머신러닝을 활용한 주식 투자 시스템 구현

남기백*, 장정식**, 오훈*** 김태형****
*강릉원주대학교 정보통신공학과
**충북대학교 소프트웨어학과
***한국외국어대학교 경영학과
****NH 선물
e-mail : qweadg@naver.com

Development of Stock Investment System Using Machine Learning

Gibaek Nam*, Jeongsik Jang**, Hun Oh***, Taehyung Kim****
*Dept. of Information and Communication Engineering, Gangneung-Wonju National University
**Dept. of Computer Science, Chungbuk National University
*** Dept. of Business Administration, Hankuk University of Foreign Studies
**** NH FUTURES Co.,Ltd

요 약

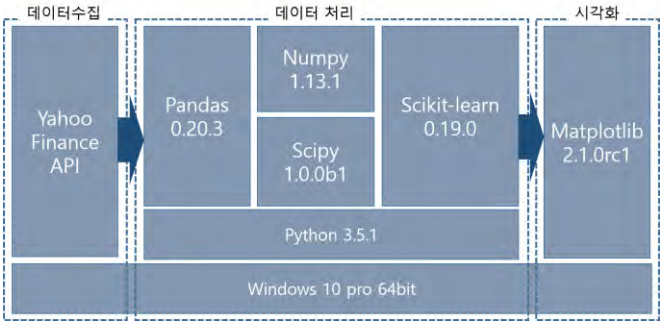
최근 기계학습에 대한 관심이 높아지면서 금융 분야에서는 인공지능을 이용하여 투자 포트폴리오를 제안하는 로보어드바이저(robo-advisor)를 출시하고 있다. 이는 고객에게 저렴한 수수료를 제공하며 높은 접근성, 인건비의 절감 등의 장점으로 이를 도입하여 다양한 상품을 개발하고 있다. 본 연구에서는 머신러닝 알고리즘인 SVM(support vector machine)과 kNN(k-nearest neighbor)을 활용하여 매월 12 개월 이전의 KOSPI 지수 데이터를 학습시킨 후 예측하는 투자 시스템을 구현하였다. 실험결과 SVM 이 2.90413 배의 성적으로 가장 우수했으며 수익률은 Precision(예측정확도)와 비례함을 보였다. 또한 수익곡선은 추세에 따라 유사한 형태를 보인 성과를 도출하였다.

1. 서론

이세돌과 알파고의 대국 이후 인공지능 열풍이 전 세계를 강타하고있다. 언론에서는 4 차 산업혁명을 이끌 것이라고 보도하며 구글, 페이스북, 아마존과 같은 공룡 IT 기업들은 인공지능에 사활을 걸고 천문학적인 자금을 쏟아 붓고있다. 이러한 열풍에 금융권에서도 개발중인 '투자 알파고' 라고 불리는 '로보어드바이저(robo-advisor)'는 저렴한 수수료, 높은 접근성 등의 장점으로 이를 도입하기 위해 다양한 협력사와 개발 및 제휴를 맺음으로써 다양한 금융상품을 개발하고 있다.

머신러닝은 기존 통계적 접근법에 비해 변수의 확률 분포에 대한 가정이나 조건이 약하거나 없는 경우도 있어 좀 더 유연한 접근 방법을 가지고 있다는 장점이 있다. 또한 이전에는 머신러닝을 위해 방대한 양의 데이터와 강력한 컴퓨팅 기술이 뒷받침되어야 했는데 최근에 비로소 이런 요소들의 진입장벽이 낮아져 이를 활용하면 대체로 통계적 기법보다 향상된 성능을 보여준다[1]. 하지만 데이터가 비슷할 경우 특정 분야에 쏠리는 현상이 생길 수 있고, 비정형적인 부분에 취약해 보험 등 종합적인 상황의 고려가 필요한 상품에서는 로보어드바이저를 제공하기 어려운 점이 있다[2]. 이러한 한계들을 보완하기 위해서 본 논문에서는 머신러닝 알고리즘을 활용하여 위기 상황과 대응방식에 대해 학습하는 시스템을 개발하고자 한다.

2. 시스템 구성



(그림 1) 시스템 워크플로우

본 시스템을 구성하는 개발환경 및 사용된 라이브러리는 (그림 1)과 같다. 전체적인 흐름은 최근 12 개월간의 주식정보 데이터를 수집한 후 파이썬 라이브러리와 머신러닝 알고리즘을 활용하여 수집된 데이터를 학습하고 모델을 만든다. 이후 학습된 모델에 적용하여 다음 달 주가를 예측하고, 마지막으로 전체적인 예측 정확도를 나타내고 이를 시각화한다.

머신러닝은 분산된 환경에서 유리한 성능을 보이지만 데이터가 크지 않고 python 의 경우 머신러닝을 위한 라이브러리 및 자료가 많으며 windows 환경을 지

원하기 때문에 위와 같이 구성하였다.

3. 데이터 수집 및 특징 추출

데이터 수집은 YAHOO API[3]를 사용하여 수집하였다. 실험에 사용할 데이터는 실제 주식환경에서 변동성이 많은 KOSPI 지수 데이터를 사용한다. 데이터는 2000 년 1 월부터 2017 년 9 월까지의 데이터를 수집하여 사용한다. 특징을 추출하는 방법으로 매달 최근 12 개월 데이터의 상승 추세 강도를 반영한 프랙탈(fractal)과 모멘텀(momentum)의 평균값과, 변동성을 특징으로 하며 다음달 수익률을 예측값으로 정의한다.

(그림 2) 프랙탈 구조를 보이는 주가

프랙탈은 부분의 모양이 전체의 모양을 닮는 자기유사성(self-similarity)을 가지면서 동일한 모양이 한없이 반복되는 순환성(recursiveness)을 보일 때, 같은 모양이 반복되어 전체를 형성하는 구조를 말한다. 주가에서 나타나는 프랙탈의 구조는 (그림 2)와 같은 형태를 보인다[4]. 모멘텀은 주식시장에서 주가 추세의 속도가 증가하고 있는지, 아니면 감소하고 있는지를 추세 운동량으로 측정하여 나타낸 지표의 뜻으로 사용된다. 이는 곡선의 한점 기울기를 계산한 후 그 변화를 선으로 그려 주가의 상승이나 하락의 강도를 미리 예측하는 기술적 분석 기법의 하나이다[5]. 변동성은 주식시장에서는 상품의 가격이 변동하는 정도를 뜻한다. 주식이나 채권, 통화 등의 시세가 비교적 일정한 방향성을 유지하면서 완만하게 움직이다가 갑자기 급등락할 경우 변동성이 확대됐다는 표현을 사용한다[6].

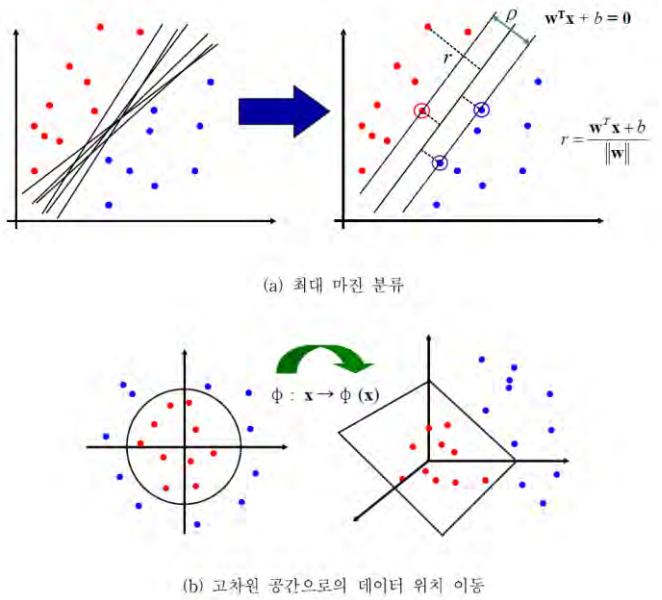
4. 분류 알고리즘

분류 알고리즘은 구분되지 않은 전체 데이터를 학습과정을 통해 비슷한 속성을 지니고 있는 데이터들로 나누는 방법을 익힌 후, 새로 받아들이는 데이터를 학습한 기준에 따라 분류하는 것을 의미한다. 본 논문에서는 분류 알고리즘 중 가장 대표적이면서도 널리 이용되는 kNN(k-nearest neighbor) 알고리즘과 SVM(support vector machine)을 사용하였다.

kNN 알고리즘은 새로 받아들인 데이터를 판단해야 할 경우, 기존에 학습한 데이터와 비교하여 거리상 가장 가까운 거리의 데이터의 분류값으로 받아들인다. 데이터들의 거리를 구할 때 다음과 같은 유클리디안 거리(Euclidean Distance)(식 1)를 사용한다.

$$\mathrm{dist}(p,q) = \sqrt{(p_1-p_1)^2+(p_2-p_2)^2+..+(p_n-p_n)^2} \quad (1)$$

SVM 은 이질적인 서로 다른 집단간의 상대적인 거리를 최대화 할 수 있는 기준면을 통해 분류한다. 두 속성의 데이터 중에서 서로 가장 바깥쪽에 위치하고 있는 경계 점을 support vector 라고 하고 이들 간의 거리를 최대화할 수 있는 경계선을 support vector machine 이라고 한다.



(a) 최대 마진 분류

(b) 고차원 공간으로의 데이터 위치 이동

(그림 3) SVM 작동원리

데이터를 한 직선으로 구분하기 힘들 때 고차원 공간으로 데이터의 위치를 이동하여 분류하기도 하는데 이 경우 커널 함수(kernel function)를 사용한다. 커널 함수 중 가장 대표적으로 사용되는 RBF(Radial Basic Function kernel)이며 수식은 다음과 같다.

$$k(x^{(i)}, x^{(j)}) = e^{-r\Sigma(x^{(i)},x^{(j)})^2} \quad (2)$$

SVM 은 명백한 이론적 근거에 기반을 두므로 결과 해석이 용이하고 적은 학습 데이터만으로도 효과가 있으며 인공 신경망 수준의 성능을 보인다는 장점이 있어 광범위하게 이용된다.

5. 실험계획

본 연구에서는 2000 년 1 월 1 일부터 매달 최근 12 개월의 KOSPI 지수를 수집하여 프랙탈, 모멘텀, 변동성의 특징을 추출한다. 이후 기계학습 방법론인 kNN, SVM 알고리즘을 활용한 모델을 만들고 다음달의 지수 상승 또는 하락에 대한 예측하여 자산을 자동으로 분배하는 시스템의 성능을 실험한다. 각 지수들의 다음 달 주가를 예측하여 수익이 플러스일 것으로 분류되면 주식 보유, 마이너스일 것으로 예측되면 현금 보유 전략을 취해 매월 이렇게 나온 수익 곡선과 현금을 1:1 로 리밸런싱한다. 현금은 연 2%의 수익을 기준으로 계산한다. 각 알고리즘의 파라미터 최적화를 위해 kNN 의 k 값에 따른 정확도를 측정하고 SVM 은 RBF 커널의 정규화 관련 인자인 C 값을 달리하여 이에 따른 정확도의 변화를 분석한다.

6. 결과

본 절에서는 앞서 설명한 방법을 통해 학습 모델의 성능을 검증한다. 각 알고리즘들의 최적의 파라미터를 찾기 위해 모델을 만들어 파라미터에 따른 성능을 비교한다. 학습된 모델의 성능측정은 예측오류를 고려하여 Accuracy, Precision, Recall, F1 score 를 측정하며 최종 수익률은 Result 에 기록한다.

<표 1> 예측값과 실제값 비교 용어표

Actual / Predict	True	False
True	True Positive(TP)	False Positive(FP)
False	False Negative(FN)	True Negative(TN)

F1 score 는 precision 과 recall 두 가지 지수를 통계적으로 종합하여 주어진다. Recall 과 Precision 에 대한 식은 다음과 같으며 식의 약어들은 <표 1>에 정의한다.

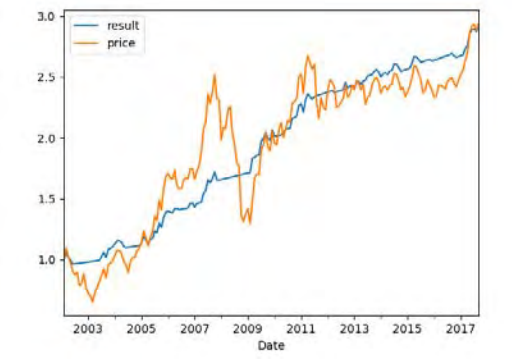
$$\text{Precision} = \frac{tp}{tp + fp} \quad (1)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (2)$$

F1 score 에 대한 식은 다음과 같다.

$$\text{F1 score} = 2 * \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} \quad (3)$$

(그림 4) kNN 알고리즘(k=2) 수익곡선



(그림 5) SVM 알고리즘 (C=70) 수익곡선

<표 2> kNN k 의 개수에 따른 결과

k	Accuracy	Precision	Recall	F1 score	Result
2	0.51064	**0.61111**	0.40741	0.48889	**2.79462**
3	0.52128	0.58333	0.58333	0.58333	2.21438
4	0.51596	0.60000	0.47222	0.52850	2.43758
5	0.51596	0.57658	0.59259	0.58447	2.26037
6	0.48936	0.56667	0.47222	0.51515	2.12833

<표 3> SVM 파라미터 C 에 따른 결과

C	Accuracy	Precision	Recall	F1 score	Result
40	0.50532	0.57282	0.54630	0.55924	2.53658
50	0.51064	0.57692	0.55556	0.56604	2.61560
60	0.51064	0.57692	0.55556	0.56604	2.61560
70	0.51596	**0.58252**	0.55556	0.56872	**2.90413**
80	0.52128	0.58054	0.56481	0.57547	2.81695

실험결과 kNN 은 k 가 2 일 때 2.79462, SVM 은 파라미터 C 가 70 일 때 2.90413 으로 SVM 의 성능이 가장 높았으며 각 알고리즘의 수익곡선 결과는 (그림 4)(그림 5)와 같다. <표 2>, <표 3>의 결과 두 알고리즘 모두 Precision 과 비례하였고 수익률은 예측정확도와 비례한다는 사실을 알 수 있었다. 전반적인 분류 정확도가 50~60%정도로 나타났지만 모멘텀을 반영하는 지표를 특징으로 선정하였기 때문에 수익 곡선도 추세에 따라 유사한 형태를 보여주는 것을 확인하였다.

앞으로의 연구에서는 독립변수를 어떻게 선정하고 어떤 종속 변수를 이용하고 알고리즘의 세부적인 파라미터의 설정 방법에 따라 무궁무진한 변형이 가능하고 본 논문에서 소개한 기본적인 전략과도 결합시키면 다양한 응용이 가능할 것으로 기대한다.

※본 논문은 2017 년 한이음 ICT 멘토링 프로젝트의 결과물입니다.

참고문헌

[1] 고윤승. "우리나라 로보어드바이저 도입을 위한 활성화 방안 탐색." 한국과학예술포럼, 25 (2016. 9): 19-33.

[2] 박재연; 유재필; 신현준. 로보어드바이저를 이용한 포트폴리오 관리. *정보기술아키텍처연구*, 2016, 13.3: 467-476.

[3] Yahoo Finance https://finance.yahoo.com/

[4] MANDELBROT, Benoit B.; PIGNONI, Roberto. *The fractal geometry of nature*. New York: WH freeman, 1983.

[5] FAMA, Eugene F.; FRENCH, Kenneth R. Size, value, and momentum in international stock returns. *Journal of financial economics*, 2012, 105.3: 457-472.

[6] GROSSMAN, Sanford J.; SHILLER, Robert J. The determinants of the variability of stock market prices. 1980.

[7] 김선웅; 안현철. Support Vector Machines 와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발. *지능정보연구*, 2010, 16.1: 71-92.