ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 September: 25(3): 1~17 http://dx.doi.org/10.13088/jiis.2019.25.3.001

시스템적인 군집 확인과 뉴스를 이용한 주가 예측

성노윤
한국과학기술원 경영공학부
(*nyseong@kaist.ac.kr*)

남기환
한국과학기술원 경영공학부
(*namkh@kaist.ac.kr*)

빅데이터 시대에 정보의 양이 급증하고, 그중 많은 부분을 차지하는 문자열 정보를 정량화하여 의미를 찾아낼 수 있는 인공지능 방법론이 함께 발전하면서, 텍스트 마이닝을 통해 주가 예측에 적용해 온라인 뉴스로 주가를 예측하려는 시도가 다양해지고 있다. 이러한 주가 예측의 방법은 대개 예측하고자 하는 기업의 뉴스로 주가를 예측하는 방식이다. 하지만 특정 회사의 뉴스만이 그 회사의 주가에 영향을 주는 것이 아니라, 그 회사와 관련성이 높은 회사들의 뉴스 또한 주가에 영향을 줄 수 있다. 그러나 관련성이 높은 기업을 찾는 것은 시장 전반의 공통적인 영향과 무작위 신호 때문에 쉽지 않다. 따라서 기존 연구들은 주로 미리 정해진 국제 산업 분류 표준에 기반을 둬 관련성이 높은 기업을 찾았다. 하지만 최근 연구에 따르면, 국제 산업 분류 표준은 섹터에 따라 동질성이 다르며, 동질성이 낮은 섹터는 그들을 모두 함께 고려하여 주가를 예측하는 것이 성능에 악영향을 줄 수 있다는 한계점을 가진다.

이러한 한계점을 극복하기 위해, 본 논문에서는 주가 예측 연구에서 처음으로 경제물리학에서 주로 사용되는 무작위 행렬 이론을 사용하여 시장 전반 효과와 무작위 신호를 제거하고 군집 분석을 시행하여 관련성이 높은 회사를 찾는 방법을 제시하였다. 또한, 이를 기반으로 관련성이 높은 회사의 뉴스를 함께 고려하며 다중 커널 학습을 사용하는 인공지능 모형을 제시한다. 본 논문의 결과는 무작위 행렬 이론을 통해 시장 전반의 효과와 무작위 신호를 제거하여 정확한 상관 계수를 찾아 군집 분석을 시행한다면 기존 연구보다 더 좋은 성능을 보여준다는 것을 보여준다.

주제어 : 온라인 뉴스, 주가 예측, 무작위 행렬 이론, 계층적 군집 분석

논문접수일 : 2019년 5월 17일 논문수정일 : 2019년 8월 31일 게재확정일 : 2019년 9월 19일
원고유형 : 일반논문 교신저자 : 남기환

1. 서론

빅데이터 시대에 정보의 양이 급증하고, 그것을 정량화하여 의미를 찾아낼 수 있는 인공지능 방법론이 함께 발전하면서, 그러한 기술을 주가 예측에 적용해 온라인 뉴스로 주가를 예측하려는 시도가 다양해지고 있다 (Seong and Nam, 2017; Seong and Nam, 2018). 이러한 주가 예측의 방법은 대개 특정 기업에 관련된 뉴스가 나오면 그 뉴스를 이용해 특정 기업의 주가의 방향성을 예측하는 것이다.

하지만 최근 연구에 따르면, 단순히 특정 회사의 뉴스만이 그 회사의 주가에 영향을 주는 것이 아니라, 그 회사와 관련성이 높은 회사들의 뉴스 또한 주가에 영향을 줄 수 있다 (Nam and Seong, 2019; Seong and Nam, 2018; Shynkevich et al., 2016). 그러나 관련성이 높은 기업을 찾기 쉽지 않다. 따라서, 기존 연구들은 주로 국제 산업 분

류 표준에 기반을 둬 관련성이 높은 기업을 찾았다 (Shynkevich et al., 2016). 하지만 연구에 따르면, 국제 산업 분류 표준은 섹터에 따라 동질성이 다르며, 동질성이 낮은 섹터는 그들을 모두 함께 고려하여 주가를 예측하는 것이 성능에 악영향을 줄 수 있다 (Seong and Nam, 2018). 즉, 관련성이 높은 회사를 정확하게 찾는 것은 중요한 문제이다.

관련성이 높은 회사를 정량적으로 찾는 방식은 주식 시장에 잡음이 많아서 어려운 것으로 알려졌다 (Bun et al., 2017). 따라서 기존 연구에서는 시장의 잡음을 제거하여 주가 예측에 사용하려는 연구가 존재하지 않았다. 이러한 연구의 빈틈을 메우기 위해, 본 논문에서는 주가 예측 연구에서 처음으로 경제물리학에서 주로 사용되는 무작위 행렬 이론을 사용하여 잡음을 제거하고 군집 분석을 시행하여 관련성이 높은 회사를 찾는 방법을 제시하고, 이를 인공지능 방법과 결합하는 방법을 제안한다.

뉴스를 통해 주가 예측을 하는 방법은 크게 사전 처리와 기계 학습이 있다. 본 논문에서는 사전 처리로는 단어 주머니 모형으로 뉴스를 숫자 벡터로 바꾸어 주며, 카이스퀘어 방법으로 필요한 단어만을 선별하였으며, TF-IDF 방법으로 가중치를 주었다. 기계 학습 방법으로는 본 논문에서는 특정 회사의 주가를 예측할 때, 특정 회사의 뉴스만이 아니라, 관련된 회사들의 뉴스 또한 함께 고려하는 방법을 사용하기 때문에, 여러 특성을 같이 포함할 수 있는 기계 학습 방법인 다중 학습 커널을 사용하였다.

본 논문의 결과는 다음과 같다. (1) 관련성이 높은 기업의 뉴스를 이용하여 주가를 예측하는 것은 효과적인 방법이라는 것을 기존 연구 흐름에 이어 확인하였다. (2) 관련 있는 기업을 찾을 때, 잘못된 방식으로 찾는다면 인공지능 예측 성능을 저하할 수 있다. (3) 무작위 행렬 이론을 통해 시장 전반의 효과와 무작위 신호를 제거하여 정확한 상관 계수를 찾아 군집 분석을 시행한다면 기존 연구보다 더 좋은 성능을 보여줄 수 있다.

본 연구의 기여는 다음과 같다. 첫 번째, 본 연구는 경제물리학에서 주로 사용되던 무작위 행렬 이론이 인공지능과 결합하면 좋은 방법론을 만들어낼 수 있다는 것을 보여주며 단순히 인공지능 알고리즘만을 발전시키는 것이 아닌 물리학 이론을 차용하여 발전시키는 것이 중요함을 시사한다. 이는 이전 엔트로피를 통한 복잡계 형성과 인공지능을 통합하여 방법론을 제시한 Nam and Seong (2019)의 연구를 확장한다. 두 번째, 본 연구는 주식 시장에서 관련성이 높은 기업을 정확하게 찾는 것이 중요한 문제임을 다시 한 번 강조하며, 인공지능 알고리즘을 연구하는 것만이 중요한 것이 아니라 입력 값을 어떻게 이론적으로 조절하는 것이 필요한 것인 지를 입증하였다.

본 논문의 뒷부분은 다음과 같이 구성된다. 2장에서는 기존 연구 논문들에 관해 서술한다. 3장에서는 무작위 행렬 이론을 이용한 금융 뉴스로 주가의 방향성을 예측하는 연구 모델을 제시한다. 4장에서는 실험 결과에 관해 설명할 것이다. 5장에서는 본 연구에 관해 결론을 내며, 한계점과 후속 연구를 위한 지침에 관해 설명할 것이다.

2. 문헌 연구

2.1 유사 기업을 고려한 주가 예측

Shynkevich et al. (2016)은 높은 관련성을 가지는 기업들의 뉴스를 통합하여 주가를 예측하는

시스템을 만들었다. 저자는 국제 산업 분류 표준 (GICS) 기반으로 관련성 있는 회사를 정의하였다. 관련 있는 회사들의 뉴스를 함께 이용하여 주가를 예측하는 경우 특정 주식에 관련된 뉴스만으로 주가를 예측하는 것보다 뛰어난 결과를 보여주었다. 하지만 저자들은 같은 GICS 섹터 체계에 있으면 관련성이 높을 것이라는 가정을 하였다. 하지만 실제로는 같은 업종에 있다고 하더라도 관련성이 높지 않을 수 있다는 한계점을 가진다.

Seong and Nam (2018)은 Shynkevich et al. (2016)의 같은 GICS 섹터 안에 있더라도 모두 관련성이 높지 않을 수 있다는 한계점을 보완하기 위해, GICS 섹터를 k-평균 군집 분석하여 관련성이 높은 회사를 고르는 방법을 고안하였다. 그 결과, 단순히 GICS 섹터 안에 있는 기업들을 관련성이 있는 회사라고 가정하여 주가를 예측하는 데 사용하는 것 보다, 군집 분석을 통해 관련성이 높은 회사들을 찾는 것이 중요하다는 것을 제시하였다.

실제로 주식 시장에서는, 대기업에 관련된 뉴스가 나오면 그 하청 업체의 주가는 영향을 받지만, 그 역 관계는 쉽게 성립하지 않는다. 이러한 점을 반영하기 위해, Nam and Seong (2019)는 실제 주식 시장은 관련성이 높은 것이 중요한 것이 아니라 인과 관계가 중요함을 경제 물리학과 인공지능 방법론을 통해 보여주었다. 저자들은 이전 엔트로피 (Transfer entropy)를 회사의 주가 사이의 인과 관계를 찾고, 그 인과 관계가 실제로 뉴스를 통해 주식을 예측할 때 유용함을 보여주었다. 위의 세 가지 연구가 모두 제시하는 바는 실제로 특정 회사에 관련된 뉴스가 나오면 관련 있는 특정 회사뿐만 아니라 회사의 주가도 함께 움직인다는 것이며, 관련된 회사를 어떻게 찾는지가 중요한 지이다. 본 논문에서는 이러한 연구 흐름을 따라, 관련된 회사를 찾는 방법을 무작위 행렬 이론을 통해 제시한다.

2.2 뉴스 사전 처리

기계 학습 방법들은 문자열 데이터를 직접 계산할 수 없으므로, 뉴스 정보를 받게 되면 문자열 데이터를 숫자로 이루어진 벡터로 변환해줄 필요가 있다. 이 문자열 데이터를 숫자 벡터로 데이터로 바꾸어 주는 것을 문자열 사전 처리 (text preprocessing) 라고 한다. 뉴스 데이터에서 주가에 영향을 미치는 변수를 추출하는 방법 중에서 가장 유명한 것은 크게 3가지 단계를 가진다: 변수 추출 (feature extraction), 변수 선택 (feature selection), 변수 표현 (feature representation) (Hagenau et al., 2013; Nam and Seong, 2019; Shynkevich et al., 2016).

변수 추출은 문자열을 숫자로 변화하여 변수들을 생성하는 과정이며, 뉴스에서 단어들과 그 조합을 추출하는 과정이다. 다양한 방법이 변수 추출 방법으로 사용되고 있지만, 기존 뉴스를 이용한 주가 예측 연구 흐름에서 가장 많이 사용되는 것은 단어 주머니 방법 (Bag of words)이다 (Groth and Muntermann, 2011; Hagenau et al., 2013; Nam and Seong, 2019). 단어 주머니 방법은 모든 단어를 형태소 분석을 하여 모든 단어를 원형으로 만들어, 그것들의 개수를 세서 이를 벡터로 표현한다. 본 논문에서는 기존 연구의 흐름을 따라 단어 주머니 모형을 사용한다.

뉴스 전반에서 나오는 단어들은 주가의 방향성에 영향을 미치지 않으며, 특정 뉴스에서 나올 때마다 주가의 방향성이 변할 때, 그중 핵심 단어들이 영향력이 있는 변수들이다. 영향력이

높은 변수들을 선택하기 위해 변수 선택을 한다. 즉, 변수 선택은 단어 주머니 모형에서 찾아낸 수많은 변수 중에서 주가에 영향을 미치는 것들을 골라내는 것이다. 변수 선택에 가장 많이 사용되는 방법은 카이 스퀘어 방법이다 (Shynkevich et al., 2016). 카이스퀘어 변수 선택은 카이 제곱 분포를 사용하여 변수의 영향력을 평가한다. 모든 단어의 기대되는 빈도는 같으므로, 관찰된 빈도 O_{ij}가 기대되는 빈도 E_{ij}와 의미있게 다른지 본다면 영향력을 평가할 수 있다. 즉, 카이스퀘어 값이 클수록 영향력이 있는 변수이며, 이를 사용하여야 한다. 이는 수식으로 (1)과 같이 표현된다.

$$\aleph^2 = \frac{\sum(O_{ij}-E_{ij})}{E_{ij}} \quad (1)$$

본 논문에서는 기존 연구의 흐름에 따라 변수 선택으로 카이 스퀘어 방법을 사용한다.

변수 추출과 변수 선택을 하면 문자열을 숫자로 바꾸고 영향력이 높은 변수들을 선택한 결과가 나온다. 하지만, 만약에 특정 단어가 자주 나올 때, 주가가 많이 오르거나, 주가가 많이 내리면 그 변수는 가중치를 주어야 한다. 또한, 항상 나오는 것이 아닌 전체 문서에서 적게 나올수록 더 의미 있는 변수이기 때문에, 이 점을 반영하여야 한다. 이때 이를 반영하기 위해 기존 연구에서 주로 사용되는 변수 표현 방식은 TF-IDF 방식이다 (Nam and Seong, 2019; Seong and Nam, 2018). 따라서, 본 논문에서는 기존 연구의 흐름에 따라 변수 표현으로 TF-IDF를 사용한다.

2.3 기계 학습 방법론

뉴스에서 특징 벡터를 추출하고 그에 해당하는 주가의 움직임을 1과 0으로 표현하면, 기계 학습 방법으로 주가를 예측할 입력 값과 출력 값이 준비된다. 기계 학습 방법론적인 측면에서 보았을 때, 뉴스를 통한 주가 예측은 전통적인 문자열 분류를 연장한 것이다. 따라서 수많은 기계 학습 방법론이 문자열 특징 벡터를 분류하는 데 적용됐다. 대표적으로 서포트 벡터 머신 (Hagenau et al., 2013), K-근접 이웃 (Groth and Muntermann, 2011) 과 인공 신경망 (Vui et al., 2013) 등이 있다. Groth and Muntermann (2011)는 기계학습 알고리즘으로 뉴스 분석을 이용한 위기 관리 방법을 제시 하였다. 저자들은 인공 신경망, 서포트 벡터 머신, 나이브 베이즈, 및 K-근접 이웃 등 다양한 방법론을 비교하였다. 이때, 서포트 벡터 머신이 가장 효율적인 방법임을 제시하였다.

최근에 기계 학습이 발전하면서, 여러 가지 기계 학습 알고리즘을 동시에 사용하여 성능을 높이고 과적합 (overfitting)을 방지하는 앙상블 방법론이 다양하게 나오고 있다. 특히 서포트 벡터 머신의 앙상블 방법의 하나로는 다중 커널 학습이 제시된다 (Aiolli and Donini, 2015). 다중 커널 학습은 서포트 벡터 머신이 한 가지 특성을 가지는 입력 값만을 잘 처리할 수 있다는 한계점을 극복하기 위해, 많은 커널을 사용하여 다양한 특성의 입력 값을 반영하고 그를 결합하는 방법론이다. 이 다중 커널 학습은 다양한 데이터 원을 가지는 곳에서 주로 사용되는 데, 뉴스를 통한 주가 예측에서도 사용된다 (Nam and Seong, 2019; Shynkevich et al., 2016).

Shynkevich et al. (2016)에서는 GICS에 기반을 두어 다른 관련성을 가지는 여러 금융 그룹들의 뉴스 데이터를 다중 커널 학습을 사용하여 주가를 예측하였다. Seong and Nam (2018)에서는 군

집 분석을 통해 특정 회사와 관련 있는 회사를 찾고, 관련 있는 기업들의 뉴스와 특정 회사의 뉴스를 다중 커널 학습을 통해 동시에 고려하여 주가를 예측하였다. Nam and Seong (2019)에서는 이전 엔트로피를 통해 인과 관계를 갖는 회사를 찾고, 인과 관계를 가지는 기업들의 뉴스와 특정 회사의 뉴스를 다중 커널 학습을 통해 동시에 고려하여 주가를 예측하였다.

2.4 무작위 행렬 이론과 시장 군집화

주식 시장이 군집으로 움직인다는 것은 오랫동안 연구되었다. Rua and Nunes (2009)는 국제 주식 시장에서 주식의 동조화 현상에 관해 연구하고, 이를 포트폴리오 구성에 어떻게 반영할지에 대해 제시하였다. Morck et al. (2000)는 주식의 동조화 현상이 시장의 효율성이 높은 곳 보다 시장의 효율성이 낮은 곳에서 더 뚜렷하게 나타난다고 제시하였다. 대한민국의 주식 시장은 미국이나 유럽의 주식 시장보다 낮은 효율성을 가지며, 그들의 주식 시장에 동조화 현상을 가지고, 한국 주식 시장 내부에서도 동조화 현상이 있다 (Cho and Mooney, 2015; Loh, 2013).

주식 시장이 군집으로 움직인다는 것은 단순히 시장의 효율성을 검증하는 것뿐 아니라, 예측 문제를 푸는 것에도 중요하다는 것이 연구되고 있다. Seong and Nam (2018)은 한국 주식 시장에서 GICS 의 소재, 제약, 음식료 섹터를 클러스터링하여 같은 군집의 회사에 대해 나온 뉴스가 같은 군집 내에 다른 회사에도 영향을 준다는 것을 제시하였다. Nam and Seong (2019)는 한국 주식 시장에서 GICS의 소재, 제약, 음식료 섹터에서 단방향 인과 관계 네트워크를 형성하여, 단순히 같은 군집에 있다고 영향을 주는 것이 아닌, 같은 군집에서도 영향을 줄 수 있는 관계가 있고, 주지 않는 관계가 있다는 것을 밝혔다. 이와 같은 연구들은 주가 예측에서 주식 시장을 군집화하는 것이 중요하다는 것을 제시한다.

하지만 주식 시장에서 정확하게 군집을 찾는 것은 주식 시장의 무작위 신호 때문에 어려운 작업이다 (Bun et al., 2017). 따라서 기존 연구에서는 경제학 기반으로 만든 GICS나 k-평균 군집 분석과 같은 인공 지능 기법으로 시장을 세분한다 (Aghabozorgi and Teh, 2014; Nam and Seong, 2019; Seong and Nam, 2018). 하지만 GICS는 단순히 경제학적으로 회사를 세분한 것일 뿐 주가 예측에 유용하게 그들을 군집화한 것이 아니며, 특정 기업이 수동으로 이를 분류하는 것이므로, 시장의 변화에 빠르게 대응하는 군집화 시스템을 구축하지 못한다는 단점이 있다. 또한, 단순한 k-평균 군집 분석은 시장의 노이즈를 고려하지 못 하기 때문에 군집 분석이 제대로 이루어지지 않을 수 있다 (Bun et al., 2017).

이를 해결하기 위해 무작위 행렬 이론이 경제물리학에서 주로 사용된다 (Bun et al., 2016; Laloux et al., 2000). 무작위 행렬 이론은 상관 계수 행렬에 존재하는 무작위 신호를 제거하여 정확한 상관 계수를 추정할 수 있게 해주는 이론이다. 이는 특히 데이터 수가 많은 곳에서 주로 사용된다 (Bun et al., 2017). Kim and Jeong (2005)는 무작위 행렬 이론을 주식 시장에 적용하는 방법을 제시하였으며, 미국 주식 시장의 주가의 상관 계수 행렬을 무작위 행렬 이론을 통해, 시장 전반의 효과와 무작위 신호를 제거하여 시장의 군집 분석을 하는 방법을 제시하였다. García (2016)는 주식 시장에서 트위터의 감성 분석 결과와 주가의 상관관계를 무작위 행렬 이론을 통해 확인하였고, 이들이 무작위 신호가 아닌 실제

상관 관계가 있음을 보여주었다.

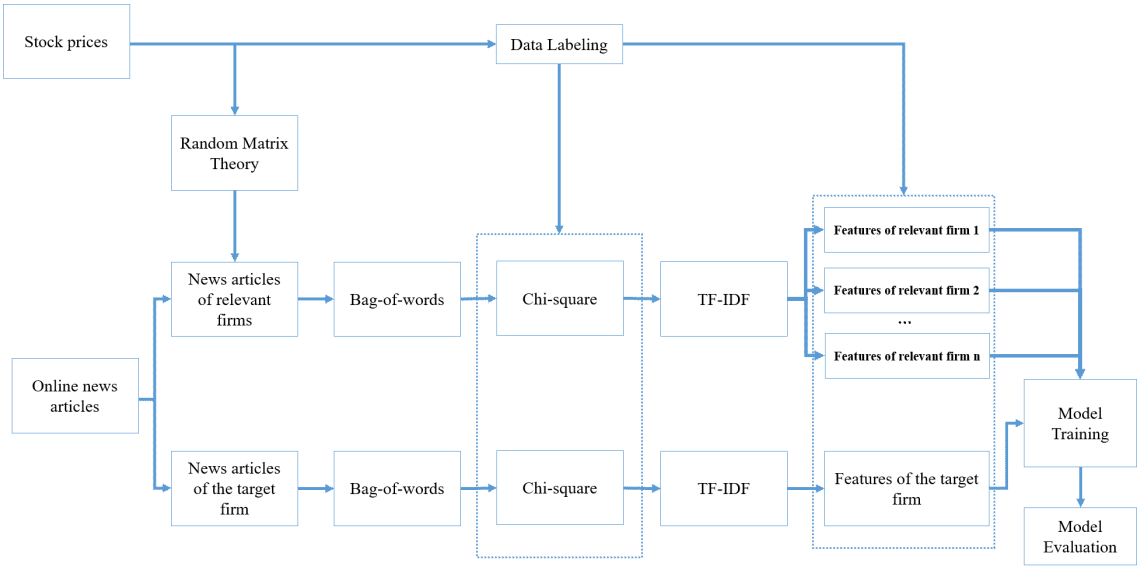
따라서, 본 논문에서는 기존의 연구의 한계를 극복하기 위해, 무작위 행렬 이론을 사용하여 GICS 섹터를 군집 분석을 하고, 그 군집이 인공 지능 알고리즘과 함께 어떤 식으로 사용될 수 있는 지에 대해 보여준다.

3. 연구 모델

이번 장에서는 무작위 행렬 이론 기반 주식 시장 군집 분석을 하여, 관련성이 높은 회사를 찾고, 그 관련성이 높은 회사들의 뉴스 기사가 주가 예측에 반영되는 방법을 제안한다. 3.1장에서는 본 논문에서 사용하는 데이터에 관해 설명한다. 3.2장에서는 무작위 행렬 이론을 통해 시장을 군집 분석하는 방법에 관해 서술한다. 3.3장에서는 텍스트 사전 처리 방식에 관해 설명하며, 3.4장에서는 본 논문에서 사용한 다중 커널 학습에 관해 설명한다. 마지막으로 3.5장에서는 평가 방법에 관해 설명한다. 이 과정에 대한 대략적인 개요는 <Figure 1>에 나와 있다.

3.1 데이터

본 논문에서는 무작위 행렬 이론을 통한 군집 분석이 주식 시장을 적절하게 나누는지와 그것이 인공 지능 방법을 통한 주가 예측에 효과적인 방법임을 검증하기 위해 실제 데이터를 통해 실험하였다. 데이터는 크게 2가지로 나뉘어있다. 그들은 각각 뉴스 데이터와 주가 데이터이다. 데이터의 범위는 2014년 1월 1일부터 2016년 12월 31일이다. 우리는 뉴스 정보를 활용하기 위해 한국 최대 포털사이트 네이버에 등록된 10개의 종합 신문과 14개의 방송 통신 신문과 9개의 경제 신문, 총 33개의 인터넷 뉴스의 모든 금융, 경제 관련 뉴스를 크롤링하였다. 이는 한국에서 대중



〈Figure 1〉 Proposed Approach

이 접할 수 있는 대다수의 금융 뉴스를 포함한 정보로, 금융 뉴스가 미치는 영향을 파악하기 좋은 데이터이다. 이 데이터는 기존 연구에서 많이 사용된 데이터이다 (Nam and Seong, 2019; Seong and Nam, 2018). 뉴스 데이터의 형식은 분류(경제, 금융, 정치 등), 제목, 작성이, 작성 시간, 내용이 있다.

인공지능을 학습하기 위해 훈련 데이터의 뉴스를 분류할 필요가 있다. 뉴스를 분류하는 기준은 주가에 미치는 영향이다. 즉 뉴스가 나왔을 때, 관련된 주가가 올랐다면 1로 표현하였으며, 관련된 주가가 내렸다면 0으로 표현하였다. 하지만 한국 주식 시장은 9시에 열고 16시에 닫는다. 따라서 16시 이후에 나온 뉴스와 주말에 나온 뉴스는 다음 영업일에 영향을 미친다고 해석하였다 (Nam and Seong, 2019).

본 논문에서는 코스피 중에서 소재 섹터에 포함된 기업들에 관해서만 연구를 진행하였다. 각각에 기업에 관련성이 높은 뉴스를 추출하는 방법으로는 본문에 회사의 이름이 포함된 뉴스를 선정 기준으로 사용하였다. 또한, 실험에 사용한 기업에 대한 정보는 <Table 1>과 같다.

3.2 무작위 행렬 이론 (Random Matrix Theory)와 군집화

상관 계수 행렬을 만들 때는 일반적으로 모든 정보를 가지고 하는 것이 아니라, 한정된 데이터를 가지고 행렬을 구성하기 때문에, 상관 계수 행렬은 경험적(empirical) 상관계수이지, 참 상관계수가 아니다. 하지만 실제로 금융 시장에서 포트폴리오를 구성하거나 시장을 세분화할 때는

〈Table 1〉 Up & Down label of the companies

company	data point	part	up label	down label	company	data point	part	up label	down label
OCI	1947	material	1080	867	Lock&Lock	652	material	381	271
Huchems Fine Chemical	175	material	94	81	Korea Petrochemical Ind.	190	material	95	95
Kukdo Chemical	85	material	45	40	SamKwang Glass	176	material	80	96
Hyundai-Steel	3741	material	1858	1883	Young Poong	970	material	441	529
NamHae Chemical	179	material	104	75	Hanwha Chemical	2088	material	1116	972
Hansol Chemical	110	material	71	39	Poongsan	1112	material	593	519
Foosung	295	material	142	153	Lotte chemical	2992	material	1515	1477
SKC	1184	material	658	526	DongKuk Steel Mill	2079	material	1081	998
SKChemical	1276	material	653	623	Taekwang Ind.	327	material	150	177
SeAh Steel	374	material	207	167	SeAh Besteel	362	material	200	162
KISWIRE	166	material	92	74	POSCO	1022	material	509	513
KiscoHolding	762	material	387	375	Kolon Ind.	972	material	481	491
Korea Zinc	619	material	378	241	LG Chem	6717	material	3592	3125
Ssangyong Cement Industrial	390	material	258	132	Lotte find Chemical Co.	178	material	108	70

참 상관계수를 구하는 것이 중요하다 (Bun et al., 2017). 이 문제를 해결하기 위해 본 논문에서는 경제 물리학에서 주로 사용하는 무작위 행렬 이론을 사용한다.

C는 참 상관 계수 행렬이며, M은 교란된 상관 계수 행렬이라고 가정할 때, 무작위 행렬 이론은 M에서 정확한 C를 찾는 것이 목표이다. 이때 C를 추정할 수 있는 함수인 Ξ(M)를 찾는 것이 핵심이다. 본 논문에서는 현재까지 최고의 성능을 보인다고 알려진 Rotational Invariant Estimator (RIE) 방식을 사용한다 (Bun et al., 2016).

$$\mathrm{C} \equiv \mathrm{n\ X\ n\ with\ eigenvalue}\ c_1 \geq \cdots \geq c_n\ and\ \overrightarrow{v_1}, \ldots, \overrightarrow{v_n} \quad (2)$$

$$\mathrm{M} \equiv \mathrm{n\ X\ n\ with\ eigenvalue}\ \lambda_1 \geq \cdots \geq \lambda_n\ and\ \overrightarrow{u_1}, \ldots, \overrightarrow{u_n} \quad (3)$$

RIE 방식은 Ξ(M) 함수를 찾는 것이 목표인데, 이는 방정식 (4)와 같이 표현되며, ξ_i를 추정 해야 한다. 이를 풀기 위해 (5)번 방정식을 손실 함수로 하는 (6)번 방정식을 풀어야 한다.

$$\Xi(\mathrm{M}) = \sum_{i=1}^{N} \xi_i \overrightarrow{u_i}\overrightarrow{u_i}^T \quad (4)$$

$$\|C - \Xi(\mathrm{M})\|_{L_2} = Tr[(\|C - \Xi(\mathrm{M})\|^2)] \quad (5)$$

$$\widehat{\Xi(\mathrm{M})} = argmin(\|C - \Xi(\mathrm{M})\|_{L_2}) \quad (6)$$

이때, 추정된 함수와 결과는 (7)과 같다.

$$\widehat{\Xi(\mathrm{M})} = \sum_{i=1}^{N} \hat{\xi}_i \overrightarrow{u_i}\overrightarrow{u_i}^T, \hat{\xi}_i = \sum_{j=1}^{N} \left(\overrightarrow{u_i}\overrightarrow{v_j}^T\right)^2 c_j \quad (7)$$

(7)에서 구한 최적의 함수로 우리는 참 상관 계수 행렬을 찾을 수 있다. 우리는 참 상관 계수 행렬을 거리 함수로 계층적 군집 분석을 시행하였다. 이때, 본 논문에서는 기존 연구와 비교하기 위하여, 계층적 군집 분석 시행 결과에서 군집을 3개로 나누어 각 군집 할당하였다.

3.3 텍스트 사전 처리

텍스트 사전 처리는 뉴스를 기계 학습 알고리즘이 인식할 수 있는 숫자 벡터로 바꿔주는 과정이며, 필요한 단어만 선택하며 핵심적인 단어에 가중치를 준다. 이 과정을 거치기 위해서 입력값이 정제되어야 한다. 이를 위해, 우리는, 뉴스에서 이메일과 HTML 태그 등 필요 없는 부분을 제거한다. 이후 변수 추출, 변수 선택, 변수 표현 3가지 과정을 거친다.

본 논문에서 변수 추출로는 단어 주머니 모형을 사용하였다. 단어 주머니 모형을 사용하기 위해서는 단어들을 모두 원형으로 만들어야 한다. 본 논문에서는 단어를 원형으로 만드는 방법으로 konlpy 패키지의 꼬꼬마 형태소 분석기를 사용하였다 (Park and Cho, 2014). 본 방법은 다양한 연구에서 사용된 방법으로 그 효율성은 입증되었다 (Nam and Seong, 2019; Seong and Nam, 2018). 본 논문에서는 기존 연구와 같이 3개 이하의 뉴스에서 언급된 원형은 모두 삭제하였으며, 이후에 각각 단어의 원형의 숫자를 센 것이 뉴스를 표현하는 벡터가 된다 (Shynkevich et al., 2016).

본 논문에서 변수 선택으로는 카이스퀘어 방법을 사용하였다. 변수 선택은 수많은 단어 중에서 실제로 주가에 영향을 미쳤을 단어를 통계적으로 찾는 방법을 의미하는 데, 카이스퀘어 방법은 단어의 분포가 기대되는 빈도보다 높을 때, 변수들을 선정하는 방법이다. 즉, 카이스퀘어 변

수 선택 방법을 시행하였을 때, 높은 카이스퀘어 값을 가지는 단어들이 영향을 많이 주는 단어이다. 본 논문에서는 데이터 원이 크게 두 가지가 있다: 특정 회사와 관련 회사들. 특정 회사의 뉴스에서 단어 선택을 할 때는 특정 회사의 주가에 영향을 준 단어들을 선별하며, 관련 회사들의 뉴스에서 단어 선택을 할 때 또한 특정 회사의 주가에 영향을 준 단어들을 선별한다. 본 논문에서는 기존 연구 흐름과 마찬가지로 변수 중 카이스퀘어가 가장 전체 변수의 상위 10%로 하여 선택을 하였다 (Nam and Seong, 2019; Seong and Nam, 2018).

카이스퀘어 변수 선택을 한 후, 특정 단어들이 특정 뉴스에서만 나오는 것에 가중치를 더 주고, 특정 단어들이 수많은 뉴스에서 나오는 것에 가중치를 덜 주기 위해 변수 표현 방법으로는 TF-IDF 방법을 사용하였다 (Nam and Seong, 2019). TF-IDF 변수 표현을 거치면 기계 학습 방법의 입력 값이 너무 작아서 기계 학습을 효율적으로 하기 위해서는 단위 조정이 필요하며, 여러 데이터 원이 같은 정도의 가중치를 갖기 위해서는 단위 조정이 필요하다 (Shynkevich et al., 2016). 따라서 본 논문에서는 기존 연구와 같이 TF-IDF로 표현된 변수에 선택한 변수의 개수를 곱해준다. 변수의 개수가 k개였다면, k*TF-IDF가 될 것이다.

3.4 다중 커널 학습(Multiple Kernel Learning)

다중 커널 학습은 다양한 서브 커널들을 양의 가중치를 부여하여 선형 조합을 한다. 각 커널은 각각 다른 데이터 원이나 각기 다른 특성을 가지는 데이터를 입력 값으로 가질 수 있다 (Gu et al., 2012). 즉, 다중 커널 학습은 방정식 (8)과 같이 나타낼 수 있다. K_s는 미리 계산된 커널 행렬이며, η_s는 행렬 K_s의 가중치이다.

$$\mathrm{K} = \sum_{s=0}^{S} \eta_s K_s \;\; s.t. \eta_s \geq 0 \qquad (8)$$

본 논문에서는 가우시안 커널을 사용하였다. 가우시안 커널은 가장 자주 사용되는 커널이면서도 다양한 비선형 관계를 처리하기에 적합하다 (Hsu et al., 2010). 또한, 가우시안 커널은 선형 커널과 시그모이드 커널을 매개 변수의 범위에 따라 포함할 수 있으므로, 본 논문에서는 가우시안 커널을 사용하였다 (Keerthi and Lin, 2003). 가우시안 커널은 다음과 같이 표현된다:

$$\mathrm{K}\left(z_i, \mathrm{z_j}\right) = e^{-\gamma \left\| z_i - z_j \right\|^2} \; s.t. \; \sigma > 0 \qquad (9)$$

매개 변수 γ는 가우시안 커널의 폭이다. γ에 따라 가우시안 커널의 속성은 크게 변하기에 적절한 값을 찾는 것은 중요하다 (Hsu et al., 2010). 최적의 매개변수를 찾기 위해, 본 논문에서는 각 커널에 서포트 벡터 머신을 사용하여 그리드 서치(grid search)를 진행하였으며,매개 변수 범위는 $\mathrm{C} = \{2^{-5},\ 2^{-3}, \ldots, 2^{15}\}, \gamma = \{2^{-15}, 2^{-13}, \ldots, 2^{3}\}$이다 (Hsu et al., 2010).

각 커널의 가중치 매개 변수는 훈련 기간 동안 조절된다. 최적의 가중치 매개 변수를 찾기 위해 다양한 다중 커널 학습 방법이 사용되었다 (Aiolli and Donini, 2015; Jain et al., 2012). 그들 중에서, 우리는 현재 최고로 알려진 EasyMKL 방법을 사용하였다 (Aiolli and Donini, 2015).

훈련 데이터가 $\mathrm{G_{tr}} = \{(x_{s,1}, y_1), \ldots, (x_{s,i}, y_i)\}$이며, 예측 데이터가 $\mathrm{G_{te}} = \{(x_{s,i+1}, y_{i+1}), \ldots, (x_{s,l}, y_l)\}$이며, $y_i \in \{-1\,(down), +1\,(up)\}$라고 가정한다.

우리는 햇(e.g. $\hat{Y}$)을 훈련 데이터에서 사용하는 서브 행렬이라고 가정한다. 이때, EasyMKL는 다음과 같다.

$$\max_{\|\eta\|=1} \min_{\gamma\in\Gamma} (1-\lambda)\gamma^T \hat{Y}\left(\sum_{s=0}^{S} \eta_s \hat{K}_s\right)\hat{Y}\gamma + \lambda\|\gamma\|^2 \quad (10)$$

$$s.t.\ \Gamma = \{\gamma \in R_+^L \mid \sum_{y_i=+1} \gamma_i = 1,\ \sum_{y_i=-1} \gamma_i = 1\}$$

$\lambda \in (0,1)$는 과적합을 방지하기 위한 정규화 매개변수이다. 만약 $\lambda = 0$이라면, 정규화가 존재하지 않으며, 결과는 서포트 벡터 머신과 같다 (Aiolli and Donini, 2015). 만약 $\lambda = 1$이라면, 최적의 해는 변수 공간에서 긍정적인 부분과 부정적인 부분의 중심의 거리의 루트 값이다 (Aiolli and Donini, 2015). 즉, λ는 훈련 데이터 특성에 따라 최적의 값이 달라진다. 최적의 λ를 찾기 위해서, 우리는 $\{0.1, 0.2, \ldots, 0.9\}$의 범위에서 그리드 서치를 하였다.

본 논문에서 제안한 방법은 특정 회사의 뉴스와 관련 있는 회사들의 뉴스를 함께 고려하는 모형이다. 관련 있는 회사가 n-1개가 있을 시, n개의 커널이 생성되며, 각 커널의 최적의 매개 변수는 서포트 벡터 머신을 그리드 서치한 결과로 나온다. 최적의 값을 찾은 뒤에는 다중 커널 학습을 시행한다.

본 논문에서 주장하고자 하는 바는 무작위 행렬 이론을 통한 시장 세분화를 하여 여러 뉴스를 사용하는 것이 k-평균 군집 분석이나 섹터 혹은 단일 특정 회사의 뉴스를 사용하는 것보다 더 좋은 결과를 보여준다는 것이다. 따라서 본 논문에서는 크게 3가지 비교 대상을 가진다. 첫 번째로는 단일 회사의 뉴스만으로 주가를 예측하는 것이며 (Hagenau et al., 2013), 두 번째는 섹터의 뉴스도 함께 고려하는 것이며 (Shynkevich et al., 2016), 세 번째는 k-평균 군집 분석을 하여 나온 관련된 회사의 뉴스도 함께 고려하는 것이다 (Seong and Nam, 2018).

3.5 평가 방법

우리는 2014년 1월부터 2016년 12월까지 총 3년간의 뉴스 데이터와 3년 간의 주가 데이터를 사용하였다. 제안된 방법의 검증을 위하여 훈련 기간, 검증 기간, 예측 기간을 각각 2년, 6개월, 6개월로 설정으로 하였다. 훈련 기간 동안 매개변수를 찾고, 검증 기간 동안 최적의 매개 변수를 그리드 서치를 통해 찾으며, 예측 기간 동안 평가를 진행한다. 실험이 끝나면 각각의 예측에 대해서, 예측이 Up이라고 예측했는데, 옳게 예측한 수를 TP라고 정의하고, 틀리게 예측한 수를 FP라고 하며, Down이라고 예측하였는데, 옳게 예측한 수를 FN, 틀리게 예측한 것을 TN이라고 한다. 이는 <Table 2>와 같이 나타낼 수 있다.

정확도는 <Table 2>에서 Accuracy $\stackrel{\text{def}}{=} \frac{TP+TN}{TP+FP+FN+TN}$로 정의가 된다.

〈Table 2〉 Confusion Matrix

		Prediction	
		Up	Down
Actual	Up	TP	FN
	Down	FP	TN

4. 결과

<Table 3>은 본 실험의 결과를 나타낸다. Reference는 비교군으로 사용한 기존의 연구들을 의미하며, Method는 각 논문에서 관련성이 있는 회사를 찾는 방식을 의미한다. 또한, Accuracy는 훈련 기간과 검증 기간에서 최상의 매개변수를 찾고 이를 기반으로 예측 기간에서 뉴스로 주가 예측을 했을 때, 정확도를 의미한다.

먼저, Shynkevich et al. (2016)가 제시한 GICS에서 Sector를 동질적인 군집으로 확인하여 같은 섹터(Sector)의 뉴스를 함께 고려하여 주가를 예측한 것과 Hagenau et al. (2013)의 Individual level, 즉 특정 회사의 뉴스만으로 주가를 예측한 것을 비교하면, 정확도가 소폭 상승했음을 알 수 있다. 이는 기존 연구 결과와 합당하게 단순히 특정 회사에 관련된 뉴스만이 아니라 관련된 회사의 뉴스를 포함하여 예측하는 것이 의미가 있다는 것을 다시 한 번 증명하며, 이 결과는 Shynkevich et al. (2016)와 일관된 결과를 보여주며, 본 연구에서 사용한 데이터가 기존 연구들과 유사하다는 것을 암시한다. 하지만 이 두 비교에서는 통계적으로 유의미한 차이를 보여주지 못하였기 때문에, 알고리즘의 복잡도에 비해 그 결과의 차이가 크지 않다.

Seong and Nam (2018)의 K-means clustering과 Shynkevich et al. (2016)의 Sector를 비교하면, 정확도가 소폭 감소하였으나 Individual level보다는 높음을 알 수 있다. 이는 군집 분석을 제대로 시행하였을 않아, 관련성이 높은 회사를 적절하게 찾지 못한다면, 오히려 악영향을 미칠 수도 있다는 것을 의미한다. 이는 기존 K-평균 군집 분석이 주식 시장에서 제대로 작동하지 못한다는 것을 의미하고, 이는 무작위 신호와 시장 전체 효과 등 때문이다. 즉, 이와 같은 무작위 신호와 시장 전체 효과를 고려하지 않고, 단순히 주식 시장에서 군집 분석을 시행하는 방식은 잘못된 결과를 낼 수 있다. 본 논문에서는 이를 해결하기 위하여, 무작위 행렬 이론을 사용하였다.

마지막으로, 본 연구에서 제시한 방법과 기존 연구들이 제시한 방법들을 비교했을 때는 본 논문에서 제시한 방법이 우수한 성능을 보여준다. 이는 기존의 방법들과 약 3%의 정확도 차이를 보이며, 통계적으로 유의미한 차이를 보인다. 기존 주가 예측 방법론이 정확도를 1% 이상 발전시키기 힘든 점에 비교하였을 때, 이는 매우 큰 차이로 보인다. 즉, 무작위 행렬 이론을 통해 관련성이 높은 기업들을 선택한 것이 k-평균 군집 분석을 시행하여 관련성이 높은 기업을 추출하는 것보다 동질성이 높은 기업들을 적절하게 찾는다는 점을 시사한다. 즉, 주식 시장에서 시장 세분화를 할 때는 무작위 신호와 시장 전체 효과에 대해서 항상 고려해야 하고 이를 제거하는 방법을 찾을 필요가 있다.

〈Table 3〉 Experimental results

Reference	Method	Accuracy
Hagenau et al. (2013)	Individual level	0.6245
Shynkevich et al. (2016)	Sector	0.6299
Seong and Nam (2018)	K-means clustering	0.6276
The proposed approach	Random Matrix Theory	**0.6586**

5. 결론

빅데이터 시대와 초연결 시대에 도입하면서, 온라인상에 저장되는 정보의 양이 무수히 많아지고 있다. 그중에서 가장 대표적인 것이 문자열 데이터이다. 문자열 데이터의 양이 급증하면서, 문자열에서 의미를 자동으로 찾고 해석하며, 이를 경영 전반에 응용하려는 경향이 뚜렷해지고 있다. 이와 같은 현상은 주가 예측에도 동일하게 나타난다.

기존의 주가 예측은 주가의 추세를 통해 미래를 예측하는 것이 대부분이었다. 하지만 문자열 처리 기술의 발전과 함께, 주가에 영향을 줄 수 있는 새로운 정보인 뉴스의 영향이 재조명되고, 이를 인공지능 기술로 자동적으로 예측하려는 시스템이 함께 발전하고 있다. 특히, 현재와 같이 정보의 양이 많을 때, 각 개인이 모든 뉴스를 읽고 주가에 영향을 미치는 정보만을 선별적으로 찾아내어 이용하는 것은 물리적으로 불가능하다. 이에 인공지능 기술로 뉴스를 이용한 주가 예측을 하려는 시도는 꾸준히 진행되어왔다.

이러한 주가 예측의 방법은 대개 A기업에 관련된 뉴스가 나오면 그 뉴스를 이용해 A 기업의 주가를 예측하는 것이다. 하지만 최근 연구에서 밝히기로는, 단순히 특정 회사의 뉴스만이 그 회사의 주가에 영향을 주는 것이 아니라, 그 회사와 관련성이 높은 회사들의 뉴스 또한 주가에 영향을 줄 수 있다는 것을 밝혔다 (Nam and Seong, 2019; Seong and Nam, 2018; Shynkevich et al., 2016). 즉, 관련성이 높은 회사를 정확하게 찾는 것은 뉴스를 이용한 주가 예측 방법에서 중요한 부분이다. 하지만 아직 이 부분에 대한 연구가 부족하기 때문에 본 논문에서는 이 연구 흐름을 경제 물리학 방법론을 통해 보완하였다.

본 연구에서는 무작위 행렬 이론을 이용하여 주가의 상관 계수 행렬의 잡음을 제거하고, 시장 전반의 효과를 제거하며 정확한 상관 계수를 찾았다. 그 상관 계수를 기반으로 군집 분석을 시행하여 관련된 회사를 찾고 그 관련된 회사들의 뉴스를 함께 고려하는 주가 예측 방법론을 제시하였다. 그 결과, 본 논문에서 제시한 방법이 각 개인의 회사의 뉴스만 가지고 주가를 예측하는 것, 혹은 섹터의 모든 뉴스를 가지고 주가를 예측하는 것, K-평균 군집 분석을 관련된 회사를 찾고 이를 이용해 주가를 예측하는 것보다 월등한 성과를 보여주었다. 이는 관련된 회사를 정확하게 찾는 것이 중요함을 다시 한 번 상기시켜주며, 제시한 방법이 기존 연구들보다 우월한 성능을 나타냄을 의미한다.

본 연구의 학문적 시사점은 다음과 같다. 첫 번째, 본 연구는 주식 시장이 이론적으로 가지는 복잡계 특성을 분석하여 이를 인공지능 알고리즘의 입력값으로 사용하는 연구를 확장하였다. 이는 단순히 모든 입력값을 통합하여 사용하는 것이 아니라, 이론적으로 그 관계를 고려해야 한다는 점을 의미한다. 두 번째, 본 연구는 주식 시장에서 관련성이 높은 기업을 정확하게 찾는 것이 중요한 문제임을 다시 한 번 강조하였다.

본 연구의 실무적 시사점은 다음과 같다. 첫 번째, 본 연구에서 사용한 무작위 행렬 이론은 주식 시장에서 정확한 상관계수를 찾는 데 사용되었으며, 그 효과를 입증하였다. 이는 포트폴리오를 구성하는 인공지능 알고리즘을 구현할 때 이러한 관계를 추가로 이용할 수 있다. 두 번째, 본 연구에서 제안한 방법으로 주가를 예측하는 방식은, 특히 시장이 정형적으로 갖춰져 있지 않은 가상화폐 시장 같은 곳에서 그들 사이의 관계를 찾고 이를 주가 예측에 반영할 때 큰 도움이

될 수 있다.

하지만 본 연구는 다음과 같은 한계점을 가진다. 첫 번째, 본 논문에서는 GICS 섹터 안에서 3개의 군집으로 군집을 나누었다. 이러한 방법은 시장 전체에서 관련성이 높은 기업들을 찾는 것이 어렵다는 단점이 있다. 또한, 현재는 3개의 군집을 가진다고 기존 연구에 따라 선택하였는데, 이는 통계적으로 분석한 것이 아니라 일반화가 힘들다는 단점이 있다. 후속 연구에서는 시장 전체에서 관련성이 높은 기업들을 찾으며, 군집의 개수를 통계적으로 분석하는 방법을 함께 사용할 필요가 있다. 두 번째, 본 논문에서는 GICS 섹터 중 소재 섹터를 선택하여 연구를 진행하였다. 하지만 이는 소재 섹터의 특성일 수도 있으니 일반화의 문제가 있다. 후속 연구에서는 일반화를 보다 잘 주장하기 위해 시장 전반에서 다양한 섹터에서 진행할 필요가 있다.

참고문헌(References)

Aghabozorgi, S., and Y. W. Teh, "Stock Market Co-Movement Assessment Using a Three-Phase Clustering Method," *Expert Systems with Applications,* Vol.11, No.4 (2014) 1301~1314.

Aiolli, F., and M. Donini, "Easymkl: A Scalable Multiple Kernel Learning Algorithm," *Neurocomputing*, Vol.169, No.1(2015), 215~224.

Bun, J., R. Allez, J.-P. Bouchaud, and M. Potters, "Rotational Invariant Estimator for General Noisy Matrices," *IEEE Transactions on Information Theory*, Vol.62, No.12(2016), 7475~7490.

Bun, J., J.-P. Bouchaud, and M. Potters, "Cleaning Large Correlation Matrices: Tools from Random Matrix Theory," *Physics Reports*, Vol.666, No.1(2017), 1~109.

Cho, C. H., and T. Mooney, "Stock Return Comovement and Korean Business Groups," *Review of Development Finance,* Vol.5, No.2 (2015), 71~81.

García, A., "Global Financial Indices and Twitter Sentiment: A Random Matrix Theory Approach," *Physica A: Statistical Mechanics and its Applications*, Vol.461, No.1(2016), 509~522.

Groth, S. S., and J. Muntermann, "An Intraday Market Risk Management Approach Based on Textual Analysis," *Decision Support Systems*, Vol.50, No.4(2011), 680~691.

Gu, Y., C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative Multiple Kernel Learning for Classification in Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, Vol.50, No.7(2012), 2852~2865.

Hagenau, M., M. Liebmann, and D. Neumann, "Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Capturing Features," *Decision Support Systems,* Vol.55, No.3(2013), 685~697.

Hsu, C., C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," Department of Computer Science National Taiwan University, 2010.

Jain, A., S. V. Vishwanathan, and M. Varma, "Spf-Gmkl: Generalized Multiple Kernel Learning with a Million Kernels," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data*

mining: ACM, (2012), 750~758.

Keerthi, S. S., and C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," *Neural computation*, Vol.15, No.7(2003), 1667~1689.

Kim, D.-H., and H. Jeong, "Systematic Analysis of Group Identification in Stock Markets," *Physical Review E*, Vol.72, No.4(2005), 046133.

Laloux, L., P. Cizeau, M. Potters, and J.-P. Bouchaud, "Random Matrix Theory and Financial Correlations," *International Journal of Theoretical and Applied Finance*, Vol.3, No.3 (2000), 391~397.

Loh, L., "Co-Movement of Asia-Pacific with European and Us Stock Market Returns: A Cross-Time-Frequency Analysis," *Research in International Business and Finance*, Vol.29, No.1(2013), 1~13.

Morck, R., B. Yeung, and W. Yu, "The Information Content of Stock Markets: Why Do Emerging Markets Have Synchronous Stock Price Movements?," *Journal of financial economics*, Vol.58, No.1-2(2000), 215~260.

Nam, K., and N. Seong, "Financial News-Based Stock Movement Prediction Using Causality Analysis of Influence in the Korean Stock Market," *Decision Support Systems*, Vol.117, No.1(2019), 100~112.

Park, E. L., and S. Cho, "Konlpy: Korean Natural Language Processing in Python," *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology,* (2014).

Rua, A., and L. C. Nunes, "International Comovement of Stock Market Returns: A Wavelet Analysis," *Journal of Empirical Finance*, Vol.16, No.4(2009), 632~639.

Seong, N., and K. Nam, "Combining Macro-economical Effects with Sentiment Analysis for Stock Index Prediction," Entrue Journal of Information Technology, Vol.16, No.2(2017), 41~54.

Seong, N., and K. Nam, "Online News-Based Stock Price Forecasting Considering Homogeneity in the Industrial Sector," *Journal of Intelligence and Information Systems*, Vol.24, No.2(2018), 1~19.

Shynkevich, Y., T. M. McGinnity, S. A. Coleman, and A. Belatreche, "Forecasting Movements of Health-Care Stock Prices Based on Different Categories of News Articles Using Multiple Kernel Learning," *Decision Support Systems*, Vol.85, No.1(2016), 74~83.

Vui, C. S., G. K. Soon, C. K. On, R. Alfred, and P. Anthony, "A Review of Stock Market Prediction with Artificial Neural Network (Ann), "*IEEE International Conference on Control System, Computing and Engineering*: IEEE, (2013) 477~482.

Abstract

Predicting stock movements based on financial news with systematic group identification

Seong NohYoon* · Nam Kihwan**

Because stock price forecasting is an important issue both academically and practically, research in stock price prediction has been actively conducted. The stock price forecasting research is classified into using structured data and using unstructured data. With structured data such as historical stock price and financial statements, past studies usually used technical analysis approach and fundamental analysis. In the big data era, the amount of information has rapidly increased, and the artificial intelligence methodology that can find meaning by quantifying string information, which is an unstructured data that takes up a large amount of information, has developed rapidly. With these developments, many attempts with unstructured data are being made to predict stock prices through online news by applying text mining to stock price forecasts.

The stock price prediction methodology adopted in many papers is to forecast stock prices with the news of the target companies to be forecasted. However, according to previous research, not only news of a target company affects its stock price, but news of companies that are related to the company can also affect the stock price. However, finding a highly relevant company is not easy because of the market-wide impact and random signs. Thus, existing studies have found highly relevant companies based primarily on pre-determined international industry classification standards. However, according to recent research, global industry classification standard has different homogeneity within the sectors, and it leads to a limitation that forecasting stock prices by taking them all together without considering only relevant companies can adversely affect predictive performance.

To overcome the limitation, we first used random matrix theory with text mining for stock prediction. Wherever the dimension of data is large, the classical limit theorems are no longer suitable, because the statistical efficiency will be reduced. Therefore, a simple correlation analysis in the financial market does

* College of Business, KAIST
** Corresponding Author: Kihwan Nam
College of Business, Korea Advanced Institute of Science and Technology (KAIST)
85 Hoegi-Ro, Dongdaemoon-Gu, Seoul, 130-722, Korea
Tel: +82-10-4930-8317, E-mail: namkh@kaist.ac.kr

not mean the true correlation. To solve the issue, we adopt random matrix theory, which is mainly used in econophysics, to remove market-wide effects and random signals and find a true correlation between companies. With the true correlation, we perform cluster analysis to find relevant companies.

Also, based on the clustering analysis, we used multiple kernel learning algorithm, which is an ensemble of support vector machine to incorporate the effects of the target firm and its relevant firms simultaneously. Each kernel was assigned to predict stock prices with features of financial news of the target firm and its relevant firms.

The results of this study are as follows. The results of this paper are as follows. (1) Following the existing research flow, we confirmed that it is an effective way to forecast stock prices using news from relevant companies. (2) When looking for a relevant company, looking for it in the wrong way can lower AI prediction performance. (3) The proposed approach with random matrix theory shows better performance than previous studies if cluster analysis is performed based on the true correlation by removing market-wide effects and random signals.

The contribution of this study is as follows. First, this study shows that random matrix theory, which is used mainly in economic physics, can be combined with artificial intelligence to produce good methodologies. This suggests that it is important not only to develop AI algorithms but also to adopt physics theory. This extends the existing research that presented the methodology by integrating artificial intelligence with complex system theory through transfer entropy. Second, this study stressed that finding the right companies in the stock market is an important issue. This suggests that it is not only important to study artificial intelligence algorithms, but how to theoretically adjust the input values. Third, we confirmed that firms classified as Global Industrial Classification Standard (GICS) might have low relevance and suggested it is necessary to theoretically define the relevance rather than simply finding it in the GICS.

Key Words : Online News, Stock prediction, Random matrix theory, hierarchical clustering

Received : May 17, 2019 Revised : August 31, 2019 Accepted : September 19, 2019
Publication Type : Regular Paper Corresponding Author : Kihwan Nam

저 자 소 개

성 노 윤
KAIST에서 물리학 학사 학위를 취득하였다. 현재 KAIST 경영대학원 경영공학부 MIS 박사 과정에 재학중이다. 주요 관심분야는 자연어 처리, 머신러닝, 빅데이터 분석, 계량경제학, 경제물리학 등이다. 기존 경제학 이론에 머신러닝을 접목하여 사회 전반적인 문제를 해결하는 데에 관심을 가지고 있다.

남 기 환
KAIST 경영대학원 경영공학부에서 MIS 박사학위를 취득하였다. 현재 KAIST 경영대학원 경영공학부와 성균관대학교 데이터 사이언스전공 겸직교수, UNIST 경영학부 초빙조교수로 재직 중이다. 주요 관심분야는 Business Analytics & Business Intelligence, Big Data Analytics, Data Mining, Statistical Analysis, Recommender Systems, Econometrics Models, Machine Learning, Deep Learning 등 이다. 관연 연구들은 Decision Support Systems, Data Mining and Knowledge Discovery 등에 논문이 게재되었다. 학문적인 연구뿐만 아니라 이론을 바탕으로 실제 기업에서 다양한 프로젝트를 성공적으로 진행함으로써 학계와 산업계 모두에 실증적인 기여를 하고 있다.