碩士學位論文

인공지능 기술을 활용한 금융시장 분석과 포트폴리오 최적화

嘉泉大學校 大學院

應用統計學科

應用統計學 專攻

朴 翰 相

碩士學位論文

인공지능 기술을 활용한 금융시장 분석과 포트폴리오 최적화

Financial market analysis and portfolio optimization using artificial intelligence technology

嘉泉大學校 大學院

應用統計學科

應用統計學 專攻

朴　　翰　　相

碩士學位論文

指導敎授 高承坤

인공지능 기술을 활용한 금융시장 분석과 포트폴리오 최적화

Financial market analysis and portfolio optimization using artificial intelligence technology

위 論文을 應用統計學 碩士學位 論文으로 提出함.

2020 年 7 月 4 日

嘉泉大學校 大學院

應用統計學科

應用統計學 專攻

朴 翰 相

이 論文을 朴 翰 相의

統計學碩士 學位論文으로 認准함

2020 年 7 月 4 日

審査委員長 고 승 곤 ㊞

審 査 委 員 김 남 형 ㊞

審 査 委 員 한 승 봉 ㊞

국문 초록

인공지능 기술을 활용한 금융시장 분석과 포트폴리오 최적화

가천대학교 대학원 응용통계학과
응용통계학 전공 박 한 상
지도교수 : 김 남 형

최근 투자자들은 단순 편리함을 넘어서 스마트한 금융서비스를 요구하고 있다. 따라서 국내·외 금융회사에서는 Big Data를 활용하여 인공지능을 구현하는 머신러닝 기술을 도입하고 있다. 대표적인 머신러닝 기술도입을 통해 기대하는 금융서비스 중 하나로 Robo-Advisor가 있다. Robo-Advisor의 경우 고도화된 알고리즘과 Big Data를 통해 포트폴리오 관리를 수행하는 온라인 자산 관리 서비스를 의미한다. 본 논문은 머신러닝 기술을 활용하여 국내주식시장에 적용할 수 있는 새로운 혼합 군집화 알고리즘을 제안하였으며, 개별 전자공시자료가 국내주가의 변동성에 유의미한 영향을 미친다는 사실을 확인하여 변동성을 추정할 때 조정하는 방법을 제안하였다. 또한 제안된 방법들이 Robo-Advisor의 기반이 되는 포트폴리오 구성 및 최적화에 활용이 될 수 있다는 것을 실증적으로 증명해보였다는 것에 의미가 있다.

중요 용어: Clustering, DART, KRX, Markowitz Model, Portfolio Optimization, Time-series

<목 차>

<표 목 차>

<그 림 목 차>

제 1 장 연구 배경 및 목적

과거 전통적인 금융 산업에서는 은행, 증권 등 금융사 지점에 직접 방문하는 것이 기본이었다. 금융서비스를 원하는 소비자가 직접 필요한 증빙서류 등을 첨부해 집이나 직장 근처의 점포를 방문해 직원을 직접 대면했다. 하지만 이렇게 시간이 많이 들던 대면 금융서비스는 정보통신기술발달에 힘입어 점차 비대면으로 진화했다. 따라서 현대사회에서는 모바일뱅킹, 모바일증권 등을 통해서 은행지점 또는 증권사에서 하던 서비스가 그대로 내 손 안의 모바일 기기를 통해 가능하게 됐다. 따라서 비밀번호・서명 이미지 등 지식기반 인증방법, SMS・E-mail 등 소지기반 인증방법, 지문・얼굴・홍채・정맥 등 바이오인식기반 인증방법 등 다양한 비대면 인증 방법을 통해 본인인증을 하고, 몇 번의 터치를 통해 각종 은행 업무부터 증권거래까지 가능해졌다. 그러나 최근 사람들은 이러한 단순한 편리함을 넘어서 스마트한 금융서비스를 원하고 있다. 즉, 나에게 적합한, 나만을 위한, 나에게 최적화된 금융상품을 추천받고 가입하고 관리받기를 원한다. 그야말로 지능화된 금융서비스를 원하는 것이다. 이러한 지능화된 금융서비스를 가능하기 위해서는 인공지능의 힘이 필요하다(김지혜 2017).

인공지능이란 인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등을 설계된 일련의 알고리즘을 통해 실현하는 기술을 의미한다. 그리고 인공지능 고도화를 위해서는 인공지능 알고리즘의 기초가 되는 Big Data가 필수적이다. 이러한 Big Data는 최근 의료, 제조, 금융, 교통, 제조 등의 여러 산업분야에서 활용되고 있다. 특히 금융 산업은 의료 산업에 이어 인공지능 기술 활용도가 두 번째로 높을 것으로 기대되는 산업이다(금융보안원 2017). 따라서 국내・외 금융회사에서는 Big

Data를 활용하여 인공지능을 구현하는 머신러닝 기술을 도입하고 있다. 머신러닝은 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 의미한다. 또한 머신러닝은 주어진 데이터를 바탕으로 새로운 질문에 대해 예측하는 것을 목적으로 한다. 이러한 머신러닝 기술의 도입은 소비자에게 양질의 금융서비스를 전달하고 기업의 생산성을 향상시키는 등 다양한 장점을 제공하여 금융 산업의 새로운 성장 동력으로 주목 받고 있다. 대표적인 머신러닝 기술도입을 통해 기대하는 금융서비스로는 시장분석, 금융보안, 신용평가, Robo-Advisor등이 있다. Robo-Advisor의 경우 고도화된 알고리즘과 Big Data를 통해 포트폴리오 관리를 수행하는 온라인 자산 관리 서비스를 의미하는데, 기존의 프라이빗 뱅커가 하던 역할을 수행할 수 있다. 따라서 Robo-Advisor에게 자신의 투자 성향 등의 정보를 알려주면 그 정보를 바탕으로 개인 맞춤형 자산 배분 전략을 짜줄 수가 있다. 여기에는 다양한 인공지능 기술들이 활용되고 있으며 현재도 관련 기술들이 꾸준히 개발되고 있다. 또한 KEB하나은행 하이로보센터에서 발간한 ‘2018 대한민국 로보어드바이저 보고서’에 따르면 Robo-Advisor가 관리하는 자산의 규모가 2018년 1조원이었던 것에 반해서 2025년에는 30조원까지 성장할 것으로 전망하고 있다. 그에 발맞추어 국내 대부분의 증권사들 또한 자체적으로 기술 개발에 나서거나 자문 업체와 제휴를 맺는 방식으로 Robo-Advisor서비스를 제공하고 있다. 서비스 분야 역시 자산관리, 투자자문, 투자추천 등 영역을 넓혀 나가고 있다.

본 논문에서는 인공지능 기술인 머신러닝을 활용하여 Robo-Advisor의 기반이라고 할 수 있는 포트폴리오 구성과 최적화에 적용하고자 한다. 포트폴리오를 구성하는 기초 자산들은 다양성을 가져야 하는데 이를 위하여 주가 자료를 비슷한 특성을 가지는 군집들로 군집화하는 알

고리즘을 제안하여 성능을 평가하였다. 기존의 주가 자료를 군집화하는 것은 주가 시계열 자료를 이용하거나 또는 기업의 재무제표와 관련된 변수들을 활용한 방법이 존재하는데 본 연구에서는 두 가지를 모두 혼합하여 군집을 만드는 새로운 알고리즘을 제안하고자 한다.

또한 포트폴리오 최적화에서는 기초자산의 변동성을 추정하여야 하는데, 과거 시계열 자료만을 이용하여 변동성을 추정하는 것의 한계를 극복하고자 전자공시자료를 활용하였다. 전자공시자료는 기업에 중요한 변동 사항이 생겼을 때 의무적으로 공시하게 되어 있는데 이러한 공시자료들 중 미래의 변동성에 유의미하게 영향을 미치는 공시자료를 찾고 이를 반영하여 변동성을 조정해줌으로써 포트폴리오 최적화에 활용하였다.

본 연구에서는 제안된 방법들을 사용해 국내 대표 주가지수인 KOSPI 200을 구성하는 기업들을 대상으로 포트폴리오 최적화에 적용해 보았다. 제안된 주가 군집화 알고리즘으로 포트폴리오를 구성하는 기초자산을 선택하는데 활용하고, 전자공시자료를 활용한 변동성 조정 방법을 적용하여 포트폴리오 최적화를 진행해본 결과 임의로 생성된 포트폴리오에 대하여 과거 시계열 자료만을 활용하여 최적화를 진행했을 때보다 더 나은 성능을 보여주는 것을 확인할 수 있었다.

이 논문은 국내주식시장에 적용할 수 있는 새로운 혼합 군집화 알고리즘을 제안하였으며, 개별 전자공시자료가 국내주가의 변동성에 유의미한 영향을 미친다는 사실을 확인하여 변동성을 추정할 때 조정하는 방법을 제안하였다. 또한 제안된 방법들이 포트폴리오 최적화에 활용이 될 수 있다는 것을 실증적으로 증명해보였다는 것에 의미가 있다.

이후의 논문 구성은 다음과 같다. 제 2 장에서는 금융시장 군집화와 전자공시자료 분석, 포트폴리오 최적화와 관련한 이론적 배경에 대하여

알아보고, 제 3 장에서는 제안된 연구 방법들에 대하여 설명하고, 제 4 장에서는 제안된 방법을 실제 자료에 적용한 결과를 제시한다. 끝으로 제 5 장에서는 결론과 시사점, 그리고 추후 과제들에 대하여 얘기한다.

제 2 장 이론적 배경

제 1 절 금융시장 군집화

금융시장인 주식시장에서 제공되는 데이터들은 해당 날짜와 시간이 바뀜에 따라 지속적으로 변화하는 시계열 데이터 형태를 띠고 있다. 시계열 데이터란 균등한 시간 간격에서 연속적으로 측정된 값들의 시퀀스를 의미한다. 따라서 주식시장에서 군집화를 실시할 때는 시계열 군집화를 가장 먼저 고려해야한다. 한편 이런 주식시장에서 제공되는 데이터들의 주체는 기업이다. 그러므로 군집화를 진행할 때 시계열 데이터뿐만 아니라 기업들이 가지고 있는 정보에 관한 것도 고려해야한다. 이러한 시계열 군집화와 기업정보를 활용한 군집화에 대해서는 아래와 같은 선행연구들이 진행돼왔다.

Lai et al 2010은 시계열 데이터 분석을 위하여 2단계 군집화 방법을 제안하였다. 첫 단계는 시계열 데이터에 SAX 변환을 취해주고 시계열 간의 SAX 기반의 유사도를 계산한다. 이때 CAST 군집화 알고리즘을 사용하여 군집들을 생성한다. 그 후 두 번째 단계에서 각 군집에 속한 시계열 데이터들의 부분 시계열 정보를 사용하여 군집을 분할하고 부분 군집들을 생성한다.

Aghabozorgi et al 2014b는 시계열 데이터의 모양을 기반으로 새로운 혼합 군집화 알고리즘을 제안하였다. 첫 단계는 시계열 데이터를 시간상의 유사도를 사용하여 부분군집들로 구성하였다. 그 후 두 번째 단계에서는 부분 군집들에 대하여 형태의 유사도와 K-Medoid 알고리즘을 사용하여 병합을 실시한다.

Aghabozorgi & Teh 2014a는 주식 시장의 동조화를 평가하기 위하여 3단계 시계열 데이터 군집화 방법을 제안하였다. 첫 단계는 시계열 데이터에 SAX 변환을 취해준다. 변환된 시계열 데이터에 K-Modes 알고리즘을 사용하여 사전 군집들을 생성한다. 다음 단계에서는 생성된 사전 군집들에 PCS 알고리즘을 사용해 정제하고, 부분 군집들을 생성해준다. 끝으로 마지막 단계에서는 형태의 유사도를 사용하여 유사도가 높은 부분 군집들을 병합하여 최종 군집을 생성한다.

안준규 & 이주홍 2017은 동조화 관계를 갖는 시계열 군집을 찾기 위하여 2단계로 구성된 CTC 군집화 알고리즘 제안하였다. 첫 단계에서는 저차원으로 축소된 시계열 데이터에 DBSCAN 알고리즘을 사용하여 사전 군집들을 생성한다. 두 번째 단계에서는 생성된 각 사전 군집에 가중거리함수를 적용하여 계층적 군집화 알고리즘을 통해 군집을 정제한다.

Nanda et al 2010은 봄베이 증권거래소에서 기업에 대한 정보를 활용하여 주식 평가지표와 주식의 수익률 등 파생변수를 만들어 기업들에 K-Means 알고리즘을 사용하여 군집을 생성한다.

Chen et al 2013은 프랑스 기업들의 재무제표에서 복잡한 시간적 행동의 간결한 시각화를 목표로 SOM 알고리즘을 사용하여 2단계 군집화 과정을 통해 몇 년 동안 회사의 재무상황을 분석하고 시각화하여 기업들의 파산 궤적 군집을 생성한다.

Momeni et al 2015는 테헤란 증권거래소에서 자동차 부품, 금속, 시멘트 등 3가지 산업에 대한 재무제표 데이터를 사용하여, 이익 기준을 선택하고 AHP를 사용하여 우선순위를 정한 후 K-Means 알고리즘을 사용하여 군집을 생성한다.

본 논문은 위와 같은 선행연구들을 참고 및 융합하여 시계열 데이터

들 간의 유사패턴을 좀 더 명확하게 군집화 할 수 있는 3단계 군집화 알고리즘을 제안하였다. 추가적으로 한발 더 나아가 군집화에 있어 주가 시계열 데이터뿐만이 아니라 기업정보 또한 활용하여 유사성을 측정하는 혼합 알고리즘을 제안했다는 점에서 기여하는 바가 있다.

제 2 절 전자공시 자료 분석

전자공시제도란 전자공시시스템(Dart ; Data Analysis, Retrieval and Transfer System)을 통해 상장법인 등이 공시서류를 인터넷으로 제출하고, 투자자 등 이용자는 제출 즉시 인터넷을 통해 해당 공시서류를 조회할 수 있도록 하는 종합적 기업공시 제도로, 1998년 4월 '전자공시제도추진종합계획'의 수립을 시작으로 개발되어 2001년 1월부터 금융감독원에 제출해야 할 공시서류를 이를 통해 제출하도록 의무화하였으므로 벌써 20년 가까이 시행돼 오고 있는 제도이다(김정은 & 남기석 2017, 이래학 2017).

공시자료가 금융시장에 미치는 영향과 관련된 연구들은 국내에서는 주로 기존의 지면기반 공시시스템에 비해 전자공시시스템을 도입함으로써 시장에 어떤 영향을 미쳤는지에 대한 연구가 주를 이루었다(라채원 & 박경진 2010, 이장건 & 정용기 2008, 정창원 2007). 이 연구결과들은 전자공시시스템이 도입되면서 개별기업의 정보가 시장에 제대로 반영이 되면서 정보비대칭 문제를 해결해주고 있다고 분석하였다. 그러나 국내에 전자공시시스템이 도입되면서 사용자들에게 기업정보에 접근하기 쉬운 환경을 제공하고 있지만 대량의 공시자료를 수집해서 텍스트 분석을 수행하기에는 보고서의 형태가 기업별로 상이하고 정형화되어 있지 않아 기계가 판독하기 불가능한 것으로 나타났다(김형준 et al 2015).

해외의 선행연구로는 바레인 기업들을 대상으로 한 연구에서 공시자료의 양과 질, 범위 등의 여러 지표들과 주가 수익률의 변동성 사이의 관계를 살펴보았을 때 유의미한 관계가 있음을 보인 결과가 있다(Mousa & Elamir 2018). 또한 미국의 S&P100 기업을 대상으로 미래 예측 공

시자료에 대한 분석 결과 자발적인 미래 예측 공시자료는 기업의 평판과 주식의 변동성에 유의한 영향을 미친다는 사실을 발견하였다(Bravo 2016). 또한 일본의 다양한 분야 기업들을 대상으로 한 연구에서도 공시자료와 언론자료가 변동성에 유의미한 영향을 미치며 공시자료는 변동성을 증가시키는 반면 언론보도는 변동성을 감소시킨다는 결과를 도출하였다(Aman & Moriyasu 2017).

위와 같이 기존의 선행연구들은 주로 전체적인 공시자료의 질이나 양과 관련하여 시장에 어떤 영향을 미치는지 거시적인 관점에서 분석한 연구들이 대다수였고, 개별 공시자료들의 유형에 따라 주가에 어떤 영향을 미치는지에 대한 연구는 거의 이루어지지 않았다. 본 논문은 개별 공시자료들의 시장에 대한 영향력을 분석하였다는 점과 이를 국내 주식시장에 대하여 분석하였다는 점에서 기여하는 바가 있다.

제 3 절 포트폴리오 최적화

포트폴리오를 최적화에 대하여 설명하기에 앞서, 포트폴리오를 구성하는 기초자산으로 주식을 선정한 세 가지 이유를 간략히 설명한다. 첫 번째는 접근성으로 정보통신기술 발달로 인해 어플리케이션만 설치하면 모바일 기기를 통해 몇 번의 터치만으로 주식을 매매하는 것이 가능해졌기 때문이다. 두 번째는 보편성으로 금융 시장에서 포트폴리오를 구성할 때, 비중이 낮거나, 배제되는 경우는 거의 존재하지 않아서이다. 세 번째는 예금/채권, 부동산 같은 다른 자산들에 비해 비교적 적은 금액으로도 높은 이익을 바라 볼 수 있는 수익성이 있다. 또한 포트폴리오를 구성하는 기초자산은 다양성을 갖는 것이 바람직한데, 주식의 경우 개별 주식마다 다양한 주가 패턴들이 존재할 것으로 예상되므로 분석 결과로 얻을 여러 형태의 군집들을 활용하여 효과적으로 포트폴리오를 구성할 수 있을 것으로 판단된다.

포트폴리오 최적화는 자산을 분산투자하여 포트폴리오를 구성하게 되면 분산투자 이전보다 위험을 줄일 수 있다는 해리 마코위츠(Harry Markowitz)의 포트폴리오 이론을 기반으로 한다. 여기서 합리적 투자자는 위험회피 성향을 가진다고 가정하며, 투자 의사결정에 고려되는 수익과 위험은 각각 평균과 분산으로 추정된다. 일반적으로 포트폴리오의 기대수익과 분산은 다음과 같은 수식 (1), (2)를 통해 계산된다.

$$\mu_p = \sum_i w_i r_i \quad - (1)$$

$$\sigma_p^2 = \sum_{i,j} w_i w_j Cov(r_i, r_j) \quad - (2)$$

여기서 식(1)의 μ_p는 포트폴리오의 수익률, w_i는 개별기초자산i의 비중을, r_i는 개별자산의 수익률을 의미하고, 식(2)의 σ_p^2은 포트폴리오의

분산, $Cov(r_i r_j)$는 공분산 행렬을 의미한다. 이때 자산 간 상관관계가 1이 아닐 경우에는 포트폴리오를 구성함으로써 위험을 감소시킬 수가 있다. 즉, 위 수식을 통해 자산 간 상관관계가 1이 아닌 경우 기초자산을 한 기업이 아니라 여러 기업에 투자하게 되면 기초자산이 갖고 있는 위험을 감소시키는 분산 효과를 얻을 수 있다. 따라서 포트폴리오 최적화는 위와 같은 수식을 통해 적은 위험성 얻으면서 수익률을 최대화하는 것을 목적으로 한다. 마코위츠가 제안한 가장 널리 사용되는 포트폴리오 최적화 모형은 평균-분산 모형으로, 그 모형은 다음과 같다(Markowitz, 1952).

$$\max_w R^T w - q \times w^T \sum w \quad - (3)$$

여기서 w는 포트폴리오를 구성하는 개별 자산의 비중으로 그 합은 1이고, R은 개별 자산의 수익률, $q \geq 0$은 투자위험감수도, $\sum$은 포트폴리오를 구성하는 자산의 수익률에 대한 공분산 행렬을 의미하고, $R^T w$와 $w^T \sum w$은 각각 포트폴리오의 수익률 μ_p와 포트폴리오의 분산 σ_p^2 의미한다.

현대에는 위와 같은 Markowitz 평균-분산 모형을 기본으로 다양한 최적화 모형이 존재한다. 따라서 최근에는 다양한 포트폴리오 최적화 모형을 활용하여 금융상품 및 서비스를 소비하는 금융소비자들의 삶의 방식이나 성격유형 또는 그들의 원하는 방식 등에 따라 알맞게 제공하는 것을 목적으로 하고 있다. 따라서 본 논문에서도 대표적인 두 가지 포트폴리오 모형을 활용하여 최근의 추세를 조금이나마 따라가 보고자하였다. 또한 아래의 관련 연구를 통해 마코위츠 포트폴리오 모형의 효과를 확인하였다.

김성문 외 2009는 국내 주식시장에서 마코위츠의 포트폴리오 선정 모형의 투자 성과를 알아보기 위하여 투자 시점으로부터 가장 가까운

일정 기간 동안 수집한 데이터를 가지고, 개별 주식의 연간 평균수익률, 분산 및 주식간의 공분산을 구하여 지속적으로 업데이트를 해주는 포트폴리오를 구성했다. 그 후 구성된 포트폴리오와 대중적으로 인기가 높고, 규모와 성과 면에서 우수한 펀드를 선정하여 비교했다. 그 결과 데이터를 기반으로 구성된 포트폴리오가 더 우수한 성과를 보임을 확인했다.

제 3 장 연구방법

제 1 절 금융시장 군집화 알고리즘

1 자료 수집 및 전처리

본 연구에서는 국내 금융 시장 대상으로 하여 분석을 진행하였다. 최종 포트폴리오를 구성하는 기초자산은 주식으로만 한정하여 국내 유가주식시장에 상장 돼있는 기업들의 자료를 수집하였다. 유가주식시장에 한하여 기업들의 자료를 수집한 이유는 다음과 같다. 한국거래소(KRX)의 2020년 2월 기준 상장돼있는 기업의 수는 총 2,354개로 포트폴리오 최적화를 위해 해당 기업들 전체에 대하여 자료를 수집해서 연구와 분석을 진행하는 것이 가장 이상적이겠지만, 연구자에게 주어진 시간과 비용에는 한계가 존재하므로 우리나라 주식시장을 대표하는 유가주식시장에서 시가총액이 크고 거래량이 많아 유가주식시장을 대표할 수 있다고 간주되는 200개 기업을 선정하여 연구와 분석을 진행하였다. 실제로 이 200개의 기업들은 유가주식시장에서 시가총액기준 약 80%이상의 비중을 차지한다. 또한 우리나라 유가시장을 대표하는 지수인 KOSPI 200은 이 200개 기업들의 주가로부터 산출하는 지수이다. 여기서 KOSPI(Korea Composite Price Index)란 종합주가지수를 뜻하는 단어로, 현재는 유가주식시장의 주가지수를 KOSPI라고 부른다. 따라서 이 KOSPI 200에 속하는 202개 기업들의 2013/01/02월부터 2017/12/28까지 5년 동안의 데이터를 한국거래소에서 수집하여 연구를 진행하였다. 이 중 5년간의 데이터를 온전히 보유하고 있지 못한 21개의 기업들은 앞으로 연구를 진

행하는데 있어 부정적인 영향을 미칠 것으로 판단하여 해당 단계에서 제외하였다. 따라서 총 181개의 기업을 가지고 데이터에 대한 탐색을 실시하였다.

금융시장 군집화는 크게 두 개의 과정으로 이루어지는데 첫 번째 과정에서는 주가 시계열자료를 이용하여 군집화를 수행하고 두 번째 과정에서는 재무제표 자료를 함께 활용하여 군집화가 이루어지게 된다. 수집된 데이터는 다음과 같이 '년/월/일, 종가, 대비, 거래량(주), 거래대금(원), 시가, 고가, 저가, 시가총액(백만), 상장주식수(주), 종목명, 업종' 총 12개의 변수로 구성돼있고, 아래의 [그림 3.1]을 통해 수집한 데이터의 전반적인 형태를 제시하였다.

	년/월/일	종가	대비	거래량(주)	거래대금(원)	시가	고가	저가	시가총액(백만)	상장주식수(주)	종목명	업종
0	2013-01-02	1,576,000	54,000	229,274	355,969,894,000	1,533,000	1,576,000	1,527,000	232,143,755	147,299,337	A005930_삼성전자	통신 및 방송장비 제조업
1	2013-01-03	1,543,000	-33,000	284,927	443,088,034,990	1,582,000	1,584,000	1,543,000	227,282,877	147,299,337	A005930_삼성전자	통신 및 방송장비 제조업
2	2013-01-04	1,525,000	-18,000	260,120	395,253,861,800	1,540,000	1,542,000	1,510,000	224,631,489	147,299,337	A005930_삼성전자	통신 및 방송장비 제조업
3	2013-01-07	1,520,000	-5,000	252,436	381,987,869,040	1,515,000	1,528,000	1,500,000	223,894,992	147,299,337	A005930_삼성전자	통신 및 방송장비 제조업
4	2013-01-08	1,500,000	-20,000	276,757	416,202,501,363	1,513,000	1,517,000	1,498,000	220,949,006	147,299,337	A005930_삼성전자	통신 및 방송장비 제조업

[그림 3.1] 2013~2017년 KOSPI200 기업의 수집된 기업별 데이터 형태

이 중 금융시장 군집화 알고리즘의 첫 번째 과정인 3단계 군집화를 통한 시계열 군집화 과정에서는 기업의 주가 정보 즉, 종가를 활용하여 비슷한 주가패턴을 띄는 기업들끼리 군집화 하는 것을 목적으로 함으로 사용할 변수는 '종가'와 '년/월/일' 두 가지 변수이다. 해당 변수들을 사용하여 데이터의 형태를 수정한 모습은 아래[그림 3.2]와 같다. 이때 '종가'의 단위는 (원)을 의미한다.

년/월/일	삼양홀딩스	하이트진로	유한양행	CJ대한통운	두산	대림산업	한국타이어월드와이드	기아차	동아쏘시오홀딩스	SK하이닉스	영원무역	GKL	락앤락	코오롱인더	한미약품
2013-01-02	71400	31400	172500	103500	131500	88500	19850	56300	116000	26600	34750	29250	23700	64100	121000
2013-01-03	71200	31750	175500	102000	132500	91700	18900	54600	115000	26650	34600	28550	22950	65100	121500
2013-01-04	71100	31700	181500	107500	132500	91000	18650	53600	118500	26350	34850	29700	23700	65400	127000
2013-01-07	70500	31500	184000	107500	132500	89700	19050	54000	123500	25900	36000	30300	24250	65200	136000
2013-01-08	69300	31450	187500	113000	132000	87600	18800	54500	121000	26250	36200	29900	24100	64500	134500

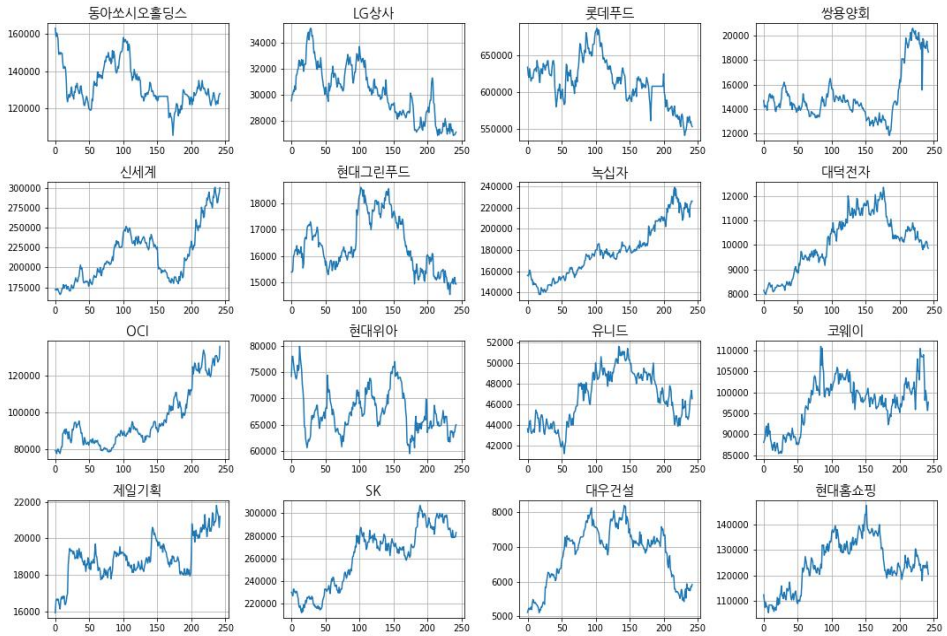
[그림 3.2] 3단계 군집화를 위해 수정된 2013~2017년 KOSPI200 기업의 데이터 형태 (단위: 원)

다음으로 아래 [표 3.1]은 2013/01/02월부터 2017/12/28까지의 '종가'에 대한 일부 기업에 대한 기초 통계량을 나타내는 표로 기업마다 종가가 다양한 양상을 띠고 있는 것을 알 수 있다.

[표 3.1] 2013~2017년 KOSPI200 일부 기업에 대한 종가의 기초 통계량

기업명 / 통계량	삼양홀딩스	하이트진로	유한양행	CJ대한통운	두산
count	1229	1229	1229	1229	1229
mean	111395	24676	221410	159966	116488
std	38651	3380	44386	40727	16537
min	65400	19900	164000	81000	70200
25%	84300	22350	184000	115000	105000
50%	102000	23800	207000	171000	117500
75%	131000	25950	253500	192500	129500
max	300000	35050	335500	228500	152000

위와 같이 5년간의 주가 시계열 데이터를 탐색한 결과, 시간의 흐름에 따라 지속적으로 주가의 패턴이 바뀌는 것을 확인할 수 있었다. 이에 따라 5년 전체 기간을 이용하여 비슷한 패턴을 갖는 기업들을 군집화하는 것은 급변하는 주식시장을 설명하기에 적절하지 않다고 판단하여 최근 1년간의 주가 시계열 데이터만을 이용하기로 결정하였다, 따라서 수집한 데이터 중 가장 최근인 2017년도 데이터만을 이용하여 금융시장 군집화의 첫 과정인 주가 시계열 군집화를 진행하였다. 이때 주가 시계열 군집화 과정의 목적은 기본적으로 비슷한 주가패턴을 띄는 기업들끼리 군집화 시키는 것을 목적으로 하므로 아래 [그림 3.3]과 같이 17년도에 대한 종가(시계열)그래프를 기업별로 그려 어떠한 패턴들을 가지고 있는지에 대하여 대략적으로 확인하였고, [그림 3.3]을 통하여 일부 기업들에 대해 나타냈다.



[그림 3.3] KOSPI200의 일부 기업에 대한 17년도 시계열 종가 그래프

다음으로는 금융시장 군집화 알고리즘의 두 번째 과정인 기업정보를 활용한 군집화 과정에서 사용되는 변수들에 대한 탐색을 실시하였다. 이 과정에서는 후에 주가 시계열 군집화 과정을 통해 결과 값으로 도출되는 Merge7(Label) 변수 외에도 기업의 주가를 예측하는데 유의미한 영향을 끼칠 것으로 예상되는 다른 변수들을 추가하여 탐색을 진행하였다. 이때 추가적으로 사용되는 변수들은 Nanda et al 2010을 참고하여 선택하였으며 총 14개로 아래 [표 3.2]와 같고, 해당 변수들은 한국거래소에서 수집한 기업들의 2017년도 데이터뿐만 아니라 해당 기업들의 2017년도 재무제표를 활용해서 만들어졌다.

[표 3.2] 기업 정보를 활용한 추가 변수

변수명	설명
수익률 (기간에 따라 6가지 변수)	1day, 1week, 1month, 3month, 6month, 1year 로그수익률에 대한 수식은 아래와 같음 $$x = \log(\frac{S_c}{S_k})$$ - S_c는 기준날짜의 종가, S_k는 알고 싶은 기간날짜의 종가, 기준날짜는 가장 최근날짜 중 수요일인 2017-12-27로 적용 - 수요일로 기준을 잡은 이유는 1주일 동안 주말의 영향을 가장 적게 받는 날이기 때문
변동성 (기간에 따라 5가지 변수)	1week, 1month, 3month, 6month, 1year 1day 로그수익률을 통한 기준 날짜로부터 알고 싶은 기간 동안의 표준편차 값으로 수식은 아래와 같음 $$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$ -x_i은 알고 싶은 기간 동안의 1day 로그수익률, $\overline{x}$는 알고 싶은 기간 동안의 1day 로그수익률의 평균, n = 알고 싶은 기간 동안의 1day 로그수익률 개수
시가 총액	상장 종목(기업)별로 그 날 종가에 상장주식수를 곱한 후 합계하여 산출
PER (주가수익비율)	주가 / 주당 순이익 - PER은 주가가 주당 순이익의 배율이 얼마인가를 나타내는 지표로 PER이 낮을 경우 해당 회사가 거둔 이익에 비해 주가가 낮고 그에 따라 기업의 가치에 비해 저평가 돼있다는 의미로 볼 수 있음. 반대로 PER이 높으면 거둔 이익에 비해 주가가 높게 평가 됐다고 판단할 수 있음
PBR (주가순자산비율)	주가 / 주당 순자산가치 - PBR은 주가를 주당 순자산가치로 나눈 비율로, 기업의 순자산에 대해 1주당 몇 배 거래되고 있는지 측정하는 지표로 PBR이 1미만인 경우 기업가치 보다 주가가 상대적으로 낮음을 의미하고, 1이상인 경우 기업가치 보다 주가가 상대적으로 높음을 의미

다음으로 추가 변수들에 의해 형성된 데이터의 기본적 형태는 아래 [그림 3.4]와 같다. 그리고 [표 3.3]을 통해 추후 분석에 사용되는 8개의 추가 변수들에 대한 기초통계량을 제시하였다. 이를 통해 특이 값들이 존재한다는 것을 알 수 있다.

company	1day_수익률	1week_수익률	1month_수익률	3month_수익률	6month_수익률	1year_수익률	1week_변동성	1month_변동성	3month_변동성	6month_변동성	1year_변동성	시가총액(단위:백만)	PER	PBR
삼양홀딩스	-0.00823	0.050858	0.18067	0.337961	-0.020451	-0.004124	0.014754	0.022626	0.018095	0.018545	0.018	1036277	25.132219	0.75565
하이트진로	-0.004115	-0.018387	0.029291	-0.061958	0.03782	0.12971	0.011322	0.015916	0.015849	0.017619	0.014926	1700740	91.313361	1.336797
CJ대한통운	0	-0.021202	-0.075637	-0.133531	-0.262364	-0.225997	0.018207	0.013364	0.013177	0.013109	0.015744	3193728	160.295052	1.375943
두산	-0.075035	-0.066971	-0.094917	-0.195937	0	0.057923	0.041572	0.030241	0.023176	0.022206	0.020803	2217369	10.811148	1.008006
대림산업	-0.012361	-0.022141	-0.030621	0.018833	-0.05916	-0.028205	0.006154	0.01265	0.017665	0.015202	0.015562	2797920	27.162741	0.602003

[그림 3.4] 추가 변수들에 의해 생성된 데이터의 기본적 형태

[표 3.3] 분석에 사용되는 8개의 추가 변수들에 대한 기초 통계량

변수명 / 통계량	1week_수익률	3month_수익률	1year_수익률	3month_변동성	1year_변동성	시가총액(단위:백만)	PER	PBR
count	143	143	143	143	143	143	143	143
mean	-0.0113	0.0329	0.0896	0.0219	0.0215	8034335	44.8384	1.8958
std	0.0438	0.1764	0.4133	0.0114	0.02	27469940	217.4241	2.225
min	-0.3463	-1.1596	-3.1892	0.0071	0.007	327172	-199.2638	0.3529
25%	-0.0247	-0.0432	-0.0861	0.0157	0.0155	1063151	10.7231	0.7759
50%	-0.0108	0.0372	0.0508	0.0194	0.0184	2369849	16.6993	1.132
75%	0.0049	0.0973	0.3113	0.0249	0.022	6154432	31.0844	1.883
max	0.0903	0.4707	1.143	0.1105	0.2048	318615100	2434.3553	13.42

다음으로는 금융시장 군집화 알고리즘의 첫 번째 과정인 주가 시계열 군집화에 들어가기에 앞서 기존의 17년도 시계열 종가 데이터를 군집화에 적합하게 가공하는 데이터 전처리 작업이 필요한데, 본 논문에서는 스케일링(Scaling)과 차원축소(Dimension Reduction) 두 가지 방식을 사용하여 해당 데이터에 대한 전처리 작업을 수행하였다.

우선 각 기업별로 종가에 대한 격차가 있으므로 해당 값들을 일정한 범위로 통합시켜주는 스케일링 작업을 실시하였다. 스케일링(Scaling)이란 자료 집합에 적용할 수 있는 데이터 전처리 과정의 하나로 모든 자료에 선형 변환을 적용하여 자료의 오버플로우나 언더플로우를 방지하고, 다차원의 값들에 대한 비교 및 분석을 용이하게 만들어주며, 독립변수의 공분산 행렬의 조건수를 감소시켜 최적화 과정에서의 안정성 및 수렴 속도를 향상시켜주는 역할을 수행한다. 특히 데이터의 스케일이 다를 경우 거리를 기반으로 분류하는 모델에 부정적인 영향을 미칠 가능성이 큼으로, 스케일링 통해 범위를 일정하게 맞추어주는 작업을 실시해주어야 한다. 이때 본 논문에서 사용된 스케일링 방법으로는 Z-변환(Z-Transformation)을 이용하는 표준화(Standardization)방법을 사용한다.

$$Y_t = \frac{X_t - \mu_X}{\sigma_X} \quad - (4)$$

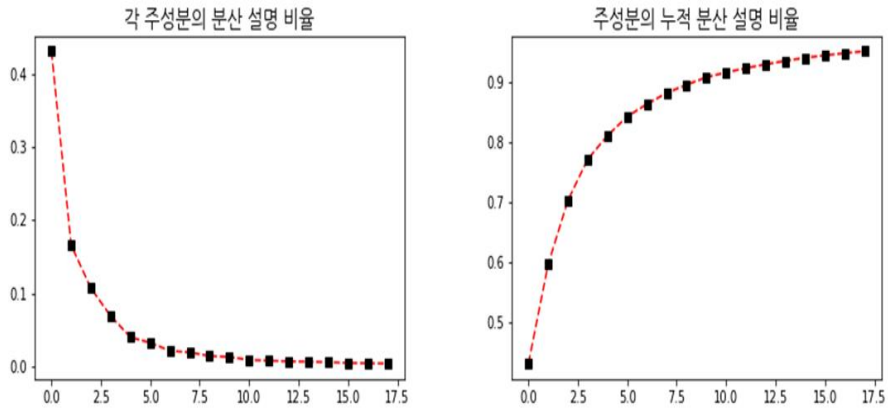
Z-변환의 수식은 (4)와 같고, 여기서 Y_t은 각 시점에서 표준화가 완료된 시계열 데이터를 의미하고, X_t은 표준화 이전의 각 시간에서 시계열 데이터를 의미한다. 또한 μ_X는 시계열 데이터의 평균, σ_X는 시계열 데이터의 표준편차를 의미한다. 즉, Z-변환이란 기업별로 각각의 시점에서의 시계열 데이터의 값과 해당 기업의 시계열 데이터의 평균의 차를 해당 기업의 시계열 데이터의 표준편차로 나누어주는 기법으로 [그림 3.5]를 통해 결과를 제시했다.

년/월/일	삼양홀딩스	하이트진로	유한양행	CJ대한통운	두산	대림산업	한국타이어 월드와이드	기아차	동아쏘시오 홀딩스	SK하이닉스	영원무역	GKL	락앤락	코오롱인더	한미약품
2017-01-02	0.529397	-1.061924	-1.424693	0.949661	-0.6979	0.961884	0.015515	1.394758	2.820028	-1.429602	-0.667272	-0.946202	-0.497396	1.257392	-1.155118
2017-01-03	0.370027	-0.866821	-1.483989	0.706192	-0.664046	1.069362	0.104226	1.887986	2.40505	-1.317495	-1.342703	-0.972609	-0.563186	1.295268	-1.128062
2017-01-04	0.2505	-0.964373	-1.424693	0.625036	-0.867172	0.61258	0.326003	2.105006	2.589485	-1.375481	-1.473432	-0.880184	-0.59608	1.484649	-1.025252
2017-01-05	0.330185	-0.866821	-1.068916	0.625036	-0.765609	0.746928	0.503424	2.02609	2.497267	-1.340689	-1.756677	-0.866981	-0.508361	1.333144	-1.01443
2017-01-06	0.210657	-0.842433	-1.246804	0.787349	-0.6979	0.720058	0.459069	1.887986	2.266724	-1.259508	-1.822042	-0.880184	-0.541256	1.408897	-1.073952

[그림 3.5] 각 기업의 시계열 종가데이터에 표준화 방법을 적용한 결과

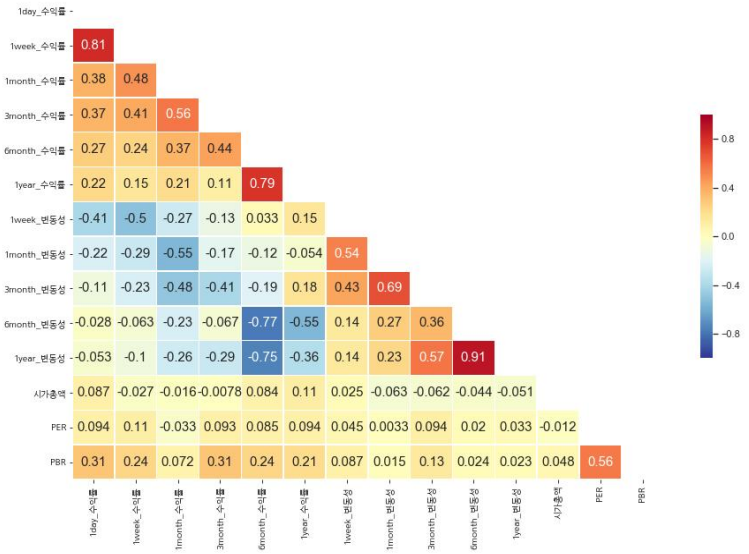
다음으로 본 논문에서 취해줄 전처리 작업은 차원축소 작업이다. 차원축소란 차원의 저주를 해소하기 위한 방법으로 변수의 차원의 수를 줄이는 과정이다. 이를 통해 차원의 저주에서 탈피할 수 있고, 시각화의 용이성이라는 이점을 얻을 수 있을 뿐만 아니라 불필요한 요소나 노이즈의 제거 또한 가능하게 해준다. 기업들에 대한 17년도 시계열 종가데이터의 경우 2017/01/02부터 2017/12/29까지 총 243개의 차원(변수)을 가지고 있는 고차원 데이터이므로 차원 축소를 적용하기에 적합하다. 본 논문에서 사용된 차원축소방법은 주성분 분석(Principal Component Analysis; PCA)방법으로 데이터의 분산을 최대한 보존하면서 서로 직교하는 새 기저(축)를 찾아, 고차원의 공간의 표본들을 선형 연관성이 없는 저차원의 공간으로 변환하는 기법이다. 여러 차원축소방법들 중에서도 주성분 분석은 스크리 도표(Scree Plot)를 사용하여 시각화를 통해 데이터의 적절한 차원의 수를 손쉽게 결정할 수 있는 장점이 있기에 해당 기법을 차원 축소 과정에서 사용하였다. 여기서 스크리 도표는 모든 고유 값들의 합에서 소수의 고유 값의 합의 비율을 나타내는 도표이다. 따라서 해당 데이터에 대하여 차원 축소를 실시한 결과는 [그림 3.6]과

같다. 이 [그림 3.6]을 살펴보면 3개의 주성분을 통해 전체의 70%이상이 설명이 가능한 것을 알 수 있다. 추가적으로 [그림3.6]의 기울기를 살펴보면 기울기가 급격하게 변하는 점이 4번째 점부터이므로 주성분은 3개로 결정하는 것이 타당하다는 것을 알 수 있다. 이때 각각의 주성분들의 설명 분산 비율은 0.4321, 0.1655, 0.1066이다.

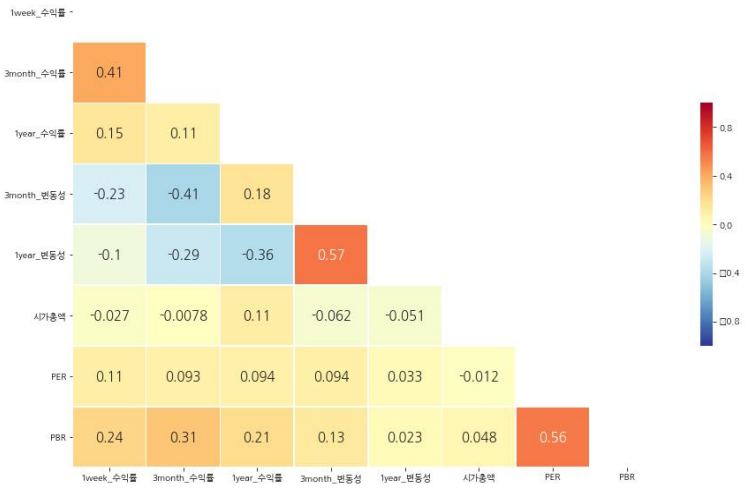


[그림 3.6] 주성분의 분산 설명 비율에 대한 스크리 도표

앞선 첫 번째 군집화 과정인 주가 군집화와 마찬가지로 두 번째 과정인 기업정보를 활용한 군집화 과정에서도 데이터에 대한 전반적인 전처리 작업이 필요하다. 따라서 [표 3.2]에 명시된 추가변수들에 대한 전처리 작업을 실시한다. 첫 번째 군집화 과정에서는 변수가 '종가' 하나였던 것과 달리 두 번째 군집화 과정에서는 3단계 군집화 과정을 통해 결과 값으로 도출된 Merge7(Label) 변수 외에 추가변수들이 총 14개이므로 해당 변수들 간의 상관관계를 파악하는 것을 우선으로 한다. 변수들의 상관관계를 파악하는 이유는 군집화를 하는데 있어 상관관계에 따라 그 영향력이 달라 질 수 있기 때문이다. 예를 들어 상관관계가 높은 변수들을 그대로 사용하여 군집화를 하면 군집화 과정에서 거리를 계산할 때 더 많은 가중치를 부여하는 것이 된다. 즉, 군집을 형성하는데 있어 상관관계가 높은 변수의 영향력이 커지는 것이므로 이는 군집화 과정에서 부정적인 영향을 끼칠 가능성이 높다. 따라서 변수들 간의 상관관계를 파악하며 그 상관관계가 큰 변수들을 제거하는 작업을 실시하였다. 그 결과 [그림 3.7]과 같이 '수익률'과 '변동성' 변수와 같이 주기의 길이가 가깝고, 맞물리는 변수들 간의 상관관계가 높게 나옴을 알 수 있다. 따라서 이러한 변수들은 주기의 간격을 두어 제거하였다. 그 결과 [그림 3.8]과 같이 8개의 변수가 선택됐다.



[그림 3.7] 추가 변수들 간의 상관관계에 대한 도표



[그림 3.8] 선택된 8개 변수들 간의 상관관계에 대한 도표

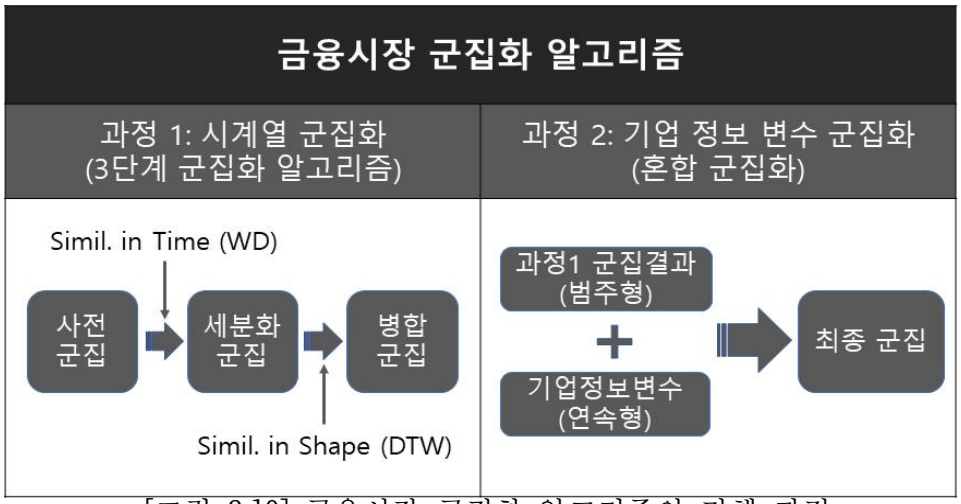
다음으로는 선택된 8개의 변수들 중 ‘수익률’과 ‘변동성’ 변수에 대하여 변환작업을 진행하였다. 그 이유는 군집화에 있어 거리를 계산할 때 변수 값이 너무 작고 몰려있는 경우 이러한 변수의 영향력이 작아질 수 있기 때문이다. 또한 ‘수익률’과 ‘변동성’ 변수는 대체적으로 0에 가까운 값을 가진다. 따라서 이 상태에서 거리를 계산할 경우 데이터를 다수 포함하고 있는 커다란 군집 하나가 생성되고 나머지 군집들 데이터를 적게 포함하고 있을 가능성이 높아진다. 따라서 이러한 변수들에 대하여 Log변환을 취하여 그 값을 퍼뜨려 변수의 영향력을 줄인다. 그 후 앞선 주가 시계열 군집화 과정과 마찬가지로 스케일링(Scaling) 작업을 통해 표준화방법 Z-변환(Z-Transformation)을 실시한 결과는 [그림 3.9]와 같다.

	1week_수익률	3month_수익률	1year_수익률	3month_변동성	1year_변동성	시가총액(단위:백만)	PER	PBR
0	1.423716	1.735235	-0.227651	-0.284035	-0.164821	-0.255649	-0.090953	-0.514215
1	-0.162235	-0.53961	0.097284	-0.631794	-0.658692	-0.231375	0.214504	-0.25211
2	-0.226721	-0.94674	-0.766338	-1.116199	-0.517965	-0.176834	0.532887	-0.234455
3	-1.274976	-1.301719	-0.077008	0.365332	0.216893	-0.212502	-0.157052	-0.400399
4	-0.248226	-0.08005	-0.286119	-0.347205	-0.548703	-0.191293	-0.081581	-0.583512

[그림 3.9] 각 기업의 추가변수 데이터에 표준화 방법을 적용한 결과

2 제안 알고리즘

이 장에서는 본 논문에서 제안한 금융시장 군집화 알고리즘에 대해 설명하도록 한다. 제안 알고리즘 과정을 도식화하면 [그림 3.10]과 같다.



[그림 3.10] 금융시장 군집화 알고리즘의 진행 과정

첫 번째 과정은 주가 시계열 자료를 군집화 하는 단계이다. 고차원의 시계열 데이터에 단순하게 군집화 알고리즘을 적용하게 될 경우 시차가 다르게 동조화 패턴이 나타나는 시계열 데이터들에 대한 군집화가 제대로 이루어지지 못한다는 단점을 보완하기 위하여 3단계 군집화 알고리즘을 적용하였다. 이를 통해 시간에 대한 유사성뿐만 아니라 형태에 대한 유사성을 이용하여 군집을 생성한다.

두 번째 과정은 첫 과정에서 나온 군집화 결과와 기업정보와 관련된 변수들을 혼합하여 군집화를 하는 단계이다. 시계열의 패턴뿐만 아니라 기업 정보를 갖고 있는 다른 변수들을 추가하여 기업 간의 유사성을 평가한 후 최종 군집을 생성한다. 이 때 범주형 변수인 군집 결과와 연속

형 변수인 기업 정보를 혼합하여 데이터의 유사성을 측정하는 것이 중요한데 여러 방법을 활용해 혼합 군집화 방법을 제안하였다.

(가) 3단계 군집화를 통한 시계열 군집화

본 논문에서는 주가 시계열 군집화를 위하여 Aghabozorgi & Teh 2014a에서 제안된 3단계 군집화 알고리즘을 변형하여 활용하였다. 이 알고리즘은 시계열 군집화에서 동조화 관계를 갖는 시계열 군집을 찾기 위해 제안되었다. 이 변형 알고리즘은 3단계로 이루어지는데 1단계에서는 고차원의 데이터를 저차원으로 유도하여 3차원에서 시각화를 통한 유사도를 활용하여 사전군집들을 만들고 2단계에서는 시간상의 유사도를 이용하여 사전군집 내에서 세분화된 군집들을 만든 다음 3단계에서 각 세분화된 군집들의 Prototype을 형태의 유사도를 통해 병합하여 최종 군집을 생성한다.

3단계 군집화 과정에서 사용하는 군집화 알고리즘은 각 단계마다 상황에 맞게 다른 알고리즘이 사용되었다. 우선 첫 단계인 사전 군집화단계에서 사용한 방법은 DBSCAN 군집 방법이다. 이 방법은 밀도기반 군집화 방법으로 기준이 되는 반경 내에 최소한의 이웃 데이터가 있으면, 하나의 군집으로 처리하는 방법이다. 사전 군집화단계에서 이 방법을 사용한 이유는 크게 두 가지이다. 우선 군집화를 시행할 때, 해당 시계열 종가데이터의 공간은 데이터마다 분포가 다를 수 있기 때문이다. 따라서 처음부터 어느 정도의 군집의 개수가 적절한지에 대하여 직관적으로 판단하는 것은 굉장히 어려운 일이다. 하지만 이러한 상황에서 위 방법을 사용하면 군집의 수를 사전에 지정해줄 필요가 없기 때문에 매우 유용하다. 다른 이유로는 군집화를 진행하는 동시에 분석에 불필요한 노이즈 데이터 또한 분류해줄 수 있어, 이상점에 의해 군집화 성능이 하락하는 현상 또한 완화해 줄 수 있다는 점이다. 다만 이 방법을 시계열 종가데이터에 원활하게 적용하기 위해서는 다음과 같은 중요 파

라미터들인 min_samples(=군집 당 최소 샘플 수)와 eps(=군집간 거리)를 적절하게 설정해주어야 하는데, 이 파라미터 값들은 KNN Distance plot을 이용하여 min_samples별 eps를 추정한 후 조합하는 방식으로 최적의 조합을 찾았다. 그 후 최적의 조합 값을 가지고 데이터 전처리를 통해 가공해놨던 3개의 주성분을 갖고 있는 3차원 데이터에 대하여 DBSCAN 군집 방법을 적용한다. 그 후 군집화 결과가 적절한지에 대한 판단은 2, 3차원 산점도 도표를 통해 진행한다. 또한 위 알고리즘 적용으로 인해 나오는 노이즈 데이터는 앞으로의 군집화 과정에서 제외한다.

두 번째 단계인 군집 세분화단계를 진행하기에 앞서 해당 사전 군집들의 데이터에 대한 정제를 실시한다. 데이터에 대한 추가 정제를 실시하는 이유는 사전 군집화 단계에서 243개의 차원을 3개의 차원으로 축소시킨 뒤 군집화를 진행했기 때문이다. 따라서 축소된 차원에서의 데이터 사이의 거리 정보가 왜곡되는 현상이 발생할 수 있다. 따라서 정제는 사전 군집들이 갖고 있는 각 데이터에 대하여 원 데이터 공간으로 데이터로 변환시켜주는 작업을 의미한다. 그 후 대략적으로 형태가 비슷한 것들끼리 군집화 됐던 사전 군집들에 대한 군집 세분화단계를 진행한다. 본 논문에서는 이 단계에 사용한 방법은 계층적 군집화 방법 중 하나인 Agglomerative 군집 방법이다. 계층적 군집화란 하나의 데이터 샘플을 하나의 군집으로 보고 가장 유사도가 높은 군집끼리 합쳐주면서 군집의 개수를 줄여가는 Bottom-Up 방법이다. 따라서 위의 사전 군집화 단계의 결과로 나온 군집 하나당 위 방법을 적용해주는 것이다. 이때 일반적인 병합 계층적 군집화 알고리즘은 매 군집화 단계마다 데이터들 사이의 거리를 유클리디안 거리함수와 같은 거리함수를 통해 계산하여 계층적 군집화를 실시하는데 이런 일반적인 거리함수로는 시간의 흐름에 따른 시계열을 생성하는 시스템의 상태변수의 변화를 고려해

주지 못하므로 가중거리함수를 사용하여 거리를 계산하고, 그 수식은 아래 (5), (6)과 같다.

$$\text{가중거리함수}(X,Y) = \sum_{i=1}^{N} \alpha_i \times \text{각 구간별 거리행렬}(X,Y) \quad \text{- (5)}$$

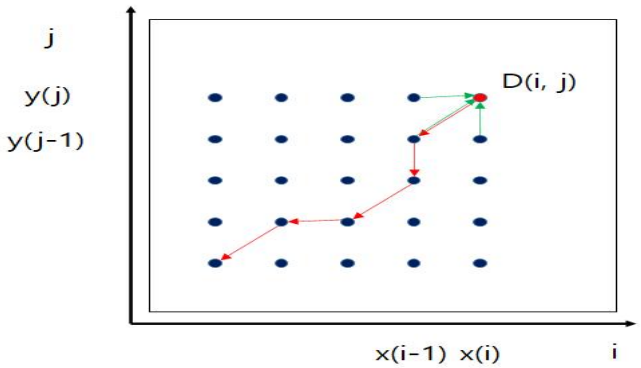
$$\text{각 구간별 거리행렬}(X,Y) = \sqrt{\sum_{i=1}^{\text{각 사전군집의 크기}} (X(t) - Y(t))^2} \quad \text{- (6)}$$

여기서 가중거리함수란 각 시계열의 시간을 구간들로 나누고 각 구간별 가중치(α_i)를 다르게 할당하여 시계열 사이의 거리를 계산하는 방법을 의미한다. 또한 가중거리함수에서 가중치에 대한 할당은 시스템의 상태변수의 변화를 고려하여 최근의 데이터로부터 과거의 데이터까지 내림차순으로 할당시켜주는 함수를 뜻한다. 또한 여기서 $X(t)$는 t시간에서 시계열 X의 값이고, $Y(t)$는 t시간에서 시계열 Y값, N은 구간의 개수이다. 따라서 17년도 시계열 종가데이터들을 가중거리함수를 통해 각 월별로 12구간으로 나누고, 각 구간별로 거리행렬을 생성한 뒤에 현 시점에서 가장 먼 구간거리행렬부터 차례대로 가중치 1~12를 곱한 후 행렬을 더해서 가중거리구간행렬 구하게 된다. 이렇게 구한 가중거리구간행렬을 통해 앞서 언급했다시피 각각의 사전 군집들에 대하여 Agglomerative 군집 방법을 실시한다. 이때 Agglomerative 군집 방법을 통해 합쳐지는 데이터는 최소 5개 이상의 경우만 고려한다.

3단계 군집화 과정의 마지막 단계는 병합 군집화 단계로, 세분화 단계에서 사전 군집별로 세분화 시켰던 군집들의 Prototype을 이용하여 군집들을 병합시키는 단계이다. 이 단계가 필요한 이유는 군집에 대한 세분화는 각 사전 군집별로 해주었으므로 병합단계를 거쳐 세분화된 군집들 간의 형태가 비슷한 시계열끼리 묶어주는 단계가 필요하기 때문이다. 이때 각각의 병합 군집에는 최소 10개 이상의 기업이 들어가 있는

경우에 한해서 선택하는 것을 목표로 한다. 본 논문에서 마지막 단계에 사용한 방법은 K-Means 군집 방법이다. 이 방법은 비계층적 방법으로 임의로 데이터를 군집화하고 군집 과정에서 중앙값의 변화에 따라 각 주어진 데이터를 k개의 군집으로 묶는 알고리즘이다. 이때 각 군집과 거리 차이의 분산을 최소화하는 방식으로 동작하고, 각 군집은 군집에 있는 데이터 객체의 평균값으로 대표된다. 헌데 병합 군집화 단계에서는 일반적인 방식처럼의 전체데이터에 대한 군집화가 아니라 세분화된 군집들에 대하여 군집들 간의 군집화가 이루어져야한다. 따라서 이들 세분화된 군집들에 대한 개별적인 Prototype을 생성하여 해당 군집들 간의 군집화를 진행하도록 한다. 여기서 사용된 Prototype이란 세분화된 군집들 내 시계열 데이터들의 각 시점별에서의 평균 시계열을 구하여 그것들 간의 거리를 동적시간워핑(Dynamic time warping; DTW)방법으로 구한 뒤 행렬화시킨 형태를 말하고, 그 수식은 아래 (7)과 같고, 아래 [그림 3.11]을 통해 수식에 대한 설명을 보충하였다.

$$D(i,j) \equiv \|x(i)-y(j)\| + \min\begin{cases} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{cases} \quad - (7)$$
$$(i=1,\ldots,n; j=1,\ldots,m;)$$



[그림 3.11] DTW 방법의 진행 방식에 대한 도식화

DTW란 시간 흐름에 따른 패턴간의 유사성을 측정하는 알고리즘으로 두 시계열 간의 거리를 최소화하는 방향으로 움직이면서 매칭시켜 각 템플릿과 누적 거리를 계산하여 최소가 되는 클래스로 인식하는 방법이다. 이 DTW를 사용한 이유는 두 시계열 데이터를 매칭시켰을 경우 유클리디안 방법과 같은 일반적인 거리 함수를 사용했을 때와는 달리 부분적으로 왜곡되거나 변형된 파형이 존재하는 시계열 데이터에 대해서도 적절하게 매칭이 가능하기 때문이다. 따라서 본 논문에서는 위 방법을 사용하여 병합 군집화를 위한 거리행렬을 생성하고, 그 거리행렬에 K-Means 군집 방법을 적용하여 군집을 생성하도록 한다.

(나) 기업정보를 활용한 군집화

금융시장 군집화의 두 번째 과정은 기업정보와 주가 자료를 함께 활용하여 군집화하는 과정이다. 따라서 첫 번째 과정인 주가 시계열 군집화에서 나온 군집화의 결과 값을 Merge7이라는 범주형 변수로 사용하여 기업정보변수와 병합시켰다. 이 때 연속형 변수인 기업정보 변수들과 범주형 변수인 Merge7변수를 병합할 수 있는 여러 가지 방법이 존재한다. 본 논문에서는 [표 3.4]를 통해 제시한 4가지 방법을 통해 Merge7을 활용했다.

위와 같이 4가지 방법을 사용한 이유는 시계열 군집화 결과를 사용하지 않고 기업정보 변수만으로 군집화를 했을 때와 시계열 군집화 결과를 추가하여 군집화를 진행하였을 때 군집화 성능이나, 추후에 포트폴리오 최적화 성능이 더 잘 나오는지를 비교 평가하기 위함에 있다. 또한 기업정보라는 변수에 시계열의 패턴으로 군집화한 정보를 추가하여 군집화를 진행하기 위하여 범주형 변수와 연속형 변수가 혼합된 자료의 유사성을 측정해야 하는 방식이 필요하다. 이를 위해 다양한 방법으로 혼합 자료를 만들고, 유사성을 측정하는 방식들을 조사하여 성능을 평가하였다. 이러한 과정을 통해 최종적으로 합리적인 투자를 위한 포트폴리오를 구성할 때 3단계 군집화 결과인 Merge7과 기업 정보를 활용한 추가변수들을 어떻게 활용해야 하는지에 대한 방법론을 제안할 수 있을 것으로 판단된다.

[표 3.4] 추가 변수들에 대한 Merge7(Label)을 병합하는 방법

방법	설 명
1. Merge7 사용 X	추가된 8개의 기업정보 변수들만을 사용하여 군집화를 실시
2. Merge7 →가변수화	Merge7을 원-핫 인코딩(One-hot encoding)방식을 통해 가변수화하여 추가 변수들과 결합 - 가변수란 범주형 변수를 0또는 1값을 가진 하나 이상의 새로운 특성으로 바꾼 것을 의미 - 0과 1로 표현된 변수는 선형이진 분류공식에 적용할 수 있어서 다음과 같이 개수에 상관없이 범주마다 하나의 변수로 표현된다.
3. Merge7 →가변수화 →표준화	두 번째 방법과 같이 Merge7을 가변수화한 후, 추가 변수들 표준화했던 것과 마찬가지로 Z-변환을 통해 가변수에 표준화 작업을 한 뒤 추가변수들과 결합 - 가변수에 표준화를 취한 이유: 각각의 가변수들이 대부분은 0값을 가지므로, 군집마다의 영향력을 다르게 해주기 위함
4. Merge7 →가변수화 →표준화 →가중치 곱	세 번째 방법과 같이 Merge7에 가변수화한 후, 가변수에 표준화작업을 취해준 후, 0~1범위로 가중치를 곱하여서 추가변수들과 결합 -표준화한 가변수에 가중치를 곱한 이유: 표준화하였을 때, 각 군집에 대한 영향력이 너무 커질 수 있어, 영향력을 줄여주기 위함

Merge7(Label) 합치는 방법에 따른 기업정보를 활용한 군집화를 진행할 때 사용된 군집화 방법은 앞선 3단계 군집화 과정 중 세분화단계에서 사용한 군집화 방법인 Agglomerative 군집방법이다. 이때 군집 선택

의 조건으로는 우선 각 군집은 적어도 5개 이상의 기업이 속하는 것을 목표로 한다. 뿐만 아니라 군집의 수가 많을수록 포트폴리오를 구성할 때 분산투자의 기회가 많아져서 투자의 안정성을 확보할 수 있으므로 군집의 수는 기본적으로 많을수록 좋다고 판단한다.

추가적으로 제 4장 연구결과 제 1절 군집화 결과에서 군집화 결과에 대한 평가는 Silhouette score, Calinski-Harabasz score, Davies-Bouldin score 3가지 평가지표를 통해 이루어진다. 각각의 지표들에 대한 설명은 [표 3.5]를 통해 대신한다.

[표 3.5] 3가지 군집 평가지표

평가 지표	설명
Silhouette score	조건의 정의는 다음 식과 같음 $s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$ -여기서 a(i)는 i번째 개체와 같은 군집에 속한 요소들 간거리들의 평균, b(i)는 I번째 개체와 다른 군집 내 개체들 간의 거리를 군집별로 구하고, 이중 가장 작은 값을 의미 -개체별로 위에 식을 계산하여 해당 개체와 내부와의 거리가 짧을수록, 해당 개체와 외부와의 거리가 길수록 s(i)는 커짐
Calinski-Harabasz score	조건의 정의는 다음 식과 같음 $\frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$ -여기서 SS_B는 군집간 전체 분산, SS_W는 군집내 전체 분산,k는 군집 개수, N은 관측치 개수를 의미 -이 비율의 값이 클수록 군집 응집도가 높아지고, 개별 군집의 구분/분리가 확실해짐
Davies-Bouldin score	조건의 정의는 다음 식과 같음 $\frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} (\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$ -여기서 n은 군집 개수,c_x 는 x 군집의 중앙, σ_x 는 x 군집내의 모든 요소들의 centroid와 c_x 까지의 평균거리, 그리고 $d(c_i, c_j)$는 c_i 와 c_j 간의 거리를 나타냄 -군집간 거리가 가깝고 군집내의 거리가 멀수록 군집을 만든 알고리즘 높은 값을 가짐

제 2 절 전자공시 자료 분석

1 자료 수집 및 전처리

전자공시시스템(DART)은 OPEN-DART를 통해 다양한 오픈 API를 제공한다. 오픈 API는 크게 공시정보, 사업보고서 주요정보, 상장기업 재무정보, 지분공시 종합정보 등 4종류 나누어진다. 본 논문에서는 그 중 공시정보의 공시검색 API를 활용하여 KOSPI 200에 속하는 202개 기업들의 2013/01/02월부터 2018/12/26까지 6년 동안의 공시자료를 OPEN-DART를 통해 수집한다. 그 중 앞선 금융시장 군집화 알고리즘의 자료수집 및 전처리 단계에서 제외된 21개의 기업들은 마찬가지로 제외하고 공시자료를 수집하도록 한다. 따라서 총 181개의 기업들에 대하여 각 기업별로 공시자료를 수집하였다. 수집한 자료는 아래와 같이 'corp_cls, corp_name, corp_code, stock_code, report_nm, report_no, flr_nm, rcept_dt' 총 8개의 변수로 구성돼있고, 아래 [그림 3.12]를 통해 수집된 데이터의 전반적인 형태를 제시한다. 그리고 해당변수들에 대한 설명을 [표 3.6]을 통해 제시한다.

	corp_cls	corp_name	corp_code	stock_code	report_nm	rcept_no	flr_nm	rcept_dt
0	Y	삼양홀딩스	126937	000070	주요사항보고서(자기주식처분결정)(자회사의 주요경영사항)	20171226800642	삼양홀딩스	2017-12-26
1	Y	삼양홀딩스	126937	000070	주요사항보고서(자기주식처분결정)	20171226000276	삼양홀딩스	2017-12-26
2	Y	삼양홀딩스	126937	000070	타법인주식및출자증권취득결정(자회사의 주요경영사항)	20171215801213	삼양홀딩스	2017-12-15
3	Y	삼양홀딩스	126937	000070	투자판단관련주요경영사항(자회사의 주요경영사항)	20171129800032	삼양홀딩스	2017-11-29
4	Y	삼양홀딩스	126937	000070	[기재정정]타법인주식및출자증권취득결정(자회사의 주요경영사항)	20171127800678	삼양홀딩스	2017-11-27

[그림 3.12] 2013~2018년 KOSPI200 기업의 수집된 공시자료 데이터 형태

[표 3.6] KOSPI200 기업의 공시자료 변수

변수명	설명
corp_cls	법인구분 : Y(유가), K(코스닥), N(코넥스), E(기타)
corp_name	공시대상회사의 종목명(상장사) 또는 법인명(기타법인)
corp_code	공시대상회사의 고유번호 (8자리)
stock_code	상장회사의 종목코드(6자리)
report_nm	공시구분+보고서명+기타정보
rcept_no	접수번호(14자리)
flr_nm	공시 제출인명
rcept_dt	공시 접수일자

본 논문에서 연구한 전자공시자료 분석에 있어 그 목적이 특정 유형의 공시 자료들이 발표됐을 때, 공시자료가 발표된 날과 공시자료가 발표되지 않은 날들 간의 변동성의 변화율에 유의미한 차이가 있는지에 대한 검정에 있으므로 수집한 공시자료에 필요한 변수는 'report_nm', 'rcept_dt' 와 'corp_name' 세 가지 변수이다. 따라서 해당 변수들을 사용하여 기업별로 공시자료의 유형에 따라 자료를 수집하여 데이터의 형태를 수정한 모습은 [그림 3.13]과 같다. 또한 이때 공시자료 유형은 A: 정기공시, B: 주요사항보고, C: 발행공시, D: 지분공시, E: 기타공시, F: 외부감사관련, G: 펀드공시, H: 자산유동화, I: 거래소공시, J: 공정위공시 총 10가지이다. 그런데 수집된 181개 기업들 중 대부분의 기업들이 F: 외부감사관련, G: 펀드공시, H: 자산유동화와 유형에 대하여 공시자료를 보유하고 있지 않으므로 해당 유형의 공시자료들이 181개 기업들을 대표할 수 없다고 판단하여 제거하였다. 또한 남은 7개의 공시자료 유형에 대한 설명은 [표 3.7](금융위원회 2015)을 통해 제시하였다.

	corp_name	report_nm	rcept_dt	type
0	삼양홀딩스	주요사항보고서(자기주식처분결정)(자회사의 주요경영사항)	2017-12-26	B
1	삼양홀딩스	타법인주식및출자증권취득결정(자회사의 주요경영사항)	2017-12-15	I
2	삼양홀딩스	투자판단관련주요경영사항(자회사의 주요경영사항)	2017-11-29	I
3	삼양홀딩스	[기재정정]타법인주식및출자증권취득결정(자회사의 주요경영사항)	2017-11-27	I
4	삼양홀딩스	분기보고서 (2017.09)	2017-11-14	A

[그림 3.13] 전자공시자료 분석을 위해 수정된 2013~2018년 공시자료 데이터의 형태

[표 3.7] 7개의 전자공시 유형

유형명	설명
정기공시	일정기간동안 기업의 사업내용, 재무상황 및 경영실적 등 기업 내용 전반에 관한 사항을 정기적으로 공시
주요사항보고	사업보고서 제출대상법인은 경영활동과 관련된 사항 중 회사존립, 조직재편성, 자본증감 등 투사의사 결정에 중요한 영향을 미치는 사실이 발생할 때 관련 내용을 공시
발행공시	증권의 공모를 위한 서류로서 증권신고서부터 증권발행실적보고서에 이르기까지 단계별로 공시
지분공시	상장회사 주식 등의 소유 및 변동 정보를 공시
기타공시	A~H 유형에 속하지 않는 나머지 공시
거래소공시	기관별 공시목적에 따른 기업공시 중 하나로 일반 투자자의 의사결정에 필요한 다양한 정보제공이 목적 - 의무공시, 자율공시, 조회공시, 공정공시로 대별
공정위공시	기관별 공시목적에 따른 기업공시 중 하나로 기업집단 및 계열사 간 불공정 내부거래 등 규율하는 것이 목적 - 대규모 내부거래, 기업집단 현황 공시, 비사장사 주요사항으로 대별

다음으로는 기업별로 공시자료 유형에 따라서 변동성을 구한다. 이를 위해 앞선 단계에서 수집했던 2013/01/02년부터 2017/12/29까지의 시계열 종가 데이터를 통하여 아래 수식(8)과 (9)를 활용하여 순간변동성을 통해 평균변동성을 구해준다.

$$\sigma_m^2 = 2(\frac{S_t - S_{t-1}}{S_t} - \ln(\frac{S_t}{S_{t-1}})) \quad - (8)$$

$$\sigma_a^2 = \frac{1}{T}\sum_{t=1}^{T}[\sigma_m^2] \quad - (9)$$

순간변동성(σ_m^2)이란 이산수익률과 로그수익률의 차이의 2배를 의미하고, 위의 수식(a)을 이용하여 현재시점(S_t)과 이전시점(S_{t-1})의 주가만으로 순간변동성을 표현한다. 따라서 2013/01/02년부터 2017/12/29까지 각 날짜별로 전 날 대비 순간변동성을 구해주고, 해당 순간변동성들을 통해서 과거 일정기간(T) 동안의 변동성을 의미하는 평균변동성(σ_m^2)을 통해 5년 동안의 기업별 공시 유형에 따른 평균변동성을 구해준다. 추가적으로 다음 단계를 위해 각 기업에 대한 공시 유형별로 공시자료가 발표된 날들의 5년 동안의 평균변동성과 해당 유형의 공시자료가 발표되지 않은 날들의 5년 동안의 평균변동성 그리고 전체공시자료가 발표되지 않은 날들의 평균변동성을 구해준다. 그 후 아래와 같은 변화율 수식(10)을 이용하여, 공시 자료들이 발표된 날과 공시자료가 발표되지 않은 날들 간의 변동성의 변화율을 구해준다.

$$\text{변동성의 변화율} = (\frac{\text{공시자료가 발표된 날 or 되지 않은 날}}{\text{전체공시자료가 발표 되지 않은 날}}) - 1) \times 100 \quad - (10)$$

2 분석 방법

본 논문에서는 해당 단계에서 쌍체(대응)표본 t-test를 활용하여 특정 유형의 공시 자료들이 발표된 날과 공시자료가 발표되지 않은 날들 간의 변동성의 변화율에 유의미한 차이가 있는지에 대한 가설 검정을 실시한다. 여기서 t-test란 두 집단 간의 평균이 통계적으로 유의미한 차이를 보이고 있는지의 여부를 검증할 때 사용되는 분석 방법을 의미한다. 따라서 쌍체(대응)표본 t-test란 동일한 항목, 사람 또는 물건에 대한 측정값이 두개인 경우에 t-test를 사용하는 분석 방법이다. 검정에 대한 절차는 우선 검정하고자 하는 목적에 따라서 귀무가설과 대립가설을 설정한다. 그 후 검정방법을 결정하고 검정통계량을 계산하고, 유의수준을 결정한다. 끝으로 귀무가설이 옳다는 전제하에 검정통계량으로 계산한 유의확률을 확인한다. 이때 유의확률이 유의수준보다 작거나, 계산된 검정통계량 값이 기각역에 속하면 해당 귀무가설이 기각된다(고승곤 et al 2003). 위 경우 설정된 두 가설은 다음과 같다. ‘H0: 공시자료들이 발표된 날과 공시자료가 발표되지 않은 날들 간의 변동성의 변화율에는 유의미한 차이가 없다.’와 ‘H1: Not H0이다.’이다. 또한 검정통계량 값은 T_O로 아래 수식(11), (12)를 통해 구해준다.

$$T_O = \frac{\overline{D}}{S_D/\sqrt{n}} \sim\ t(n-1) \quad \text{- (11)}$$

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i),\ \ S_D^2 = \frac{1}{n-1}\sum_{n-1}^{n}(D_i - \overline{D})^2 \quad \text{- (12)}$$

여기서 $\overline{D}$와S_D은 두 집단 간의 평균 차와 표준편차 차를, n은 집단의 개수, 그리고 X_i, Y_i는 각 집단 별 해당 기업의 변동성의 변화율을 의미하고, 자유도는 $t(n-1)$이다. 또한 유의수준은 0.05로 설정한다.

제 3 절 포트폴리오 최적화

본 논문에서는 앞선 단계인 금융시장 군집화에서 얻을 여러 형태의 군집들과 전자공시자료 분석을 통해 얻을 유의미한 전자공시 유형을 활용하여 포트폴리오 최적화를 위해 해리 마코위츠가 만든 현대 포트폴리오 이론을 통해 각기 다른 목적을 갖고 있는 두 가지의 Markowitz모형을 사용한다. 이러한 두 모형 간에는 차이가 존재하는데 첫 번째 Markowitz모형은 분산을 최소화하여 포트폴리오 구성하는 것을 목표로 하는 반면에 두 번째 Markowitz 모형은 앞선 이론적 배경 단계에서 설명했던 것으로 평균을 사용하여 평균과 분산의 차를 최대화시켜 포트폴리오를 구성하는 것을 목표로 한다. 모형의 수식에 대한 설명은 (13), (14)를 통해 제시한다.

분산최소화 Markowitz 모형: $\min_w w^T \sum w$ - (13)

Subject to : $R^T w \geq R,\ w^T e = 1,\ Aw = b,\ Cw \geq d$

여기서 w는 포트폴리오를 구성하는 개별 자산의 비중으로 그 합은 1이고, R은 개별 자산의 수익률, $\sum$은 포트폴리오를 구성하는 자산의 수익률에 대한 공분산 행렬을 의미하고, $R^T w$와 $w^T \sum w$은 각각 포트폴리오의 수익률 μ_p와 포트폴리오의 분산 σ_p^2 의미한다.

평균-분산 Markowitz 모형: $\max_w R^T w - \frac{\lambda}{2} \times w^T \sum w$ - (14)

Subject to : $Aw = b,\ Cw \geq d$

여기서 w는 포트폴리오를 구성하는 개별 자산의 비중으로 그 합은 1이고, R은 개별 자산의 수익률, λ는 위험회피 값, $\sum$은 포트폴리오를 구성하는 자산의 수익률에 대한 공분산 행렬을 의미하고, $R^T w$와 $w^T \sum w$은 각각 포트폴리오의 수익률 μ_p와 포트폴리오의 분산 σ_p^2 의미

한다. 여기서 위험회피 값인 λ은 주로 1~5사이로 설정하므로 그 중간 값인 3으로 설정한다(Kim & Sra 2014).

위와 같은 포트폴리오 모형은 금융시장 군집화의 결과로 나올 다른 특징을 가진 여러 형태의 군집들에 의하여 구성된다. 여기서 개별 군집들의 객체인 기업들이 개별 기초자산의 투자 대상이 되고, 모형의 효과를 일반화를 위해 개별 군집들을 구성하고 있는 객체인 기업들에 대한 선택은 랜덤하게하고, 그 과정을 1000번 반복한다.

두 포트폴리오 최적화 모형을 가지고 투자하기에 앞서, 해당 포트폴리오의 업데이트 주기를 정해주어야 한다. 완성된 포트폴리오를 주기적으로 업데이틀 해주는 이유는 금융 시장은 멈추어있는 것이 아니라 계속 꾸준히 변하고 있기 때문이다. 따라서 주기적으로 포트폴리오를 업데이트하여 이러한 금융 시장의 변화를 쫓아 가야한다. 본 논문에서는 1개월 주기로 공분산 행렬에 업데이트를 실시하여 이러한 변화를 따라가고자 한다. 최종적으로 1개월 주기로 업데이틀 실시할 때, 금융시장 군집화 알고리즘에서는 과거의 시계열 데이터의 수익률을 이용하여 변동성을 조정해주는 것을 목적으로 하고, 전자공시 자료 분석에서는 변동성에 유의미한 영향을 미치는 유형의 공시자료를 찾아 해당 공시자료가 발표된 날의 평균변화량을 이용하여 미래의 변동성을 조정하여 포트폴리오를 업데이트 하는 것을 목적으로 한다.

추가적으로 제 4장 연구결과 제 3절 포트폴리오 최적화 결과에 대한 평가는 Rate of Return(수익률), Volatility(변동성), Sharp Ratio 그리고 Turnover 4가지 평가지표를 통해 이루어진다. 각각의 지표들에 대한 설명은 [표 3.8]을 통해 대신한다.

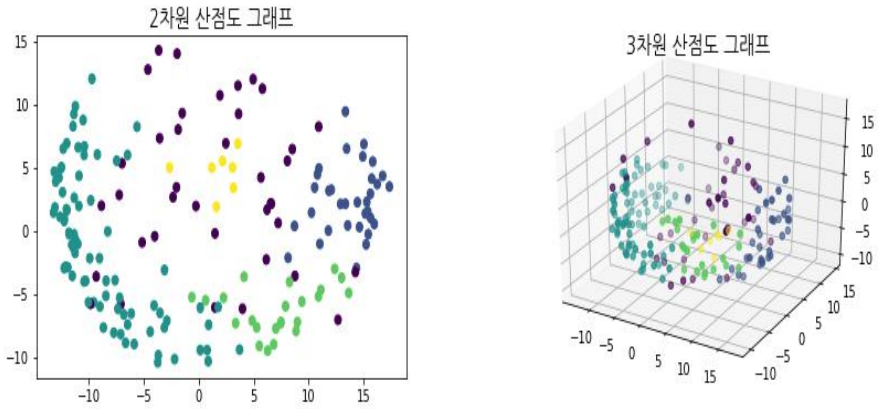
[표 3.8] 4가지 포트폴리오 평가지표

방법	설 명
Rate of Return	$$\mu_p = \frac{1}{T}\sum_{i=1}^{T-1} w_i r_{i+1}$$ μ_p는 포트폴리오의 평균 수익률 w_i는 i시점에서의 포트폴리오 비중을 의미하고, r_i는 i시점에서의 각 자산의 수익률을 의미 T 는 포트폴리오 업데이트 횟수를 의미
Volatility	$$\sigma_p^2 = \frac{1}{T}\sum_{i=1}^{T-1} [w_i r_{i+1} - \mu_p]^2$$ σ_p^2는 포트폴리오의 평균 변동성
Sharp Ratio	$$SR = \frac{\mu_p}{\sigma_p}$$ Sharp Ratio란 위험 자산에 투자함으로써 얻은 초과 수익의 정도를 나타내는 지표로 위험을 얼마나 잘 활용하여 수익을 달성 하는 가를 평가한 지표라 할 수 있음 Sharp Ratio가 높으면 감수한 위험 대비 수익이 좋다는 의미이기 때문에 Sharp Ratio가 높을수록 좋음
Turnover	$$Turnover = \frac{1}{T-1}\sum_{i=1}^{T-1}\sum_{j}^{N}(\lvert w_{j,t+1} - w_{j,t}\rvert)$$ $w_{j,t}$는 j자산의 t시점의 비중을 의미 N은 포트폴리오를 구성하는 자산의 수 Turnover는 포트폴리오 업데이트 과정에서 포트폴리오 구성이 얼마나 변경이 되었는지를 의미 Turnover값이 클수록 포트폴리오를 구성하기 위한 거래 비용 등이 증가함을 의미하므로 작을 값을 가질수록 좋다고 볼 수 있음

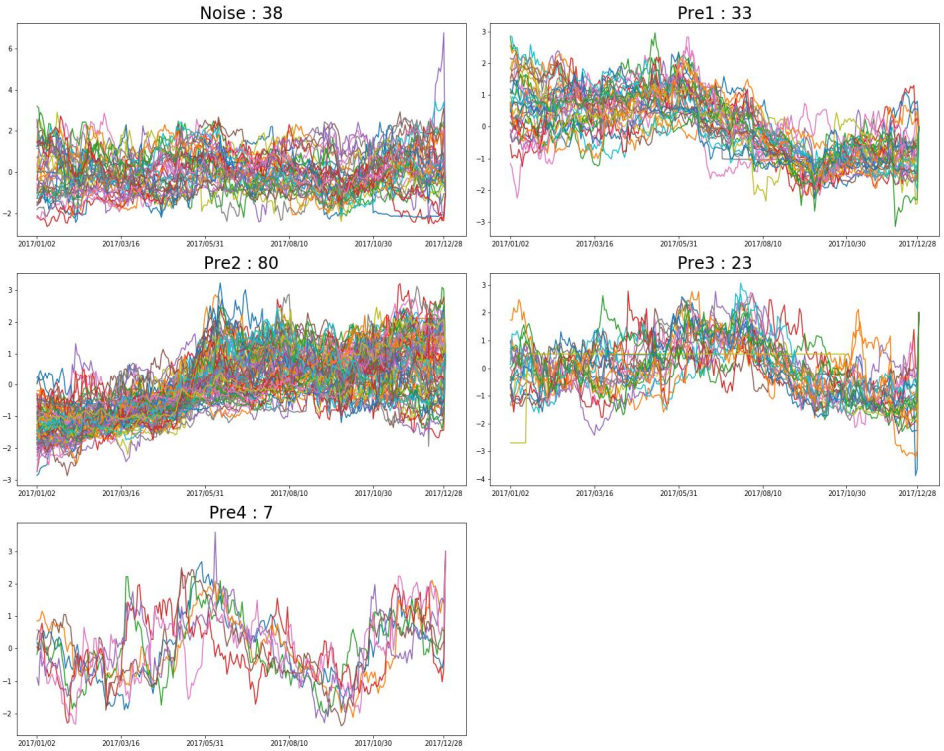
제 4 장 연구결과

제 1 절 군집화 결과

제 3장 제 1절의 2-(가) 3단계 군집화를 통한 시계열 군집화의 첫 단계인 사전 군집화 단계에서의 최적의 파라미터 조합은 min_samples = 6, eps = 4.08이다. 따라서 DBSCAN 군집 방법을 적용한 결과 총 5개의 군집이 선택됐다. 이때 군집별 기업의 개수는 Noise_cluster - 38개, Pre_cluster1 - 33개, Pre_cluster2 - 80개, Pre_cluster3 - 23개, Pre_cluster4 - 7개이다. 군집화 결과가 적절한지에 대한 판단은 2·3차원 산점도 도표를 통해 진행한다. 이를 2·3차원 산점도 도표로 통해 판단한 이유는 주성분분석을 통해 주성분을 3개로 선택했으므로 해당 군집화 과정이 3차원에서 진행됐기 때문이다. 따라서 해당 2·3차원 산점도 도표의 형태는 [그림 4.1]과 같다. 여기서 보라색의 데이터들이 Noise_cluster를 의미하고, 이를 통해 비교적 사전 군집화가 잘 이루어졌다고 판단 할 수 있었다. 다음으로는 4개의 Pre_cluster와 1개의 Noise_cluster로 이루어진 5개의 군집에 대한 시계열 종가데이터 형태는 [그림 4.2]와 같고, 4개의 Pre_cluster들의 시계열 종가데이터의 형태를 살펴보면 대략적으로 비슷한 형태의 데이터들끼리 군집화가 이루어졌음을 알 수 있고, 4개의 사전 군집들 간의 형태의 차이가 존재함을 알 수 있다. 또한 위 결과로 인해 38개의 노이즈 데이터는 앞으로의 군집화 과정에서 제외한다. 따라서 총 143개의 기업을 통해 다음 군집화 과정을 진행한다.

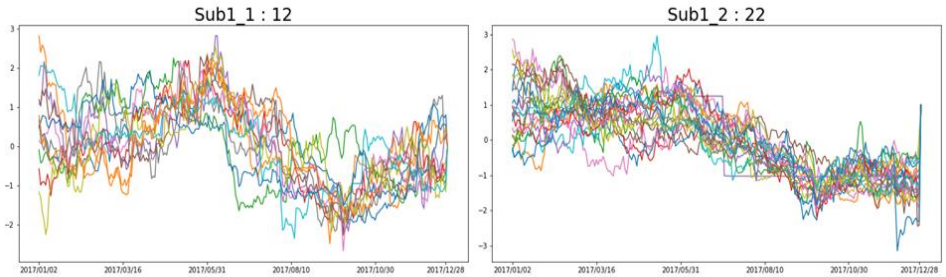


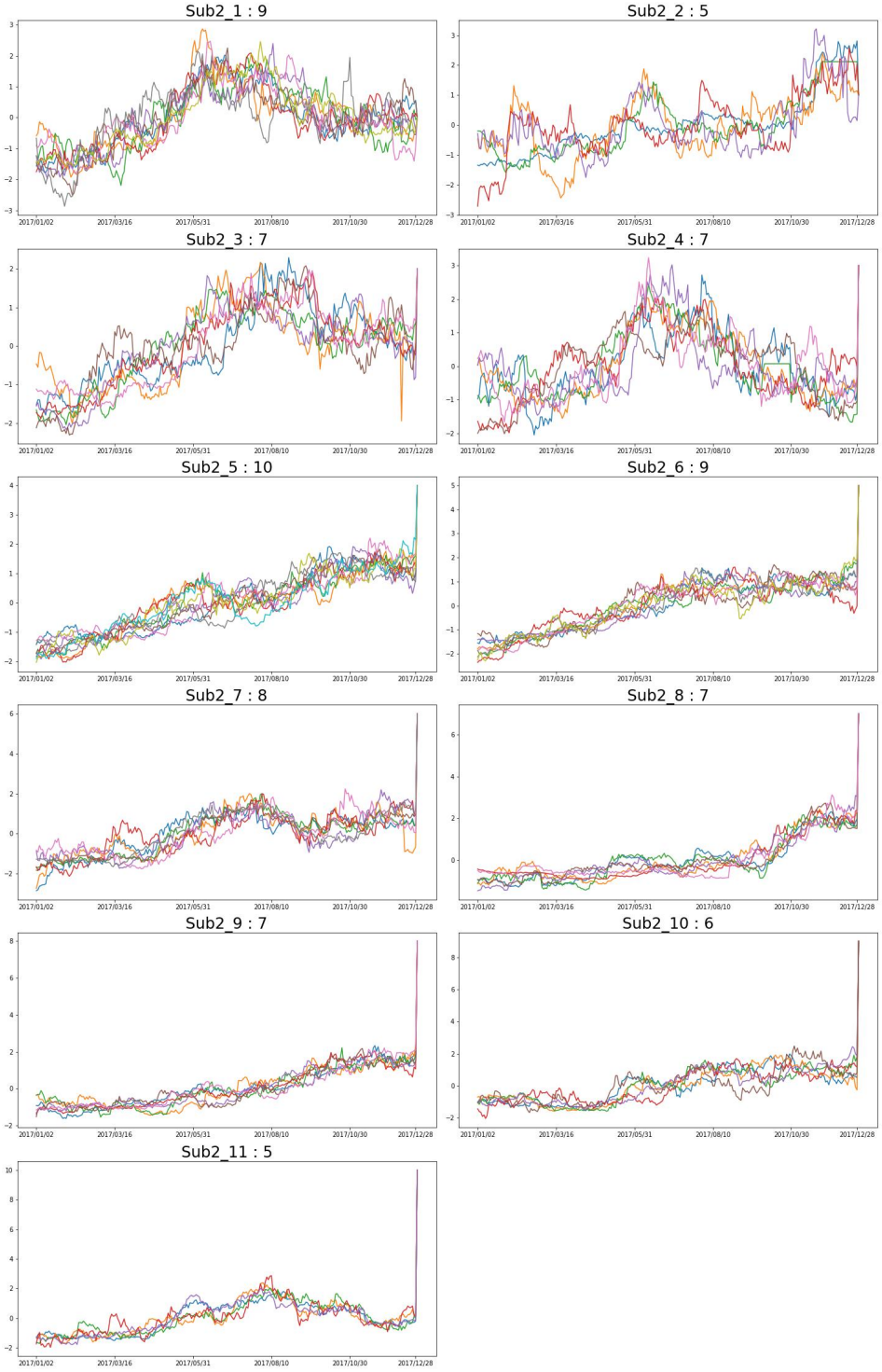
[그림 4.1] DBSCAN 군집화 방법의 결과로 제시된 5개 군집에 대한 산점도

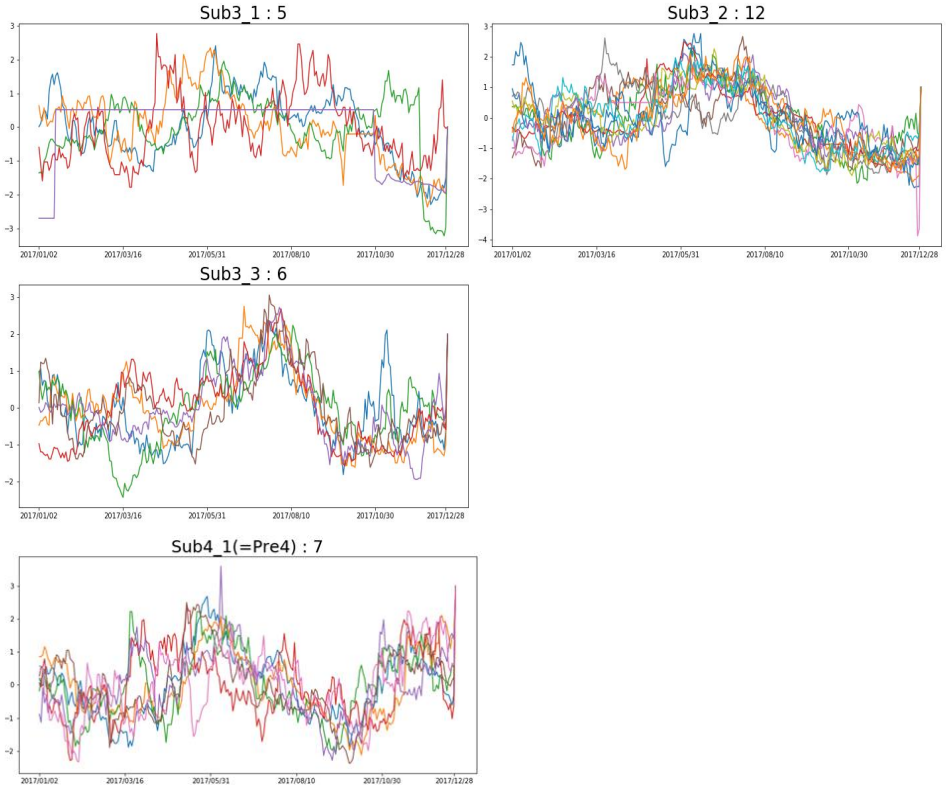


[그림 4.2] 1개의 노이즈 군집과 4개의 사전 군집에 대한 시계열 종가데이터 형태

두 번째 단계인 세분화 단계에서는 정제과정 후, 각 사전 군집들에게 Agglomerative 군집 방법을 적용했다. 그 결과 17개의 세분화된 Sub_cluster가 생성된다. 이렇게 생성된 Sub_cluster는 Pre_cluster1에서 2개, Pre_cluster2에서 11개, Pre_cluster3에서 3개, Pre_cluster4에서 1개이다. 이때 Pre_cluster4는 사전 군집화 단계에서 이미 해당 데이터의 개수가 7개라 따로 Sub_cluster로 세분화하는 단계의 과정을 거치지 않았다. 다음으로 4개의 사전 군집에서 세분화된 17개의 세분화 군집에 대한 시계열 종가데이터 형태는 아래 [그림 4.3]과 같고, 이들 17개의 세분화 군집에 대한 시계열 종가데이터의 형태를 살펴보면 사전 군집화 단계보다 훨씬 더 비슷한 형태의 데이터들끼리 군집화와 세분화가 잘 이루어졌음을 알 수 있다.




Sub2_1 : 9
Sub2_2 : 5
Sub2_3 : 7
Sub2_4 : 7
Sub2_5 : 10
Sub2_6 : 9
Sub2_7 : 8
Sub2_8 : 7
Sub2_9 : 7
Sub2_10 : 6
Sub2_11 : 5
2017/01/02
2017/03/16
2017/05/31
2017/08/10
2017/10/30
2017/12/28



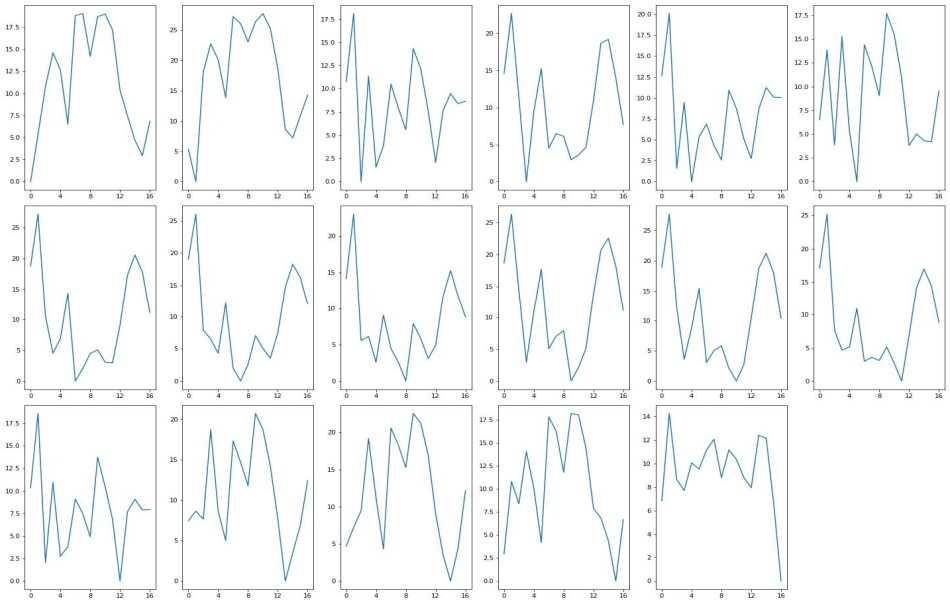
[그림 4.3] 17개의 세분화 군집에 대한 시계열 종가데이터 형태

세 번째 단계인 병합 군집화 단계에서는 해당 단계에서 생성된 Prototype에 K-Means 군집 방법을 적용하여 군집 평가지표들인 Silhouette score, Calinski-Harabasz score, Davies-Bouldin score등을 평가기준 삼아 군집의 수를 5개부터 16개까지 다양하게 조정해가며 해당 과정을 수행하였다. 따라서 위와 같은 평가 기준과 앞서 언급했던 각 군집은 적어도 10개 이상의 기업이 속하는 것을 목표로 군집 판단 조건을 만족하는 최적의 군집의 수를 탐색한 결과를 아래 [표 4.1]을 통해 제시하였다. 이때 [표 4.1]에 대하여 평가지표로만 봤을 때는 군집의 수가 6개나 9개에서 더 좋다고 판단을 내릴 수도 있다. 하지만 군집의 수가 9개인 경우 군집 당 최소 기업의 수인 10개를 만족하지 못하였고, 군집의 수가

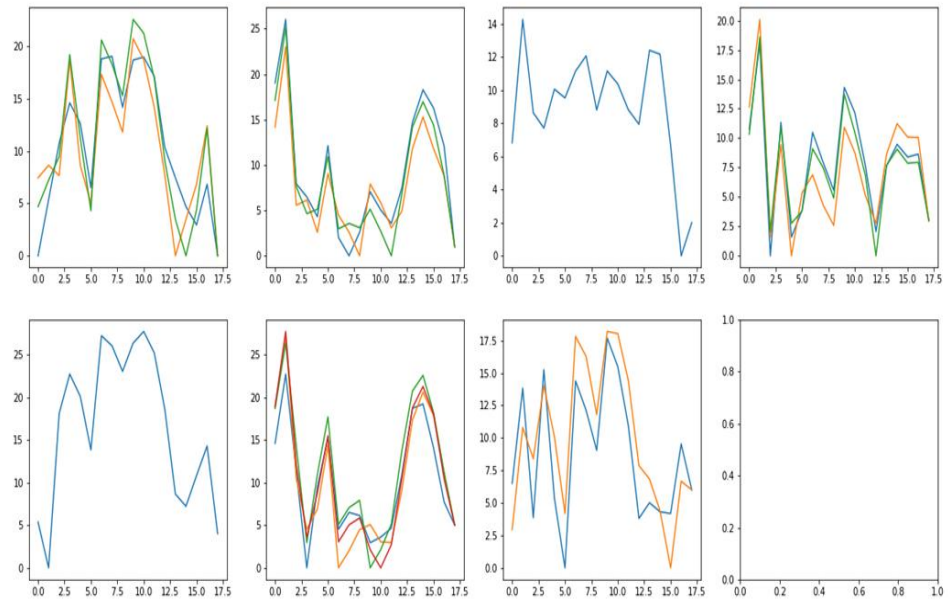
6개인 경우에는 DTW 거리 형태를 확인해본 결과 형태에 있어 명확히 유사하지 않은 부분이 있음에도 불구하고 군집으로 묶였음을 확인했다. 이에 대한 판단은 [그림 4.4]를 통해 각 군집의 개수 별로 [그림 4.5]와 같은 방식으로 통해 각 세분화된 군집들의 DTW 거리 형태와 군집 수에 따른 군집 결과를 보고 결정하였다. 그 결과 [그림 4.5]와 같이 군집 수가 7개일 때, 가장 적절하게 군집화가 이루어졌음 파악할 수 있었다.

[표 4.1] 평가지표를 통한 3단계 군집화 결과

군집 수 / 평가지표	5	6	7	8	9	10
Silhouette score	0.359	0.36	0.283	0.209	0.244	0.203
Calinski score	27.71	29.49	29.847	27.365	32.172	31.589
Davies score	0.649	0.533	0.709	0.683	0.585	0.482
군집 수 / 평가지표	**11**	**12**	**13**	**14**	**15**	**16**
Silhouette score	0.187	0.153	0.124	0.104	0.073	0.063
Calinski score	30.603	30.999	32.35	38.622	45.324	82.658
Davies score	0.448	0.365	0.258	0.266	0.197	0.097

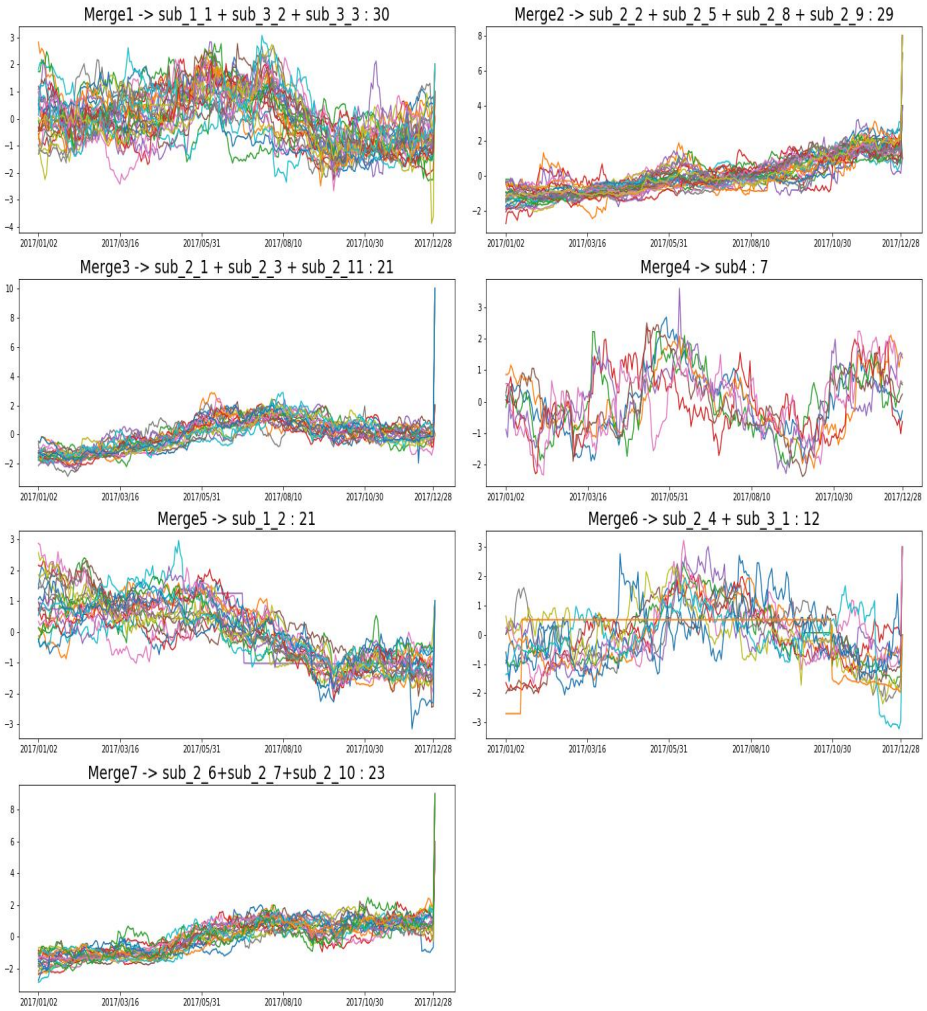


[그림 4.4] 17개의 세분화 군집에 대한 DTW 거리 형태



[그림 4.5] 17개의 세분화 군집에 대한 DTW 거리 형태에 K-Means 군집 방법(k=7)을 적용한 결과

끝으로 17개의 세분화된 군집에서 병합된 7개의 병합 군집에 대한 시계열 종가데이터 형태는 아래 [그림 4.6]과 같고 이들 7개의 병합 군집에 대한 시계열 종가데이터의 형태를 살펴보면 비슷한 형태의 데이터들끼리 군집화가 잘 이루어졌음을 알 수 있다. 또한 [그림 4.6]을 통해 어떤 세분화된 군집들끼리 군집화가 됐는지, 또 해당 병합 군집들이 몇 개의 기업들로 이루어졌는지에 대하여 제시하였다.



[그림 4.6] 7개의 병합 군집에 대한 시계열 종가데이터 형태

다음으로 [표 3.4]와 같은 4가지 방법을 가지고 군집의 개수를 다양하게 조정해가며 군집화를 진행한 결과는 다음과 같다. 우선 방법1과 방법2의 군집화 결과는 상당히 유사하였다. 이를 통해 Merge7을 단순히 가변수화하여 추가변수들과 결합하는 것은 군집형성에 있어 거의 영향을 끼치지 않는 다는 것을 알 수 있다. 추가적으로 이 두 방법의 경우 군집들 간의 데이터 개수의 격차가 상당했고, 큰 2개의 군집이 143개의 기업들 중 약 100개의 기업을 데이터로 갖고 있음을 알 수 있었다. 따라서 방법1과 방법2를 통한 군집화는 적절하지 않다고 판단하였다. 다음으로 방법3을 적용한 결과 군집 선택의 조건에 만족하는 결과들이 나옴을 알 수 있었다. 그러나 군집의 개수를 7개로 지정했을 때 결과를 보면 3단계 군집화를 통한 시계열 군집화 결과와 거의 비슷하게 개수가 나누어지고, K-Means 군집 방법을 적용하면 3단계 군집화와 결과가 동일하게 나옴을 알 수 있다. 이는 Merge7 가변수가 대부분 0을 가지고 있으므로, 그 영향력을 군집마다 다르게 주기위한 표준화 때문이다. 즉, 가변수들의 영향력이 너무 켜졌음을 알 수 있다. 따라서 방법4를 활용하여 가변수들에게 0부터 1까지 범위로 적절한 가중치를 부여하여 영향력을 줄였다. 이때 적절한 가중치의 값을 파악하기 위하여 3단계 군집화를 통한 시계열 군집화 결과인 Merge7과 마찬가지로 군집의 개수를 7개로 고정한 후 가중치를 바꾸어가며 Agglomerative 군집 방법을 실시하였다. 그 결과 가중치를 0.5미만으로 부여하면 방법1과 결과가 비슷하게 나오고, 0.8보다 높게 부여하면 방법3과 비슷한 결과가 나오는 것을 알 수 있었다. 따라서 가중치의 범위는 0.5부터 0.8까지가 적당하다고 판단하였다. 결과적으로 Merge7을 여러 가지 방법으로 합쳐본 결과 Merge7을 가변수화한 후 표준화하고 적당한 가중치를 부여하는 방법인 방법4가 Merge7과 추가변수들 양쪽 모두의 영향력을 적절하게

해주는 것으로 판단됐다. 따라서 포트폴리오 구성에 있어 3단계 군집화를 통한 시계열 군집화와 기업정보를 활용한 군집화 두 군집화과정 모두 영향을 끼칠 수 있는 혼합 알고리즘이 만들어졌다. 한편 가중치의 범위로 정한 최솟값 0.5부터 최댓값 0.8까지에 따른 군집 형성 결과를 비교하던 중 대부분의 경우 효과적인 분산투자를 위한 포트폴리오 구성에 맞게 지정한 군집의 수가 커질수록 군집이 골고루 형성되는 것을 알 수 있었다. 하지만 반복적으로 1개의 기업으로만 이루어지는 군집들이 있음을 발견하였다. 이를 가중치의 범위의 극값인 0.5와 0.8인 경우를 예시로 들어 [표 4.2]를 통해 제시했다.

[표 4.2] 가중치를 양극 값으로 조정했을 때 방법4 군집결과

가중치	각 군집에 속하는 기업의 수
Min 0.5	군집 수 7 : [22 2 101 1 1 1 15] 군집 수 8 : [2 7 101 15 1 1 15 1] 군집 수 9 : [101 7 15 15 1 1 1 1 1] 군집 수 10 : [56 45 15 15 7 1 1 1 1 1] 군집 수 11 : [45 27 15 15 7 1 1 1 1 1 29] 군집 수 12 : [27 39 15 15 7 6 1 1 1 1 29 1] 군집 수 13 : [39 11 15 15 7 16 1 1 1 1 29 1 6] 군집 수 14 : [15 11 24 15 7 16 15 1 1 1 29 1 6 1] 군집 수 15 : [11 7 24 15 29 16 2 1 1 1 13 1 6 1 15] 군집 수 16 : [7 29 24 15 13 16 2 6 1 1 5 1 6 1 15 1]
Max 0.8	군집 수 7 : [29 31 56 1 1 5 20] 군집 수 8 : [56 31 5 7 1 22 20 1] 군집 수 9 : [27 31 5 29 1 22 20 1 7] 군집 수 10 : [5 31 20 29 17 22 10 1 7 1] 군집 수 11 : [31 4 20 29 17 22 10 1 7 1 1] 군집 수 12 : [4 25 6 29 17 20 10 1 7 1 1 22] 군집 수 13 : [3 25 6 29 17 20 10 1 7 1 1 22 1] 군집 수 14 : [25 29 6 2 17 20 10 1 7 1 1 22 1 1] 군집 수 15 : [6 29 20 2 17 14 10 1 7 1 1 22 1 1 11] 군집 수 16 : [20 29 5 2 17 14 10 1 7 1 1 22 1 1 11 1]

해당 군집들의 개수를 파악한 결과 3~6개의 군집이 반복적으로 1개의 기업으로만 형성됨을 알 수 있다. 이렇게 단일로만 지속적으로 형성된다는 것은 해당 기업들이 극단 값을 가지고 있어서라고 판단하였고, [표 3.3] 추가 변수들의 기초 통계량과 비교해본 결과 해당 기업들마다 특정 추가 변수에서 극단 값을 가지는 것을 확인됐다. 따라서 해당 기업들은 이상치를 의미하는 것으로 판단하여 제거하고, 아래 [표 4.3]을 통해 그 값들을 제시했다.

[표 4.3] 극단 값이 존재한다고 판단되는 기업의 추가 변수 기초통계량

변수명 / 기업명	1week_수익률	3month_수익률	1year_변동성	3month_변동성	1year_변동성	시가총액(단위:백만)	PER	PBR
오리온홀딩스	-0.0111	0.1733	-3.1892	-3.7133	-1.5859	1672596	0.7504	0.8789
영진약품	-0.0068	-0.1006	-0.0045	-3.7849	-3.5006	1605798	859.76	13.4208
삼성전자	-0.0303	-0.0459	0.3223	-4.0623	-4.2097	318615100	11.0627	2.1021
한미사이언스	0.03218	0.1985	0.4723	-3.3105	-3.5198	6881926	2434.35	12.1739
현대중공업	-0.3463	-0.3821	-0.4201	-3.0007	-3.4988	5490880	1.3417	0.5231
대우조선해양	-0.0786	-1.1595	1.1429	-2.2029	-1.8494	1475655	1.9964	0.5377

[표 4.3]을 살펴보면 [표 3.3]의 최솟값이나 최댓값을 앞서 언급한 이상 값으로 판단한 6개의 기업들이 가지고 있음을 알 수 있다. 따라서 이들 기업을 제거한 후 총 137개의 기업에 대하여 군집 평가 지표들인 Silhouette score, Calinski-Harabasz score, Davies-Bouldin score등을 평가 기준 삼아, 가중치는 0.5부터 0.8까지, 군집의 개수는 5개부터 16개까지 조절해가며 군집화를 진행하였다. 따라서 위와 같은 평가 기준과 앞서 언급했던 각 군집은 적어도 5개 이상의 기업이 속하는 것을 목표

로 해당 조건을 만족하는 최적의 조합을 Agglomerative 군집 방법을 통해 탐색한 결과를 [표 4.4]를 통해 제시하였고, [표 4.4]의 경우 각 군집에 적어도 5개 이상 속하는 군집들만을 제시하였다. 그 결과 군집의 수가 13개이고 가중치를 0.6으로 부여한 경우가 가장 적합하다는 것을 알 수 있었다. 물론 이때 Calinski-Harabasz가 가장 높지는 않지만 나머지 두 개의 평가 지표에서 가장 뛰어난 성능을 보이므로 최적의 군집 형태라고 판단했다. 또한 3단계 군집화를 통해 나온 7개의 군집과 13개의 군집의 형태를 비교해본 결과 3단계 군집 결과와 비교하여 몇몇 기업들의 군집이 바뀜으로써 적절하게 세분화된 군집이 형성됐음을 알 수 있었다.

[표 4.4] 평가지표를 통한 방법4 군집화 결과

군집 수 / 평가지표	5	6	7	8	9	5	6
Weight	0.5	0.5	0.5	0.5	0.5	0.6	0.6
Silhouette score	0.104	0.095	0.105	0.122	0.14	0.15	0.143
Calinski score	17.716	16.678	16.159	15.711	15.541	18.259	17.143
Davies score	2.06	1.869	1.779	1.679	1.644	1.777	1.706
군집 수 / 평가지표	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Weight	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Silhouette score	0.152	0.147	0.156	0.167	0.187	0.205	0.223
Calinski score	16.441	16.143	16.002	15.883	15.874	15.96	16.08
Davies score	1.638	1.548	1.532	1.536	1.478	1.444	1.373

추가적으로 6개의 기업을 제거한 상태에서 [표 3.4]에 제시된 방법2, 3, 4,를 제외한 나머지 방법인 방법1을 통해 군집을 형성해본 결과, 군집 판단의 조건에 부합함을 확인하였다. 따라서 위와 같은 평가 기준과 앞서 언급했던 군집 판단 조건을 만족하는 최적의 조합을 Agglomerative 군집 방법을 같은 방식으로 탐색한 결과를 [표 4.5]를 통해 제시하였다. 그 결과 군집 수를 8개로 지정하였을 때, 제일 적합한 결과를 보였다. [표 4.5]의 경우도 [표 4.4]와 마찬가지로 각 군집에 적어도 5개 이상 속하는 군집들만을 제시하였다.

[표 4.5] 평가지표를 통한 방법1 군집화 결과

군집 수 / 평가지표	5	6	7	8
Silhouette score	0.112	0.126	0.135	0.147
Calinski score	23.052	21.923	21.319	20.984
Davies score	1.765	1.647	1.615	1.526

결과적으로 금융시장 군집화 과정을 거치면서 포트폴리오 최적화 하는데 있어 적용할 수 있는 총 3가지 군집 결과가 도출 됐다. 하나는 3단계를 군집화를 통해 시계열 종가로만 군집화를 이룬 Mer7이다. 두 번째는 기업정보를 활용한 군집화 단계에서 추가된 기업정보 변수들로만 이루어진 Agg8이다. 마지막은 이 두 과정을 모두 활용하여 도출된 Agg13이다. 해당 군집 결과들에 대한 상세한 설명은 [표 4.6]을 통해 제시했다.

[표 4.6] 금융시장 군집화 과정을 거쳐 나온 3가지 군집 결과

군집 결과	설 명
3단계 군집화를 통한 시계열 군집화 → 7개 군집선택 **(Mer7)**	1단계: 군집분석에 앞서 일별 종가 시계열 데이터에 대한 표준화 및 차원축소 전처리과정을 거친 후, 전처리된 데이터에 군집 알고리즘 DBSCAN 방법을 활용하여 Pre-Cluster를 생성 2단계: 생성된 Pre-Cluster을 세분화하기 위해 정제 단계를 거친 후, 가중거리함수를 이용해 각각의 사전 군집들에 계층적 군집 알고리즘인 Agglomerative 방법을 활용하여 Sub-Cluster를 생성 3단계: 2단계에서 생성한 Sub-Cluster를 병합하기 위하여 시간의 흐름에 따른 패턴 간 유사성을 기초로 하는 DTW거리 행렬을 생성 후, DTW거리 행렬에 군집 알고리즘인 K-Means를 활용하여 7개 군집을 생성 (Merge-Cluster)
기업정보를 활용한 군집화 →8개 군집선택 **(Agg8)**	기업정보를 활용한 군집화 과정에서 생성된 추가 변수들인 수익률(3), 수익률의 변동성(2), 시가총액, PER, PBR 등 8개 변수를 가지고 표준화 시킨 뒤, 군집을 형성하는 방법 위에서 언급한 이상 값으로 판단되는 6개의 기업을 제거하고 군집화를 실시, Agglomerative 방법을 통해 8개의 군집을 생성
기업정보 활용 8개 변수 + 3단계의 군집 라벨 더미변수 군집화 →13개 군집선택 **(Agg13)**	기업정보 변수에 대해서 이상 값으로 판단되는 6개 기업을 제외하고, 3단계 군집 Label인 Merge7을 가변수화한 후 표준화를 실시하고 가변수의 영향을 줄이기 위해 적절한 가중치(0.6)를 부여해서, 표준화된 기업정보 변수와 결합하여 군집화를 실시, Agglomerative 방법을 통해 13개의 군집을 생성

제 2 절 공시자료 분석 결과

앞 선 단계의 군집화 결과로 남은 총 137개의 기업에 대하여 공시 유형별로 공시자료가 발표된 날과 발표되지 않은 날들 간의 변동성의 변화율에 유의미한 차이가 있는지에 대하여 쌍체(대응)표본 t-test로 검정한 결과를 [표 4.7]을 통해 제시하였고, [표 4.8]을 통해 유형별로 5년간 발표된 공시의 개수를 나타냈다. 이때 쌍체(대응)표본 t-test의 기본 가정인 정규성을 만족하지 못하였으나, 집단의 크기가 30개 이상이므로 중심극한정리에 의하여 정규성 검정을 만족한다고 판단할 수 있었다. 그 결과 B: 주요상항보고, E: 기타공시, I: 거래소공시 3가지 유형에서 각각 유의수준 0.05하에서 각각 유의확률 0.00083, 0.0099, 0.0000019로 귀무가설 H0를 기각하므로 해당 유형들의 경우공시자료가 발표된 날과 발표되지 않은 날들 간의 변동성의 변화율에 유의미한 차이가 존재함 알 수 있었다. 위와 같은 3가지 공시가 유의미하게 나온 이유로는 우선 B: 주요사항보고의 경우 법인의 경영·재산 등에 중대한 영향을 미치는 상항이 발생한 경우에 발표하는 공시이기에 투자의사 결정에 중요 영향을 미치기 때문이라고 판단하였다. 다음으로 I: 거래소공시의 경우 주로 주가 등에 큰 영향을 미치거나, 기업 스스로 판단하여 중요하다고 생각하는 항목들이 발표하는 공시이기기 때문에 이 역시 투자의사 결정에 있어 중요 영향을 미친다고 생각하였다. 즉, 해당 유형의 공시들이 발표된 날의 변동성의 변화율이 발표되지 않은 날보다 크다는 것을 알 수 있었다. 반면에 E: 기타공시의 경우 A~H유형에 속하지 않는 나머지 공시들, 즉 별로 중요하지 않은 공시들로 구성됐기에 오히려 해당 유형의 공시가 발표된 날의 변동성의 변화율이 발표되지 않은 날보다 작다는 것을 알 수 있었다.

[표 4.7] 7가지 공시유형에 대한 대응표본 t-test 결과 (유의수준 = 0.05)

유형명 / 결과	정기공시	주요사항 보고	발행공시	지분공시	기타공시	거래소공시	공정위공시
검정통계량	-0.246	3.431	-0.972	0.365	-2.615	4.969	-1.351
p-value	0.805	0.00083	0.333	0.715	0.0099	0.0000019	0.18

[표 4.8] 137개 기업의 공시유형별 5년간의 발표된 공시 수

유형명	정기공시	주요사항 보고	발행공시	지분공시	기타공시	거래소공시	공정위공시
개수	2918	885	1996	6285	1988	9541	3413

제 3 절 포트폴리오 최적화 결과

본 논문에서 제시한 결과 값들은 해당 단계에서 수행되는 알고리즘을 1000번 반복한 값의 평균이다. 위 알고리즘은 금융시장 군집화와 전자공시자료 분석 과정을 거쳐 남은 137개 기업들의 18년도 데이터에 대하여 수행됐다. 이때 18년도에 약 6개월간 거래중지 기간이 있는 한일시멘트 기업의 경우 데이터 자체가 존재하지 않아 분석이 불가능하므로 추가적으로 제거하였다. 따라서 총 136개 기업들에 대하여 해당 군집 결과에 개수에 맞게 포트폴리오를 구성하고 최적화를 두 단계를 나누어 진행하였다.

첫 단계에서는 분산최소화 Markowitz모형과 평균-분산 Markowitz모형에서 금융시장 군집화의 결과로 얻은 Agg13, Agg8, Mer7로 포트폴리오를 구성하여 최적화를 실시한다. 그 후 두 모형 중 성능이 더 우수한 모형과 어떤 군집 결과로 포트폴리오 구성하는 것이 최적화에 더 적합한지 평가하였다. 그 후 우수한 모형과 더 적합한 군집 결과에 대한 추가 비교를 위해, 우수한 군집 결과와 같은 개수로 포트폴리오를 구성하는 랜덤모형을 만들어 포트폴리오를 구성하고 최적화 시킨 뒤 평가하였다. 그 다음 단계에서는 앞서 얻은 최적의 모형과 군집 결과에 전자공시자료 분석을 결과를 추가한 혼합 Markowitz모형을 만들어 공분산행렬 조정해주어 최적화를 진행하였고, 그 결과를 일반 Markowitz모형 비교하였다

그 첫 단계의 결과를 아래 [표 4.8]과 [표 4.9]를 통해 제시하였다. 포트폴리오 최적화에 대한 평가지표들을 통해 그 결과를 살펴보았을 때, 두 Markowitz모형 모두에서 전반적으로 군집 개수가 13개인 Agg13의 성능이 우수함을 알 수 있었다. 이는 13개의 기업들을 임의로 뽑은 랜

덤모형과도 비교해서도 마찬가지였다. 다음으로 두 Markowitz모형을 비교한 결과 다른 모든 지표에서는 전반적으로 분산최소화 Markowitz모형이 우수하였지만, 수익률 지표에서 만큼은 평균-분산 Markowitz모형이 더 우수함을 알 수 있었다.

[표 4.9] 분산최소화 Markowitz모형의 포트폴리오 최적화 결과

군집 결과 / 평가지표	Agg13	Agg8	Mer7	Random13
Volatility	0.36791	0.44480	0.43983	0.41914
Rate of return	0.19452	0.21419	0.20222	0.18988
Sharpe ratio	0.34584	0.33643	0.32259	0.30781
Tunover	0.41368	0.43852	0.44455	0.45529

[표 4.10] 평균-분산 Markowitz모형의 포트폴리오 최적화 결과

군집 결과 / 평가지표	Agg13	Agg8	Mer7	Random13
Volatility	2.13622	1.77708	2.16603	2.02071
Rate of return	0.27287	0.13603	0.27734	0.23235
Sharpe ratio	0.19369	0.10962	0.20220	0.17666
Tunover	1.17081	1.12315	1.20140	1.17679

그 결과 분산최소화 Markowitz모형에서의 Agg13군집 결과를 대상으로 하여 다음 단계를 진행하였다. 따라서 공시 자료들이 발표된 날과 발표되지 않은 날들 간의 변동성의 변화율에 유의미한 차이가 있던 3가지 유형에 대하여 공시자료가 발표된 날의 평균 변화량 값을 통해 분산

최소화 Markowitz모형의 공분산행렬을 조정하여 주었고, 그 값은 B:주요사항보고의 경우 1.136, E:기타공시의 경우 0.957, I:거래소공시의 경우 1.028이었다. 결과는 아래 [표 4.10]을 통해 제시했다. 그 결과 금융시장 군집화 결과에 전자공시 자료분석의 결과를 적용해준 혼합-분산최소화 Markowitz모형이 금융시장 군집화 결과만 가지고 포트폴리오 최적화를 진행한 것보다 Volatility와 Turnover지표에서 확실하게 유의미한 차이를 보여주며 성능이 우수함을 보였고, Sharpe Ratio와 Rate of Return에서도 분산최소화 Markowitz모형보다 조금 낮았지만 그 차이가 크지 않았다.

[표 4.11] 두 분산최소화 Markowitz모형의 포트폴리오 최적화 결과

모형명 / 평가지표	분산최소화 Markowitz모형	혼합-분산최소화 Markowitz모형
Volatility	0.36791	0.31255
Rate of return	0.19452	0.17526
Sharpe ratio	0.34584	0.33671
Tunover	0.41368	0.36415

제 5 장 결론

제 1 절 결론 및 시사점

본 연구에서 제안한 방법을 KOSPI200 자료에 적용하여 포트폴리오 최적화의 성능을 비교해보았을 때, 최종적으로 혼합-분산최소화 Markowitz모형의 결과가 가장 우수하였고, 결과적으로 수익률 지표를 제외하면 분산 최소화 Markowitz모형이 평균-분산 Markowitz모형보다 다른 평가 지표에서는 우수하다는 것을 알 수 있었다. 하지만 수익률 지표에서만큼은 평균-분산 Markowitz모형이 우수했으므로 High-risk & High-return의 금융 상품을 원하는 투자자에게는 평균-분산 Markowitz모형을 추천해주는 것도 가능다고 판단하였다.

본 논문은 새로운 금융시장 군집화 알고리즘에 대한 제안과 전자공시자료가 주가의 변동성에 유의미한 영향을 미친다는 사실을 확인하여 포트폴리오를 구성하고 최적화는데 있어 새로운 방향을 제시했다는 것에 의의가 있다. 좀 더 구체적으로는 포트폴리오 구성할 때는 금융시장 군집화 알고리즘 결과를, 최적화시에는 전자공시자료 분석을 통해 미래의 변동성을 조정하는 방법을 제안하였다는 것에 의의가 있다고 할 수 있다.

제 2 절 추후 과제

우선 [표 4.10]을 보았을 때, 혼합-분산최소화 Markowitz모형을 통해 Volatility와 Turnover 지표는 분산최소화 Markowitz모형보다는 향상 시킬 수 있었지만 Rate of return과 Sharpe ratio 지표는 작게나마 저하됐음을 알 수 있었다. 이는 전자공시자료 분석을 통해 Markowitz모형 공분산 행렬 조정해주는 행위를 통해 변동성만이 수정됐기 때문이라고 판단 할 수 있었다. 따라서 수익성도 조절 할 수 있는 알고리즘을 추가적으로 보완할 경우 Rate of return과 Sharpe ratio지표 또한 조금 더 향상 시킬 수 있을 것으로 예상된다.

또한 본 논문의 경우 포트폴리오를 구성할 때, 여러 금융시장들 중에서 주식시장에 한해서만 포트폴리오를 구성했다는 점에서 한계점을 가진다. 따라서 추후 연구에 있어서는 주식시장뿐만 아니라 다른 직접금융시장인 채권시장 또는 간접금융시장인 은행 등을 포트폴리오 구성에 있어 추가하는 것이 바람직할 것으로 예상된다.

추가적으로 공시 유형에 대한 변동성을 규정할 때, 중복공시에 대한 문제가 존재하였다. 중복공시란 크게 두 가지 경우로 나누어질 수 있다. 첫 번째는 특정한 날에 발표된 여러 공시가 다른 유형의 경우로 각각의 영향력이 어떠한지에 대하여 종가데이터만을 가지고는 판단하기가 어려웠다. 두 번째는 특정한 날에 발표된 공시가 같은 유형의 경우에도 마찬가지로 어려움이 존재하였다. 하나의 공시자료로 보기에도 애매하고, 그렇다고 해당 공시자료의 영향력을 나누는 것도 어려웠다. 따라서 추후 연구에 있어서는 위와 같은 중복 공시의 영향력 문제를 고려해주어 조금 더 명확한 기준이 존재해야 할 것으로 판단하였다.

끝으로 차후 혼합 알고리즘을 통하여 포트폴리오 구성 및 최적화를

유가증권시장 뿐만 아니라 코스닥시장에도 적용하여 그 효과성을 확인해 볼 필요가 있을 것으로 예상된다. 코스닥 시장의 경우 유가증권시장보다 주가가 소액이기 때문에 코스닥시장에 대한 분석이 추가로 이루어진다면 큰 자산을 보유하고 있지 않은 투자자들도 해당 알고리즘을 사용하여 효율적인 최적의 포트폴리오를 구성할 수 있는 가능성을 제시할 수 있을 것으로 기대된다.

참고문헌

고승곤, 양완연, and 오현숙. “일반통계학.” 서울: 교우사 (2003).

금융보안원 보안기술연구팀. “국내・외 금융권 머신러닝 도입 현황”, 2017

금융위원회 공정시장과. “기업공시제도 - 규제선진화 참고자료”, 2015

김성문, 김홍선. “한국 주식시장에서 비선형계획법을 이용한 마코위츠의 포트폴리오 선정 모형의 투자 성과에 관한 연구.“ 경영과학 26.2 (2009): 19-35.

김정은, 남기석. “우리나라 전자공시시스템의 현황 분석.“ 글로벌경영연구 29 (2017): 1-37.

김지혜. “로보 파이낸스가 만드는 미래 금융 지도.” 한스미디어 (2017).

김형준, 박종원, 이재원. “전자공시시스템 (DART) 을 활용한 국내 텍스트 분석 (Textual Analysis) 환경에 관한 연구.“ 회계저널 24.4 (2015): 199-221.

라채원, 박경진. “전자공시시스템이 개별기업의 정보반영에 미친 영향에 관한 연구.“ 회계저널 19.1 (2010): 203-231.

안준규, 이주홍. “동조화 관계를 갖는 시계열을 위한 군집화 알고리즘.“ 한국지능시스템학회 논문지 27.6 (2017): 552-559.

이래학. “전자 공시 100% 활용법.” 이레미디어 (2017).

이장건, 정용기. “기업 이익정보에 대한 전자공시의 정보효과.“ 회계정보연구 26.1 (2008): 313-347.

정창원. “전자공시시스템(DART)이 자본시장의 정보비대칭에 미치는 영향.” 연세대 학위논문(석사) (2007).

KEB하나은행 하이로보센터. “2018 대한민국 로보어드바이저 보고서”, 2018

Aghabozorgi, S. and The, Y. W. “Stock market co-movement assessment using a three-phase clustering method.“ Expert Systems with Applications 41.4 (2014a): 1301-1314.

Aghabozorgi, S., Teh, Y. W., Tutut, H., Hamid, A, J., Mohammad, A. S. and Alireza, J. “A hybrid algorithm for clustering of time series data based on affinity search technique.“ The Scientific World Journal 2014 (2014b).

Aman, H. and Moriyasu, H. “Volatility and public information flows:

Evidence from disclosure and media coverage in the Japanese stock market.“ International Review of Economics & Finance 51 (2017): 660-676.

Bravo, F. “Forward-looking disclosure and corporate reputation as mechanisms to reduce stock return volatility.“ Revista de Contabilidad 19.1 (2016): 122-131.

Chen, N,, Ribeiro, B., Vieira, A. and Chen, A. “Clustering and visualization of bankruptcy trajectory using self-organizing map.“ Expert Systems with Applications 40.1 (2013): 385-393.

Kim, N. and Sra, S. “Portfolio optimization with groupwise selection.“ Industrial Engineering & Management Systems 13.4 (2014): 442-448.

Lai, C. P., Chung, P. C. and Tseng, V. S. “A novel two-level clustering method for time series data analysis.“ Expert Systems with Applications 37.9 (2010): 6319-6326.

Markowitz, Harry M. “(1952). Portfolio selection.“ Journal of Finance 7.1 (1952): 77-91.

Momeni, M., Mohseni, M. and Soofi, M. “Clustering stock market companies via K-means algorithm.“ Kuwait Chapter of the Arabian Journal of Business and Management Review 4.5 (2015): 1.

Mousa, G. A. and Elamir, E. A. H. “The relationship between corporate forward-looking disclosure and stock return volatility.“ Problems and perspectives in management 16, Iss. 3 (2018): 130-149.

Nanda, S. R., Mahanty, B. and Tiwari, M. K. “Clustering Indian stock market data for portfolio management.“ Expert Systems with Applications 37.12 (2010): 8793-8798.

ABSTRACT

Financial market analysis and portfolio optimization using artificial intelligence technology

Park, Han Sang
Advised by Prof. Kim, Nam-hyoung
Dept. of Applied Statistics
Graduate School of
Gachon University

Recently, investors need smart financial services beyond simple convenience. Therefore, domestic and foreign financial companies are introducing machine learning technology to realize artificial intelligence using Big Data. Robo-Advisor is a representative financial service expected to grow through the introduction of machine learning technology. In the case of Robo-Advisor, it means an online asset management service that performs portfolio management through advanced algorithms and Big Data. This paper proposed a new mixed clustering algorithm that can be applied to the domestic stock market using machine learning technology. It was also confirmed that individual electronic disclosure data have a significant effect on the volatility of domestic stock price. Based on these facts, this paper proposed a new method to adjust predictive volatility. It is also meaningful that the proposed methods have been proved empirically that they can be used for portfolio composition and optimization, which is the basis of Robo-Advisor.

Keywords: Clustering, DART, KRX, Markowitz Model, Portfolio

Optimization, Time-series

인공지능 기술을 활용한 금융시장 분석과 포트폴리오 최적화

朴 翰 相