R과 분석 알고리즘을 활용한 기업의 성장성 예측에 관한 연구

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

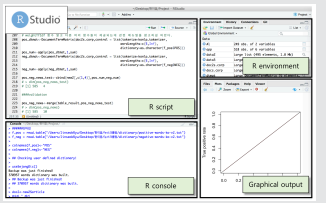
류지승, 이화여자대학교

요약

기업의 성장성과 가치를 매출, 매출원가, 영업이익율 등의 정형데이터와 경제, 경영관련 뉴스 등 비정형 데이터를 기반으로 다양한 알고리즘을 활용해 분석하고, 각 알고리즘이 출력하는 그 결과의 유의성을 검증한다. 오픈소스 도구인 R을 활용해 정형데이터 분석에는 주성분회귀분석, 인공신경망 기법을 사용하고 비정형데이터 분석에는 나이브 베이지안 분류자, 긍/부정 사전분석 모델을 사용해 분석한다. 분석된 결과를 검토하여 각 분석모델별 성능을 확인하고, 기업 성장성 예측을 위해 가장 적합한 모델과 필요한 데이터를 제시한다.

연구 목적

기업의 성장성은 해당 기업의 이해관계자 및 다양한 사람들에게 매우 큰 관심사항이다. 그러나 해당 기업의 방대한 데이터를 분석하고 의사결정을 내리는 것은 일반적으로 매우 어려운 일이다. 그러나 최근 빅데이터 분석, 공공데이터 개방 트렌드의 확산과 R과 같은 오픈소스 도구의 사용확대 등으로 적절한 분석기법과 모델을 선택할 수 있다면 누구나 신뢰성 높은 의사결정을 진행할 수 있는 시대가 되었다. 이에 본 연구는 오픈소스 도구 R을기반으로 어떤 알고리즘을 적용하는 것이 가장 바람직한 결과를 도출하여 주는지 평가하고 판단해 보고자 한다.



연구 방법

기업의 정형 데이터인 재무제표상의 데이터를 주성분회귀분석, 인공신경망을 통해 분석해 보고 비정형 데이터인 기업관련 뉴스를 나이브 베이지안, 긍/부정 사전을 통해 분석해 본다. 최종적으로 각 알고리즘별 도출결과의 유의성을 판단해 본다.

Step 1 : 정형데이터 분석 - 주성분 회귀분석(Principle Component Analysis)

- 5년간 분기별 재무제표 데이터를 웹 크롤링(Crawling)을 통해 수집
- 각 변수별 값에 대한 결측치 처리, 스케일링 등 전처리 수행
- 영업이익률, ROA, ROE, 자산총계, 부채총계 등을 주성분으로 범주화
- 컬럼명 변환, 단위 통일 등 분석을 위한 표준화(Nomalization)
- 주성분 회귀분석을 활용하여 모델 적합

Step 2 : 정형데이터 분석 - 인공신경망(Artficial Neural Network)

- 7년간 일별 주가 데이터를 웹 크롤링(Crawling)을 통해 수집
- 약 1000개의 학습데이터를 훈련집합(Training Set)과 검증집합(Test Set)으로 구분
- 훈련집합에의 함수 처리를 통하여 인공신경망 모델 학습
- predict를 호출하여 주가 예측 결과 도출

Step 3 : 비정형데이터 분석 - 나이브 베이지안 분류자(Naïve Bayesian Classifier)

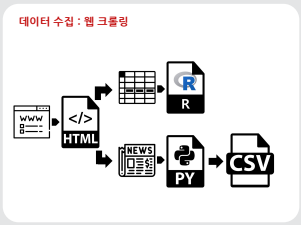
- 1년간 '00홈쇼핑' 뉴스기사를 웹 크롤링(Crawling)을 통해 수집
- 각 변수별 값에 대한 전처리 수행
- 뉴스기사를 명사 단위로 잘라 10번이상 나온 단어들을 이용해 희소행렬로 변환
- 나이브 베이지안 분류자를 활용하여 모델 적합

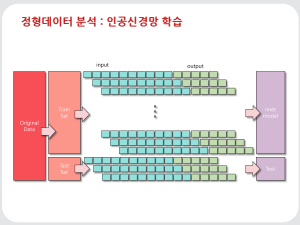
Step 4 : 비정형데이터 분석 - 긍/부정 사전(Positive/Negative Sentiment Dictionary)

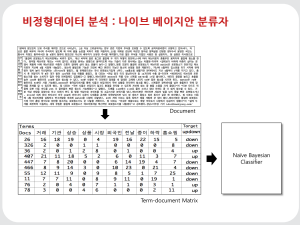
- 영어 긍부정 사전을 번역한 한글 긍부정 사전을 이용하여 희소행렬로 변환
- 파생변수 생성 - 기사별 긍정단어개수, 부정단어개수
- 긍정단어 개수가 많은 기사를 긍정기사라고 분류
- 긍정기사가 많이 나온 날은 주가가 오를 것이라 가설을 세우고 모델 적합

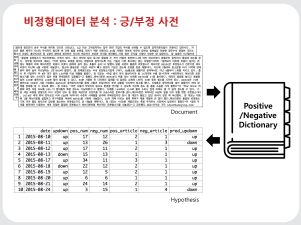
연구 진행과정

분석을 위한 데이터 수집을 위해 웹 크롤링을 수행하고 정형데이터 분석, 비정형 데이터 분석을 수행한다.









결론

본 연구에서는 기업의 성장성을 예측하기 위해 오픈 소스 통계분석 도구인 R을 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 SNS의 반응에 대한 강도(Strength)분석이 추가로 필요할 것으로 보이며 문장과 전체적인 맥락을 인식하기 위한 연구를 차후 진행할 계획이다.