수치해석 4차과제

주가 데이터 분류

과목명	수치해석
담당교수	심동규 교수님
학과	컴퓨터정보공학부
학년	3학년
학번	2019202009
이름	서여지
제출일	2021.12.1 (수)

KWANGWOON UNIVERSITY

1. 과제 개요

이번 과제는 KRX300종목과 ETF종목의 1년치 종가에 대해서 K—means clustering을 수행하는 것이다. 자료로 주어진 종목은 모두 584개 종목이고, 개장일은 248일이다. 입력 정보에 대하여 종가 전처리 과정을 통해 절대 가격대 차이를 보정하고, Feature 추출 과정을 통해 차원을 축소한다. 그 결과를 이용하여 비슷한 형태를 갖는 종목을 분류한다.

2. 과제 수행 방법

- 종가 전처리

  주식 종목 사이의 절대 가격대 차이를 보정하기 위하여 종가 데이터에 대한 전처리를 수행한다. 당일 종가와 전날의 종가의 비율을 이용하여 일간 변화율을 구하여 종가에 대한 전처리를 수행하였다.

- feature추출

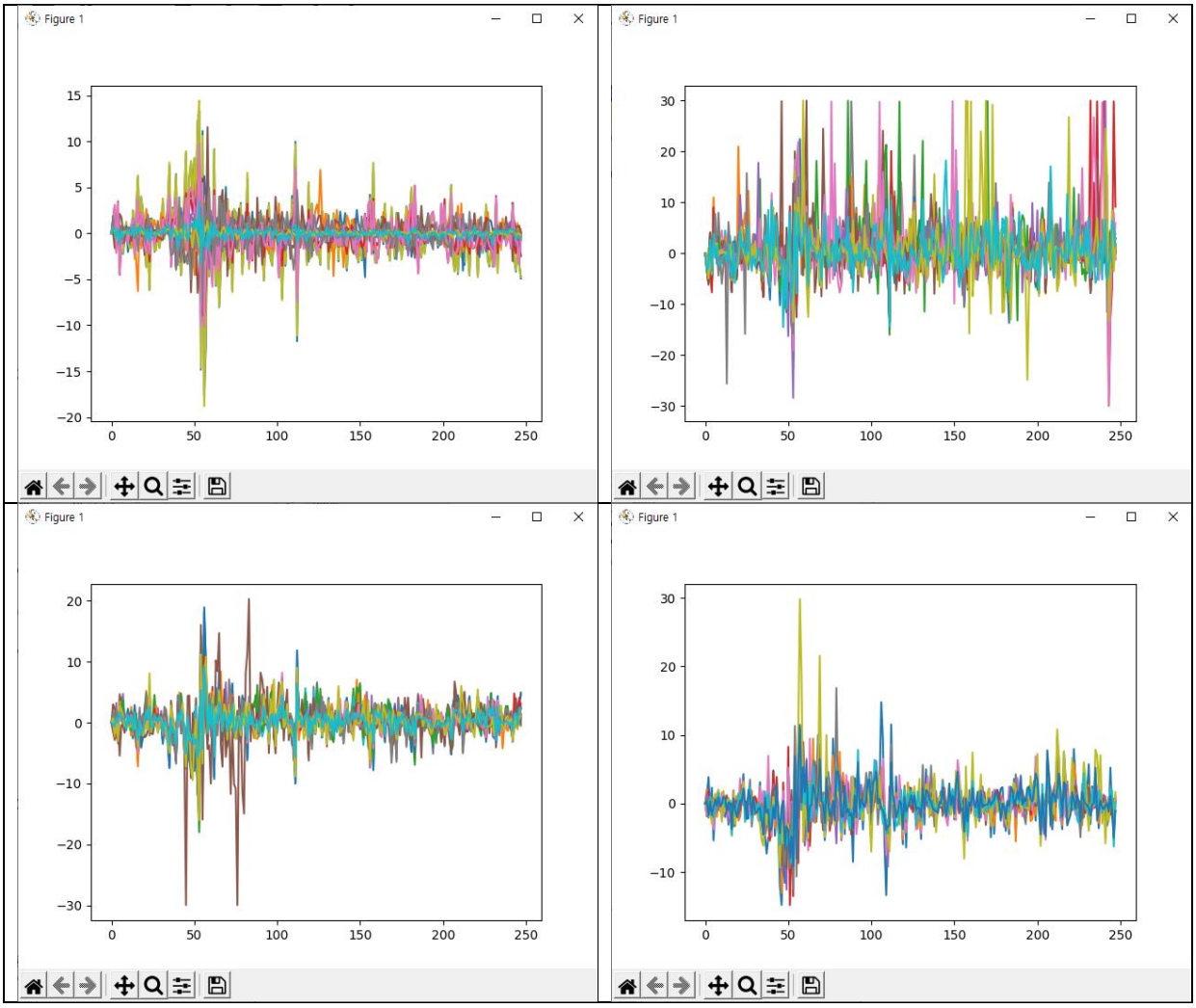
  전처리가 완료된 종가 정보를 갖고있는 data 리스트는 584*248 크기를 갖는다. 정보의 수가 많아서 처리에 사용되는 시간이 늘어나고, 불필요한 세부사항에 의해 clustering의 과정에서 어려움이 나타날 수 있다. 이것을 방지하기 위해 PCA를 이용하여 data의 차원을 축소하였다. PCA는 sklearn 모듈의 함수를 import하여 진행하였다. 또한 해당 모듈에서 제공하는 pca의 variance_ratio의 합을 계산하여 80%의 값을 대표할 수 있는 60개의 주성분을 사용하여 진행하였다.

- K-means clustering

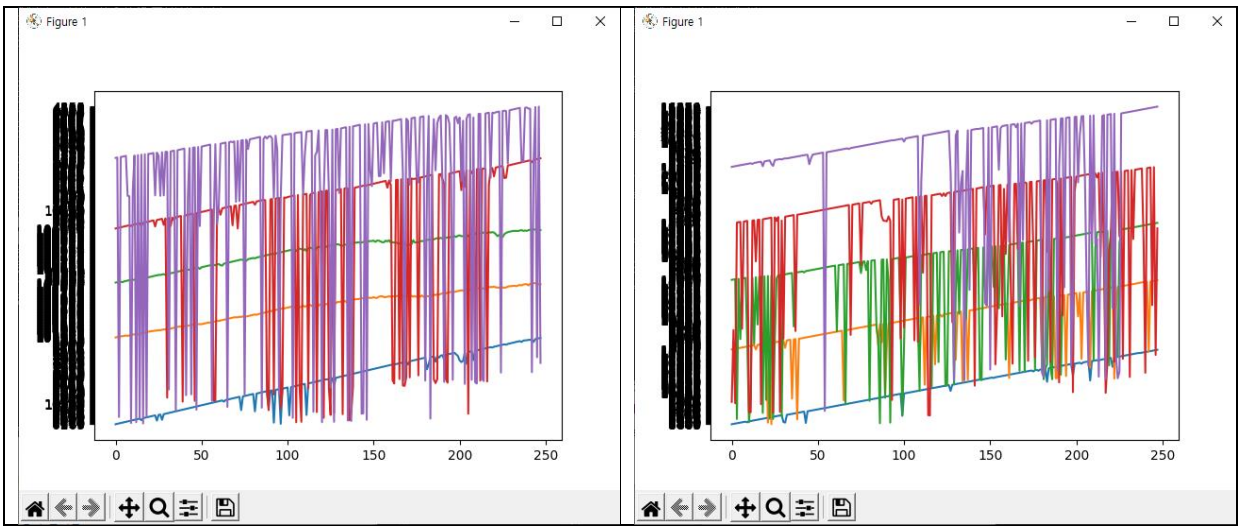
  K means clustering을 이용하여 종목을 K개의 cluster로 분류하였다. 초기 60차원 정보는 모두 0으로 초기화 했으며, numpy모듈의 array 계산을 이용하여 각 종목의 종가 데이터의 일간 변화율에 대하여 유클리드 거리를 계산하여 분류하였다. 또한 새로운 벡터의 값을 결정하는 과정 또한 numpy.array를 이용하여 평균을 계산했다.

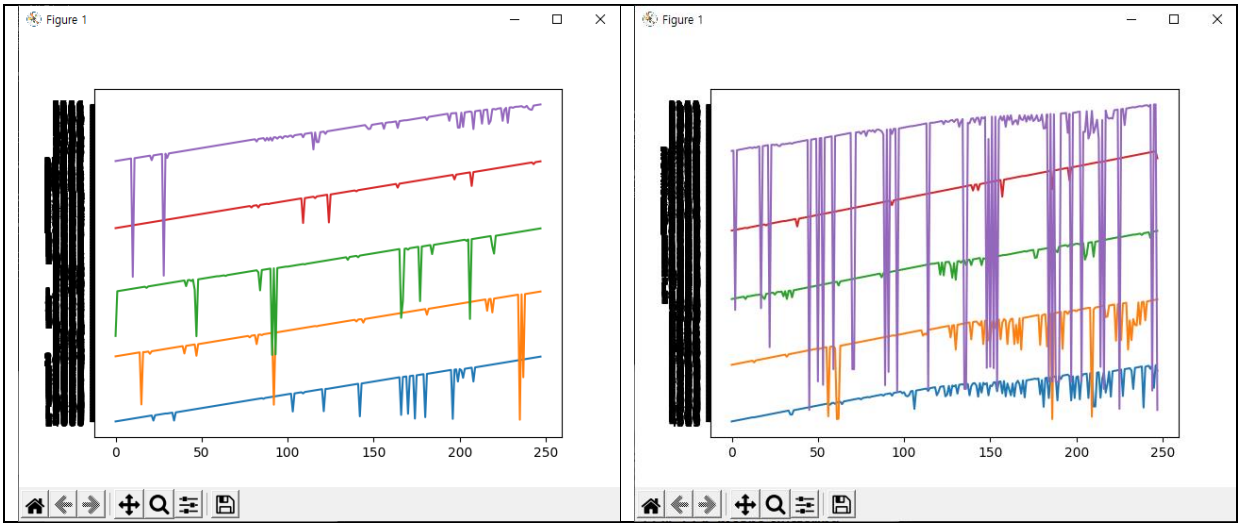
3. 결과

전체 종목을 4개의 cluster로 분류하였고, 최대 20개의 일간 변화율을 그래프로 표시하였다.

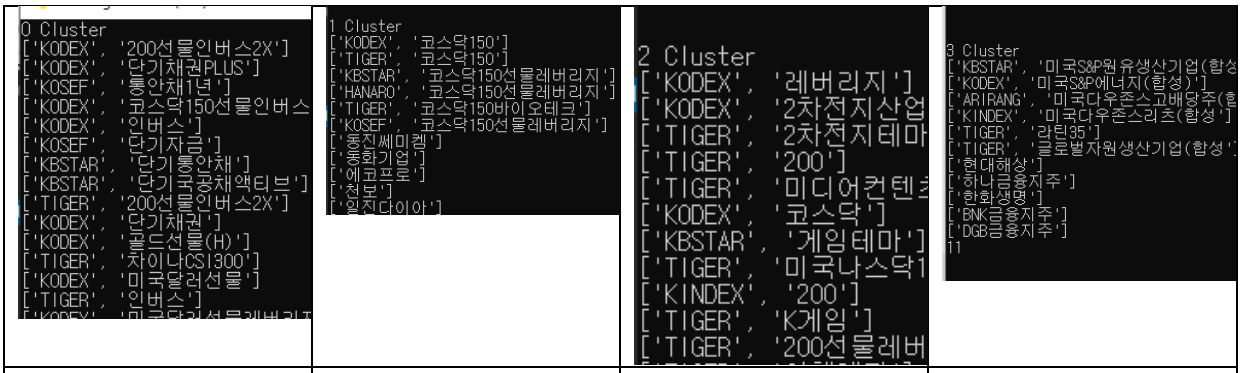


같은 조건에서 cluster별 5개 종목에 대한 원본의 종가데이터를 출력하였다.





cluster별로 분류된 종목의 수는 다음과 같이 나타났다.



119	43	411	11

4. 고찰

분류 결과를 그래프로 확인해도 cluster간의 차이는 보이지 않았다. 또한 특정 cluster에 과다한 종목이 분리되기도 했다. 성능 개선을 위해 k means 알고리즘의 반복 횟수를 증가시키고, data의 차원을 축소하는 것에 PCA가 아닌 다른 방법을 사용할 수 있으며, 종가 데이터를 전처리 하는 과정에서 일간 변화율이 아닌 누적변화율을 사용하거나 적절한 cluster의 수를 구하여 적용하는 등의 방법을 시도할 수 있을 것같다. python의 다양한 라이브러리를 이용하여 많은 데이터를 간단하게 처리할 수 있는 경험이되었다. 방학기간을 이용하여 프로그램을 더 제대로 작성할 수 있을 것이다. 흥미로운 주제의 과제를 제대로 수행하지 못하여 아쉬움이 남는다.