R과 분석 알고리즘을 활용한 기업의 성장성 예측에 관한 연구

강희석*, 김경수**, 류지승***, 이가연****, 이민정****
*아시아나 IDT
**숙명여자대학교 통계학과
***이화여자대학교 통계학과
****동덕여자대학교 정보통계학과
e-mail : crazypterpan@naver.com

A Study of Prediction on Company's Growth with R and Analysis Algoritnm

Hui-Seok Kang*, Kyung-Su Kim**, Ji-Seung Ryu***
Ga-Yeon Lee****, Min-Jung Lee****

*Asiana IDT
**Dept. of Statistics, Sook-Myung Woman's University
***Dept. of Statistics, Ewha Woman's University
**** Dept. of Information Statistics, Dong-Duk Woman's University

요 약

기업의 성장성과 기업 주식가치를 매출, 매출원가, 영업이익율 등의 정형데이터와 경제, 경영 관련 뉴스 등 비정형 데이터를 토대로 다양한 알고리즘을 활용해 분석하고, 그 결과의 유의성을 검증한다. 주성분회귀분석, 인공신경망, 나이브 베이지안 분류자, 긍/부정 사전분석 모델을 통해 분석된 결과를 검토하여 각 분석모델 별 성능을 확인하고, 기업 성장성 예측을 위해 활용 가능한 모델과 필요한 데이터를 제시한다.

1. 서론

기업의 성장성은 해당 기업의 이해관계자뿐만 아니라 잠재적 투자자, 일반 대중에 이르기까지 다양한 계층에게 매우 큰 관심사항이라고 할 수 있다. 이렇게 지대한 관심을 가지고 있음에도 불구하고 증권사나 펀드사와 같은 기업체가 아닌 경우, 다양한 기업관련 데이터를 분석하고 해당 기업의 성장성에 대해 의사결정을 내리는 것은 일반 투자자에게는 매우 어려운 일이다. 그러나 최근에는 빅데이터 분석, 공공데이터 개방 트렌드 의 확산과 각종 표준 API 의 배포 등으로 막대한 양의 기업 관련 데이터를 활용할 수 있어 적절한 분석기법과 모델이 제공된다면 누구나 정확한 기업 데이터를 기반으로 신뢰성 높은 의사결정을 진행할 수 있는 시대가 되었다. 데이터 분석을 위한 도구 측면에서도 오픈소스 기반의 데이터 분석 도구인 R 등이 대중화되면서 간단한 스크립트 언어인 파이선(Python)을 기반으로 여러 가지 복잡한 통계 분석을 비교적 쉽게 활용할 수 있다.

따라서 이러한 데이터들과 도구를 잘 활용한다면 자의적이고 비효율적인 의사결정 에서 벗어날 수 있을 것으로 판단하였다.

본 고에서는 기업의 정형 데이터인 매출, 매출원가, 영업이익, 영업이익율, 당기순이익 등 재무적인 정형데이터와 비정형 데이터인 기업관련 경제, 경영뉴스를 웹 상에서 추출하고 수집, 사전 가공한 후 이를 토대로 분석모델을 적용할 경우 어떠한 알고리즘과 프로세스를 적용하는 것이 기업의 성장성 분석을 위해 가장 바람직한 결과를 도출하여 주는지 평가하고 판단하는 과정을 기술한다.

본 논문의 구성은 다음과 같다. 2 장에서는 정형데이터인 기업의 재무정보를 주성분 회귀분석과 인공신경망을 기반으로 분석하는 과정을 설명한다. 3 장에서는 비정형 데이터인 기업관련 뉴스를 웹 크롤링 기법을 통해 수집하고 나이브 베이지안 알고리즘을 통해 1 차 분류하며 최종적으로 긍/부정 사전을 통해 뉴스의 성향을 분석하는 과정을 설명한다. 4 장에서는 각 알고리즘 적용을 통해 얻어진 결론을 제시한다.

2. 정형데이터 분석

주식시장에서의 주가 예측은 주가를 형성하는 변수를 중심으로 한 회귀분석과 과거 주가의 추세를 바탕으로 변동을 제거하여 주가를 예측하는 시계열 방법이 주로 사용되었다. 회귀분석의 경우 주어진 변수들을 토대로 분석하는데 시계열 방식보다 난이도가 낮아 적용하기 쉬운 장점이 있으며 시계열 방법의 경우 정확성에서는 회귀분석보다 나은 결과를 보여주나 시간의 흐름에 따른 데이터가 대량으로 필요한 단점이 있다. 기업의 재무제표 데이터는 분기에 한 번 공시되는 특성에 따라 데이터가 한정적인 특성을 고려하여 우선 재무정보를 중심으로 한 주성분회귀분석을 실시하기로 하였다.

2-1. 주성분회귀분석

일반적으로 다중회귀모형에서 변수들 간의 상관관계가 높을 경우 다중 공선성이 발생할 수 있다. 이 문제를 해결하기 위하여 주성분분석(PCA, Principal Component Analysis)을 실시하고 결과로 생성된 주성분을 설명변수로 사용하여 회귀모형을 적합 시키고자 하였다.

본 연구에서는 예측을 위하여 2012 년에서 2017 년까지의 기업의 분기별 재무제표 데이터를 크롤링(Crawling)하고 원활한 분석을 위하여 각 변수별 값에 대해 전처리 과정을 수행하였다. 먼저, 결측치가 많은 변수와 연도별로 계속 같은 값을 갖는 변수 및 2017 년 이후 날짜를 제거하였다.

재무제표의 특성에 따라 변수가 수익성, 안정성, 성장성의 3 가지 범주로 나뉠 것이라고 예상하고 첫 번째 주성분은 영업이익률, ROA, ROE, 부채비율, BIS 로 정의해 수익성과 안정성에 관련된 지표로 구성하였다. 두 번째 주성분은 FCF, 자산총계, 부채총계, 현금 DPS 로 정의해 성장성과 관련되도록 구성하였다.

그 다음으로는, 분석 편의성을 위해 알파벳으로 각 데이터 별 열(Column)의 이름을 변환해주고, 단위를 통일하기 위하여 표준화 과정을 수행하였다.

마지막으로 기업현황을 통일된 하나의 지표로 판단하기 위하여 기존 변수들을 활용해 Bis(자기자본비율)=자본총계/자산총계, SP(stock price, 주가, 이하 SP 로 표기)=EPS*PER 이라는 새로운 변수를 생성하여 추가하였다. SP 는 예측하고자 하는 최종 Target 으로 설정하였다.

1) Bis(자기자본비율)=자본총계/자산총계
2) SP(stock price)=EPS*PER

(그림 1) 성장성 예측 Target 산식

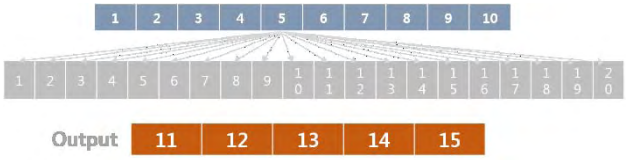
주성분 분석을 실시하는 것이 유의미한지 사전에 파악해 보기 위하여 분산의 균일성을 검사하는 Bartlet 검정을 실시한 결과, 도출된 유의확률은 2.2e-16 보다 작아 유의함을 확인하여 주성분 분석 시행이 의미가 있음을 확인하였다. 또한 주성분 분석을 실시하기 위한 함수로 자료 행렬의 특이값 분해 알고리즘을 통하여 분석을 시행하는 prcomp 함수를 이용한 결과, 누적 설명력이 두 번째 주성분인 기업 성장성 관련 지표에서 90 이 넘는 것을 확인하여 두 번째 주성분까지 사용하여 분석하는 것이 타당하다고 판단되었다.

2 가지 주성분을 독립변수로 하는 회귀분석은 R 에 내장된 lm 함수를 이용하였다. 주성분을 독립변수로 하고 주식의 종가를 종속변수로 하여 회귀 분석을 실시한 결과 회귀식이 Sp= 204104-8455PC1-10838PC2 로 추정되었다. 회귀 모형 적합 결과 유의수준 0.05 하에 유의하지 않은 결과가 산출되었기에 주성분회귀분석과 같은 선형적 방법은 성장성을 시계열에 따라 분석하기에는 부적합한 부분이 있다고 판단하였고, 이를 기반으로 비선형적 방법을 시도할 필요가 있음을 확인하였다.

2-2. 인공신경망

주성분회귀분석, 다중회귀분석과 같은 선형 분석기법들은 결과의 정확도가 상대적으로 떨어지며 데이터에 따라 적용이 제한되는 문제점이 있다. 이를 보완하기 위하여 인공신경망(ANN, Artificial Neural Network) 기법을 사용하였다. 인공신경망은 비모수적 분석 방법으로서 분류 및 예측 분야에서 폭넓게 사용되고 있으며, 경험을 통해 알고리즘 스스로 학습이 가능하여 다양한 분야에서 활용이 증가하고 있다.

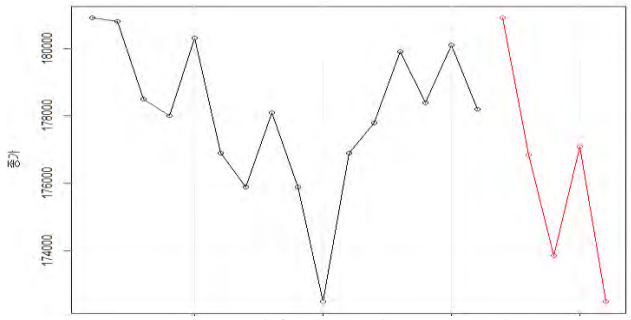
본 연구에서는 예측을 위하여 2010 년 9 월부터 2017 년 9 월까지의 기업의 주식가격 데이터를 사용하였다. 약 1000 개의 학습 데이터를 수집하였고 500 개는 훈련 집합(Training Set)으로 나머지 500 개는 검증집합(Test Set)으로 분류하였다. 실제 분석은 R 의 nnet 패키지와 파이썬의 bpnn 라이브러리를 사용하여 인공신경망을 기반으로 한 분석을 시행하였다. 위와 같이 분석모델을 구성한 후 nnet 함수를 호출하여 인공신경망에 모델을 학습시킨 후 predict 를 호출하여 예측결과를 도출하는 방식으로 진행하였다. 또한 정확한 검증을 위하여 테스트 집합을 통해 교차 오분류표를 구하고 최적의 은닉 노드 수와 가중치 모수값을 선택할 수 있도록 구성하였다.



(그림 2) 신경망 구조(입력, 은닉, 출력노드)

구축된 신경망 모델 구조상에서 은닉층에는 20 개의 은닉노드가 존재하고 있다. 10 개의 데이터를 입력노드에 넣으면 은닉층을 거치면서 5 개의 예측값이 출력된다. 주성분회귀분석을 통한 선형회귀모형보다 비선형회귀 모형인 인공신경망의 예측 오차 비교 결과 주성분회귀분석 모형이 약 10% 더 높은 오차율을 보였으며 인공신경망 모형의 표준오차(RMSE)가 5000 정도 더 낮은 것을 확인하였다. 결론적으로 주가를 기반으로 한 기업의 성장성 분석 시에는 주성분회귀분

석 보다는 인공신경망을 통한 예측이 더 유의하다고 결론 내릴 수 있었다.



(그림 3) 입력층의 주가변동 그래프

3. 비정형데이터 분석

3-1. 웹 크롤링(Web Crawling)

웹 페이지 내의 Dom 구조를 분석하여 필요로 하는 정보를 추출해 내는 과정을 웹 크롤링이라 한다. 본 연구에서는 기업의 평판, 현황, 향후 가능성 등의 데이터를 추출하여 분석하기 위해 사용되었다.
기본적인 과정은 크롤링 대상 페이지를 url 을 통해 접속하고 HTML 파일을 파싱(Parsing)하여 연구자가 원하는 내용을 가져오는 과정이다.
본 연구에서 Python scrapy 패키지를 활용하여 다음 뉴스에서 뉴스 데이터를, R rvest 패키지를 활용하여 구글 파이낸스에서 주가 데이터를 크롤링하였다. 이후 모든 분석 작업은 R 에서 진행하였다.

	Date	Open	High	Low	Close	Volume
1	11.Aug.17	227000	234000	227000	233900	10095
2	10.Aug.17	231500	231500	228500	228600	6917
3	9.Aug.17	231800	232900	230100	230800	4025
4	8.Aug.17	228200	234500	228200	233900	6076
5	7.Aug.17	232000	232000	228600	230500	6723
6	4.Aug.17	229400	237000	228100	231600	15969

(그림 4)크롤링 결과 : Google Finance 데이터

date	title	article
2017-07-12	GS홈쇼핑 식기 브랜드 '코렐' 지분 투자…아시아 주방용품 공략	GS홈쇼핑은 코렐 비전 파이렉스 등의 식기 브랜드를 갖고 있는 기…
2017-07-16	물걸레 로봇청소기 '에브리봇' "바퀴 없이 걸레에 하중 실어 청소실…	물걸레 로봇청소기 에브리봇에는 바퀴가 없다. 대신 바닥에 닿린 동…
2017-06-30	청와대 참모들 보유주식 모두 매각…장하성 실장 54억 '최대'	청와대 참모진이 보유 주식을 매각한 것으로 나타났다. 1급 이상 …
2017-07-14	탈모닷컴 TS삼푸 1000억원 판매와 100만 고객 돌파	탈모방지에 도움을 주는 스테디셀러 TS삼푸를 선보이는 탈모닷컴(…
2017-07-19	나노스 어느새 시총 8위…닷세째 '급등'	가 5거래일째 급등하며 시가총액 8위로 올라섰다.19일 오전 9시7…
2017-07-19	허창수 "불확실할수록 선제적 투자"	"지난 일을 잊지 않고 잘 살펴서 일의 지침으로 삼는다."허창수 GS…
2017-07-17	GS '열린 소통공간' 마련해 직원 참의성 극대화	GS그룹은 구성원들이 원활하게 소통할 수 있는 '열린 조직문화' 정…
2017-07-16	집밥같은 간편식에 빠진 4050 주부들	40~50대 주부들의 가정간편식(HMR) 구매가 빠르게 늘고 있다. …
2017-06-30	코스피 外人·기관 '팔자'에 하락…사상 첫 7개월 연속 상승	코스피지수가 사상 처음으로 7개월 연속 상승했다.30일 코스피지…
2017-07-03	코스피 2380선으로 후퇴…코스닥 660선 위태	코스피지수가 2380선으로 밀려나 하락 흐름을 이어가고 있다.3일…
2017-07-14	코스피 2410선에서 상승세…삼성전자·SK하이닉스 ↑	코스피가 개인의 매수세에 힘입어 이틀째 2410선에서 상승세를 …
2017-07-03	GS 허창수 회장 "부단한 혁신 새 가치 창출"	허창수 GS 회장은 "고객 니즈의 변화를 빠르게 파악하고 유연하게 …

(그림 5) 크롤링 결과 : 다음 뉴스 데이터

3-2. 나이브 베이지안 분류자 (Naïve Bayesian Classifier)

나이브 베이지안 분류자(Naïve Bayesian Classifier)는 텍스트 분류에 사용됨으로써 문서를 여러 범주 중 하나로 판단하는 문제에 대한 대중적인 방법이다. 뉴스의 경우에도 분문에 포함된 단어들을 분류하고 분석에 활용하여 향후 기업의 성장성이 긍정적 방향일지, 부정적 방향일지를 예측하는 것이 가능해진다.
본 연구에서는 크롤링한 전체 기사에서 10 번 이상 반복하여 노출되는 주요 명사데이터들을 활용하여 나이브 베이지안 분류자(Naïve Bayesian Classifier)를 이용한 주가 예측 모형을 제안하고, 10 fold-Cross-Validation 을 사용하여 나이브 베이지안 분류자 알고리즘의 분석성능을 평가하였다. 평가지표로서는 F1-measure, Recall, Precision, Accuracy 의 4 가지를 평가 지표로 활용하였다.

3-3. 긍/부정사전

단어사전 Github 에 오픈소스로 공개되어 있는 한글 긍/부정사전을 기사의 긍/부정 반응평가에 활용하였다. 이 자료는 영어 긍/부정사전을 한국어로 번역한 자료로서, 한글 긍/부정 사전을 활용하여 한 기사당 긍정적인 단어가 나온 횟수, 부정적인 단어가 나온 횟수를 집계하였고, 긍정적인 단어가 더 많이 나왔다면 그 기사는 긍정적인 기사로 분류하였다.
분석을 수행하는 당일, '긍정적인 기사가 많이 나오면 다음날은 해당 기업의 주가가 상승할 것이다'는 가설을 전제로 긍정적인 기사가 부정적인 기사보다 많으면 주가가 상승할 것이라 예측하였다.
긍/부정사전을 활용한 예측방법의 지표는 나이브 베이지안 분류자와 동일하게 사용하였다.

<표 1> 분석 대상 데이터의 사례

구분	주가 데이터	뉴스 데이터
소재	구글 finance	다음 뉴스
데이터 수	495	585
크롤링 도구	R	Python
크롤링 패키지	rvest	scrapy
기간	2015-08-10 ~ 2017-08-10	2015-08-10 ~ 2017-08-10
대상	oo 홈쇼핑	oo 홈쇼핑

전체 뉴스 기사에서 10 번 이상 등장한 단어들을 독립변수로 두고 뉴스 주가가 오를지 내릴지를 종속변수로 하여 나이브 베이지안 분류자 모델에 적용시켰다. 그 결과 F1-measures 는 0.40, Balanced Accuracy 는 0.50 의 값을 얻을 수 있었다.
여기에 긍/부정 사전에 등장한 긍정적인 단어가 나온 횟수, 부정적인 단어가 나온 횟수, 긍정적인 기사 개수, 부정적인 기사 개수를 새로운 변수로 추가하였다. 긍정적인 뉘앙스의 기사와 부정적인 뉘앙스의 기사의 등장 빈도를 비교하여 긍정 뉘앙스의 기사가 더 많으면 기업성장성이 긍정적이며 주가는 오를 것이라고 예측하고, 부정 뉘앙스의 기사가 더 많으면 기업성장성이 부정적이며 주가는 내릴 것이라 가설을 설정하였다. 이러한 가설설정을 토대로 실제 시뮬레이션을 진행한 결과 F1-measures 는 0.60, Balanced Accuracy 는 0.49 를 얻을 수 있었다. 즉, Balanced

Accuracy 는 5%정도 낮았지만 F1-measure 는 20%나 높은 결과를 보여주었다. 결론적으로 가설은 옳다고 결론내릴 수 있었으며 뉴스 기사를 활용한 기업의 성장성 예측 모형을 활용하는 경우에는 나이브 베이지안 분류자 알골리즘을 활용하기 보다는 긍/부정 사전을 활용하여 기사의 뉘앙스를 바탕으로 한 예측이 더 좋은 결과를 얻을 수 있음을 확인할 수 있었다.

<표 2> 성장성 예측 모델별 비교

구분	나이브 베이지안 분류자	긍부전 사전
새로운 변수	10 번 이상 등장한 단어 모두	긍부전 사전 단어
예측 방법	모델링 (나이브 베이즈)	가설 (전 날에 긍정 기사가 많이 나오면 주가가 상승할 것이다.)
F1-measure	0.4084	0.6076
Balanced Accuracy	0.5028	0.4968
Recall	0.3615	0.8432
Precision	0.4958	0.4749

4. 결론

본 연구에서는 기업의 성장성을 예측하기 위해 오픈소스 통계분석 도구인 R 을 활용해 여러가지 분석 기법들을 실제 적용해 보면서 그 성능을 확인해 보았다. 정형데이터를 기반으로 일반적인 회귀분석을 수행했을 경우, 유의한 결론을 얻기 어렵다는 점을 실제 수행을 통해 확인하였으며 인공신경망을 기반으로 데이터를 학습시키고 분석하는 기법이 좀 더 나은 의사결정을 지원할 수 있음을 확인하였다. 또한 여기에 추가로 비정형 데이터인 뉴스의 긍/부정 뉘앙스를 분석하여 의사결정 시에 참고한다면 보다 안정적인 수준의 결론에 도달할 수 있음을 확인할 수 있었다. 향후에는 인공신경망의 정확성을 더욱 보강하기 위하여 보다 많은 훈련 데이터(Training Set)가 필요할 것으로 판단되며 훈련시에 과적합(Overfit)이 발생하지 않도록 유의해야 할 것으로 판단된다. 또한 비정형데이터는 보다 다양한 반응을 분석하기 위해 SNS 의 반응에 대한 강도(Strength)분석이 추가로 필요할 것으로 보이며 문장과 전체적인 맥락을 인식하기 위한 연구를 차후 진행할 계획이다.

참고문헌

[1] 범주형 자료분석 개론. Alan Agresti(2009.08)
[2] 빅데이터 분석을 위한 데이터마이닝 방법론. 강현철, 한상태, 최종후, 이성건, 김은석(2014.03)
[3] Probability and Statistical Inference, Global Edition. RRobert Hogg , Elliot Tanis , Dale Zimmerman(2014.12)
[4] 통계학. 류근관(2013.02)
[5] Statistics 4th Edition. David Freedman(2015.08)
[6] Python Programming: An Introduction to Computer Science. John Zelle(2014.09)
[7] 긍부전 사전 github 자료 : https://github.com