주성분분석와 인공신경망망을 활용한

7일 이내 주가 변동 범위 예측

허지혜[1], 유지훈[2]

경상국립립대학교 정보통계학과[1,2]

Abstract

주식 시장에서 미래의 주가를 예측하려는 시도는 꾸준히 이뤄지고 있다. 그럼에도 불구하고 시장의 내부 및 외부에서 주가에 영향을 미치는 요소들이 너무나도 다양하여 '주가 예측' 분야는 항상 재무 전문가들에게 도전적인 영역으로 남아있다.

최근 들어 기계학습 분야의 발전으로 기존에는 시행되기 어려웠던 수많은 변수들을 활용한 분석 방법이 꾸준히 개발되고 있으며, 이러한 상황은 고전적인 재무 분석 방법으로는 한계에 부딪혔던 주식 시장의 변동 예측 확률을 높여줄 것으로 기대된다.

이번 분석에서는 2가지 데이터셋에 대하여 43개의 재무 변수를 활용하여 7 거래일 이내의 주가 변동 범위를 예측해보았다. 분석에 활용한 기법은 주성분분석(Principal Component Analysis)와 인공신경망(Artificial Neural Network) 이며, 주성분분석을 통하여 주성분을 추출한 후 인공신경망을 시행하였다. 인공신경망 시행의 효과성을 확인하기 위하여 인공신경망과 SVM(Support Vector Machine)만 단독으로 시행한 모형을 비교 모형으로 생성하였다.

혼동 행렬(confusion matrix)을 통하여 각 모형에 대하여 분석한 결과, SVM은 종속 변수를 대부분 동일하게 분류하였음에도 불구하고 종속 변수의 불균등 분포로 인하여 정확도와 재현율은 높게 산출되었다. 하지만 투자 성과에 영향을 미치는 정밀도는 인공신경망에서 더욱 높게 산출되었으며, 주성분 분석 후 변수를 기준으로 인공신경망을 시행하였을 때 정밀도가 향상되는 것을 확인하였다.

1. 서론

"주식 시장은 살아있는 생물과 같다"라는 말이 있다. 이는 수많은 내외부 요인들과 지속적으로 상호작용하며 항상 변하고 움직이는 주식 시장의 모습 때문일 것이다. 이러한 주식 시장에 영향을 미치는 요인들은 국내외 정치·경제 상황, 물가 상승률, 금리, 환율, 유가, 각종 자산의 가격 변동 등등 셀 수 없을 만큼 다양하며, 동일한 요인이라고 하더라도 항상 주식 시장에 똑같은 영향을 미치는 것은 아니다.(ex) 이전까지 미국의 낮은 실업률은 주식 시장에 긍정적인 신호였으나, 최근에는 악재로 인식됨)

이러한 주식 시장에서 미래의 주가를 예측하려는 전문가들의 시도는 꾸준히 이뤄지고 있는데, 주가의 변동을 예측하기 위한 다양한 보조 지표를 개발하여 주식 투자에 널리 활용하고 있고 특정 조건에 도달하면 주식을 자동으로 매매하는 프로그램도 흔히 볼 수 있는 시대가 되었다.

하지만 다양한 보조 지표의 개발에도 불구하고 무수히 많은 요인으로부터 지속적으로 영향을 받는 주식 시장의 유기체적인 성격으로 인하여, 전통적이고 고전적인 분석 방법으로는 주식 시장 예측의 한계에 부딪힐 수 밖에 없다.

이러한 상황에서 우리는 최근 빠른 속도로 발전하고 있는 기계학습습 분야 중 생물체의 뇌 신경의 정보처리 메커니즘을 모방한 인공신경망 알고리즘을 활용하여 미래의 주가 변동을 예측하는 모형을 만들어보고자 한다.

2. 이론적 배경

1) SVM

SVM[1]은 1992년 Boser 등에 의해 제안된 기계학습 알고리즘으로, 커널 매핑 개념과 최적화 기술을 통계적 학습의 원리에 통합한 알고리즘이다. 이 알고리즘의 단순한 형태는 두 집합을 분리하기 위해 평면과 가장 가까운 좌표 간의 거리인 마진(margin)을 가장 크게 나타내는 최적화된 초평면(hyperplane)을 찾음으로써 두 개의 클래스(class)로 나누는 것이다. 여기서, 초평면이란 p차원의 공간에서 p-1차원인 평평한 아핀(affine) 부분 공간을 의미한다. SVM은 가장 최적화된 초평면을 찾는것이 목표이며, 학습 데이터를 통해 초평면을 찾고 이후 입력된 벡터 값의 좌표의 위치에 따라 해당 클래스를 반환한다. SVM은 데이터가 선형 분리가 가능 한 상태이어야 한다. 하지만 현실에서 수집되는 대부분의 데이터는 선형 분리가 불가능한 경우가 많다. 이 경우 커널 트릭을 적용하여 문제를 해결하였다.

2) 인공신경망

[1] Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh (1992)

인공신경망[2]은 1943년 Walter Pitts 등에 의해 제안된 기계학습 알고리즘으로, 사람의 두뇌가 반복적 경험을 통해 학습해가는 것을 모방한 알고리즘으로, 주어진 학습 데이터를 통해 반복적인 학습을 함으로써 특정한 패턴을 찾아내고 이를 통해 미래의 값을 예측한다. 인공신경망은 특히 독립변수(independent variable)와 종속변수(dependent variable) 사이의 관계를 명확히 설명하기 어렵고 복잡한 경우에도 비교적 좋은 결과를 제공한다는 장점이 있다. 인공신경망의 구조는 입력 층(input layer)과 출력 층(output layer) 그리고 은닉 층(hidden layer)으로 이루어져 있다. 입력 노드(node)를 통해 들어온 독립변수를 은닉 노드가 전달받아 선형결합을 한다.

3) 주성분분석

주성분분석[3]은 1901년 Karl Pearson 등에 의해 제안된 차원축소 방법으로 여러 변수들(x1, x2, x3,…x)이 관측되었을 때, 상관성이 높은 변수들을 공동요인으로 묶어 기존 변수들이 가지고 있는 정보들을 최대한 확보하는 적은 수의 새로운 변수들을 생성하는 방법이다. 이와 같은 주성분분석의 특성 때문에 다양한 방면의 텍스트 마이닝과 다변량 통계분석 관련 연구에서 널리 사용된다.

3. 데이터셋 설명

우리는 2가지 실제 기업의 주식 관련련 데이터셋을 가지고 실험하였다.

1) 삼성전자 주식 데이터셋

삼성전자 주식 데이터셋은 2019년 1월 3일부터 2022년 12월 5일까지 약 3년간의 삼성전자 주식과 관련된 지표와 달러 환율, 코스피, 나스닥의 시장 상황에 대한 43개의 지표이다. 해당 지표를 독립변수로 활용하고 7일 이내의 주가 변동폭을 종속변수로 하여 예측하고자 한다.

Table 1. 44가지 변수에 대한 설명

구분	지표명	설명
삼성전자 관련 지표	시가	당일 주식 시장에서 처음 거래된 가격(원)
	고가	당일 주식 시장에서 거래된 가장 높은 가격(원)
	저가	당일 주식 시장에서 거래된 가장 낮은 가격(원)
	종가	당일 주식 거래에서 가장 마지막에 거래된 가격(원)

[2] McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 115–133 (1943). https://doi.org/10.1007/BF02478259
[3] Karl Pearson F.R.S. (1901) LIII. *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2:11, 559-572, DOI: 10.1080/14786440109462720

구분	지표명	설명
	전일비	전일 종가 대비 당일 종가 상승 여부 0 : 전일 종가 대비 당일 종가 미상승 1: 전일 종가 대비 당일 종가 상승
	등락	전일 종가 대비 당일 종가 변동 가격(원)
	주가 등락률	전일 종가 대비 당일 종가 변동률(%)
	거래량	당일 주식 거래량(주)
	장전거래	당일 장전 시간 주식 거래량(주)
	장중거래	당일 장중 주식 거래량(주)
	장후거래	당일 장후 시간 주식 거래량(주)
	금액(백만)	당일 주식 거래 대금
	신용비	발행 주식 중 증권사의 돈을 빌려 매수한 주식의 전체 비율
	개인	당일 주식 거래량 중 개인 거래량(주)
	기관	당일 주식 거래량 중 국내 기관 거래량(주)
	외인	당일 주식 거래량 중 외국인 거래량(주)
	외국계	당일 주식 거래량 중 외국계 기관 거래량(주)
	프로그램	당일 주식 거래량 중 프로그램 거래량(주)
	외인비	발행 주식 중 외인, 외국계 보유 비율(%)
국내 시장 관련 지표	KOSPI	KOSPI 시장에 상장된 상장기업의 주식 변동을 기준시점과 비교시점을 비교하여 작성한 지표의 당일 종가
	전일비(KOSPI)	전일 대비 당일 KOSPI 상승 여부 0: 전일 대비 미상승 1: 전일 대비 상승
	등락(KOSPI)	전일 대비 당일 KOSPI 변동값
	등락률(KOSPI)	전일 대비 당일 KOSPI 변동률
	거래대금(KOSPI)	당일 KOSPI 시장에 상장된 상장기업의 총 주식 거래대금(억원)
	개인(KOSPI)	당일 거래대금(KOSPI) 중 개인의 거래대금
	외국인(KOSPI)	당일 거래대금(KOSPI) 중 외국인의 거래대금
	기관계(KOSPI)	당일 거래대금(KOSPI) 중 국내 기관의 거래대금
	금융투자(KOSPI)	당일 거래대금(KOSPI) 중 금융투자업체의 거래대금
	보험(KOSPI)	당일 거래대금(KOSPI) 중 보험업체의 거래대금

구분	지표명	설명
	투신(KOSPI)	당일 거래대금(KOSPI) 중 투자신탁업체의 거래대금
	기타금융(KOSPI)	당일 거래대금(KOSPI) 중 기타금융업체의 거래대금
	은행(KOSPI)	당일 거래대금(KOSPI) 중 은행의 거래대금
	연기금등(KOSPI)	당일 거래대금(KOSPI) 중 연기금 등의 거래대금
	사모펀드(KOSPI)	당일 거래대금(KOSPI) 중 사모펀드의 거래대금
	기타법인(KOSPI)	당일 거래대금(KOSPI) 중 기타법인의 거래대금
환율	기준환율	당일 USD 1달러 당 원화 환율
	환율 증감	전일 대비 당일 환율 변동값
미국 시장 관련 지표	Open(NASDAQ)	당일 나스닥 종합주가지수의 시가
	High(NASDAQ)	당일 나스닥 종합주가지수의 고가
	Low(NASDAQ)	당일 나스닥 종합주가지수의 저가
	Close(NASDAQ)	당일 나스닥 종합주가지수의 종가
	Volume(NASDAQ)	당일 나스닥 종합주가지수의 거래량
	나스닥 등락율	전일 종가 대비 당일 나스닥 종합주가지수 변동률
삼성전자 관련 지표	7 거래일 이내 최고 상승률	0: 7거래일이 경과하지 않아 자료 산출 불가 1: 7거래일 이내 주가 상승률 5% 미만 2: 7거래일 이내 주가 상승률 5% 이상 ~ 10% 미만 3: 7거래일 이내 주가 상승률 10% 이상 ~ 15% 미만 4: 7거래일 이내 주가 상승률 15% 이상

종속변수는 7 거래일 이내 최고 상승률이다. 이는 7 거래일 이내 주가 중 최고 높은 때의 상승률을 0부터 4까지 클래스를 분류하려한다. 클래스에 대한 각각의 설명은 Table 2와 같다.

Table 2. Counts of dependent variable for samsung

7 거래일 이내 최고 상승률	**COUNT**
0	7
1	789
2	146
3	25
4	2

TOTAL	969

삼성전자 주식 데이터셋의 종속변수의 갯수를 세어보면 Table 2와 같이 1과 2와 매우 많은 불균형 데이터(imbalanced data)이다.

삼성전자 주식 데이터셋은 총 969개로 이 중에서 학습 데이터는 775개 테스트 데이터는 194개로 이루어져 있다.

2) 현대차 주식 데이터셋

현대차도 삼성전자와 마찬가지로 2019년 1월 3일부터 2022년 12월 5일까지 약 3년간의 현대차 주식과 관련된 지표와 달러 환율, 코스피, 나스닥의 시장 상황에 대한 43개의 지표이다. 지표에 관한 설명은 Table 1과 동일하며 마찬가지로 7일 이내의 주가 변동폭을 예측하고자 한다. 종속변수 클래스의 값은 Table 3과 같다.

Table 3. Counts of dependent variable for hyundai car

7 거래일 이내 최고 상승률	**COUNT**
0	7
1	834
2	107
3	14
4	7
TOTAL	969

삼성전자 주식 데이터셋과 마찬가지로 현대차 주식 데이터셋도 1과 2가 매우 많은 불균형 데이터이다. 또한, 마찬가지로 총 969개로 이 중에서 학습 데이터는 775개 테스트 데이터는 194개로 이루어져 있다.

4. 결과

1) 주성분분석 결과

본 논문에서는 7일 이내의 주가 변동폭을 예측변수로 주로 활용되는 삼성전자 주식과 관련된 지표를 대상으로 주성분분석을 통해 새로운 예측변수를 생성하고 분석하고자 한다.

삼성전자 주식 데이터셋을 주성분분석을 이용하여 변수 각각의 고유값(eigenvalue)을 구하였다. Fig 2.은 고유값을 이용하여 plot을 그렸을 때를 나타낸다.

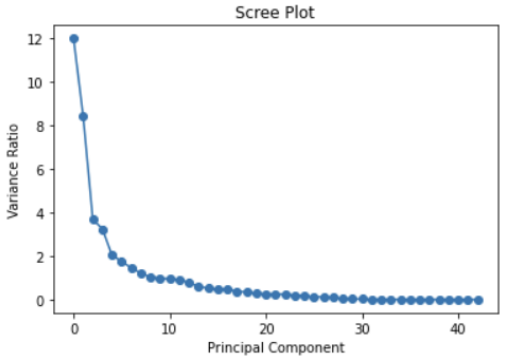

Fig 1. Eigenvalue visualization for samsung

삼성전자 주식 데이터셋은 누적기여율이 97%되는 선에서 변수의 개수를 지정하였다. 따라서 22개의 주성분으로 축소하여 분류 성능을 비교하고자 한다.

삼성전자 주식 데이터셋을 주성분분석을 이용하여 변수 각각의 고유값을 구하였다. Fig 2.은 고유값을 이용하여 plot을 그렸을 때를 나타낸다.

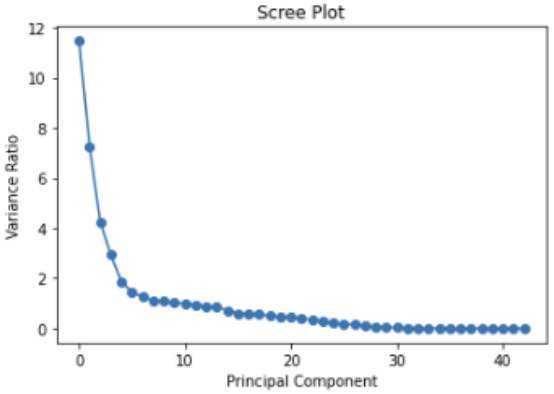

Fig 2. Eigenvalue visualization for hyundai car

현대차 주식 데이터셋도 마찬가지로 누적기여율이 97%되는 선에서 변수의 개수를 지정하였다. 따라서 23개의 주성분으로 축소하여 분류 성능을 비교하고자 한다.

2) 인공신경망 구조

본 논문에서 사용한 인공신경망의 구조는 다음과 같다.

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_2 (Dense)              (None, 20)                480

 dense_3 (Dense)              (None, 5)                 105

=================================================================
Total params: 585
Trainable params: 585
Non-trainable params: 0
_________________________________________________________________
```

Fig 3. ANN architecture

3) 판단 척도

여러 모형의 성능을 평가하기 위한 성능 평가지표로서 정확도(accuracy), 정밀도(precision), 민감도(sensitivity), F1-score 등이 있다. 분류 모형의 성능 평가지표는 혼동 행렬(confusion matrix) 계산으로 얻어진다. 혼동 행렬은 실제 데이터에서 실제 클래스(actual class)와 모형이 예측한 예측 클래스(predicted class)가 일치하는지를 갯수로 나타낸 표이다. 실제 클래스는 행으로 나타내고 예측 클래스는 열로 나타낸다. 여기서는 5개의 클래스를 분류하는 다중 분류(multi-label classification)에서 모형의 성능평가 지표에 대해 알아보고자 한다[4].

데이터셋에 대하여 독립변수를 x_i, 예측 클래스를 $MLC(x_i)$ 라고 정의했을 때 정확도, 정밀도, 민감도, F1-score는 다음과 같다.

다중 분류에서 정확도는 다음과 같이 정의된다.

$$Accuracy = \frac{1}{N}\left(\sum_{i=1}^{N} \left|\frac{MLC(x_i) \cap Y_i}{MLC(x_i) \cup Y_i}\right|\right),$$

실제 데이터에 대해 옳게 예측하는 비율을 의미한다. 정확도의 경우, 분류 클래스가 균일하지 못하면 성능을 제대로 나타낼 수 없다.

정밀도는 실제 i 클래스라고 예측한 것 중에 실제 i 클래스의 비율을 의미한다.

---

[4] R. Venkatesan and M. J. Er, "Multi-label classification method based on extreme learning machines," *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 619-624, doi: 10.1109/ICARCV.2014.7064375.

$$Precision = \frac{1}{N}(\sum_{i=1}^{N} |\frac{MLC(x_i) \cap Y_i}{MLC(x_i)}|),$$

민감도는 실제 i 클래스인 것 중 실제 i 클래스라고 예측한 것의 비율을 의미한다. 다른말로 재현율(recall)이라고도 한다.

$$Sensitivity = \frac{1}{N}(\sum_{i=1}^{N} |\frac{MLC(x_i) \cap Y_i}{|Y_i|}|),$$

F1 - score는 정밀도와 민감도의 조화평균이다.

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

F1-score는 주로 분류 클래스 간 데이터 불균형이 심할 때 사용한다.

4) 실험 결과

본 논문에서 모형의 성능을 평가하기 위해 기존의 기계학습 기법 SVM과 주성분분석을 적용 유무를 가진 인공신경망과 비교하고자 한다.

Table 4는 삼성전자 주식 데이터셋에서의 성능 실험결과이다.

Table 4. Experiment for samsung dataset

	Accuracy	Precision	Sensitivity	F1-score
SVM	0.8076	0.6522	0.8076	0.7216
ANN	0.8041	0.7350	0.8041	0.7680
PCA(22)+ANN	0.8007	0.6853	0.8007	0.7384

Table 4에서 보면, 주성분분석을 이용하여 22개의 변수를 추출하고 인공신경망을 통해 결과를 낸 모형은 고려된 4가지 성능지표 모두에서 기존의 인공신경망보다 낮은 수치를 보여준다. 특히, 정밀도에서 기존의 SVM보다는 좋은 수치를 내고 있지만 기존의 인공신경망보다는 현저히 떨어지는 수치를 보여준다.

Table 5는 현대자 주식 데이터셋에서 성능 실험결과이다.

Table 5. Experiment for hyundai car dataset

	Accuracy	Precision	Sensitivity	F1-score
SVM	0.8522	0.7263	0.8522	0.7842
ANN	0.8488	0.7872	0.8488	0.8168
PCA(23)+ANN	0.8488	0.8043	0.8488	0.8260

Table 5에서 보면, 주성분분석을 이용하여 23개의 주성분을 이용하여 인공신경망 결과를 낸 모형이 삼성전자 주식 데이터셋에서 불균형 데이터임에도 불구하고 고려된 4가지 성능지표 모두에서 매우 높은 수치를 보여주는 반면에, 기존의 SVM은 정확도나 민감도에서는 높은 수치를 보이나 정밀도나 F1-score에서는 낮은 수치 즉, 0.7263과 0.7842를 보이는 것으로 나타났다. 또한, 기존의 인공신경망은 주성분분석을 통한 후 인공신경망에 적용했을 때보다 대부분 수치는 비슷하나 정밀도가 낮음을 확인할 수 있다.

5. 결론

주식 시장은 다양한 요인들과 복잡하게 상호 작용하여, 전통적인 분석 방법으로는 각각의 요인들이 주식 시장에 미치는 영향을 판단하기 매우 힘든 분야 중 하나이다. 하지만 머신 러닝의 발전으로 비록 각 요인이 가진 특성을 분석하는 데에는 한계가 있지만, 비교적 적은 비용으로 빠르게 변화하는 환경에 즉시 적용 가능한 예측 모형을 만들 수 있다는 장점이 있다.

각 분석에는 획득이 용이하고 단순한 재무 지표들을 활용하여 비교적 짧은 기간인 7일 이내의 삼성전자와 현대차의 주가 범위를 예측하는 모형을 SVM, 인공신경망, 주성분분석을 활용하여 만들보았으며, 각 모형들 간의 유의미한 차이가 있음을 확인하였다. 삼성전자의 경우 인공신경망만을 적용한 모형에서 정밀도(precision)가 가장 높게 산출되었으나, 현대차의 경우 주성분분석 이후 인공신경망을 적용한 모형에서 정밀도가 가장 높게 산출되었는데, 이는 향후에 더욱 상세히 다뤄볼 예정이다.

이번 모형 분석에서는 2019년 1월 3일부터 2022년 12월 5일 동안 발생한 재무 관련 지표를 기계학습에 활용하였으나, 해당 기간 동안의 거래일이 969일에 불과하여 기계학습의 활용도를 높이기 위해서 분석 데이터의 기간을 늘리고 라벨링을 보다 세분화하여 분류 클래스 간 데이터 블균형의 문제를 해결할 필요가 있다. 또한 기존에 사용된 전통적 재무지표 뿐만 아니라 이동평균선, 스토캐스틱(Stochastic), 볼린저밴드(Bollinger band) 등의 통계적 보조 지표를 독립변수에 추가하여 해당 모형의 예측성을 향상시키는 방향으로 모형을 개선·보완할 필요가 있다.