베이지안 신경망을 활용한 주가상승 예측과 영향변수 분석

Stock market anaylsis via Bayesian Neural Network

저자 (Authors)
장수연, 신규용
Suyeon Jang, Kyuyong Shin

출처 (Source)
한국정보과학회 학술발표논문집 , 2020.7, 1652-1654 (3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874877

APA Style
장수연, 신규용 (2020). 베이지안 신경망을 활용한 주가상승 예측과 영향변수 분석. 한국정보과학회 학술발표논문집, 1652-1654.

베이지안 신경망을 활용한 주가상승 예측과 영향변수 분석

장수연, 신규용
성균관대학교 통계학과, 홍익대학교 도시공학과
{jangsuyeon486, p37329}@gmail.com

Stock market anaylsis via Bayesian Neural Network

Suyeon Jang, Kyuyong Shin
University of Sungkyunkwan, University of Hongik

요 약

이 논문은 주가상승 예측 알고리즘과 주식시장 분석에 쓰이는 다양한 변수들의 영향도 파악, 그리고 예측 모델의 예측 신뢰도 측정 방법에 목적을 두고 있다. 기존의 논문들이 이러한 부분에 초점을 두고 변수 영향도 측정과 모델 신뢰도 파악을 실행하고 있었지만, 인공신경망(Artificial Neural Network) 기반의 분석은 미미했다. 이 논문에서는 베이지안 신경망기반(Bayesian Neural Networks)의 주가상승 예측 모델을 모델링하고, 이를 이용해 모델 불확실성(uncertainty)을 측정한다. 더불어 베이지안 신경망에 적용될 수 있는 영향변수 파악 방법 RelAtive Centrality Measure(이하 RATE)를 활용한 주식시장 변수 분석을 진행한다.

1. 서 론

주가 예측과 그에 이용되는 변수 영향력 평가는 기존에 많은 연구들에서 중요하게 다뤄지고 있다. 이에 과거부터 현재까지 다양한 기계학습(Machine learning) 방법이 고르게 활용되었으며, 변수로는 각 기업의 경영 성과, 재무제표, 그리고 거시적 경제변수들이 사용되어왔다. [1]에서는 주식시장에 미치는 경제적 요인을 강조, 거시경제 변수들과 주가 간 관계를 공적분 벡터를 이용해 정립하였으며, 이들 상관관계를 토대로 벡터오차수정모형을 통해 주가를 예측하였다. [2]에서는 한국 증시 시장에 영향을 줄 수 있는 거시경제 변수들을 활용해 분위 회귀 분석으로 주가를 예측했으며, 마찬가지로 이 중 주가에 가장 크게 영향을 주는 경제지표를 분석했다.

그러나 최근 컴퓨터 기술의 발달로 딥러닝(Deep leraning)이 대두되며 다양한 기법들이 주식시장을 분석하는데 활발히 이용되고 있다. 이러한 모델들은 기본적으로 sequence to sequence 데이터에 대해서 효과적인 inductive bias를 가정하고 있으며, 이를 통해 time series classification, stock market prediction 등에서 우수한 성과를 보였다. 하지만 딥러닝의 인상적인 성능에도 불구하고, 설명력이 떨어진다는 단점이 산업 전반에서 이들의 취약점 중 하나로 지적됐다.

이에 이 논문에서는, 베이지안 신경망(이하 BNNs)을 활용하여 주가상승 예측과 변수 분석을 진행한다. BNNs 모델은 통계모델이 지니고 있는 예측 신뢰도검정과 딥러닝 모델의 우수한 예측 정확도 두 가지 특성을 모두 활용할 수 있게 해준다. 더불어 이를 기반으로, 최근에 제안돼 여러 분야의 데이터셋에서 우수한 설명력을 입증한 RATE [3]를 활용해 주가상승 예측에 영향을 준 변수 영향력 평가를 진행한다.

이 논문의 중요 기여는 다음과 같다.

- BNNs을 활용하여 기존 통계모델들의 부족한 예측 정확도를 보완하고, 딥러닝 모델의 예측 신뢰도 검정이 불가능하다는 점을 해결한다.
- 최근에 제안된 RATE를 활용하여 주가상승 예측에 영향을 준 변수들이 무엇인지 각 종목별로 파악하고, 이를 해석한다.
- Embedding visualization을 통하여 각 종목별 군집을 확인하고, 이를 해석한다.

2. 선행 연구

2.1 주가상승 예측 및 영향변수 분석

[4]와 같은 연구들은 기존의 기계학습 모델을 수정하여 주식시장 모델링을 시도하고 있지만, 최근의 많은 딥러닝 주식시장 예측 모델에 비해 뒤떨어지는 성능을 보인다. 그럼에도 불구하고, 특히 전처리나 변수 해석에서는 아직까지 기존의 기계학습이나 통계적 방법론이 지배적인데, 이는 딥러닝의 태생적 한계와 맞물려 기존 통계모델이 우수한 설명력과 정밀한 수학적 모델링을 기반으로 한다는 점에 기인한다.

하지만, 주가상승 예측 모델 성능에 의존하는 대부분의 영향변수 분석 연구의 경우, 예측 모델인 통계 모델 자체의 성능이 미흡하다는 점은 간과할 수 없는 부분이다. [2]에서 실시한 분위 회귀 분석 방법이나 [4]의 군집 모형이 그러한 예시

이다. 이들 모델의 경우 결국 데이터 특성에 따라 일반화가 쉽지 않다는 점과, 비선형 데이터를 다룰 때 성능이 떨어진다는 단점이 존재한다.

2.2 베이지안 신경망 (Bayesian Neural Network)

베이지안 신경망은 기본적으로 사후확률을 활용하는 베이지안 회귀분석(Bayesian regression)의 아이디어를 따른다. 베이지안 회귀분석의 구체적 수식은 아래와 같다.

$$p(y^*|x^*, D) = \int_w p(y^*|x^*, w)\, p(w|D) dw \qquad (1)$$

베이지안 신경망 모델의 경우 θ로 파라미터화 된 새로운 $q_\theta(w)$를 사후확률로 정의하여 사후확률 $p(w|D)$를 추정하는 방법을 사용한다. 이 논문에서는 이러한 추정방법 중, 현재 가장 많이 쓰이는 변분 추론(Variational Inference)을 사용한다. 자세한 수식은 아래와 같다.

$$\begin{aligned} KL(q_\theta(w)||p(w|D)) &= \int_w q_\theta(w) \log \frac{q_\theta(w)}{p(w|D)} dw \\ &= \int_w q_\theta(w) \log \frac{q_\theta(w)p(D)}{p(D|w)p(w)} dw \\ &= E_{q_\theta(w)}\left[\log \frac{q_\theta(w)}{p(w)} - \log p(D|w) + \log p(D)\right] \qquad (2) \\ &= KL(q_\theta(w)||p(w)) - E_{q_\theta(w)}[\log p(D|w)] + C, \\ ELBO &= E_{q_\theta(w)}[\log p(D|w)] - KL(q_\theta(w)||p(w)) \end{aligned}$$

결과적으로, $ELBO$를 최소화시키는 $q_\theta(w)$는 우리가 찾는 사후확률 분포라고 할 수 있다. 이 논문에서는 $q_\theta(w)$를 활용하여 베이지안 신경망의 전파를 진행하며, 이후 이 파라미터가 확률변수로 설정돼 있음을 이용, 모델의 예측 신뢰도를 계산한다.

2.3 Relative Centrality Measure (RATE)

RATE는 black-box 모델에 대해 전역적인 변수 영향력을 산출하는 방법이다. 이는 Bioinformatic, Recommendation등 여러 분야의 데이터셋에서 안정적이고 인상적인 결과를 보였다. 아래는 하나의 독립변수에 해당하는 가중치 파라미터 W_j에 대한 RATE 계산식이다.

$$\begin{aligned} RATE(W_j) &= KLD(W_j) / \sum_l KLD(W_j), \\ KLD(W_j) &\triangleq KL(p(W_{-j})||p(W_{-j}|W_j = 0)) \qquad (3) \end{aligned}$$

$KLD(W_j)$는 W_j가 0으로 설정돼 있을 때 W_j를 제외한 나머지 파라미터, 즉 W_{-j}의 분포가 기존의 W_{-j}분포와 비교해서 얼마나 달라지는지 측정하는 식이다. 그리고 최종적 $RATE$값은 이러한 KLD 식을 모든 W에 대해 계산하여 표준화한 값이다. 이 연구에서는 주식시장 독립변수 454개의 영향 정도를 파악, 더 가치 있는 변수가 무엇인지 알아본다.

3. 방법론

3.1 데이터셋 및 데이터 전처리

이 연구에선 크롤링한 데이터를 사용하였다. 사용한 입력 데이터셋으로는 지수지가, 시가총액, 자본금, 배당 수익율 등의 한국증시를 나타내는 30여개의 변수와 전반적인 세계 경제 상황을 나타내는 세계 각국의 생산자 물가, 본원통화, 정책금리 등 400개의 변수를 선정했다. 더불어 한국과 교류가 많은 국가의 환율 역시 변수로 추가하였으며, 실물자산 지표로는 금 가격, 원유 가격, 구리 가격 등의 변수를 사용하였다. 종속변수로는 제조업, 섬유의복, 의약품 등 총 13개 주식종목을 사용하였으며, 데이터셋의 기간은 2005년 1월부터 2018년 5월까지로 정하였다.

데이터 크롤링 과정에서 결측치가 발생하여, 다중대치를 통해 데이터 결측치 처리를 해주었다. 이 과정에서 우리는 missing at random(MAR)가정과 predictive mean matching(PMM) 방법론을 사용하여 다중대치를 진행하였다.

3.2 알고리즘 및 트레이닝

모델의 경우 추천모델의 아키텍쳐를 따른다. 따라서 입력 값의 마지막 단 hidden vector와 추천 종목의 embedding vector을 내적 하여 logit을 산출한다. 이는 해당 날짜의 주식시장의 여러 지표 값들을 보고 오늘 어떤 종목을 사야 하는지 추천하는 알고리즘으로, 수식으로 나타내면 아래와 같다.

$$\begin{aligned} p_{d,i} &= sigmoid(h_d^T z_i), \\ h_d &= \sigma(w_1^T \sigma(w_0^T x_d + b_0) + b_1), \qquad w_1^T \sim \mathcal{N}(0, I) \qquad (4) \end{aligned}$$

이때 모델 마지막단 hiddnen vector h는 BNNs을 통과한 결과값으로써, 이후 모델의 불확실성과 RATE값을 측정하는데 사용된다. 추천할 종목의 embedding vector z는 모델을 충분히 학습한 이후 그림 2와 같이 embedding space 시각화에 사용된다.

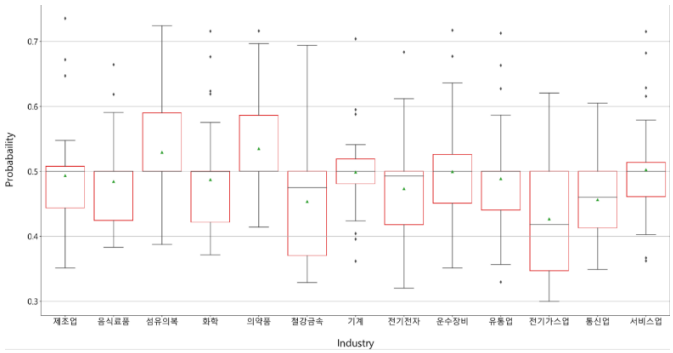

그림 1: 특정한 날 모델의 예측 신뢰수준을 시각화한 그림이다. 빨간색 박스 안 검은 선은 중위 값을 의미하며, 초록 점은 평균값, 바깥의 검은 점들은 이상치이다.

4. 결과

4.1 종목추천 및 변수분석

표 1: 주가상승 예측에 대한 방법론별 F1-score.

Methods	F1-micro score
Random Forest	0.548
Decision Tree	0.542
KNN Classifier	0.523
Ours	0.597

표 1 에서 확인할 수 있듯 대조군인 여타 기계학습 방법들에 비해 우수한 성능을 보임을 알 수 있다. 더불어 그림 1에서 확인할 수 있듯, 우리 모델은 예측 신뢰도 검정이 가능하다. 그림 1의 결과는 특정한 날의 모델 예측 신뢰도를 시각화한 것으로, 주식 투자 전략을 세울 때 도움을 줄 수 있는 자료이다. 예를 들어 해당 날짜의 주식시장 지표를 보고 내일의 투자전략을 세운다고 할 때, 그림 1의 경우 의약품과 섬유의복에 투자하는 것이 가장 안정적인 결과를 보일 것이라고 예상할 수 있다. 이는 투자 리스크를 줄여주는데 도움이 된다.

표 2: 각 주식종목별 중요변수 (쪽수 제한으로 일부만 표시)

	Rank1	Rank2	Rank3	Rank4	Rank5
화학	배당 수익율	통화량 (중국)	본원통화 (미국)	미국 정책금리	미국 채권 1년
서비스업	배당 수익율	산업생산 (캐나다)	수입국 (파라과이)	소비자물가(베트남)	전력,가스및수도 (총지수)

표 2는 RATE를 이용해, 모델이 예측을 하는데 가장 영향을 많이 준 영향변수를 해당 주식 종목별 순서대로 나열한 것이다. 이는 경제적 영향력이 높은 국가의 경제정책 조정이 상호의존도가 높은 다른 국가의 실물 및 금융지표 변동에 어떠한 영향을 끼치는지 알아볼 수 있는 자료이다. 가장 먼저, 표 2에 산업별 중요 변수를 보면 대체적으로 한국증시와 상호의존이 높은 나라를 알 수 있다. 다양한 국가가 있지만 그 중 미국이 가장 많은 영향력을 주는 나라인 것을 확인할 수 있다.

화학과 서비스업은 배당 수익률에 제일 큰 영향을 받는다. 이는 배당평가 모형으로 설명할 수 있다. 배당평가 모형은 주식 가격 결정이론으로 주가와 거시 경제 변수들과의 관계를 적절하게 나타낼 수 있는 대표적인 모형이다. 이 모형에 따르면, 주가는 주식을 소유함으로써 획득할 수 있다고 기대되는 가치로, 미래의 현금흐름을 적절한 할인율로 할인한 값이다 [4].

4.2 Embedding Visualization

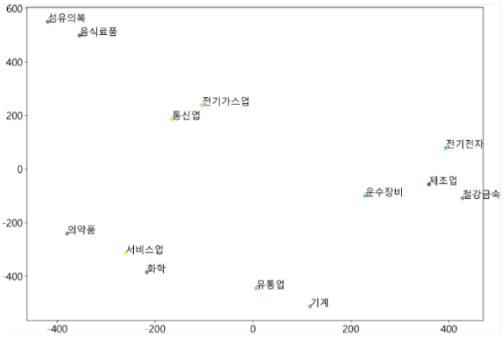

그림 2: 13개 투자 종목에 대해 t-SNE를 사용해 embedding space를 시각화한 그림이다.

표 2와 그림 2에서 보이듯, 거리가 가까운 산업군끼리 영향을 받는 변수들이 공통되는 것을 확인할 수 있다. 섬유의복과 음식료품의 경우 미국 정책금리, 미국채권금리, 소비자물가 등 네 개의 변수를 공통적으로 고려하고 있다. 또 전기가스업과 통신업은 미국의 본원통화, 룩셈부르크의 외환보유액 두개의 변수를 공통적으로 포함하며, 전기전자, 제조업, 운수장비, 철강금속 산업도 중요 변수로 미국의 본원통화를 공통적으로 고려하고 있는 것을 확인할 수 있다. 우리는 embedding visualization를 통해서 어떤 주식종목군이 비슷한 성질을 지니고 있는지 확인할 수 있었다.

5. 결론

기존 딥러닝 주가예측 모형의 우수한 예측 성능에도 불구하고, 딥러닝 모델은 예측 신뢰수준 파악이나 영향변수 분석에서 통계적 모델의 이점을 취하지 못했다. 따라서 우리는 이것을 해결할 수 있는 BNNs의 활용을 제시하고, 최근 가장 우수한 변수분석 방법론인 RATE을 이용해 이 문제를 해결했다. 우리 방법론은 투자자가 리스크를 고려할 수 있게 하며, 좋은 입력 변수 선택을 가능케 해 해석 가능하고 우수한 성능의 딥러닝 모델 구축에 도움을 준다. 이는 논문의 다양한 실험 결과를 통해 확인할 수 있었다.

6. 참고 문헌

[1] 김성희, “주식가격과 거시경제변수의 관계에 대한 분석: 공적 분과 vecm 모형을 중심으로,” 2002.

[2] 이소영, “거시경제변수와 주식수익률의 관계분석,” 2013.

[3] J. Ish-Horowicz, D. Udwin, S. Flaxman, S. Filippi, and L. Crawford, “Interpreting deep neural networks through variable importance,” arXiv preprint arXiv:1901.09839, 2019.

[4] R. Luss and A. d’Aspremont, “Clustering and feature selection using sparse principal component analysis,” Optimization and Engineering, 145-157, 2010.