인공지능 기반 주식시장 변동성 이상탐지모델 개발

김현정 [1], 유헌창 [2]
[1] 고려대학교 컴퓨터정보통신대학원 인공지능융합학과
[2] 고려대학교 정보대학 컴퓨터학과

hash@korea.ac.kr, yuhc@korea.ac.kr

Development of a Stock Volatility Detection Model Using Artificial Intelligence

HyunJung Kim[1], Heonchang Yu[2]
[1]Dept. of Applied Artificial Intelligence, Graduate School of Computer & Information Tech., Korea University
[2]Dept. of Computer Science and Engineering, Korea University

요 약

경제 위기 대비를 위해 인공지능을 활용한 주식시장 변동성 이상을 탐지하는 목적을 가지고 있다. 글로벌 이슈와 경제 위기 대비를 위해 주식시장 변동성 예측의 중요성이 부각되고 있으며, 기존의 주식시장 변동성 지수인 VIX 의 한계로 인해 더 복잡한 모델 및 인공지능을 활용한 연구에 관심이 집중되고 있다. 기존의 주식시장 변동성 예측에 관한 연구들은 통계적인 방법을 사용했으며 인공지능을 이용한 연구 또한 대부분 이상치 구간을 표시하여 예측을 목표로 하고 있으나 이러한 접근법은 라벨이 있는 데이터 수집 어려움, 클래스 불균형 문제가 있다. 본 연구는 인공지능을 활용한 주식시장 변동성 탐지에 기여하고 지도 학습 방식 대신 비지도 학습 기반의 이상탐지모델을 사용하여 주식시장 변동성을 예측하는 새로운 방법론을 제안한다. 본 연구에서 개발한 인공지능 모델은 IsolationForest 모델을 활용하며, 시계열 데이터를 전처리한 후 정상성을 확보하는 등의 과정을 거친다. 실험 결과로 인공지능 모델이 주요 경제이슈를 이상치로 검출하는 성능을 확인하였으며 재현율 약 93.6%, 정밀도 100%로 높은 성능을 달성했다.

1. 서론

자본주의의 특징과 정보통신기술, 글로벌화로 인한 금융시장 규제 완화로 주식시장의 흐름과 이동이 자유롭게 이루어지고 있으며, 이로 인해 주식시장은 경제 및 금융시장 상황을 제공하는 중요한 역할을 하고 있다[1]. 최근 글로벌 이슈와 경제 위기에 대비하기 위해 주식시장 변동성 탐지는 더욱 중요해지고 있다. 주식 변동성은 주가의 상승 또는 하락 정도를 나타내는 용어이며 이를 예측하는 것은 투자자에게 수익을 창출할 수 있는 기회를 제공한다. 그러나 주식시장 변동성의 대중적인 지수인 VIX 가 예측력 부족으로 비판을 받아 현재에도 높은 예측력을 갖는 더 복잡한 모델들이 연구되고 있으며, 최근에는 인공지능을 활용하여 주식시장의 위험을 예측하는데 많은 관심이 집중되고 있다.

본 연구의 목적은 변화하는 금융 환경에서 다양한 시계열 데이터를 기반으로 인공지능을 이용하여 경제 위기 대비를 위해 주식시장 변동성을 탐지하는 것이다. 기존의 주식시장 변동성 예측에 관한 연구들은 대부분 통계적인 방법을 사용했으며 AI 이상탐지연구 또한 대부분 이상치 구간을 기준으로 현재 금융시장과 비교하는 한계점이 있으며[1], 라벨이 있는 데이터 수집 어려움, 클래스 불균형 문제가 있다. 따라서 본 연구에서는 기존의 지도 학습 방식 대신 비지도 학습 기반 이상탐지모델을 제안하여 새로운 방법론을 연구한다. 본 논문에서 제안하는 주식시장 변동성 이상탐지모델은 실험 결과 재현율은 약 93.6%, 정밀도는 100%로 높은 성능을 달성했다. 과거 경제위기나 이슈 시점을 이상치로 탐지했으며 주식시장의 패턴 인식 가능성이 있음을 보였다.

2. 관련연구 및 이론

2.1 주식시장 변동성 예측관련 이전 연구

주식시장 변동성 예측 관련 선행 연구에서는 예측

모델을 개발하거나 새로운 예측 변수를 찾아 변동성 예측의 정확성을 향상시키기 위해 노력해왔다[2][3][4][5]. 최근에는 Ma et al.(2023)이 Markov 정권전환(MRS) 기법과 변수 선택 방법론을 통합한 MRS-LASSO 모델을 제안하여 미국 주식시장의 변동성을 예측했다[5]. 이러한 이전 연구 모델은 통계 및 경제적 관점에서 미래 변동성을 성공적으로 예측할 수 있었지만, 복잡한 수식과 이론을 기반으로 하기 때문에 특정 패턴 이외의 예측에 한계가 있을 수 있다.

2.2 IsolationForest 기반 이상탐지모델

IsolationForest 는 Decision Tree 를 기반으로 하며, 정상과 이상값을 분리할 때 Decision Tree 를 깊이 타고 내려가는 방식을 활용한다. 이상값은 상단에서 분리할 수 있어 계산량이 적고, 특성을 랜덤하게 선택하여 최대값과 최소값 사이의 임계값으로 관측치를 분리한다. 이 모델의 주요 특징은 비지도(Unsupervised) 이며, 트리 기반 비모수(Non-parametric based on Tree) 구조이다. 주요 장점으로는 모든 점 간 거리 계산이 필요한 군집기반 알고리즘 대비 계산량이 적고, 고차원 데이터에서도 효과적이며, Random sampling 과 Ensemble 을 활용하여 견고한 모델을 형성할 수 있다[6].

2.3 AutoEncoder 기반 이상탐지모델

AutoEncoder 는 시계열 데이터 분석에 효과적인 비지도 이상탐지 딥러닝 모델로, 모델링한 분포의 범위를 벗어난 값을 이상치로 간주한다. 이 모델은 정상 데이터를 저차원 잠재 공간으로 압축한 후 다시 복원하여 이상 탐지를 수행한다. 인코더는 정상 데이터를 저차원으로 압축하며, 디코더는 압축된 샘플을 다시 원래 차원으로 복원한다. 이 과정에서 CNN (Convolutional Neural Network) 신경망을 사용하여 데이터의 내제된 지역적 특징을 추출하고 이를 기반으로 모델을 형성한다.

3. 인공지능 주식시장 변동성 이상탐지모델 개발

3.1 데이터셋

주식시장 변동성을 탐지하기 위해 적절한 입력 변수를 선택하는 것이 중요하다. <표 1>과 같이 선택된 8 개의 변수(VIX, SPY 등)는 주식시장의 불안정성을 사전에 알려주고, 내외적 충격을 잘 반영하며, 금융시장 모니터링을 위한 효율적인 변수로 판단되었다[1].

이 데이터는 investing.com 에서 수집되었고, S&P500 지수, VIX 지수, 달러원환율 등 최근 18 년 동안(06-01-01~23-12-31)의 일별 종가 데이터로 이루어져 있다.

<표 1> 8 개 입력 변수 목록

Name	description
SPY	S&P500(시가총액 상위 500 개 기업) 추종 펀드
DIA	다우존스(각 산업군 대표 우량주 30 개 기업) 추종 펀드
QQQ	나스닥(시가총액 상위 100 개 기업, 기술주) 추종 펀드
VIX	VIX 지수(CBOE Volatility Index), S&P500 지수 옵션변동성
GOLD	금선물(Gold Future), 대표적인 인플레이션 헷지 옵션
USD/KRW	환율(USD/KRW), 환율과 주가는 양의 상관관계
USD_10	미국국채(USD 10 Year Bond Yield), 대표 안전자산
OIL	국제유가 종가

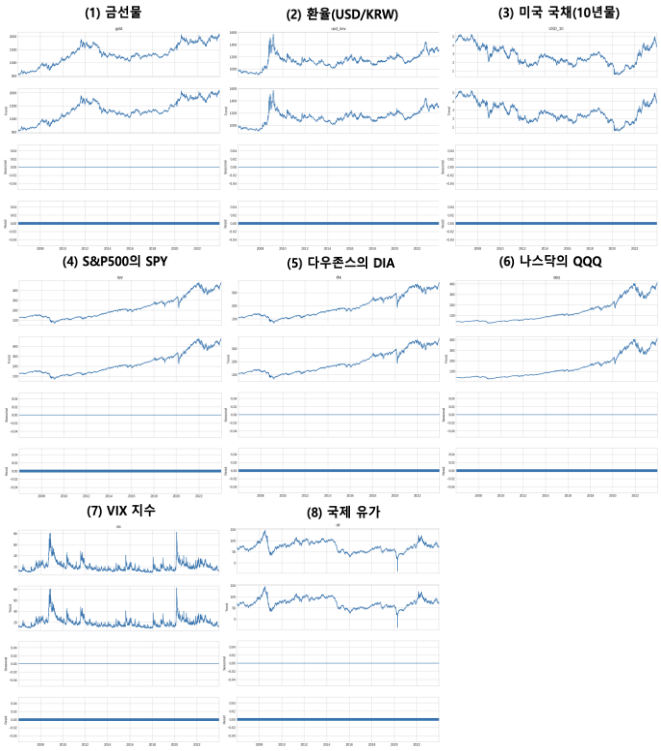
3.2 시계열 데이터 전처리

시계열 데이터 전처리 절차를 통해 정상성을 확보하고, 이상치를 탐지하기 위한 모델을 구축하였다.

(1) 결측치 처리: 시계열 데이터에서 발생한 약 8 백여개의 결측치는 일자를 인덱스로 하여 일자에 가중치를 부여한 보간법을 통해 처리되었다.

(2) 노이즈 처리: 결측치 처리 후 남은 데이터에서 노이즈를 제거하기 위해 칼만 필터를 사용하였다. 이는 노이즈가 통계적 특성을 왜곡시키는 것을 방지하기 위한 조치이다.

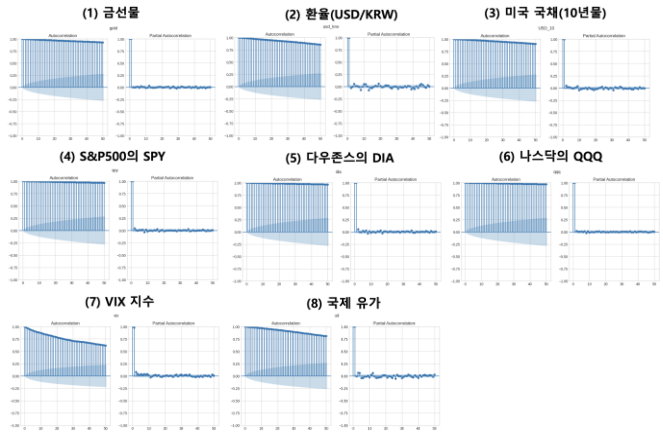
(3) 시계열 분해: 시계열에서 각각의 성분을 구할 수 있다면, 원래 시계열에서 특정 성분을 제거하는 것이 가능하기 때문에 시계열분해를 수행한다. (그림 1)은 각 지표를 계절(특정 계절에 영향 받음), 트랜드 (상승 혹은 하강 기울기), 패턴(주기적 패턴) 3 가지 속성으로 분해한 결과이다. 8 개 지표의 경우 모두 추세(Trend)만 있다.



(그림 1) 각 변수의 시계열 분해 (Trend(추세), Seasonal(계절), Resid(불규칙)).

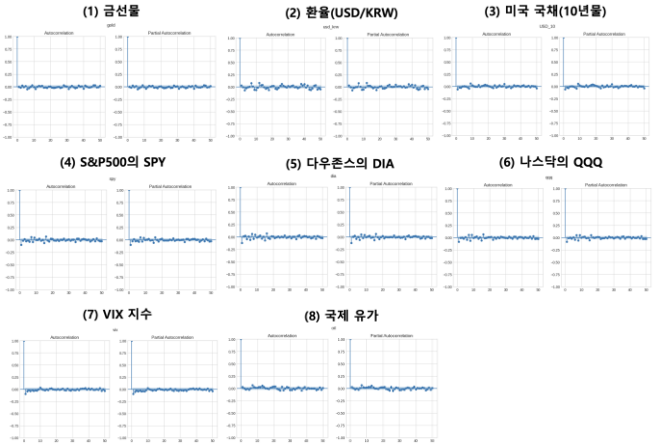
(4) ACF 및 PACF 분석: 정상 시계열의 경우

ACF(Autocorrelation)가 빠르게 0 으로 감소해야 하는데, (그림 2)와 같이 천천히 감소하고 있으므로 비정상 시계열임을 알 수 있다. 따라서, 정상 시계열로 변환이 필요하다.



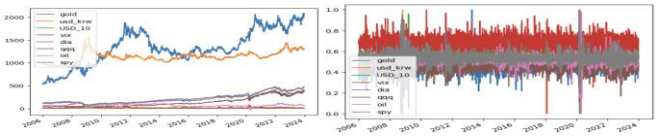
(그림 2) 각 변수의 ACF 와 PACF.

(5) 정상화: 로그와 차분을 수행한 데이터의 경우, (그림 3)과 같이 ACF 가 빠르게 0 으로 감소하므로 정상 시계열임을 알 수 있다.



(그림 3) 로그와 차분을 수행한 각 변수의 ACF 와 PACF.

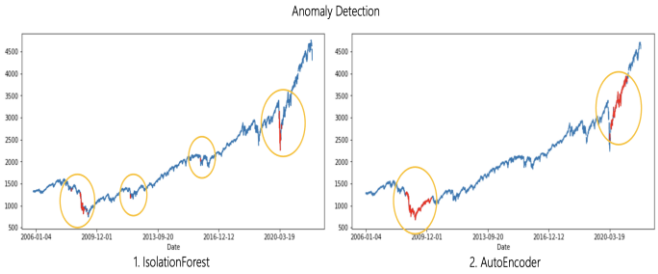
(6) 특성 스케일링: 인공지능 모델 구축을 위해 (그림 4)의 우측과 같이 정상 시계열 데이터에 Minmaxscaler 를 적용하여 값의 분포를 유지하면서 정규화하였다.



(그림 4) 정상 시계열 및 Minmaxscaler 수행 전과 후

3.3 알고리즘 선택

본 연구는 주식시장의 변동성을 다루는 데 있어서 데이터 특징을 고려하여 비지도 방식의 이상탐지 알고리즘을 선택했다. 변동성이 높은 데이터는 정답이 없는 형태로 비지도 학습에 해당하고 주식시장 변동성이 높은 데이터는 정상 일자 데이터와 비교 시 매우 적은 비중을 차지하는 아웃라이어(Outlier)에 해당하므로 이상탐지(Anomaly Detection)에 해당한다. 비지도 학습 기반 이상탐지모델 중 대표적인 알고리즘인 IsolationForest 와 AutoEncoder 를 활용한 인공지능 모델을 구축하여 이상치(anomaly)를 비교한 결과 IsolationForest 가 주식시장 관련 금융 지표 데이터에 대한 이상치 분별력이 우수했다.

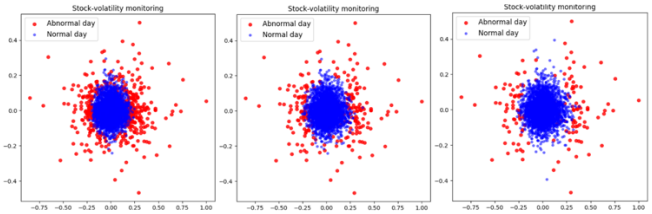


(그림 5) SPY 데이터 이상탐지결과 비교 : IsolationForest vs AutoEncoder.

(그림 5)는 SPY(S&P500) 데이터에 이상치를 붉은색으로 표시하여 IsolationForest 와 AutoEncoder 모델을 비교한 것이다. IsolationForest 는 사전에 검출하고자 했던 글로벌 금융위기, 코로나 19 시점 등 경제 이벤트를 탐지했고, AutoEncoder 는 2020 년 초 코로나 19 팬더믹 발생으로 인한 급락 이후 급등 시기인 20 년 하반기~21 년 상반기 주가를 이상치로 탐지하고 있어 2 개 모델의 이상치 검출 결과가 차이가 있었다. 이상탐지에 사용한 CNN 기반 AutoEncoder 모델의 경우, 시계열 데이터상 지역적 특징이 있다는 가정이 충족되어야 해서 정상 시계열 데이터를 사용하는 경우, 이상 탐지가 되지 않아 본 연구에 적합하지 않다.

3.4 하이퍼파라미터 튜닝

본 연구에서 활용한 IsolationForest 모델은 하이퍼파라미터로 이상치 임계값을 정의한다. (그림 6)와 같이 임계값이 커질수록 이상치가 많아지며 분석이 달라지므로 적절하게 설정하는 것이 중요하다. 여러 번의 실험으로 AI 모델이 주요 경제위기만을 이상치로 검출하도록 임계치를 조정하였으며 0.03(3% 이하는 이상치)에서 성능이 가장 우수하다고 판단되었다.



(그림 6) 임계값에 따른 이상치 변화(0.1, 0.05, 0.03).

4. 연구 결과 및 모델 성능 평가

본 연구에서는 정답 데이터가 없어 AI 모델의 성능

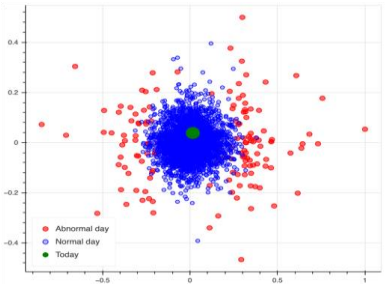
을 확인하기 위해 경제 위기와 같은 사전에 검출하고자 했던 일자들이 검출되었는지 확인한다.

2006 년 1 월부터 2023 년 12 월까지의 4,695 건의 일자 중 141 건이 이상치로 검출되었다. 본 연구의 목적이 주식시장 변동성을 탐지하여 경제 위기를 사전에 대비하기 위함 이기 때문에, <표 2> 와 같이 월별 경제이슈 관련 사건 요약표를 사전에 준비하고 이에 대응하는 AI 모델 검출 일자를 표기한다. AI 모델의 이상탐지 예측 일자 141 건 중 실제 이상 일자는 132 건으로 재현율(recall)은 약 93.6% 이었다. 또한 2020 년 3 월 주식급락, 2011 년 유럽발 금융위기, 2008 년 금융위기와 같은 사전에 검출하고자 했던 일자들이 모두 검출되어 정밀도(precision)는 100%로 높은 성능을 달성했다.

<표 2> 월별 경제이슈 관련 사건 요약 및 AI 검출일자 (2006~2023 년)

년	월	내용	AI 검출 일자
2007	11	중국증시 거품붕괴	07.11.12
2008	03	짐바브웨 초인플레이션	08.3.10 외 3 건
	09	글로벌 금융위기(2007~2009 년), 리먼브라더스 대공황	08.9.12~09.7.14 (61 건)
	10	불가리아, 중앙아시아 에너지위기	08.10.1
2010	02	포르투갈 금융위기	10.2.3
2011	08	증시하락	11.8.3 외 6 건
	09	유럽 재정위기	11.9.21 외 4 건
2012	06	키프로스 금융위기	12.6.28
2013	06	중국 은행유동성위기, 베네수엘라 경제위기	13.6.19
2015	08	중국 증시변동	15.8.20 외 2 건
2016	06	브렉시트 증시붕괴	16.6.23
2018	01	암호화폐 붕괴	18.2.2 외 1 건
	10	터키 외환부채위기	18.10.9
2019	08	미중 무역전쟁	19.8.2 외 1 건
2020	02	주가 대폭락(코로나 팬데믹), 러시아-사우디 유가 전쟁	20.2.21~20.3.31(21 건)
	03	증시 반등 및 불안정	20.4.1~20.10.27 (9 건)
	11	미국 대선	20.11.3 외 1 건
2022	03	인플레이션 대비 미국 기준금리 인상 1 차	22.3.8
	05	미국 기준금리 인상 2 차,환율상승	22.5.3 외 2 건
	07	미국 기준금리 인상 3 차	22.6.10
	09	미국 기준금리 인상 4 차	22.8.25 외 1 건
	11	미국 기준금리 인상 5 차	22.11.9
	12	미국 기준금리 인상 6 차	22.11.29

본 연구의 AI 모델은 주성분 분석(PCA)을 활용하여 주식시장 그리고 경제 환경이나 금융시장에 대해 일별 모니터링 기능을 제공할 수 있다. 이를 통해 최근 상황을 진단해 보면, (그림 7)은 특정 일자(2023-12-28)를 녹색으로 표시한 PCA 결과인데, 해당 일자는 변동성이 낮은 주식 환경에 위치하고, 앞으로 변동성이 낮고 위험이 있을 확률이 낮다고 해석할 수 있다.



(그림 7) IsolationForest 이상탐지 결과의 2 차원 PCA (녹색:특정일자, 파랑:정상, 적색:이상치).

5. 결론

본 연구는 인공지능을 주식시장 변동성 이상탐지에 적용하는 사례로서, 점점 방대해지고 복잡해지는 금융데이터를 활용하는 인공지능 기반 주식시장 변동성 이상탐지모델 구현을 궁극적인 목표로 하고 있다. AI 이상탐지기술을 통해 대규모의 자료를 검토·분석하고 위험의 소지가 있는 부분을 조기에 발견하는 비즈니스 문제 해결에 의미 있는 기여를 하였다. AI 이상탐지모델을 통해 기존 주식시장 변동성 예측 한계를 극복하고 복잡한 패턴을 이해하고 대량의 데이터를 활용한 이상탐지가 가능해졌다. 또한, 본 연구는 비지도 학습 방법론을 통해 새로운 모델 방법론을 제안한다. 실험 결과에서는 IsolationForest 가 딥러닝 모델인 AutoEncoder 에 비해 높은 성능을 보였으며 다양한 금융 지표를 학습하고 이상치를 효과적으로 탐지하였다. 향후에는 더 많은 훈련 데이터를 이용하여 모델 성능을 향상시키고, 정량적인 비교를 위해 다양한 평가지표를 도입할 필요가 있다.

참고문헌

[1] 오경주, et al. "효율적 금융시장 모니터링을 위한 주식시장 불안정성 지수 개발과 이를 활용한 조기경보시스템의 구축." 한국은행 금융안정분석국 금융안정관련 외부연구용역사업. (2009).

[2] 김도현, et al. "VaR 에 근거한 주식시장 변동성 예측성과 평가." 국제지역연구 19(2): 207-229. (2015).

[3] Dueker, M. J. "Markov switching in GARCH processes and mean-reverting stock-market volatility." Journal of Business & Economic Statistics 15(1): 26-34. (1997).

[4] Fernandes, M., et al. "Modeling and predicting the CBOE market volatility index." Journal of Banking & Finance 40: 1-10. (2014).

[5] Ma, F., et al. "Stock market volatility predictability in a data-rich world: A new insight." International Journal of Forecasting 39(4): 1804-1819. (2023).

[6] FT Liu, et al. "Isolation Forest" 2008 Eighth IEEE International Conference on Data Mining: 413-422. (2008)