ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2018 June: 24(2): 1~19 http://dx.doi.org/10.13088/jiis.2018.24.2.001

산업군 내 동질성을 고려한 온라인 뉴스 기반 주가예측*

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

남기환
한양대학교 경영대학 경영학부
(*namkh@kaist.ac.kr*)

주가 예측은 학문적으로나 실용적으로나 중요한 문제이기에, 주가 예측에 관련된 연구가 활발히 진행되었다. 빅 데이터 시대에 도입하면서, 빅 데이터를 결합한 주가 예측 연구도 활발히 진행되고 있다. 다수의 데이터를 기반으로 기계 학습을 이용한 연구가 주를 이룬다. 특히 언론의 효과를 접목한 연구 방법들이 주목을 받고 있는데, 그중 온라인 뉴스를 분석하여 주가 예측에 활용하는 연구가 주를 이루고 있다.

기존 연구들은 온라인 뉴스가 개별 회사에 대한 미치는 영향을 주로 살펴보았다. 또한, 관련성이 높은 기업끼리 서로 영향을 주는 것을 고려하는 방법도 최근에 연구되고 있다. 이는 동질성을 가지는 산업군에 대한 효과를 살펴본 것인데, 기존 연구에서 동질성을 가지는 산업군은 국제 산업 분류 표준에 따른다. 즉, 기존 연구들은 국제 산업 분류 표준으로 나뉜 산업군이 동질성을 가진다는 가정하에서 분석을 시행하였다. 하지만 기존 연구들은 영향력을 가지는 회사를 고려하지 못한 채 예측하였거나 산업군 내에서 이질성이 존재하는 점을 반영하지 못했다는 한계점을 가진다. 본 연구는 산업군 내에 이질성이 존재함을 밝히고, 이질성을 반영하지 못한 기존 연구의 한계점을 K-평균 군집 분석을 적용하여, 주가에 영향을 미치는 산업군의 동질적인 효과를 반영할 수 있는 방법론을 제안하였다. 방법론이 적합하다는 것을 증명하기 위해 3년간의 온라인 뉴스와 주가를 통해 실험한 결과, 다수의 경우에서 본 논문에서 제시한 방법이 좋은 결과를 나타냄을 확인할 수 있었으며, 국제 산업 분류 표준 산업군 내에서 이질성이 클수록 본 논문에서 제시한 방법이 좋은 효과를 보인다는 것을 확인할 수 있었다. 본 연구는 국제 산업 분류 표준으로 나누어진 기업들이 높은 동질성을 가지지 않는 다는것을 밝히고 이를 반영한 예측 모형의 효율성을 입증하였다는 점에서 의의를 가진다.

주제어 : 주가 예측, 텍스트 마이닝, 기계 학습, 다중 커널 학습, 군집 분석

논문접수일 : 2017년 11월 10일 논문수정일 : 2018년 4월 30일 게재확정일 : 2018년 5월 24일
원고유형 : 일반논문 교신저자 : 남기환

1. 서론

빅 데이터 시대에 돌입하면서, 다양한 데이터를 기반으로 다양한 연구들이 진행되고 있다. 이러한 연구 흐름에 맞추어 기존에 다양한 시도를 해왔던 주가예측 분야도, 빅데이터 분석을 활용한 연구가 활발히 진행됐다. 가장 활발히 진행되고 있는 분야는 온라인 뉴스를 기반으로 각 기업의 주가를 예측하는 연구들이었다 (Kim et al., 2012; Jeong et al., 2015; Lee and Lee, 2017).

하지만, 기존의 연구들을 보면, 대다수 각 기업에 관련성이 가장 높은 기사들만을 선택하여

* 본 연구는 한국과학기술원의 미래선도형 특성화 연구사업 중, IoT기반 초연결 사회를 위한 미래 비젼에 관한 연구의 일환으로 이루어졌음.

연구한다. 관련성이 높은 기사들을 선택하는 방법은 일반적으로, 틱커(ticker)를 기반으로 선택하거나, 제목에 회사 이름이 포함되어 있거나, 본문에 대상 기업이 포함되는 경우이다. 하지만 카메라 제조 회사의 호재 뉴스가 나온다면 렌즈 부품 회사의 주가도 함께 오르듯이, 관련성이 높은 기업의 뉴스 또한 그 회사에 영향을 줄 수 있다. 이 사실을 이용하여, Shynkevich et al. (2016)은 국제 산업 분류 표준에서 같은 산업군에 있는 기업끼리는 관련성이 높은 기업이라고 정의를 하고, 예측할 기업의 뉴스와 관련성이 높은 기업의 뉴스를 통합하여 주가의 움직임을 예측하였다. 이때, 단순히 회사를 기반으로 하여 예측을 하는 것보다 더 정확하게 예측한다는 것을 보여주었다.

하지만 국제 산업 분류 표준은 산업 체계를 반영하여 체계적으로 구성한 시스템이지만, 실제로 주가 간의 관련성을 반영한 체계라고 보기는 어렵다. 하지만 기존 연구의 같은 산업군 내의 효과를 반영한 연구는 모든 산업군 내에는 모두 동질성을 가진다는 가정에서 출발하였다. 산업군끼리 비교하였을 때보다 산업군 내에서 이질적인 성향을 띄는 산업군이 있고 더욱 동질적인 성향을 띄는 산업군은 존재할 것이다. 하지만 기존 연구에서는 이러한 다양한 형태의 산업군의 특성을 모두 살펴보지 못하고 동질성을 띠는 산업군에 관해서만 연구를 진행하였다. 다양한 산업군을 확인해본 결과 기존 연구의 주장과는 다른 결과를 나타내는 산업군 또한 존재함을 알 수 있었다. 본 연구는 이러한 특성을 기계 학습 기법을 접목해 산업군 내의 동질적인 경향성을 찾고 명확하게 적용할 수 있는 방법을 제안한다. 따라서 본 논문에서는 국제 산업 분류 표준으로 뉴스의 관련성을 측정하는 것을 대체할 방법을 고안한다. 이를 위해 주식의 동향에 따라 군집 분석을 하여, 관련성이 높은 기업들을 선별하여, 이를 기반으로 하여 주가를 예측하였다.

문자열 정보로 주가 방향성 예측을 할 때는 주로 서포트 벡터 머신(Support Vector Machine)이 사용된다 (Hagenau et al., 2013). 하지만 본 논문에서는 예측 대상 기업 외에도 관련성이 있는 기업의 뉴스 정보 또한 반영하여야 하므로, 서포트 벡터 머신이 적합하지 않다. 여러 가지 특성을 가지는 데이터를 통합하는 방법으로는 다중커널 학습 (Multiple Kernel Learning)이 주로 사용된다 (Deng et al., 2011; Li et al., 2011; Yeh et al., 2011; Wang et al., 2012; Shynkevich et al., 2016). 다중커널학습은 여러 가지 커널을 가져, 각각의 커널이 다른 데이터를 받아들여 예측한다. 이때, 각 커널의 가중치를 잘 조절해 주는 것이 중요한데, 이를 해결하기 위해서, 본 논문에서는 EasyMKL 방법을 사용하였다 (Aiolli and Donini, 2015). 각각의 커널은 대상 기업, K-평균 군집 분석으로 나눈 산업군의 금융 뉴스의 변수(feature)들로 주가를 예측하는데 할당되었다.

본 논문의 결과는 다음과 같다. (1) 자신과 관련 있는 회사들로 묶인 산업군의 정보 또한 유의미한 정보를 포함한다는 것을 확인하였으며, 관련성 있는 기업의 뉴스와 자신의 뉴스를 함께 고려할 때 더 뛰어난 예측력을 가진다는 것을 확인하였다. (2) 산업군 내에 주가가 어느 수준의 동질성을 가지는지에 따라, 군집의 수를 다르게 하여 예측하는 것이 중요하다는 것이다. 즉, 산업군 내에서 주가가 동질적일 때 군집 분석을 하지 않고 산업군 수준으로 관련성 효과를 사용하거나 적은 수의 군집으로 나누어 사용하는 것이 중요하며, 산업군 내에서 주가가 이질적일 때 군집 분석을 하여 기업들을 동질적인 그룹으로 묶어

예측하는 것이 중요하다는 것이다.

본 연구의 기여는 다음과 같다. 첫 번째, 본 연구는 기존에 국제 산업 분류 표준에서 같은 산업군으로 나누어진 기업들이 이질성을 가진다는 것을 밝힘으로써, 관련성을 단순히 국제 산업 분류 표준에서 정의하는 것이 아닌, 기계 학습 및 통계적 분석 방법론을 통해 정의하는 것이 필요하다는 것을 밝혀냈다. 두 번째, 산업군 내에 이질성이 클수록 더 많은 군집으로 나누어 예측해야 한다는 것을 밝힘으로써, 이질성을 반영한 예측 모형의 효율성을 입증하였다.

본 논문의 뒷부분은 다음과 같이 구성된다. 2장에서는 정성적 정보의 주가에 대한 영향을 예측하는 기존의 관련 논문들에 대한 전반적인 흐름을 볼 것이다. 3장에서는 금융 뉴스 데이터로 주식을 예측하는 연구 모델을 제시하며, 국제 산업 분류 표준에 관한 기존 연구 (Schumaker and Chen, 2009; Shynkevich et al., 2016)와 비교할 것이다. 4장에서는 실험 결과에 대해 설명할 것이다. 5장에서는 본 연구에 관해 결론을 내며, 한계점과 후속 연구를 위한 지침에 관해 설명할 것이다.

2. 관련 연구

2.1 핵심 관련 연구

Deng et al. (2011) 에서는 소셜 네트워크 서비스 (Social Network Service)의 감성 분석, 기술적 분석, 뉴스의 수치적 특성과 같은 다양한 특성을 가지는 데이터를 이용하여 주가를 예측하는 알고리즘을 고안하였다. 이때, 서포트 벡터 회귀 (Support Vector Regression)는 하나의 특성을 가지는 데이터만을 사용할 수 있으므로, 이 한계점을 극복하고자 앙상블 방법의 하나인 다중커널 학습을 하여 단일의 커널을 사용한 서포트 벡터 회귀보다 더 좋은 결과를 보여주었다. 이에 따라, 본 논문에서도 다양한 특성을 가지는 데이터를 통합하기 위해 다중커널학습을 사용한다.

Hageneu et al. (2013) 에서는 기업 발표와 금융 뉴스를 자동으로 받아와서 텍스트 마이닝을 통해서 주가를 예측하는 시스템을 구축하였다. 이때, 변수 선택을 위해서 시장 반응을 이용한 카이 제곱 (Chi-square), 이중 정상 분리 (Bi normal separation)을 하였고, 단어 주머니 모형 (Bag of Word), 2-Gram 등 여러 가지 변수 추출 방법을 사용하여 높은 예측률을 보여주었다. 본 논문에서는 Hageneu et al. (2013) 에서 사용한 단어 주머니 모형에 카이제곱 변수 선택과 TF-IDF (Term Frequency – Inverse Document Frequency) 가중치를 사용하였다.

Schumaker and Chen(2009)은 뉴스 기사와 주식 거래 전문가의 의견 그리고 주식 시세로 데이터를 구성한 후, 서포트 벡터 회귀(Support Vector Regression)로 예측하는 시스템 Arizona Financial Text System(AZFinText)을 제안하였다. 저자는 AZFinText를 사용하여, 효율적으로 데이터를 모으고 문자열 정보로 체계적으로 주가를 예측하는 방법에 대하여 논의하였다. 저자들은 주가에 영향을 미치는 뉴스를 국제 산업 분퓨 표준에 따라 여러 단계로 분리하였다. 저자는 기업과 관련있는 산업군, 산업그룹, 산업, 하위산업, 특정 회사의 뉴스 기사를 전부 사용하여, 각각의 문자열 데이터로 주가를 예측하였다. 그 결과 특정 회사에 관련있는 뉴스뿐만 아니라, 산업군 기반 뉴스 데이터 등의 데이터도 주가를 예측함에 유효함을 보여주었다. 하지만 이 논문에서는 여

러 단계에서의 변수들을 동시에 사용하여 예측하지 않아 예측력을 높이지 못했다는 한계점이 있다. 따라서 본 논문에서는 산업군과 특정 회사 등 여러 단계의 변수들을 통합하여 사용하기 위하여, 다중커널학습을 사용하였다.

Shynkevich et al. (2016)은 서로 다른 단계의 관련성을 가지는 기업들의 뉴스를 통합하여 주가를 예측하는 시스템을 만들었다. 저자는 국제 산업 분류 표준 기반의 뉴스 그룹을 만들어 주가를 예측하였다. 즉, 기업과 관련이 있는 산업군 기반, 산업그룹 기반, 하위 산업 기반, 그룹 기반, 특정 회사의 뉴스 기사를 전부 사용하여 다중커널학습을 사용하여 비교하였다. 그 결과 다른 수준의 관련성을 함께 고려하여 주가를 예측하는 경우 주가를 특정 주식만으로 예측하는 것보다 뛰어난 결과를 보여주었다. 하지만, Shynkevich et al. (2016) 에서는 같은 국제 산업 분류 표준 체계에 있으면 관련성이 높을 것이라는 가정하에서 시스템을 고안하였으나, 실제로는 같은 업종에 있다고 하더라도 관련성이 높지 않을 수 있다. 따라서 본 논문에서는 단순히 국제 산업 분류 표준 체계를 사용하는 것이 아닌 기계 학습 기법을 사용하여 기업 간의 관련성을 반영하는 방법을 찾았다.

2.2 텍스트 사전 처리

뉴스 정보를 받게 되면 문자열 데이터를 기계 학습 알고리즘이 사용할 수 있는 의미 있는 형태로 변환해줄 필요가 있다. 즉, 뉴스 데이터에서 주가에 영향을 미치는 변수들을 추출 해야 한다. 예를 들어, 유상 증자라는 단어는 그 회사의 주가에 지대한 영향을 미친다. 즉, 단어들과 단어들의 조합, 문장들이 주가에 영향을 미치며, 이 정보들이 적절하게 표현되어야 한다. Mittermayer(2004)는 텍스트 사전 처리를 3가지로 나타내었다. 변수 추출(feature extraction), 변수 선택(feature selection), 변수 표현(feature representation)이다. 문자열 사전 처리은 Hagenau et al. (2013) 연구에서 제시한 방법을 사용하였다.

변수 추출은 영향을 미칠 수 있는 변수들을 생성하는 과정이다. 문서에서 단어들과 그 조합을 추출하는 과정이 이 부분에 속한다. Nassirtoussi et al. (2014) 에 따르면, 단어 주머니 접근법이 텍스트 마이닝을 통한 주가 예측분야에서 가장 많이 사용되는 변수 추출방법이며, Hagenau et al. (2013) 에서 그 효율성과 정확도를 입증하였다. 따라서 본 논문에서는 단어 주머니 모형을 사용한다. 단어 주머니 모형을 수행하기 위해서, 이메일 등 의미 없는 문자열을 먼저 제거하고, 형태소 분석을 통해 구두점 등을 제거하며 단어들의 원형을 찾는다. 이때, 나온 원형의 단어들이 기사를 나타내는 변수다.

변수 선택은 단어 주머니 모형에서 찾아낸 수많은 변수 중에서 주가에 영향을 미치는 것들을 골라내는 것이다. 예를 들어, 형태소 분석을 통해 나온 결과가 변수 추출 단계의 결과인데, ‘를’, ‘을’ 등은 주가의 방향성을 예측하는 데 도움을 주지 않고, ‘호재’와 같은 단어는 영향을 줄 것이다. 주가의 방향성을 예측하는 데에서 설명력을 가지는 것을 판단할 때는 주로 카이 스퀘어를 사용한다 (Groth and Muntermann, 2011; Hageneu et al., 2013; Shynkevich et al., 2016). 카이스퀘어는 카이제곱 분포에 기초한 통계적 방법으로, 관찰된 빈도 O_{ij}가 기대되는 빈도, E_{ij}와 의미 있게 다른 지 여부를 검증할 때 사용된다. 이때 카이스퀘어 값은 다음과 같다.

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

카이스퀘어 값이 높을 수록 변수가 관찰 빈도와 기대 빈도가 통계적으로 유의미하게 의미가 있으며, 변수의 예측력이 높다는 의미이다. 따라서 카이스퀘어 값이 높은 변수들을 선택한다.

변수 선택을 거쳐 영향력이 높은 변수들을 선택한 후에, 기계 학습에 적용하기에 적합한 형태로 변경을 해주어야 한다. 이때, 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. 따라서 주로 사용되는 형태는 TF-IDF 가중치를 사용한다 (Fung et al., 2003; Zhai et al., 2007; Groth and Muntermann, 2011; Hageneu et al., 2013).

2.3 기계 학습 예측 기법

모든 준비과정이 끝나면, 기계 학습 방법으로 문자열 정보에서 표현된 변수를 가지고 주가를 예측한다. 기존의 많은 연구자가 문자열 정보에 기계 학습을 접목하였기에, 수많은 방법이 사용됐다. 대표적으로 서포트 벡터 머신(Support Vector Machine) (Fung et al., 2005; Hagenau et al., 2013; Groth and Muntermann, 2011), K-근접 이웃 (K-Nearest Neighbors) (Groth and Muntermann, 2011), 나이브 베이즈 (Naïve Bayes) (Gidofalvi and Elkan, 2001) 등이 사용 되어져 왔다. 특히, Groth and Muntermann (2011)에서는 인공신경망, 서포트 벡터 머신, 나이브 베이즈, 그리고 K-근접 이웃을 이용하여 결과를 비교하였다. 저자는 기계학습 알고리즘으로 텍스트 분석를 이용한 위기 관리와 투자 의사 결정을 하였다. 저자는 결과와 시간의 효율성을 모두 고려할 때, 서포트 벡터 머신을 추천하였다. 특히 Hagenau et al. (2013) 에서는 시장의 흐름을 긍정, 부정으로 나누고 금융 메시지를 이용하여 예측하는 알고리즘을 구축하였다. 저자는 서포트 벡터 머신, 인공신경망, 나이브 베이즈를 사용하였다. 이때, 서포트 벡터 머신이 다른 알고리즘보다 더 뛰어난 성능을 보였다. 이러한 연구의 흐름을 볼 때, 텍스트 데이터를 가지고 주가를 예측할 때, 서포트 벡터 머신이 가장 보편적이고 정확한 알고리즘 중 하나라고 생각할 수 있다.

최근에 앙상블 기법이 주식 시장을 예측하는 데 많이 사용되고 있다. 앙상블 기법은 여러 가지 기계 학습 방법을 동시에 사용하여 과적합을 방지하여 알고리즘이 여러 상황에 적용하기 용이하기 하며, 정확도를 높이는 방법이다. 그중에서도 특히, 텍스트 마이닝으로 주가를 예측하는 연구에서는 주로 서포트 벡터 머신 알고리즘의 연장선인 다중커널학습을 주로 사용한다(Deng et al., 2011; Wang et al., 2012; Shynkevich et al., 2016). 서포트 벡터 머신과 다중커널학습의 가장 큰 차이는 서포트 벡터 머신은 하나의 특성을 가지는 데이터만을 사용할 수 있으나, 다중커널학습은 서포트 벡터 머신의 다양한 커널을 결합하여 예측하는 개념으로 다양한 데이터를 동시에 사용할 수 있다는 점이다.

다중커널학습을 이용한 연구들의 예시는 다음과 같다. Deng et al. (2011) 에서는 여러 가지의 정보를 통합하여 주가를 예측하는 시스템을 구축하였다. 뉴스 텍스트 데이터, 댓글 수 등을 사용하였다. 각각의 데이터에서 다양한 변수 추출을 하여 다중커널학습을 하였다. 이때, 여러 다

른 변수들을 사용하여 합치는 것이 단일 커널들보다 더 뛰어난 효과를 보였다. 또한, Wang et al. (2012)에서는 레디얼 기저 커널(Radial basis kernel)을 이용한 다중커널학습을 사용하여 예측하였으며 이는 단일 커널보다 더 뛰어난 알고리즘을 보여주었다.

Shynkevieh et al. (2016) 에서는 국제 산업 분류 표준 체계에 기반을 두어 다른 관련성을 가지는 여러 금융 그룹들의 뉴스 데이터를 다중커널학습을 사용하여 주가를 예측하였다. 이때, 기업의 뉴스 데이터만을 사용한 것 보다, 관련성을 가지는 기업들의 뉴스들의 텍스트를 사용하여 예측하는 것이 더 뛰어난 예측 결과를 나타내었다. 하지만 복잡한 시스템의 이질성을 줄이기 위해 기업들을 군집 분석으로 동기화를 높여 예측하는 논문은 없었다. 따라서 본 논문에서는 기존 연구에서 사용한 것과 같이 단어 주머니 모형으로 변수를 추출하고 카이스퀘어 변수 선택과 TF-IDF 가중치 준 뒤에 다중커널학습방법을 사용하였다.

2.4 동종 그룹 로컬 모델

금융 시장은 다른 특성을 가지는 기업들이 서로에게 영향을 주고받는 가장 대표적인 복합 시스템 중 하나이다 (May et al., 2008). 복합시스템은 대체로 이질적인 특성을 많이 가지는 데, 이와 같은 특성은 각 시스템 마디들의 동기화를 방해한다. 이때, 군집 분석을 사용하면 동기화를 높일 수 있고, 복합시스템을 좀 더 잘 분석하고 예측할 수 있다 (Motter et al., 2005).

이를 입증하는 실례로 Cherif et al. (2011)에서는 자기 구성 지도를 이용한 시계열 군집화를 하여, 시계열을 예측하는 모델을 실험하였다. 이때, 이질적인 특성을 가지는 시계열을 군집화를 하지 않은 것 보다, 군집화를 하여 같은 알고리즘으로 예측한 것이 월등한 결과를 보였다.

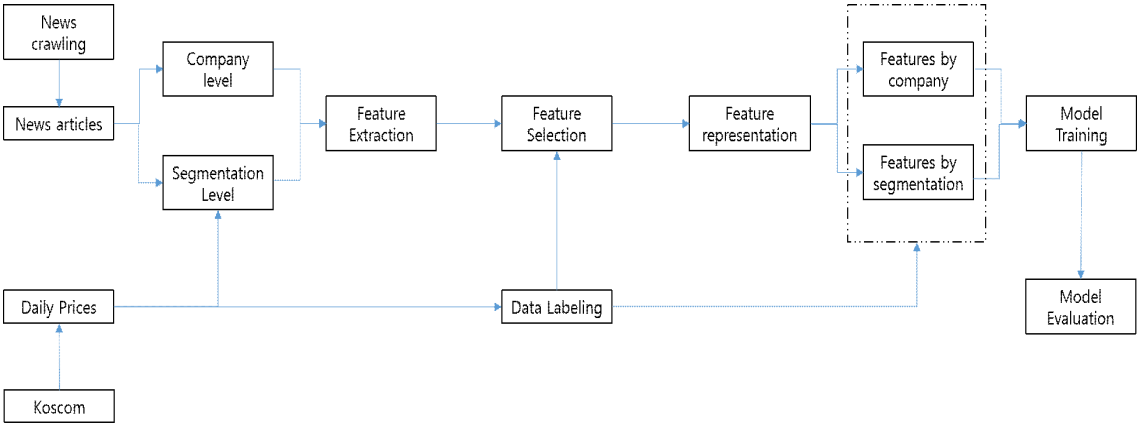
따라서 본 논문에서는 Shynkevich et al. (2016)에서처럼 단순히 국제 산업 분류 표준 체계로 주가를 예측하는 것이 아닌, 동질적인 그룹으로 주가 예측을 하는 것이 아닌, 군집 분석을 통해 그룹의 동기화를 높이는 방법을 고안하여, 더 높은 관련성을 가지는 군집으로 주가를 예측하였다.

3. 연구 모델

이번 장에서는 K-평균 군집 분석 기반 주식 시장 세분화를 적용하여, 뉴스 기사로 주식 동향 예측 시스템을 구축하는 방법에 관해 자세한 설명을 한다. 먼저, 3.1장에서는 본 논문에서 사용하는 데이터에 관해 설명한다. 3.2장에서는 K-평균 군집 분석을 통해 주식 시장 세분화를 하는 방식에 관해 설명한다. 3.3장에서는 텍스트 사전처리 방식에 관해 설명하며, 3.4장에서는 본 논문에서 사용한 EasyMKL 알고리즘에 관해 설명한다. 마지막으로 3.5장에서는 평가 방법에 관해 설명한다. 이 과정에 대한 대략적인 개요는 <Figure 1>에 나와있다.

3.1 뉴스 데이터

본 논문에서 제시하는 방법이 실제로 효과가 있는 지를 검증하기 위하여 실제 데이터를 가지고 실험을 하였다. 데이터는 2014년 1월 1일부터 2016년 12월 31일의 금융 뉴스와 주가 데이터로 이루어져 있다. 뉴스 정보를 구축하기 위해서 한국 최대 포털사이트 네이버에 등록된 10개의 종



〈Figure 1〉 Proposed Approach

합 신문과 14개의 방송 통신 신문과 9개의 경제 신문, 총 33개의 인터넷 뉴스의 모든 금융, 경제 관련 뉴스를 크롤링하였다. 이는 한국에서 대중이 접할 수 있는 대다수의 금융 뉴스를 포함한 정보로, 금융 뉴스가 미치는 영향을 파악하기 좋은 데이터이다. 이 기간에 크롤링된 뉴스는 중복된 것을 제외하자 데이터는 총 1,397,800개 있었다. 뉴스 데이터의 형식은 분류(경제, 금융, 정치), 제목, 작성이, 작성 시간, 내용이 있다.

뉴스가 주가에 미치는 영향에 따라 뉴스를 분류하는데, 각각의 뉴스는 모두 주가가 오르고 내림에 따라 표현된다. 예를 들어 월요일 11시에 작성된 A라는 기업에 대한 뉴스가 수익률이 1보다 크거나 같다면 1, 1보다 작다면 0으로 표현한다. 하지만 한국의 주식 시장은 9시에 열고 16시에 닫기 때문에, 16시 이후에 나온 뉴스는 다음 날에 영향을 미친다고 해석을 하며, 휴일에 나온 뉴스는 다음 주식시장 영업일에 영향을 미치게 하였다 (Li et al., 2014).

각각에 기업에 관련성이 높은 뉴스를 추출하기 위해서, 본문에 회사의 이름이 포함된 뉴스를 할당한다. 이때, 실험에 사용한 모든 기업과 각 기업에 해당하는 등락의 개수는 <Table 1>과 같다.

3.2 관련 회사의 식별

본 연구에서는 산업군 내 이질성을 해결하기 위해 산업군 내에서 군집 분석을 시행하여 동질한 군집을 찾는 과정을 시행한다. 군집 분석을 시행하는 데는 다양한 방법이 있다. K-평균은 주어진 데이터를 k개의 군집으로 묶는 알고리즘으로 각 군집과의 거리 차이의 분산을 최소화하는 방식으로 동작하며 (MacQueen, 1967), DBSCAN은 군집들이 일정 이상의 밀도를 가지도록 하는 방법으로 군집 분석을 한다(Ester et al., 1996). 또한, CLARANS (Ng and Han, 1994), BIRCH (Zhang et al., 1996) 등이 주로 사용된다. 그중에서도 사용하기 간단하고, 많은 데이터를 처리하기 좋으며, 널리 사용되는 알고리즘은 k-평균이다. 따라서 본 논문에서는 산업군 내 이질성을 해결하기 위해 K-평균 컬러스터링

〈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	Daewoong Pharm.	873	pharmacy	436	437
Huchems Fine Chemical	175	material	94	81	Green Cross	1583	pharmacy	830	753
Kukdo Chemical	85	material	45	40	Yuhan	854	pharmacy	444	410
Hyundai-Steel	3741	material	1858	1883	Jeil Pharmaceutical	155	pharmacy	89	66
NamHae Chemical	179	material	104	75	Bukwang Pharm.	218	pharmacy	110	108
Hansol Chemical	110	material	71	39	Hanmi Pharm.	3051	pharmacy	1212	1839
Foosung	295	material	142	153	Dong-A ST	497	pharmacy	263	234
SKC	1184	material	658	526	Boryung Pharm.	583	pharmacy	312	271
SKChemical	1276	material	653	623	Hanall BioPharma	193	pharmacy	99	94
SeAh Steel	374	material	207	167	JW Pharm.	489	pharmacy	231	258
KISWIRE	166	material	92	74	C.K.D	1200	pharmacy	625	575
KiscoHolding	762	material	387	375	Yungjin Pharm.	171	pharmacy	84	87
Korea Zinc	619	material	378	241	Ildong Holdings	46	pharmacy	30	16
Ssangyong Cement Industrial	390	material	258	132	Hanmi Science	610	pharmacy	267	343
Lock&Lock	652	material	381	271	Dong-A Socio Holdings	309	pharmacy	126	183
Korea Petrochemical Ind.	190	material	95	95	Il-Yang Pharm	303	pharmacy	180	123
SamKwang Glass	176	material	80	96	Kwang dong Pharm.	721	pharmacy	410	311
Young Poong	970	material	441	529	CJ CheilJedang	5828	food	3019	2809
Hanwha Chemical	2088	material	1116	972	Samyang	342	food	190	152
Poongsan	1112	material	593	519	Ottogi	2087	food	1100	987
Lotte chemical	2992	material	1515	1477	Hitejinro	3764	food	2008	1756
DongKuk Steel Mill	2079	material	1081	998	Namyang	1375	food	800	575
Taekwang Ind.	327	material	150	177	Muhak	1981	food	926	1055
SeAh Besteel	362	material	200	162	KT&G	3349	food	1803	1546
POSCO	1022	material	509	513	Nonhshim	4093	food	1990	2103
Kolon Ind.	972	material	481	491	Farmsco	236	food	114	122
LG Chem	6717	material	3592	3125	Orion	2915	food	1473	1442
Lotte find Chemical Co.	178	material	108	70	Samyang Holdings	309	food	207	102
-	-	-	-	-	Dongwon F&B	1528	food	752	776
-	-	-	-	-	Lotte Chilsung	2932	food	1472	1460
-	-	-	-	-	Binggrae	1205	food	544	661
-	-	-	-	-	HiteJinro Holdings	127	food	61	66
-	-	-	-	-	Lotte Food	1737	food	893	844
-	-	-	-	-	Lotte Confectionery	3937	food	2154	1783

을 사용하였다.

K-평균 군집 분석은 n개의 데이터를 k개의 군집으로 나누는 것이다. 데이터를 $x_1, x_2, x_3, \dots x_n$이라고 하고, 군집들의 중심을 $\mu_1, \mu_2, \dots, \mu_k$라고 하며, K-평균 컬러스터링에 의해 나뉜 군집을 $\vec{S} = \{S_1, S_2, S_3, \dots, S_k\}$라고 할 때, 알고리즘은 다음과 같다 (MacQueen, 1967).

$$\arg\min_{\vec{S}} \sum_{i=1}^{k} \sum_{S_i} \|x - \mu_i\|^2 \quad (2)$$

즉, K-평균 군집 분석은 군집의 중심과 그 군집에 포함된 데이터들의 거리를 최소화하는 방식이다. 알고리즘은 초깃값 설정을 한 후에 중심점을 찾아가는 방법을 적용한다.

초깃값 설정은 데이터에서 무작위로 하나의 중심을 고르고, 그 중심으로부터 가장 가까운 데이터들과의 거리 D(x)를 구한 뒤, 다른 데이터를 $D(x)^2$에 비례하는 가중 확률을 주어 무작위로 새로운 중심으로 구한다. 이와 같은 과정을 반복하며 k개의 중심을 고른다. 이와 같은 초깃값 설정 알고리즘을 k-평균++(Arthur and Vassilvitskii, 2007)이라 하는데, k-평균++는 데이터가 많아질수록 시간이 기하급수적으로 많이 걸리는 k-평균 군집 분석의 문제점을 해결하였기에 본 논문에서는 효율적인 계산을 위해 k-평균++을 사용하였다.

초깃값을 설정한 이후, 데이터들은 다음과 같은 조건을 만족하는 군집에 할당된다.

$$S_i = \{x_p : \|x_p - \mu_i\| \leq \|x_p - \mu_j\|, \forall j, i \neq j\} \quad (3)$$

이후, 다시 중심을 계산하고, 군집에 다시 할당되며, (2)를 만족하는 배열을 찾아간다.

K-평균 군집 분석은 시행하기 전에, 몇 개의 군집을 가질 지를 선정해야 하는데, 이것이 군집 분석의 매개변수들 중에서 가장 중요한 문제이다(Jain, 2010). 잠재계층분석 (Latent Class Analysis)과 같이 통계적으로 유의미한 군집의 개수를 선정할 수도 있으며 (Lazarsfeld and Henry, 1968), 순차적으로 K를 늘려가며 다른 군집에 비해 얼마나 자기 군집과 비슷한지 측정을 하며 최적의 K를 설정할 수 있다(Rousseeuw, 1987). 또한 순차적으로 K를 늘려가며 베이지안 정보 기준을(Bayesian Information Criterion) 최대화시키는 방법 또한 있다하지만 본 연구는 일차적으로 동질 산업으로 묶여 있는 상태에서 산업군 내에서 이질적인 점을 잡아내는 것이므로 많은 수의 K는 필요 없는 실정이다. 따라서 본 연구에서는 최근 데이터 마이닝 분야에서 많이 활용 되는 그리드 서치(Grid Search)와 같이 K를 2부터 적용해 봄으로써 최적일 때의 결과를 적용한다. 분석 결과 이미 한 차례 동질적인 산업군으로 구분되었기 때문에 K가 4를 넘어야 최적의 결과를 보이는 것은 없었다. 따라서 본 연구에서는 K를 4까지로 결과를 보여주며 해석하기로 한다.

3.3 텍스트 사전 처리

텍스트 전처리는 텍스트 마이닝에서 가장 중요한 부분이며, 뉴스 기반의 주가 예측 시스템을 구현하는 데에 있어 가장 핵심적인 부분 중 하나이다. 먼저, 뉴스에서 필요 없는 부분을 제거해야 한다. 이메일과 HTML 태그와 같은 불필요한 정보들을 모두 제거한다. 2장에서 설명한 것과 마찬가지로 단어 주머니 모형을 사용하기위해, 단어들을 모두 원형으로 만들어야 한다. 따라서 한국어 형태소 분석기 중 가장 많이 사용되는 꼬꼬마 형태소 분석기 (Lee et al., 2010)를 이용하여 모두 단어의 원형으로 만든다. 각각의 원형의 단어는 하나의 변수를 의미하며, Shynkevich et al. (2016)과 마찬가지로 3개 이사의 기사에서 언급된 단어의 원형들은 모두 제거한다.

변수 선택은 2장에서 언급한 카이스퀘어 방법을 사용하여 기업별로 영향을 미치는 변수 중 카

이스퀘어가 가장 전체 변수의 상위 10%로 하여 선택을 하여 기계 학습의 입력 값을 생성한다. 이때 나온 변수들은 500~1000개로 Hagenau et al. (2013)의 567개 단어 주머니를 위한 변수 선택과 Shynkevich et al. (2016)의 500개와 일치하는 값을 가진다.

변수 선택을 거친 후, 각각의 변수에 가중치를 주게 되는 데, 이는 2장에서 언급한 TF-IDF 방법을 사용한다. 단 이때, TF-IDF를 거치면 변수의 값들이 너무 작게 되어 기계 학습을 효율적으로 하기 위해서 단위 조정이 필요하다. Shynkevich et al. (2016)고나 같은 방법으로 TF-IDF로 표현된 변수에 선택한 변수의 개수 만큼을 곱해준다. 즉, 카이스퀘어 분석 결과 상위 10%의 변수의 개수가 k개였다면, k*TF-IDF가 될 것이다. 본 논문에서는 여러 가지의 변수를 사용하기 때문에 그 영향력이 한쪽으로 편향되어있지 않기 위해서는 필수적인 단위 조정 작업이다.

3.4 다중커널학습(Multiple Kernel Learning)

다중커널학습은 여러 가지 특성을 고려한 텍스트 마이닝에서 주로 쓰이는 기계학습 알고리즘이다 (Deng et al., 2011; Wang et al., 2012; Shynkevich et al., 2016). 다중커널학습은 약한 커널 $K_1, K_2, K_3, \dots K_n$을 정의한 뒤에, 이를 선형으로 조합하여 계산하는 알고리즘이다.

이를, 수식으로 표현하면 다음과 같다.

$$K = \sum \beta_i K_i, \forall \beta_i \geq 0 \quad (4)$$

하지만 다중커널학습을 최적화하는 문제는 어려운 문제이다. 약한 커널들의 단순한 평균보다 뛰어난 값들을 가지는 것을 찾는 것이 매우 어려우며, 계산상의 복잡함을 가지고 있다 (Aiolli et al., 2015). 따라서, 이를 계산하는 것은 많은 시간과 메모리를 소요하며, 이를 해결하기 위해 많은 노력이 있었다 (Sun et al., 2010; Jain et al., 2012; Aiolli and Donini, 2015). 그중에서도 특히, Aiolli and Donini(2015)가 제안한 EasyMKL은 힐버트 공간에서 약한 커널 조합 벡터 $\vec{\beta}$와 확률 분포 $\vec{\gamma}$를 조절하여 긍정과 부정 표본들의 거리를 최대화시키는 최적화 알고리즘을 개발하였다. 이를 수식으로 나타내면 다음과 같다.

$$\max_{|\beta|=1} \min(1-\mu)\,\beta^T \hat{Y}\left(\sum_r^R \beta_i \hat{K}_i\right)\hat{Y}\beta + \mu|\beta|^2 \quad (5)$$

이 방법은 기존에 최고의 방법으로 생각되는 SPF-GMKL(Jain et al., 2012)보다 여러 가지 데이터에서 월등한 AUC 점수를 보였을 뿐만 아니라, 더 효율적인 메모리 사용량을 보이기에 본 연구에서는 EasyMKL을 사용하였다 (Aiolli and Donini, 2015).

본 연구에서는 각각의 커널은 대상이 되는 기업과 그 기업과 동질적인 특성을 가지는 기업들을 군집 분석으로 구한 군집에 할당을 한다. 우리의 텍스트 데이터에 어떤 커널이 가장 좋은 성능을 보일지 알지 못해, 선형 커널, 3차 다항 커널, 가우시안 커널과 그 조합들, 선형 가우시안, 다항 가우시안, 선형과 다항을 사용하였다. 즉, 각각의 회사는 선형 커널, 가우시안 커널, 다항 커널 중 1개 혹은 2개와 자신과 자신이 속한 그룹의 조합으로 최소 2개, 최대 4개의 커널을 할당받는다.

이때 비교군은 기존 연구에서 사용되던 두 가지 방법이다. (1) 개별 기업만을 가지고 3가지 커널로 서포트 벡터 머신을 해서 비교를 한다. (2)

군집으로 동질적인 특성을 가지는 그룹을 찾아주는 것이 좋은 결과를 나타낸다는 것을 보여주기 위하여 국제 산업 분류 표준 체계에 기반을 둔 다중커널학습 알고리즘으로 예측하여 비교를 하였다.

3.5 평가 방법

본 연구에서는 2014년 1월부터 2016년 12월까지 총 3년 간의 뉴스 데이터와 3년 간의 주가 데이터를 사용하였다. 이때 검증을 하기 위하여 훈련 기간과 예측 기간을 2년 6개월, 6개월로 설정으로 하여, 훈련 기간 동안 매개 변수를 찾고, 예측 기간 동안 그 결과가 실제 상황에서 잘 맞는지를 확인을 한다. 이때, 실험의 평가 방법은 정확도로 한다. 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. 이때 이를 표로 표현하면 다음 <Table 2>와 같고 이와 같은 표를 오차 행렬(Confusion Matrix)이라고 한다.

〈Table 2〉 Confusion Matrix

		Prediction	
		Up	Down
Actual	Up	TP	FN
	Down	FP	TN

정확도는 오차 행렬에서 $\text{Accuracy} \stackrel{\text{def}}{=} \frac{TP+TN}{TP+FP+FN+TN}$로 정의가 된다.

4. 결과

<Table 3>은 본 실험의 결과를 나타낸다. 먼저, 3장에서 언급한 것과 마찬가지로, 어떤 커널이 텍스트 데이터를 처리하기에 가장 적합한지 알 수 없으므로, 본 논문에서는 Shynkevich et al. (2016)과 마찬가지로 3가지 커널을 조합하여 단일 커널과 2가지 조합 커널을 사용하는 방식을 채택하였다. 이때, <Table 3>의 ‘poly’는 3차 다항 커널, ‘rbf’는 가우시안 커널, ‘lin’은 선형 커널을 의미하며, ‘lp’은 선형 커널과 다항 커널이 사용된 것, ‘rp’는 가우시안 커널, 다항 커널이 사용된 것을 의미하고, ‘lr’은 선형 커널과 가우시안 커널이 된 것을 의미한다.

<Table 3>에서 첫 번째 줄은, 소재, 제약, 음식료 세 개 군집에서 Hagenau et al. (2013) 의 방법대로 각각의 회사를 각각의 회사에 할당된 뉴스의 데이터만을 가지고 서포트 벡터 머신으로 예측하거나, 하나의 데이터를 두 가지 커널에 할당하여 다중커널학습으로 예측한 결과를 의미한다. 이때, ‘poly’, ‘rbf’, ‘lin’은 서포트 벡터 머신을 한 것이며, ‘lr’, ‘rp’, ‘lp’는 하나의 데이터에 대하여 다중커널학습을 한 것이다. 두 번째 줄은 각각의 회사와 그에 해당하는 국제 산업 분류 표준 체계의 그룹을 함께 고려한 것을 의미한다. Shynkevich et al. (2016)의 방법과 같이, ‘poly’, ‘rbf’, ‘lin’은 개별 수준과 산업군 수준의 영향력의 정도를 확인하기 위하여, 회사와 산업군에 같은 커널을 할당한 것이며, ‘lr’, ‘rp’, ‘lp’는 그 영향력의 결합을 보기 위하여, ‘lp’의 경우 회사에 선형 커널, 다항 커널을 할당하며 산업군에도 마찬가지로 두 개의 커널을 할당한다. 3-5 번째 줄은 본 연구에서 제시하는 방법으로, 산업군들을 K-평균 군집 분석을 한 뒤에, ‘Sector’에서와 같

〈Table 3〉 Experimental results for the MKL approach

		Food Expenses		Pharmacy		Material	
Individual level	poly	0.6060	0.6009	0.5964	0.5961	0.6262	0.6224
	rbf	0.6004		0.5881		0.6245	
	lin	0.5937		0.5995		0.6155	
	lr	0.5933		0.5995		0.6155	
	rp	0.6060		0.5964		0.6262	
	lp	0.6061		0.5959		0.6266	
Sector	poly	0.6124	**0.6075**	0.5935	0.5979	0.6242	0.6236
	rbf	0.6025		0.6097		0.6299	
	lin	0.6025		0.5992		0.6198	
	lr	0.6028		0.5991		0.6198	
	rp	0.6124		0.5935		0.6242	
	lp	0.6128		0.5924		0.6240	
Group K= 2	poly	0.6044	0.5997	0.6006	**0.6021**	0.6249	0.6199
	rbf	0.6019		0.6081		0.6224	
	lin	0.5906		0.6016		0.6114	
	lr	0.5904		0.6016		0.6114	
	rp	0.6044		0.6006		0.6249	
	lp	0.6030		0.6003		0.6246	
Group K= 3	poly	0.6054	0.6021	0.5988	**0.6021**	0.6343	0.6291
	rbf	0.5974		0.6027		0.6276	
	lin	0.5996		0.6067		0.6218	
	lr	0.5996		0.6067		0.6216	
	rp	0.6054		0.5988		0.6343	
	lp	0.6046		0.5990		0.6340	
Group K= 4	poly	0.6079	0.6002	0.5992	0.5972	0.6468	**0.6348**
	rbf	0.6011		0.5993		0.6241	
	lin	0.5896		0.5941		0.6223	
	lr	0.5896		0.5941		0.6223	
	rp	0.6079		0.5992		0.6468	
	lp	0.6051		0.5971		0.6467	

은 방식으로 할당한 값들이다. 그룹 K=2는 군집의 개수를 2개로 한 것, 그룹 K=3는 군집의 개수를 3개로 한 것, 그룹 K=4는 군집의 개수를 4개로 한 것을 의미한다. 이때, 산업군 - 소재, 제약, 음식료 - 별로 가장 좋은 결과를 보여주는 값을 진하게 표시해놓았다. 각 실험은 세그멘 테이션의 초기값을 변경하고 각종 매개 변수 값들을 변경해 가며 100차례 실험을 수행하였으며, 이를

기반으로 통계적 유의미한 차이가 존재하는지를 티 검증(t-test)을 통해 확인할 수 있었다.

먼저, 산업군 별로 개별 수준과 'Sector'를 본다면 음식료 산업군의 개별 수준에서는 0.60086의 평균값을 보이며, 'Sector' 에서는 0.60749의 평균값을 보인다. 또한, 제약 산업군의 개별 수준에서는 0.59592의 평균값을 보이며, 'Sector' 에서는 각 커널 별로 0.59784의 평균값을 보이며, 소개 산업군의 개별 수준에서는 0.62235의 평균값을 보이며, 섹터 수준에서는 0.62357의 평균값을 보인다. 이는 개별 수준으로 예측을 하는 것 보다, 관련성을 가지는 다른 데이터들로도 함께 예측을 하는 것이 더 높은 예측력을 보이며, 유의미함을 알려준다. 이 결과는 Schumaker and Chen (2009)에서 실험하였던 것과 마찬가지로 자신과 관련 있는 회사들로 묶인 산업군의 정보 또한 유의미한 정보를 포함한다는 것을 다시 한 번 증명해주며, Shynkevich et al. (2016)처럼 관련성 있는 기업의 뉴스와 자신의 뉴스를 함께 고려할 때 더 뛰어난 영향력을 가진다는 것을 확인할 수 있다.

'Sector'와 본 논문에서 제시한 군집 분석 방식을 비교한다면, 음식료 산업군에서는 'Sector'가 0.60749로 가장 좋은 값을 보여줌을 알 수 있고, 제약 산업군에서는 그룹 K= 2와 그룹 K= 3이 평균적으로 같은 예측력을 보인다는 것을 알 수 있으며, 소재 산업군에서는 그룹 K= 4가 가장 좋은 예측력을 보여준다는 것을 알 수 있다. 이때 이 결과는 소재와 같이 산업군 내 이질성이 높은 산업군에서는 군집 분석을 통해 각각의 요소, 기업들이 서로에게 영향력을 가지는 복잡계 내에서 군집 분석을 통해 동질적인 그룹으로 예측하는 것이 요소 간의 동기화를 높여 더 높은 예측력을 보여주기 때문이며, 음식료 산업군이나 제약 산업군과 같이 이미 동질적인 특징을 가지는 그룹에서는 굳이 군집 분석을 하여 다른 기업들의 효과들을 덜 고려하는 것이 의미있는 관계성을 제거하는 것이므로, 작은 군집 수로 군집 분석을 하거나, 군집 분석을 하지 않는 것이 더 유리함을 알 수 있다. 즉, 산업군 내에서 주가가 동질적일 때 군집 분석을 하지 않고 산업군 수준으로 관련성 효과를 사용하거나 적은 수의 군집으로 나누어 사용하는 것이 중요하며, 산업군 내에서 주가가 이질적인 경우에는 군집 분석을 하여 기업들을 동질적인 그룹으로 묶어 예측하는 것이 예측력을 높이는 방법이다.

5. 결론

정보의 양이 급증하게 되어 주식에 관련된 정보의 양이 무수히 많아지게 되는 빅데이터 시대에, 개인이 모든 뉴스를 읽고 주가에 영향을 미치는 정보만을 선별적으로 찾아내어 이용하는 것은 물리적으로 불가능해졌다. 이와 함께, 수많은 텍스트 정보들을 자동적으로 처리하고 예측할 수 있는 알고리즘을 개발하는 것이 주요한 과제가 되었다. 특히, 수 많은 정보들 중에 영향을 주는 정보를 어떻게 선정하는 지도 주요한 과제가 되었다. 일반적으로는 제목에 회사의 이름이 있거나, 뉴스의 태그의 그 회사의 틱커가 있으면 그 회사에 영향을 주는 정보라고 인식을 한다. 본 연구에서는 영향의 범위를 각 회사와 동질적인 패턴을 보이는 그룹으로 확장하여 각 개별뿐만 아니라 영향력을 줄 수 있는 기업들도 함께 고려하여 예측을 할 때 어떻게 성능이 좋아지는 지에 관하여 연구를 하였다. 본 연구에서 제시한 방법을 다중커널학습방법을 사용하여 예측

한 결과 기존의 국제 산업 분류 표준 체계로 예측하거나, 개별 회사 단위로 예측하는 것보다 더 높은 예측률을 보였다.

본 연구에서는 동질적인 패턴을 보이는 그룹을 찾기 위하여, 주가의 흐름이 비슷한 기업들을 찾았고, 이를 군집 분석을 하여 동질적인 그룹을 구성하였다. 이때, 대상이 된 산업군은 국제 산업 분류 표준 체계에서 소재, 음식료, 제약이었다. 이를 선정한 기준은 산업군 내에서 이질적인 큰 산업군와 작은 산업군들을 선별적으로 선택하는 것이었다. 소재 산업군는 산업군 내에서 기업들의 주가 분산이 커, 이질성이 크며, 음식료와 제약은 이질성이 작았다. 이때, 본 연구에서 주장하는 바와 같이 산업군 내에서 이질성이 큰 그룹에서는 군집 분석을 하여 동질적인 그룹을 만들어 예측하는 것이 도움되며, 이질적인 작은 그룹에서는 산업군 자체로 예측을 하는 것이 도움된다는 것을 볼 수 있었다.

하지만 본 연구는 다음과 같은 한계점을 가진다. 첫 번째, 본 논문에서는 K-평균 군집 분석을 시행하였는데, 이는 동질한 그룹을 찾을 수 있지만, 그 안에서 계층적 구조를 알기 어렵고, 기업간의 거리를 반영하기 힘들다는 단점이 있다. 후속 연구에서 동질한 군집 내에 있더라도, 다른 관련성을 가지는 형태를 반영하는 연구를 할 수 있을 것이다. 두 번째, 본 논문에서는 주가의 동향을 가지고 군집 분석을 시행하였는데, 이는 주가가 비슷하게 움직인다면 동질성이 높은 주식이라는 가정하에 있다. 하지만, 같은 뉴스가 나올 때 혹은 같은 키워드가 나올 때, 같은 방향성의 움직임을 가지는 주식들이 뉴스로 주가를 예측함에서 동질성이 높은 주식이라고 생각할 수 있다. 이 점을 반영한 연구를 후속 연구로 진행할 수 있을 것이다.

참고문헌(References)

Aiolli, F., and M. Donini, “EasyMKL: a scalable multiple kernel learning algorithm,” *Neurocomputing*, Vol. 169, (2015), 215-224.

Arthur, D. and S. Vassilvitskii, "k-means++: the advantages of careful seeding". *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. (2007), 1027–1035.

Cherif, A., H. Cardot, and R. Boné, “SOM time series clustering and prediction with recurrent neural networks,” *Neurocomputing*, Vol. 74, No. 11(2011), 1936-1944.

Deng, S., T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, “Combining technical analysis with sentiment analysis for stock price prediction,” *In Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on* (2011), 800-807.

Ester, M., H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” *In Kdd*, Vol. 96, No. 34(1996), 226-231.

Fung, G. P. C., J. X. Yu, and H. Lu, “The Predicting Power of Textual Information on Financial Markets,” *IEEE Intelligent Informatics Bulletin*, Vol. 5, No. 1(2005), 1-10.

Gidofalvi, G., and C. Elkan, “Using news articles to predict stock price movements,” *Department of Computer Science and Engineering, University of California, San Diego*, (2001).

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.

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.

Jain, A. K., "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, Vol. 31, No. 8(2010), 651-666.

Jain, A., S. V. Vishwanathan, and M. Varma, "SPF-GMKL: generalized multiple kernel learning with a million kernels," *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining,* (2012), 750-758.

Jeong, J. S., D. S. Kim, and J. W. Kim, "Influence analysis of Internet buzz to corporate performance: Individual stock price prediction using sentiment analysis of online news", Journal of Intelligence and Information Systems, Vol. 21, No. 4 (2015), 37~51.

Kim, Y.-S., N.-G. Kim, and S.-R. Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining", Journal of Intelligence and Information Systems, Vol. 18, No. 2 (2012), 143~156.

Lazarsfeld, P.F. and Henry, N.W., "Latent structure analysis", Boston: Houghton Miffli, (1968)

Lee, D. J., J. H. Yeon, I. B. Hwang, and S. G. Lee, "KKMA: a tool for utilizing Sejong corpus based on relational database," *Journal of KIISE: Computing Practices and Letters*, Vol. 16, No. 11(2010), 1046-1050.

Lee, M. and H. J. Lee, "Stock Price Prediction by Utilizing Category Neutral Terms: Text Mining Approach", Journal of Intelligence and Information Systems, Vol. 23, No. 2 (2017), 123~138.

Li, Q., T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements," *Information Sciences*, Vol. 278, (2014), 826-840.

Li, X., C. Wang, J. Dong, and F. Wang, "Improving stock market prediction by integrating both market news and stock prices," *Database and Expert Systems Applications*, *Lecture Notes in Computer Science,* Vol. 6861 (2011), 279–293.

MacQueen, J., "Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability," Vol. 1, No. 14(1967) 281-297.

Mittermayer, M., "Forecasting intraday stock price trends with text mining techniques," *Proceedings of the 37th Annual Hawaii International Conference on System Sciences,* (2004), 1–10.

Motter, A. E., C. S. Zhou, and J. Kurths, "Enhancing complex-network synchronization," *EPL(Europhysics Letters)*, Vol. 69, No. 3 (2005), 334.

Nassirtoussi, A.K., T.Y. Wah, S.R. Aghabozorgi, and D.N.C. Ling, "Text mining for market prediction: a systematic review," *Expert Systems with Applications*, Vol. 41, No. 16(2014), 7653–7670.

Ng, R. T., and J. Han, "Efficient and effective clustering method for spatial data mining," *In Proceedings of VLDB* (1994), 144~155.

Rousseeuw, P. J., "Silhouettes: a graphical aid to

the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, Vol. 20 (1987), 53-65.

Schumaker, R. P., and H. Chen, "A quantitative stock prediction system based on financial news," *Information Processing & Management*, Vol. 45, No. 5(2009), 571-583.

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, (2016), 74-83.

Sun, Z., N. Ampornpunt, M. Varma, and S. Vishwanathan, "Multiple kernel learning and the SMO algorithm," *In Advances in neural information processing systems*, (2010), 2361-2369.

Wang, F., L. Liu, and C. Dou, "Stock market volatility prediction: a service-oriented multi-kernel learning approach," *2012 IEEE Ninth International Conference on In Services Computing (SCC)* (2012), 49-56.

Yeh, C.-Y., C.-W. Huang, and S.-J. Lee, A multiple-kernel support vector regression approach for stock market price forecasting, *Expert Systems with Applications,* Vol. 38, No. 3(2011), 2177–2186.

Zhai, Y., A. Hsu, and S. K. Halgamuge, "Combining news and technical indicators in daily stock price trends prediction," *In Proceedings of the 4th international symposium on neural networks: advances in neural networks, Part III* (2007), 1087–1096.

Zhang, T., R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *In ACM Sigmod Record* Vol. 25, No. 2(1996), 103-114.

Abstract

Online news-based stock price forecasting considering homogeneity in the industrial sector

Nohyoon Seong* · Kihwan Nam**

Since stock movements forecasting is an important issue both academically and practically, studies related to stock price prediction have been actively conducted. The stock price forecasting research is classified into structured data and unstructured data, and it is divided into technical analysis, fundamental analysis and media effect analysis in detail.

In the big data era, research on stock price prediction combining big data is actively underway. Based on a large number of data, stock prediction research mainly focuses on machine learning techniques. Especially, research methods that combine the effects of media are attracting attention recently, among which researches that analyze online news and utilize online news to forecast stock prices are becoming main.

Previous studies predicting stock prices through online news are mostly sentiment analysis of news, making different corpus for each company, and making a dictionary that predicts stock prices by recording responses according to the past stock price. Therefore, existing studies have examined the impact of online news on individual companies. For example, stock movements of Samsung Electronics are predicted with only online news of Samsung Electronics. In addition, a method of considering influences among highly relevant companies has also been studied recently. For example, stock movements of Samsung Electronics are predicted with news of Samsung Electronics and a highly related company like LG Electronics.These previous studies examine the effects of news of industrial sector with homogeneity on the individual company. In the previous studies, homogeneous industries are classified according to the Global Industrial Classification Standard. In other words, the existing studies were analyzed under the assumption that industries divided into Global Industrial Classification Standard have homogeneity.

However, existing studies have limitations in that they do not take into account influential companies

* KAIST College of Business, Korea Advanced Institute of Science and Technology (KAIST)
** Corresponding Author: Kihwan Nam
College of Business, Hanyang University
222-1, Wangsimi-ro, Seongdong-gu, Seoul, South Korea 133-792
Tel: +82-10-4930-8317, E-mail: namkh@hanyang.ac.kr

with high relevance or reflect the existence of heterogeneity within the same Global Industrial Classification Standard sectors. As a result of our examining the various sectors, it can be seen that there are sectors that show the industrial sectors are not a homogeneous group. To overcome these limitations of existing studies that do not reflect heterogeneity, our study suggests a methodology that reflects the heterogeneous effects of the industrial sector that affect the stock price by applying k-means clustering. Multiple Kernel Learning is mainly used to integrate data with various characteristics. Multiple Kernel Learning has several kernels, each of which receives and predicts different data. To incorporate effects of target firm and its relevant firms simultaneously, we used Multiple Kernel Learning. Each kernel was assigned to predict stock prices with variables of financial news of the industrial group divided by the target firm, K-means cluster analysis.

In order to prove that the suggested methodology is appropriate, experiments were conducted through three years of online news and stock prices.

The results of this study are as follows. (1) We confirmed that the information of the industrial sectors related to target company also contains meaningful information to predict stock movements of target company and confirmed that machine learning algorithm has better predictive power when considering the news of the relevant companies and target company's news together. (2) It is important to predict stock movements with varying number of clusters according to the level of homogeneity in the industrial sector. In other words, when stock prices are homogeneous in industrial sectors, it is important to use relational effect at the level of industry group without analyzing clusters or to use it in small number of clusters. When the stock price is heterogeneous in industry group, it is important to cluster them into groups.

This study has a contribution that we testified firms classified as Global Industrial Classification Standard have heterogeneity and suggested it is necessary to define the relevance through machine learning and statistical analysis methodology rather than simply defining it in the Global Industrial Classification Standard. It has also contribution that we proved the efficiency of the prediction model reflecting heterogeneity.

Key Words : Stock prediction, Text Mining, Machine Learning, Multiple Kernel Learning, Clustering

Received : November 10, 2017 Revised : April 40, 2018 Accepted : May 24, 2018
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 경영공학부에서 MIS 박사학위를 취득하였다. 현재 한양대학교 경영대학 겸임교수로 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. IT기업들의 경영 전반에 걸친 문제들을 데이터 기반으로 접근함으로써 보다 효과적인 마케팅 및 경영전략을 수립하는데 관심을 가지고 있다.