ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2022 December: 28(4): 157~177 http://dx.doi.org/10.13088/jiis.2022.28.4.157

분류 알고리즘 기반 주문 불균형 정보의 단기 주가 예측 성과

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

투자자들은 증권회사가 제공하는 시세표인 Limit Order Book 정보를 통해 국내외 투자자들이 제출하는 주문 정보를 실시간으로 파악하면서 거래에 참여하고 있다. Limit Order Book에 실시간으로 공개되고 있는 주문 정보가 주가 예측에서 유용성이 있을까? 본 연구는 장 중 투자자들의 매수와 매도 주문이 어느 한쪽으로 쏠리면서 주문 불균형이 나타나는 경우 미래 주가 등락의 예측 변수로서 유의성이 있는지를 분석하는 것이다. 분류 알고리즘을 이용하여 주문 불균형 정보의 당일 종가 등락에 대한 예측 정확도를 높이고, 예측 결과를 이용한 데이트레이딩 전략을 제안하며 실증분석을 통해 투자 성과를 분석한다. 자료는 2004년 1월 19일부터 2022년 6월 30일까지의 4,564일 동안의 코스피200 주가지수선물 5 분봉 주가를 분석하였다. 실증분석 결과는 다음과 같다. 첫째, 총매수 주문량과 총매도 주문량의 불균형 정도로 측정하는 주문 불균형지수와 주가는 유의적 상관성을 보인다. 둘째, 주문 불균형 정보는 당일 종가까지의 미래 주가 등락에 대해서도 유의적인 영향력이 나타났다. 셋째, 주문 불균형 정보를 이용한 당일 종가 등락의 예측 정확도는 Support Vector Machines 알고리즘이 54.1%로 가장 높게 나타났다. 넷째, 하루 중 이른 시점에서 측정한 주문 불균형지수가 늦은 시점에서 측정한 주문 불균형지수보다 예측 정확성이 더 높았다. 다섯째, 종가 등락 예측 결과를 이용한 데이트레이딩 전략의 투자 성과는 비교모형의 투자 성과보다 높게 나타났다. 여섯째, 분류 알고리즘을 이용한 투자 성과는 K-Nearest Neighbor 알고리즘을 제외하면 모두 비교모형보다 총수익 평균이 높게 나타났다. 일곱째, Logistic Regression, Random Forest, Support Vector Machines, XGBoost 알고리즘의 예측 결과를 이용한 데이트레이딩 전략의 투자 성과는 수익성과 위험성을 동시에 평가하는 샤프비율에서도 비교모형보다 높은 결과를 보여주었다. 본 연구는 Limit Order Book 정보 중 총매수 주문량과 총매도 주문량 정보의 경제적 가치가 존재함을 밝혔다는 점에서 기존의 연구와 학술적 차별점을 갖는다. 본 연구의 실증분석 결과는 시장 참여자들에게 투자 전략적 측면에서 함의가 있다고 판단된다. 향후 연구에서는 최근 활발히 연구가 진행되고 있는 딥러닝 모형 등으로의 확장을 통해 주가 예측의 정확도를 높임으로써 데이트레이딩 투자전략의 성과를 개선할 필요가 있다.

주제어 : 시세표, 주문불균형정보, 분류알고리즘, 코스피200 지수선물, 데이트레이딩

논문접수일 : 2022년 10월 14일 논문수정일 : 2022년 11월 6일 게재확정일 : 2022년 11월 11일
원고유형 : Regular Track 교신저자 : 김선웅

1. 서론

우리나라를 포함한 대부분 주식시장의 거래시스템은 거래소로 전달되는 투자자들의 매수와 매도 주문 상황을 Limit Order Book(LOB)에 공표하여 해당 주식에 대한 투자자들의 수요(demand)와 공급(supply) 정보를 실시간으로 투명하게 공개하고 있다. 시장에서 '시세표'로 불리는 LOB는 투자자들의 신규 주문, 주문 취소, 거래 체결의 상호 작용에 따라 동적으로 변동하고 있으며

이는 다시 투자자들의 거래전략에 영향을 미칠 수 있다. 대부분의 시장 참여자들이 가장 관심 있게 관찰하고 있는 LOB 정보를 이용하면 미래 주가를 예측할 수 있을까?

Cao et al.(2009)은 호주 주식시장을 분석한 결과, 시세표 LOB에 공개되는 매수, 매도의 주문 불균형(order imbalance) 정보를 이용하면 주가의 단기 예측에서 유의적인 성과가 있음을 밝혔다. Cenesizoglu et al.(2022)은 LOB에서 매수 주문과 매도 주문의 변화 기울기를 분석한 결과, 기울기가 가파를수록 가격 변동이 커짐을 보여주었다. Kim(2019)은 한국의 코스피200 주가지수선물 시장에서 매수호가 총 잔량과 매도호가 총 잔량을 이용한 데이트레이딩의 투자 성과가 비교모형보다 높음을 보여주었다. 한편, Griffiths et al.(2000)은 토론토 증권거래소의 주가 예측에서 LOB 정보의 유의미한 정보효과를 찾지 못하였다. Kozhan and Salmon(2012)는 LOB 정보가 미래 주가 예측에서 유의적인 결과가 나타났지만, 이를 이용한 투자전략의 성과는 경제적 가치가 없다고 주장하였다. 투자자들에게 공개되고 있는 LOB 정보의 경제적 가치 관련 연구들은 혼재된 결과를 보여주고 있다.

한국거래소(Korea Exchange) 역시 투자자들이 제출하는 주식, 선물, 옵션 등의 주문 상황을 LOB 화면으로 실시간으로 공개하고 있다. 투자자들은 실시간으로 LOB 정보를 보면서 매수나 매도 주문을 거래소에 제출하고, 거래가 체결되면 가격과 주문 상황이 변동하게 된다. 오늘날은 온라인 플랫폼을 통해 빠른 속도로 전 세계 투자자들의 주문 상황을 파악하고 자신의 거래전략을 세워 적절한 주문을 거래소에 다시 제출할 수 있게 되었다.

본 연구의 목적은 세계적인 선물시장으로 발전한 코스피200 주가지수 선물시장(KOSPI200 Index Futures)에서 투자자들에게 실시간으로 공개되고 있는 LOB 정보 중 총매수 주문량과 총매도 주문량의 불균형 정보가 주가의 단기적 방향성에 대한 유의적인 예측정보효과가 존재하는지를 밝히고, 이를 이용한 투자전략의 수익성을 분석하는 것이다. 특히, 선물 거래자들은 하루에도 여러 번 거래하는 단기 거래가 일반적이기 때문에 LOB 정보는 선물 거래자들에게는 거래의 중요한 판단 기준이 된다.

본 연구는 매수와 매도 주문 수량 정보를 이용하면 돈을 벌 수 있다는 시장 참여자들의 믿음에 대한 궁금증에서 출발하였다. 특히, 한국거래소는 매수 희망자와 매도 희망자가 제출하는 주문에 의해서만 거래가 체결되는 주문자 주도의 순수 경매시장(pure auction market)이므로 LOB의 시장 영향력은 크다고 할 수 있다. 실제 대부분의 단기 거래자들은 증권회사 HTS(Home Trading System)에서 제공하는 시세표인 LOB 정보를 가장 많이 참고하면서 거래하고 있다.

2. 이론적 배경

2.1 코스피200 주가지수선물 시장

코스피200 주가지수선물은 유가증권시장에 상장된 기업 200종목의 시가총액 기준으로 산출된 코스피200 주가지수를 기초자산으로 하는 선물상품이다. 코스피200 주가지수선물 시장은 1996년 개설된 이래 낮은 거래비용, 양방향 거래, 시장참여의 용이 등으로 거래가 활발하게 이루어지며 오늘날은 세계적인 선물시장으로 발전하였다(Park et al., 2019).

Yang(2021)의 연구에 의하면 코스피200 주가지수선물 시장은 정보력에서 우위에 있는 글로벌 투자자들의 거래 참여가 확대되면서 2016년~2019년 외국인 투자자가 차지하는 거래량 비중이 전체 거래량의 65% 이상을 차지하는 것으로 나타났다. 그만큼 코스피200 주가지수선물 시장에서 외국인 투자자의 영향력은 크다고 할 수 있다.

Ryu(2013)는 코스피200 주가지수선물 시장에서 대량 주문을 제출하는 큰 손의 영향력이 큰 것으로 밝혔지만 기존의 연구 결과와 달리 매수자 주도 주문보다는 매도자 주도 주문의 가격 영향력이 더 크다고 주장하였다. Kang and Ryu(2010)는 코스피200 주가지수선물 거래에 참여하는 투자자를 외국인 투자자와 국내 투자자로 구분하여 투자 성과를 분석한 결과 외국인 투자자가 최고의 수익성을 보여주고 있음을 밝혔다. Lee and Kim(2013)은 코스피200 주가지수선물 거래의 고빈도자료와 Cont et al.(2010)의 CST 모형을 이용하여 높은 유동성을 보이는 시장에서 LOB 정보를 분석하고 안정적인 수익성이 가능함을 보여주었다.

2.2 Limit Order Book의 구조와 정보효과

스페셜리스트(specilaist)가 중간자 역할을 하는 미국 등의 증권시장과 달리 한국 증권시장은 매수와 매도 주문자가 직접 연결되는 순수 경매 시스템 방식으로 거래가 이루어지고 있다. 이러한 주문자 중심 거래시스템에서는 주가의 상승이나 하락을 예상하는 투자자들이 자신의 매수 주문이나 매도 주문을 증권회사의 홈트레이딩시스템을 통해 전달하면 한국거래소의 매매체결시스템인 LOB에 실시간으로 공개되어 전 세계 모든 시장 참여자들이 볼 수 있는 정보가 된다. Table 1은 코스피200 주가지수선물 시장의 장중 LOB 사례를 보여주고 있다.

〈Table 1〉 LOB Example

Ask	Quotation	Bid
162	291.10	
150	291.05	
155	291.00	
17	290.95	
52	290.90	
	290.85	87
	290.80	149
	290.75	44
	290.70	32
	290.65	11
6,377	-1,304	5,073

Table 1은 코스피200 주가지수선물의 장 중 특정 시점에서 투자자들의 주문 상황을 종합적으로 보여주는 LOB 사례이다. Bid 행은 매수 주문, Ask 행은 매도 주문을 나타낸다. 현재 주문 내용 중 최우선 매수 호가(best bid quotation)는 290.85, 최우선 매도 호가(best ask quotation)는 290.90이며, 각각 87계약 사자, 52계약 팔자 호가 상황이다. 최우선 호가를 포함하여 매수 주문 총수량은 5,073계약, 매도 주문 총수량은 6,377계약이다. 현재 주문 상황은 매도 주문 수량이 매수 주문 수량보다 1,304계약, 즉 1.2배 이상 많은 주문 불균형 상황이다. 총매도 주문 수량이 총매수 주문 수량보다 많은 현재의 LOB 정보를 이용하면 향후 주가의 상승이나 하락의 방향을 예측할 수 있을까?

LOB 정보는 거래가 체결되기 전의 주문 상황을 모아서 보여주고 있으며 신규 주문, 주문의 취소, 주문의 거래 체결 등에 따라 동적으로 변

동한다. 이러한 LOB 정보 관련 연구는 방대한 양의 데이터 처리가 필요하므로 컴퓨팅 능력의 확대에 따라 연구자들의 많은 관심을 받고 있다. Harris and Panchapagesan(2005)는 매수와 매도호가 불균형 정보가 미래 주가 예측에서 정보효과가 존재한다고 주장하였다. 주문이 한 쪽 방향으로 쏠리면서 불균형이 나타나면 투자자들은 이익을 극대화하기 위해 주문이 많은 방향으로 시장가 주문을 통해 거래를 체결하려는 수요가 커질 것이다. 이는 결국 시장의 주문 불균형 정보가 미래 주가의 상승과 하락에 대한 정보효과가 있음을 밝혔다. 그러나 미국 주식시장의 미시구조(micro structure)에서 존재하는 스페셜리스트들의 행동을 중심으로 분석한 연구로서 우리나라의 순수 경매시장의 미시구조와는 다른 환경이기 때문에 직접 비교하기에는 무리가 있다. Cao et al.(2009)은 호주의 주식시장에서 LOB에 공개되는 주문 정보의 정보효과가 존재함을 밝혔으나, 정보효과가 아주 짧은 시간 동안만 나타나고 있어 전문 투자자가 아니면 활용 가치가 없다. LOB의 정보효과에 동의하고 있는 연구들은 대부분 짧은 시간 동안의 가격 예측력이 존재한다고 밝히고 있지만, 이 정보를 이용한 투자전략의 수익성 분석까지는 다루지 않았다.

Kozhan and Salmon(2012)은 외환시장에서 달러와 스털링의 주문 정보를 이용하여 환율 거래전략을 제안하고 거래전략의 수익성을 실증 분석한 결과, 유의미한 수익성이 없음을 밝혔다. 구체적으로, 2003년 데이터에서는 투자전략의 수익성을 찾았지만 2008년 데이터에서는 수익이 대부분 사라지고 있음을 보여주었다. 2003년 이후 증가하고 있는 빅데이터를 이용한 다양한 거래 알고리즘으로 인해 시장의 효율성이 강화되면서 실시간으로 공개되고 있는 주문 정보만 가지고는 가격 예측력이 없다고 주장하였다.

미국 주식시장에서 빅데이터를 기반으로 하는 알고리즘 거래량은 전체 시장 거래량의 상당 부분을 차지할 정도로 큰 편이지만, 한국 주식시장에서는 아직 까지는 그 비중이 높지 않기 때문에 주문 불균형 정보의 가격 예측력에 대한 분석은 의의가 크다고 판단된다. 최근에 와서 국내에서도 LOB 정보 관련 연구가 활발히 진행되고 있다. Lee and Choi(2007)은 주문 정보의 호가 단계 중 매수와 매도의 5호가 까지는 유의적인 예측력이 나타났지만 10분 후에는 예측력이 사라지는 것으로 밝혀져 정보효과가 초단기적임을 보여주었다. Han(2017)의 연구에서는 외국인 투자자의 공매도(short-selling) 거래가 주가에 미치는 영향을 분석하면서 주문 체결의 불균형 정보를 이용하고 있지만, LOB 정보의 Bid와 Ask 정보를 분석하지는 않았다. Kim(2019)은 국내 주식시장에서 호가 잔량의 호가 범위를 매수호가 총잔량과 매도호가 총잔량으로 확대하고 통계적 투자전략을 제안하고 실증분석을 통해 수익률을 분석한 결과 비교모형보다 우수한 투자 성과가 나타남을 보여주었다.

2.3 분류 알고리즘과 주가 등락 예측

불규칙하고 비선형적으로 움직이는 주가와 같은 시계열 자료의 예측 연구에서 인공지능 기법은 전통적인 통계적 기법과 비교하여 좋은 예측 성과를 보여주고 있다. 초기의 신경망 모형을 이용한 주가 예측에서 시작된 인공지능 모형의 주가 예측 활용은 Cao and Tay(2001)에 와서 Support Vector Machines(SVM) 모형을 이용하여 미국 주가지수를 예측한 결과 신경망 모형의 결과보다 우수한 성과를 보여주면서 다양한 기계학습 모

형들이 활용되기 시작하였다.

최근 주가 연구에서는 고빈도 자료(high frequency data)의 다양화와 컴퓨팅 능력의 향상으로 우수한 예측 성능을 보이는 분류 알고리즘의 활용 연구가 활발히 진행되고 있다. 대표적인 분류 알고리즘에는 Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Machines, XGBoost 모형 등이 포함된다.

Logistic Regression(LOGIT) 모형은 선형회귀모형과 유사한 모형이지만 종속변수가 Up, Down과 같은 범주형 자료를 대상으로 하는 통계적 분류 모형이다. K-Nearest Neighbor(KNN) 알고리즘은 사례에 기반한 단순한 학습 형태의 기계학습모형으로서 새로운 자료가 주어 지면 그 주변의 k개의 자료에 가장 많이 포함된 범주로 분류하는 대표적인 분류 모형이다. 하이퍼파라미터는 탐색할 이웃의 수 k와 거리 측정 방법이다. k가 작으면 과대 적합되는 경향이 있고 반대로 k가 크면 과소 적합되는경향이 있다. Random Forest(RF) 알고리즘은 앙상블 학습 모형(ensemble learning model)으로서 여러 개의 의사결정수(decision tree)를 형성하고 새로운 자료를 각 의사결정수에 통과시켜 각 결정수에서 분류한 결과를 최종 분류 결과로 산출한다. SVM 알고리즘은 분류를 위한 기준선인 decision boundary를 결정하는 모형으로서 새로운 자료가 주어지면 이 점이 어느 경계면에 속하는지를 판단하는 분류 모형이다(Song et al., 2017).

XGBoost(XGB)는 앙상블의 부스팅 기법으로서 이전 모형의 오류를 보완해나가는 방법으로 모형을 생성하는 분류기법이다(Kim et al., 2020).

Lohrmann and Luukka(2019)는 Random Forest 알고리즘을 이용하여 미국 주가지수의 장 중 가격 변동을 예측하고 투자전략을 통해 비교 전략보다 우수한 투자 성과가 나타남을 보여주었다. Zhang et al.(2017)은 KNN 모형을 이용하여 미국 주가지수 예측에 적용한 결과 주가의 단기 예측에서 좋은 성과를 보여주었다. Malagrino et al.(2018)은 브라질 주식시장에서 베이즈 모형의 주가 예측 성과가 우수하게 나타남을 밝혔다. Zhang and Lou(2021)는 Backpropagation 알고리즘을 이용하여 중국 주식시장의 주가를 예측한 결과, 73.29%의 예측 정확도를 보여 비교모형인 딥러닝 기반 fuzzy 알고리즘의 예측 정확도 62.12%보다 높은 결과를 보여주었다.

분류 알고리즘을 결합하여 예측 성과를 높이려는 시도도 이루어지고 있다. Cao et al.(2019)과 Chen and Hao(2017)은 KNN과 SVM 모형을 결합하여 주가의 예측 성과를 높이고 있으며, Weng et al.(2017)은 Artificial Neural Network과 SVM의 결합을 통해 주가의 예측 성과를 높였다. Thakkar and Chaudhari(2022)는 유전자 알고리즘(genetic algorithm) 최적화기반 LSTM 모형을 이용한 코스피 주가지수와 6개의 대형주 주가를 예측한 결과 설명력이 61% 이상 증가함을 보여주었다. Yun et a.(2021)은 유전자 알고리즘과 XGBoost를 결합한 하이브리드 예측모형을 제안하고, 코스피200 주가지수의 일별 종가를 예측한 결과 정확도가 93.28%까지 증가함을 보여주었으며, 특히 특성 변수 중요도를 기준으로 71개의 전체 특성 변수 중 33개만 선택하여 예측한 결과도 정확도가 93.15%로 나타나 차원의 저주(curse of dimensionality)를 피하면서도 우수한 예측 결과를 얻을 수 있었다.

기계학습모형을 이용한 주가 예측 연구는 대부분 입력변수로 주가나 기술적 지표 또는 경제변수 등이 활용되고 있다(Kim and Ahn, 2010; Kumbure et al., 2022). 기계학습모형을 활용한 주가

예측 관련 138개의 연구 논문을 분석한 Kumbure et al.(2022)에 의하면 2,173개의 특성 변수 중 기술적 지표 활용이 1,208개로 전체의 55.6%를 차지하고 있으며 경제 변수는 전체의 12.8%를 차지하여 두 변수가 대부분을 차지하고 있음을 확인하였다. 특히, 본 연구에서 분석대상으로 하는 LOB 주문 정보 중 매수 호가 총수량과 매도 호가 총수량의 정보를 이용한 분류 알고리즘 기반 주가 예측 연구는 찾기가 어렵다(Kim, 2019).

3. 자료와 주문 불균형지수

3.1 자료 소개

본 연구에서는 코스피200 주가지수선물 시장을 분석대상으로 한다. 코스피200 선물거래는 매수와 매도의 양방향 거래가 가능한 상품으로서 짧은 만기, 낮은 증거금률, 풍부한 유동성 등으로 인해 하루 중 장 마감 시점에서 보유 포지션을 모두 청산하는 데이트레이딩과 같은 단기 거래가 대부분을 차지하고 있다. 코스피200 선물거래에는 개인, 기관, 외국인 등 다양한 투자자들이 거래에 참여하고 있으며, 자본력과 경험이 풍부한 외국인 투자자의 주가 영향력이 큰 것으로 밝혀지고 있다(Kim and Ok, 2015).

본 연구에서는 LOB의 주문 정보를 이용한 단기적 주가 변동 예측을 위해 장 시작 초반 시점의 코스피200 선물 정보를 수집하였다. 구체적으로, 9시 시가부터 9시 20분까지의 5분 간격 가격, 해당 시점의 매수 주문 총수량, 매도 주문 총수량, 해당 시점의 외국인 순매수 수량, 해당 시점의 총거래량, 그리고 해당 시점부터 장 마감 종가까지의 주가 변동 폭이다. 분석 자료는 증권회사 홈트레이딩시스템인 YesTrader 4.0에서 구하였다. 분석 기간은 LOB 주문 정보와 외국인 거래량 정보를 모두 구할 수 있는 최초일인 2004년 1월 19일부터 2022년 6월 30일까지의 4,564일이다.

3.2 주문 불균형지수

실시간 LOB 정보를 파악하면 투자자들의 매수 주문이나 매도 주문 상황을 파악할 수 있다. Table 1의 사례에서 현재 시점의 코스피200 주가지수선물에 대한 주문 정보는 매도 주문이 매수 주문보다 1.2배 이상 높은 상황이다. 일반적으로 주가가 상승하기 시작하면 추가적인 주가 상승을 예상하는 투자자는 매수 기회를 놓치지 않기 위해 공격적으로 신규 매수 주문을 제출할 것이며, 매도 관점을 가지고 있는 반대편 투자자들은 역선택 위험을 회피하기 위해 매도 주문 제출을 망설이거나 제출된 매도 주문을 취소함에 따라 매수 주문 총수량은 증가하고 매도 주문 총수량은 감소하는 현상을 보일 것이다(Ranaldo, 2004; Wang et al., 2008; Kim, 2019).

시장의 주문 불균형 상황을 지수화하기 위하여 Eq. (1)과 같은 주문 불균형지수 OIB(order imbalance index)를 제안한다.

$$OIB_t = \frac{bids_t - asks_t}{bids_t + asks_t}, \quad (1)$$

$where\ OIB_t\ is\ Order\ Imbalance\ Index\ at\ time\ t,$
$bids_t$ and $asks_t$ are

OIB_t는 -1부터 +1 사이의 값을 가지며, 양수는 매수 총수량이 매도 총수량보다 많은 상황, 음수는 매도 총수량이 매수 총수량보다 많은 주문의 불균형 상황을 말한다. 특정 시점 t에서 OIB를 이용하여 당일 종가까지의 주가의 상승이나 하

락을 예상할 수 있을 것인가? 예상할 수 있다면 주문 불균형지수를 이용한 거래전략은 경제적 가치가 있는 수익성을 보일 것인가? 본 논문의 목적은 실제 주가 자료의 분석을 통해 이러한 물음에 대한 답을 얻는 것이다.

코스피200 주가지수선물 시장에서 자금력과 정보력의 우위를 가진 외국인 투자자들의 시장 참여 강도를 측정하기 위해서 Eq. (2)와 같은 외국인 거래량 비중 지수(Foreigners' Trading Ratio: FTR)를 구한다.

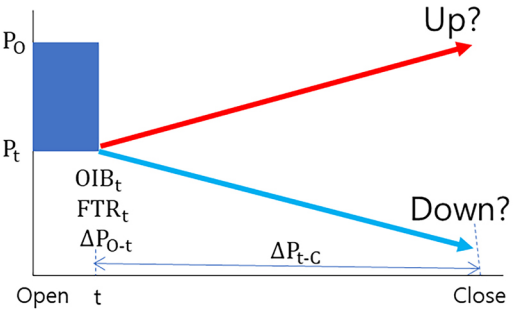
$$FTR_t = \frac{buy\ volume_t - sell\ volume_t}{} \qquad (2)$$

Eq. (2)는 아침 9시 시작 시점에서 특정 t 시점까지의 시장 전체 거래량 대비 외국인 투자자의 상대적 거래량 비중을 측정하는 지수로서, 이 값이 크다면 외국인의 집중 매수가 이루어졌다는 의미이고, 반대로 큰 음의 값을 갖는다면 이 시간 동안 외국인의 집중적인 매도가 이루어졌다는 의미이다.

4. 분류 알고리즘을 이용한 주가 예측

4.1 주문 불균형과 주가 변동

Figure 1은 주식시장이 9시에 개장(시가)하여 t 시점까지 관찰된 시장 상황의 변동과 t 시점부터 코스피200 주가지수선물의 종가 시간인 15시 45분까지의 주가 변동을 보여주고 있다.



〈Figure 1〉 Order Imbalance and Intra-day Price Change

Figure 1에서 t 시점이 9시 5분이라면, OIB_t는 9시 5분에 관찰된 주문 불균형지수, FTR_t는 9시 5분에 관찰된 코스피200 주가지수선물 총거래량 대비 외국인 순매수(총매수량-총매도량) 거래량 비율, $\triangle P_{O-t}$는 9시 시가 대비 9시 5분 주가 변동량을 측정하며, $\triangle P_{t-C}$는 9시 5분부터 15시 45분 종가까지의 주가 변동을 측정한다. 9시 5분 대비 당일 코스피200 주가지수선물 종가가 더 높게 끝나면 Up, 종가가 더 낮게 끝나면 Down으로 분류한다. 구체적으로, 시점 t에서의 Up, Down의 정의는 Eq. (3)과 같다.

$$\begin{aligned} &If\ \triangle P_{t-C} > 0\ then\ market\ trend = Up \\ &If\ \triangle P_{t-C} \leq 0\ then\ market\ trend = Down \end{aligned} \qquad (3)$$

$$where\ \triangle P_{t-C}\ is\ price\ change\ from\ t\ to\ market\ close(C).$$

Table 2는 자료의 전체 기간에서 시점 t에서 관찰한 주문 불균형지수, 외국인 투자자 거래량 비중, 9시에서 t 시점까지의 주가 변동 폭을 보여주고 있다.

〈Table 2〉 Market Variables Observed at Time t

Vars	Time	9:05	9:10	9:15	9:20
OIB_t	Avg	0.52	0.71	0.86	0.91
	Min	-70.40	-59.70	-61.40	-61.2
	Max	60.5	66.4	61.0	82.8
FTR_t	Avg	-0.35	-0.28	-0.24	-0.21
	Min	-85.8	-51.3	-41.3	-34.3
	Max	30.9	24.6	20.5	19.9
$\triangle P_{O-t}$	Avg	-0.02	-0.01	-0.01	-0.00
	Min	-3.7	-6.3	-5.6	-6.35
	Max	4.0	7.85	5.75	6.1

주문 불균형과 주가 변동 사이에 유의적인 상관관계가 있는지를 판단하기 위하여 시점 t에서 Eq. (4)의 회귀분석 모형을 통해 분석한다.

$$\triangle P_{O-t} = \alpha_t + \beta_t OIB_t + \gamma_t FTR_t + \epsilon_t \qquad (4)$$

시점 t가 9시 5분인 경우, Eq. (4)는 시가부터 9시 5분까지의 주가 변동이 9시 5분에서 관찰된 주문 불균형지수와 외국인 투자자 거래량 비중에 영향을 받는지를 분석하며, Table 3은 분석 결과를 요약하여 보여주고 있다.

〈Table 3〉 Estimations on Eq. (4)

Estimators	9:05	9:10	9:15	9:20
α	-0.023 (-3.09)***	-0.028 (-3.17)***	-0.035 (-3.74)***	-0.035 (-3.52)***
β	0.026 (51.42)***	0.035 (59.99)***	0.039 (63.25)***	0.042 (66.41)***
γ	0.017 (16.05)***	0.027 (17.55)***	0.034 (18.66)***	0.041 (19.43)***

*** : *significant at* 1%

Table 3에서 주문 불균형지수와 외국인 투자자 거래량 비중의 영향력에 대한 회귀분석 결과는 유의적인 관계를 잘 보여주고 있다. 주문 불균형지수와 외국인 투자자 거래량 비중 변수는 모두 주가 변동에 양의 영향력을 미치고 있으며, 특히 주문 불균형지수의 영향력은 매우 높게 나타나고 있다. 코스피200 주가지수선물 가격 결정에서 외국인 투자자의 시장 참여도와 주문 불균형 정도는 주가의 방향과 강한 양의 관계가 있음을 알 수 있다. 그렇다면 특정 시점 t에서 관찰된 주문 불균형지수는 t 시점부터 당일의 15시 45분 종가까지의 단기적 주가의 방향을 예측할 수 있을까?

이번에는 시점 t에서 관찰된 주문 불균형 정보, 외국인 투자자 거래량 비중, 시점 t까지의 주가 변동 폭을 알고 있을 때 시점 t에서 당일 종가까지의 미래 주가 변동 방향에 대한 영향력을 분석하기 위하여 Eq. (5)의 회귀모형을 이용한다.

$$\triangle P_{t-C} = \alpha_t + \beta_t OIB_t + \gamma_t FTR_t + \delta \triangle P_{O-t} + \epsilon_t \qquad (5)$$

Eq. (5)는 t 시점에서 관찰된 주문 불균형지수, 외국인 투자자 거래량 비중, 최근 주가 변동 정보의 t 시점부터 당일 종가까지의 주가 변동 $\triangle P_{t-C}$에 대한 영향력을 분석하고 있다. Table 4는 Eq. (5)에 대한 추정 결과를 보여주고 있다.

〈Table 4〉 Estimations on Eq. (5)

Estimates	9:05	9:10	9:15	9:20
α	-0.029 (-0.85)	-0.036 (-1.09)	-0.036 (-1.08)	-0.041 (-1.27)
β	0.012 (4.04)***	0.011 (3.65)***	0.009 (2.85)***	0.009 (3.11)***
γ	-0.004 (-0.74)	-0.009 (-1.58)	-0.008 (-1.22)	-0.007 (-0.93)
δ	-0.055 (-0.82)	-0.000 (-0.01)	0.011 (0.22)	-0.012 (-0.24)

*** : *significant at* 1%

Table 4에서 주문 불균형지수의 단기 주가 예측력은 여전히 양의 영향력이 나타나고 있으나, 외국인 투자자 거래량 비중과 주가 변동 폭 정보는 당일의 종가 예측에서 반대로 음의 영향력이 나타나고 있다. 특히, 주문 불균형지수는 종가 예측에서도 강한 유의성이 존재하지만 외국인 투자자 거래량 비중과 단기 주가 변동 폭 정보는 통계적 유의성이 없다. 외국인 투자자 거래량 비중의 영향력이 Table 3의 결과와 반대로 나타남에 따라 코스피200 주가지수 선물가격은 장 초반 외국인 투자자 거래 정보에 과잉반응한 후 다시 균형 가격으로 회귀하면서 반전되는 것으로 판단된다.

4.2 분류 알고리즘을 이용한 주가 등락 예측

4.2.1 실험 설계

본 연구에서는 대표적 분류 알고리즘인 Logistic Regression(LOGIT), K-Nearest Neighbor(KNN), Random Forest(RF), Support Vector Machines(SVM), XGBoost(XGB) 등을 이용하여 특정 시점 t에서 당일 코스피200 주가지수선물 시장 마감 종가까지의 선물가격의 추세에 대한 Up과 Down을 예측하고자 한다.

코스피200 주가지수선물의 당일의 종가 방향 예측을 위한 분류 알고리즘 학습을 위한 입력변수(input variables)로는 주문 불균형지수, 외국인 투자자 거래량 비중, 주가 변동 폭이며, 예측 대상 변수인 출력변수(target variable)는 특정 시점 t에서 코스피200 주가지수선물 시장 마감 종가까지의 코스피200 주가지수선물 가격의 Up과 Down 분류이다.

비교모형은 주문 불균형 정보를 이용하지 않는 오버나이트 퍼즐 전략(overnight puzzle strategy)을 제안한다. 오버나이트 퍼즐 현상은 주식시장이 오후에 폐장한 후 야간 시간 동안 거래가 발생하지 않다가 다음 날 오전 다시 거래가 시작될 때 주가가 과잉반응함에 따라 장 중에는 주가가 균형 가격으로 환원되면서 시가 대비 당일 종가가 하락하는 현상을 말한다(Berkman et al. 2012). 이를 이용한 적절한 전략은 오전 시점 t에 코스피200 주가지수선물을 매도한 후 당일 종가에 포지션을 청산하는 데이트레이딩 전략이며 이를 비교 전략으로 하여 분석한다.

자료를 학습 자료와 예측 자료로 구분하기 위하여 2004년 1월 19일부터 2016년 12월 2일까지의 기간을 학습 기간, 2016년 12월 5일부터 2022년 6월 30일까지의 기간을 테스트 기간으로 구분하였다. 테스트 기간은 전체 기간의 30%에 해당하며, 학습 기간에서 Up과 Down 목표 변수의 어느 한쪽의 과부족이 발생하면 무작위 추출을 통해 Up과 Down의 발생 빈도가 같게 만든 후 학습을 진행하였다.

제안된 분류 알고리즘에 대한 하이퍼파라미터(hyper parameters)는 과 최적화 가능성을 회피하기 위해 파라미터의 튜닝 과정 없이 각 분류 알고리즘의 기본값을 채택하였다. 학습 자료를 이용하여 알고리즘을 학습한 후에는 테스트 자료를 학습된 모형에 적용하여 당일 종가의 Up과 Down의 예측 결과를 산출하였다.

예측 결과의 성과 평가지표는 예측의 정확성 지표 Accuracy를 산출하였다. 정확성 지표는 예측 결과에 대한 confusion matrix를 이용하여 다음 Eq. (6)과 같이 산출한다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (6)$$
$$where\ TP,\ TN,\ FP,\ FN\ mean\ True\ Positive,\ True\ Negative,\ False\ Positive,\ False\ Negative.$$

Accuracy는 전체 예측 사례에 대한 올바른 예측 사례의 비율을 측정하는 지표로서 주가의 방향성을 예측하는 본 연구의 경우 중요한 예측의 정확도를 판단할 수 있다.

4.2.2 실험 결과

주문 불균형지수를 이용한 분류 알고리즘의 당일 종가에 대한 예측 성과는 시점 t 별로 Table 5에 제시하였다.

〈Table 5〉 Accuracy on Test Data Prediction

Time	9:05	9:10	9:15	9:20
LOGIT	0.528	**0.528**	**0.520**	**0.527**
KNN	0.533	0.512	0.501	0.500
RF	0.501	0.519	0.517	0.497
SVM	**0.541**	0.518	0.512	0.519
XGB	0.520	0.527	0.518	0.520
Average	0.525	0.521	0.514	0.513

* Boldface marks the highest values.

Table 5에서 보면 전체 결과에서는 9시 5분 시점의 SVM 기반 당일 종가 예측 모형의 정확도가 54.1%로 가장 높게 나타났고, 9시 10분, 9시 15분, 9시 20분에서는 LOGIT 기반 당일 종가 예측 모형의 정확도가 다른 모형보다 높게 나타났다. 분류 알고리즘 전체의 평균 예측 정확도는 9시 5분에서 52.5%로 가장 높게 나타났고, 대부분의 분류 알고리즘에서 장 초반일수록 예측력이 더 높은 것으로 나타났다. 이는 데이트레이딩을 위한 하루 중의 주가 예측에서는 빠른 진입 시점이 실제 투자에서 유리한 것으로 알려진 시장 참여자들의 투자 아이디어와도 맞아떨어지는 결과이다.

5. 예측 결과를 이용한 데이트레이딩 전략

5.1 데이트레이딩 전략의 설계

각 시점 t에서 관찰된 주문 불균형지수, 외국인 투자자 거래량 비중, 주가 변동을 입력변수로 하는 분류 알고리즘의 예측 결과는 시점 t에서 당일 종가까지의 코스피200 주가지수선물의 가격 변동에 대한 Up과 Down의 분류로 나타난다. 분류 결과가 Up이라면 시점 t에서 코스피200 주가지수선물 1계약을 매수하여 당일 종가까지 보유 후 종가에 포지션을 청산한다. 분류 결과가 반대로 Down으로 나타나면 시점 t에서 매도 포지션을 진입한 후 당일 종가에 포지션을 청산한다. 이러한 데이트레이딩 전략은 장 마감 동시호가에 포지션을 청산하고 no position으로 야간 시간을 지나기 때문에 밤사이 큰 호재나 악재 발생에 따른 다음 날의 시가 갭(opening gap)의 위험을 부담하지 않는 비교적 위험성이 적은 투자전략이다.

제안된 데이트레이딩 전략 식은 다음 Eq. (7)과 같다.

$$\begin{aligned} &Day\ Trading\ Strategy(DTS) \\ &If\ \text{Predicted Output}_t = Up\ then\ Buy\ at\ P_t; \\ &If\ \text{Predicted Output}_t = Down\ then\ Sell\ at\ P_t; \\ &Exit\ Position\ at\ Market\ Close; \\ &\quad where\ \text{Predicted Output}_t\ is\ Classification \\ &\quad Algorithm\ Result\ at\ t,\ P_t\ is\ KOSPI200 \\ &\quad Futures\ \text{Price at } t,\ t\ is\ 9:05, 9:10, 9:15, 9:20. \end{aligned} \tag{7}$$

시점 t가 9시 5분이고 예측 결과가 Up이라면 9시 5분에 코스피200 주가지수선물 매수 포지션을 진입하고 보유 후 코스피200 주가지수선물 거래 마감 시간인 오후 3시 45분에 매도 주문을 실행하여 보유 포지션을 청산한다. 이 경우 포지

션의 총 보유시간은 6시간 40분이며, 시점 t가 9시 20분이라면 포지션의 총 보유시간은 6시간 25분이다. 코스피200 주가지수선물의 만기일인 3월, 6월, 9월, 12월의 두 번째 목요일의 경우는 코스피200 주가지수선물의 마지막 거래시간인 오후 3시 20분에 보유 포지션을 청산하는 것으로 설정하였다.

본 연구에서 제안된 데이트레이딩 전략 DTS의 성과 평가는 2016년 12월 5일부터 2022년 6월 30일까지의 1,369일 동안의 테스트 기간의 거래 결과를 대상으로 한다. 성과 평가지표는 총 손익(Total Profit: TP), 누적 최대 손실 폭(Maximum Draw-Down: MDD), 그리고 샤프비율(Sharpe Ratio: SR)이다.

각각의 성과 평가지표는 Eq. (8), Eq. (9), Eq. (10)과 같이 계산하여 산출한다.

$$TP = \sum_{n=2016.12.05}^{2022.06.30} PL_n, \quad \text{where } TP \text{ is} \tag{8}$$

$$MDD = \text{maximum of } draw-downs, \quad \text{where } draw-down \text{ is a } peak-to-trough \text{ decline during a specific period for a trading account.} \tag{9}$$

$$SR = \frac{Average\ Profit\ per\ Trade}{Standard\ Deviation\ of\ Profit\ per\ Trade}, \quad SR \text{ is Sharpe Ratio.} \tag{10}$$

Eq. (8)의 총손익(TP)은 2016년 12월 5일부터 2022년 6월 30일까지의 거래 기간에서 분류 알고리즘의 예측 결과를 이용한 데이트레이딩 전략 DTS의 누적 수익을 산출한다. TP가 크다면 해당 기간에서 높은 수익이 발생했음을 의미한다. MDD는 해당 기간의 거래에서 누적 수익이 줄어드는 경우의 draw-down 중 최대 크기를 측정하며, 거래전략의 투자 위험 지표로 인식된다. MDD가 크다면 투자자가 큰 스트레스에 직면하므로 해당 전략을 계속해서 유지하기가 어려워진다. SR는 Sharpe(1966)가 제안한 지표이다. 이 지표는 위험과 수익의 상충관계(risk-return trade-off)로부터 위험 대비 수익의 상대적 크기를 측정하는 지표로서 투자 성과의 통합 평가지표이다(Kim, 2022).

5.2 데이트레이딩 전략의 성과 분석

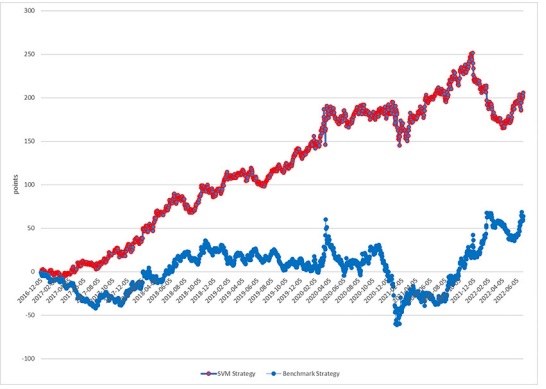
분류 알고리즘 기반 주문 불균형지수를 이용한 코스피200 주가지수선물의 종가 등락 예측을 이용한 데이트레이딩 전략의 투자 성과를 실증분석하였다.

Table 6과 Figure 2는 9시 5분 시점에서의 분류 알고리즘에 따른 데이트레이딩 전략의 성과 요약과 가장 높은 총수익을 가져오는 분류 알고리즘 데이트레이딩 전략의 수익 곡선(equity curve)을 보여주고 있다. 비교를 위해 비교모형 전략의 수익 곡선도 Figure 2에 표시하였다.

〈Table 6〉 Trading Performance (9:05)

Algorithm	TP	MDD	SR
LOGIT	190.3	74.95	0.048
KNN	225.8	**44.55**	0.057
RF	81.7	69.15	0.021
SVM	**296.1**	74.55	**0.075**
XGB	222.1	74.45	0.056
BM	64.0	121.15	0.016

* Boldface marks the best values.



〈Figure 2〉 Equity Curve of SVM Strategy (9:05)

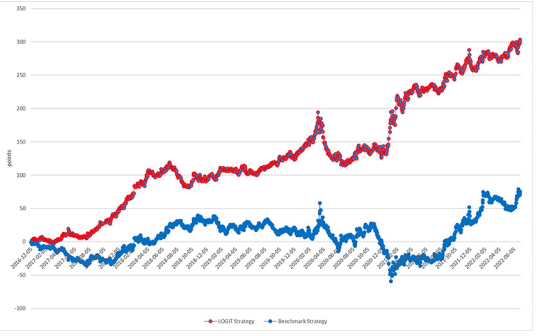
9시 5분 시점에서 관찰된 주문 불균형지수, 외국인 투자자 거래량 비중, 주가 변동 정보를 입력변수로 하는 당일의 종가 등락 예측을 이용한 데이트레이딩 전략의 수익성은 SVM 알고리즘이 296.1포인트로 가장 높은 총수익을 보여주고 있다. 투자 스트레스의 정도를 측정하는 MDD는 KNN 알고리즘이 44.55포인트로 가장 낮았고, 투자 위험 대비 수익의 상대적 크기를 측정하는 샤프비율(SR)은 SVM 알고리즘이 0.075로 가장 높은 비율을 보여주었다.

Table 7과 Figure 3는 9시 10분 시점에서의 분류 알고리즘에 따른 데이트레이딩 전략의 성과 요약과 가장 높은 총수익을 가져오는 분류 알고리즘 데이트레이딩 전략의 수익 곡선(equity curve)을 보여주고 있다. 비교를 위해 비교모형 전략의 수익 곡선도 Figure 3에 표시하였다.

〈Table 7〉 Trading Performance (9:10)

Algorithm	TP	MDD	SR
LOGIT	**303.65**	80.05	**0.080**
KNN	-14.35	92.05	-0.004
RF	85.55	79.0	0.022
SVM	238.05	87.85	0.062
XGB	262.95	**78.2**	0.069
BM	75.05	117.1	0.020

* Boldface marks the best values.



〈Figure 3〉 Equity Curve of LOGIT Strategy (9:10)

9시 10분 시점에서 관찰된 주문 불균형지수, 외국인 투자자 거래 비중, 주가 변동 정보를 입력변수로 하는 당일의 종가 등락 예측을 이용한 데이트레이딩 전략의 수익성은 LOGIT 알고리즘이 303.65포인트로 가장 높은 총수익을 보여주고 있다. 투자 스트레스의 정도를 측정하는 MDD는 XGB 알고리즘이 78.2포인트로 가장 낮게 나왔고, 투자 위험 대비 수익의 상대적 크기를 측정하는 샤프비율(SR)은 LOGIT 알고리즘이 0.080으로 가장 높은 비율을 보여주었다.

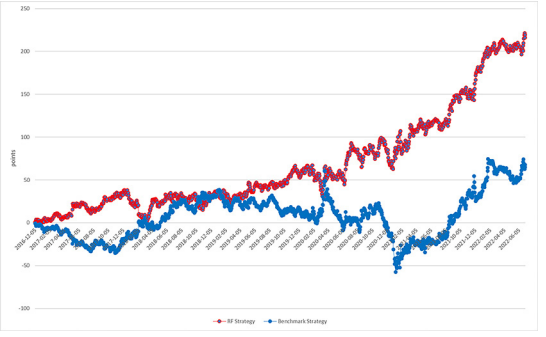
Table 8과 Figure 4는 9시 15분 시점에서의 분류 알고리즘에 따른 데이트레이딩 전략의 성과 요약과 가장 높은 총수익을 가져오는 분류 알고리

즘 데이트레이딩 전략의 수익 곡선(equity curve)을 보여주고 있다. 비교를 위해 비교모형 전략의 수익 곡선도 Figure 4에 표시하였다.

〈Table 8〉 Trading Performance (9:15)

Algorithm	TP	MDD	SR
LOGIT	112.55	69.25	0.030
KNN	40.45	86.1	0.011
RF	**216.15**	**36.35**	**0.058**
SVM	115.35	84.8	0.031
XGB	189.65	56.1	0.051
BM	67.75	117.8	0.018

* Boldface marks the best values.



〈Figure 4〉 Equity Curve of RF Strategy (9:15)

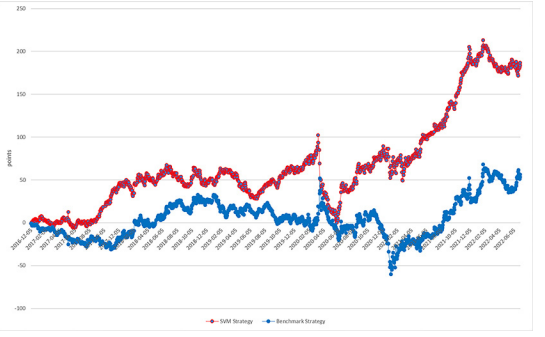
9시 15분 시점에서 관찰된 주문 불균형지수, 외국인 투자자 거래 비중, 주가 변동 정보를 입력변수로 하는 당일의 종가 등락 예측을 이용한 데이트레이딩 전략의 수익성은 RF 알고리즘이 216.15포인트로 가장 높은 총수익을 보여주고 있다. 투자 스트레스의 정도를 측정하는 MDD는 RF 알고리즘이 36.35포인트로 가장 낮게 나왔고, 투자 위험 대비 수익의 상대적 크기를 측정하는 샤프비율(SR) 역시 RF 알고리즘이 0.058로 높은 비율을 보여주었다.

Table 9와 Figure 5는 9시 20분 시점에서의 분류 알고리즘에 따른 데이트레이딩 전략의 성과 요약과 가장 높은 총수익을 가져오는 분류 알고리즘 데이트레이딩 전략의 수익 곡선(equity curve)을 보여주고 있다. 비교를 위해 비교모형 전략의 수익 곡선도 Figure 5에 표시하였다.

〈Table 9〉 Trading Performance (9:20)

Algorithm	TP	MDD	SR
LOGIT	181.45	**66.2**	0.049
KNN	-40.15	111.75	-0.011
RF	-82.75	200.25	-0.022
SVM	**186.95**	104.95	**0.051**
XGB	96.85	106.35	0.026
BM	56.15	117.8	0.015

* Boldface marks the best values.



〈Figure 5〉 Equity Curve of RF Strategy (9:20)

9시 20분 시점에서 관찰된 주문 불균형지수, 외국인 투자자 거래 비중, 주가 변동 정보를 입력변수로 하는 당일의 종가 등락 예측을 이용한 데이트레이딩 전략의 수익성은 SVM 알고리즘이 186.95포인트로 가장 높은 총수익을 보여주고

있다. 투자 스트레스의 정도를 측정하는 MDD는 LOGIT 알고리즘이 66.2포인트로 가장 낮게 나왔고, 투자 위험 대비 수익의 상대적 크기를 측정하는 샤프비율(SR)은 SVM 알고리즘이 0.051로 높은 비율을 보여주었다.

9시 5분, 9시 10분, 9시 15분, 9시 20분에서의 투자 성과를 평균하면 Table 10과 같은 결과를 얻을 수 있다.

〈Table 10〉 Trading Performance Averages

Algorithm	TP	MDD	SR
LOGIT	196.99	**72.61**	0.052
KNN	52.94	83.61	0.013
RF	75.16	96.19	0.020
SVM	**209.11**	88.04	**0.055**
XGB	192.89	78.78	0.051
Average	145.42	83.85	0.038
BM	65.74	118.46	0.017

* Boldface marks the best values.

코스피200 주가지수선물 데이트레이딩의 9시 5분, 9시 10분, 9시 15분, 9시 20분 진입 시점 전체에 대한 평균적인 투자 성과는 SVM 알고리즘이 209.11포인트로 가장 높은 총수익 평균을 보여주고 있으며, 위험을 고려하더라도 SVM 알고리즘의 샤프비율이 평균 0.055로 가장 높게 나타나 위험과 수익 측면에서 모두 SVM 알고리즘의 결과가 가장 좋은 투자 성과를 보여주었다.

Figure 6은 데이트레이딩의 진입 시점별로 전체 분류 알고리즘의 예측 종가를 이용한 데이트레이딩 전략의 투자 성과를 비교하여 보여주고 있다.

〈Figure 6〉 Average Performance of All Strategies

진입 시점별로 전체 분류 알고리즘 진입전략의 총수익 평균은 9시 5분 진입 시점에서 203.2포인트로 가장 높게 나타났고, 진입 시점이 늦어질수록 총수익은 낮아지는 특성을 보인다. 총수익과 더불어 위험을 고려한 투자 성과의 평가지표인 샤프비율 역시 데이트레이딩의 진입 시점이 가장 빠른 9시 5분의 경우 가장 높은 성과를 보여주었고 진입 시점이 늦어질수록 투자 성과는 낮아지는 특징을 보여주고 있다. Figure 6의 결과는 위험을 최소화하고 수익을 극대화하려는 데이트레이더에게 실전 투자에서 진입 시점 선택의 중요성을 잘 보여주고 있다.]

실제 선물거래에는 거래에 따른 거래비용(transaction cost)이 발생한다. 거래비용에는 증권회사에 지출하는 거래수수료(brokerage commission)와 거래 실행에 따른 시장 충격비용인 슬리피지 비용(slippage cost)을 포함한다. 유동성이 풍부한 코스피200 주가지수선물 거래에서 종가에 포지션을 청산하는 소규모 데이트레이딩의 경우 1틱(tick) 또는 2틱을 고려하면 적절하다. Table 11은 거래에 따른 거래비용 1틱을 가정하는 경우, 거래비용 고려 후의 전략별 총 손익을 보여주고 있다.

<Table 11> Total Profits after Transaction Costs

Algorithm	9:05	9:10	9:15	9:20
LOGIT	121.85	**235.20**	44.10	113.00
KNN	157.35	-82.80	-28.00	-108.60
RF	13.25	17.10	**147.70**	-151.20
SVM	**227.65**	169.60	46.90	**118.50**
XGB	153.65	194.50	121.20	28.40
BM	-4.45	6.60	-0.70	-12.30

* 1 tick transaction costs are assumed.
* Boldface marks the best values.

거래비용을 고려하지 않은 경우는 9시 10분과 9시 20분 진입시간에서만 손실이 발생하였지만, 거래비용을 고려한 경우는 9시 15분에서도 손실이 발생하고 있다. 전체적으로 LOGIT, SVM, XGB 전략은 모든 진입 시간대에서 양의 수익을 가져다주고 있으며, 비교전략 BM의 경우는 9시 10분을 제외하면 모두 손실이 발생하고 있다.

6. 결론

투자자들이 제출하는 매수나 매도의 주문 상황을 보여주는 Limit Order Book 정보는 시세표라는 이름으로 투자자들에게 실시간으로 제공되고 있다. 실제로 투자자들은 주식이나 선물 등의 주문을 제출할 때 이러한 시세표를 참조하며 특히 시세표의 변동에 민감하게 반응하는 투자자들이 대부분이다.

본 연구는 전 세계 투자자들에게 실시간으로 공개되고 있는 코스피200 주가지수선물 시장의 LOB 정보를 이용하면 안정적인 수익이 가능한지를 분석하는 것을 목적으로 하였다. 구체적으로, 장 시작 초반의 코스피200 주가지수선물 시장의 LOB 정보에서 총매수 주문량과 총매도 주문량의 한쪽 쏠림을 측정하는 주문 불균형지수를 산출하고, 분류 알고리즘을 이용하여 당일의 장 마감 종가까지의 코스피200 주가지수선물 가격의 등락을 예측하였다. 실증분석을 통해 예측의 정확성을 평가하고 예측 결과를 이용하여 장 마감 종가로 포지션을 청산하는 데이트레이딩 전략의 성과를 평가하였다. 분석 기간은 2004년 1월 19일부터 2022년 6월 30일까지의 4,564일간의 5 분봉의 장기 자료이다.

실증분석 결과는 다음과 같다. 첫째, 주문 불균형 정보와 외국인 투자자 거래 비중 정보는 코스피200 주가지수선물 가격에 강한 양의 영향력을 미치고 있다. 둘째, 외국인 투자자 거래 비중 정보와 달리 주문 불균형 정보는 코스피200 주가지수선물의 당일 종가까지의 미래 가격에도 유의적인 양의 영향을 미치는 것으로 나타났다. 셋째, 분류 알고리즘을 이용한 당일 종가의 등락 예측 성공률의 최고치는 SVM 알고리즘에 기반한 54.1%로 나타났다. 특히, 진입 시점이 빠를수록 평균적인 예측 성공률도 높게 나타나는 특성을 보여주었다. 넷째, 예측 결과를 이용한 데이트레이딩 전략의 투자 성과는 비교모형인 오버나이트 퍼즐 전략보다 높게 나타났다. 분류 알고리즘에 기반한 데이트레이딩 전략의 성과는 KNN 알고리즘을 제외하면 모두 비교모형보다 총수익 평균이 높게 나타났고, 위험도를 반영하는 샤프비율 역시 비교모형보다 높게 나타났다.

본 연구는 투자자들에게 실시간으로 공개되는 LOB 정보 중 총매수주문 수량과 총매도주문 수량 정보의 경제적 가치(economic value)가 존재함을 밝혔다는 점에서 기존의 연구와는 학술적 차별점을 갖는다. 데이트레이딩 전략의 실증분석 결과는 실제 시장 참여자들에게도 실전적 측

면에서 도움이 될 것으로 판단된다. 향후 연구에서는 거래 시장을 주식시장으로 확장하여 주문 불균형 정보의 경제적 가치가 나타나는지를 분석할 필요가 있다. 또한, 최근 주가 예측에서 활발히 연구되고 있는 딥러닝 모형 등으로의 확장을 통해 주가 예측의 정확성을 높임으로써 투자 전략의 성과 개선도 가능할 것으로 판단된다.

참고문헌(References)

Berkman, H., Koch, P. D., Tuttle, L., & Zhang, Y. J. (2012). Paying attention: Overnight returns and the hidden cost of buying at the open. *Journal of Financial and Quantitative Analysis,* 47(4), 715-741. https://doi.org/10.1017/S0022109012000270

Cao, C., Hansch, O., & Wang, X. (2009). The information content of an open limit-order book. *The Journal of Futures Markets,* 29(1), 16-41. https://doi.org/10.1002/fut.20334

Cao, H., Lin, T., & Zhang, H. (2019). Stock price pattern prediction based on complex network and machine learning. *Complexity,* 2019(10), 1-12. https://doi.org/10.1155/2019/4132485

Cao, L., & Tay, F. E. H. (2001). Financial forecasting using support vector machines. *Neural Computing & Applications,* 10, 184-192. https://doi.org/10.1007/s005210170010

Cenesizoglu, Dionne, T., G., & Zhou, X. (2022). Asymmetric effects of the limit order book on price dynamics. *Journal of Empirical Finance,* 65, 77-98. https://doi.org/10.1016/j.jempfin.2021.11.002

Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications,* 80, 340-355. https://doi.org/10.1016/j.eswa.2017.02.044

Cont, R., Stoikov, & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research,* 58(3), 549-563. https://doi.org/10.1287/opre.1090.0780

Griffiths, M. D., Smith, B. F., Turnbull, D. A. S., & White, R. W. (2000). The costs and determinants of order aggressiveness. J*ournal of Financial Economics,* 56(1), 65-88. https://doi.org/10.1016/S0304-405X(99)00059-8

Han, S. B. (2017). Foreigners' short selling and price pressures in the Korean stock market. *Journal of Industrial Economics and Business,* 30(6), 2119-2139. https://doi.org/10.22558/jieb.2017.12.31.6.2119

Harris, L., & Panchapagesan, V. (2005). The information content of the limit order book: Evidence from NYSE specialist trading decisions. *Journal of Financial Markets,* 18(1), 25-67. https://doi.org/10.1016/j.finmar.2004.07.001

Kang, J., & Ryu, D. (2010). Which trades move asset prices? An analysis of futures trading data. *Emerging Markets Finance & Trade,* 46, 7-22. https://doi.org/10.2753/REE1540-496X4603S101

Kim, S. W. (2022). Performance on Altcoin investment using technical trading rules. *Journal of the Korean Academia-Industrial,* 23(6), 198-207. https://doi.org/10.5762/KAIS.2022.23.6.198

Kim, S. W. (2019). Performance analysis on day trading strategy with bid-ask volume. The *Journal of the Korea Contents Association,* 19(7), 36-46. https://doi.org/10.5392/JKCA.2019.19.07.036

Kim, S. W., & Ahn, H. (2010). Development of an intelligent trading system using support vector

machines and genetic algorithms. *Journal of Intelligence and Information Systems,* 16(1), 71-92.

Kim, T. D., & Ok, K. (2015). Private information and trading behavior: KOSPI200 Futures Markets. *Journal of Derivatives and Quantitative Studies,* 23(2), 207-241. https://doi.org/10.1108/JDQS-02-2015-B0003

Kim, Y., Choi, H. S., & Kim, S. W. (2020). A study on risk parity asset allocation model with XGBoost. *Journal of Intelligence and Information Systems,* 26(1), 135-149. https://doi.org/10.13088/jiis.2020.26.1.135

Kozhan, R., & Salmon, M. (2012). The information content of a limit order book: The case of an FX market. *Journal of Financial Markets,* 15, 1-28. https://doi.org/10.1016/j.finmar.2011.07.002

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications,* 197, 1-41. https://doi.org/10.1016/j.eswa.2022.116659

Lee, W. B., & Choe, H. (2007). Short-term return predictability of information in the open limit order book. *Asia-Pacific Journal of Financial Studies,* 36(6), 963-1008.

Lee, Y., & Kim, W. C. (2013). A stochastic model for order book dynamics: An application to Korean stock index futures. *Management Science and Financial Engineering,* 19(1), 37-41. https://doi.org/10.7737/MSFE.2013.19.1.037

Lohrmann, C., & Luukka, P. (2019). Classification of intraday S&P500 returns with a random forest. *International Journal of Forecasting,* 35, 390-407. https://doi.org/10.1016/j.ijforecast.2018.08.004

Malagrino, L., Roman, N. T., & Monteiro, A. M. (2018). Forecasting stock market index daily direction: A Bayesian network approach. *Expert Systems with Applications,* 105, 11-22. https://doi.org/10.1016/j.eswa.2018.03.039

Park, Y. J., Kutan, A. M., & Ryu, D. (2019). The impacts of overseas market shocks on the CDS-option basis. *The North American Journal of Economics and Finance,* 47, 622-636. https://doi.org/10.1016/j.najef.2018.07.003

Ranaldo, A. (2004). Order aggressiveness in limit order book markets. *Journal of Financial Markets,* 7(1), 53-74. https://doi.org/10.1016/S1386-4181(02)00069-1

Ryu, D. (2013). Price impact asymmetry of futures trades: Trade direction and trade size. *Emerging Markets Review,* 14, 110-130. https://doi.org/10.1016/j.ememar.2012.11.005

Sharpe, W. F. (1966). Mutual Fund Performance. *The Journal of Business,* 39(1), 119-138.

Song, J. H., Choi, H. S., & Kim, S. W. (2017). A study on commodity asset investment model based on machine learning technique. *Journal of Intelligence and Information Systems,* 23(4), 127-146. https://doi.org/10.13088/jiis.2017.23.4.127

Thakkar, A., & Chaudhari, K. (2022). Information fusion-based genetic algorithm with long short-term memory for stock price and trend prediction. *Applied Soft Computing,* 128, 1-20. https://doi.org/10.1016/j.asoc.2022.109428

Wang, M. C., Zu, L. P., & Kuo, C. J. (2008). The state of the electronic limit order book, order aggressiveness and price formation. *Asia-Pacific Journal of Financial Studies,* 37(2), 245-296.

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems*

with Applications, 79, 153-163. https://doi.org/10.1016/j.eswa.2017.02.041

Yang, H. (2021). Investor sentiment and market dynamics: Evidence from index futures markets. *Investment Analysts Journal,* 50, 258-272. https://doi.org/10.1080/10293523.2021.2010376

Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications,* 186, 1-21. https://doi.org/10.1016/j.eswa.2021.115716

Zhang, N., Lin, A., & Shang, P. (2017). Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting. *Physica A,* 477, 161-173. https://doi.org/10.1016/j.physa.2017.02.072

Zhang, D., & Lou, S. (2021). The application research of neural network and BP algorithm in stock price pattern classification and prediction. F*uture Generation Computer Systems*, 115, 872-879. https://doi.org/10.1016/j.future.2020.10.009

Abstract

Classification Algorithm-based Prediction Performance of Order Imbalance Information on Short-Term Stock Price

S. W. Kim*

Investors are trading stocks by keeping a close watch on the order information submitted by domestic and foreign investors in real time through Limit Order Book information, so-called price current provided by securities firms. Will order information released in the Limit Order Book be useful in stock price prediction? This study analyzes whether it is significant as a predictor of future stock price up or down when order imbalances appear as investors' buying and selling orders are concentrated to one side during intra-day trading time. Using classification algorithms, this study improved the prediction accuracy of the order imbalance information on the short-term price up and down trend, that is the closing price up and down of the day. Day trading strategies are proposed using the predicted price trends of the classification algorithms and the trading performances are analyzed through empirical analysis. The 5-minute KOSPI200 Index Futures data were analyzed for 4,564 days from January 19, 2004 to June 30, 2022. The results of the empirical analysis are as follows. First, order imbalance information has a significant impact on the current stock prices. Second, the order imbalance information observed in the early morning has a significant forecasting power on the price trends from the early morning to the market closing time. Third, the Support Vector Machines algorithm showed the highest prediction accuracy on the day's closing price trends using the order imbalance information at 54.1%. Fourth, the order imbalance information measured at an early time of day had higher prediction accuracy than the order imbalance information measured at a later time of day. Fifth, the trading performances of the day trading strategies using the prediction results of the classification algorithms on the price up and down trends were higher than that of the benchmark trading strategy. Sixth, except for the K-Nearest Neighbor algorithm, all investment performances using the classification algorithms showed average higher total profits than that of the benchmark strategy. Seventh, the trading performances using the predictive results of the Logical Regression, Random Forest, Support

* Corresponding Author: Sun Woong Kim
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, Korea
Tel: +82-2-910-5471, E-mail: swkim@kookmin.ac.kr

Vector Machines, and XGBoost algorithms showed higher results than the benchmark strategy in the Sharpe Ratio, which evaluates both profitability and risk. This study has an academic difference from existing studies in that it documented the economic value of the total buy & sell order volume information among the Limit Order Book information. The empirical results of this study are also valuable to the market participants from a trading perspective. In future studies, it is necessary to improve the performance of the trading strategy using more accurate price prediction results by expanding to deep learning models which are actively being studied for predicting stock prices recently.

Key Words : Limit Order Book, Order Imbalance Information, Classification Algorithms, KOSPI200 Index Futures, Day Trading

Received : October 14, 2022 Revised : November 6, 2022 Accepted : November 11, 2022
Corresponding Author : S. W. Kim

저 자 소 개

김선웅
현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. 서울대학교 경영학과에서 경영학사를 취득하고, KAIST 경영과학과에서 투자론을 전공하여 공학석사와 공학박사 학위를 취득하였다. 주요 관심분야는 트레이딩시스템, 투자공학, AI증권이다.