주가 데이터 Back-Testing을 통한 투자 기법 검증 및 주가 예측

김수영
경희대학교 산업경영공학과

1. 서 론

코로나 사태 이후 증시와 은행 예금으로 동시에 자금이 대거 유입되고 있다. 정부와 한국은행의 적극적인 유동성 공급 정책으로 시중 자금이 급증한 가운데, 이 돈이 '모(투자 위험 큰 주식) 아니면 도(안전한 예금)'로 몰리고 있는 것이다. 특히 최근 증시로의 민간 자금 유입은 3 월 29 일 45 조 원(개인 투자자의 예탁금 규모)으로 역대 최대기록을 달성하고, 언론에서는 이를 '동학개미운동' 이라는 말로 대서특필할 정도로 전례 없던 규모이다. 그러나 엄청나게 커진 증시에 대한 개인 투자자들의 관심과 자금 유입에 비해, 개인 투자자들의 증시 관련 정보나 지식 수준은 그에 미치지 못하는 실정이다.

2. 데이터 셋 확보

2.1 주가 데이터

향후 클러스터링, 모델링 등에 필요한 주가 데이터는 2016 년 1월 초부터 현 시점까지의 약 5년 분량의 주가 데이터를 수집하였다. (2016.01.04 ~ 2020.09.29) Yahoo Finance API를 이용하여 코스피 전 종목(795개)에 대한 데이터를 수집할 수 있었다.

2.2 거시경제 데이터

거시경제 데이터 또한 유사한 방식으로 수집할 수 있었다. 그 내역은 아래와 같다.

1) 원/달러 환율 : 일별 원/달러 간 환율
2) 거래대금 : KOSPI 시장 전체에서 거래된 주식의 일별 총 거래대금 (단위 : 1억원)
3) 외국인 순매수 : KOSPI 시장 전체에서 외국인의 일별 총 순매수량 (단위 : 1억원)
4) 전체 시총 : KOSPI 시장의 일별 전체 시가총액 (단위 : 1억원)

2.3 기본적 분석지표 데이터

	일자	종목코드	종목명	관리여부	종가	EPS	PER	BPS	PBR	주당배당금	배당수익률	게시물 일련번호	종카운트
0	2016/01/04	000020	동화약품	0	8,140	177	45.99	8,225	0.99	80	0.98	1	887.0
1	2016/01/04	000030	우리은행	0	8,600	1,621	5.31	26,592	0.32	500	5.81	2	0.0
2	2016/01/04	000040	KR모터스	0	1,305	0	0	385	3.39	0	0.00	3	0.0
3	2016/01/04	000050	경방	0	180,000	4,190	42.96	236,250	0.76	500	0.28	4	0.0
4	2016/01/04	000060	메리츠화재	0	15,900	1,131	14.06	13,190	1.21	380	2.39	5	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
49751	2020/09/01	33637K	두산솔루스1우	0	16,200	0	0	0	0	0	0.00	887	0.0
49752	2020/09/01	33637L	두산솔루스2우B	0	32,450	0	0	0	0	0	0.00	888	0.0
49753	2020/09/01	344820	케이씨씨글라스	0	28,600	0	0	0	0	0	0.00	889	0.0
49754	2020/09/01	353200	대덕전자	0	11,000	0	0	0	0	0	0.00	890	0.0
49755	2020/09/01	35320K	대덕전자1우	0	6,790	0	0	0	0	0	0.00	891	0.0

Fig. 1 기본적 분석지표 데이터 셋

3. 전처리

3.1 주가 데이터 전처리

주가 데이터의 경우, Min-Max 정규화는 수행하지 않았다. 종목 Segmentation을 위해 이용되는 방법론인 K-Means Clustering의 경우 Data point의 거리를 기준으로 클러스터가 형성되기 때문에, 주식의 단위 가격에 영향을 받지 않게 scaling 하였다. 대신 좋은 시각화 결과를 얻기 위해 모든 종목의 시작 시점, 즉 2016년 1월 4일의 가격을 100원이 되게끔 변형하였으며 분석에 있어서는 raw data를 사용하였다.

종목명	2016.01.04		2020.09.29	
	전처리 전	전처리 후	전처리 전	전처리 후
한화생명	7220원	100원	1525원	21.12원

Table 1 주가 전처리 예시

3.2 거시경제 및 기본적지표 분석 데이터 전처리

기본적 분석 데이터는 KRX(한국거래소) 사이트에서 이미 전처리가 수행된 데이터를 사용하였기 때문에 결측치가 발생하는 경우는 거의 없었다. 다만 PER(주가수익비율)에서 결측치가 생기는 경우가 있었는데, 이 경우는 해당 분기의 수익이 적자라 음수를 나누고자 할 때 생기는 오류로 0으로 치환할 수 있었다.

4. 클러스터링

4.1 개요

주가 데이터는 일별 가격을 기준으로 클러스터링을 진행할 수 있었다. 날짜 하나하나가 클러스터링에 있어 attribute 가 됨으로써 주가의 이동이 유사한 종목들끼리 segment를 구성할 수 있게끔 하였다.

4.2 방법론 소개

- k-Means Clustering

k-Means Clustering 알고리즘은 클러스터링 방법 중 분할법에 속한다. 분할법은 주어진 데이터를 여러 파티션 (그룹) 으로 나누는 방법이다. 예를 들어 n개의 데이터 오브젝트를 입력받았다고 가정하자. 이 때 분할법은 입력 데이터를 n보다 작거나 같은 k개의 그룹으로 나누는데, 이 때 각 그룹은 클러스터를 형성하게 된다. 다시 말해, 데이터를 한 개 이상의 데이터 오브젝트로 구성된 k개의 그룹으로 나누는 것이다. 이 때 그룹을 나누는 과정은 거리 기반의 그룹간 비유사도 (dissimilarity) 와 같은 비용 함수 (cost function) 을 최소화하는 방식으로 이루어지며, 이 과정에서 같은 그룹 내 데이터 오브젝트 끼리의 유사도는 증가하고, 다른 그룹에 있는 데이터 오브젝트와의 유사도는 감소하게 된다. k-평균 알고리즘은 각 그룹의 중심 (centroid)과 그룹 내의 데이터 오브젝트와의 거리의 제곱합을 비용 함수로 정하고, 이 함수값을 최소화하는 방향으로 각 데이터 오브젝트의 소속 그룹을 업데이트 해 줌으로써 클러스터링을 수행하게 된다.

-Elbow method

군집 분석에서 Elbow method은 데이터 세트의 군집 수를 결정하는 데 사용되는 휴리스틱 방법론이다. 이 방법은 군집 수를 측정하는 함수로 plotting한 후 곡선의 변곡점 위치를 통해 사용할 군집 수를 선택할 수 있다.

-Silhouette Analysis

Silhouette Analysis(실루엣 분석)는 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지를 나타낸다. 효율적으로 잘 분리

됐다는 것은 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐 있다는 의미이다. 군집화가 잘 될수록 개별 군집은 비슷한 정도의 여유공간을 가지고 떨어져 있을 것이다.

4.3 클러스터링 및 클러스터 평가

주가 데이터의 클러스터링에는 K-Means Clustering이 이용되었으며, 최적의 클러스터 개수를 선정하는 데에는 SSE를 최소화하는 방식으로 heuristic하게 개수를 선정할 수 있는 Elbow Model을 이용하였다.

클러스터링에 앞서 Elbow Model을 통한 클러스터 개수의 평가 결과, SSE 감소 비율이 급격히 작아지는 시점인 3을 선택하여, 클러스터의 군집 수를 3개로 선정하였다.

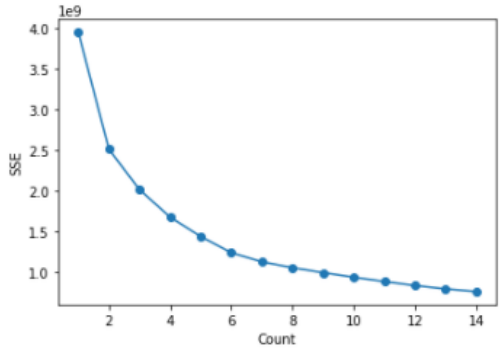

Fig. 2 Elbow Model 적용 결과

최적의 클러스터링 결과를 얻기 위한 Silhouette Analysis 또한 수행되었으며, 그 결과 비슷한 특징을 가진 군집끼리 잘 뭉쳐 있는 결과가 형성되었음을 확인할 수 있었다.

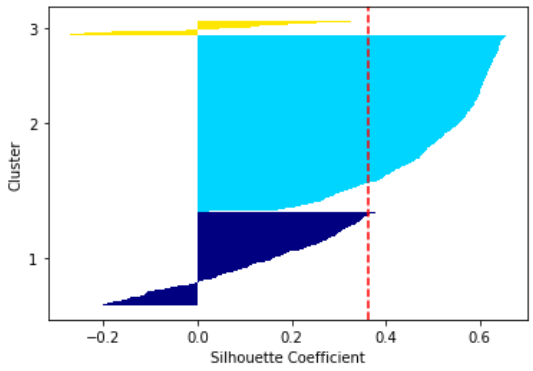

Fig. 3 Silhouette Analysis 결과

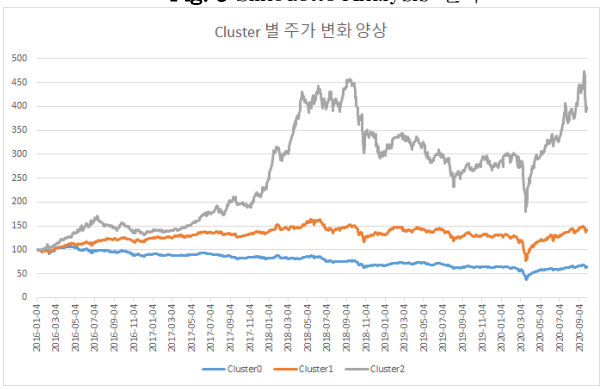

Fig. 4 K-Means Clustering 결과

Fig.4는 클러스터 별로 각 클러스터 내 종목들의 일별 주가를 평균낸 후 하나의 그래프에 도식한 결과이다. 각 클러스터별로 상당히 다른 주가 흐름을 보이고 있으며 이를 정리하면 아래와 같다.

2016년 1월 4일의 클러스터 평균 주가를 100원이라고 하였을 때 (총 3개 클러스터, 719개 종목)

Cluster 0 : 2020년 9월 평균 주가 약 65원 (-35%)
➔ "하락주", 439개 종목

Cluster 1 : 2020년 9월 평균 주가 약 142원 (+42%)
➔ "성장주", 246개 종목

Cluster 2 : 2020년 9월 평균 주가 약 397원 (+297%)
➔ "급등주", 34개 종목

이후 기술적, 기본적 분석에 이용될 수 있게끔 분류된 Cluster와 분석 지표 데이터를 통합하였다.

	일자	종목명	클러스터	관리여부	종가	EPS	PER	BPS	PBR	주당배당금	배당수익률
0	2016년 1월	동화약품	1	0	8140	177	46	8225	1	80	1
1	2016년 1월	KR모터스	0	0	1305	0	0	385	3	0	0
2	2016년 1월	경방	0	0	180000	4190	43	236250	1	500	0
3	2016년 1월	메리츠화재	1	0	15900	1131	14	13190	1	380	2
4	2016년 1월	삼양홀딩스	0	0	151500	327	463	148290	1	1500	1
.	...	...	...	...	...	...	...	...	...	...	...
3	2020년 9월	금호에이치티	0	0	4205	0	0	2439	2	0	0
4	2020년 9월	경보제약	0	0	11800	221	53	6239	2	100	1
5	2020년 9월	토니모리	0	0	10050	0	0	5600	2	0	0

Fig. 5 통합 데이터 셋

5. Back-Testing을 통한 투자 기법 검증

5.0 개요

앞서 분류한 클러스터들을 통해, 각 클러스터 별로 어떤 투자 전략이 유효한 지 검증하기 위하여 Back-Testing을 진행하였다. 클러스터 별로 어떤 기본적 지표에 대해 어떤 기준을 적용했던 종목이 최대의 수익을 달성하였는 지 확인하였다.

또한 투자 기법 별로 Grid-Search를 수행하여 최적의 전략과 최적의 투자 기간을 산출할 수 있었다.

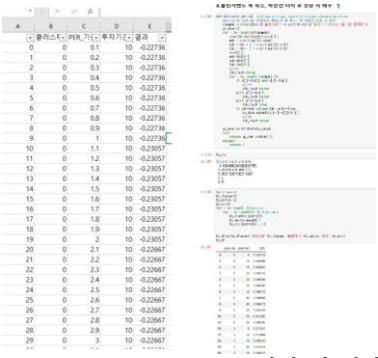

Fig. 6 Back-Testing 과정 및 결과

5.1 Cluster 0 (하락주 클러스터)

- PBR ≤ 1 인 종목에 대해 투자
- 5 일 연속 하락한 종목에 대해 투자
- 볼린저 밴드 하단에서 15~50 일 간 횡보 후 상승 시 투자

- 580 일 이상 투자

5.1 Cluster 1 (성장주 클러스터)

- 5≤ PBR ≤ 19 인 종목에 대해 투자
- 0.1 ≤ PER ≤ 1.0 인 종목에 대해 간 투자
- 4 일 연속 하락한 종목에 대해 투자
- 볼린저 밴드 하단에서 45~55 일 간 횡보 후 상승 시 투자
- 580 ~ 610 일 간 투자

5.1 Cluster 2 (급등주 클러스터)

- PBR ≤ 3 인 종목에 대해 투자
- PER ≤ 4 인 종목에 대해 투자
- 2 일 연속 하락한 종목에 대해 투자
- 볼린저 밴드 하단에서 30~50 일 간 횡보 후 상승 시 투자
- 580 ~ 670 일 간 투자

6. RNN 모형 구성 및 주가 예측

6.1 개요

앞서 분류한 클러스터들을 바탕으로 효과적인 주가 예측을 수행하기 위해 RNN 모형을 구축하였다. RNN(순환 신경망) 은 인공 신경망의 한 종류로, 유닛간의 연결이 순환적 구조를 갖는 특징을 갖고 있다. 이러한 구조는 시변적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해주므로, 순방향 신경망과 달리 내부의 메모리를 이용해 시퀀스 형태의 입력을 처리할 수 있다. 특히 RNN의 한 종류인 LSTM의 경우 RNN의 문제점 중 하나인 장기 의존성 문제를 해결할 수 있어 주가와 같은 시계열 데이터 예측에 적합하다. 이번 연구에서는 LSTM을 사용하였다.

6.2 모델 구성

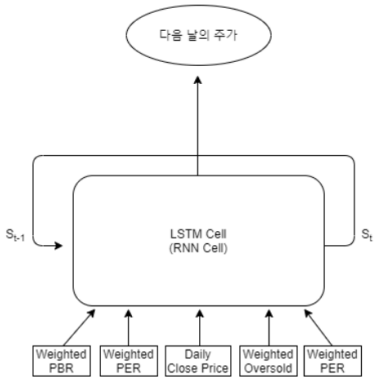

Fig. 7 LSTM Model 개요

모델의 input으로는 당일의 종가, 기본적 지표들인 PBR과 PER, 과매도 데이터가 들어갔다. 이를 바탕으로 다음 날의 주가를 예측할 수 있었다.

Model의 구조는 휴리스틱하게 여러 모델을 시도하였으나, 한 개의 층으로 구성된 Simple한 LSTM 모델의 성능이 가장 좋아 이를 차용하였다.

```
from keras.models import Sequential
from keras.layers import Dense
from keras.callbacks import EarlyStopping, ModelCheckpoint
from keras.layers import LSTM

model = Sequential()
model.add(LSTM(16,
               input_shape=(train_feature.shape[1], train_feature.shape[2]),
               activation='relu',
               return_sequences=False)
          )
model.add(Dense(1))
```

Fig. 8 LSTM Model 구성

6.3 예측 결과

2020년 8월과 9월의 데이터를 test data로 하여 예측을 수행하였다. 모델 예측의 평가 척도는 MAPE(Mean Absolute Percentage Error)를 사용하였으며 이는 오차가 예측값에서 차지하는 정도를 의미한다.

-가격

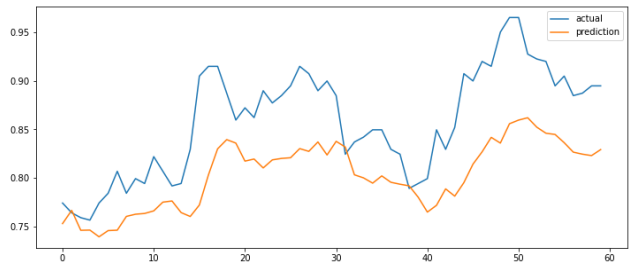

Fig. 8 가격 예측

Window=5로 설정한 후 다음 날의 가격을 예측한 결과, Fig.8과 같은 결과를 얻을 수 있었다. 이 때 MAPE는 6.34이다.

-추세

향후 n일 간의 평균을 예측함으로써 향후 주가의 추세를 예측한다면 더 높은 정확도를 얻을 수 있을 것이라는 판단 하에, 주가의 추세를 예측하고자 하였다.

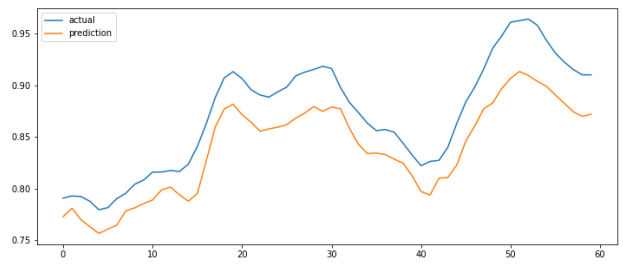

Fig. 9 추세 예측

5일 간의 추세를 예측하였을 때를 시각화한 Fig.9의 경우, 가격을 예측한 Fig.8과 다르게 가격에 비해 추세를 더 잘 예측함을 한눈에 확인할 수 있다. 예측력은 몇 일치를 예측하는 지에 따라 달라지며, 이 차이를 MAPE로 수치화한 값은 아래와 같다.

예측일수	평균 MAPE
3	4.8528
5	5.5176
7	8.1468
20	3.0712
30	5.6121
50	5.3146
100	3.0413

Table 2 추세 예측 결과

-Cluster 예측

앞서 가격과 추세 예측을 통해, 20일 간의 평균을 예측하는 것이 가장 좋은 추정 결과를 얻게 해 줌을 알게 되었다. 이를 바탕으로 각 클러스터 별로 20일 간의 추세 예측을 수행하였고, 그 결과는 아래와 같다.

Cluster	C0	C1	C2
Count	439	246	34
Mean	16.53	10.13	11.42
Median	8.91	7.56	8.34
Std .Dev	73.58	8.54	10.21

Table 3 Cluster 별 예측 결과

6. 결 론

본 연구에서는 5년 간의 주가 데이터를 통해 투자 기법 검증과 주가의 추세 예측을 수행할 수 있었다. 각 종목의 주가 흐름을 클러스터링 하여 성장주, 하락주, 급등주 등을 구분해낼 수 있었고, 이를 바탕으로 최적의 투자 전략과 투자 기간을 요약하고 최종적으로 RNN 모형을 통해 주가의 추세를 예측할 수 있었다. 한편으로는 차후 연구에서 개선되어야 할 사항도 존재한다. 모델과 모델의 파라미터를 개선함과 동시에, 수행한 추세 예측이 어떤 방식으로 활용될 수 있을 지에 대한 활용 방안 역시 강구되어야 할 것이다.

참 고 문 헌

Ruoxuan Xiong and Eric P. Nichols and Yuan Shen(2016), Deep Learning Stock Volatility with Google Domestic Trends

Adil Moghar, Mhamed Hamiche, Stock Market Prediction Using LSTM Recurrent Neural Network, Procedia Computer Science, Miles, M.D. (1986). “Measurement of Six Degree of Freedom Model Motions Using Strapdown Accelerometers”, Proc. of the 21st ATTC, Vol 2, No 2, pp 369-375.

Ankit Thakkar, Kinjal Chaudhari, Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions, Information Fusion, Schlichting, H. (1968). Boundary Layer Theory, 6th ed., McGraw-Hill, New York.

Hiransha M, Gopalakrishnan E.A., Vijay Krishna Menon, Soman K.P., NSE Stock Market Prediction Using Deep-Learning Models, Procedia Computer Science

Si Woon Lee, Ha Young Kim, Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation, Expert Systems with Applications