mathematics, London School of Economics, London, U. K.

신동훈(Dong-Hoon Shin)

・교신저자(Corresponding author)
・dhshin@inha.ac.kr
・인하대학교 글로벌 금융학과 교수
・관심분야 : Financial Engineering, Financial Derivatives, Optimal Control Theory
・Professor, Department of Global Finance and Banking, Inha University, Incheon, Korea