



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

박사학위논문

빅데이터와 인공지능을 이용한
포트폴리오 관리

Portfolio Management Using
Big Data and Artificial Intelligence

상명대학교 대학원

경영공학과

한 창 훈

2018년 2월

박사학위논문

빅데이터와 인공지능을 이용한
포트폴리오 관리

Portfolio Management Using
Big Data and Artificial Intelligence

상명대학교 대학원

경영공학과

한 창 훈

2018년 2월

빅데이터와 인공지능을 이용한
포트폴리오 관리

Portfolio Management Using
Big Data and Artificial Intelligence

지도교수 신 현 준

본 논문을 박사학위 논문으로 제출함

상명대학교 대학원

경영공학과

한 창 훈

2018년 2월

한 창 훈의

박사학위 논문을 인준함

심사위원장 서 광 규 ①

심사위원 안 범 준 ①

심사위원 신 현 준 ①

심사위원 백 종 관 ①

심사위원 이 근 철 ①

상명대학교 대학원

2018년 2월

차 례

표차례	i
그림차례	ii
국문 요약	iii
제 1장 서론	1
제 1절 연구의 배경 및 목적	1
제 2절 논문의 구성	4
제 2장 포트폴리오 구성 전략	5
제 1절 문제의 정의	8
제 2절 포트폴리오 구성	9
2.1 마코위츠 모형의 이론적 고찰	9
2.2 마코위츠 모형의 한계점	11
2.3 경영 효율성 평가의 이론적 배경	12
2.4 경영 효율성을 고려한 포트폴리오 선택	15
제 3절 실험결과 및 분석	17
3.1 실험 계획	17
3.2 포트폴리오 수익률 분석	18
3.3 성과 측정	21
제 3장 섹터 투자 전략	24
제 1절 문제의 정의	24
제 2절 연구의 자료	29
2.1 섹터 선정	29
2.2 텍스트마이닝	30

제 3절 섹터 투자 전략	34
제 4절 실험결과 및 분석	37
4.1 실험 계획	37
4.2 실험 결과 및 분석	38
4.3 성과 측정	41
제 5절 응용 사례	43
5.1 사례 적용 배경	43
5.2 클러스터링을 이용한 패턴 분석	46
5.2.1 클러스터링의 이론적 배경	47
5.2.2 SOM의 이론적 배경	49
5.3 실험 결과 및 분석	51
5.3.1 실험 계획	51
5.3.2 클러스터링 결과 및 분석	53
제 4장 결론	59
참고문헌	63
ABSTRACT	67

표 차 례

<표 2-1> 선행연구에서의 투입 및 산출 요소	14
<표 2-2> 실험 계획	17
<표 2-3> 분기 별 수익률 결과(단위, %)	20
<표 2-4> 성과 측정(단위, %)	23
<표 3-1> KOSPI와 섹터간의 상관계수	30
<표 3-2> 선정된 에너지화학 섹터 키워드	33
<표 3-3> 실험 계획	37
<표 3-4> 11개 섹터의 투자 수익률	40
<표 3-5> TM-T의 성과 측정	42
<표 3-6> 수집 자료	52
<표 3-7> 실험 계획	53
<표 3-8> 클러스터 별 주가 움직임	55

그 립 차 례

<그림 2-1> 국내 ETF 순자산 추이	6
<그림 2-2> 효율성 점수 산출 사례	15
<그림 2-3> PM_{DEA} , PM_{ETF1} , PM_{ETF2} 의 누적 수익률	20
<그림 2-4> PM_{DEA} , PM_{ETF1} , PM_{ETF2} 의 누적 수익률	21
<그림 3-1> 텍스트마이닝을 이용한 ETF 투자 전략	28
<그림 3-2> R을 이용한 키워드 선정	33
<그림 3-3> 검색상대변화량과 ETF 가격과의 관계	36
<그림 3-4> 패턴 분석 과정	46
<그림 3-5> 클러스터링의 개념	47
<그림 3-6> 학습 과정	54
<그림 3-7> 클러스터 별 포함된 일 수	57
<그림 3-8> 클러스터 별 대표 패턴	57

국 문 요 약

빅데이터와 인공지능을 이용한 포트폴리오 관리

본 논문은 금융 시장에 존재하는 다양한 빅데이터를 계량적인 방법론을 통해서 포트폴리오 관리 및 섹터 투자 전략에 대해서 연구하였다. 본 논문은 크게 2장과 3장에서 앞서 기술한 방법론을 설명하고 있는데, 2장에서는 포트폴리오 구성에 있어서 기업의 다양한 재무 자료를 자료포락분석 기법을 이용해 포트폴리오를 구성하는 전략에 대해서 설명하고 있다. 특히 최근 글로벌 경제 위기에 대한 불안감이 높아지면서 많은 투자 기관들은 액티브(active) 시장에 대한 불신이 커지고 반면 ETF 상품과 같은 패시브(passive) 시장에 대한 선호도가 높아지고 있다. 그러나 국내 패시브 시장의 대표라고 할 수 있는 ETF는 짧은 기간 동안에 많은 ETF 종목이 상장하였지만 대부분 시가총액이 높은 우량 종목들로 포트폴리오를 구성하고, 구성 비율 또한 특정 기업에 편중되는 한계점을 내포하고 있다. 때문에 성장 가능성이 높거나 기업의 가치에 비해서 저평가된 종목들이 포함되지 않기 때문에 시장 수익률 대비 초과 수익률이 크게 높지 않다. 따라서 2장에서는 기업의 경영 효율성을 평가하는 자료포락분석법을 이용해서 포트폴리오를 구성하고 이를 실제 상장되어있는 ETF 상품과 비교 분석하고자 한다. ETF 상품의 특성 상 특정한 섹터(sector)를 선정해야 하는데 본 연구에서는 IT 업종을 선정한다. 따라서 KOSPI와 KOSDAQ에 상장되어있는 IT 기업들을 대상으로 자료포락분석법을 이용하여 포트폴리오를

구성하고, 이와 가장 유사한 섹터를 구성하고 있는 TIGER 200 IT와 TIGER 소프트웨어라는 ETF 상품과 수익률 분석 및 포트폴리오 성과 측정을 하였다. 그 결과 수익률 측면과 성과 측정에 있어서 본 연구에서 제안하는 자료포락분석법을 이용한 ETF 상품 운용이 실제 시장에 상장된 벤치마크 ETF 상품에 비해서 월등하게 우수한 성과를 보이는 것으로 나타났다.

다음 3장에서는 금융 시장에 존재하는 다양한 키워드인 빅데이터를 이용해 2장에서 소개한 ETF를 정량적으로 매매하는 전략에 대해서 설명한다. 이처럼 빅데이터는 다양한 산업 분야에서 현업에 적용하고자 하는 시도가 증가하고 있으며, 특히 금융 시장에 활용하려는 연구들이 활발하게 진행되고 있다. 더불어 금융시장 관련 빅데이터 트렌드가 주식시장의 움직임을 선 반영할 수 있다는 사실이 최근 연구들에 의해 입증되고 있다. 기존 연구와 달리 본 연구에서는 주식 시장의 트렌드를 보다 세분화하여 포착하기 위해 주식 시장을 11개의 섹터로 세분화한다. 각 섹터의 트렌드를 대표하는 키워드들을 텍스트마이닝과 브레인스토밍 기법을 통해 각각 선정하고 5년간의 트렌드 데이터를 수집함으로써 섹터별 상장지수펀드(ETF) 투자 포트폴리오 전략을 수립한다. 섹터별 투자성과를 누적수익률 및 연도별 수익률 관점에서 비교한 결과 텍스트마이닝 기법에 기반을 둔 섹터 트렌드 전략이 보다 우수한 성과를 보이는 것으로 나타났다.

또한 본 논문에서 제안하는 다양한 방법론은 정성적이 아닌 과학적이고 정량적인 방법론에 의존하기 때문에 향후 기계 학습을 적용한

인공 지능적인 포트폴리오 관리와 매매 전략 시스템을 개발하는데 큰
기여를 할 것으로 사료된다.

제 1 장. 서 론

제 1 절 연구의 배경 및 목적

국내·외 금융상품(financial instruments)을 운용하는 기관들은 개인은 물론 연기금(pension funds), 뮤추얼펀드 등 다양한 기관에서 투자된 자금을 관리하는데 막대한 책임감을 갖는다. 따라서 이러한 투자자금을 효율적으로 관리할 수 있는 방안은 운용 기관들에 있어서 매우 중요한 사안이다. 그러나 대부분의 투자 결정은 인간의 감정이 내포된 정성적인 의사결정 또는 비합리적인 정량적 시스템을 투자 판단에 적용하는 경우가 존재하기 마련이다.

일반적으로 막대한 자금을 정량적인 시스템을 이용해서 운용을 하는 경우에는 해리 마코위츠(Harry Max Markowitz)의 포트폴리오 선정모형을 활용한다. 그러나 종목들 간의 상관관계를 최소화하는 마코위츠 포트폴리오 선정 모형은 종목들의 과거 자료에만 의존하기 때문에 기업의 현재 가치를 적절하게 판단하고 이를 투자 결정에 반영하지 못한다는 한계가 있다.

이처럼 포트폴리오를 운용하는 전략에 있어서 주관적인 오판(misjudgment)을 최소화하고 과학적이고 정량적인 투자 시스템을 수립하는 것이 매우 중요한데, 최근에는 알파고(AlphaGo)의 등장 이후에 금융 분야에도 인간의 주관적 판단을 최소화하고 빅데이터(big data)가 반영된 인공지능적인 로보어드바이저(robo-advisor)가 관심을 받고

있다. 따라서 본 논문에서는 크게 두 가지로 나눠 금융 분야에 적용 가능한 정량적이고 과학적인 방법론을 제안하고자 한다.

하나는 기업의 방대한 재무 자료를 자료포락분석(Data Envelopment Analysis; 이하 DEA)에 적용하여 포트폴리오를 구성하는 전략에 대해서 제안한다. DEA 기법은 정량적으로 기업의 경영효율성평가를 하고 해당 기업의 경영 효율성을 계량적인 수치로 나타낼 수 있는 방법론인데 본 연구에서는 이 기법을 통해서 포트폴리오의 구성 방법론과 구성 비율 등에 대해서 설명하고자 한다. 또한 이러한 방법론의 성과를 분석하기 위해서 최근 자산 규모가 커지고 있는 상장지수펀드(exchange traded fund; 이하 ETF)와 비교·분석하고자 한다. ETF는 기업의 산업군, 즉 특정 섹터로 구분하여 개별 종목들이 구성된 지수인데 일반적으로 ETF에 구성되는 종목과 비율들은 개별 종목의 시가총액에 의해 구성된다. 또한 DEA 기법은 동일한 섹터를 갖는 기업들을 대상으로 경영효율성평가를 산출하는 것이 이론적으로 더 타당하기 때문에 본 연구에서는 DEA 기법을 이용해서 상장된 IT기업들을 대상으로 포트폴리오 구성하고 실제 상장된 TIGER 200 IT와 TIGER 소프트웨어라는 ETF 상품과 수익률 분석 및 포트폴리오 성과 측정을 하고자 한다.

다음으로는 좀 더 빅데이터를 활용하여 금융 상품 매매 전략에 적용하고자 투자 판단에 영향을 끼치는 다양하고 방대한 자료들을 수집하고 이를 텍스트마이닝(Text Mining) 기법을 통해서 금융 상품의 매매 전략에 적용하고자 한다. 앞서 설명했듯이 빅데이터에 대한 관심은 매

우 커지고 있는 실정이다. 실제 빅데이터는 민간 및 공공부문에 있어서 다양한 부가가치를 창출하고 효과적인 활용으로 생산성 향상, 기업의 경쟁력 제고 그리고 국가 미래전략 지원 및 공공 서비스의 혁신이라는 기대감을 주고 있다. 이처럼 빅데이터가 주목받는 가장 주된 요인은 대용량 데이터의 분석과 추론을 통해서 새로운 서비스를 제공할 수 있고 금융업, 제조업, 유통업 등 다양한 산업에 적용가능하다는 것이다. 또한 온라인 산업과 서비스가 발달하면서 온라인 네트워크 이용자들의 트렌드(trend)에 대한 분석이 빅데이터 분석을 통해서 가능하고 이는 기업에서 다양한 의사결정을 하는데 있어서 중요한 척도가 되고 있다. 특히 포털 빅데이터는 온라인 네트워크를 이용하는 사람들의 일상생활에 내재된 행동양식이 반영되어 있으며 복잡한 현상에 대한 근본적인 문제를 해결하기 위한 새로운 기회가 된다. 이러한 포털 빅데이터는 금융 시장과도 밀접한 관련이 있으며, 빅데이터를 이용한 인공지능 시스템을 개발하고자 하는 수요는 향후 계속 증가할 것으로 판단된다. 따라서 본 논문에서는 주식 시장의 트렌드를 세분화하기 위해서 총 11개의 섹터로 나누고 각 섹터들에게 영향을 끼치는 키워드들을 텍스트마이닝과 브레인스토밍 기법을 통해서 선정하고자 한다. 이렇게 선정된 키워드들을 11개의 상장된 ETF 상품의 매매 전략에 적용하여 그 우수성과 실효성 등을 분석하고자 한다.

제 2 절 논문의 구성

본 논문은 지금까지 기술한 제 1장을 포함하여 전체 4장으로 구성되어 있으며, 각 장에서 설명하는 내용은 다음과 같다.

제 1장에서는 본 연구의 배경과 목적을 설명하고 있다. 크게 정량적인 포트폴리오 구성 방안의 필요성과 ETF와 같은 섹터로 구성된 상품의 투자 전략에 있어서 발생하는 난제에 대해서 기술하고 있다.

제 2장에서는 개별 종목들에 대한 포트폴리오 구성에 있어서 경영과 학 분야의 경영효율성평가를 적용하는 방법론을 제안하고, 이에 대한 실험 결과 등을 기술한다.

제 3장에서는 빅데이터를 이용하여 ETF를 정량적으로 매매하는 방법론과 함께 인공지능이 적용가능 한 패턴 분석을 통한 금융 상품 투자전략에 대한 방법론에 대해서 응용 사례를 통해서 설명하고자 한다.

마지막으로 4장에서는 본 연구의 결론을 정리하여 기술하고 추후 연구에 대해서 기술한다.

제 2 장. 포트폴리오 구성 전략

제 1 절 문제의 정의

한국 증시는 국내뿐만 아니라 글로벌(global) 금융 시장의 변수에 노출되어 있다. 예컨대 2016년 6월 24일 브렉시트(brexit) 사태로 인해서 KOSPI는 장중 1,892.75pt까지 하락했으며, 2016년 11월 9일에는 트럼프 미국 대통령이 당선되면서 KOSPI가 장중 1,931.07pt까지 하락했다.

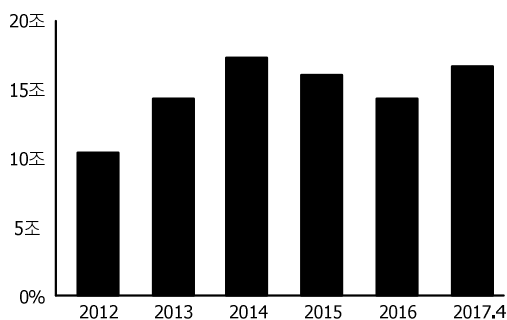
이처럼 금융 시장의 변동성(variability)과 위험(risk)이 높아지고 불안한 증시가 지속됨에 따라 투자자들은 이러한 위험을 최소화하고 안정적인 투자 상품을 선호한다. 또한 사회는 낮은 출산율과 급격한 노령화로 인해서 퇴직연금과 같은 장기적 관점의 투자가 많이 이뤄진다.

이와 같은 시장 상황에서 투자 위험을 최소화하면서 낮은 비용으로 시장대비 손실위험을 축소할 수 있는 대표적인 투자 대안으로 ETF와 같은 패시브펀드(passive fund)가 주목을 받고 있다. 개인 투자자뿐만 아니라 연기금과 공제회와 같은 기관에서도 종목을 선별해서 투자하는 액티브펀드(active fund) 운용 비중을 대폭 낮추고 지수나 자산 군에 투자하는 ETF의 투자 비중을 늘리고 있다. KOSPI보다 낮은 성과를 내는 액티브펀드에 돈을 맡기는 것보다 경기 흐름과 시장에 투자하는 패시브펀드에 투자하는 것이 낫다는 판단에서다. 실제로 2016년도 국내 펀드매니저가 운용하는 주식형 공모 펀드의 평균 수익률은 약 -1%였고 동일 기간의 KOSPI 수익률은 약 5%로 시장 수익률이

더욱 높았다.

이처럼 액티브펀드의 평균 수익률이 시중은행의 정기 예금 금리에도 못 미치면서 국내 기관들은 ETF와 같은 패시브펀드에 집중 투자를 늘려가고 있다. 예컨대 자산 규모가 약 22조원에 달하는 한국교직원공제회는 2017년도에 ETF 투자 금액을 2016년도에 비해서 약 60% 이상 늘리기로 했으며, 대한지방행정공제회는 2016년도에 운용 자금 5,000억 중에서 약 4,000억을 ETF에 투자했다. 그 밖에 군인공제회, 사립학교교직원연금공단, 공무원연금공단 등의 기관들도 2017년도부터 ETF 투자를 크게 늘리고 있다.

<그림 2-1>은 2012년부터 2017년 4월까지의 국내 ETF의 순자산 추이를 나타내고 있다. 2012년도에 ETF 순자산은 약 10조에 불과한 ETF 자산이 2017년 4월에는 약 16조 이상인데 이러한 추세라면 2018년에는 약 20조를 돌파할 가능성이 매우 높다. 반면 주식형 공모펀드는 2012년도에 약 60조인 반면 2017년 4월에는 약 40조로써 약 20조가 줄어들었다.



<그림 2-1> 국내 ETF 순자산 추이

이러한 현상은 국내 시장 뿐만 아니라 미국의 경우도 마찬가지이다. 미국 네바다공무원연금은 패시브펀드에 대한 투자를 지속적으로 늘리면서 패시브펀드에 투자하는 금액이 약 39조원을 넘어섰다. 또한 2017년도 1분기에 전 세계 투자자들이 ETF에 투자한 자금이 약 230조원으로 이는 분기 기준 사상 최대치이며, 2016년도 한 해 동안 ETF로 유입된 자금은 약 4천억 달러로서 ETF에 유입되는 투자 금액이 크게 늘어나고 있는 실정이다.

반면 지난해 액티브펀드는 전 세계적으로 약 5천억 달러의 투자 자금이 유출되었다. 또한 미국의 대표적인 로보어드바이저 운용사인 Betterment도 포트폴리오 자산의 대부분이 ETF 상품이다(박재연, 2016). 그러나 ETF가 이처럼 투자자들에게 많은 주목을 받고 있는 만큼이나 좀 더 개선해야 할 사안들도 있다.

ETF는 평균적으로 20개 이상의 종목들을 포함하고 있으며 일반적으로 해당 ETF 상품들의 성격에 맞는 종목들 중에서 우량한 기업의 주식을 위주로 포트폴리오를 구성하고 있다. 예컨대 IT 업종의 ETF는 상장된 IT기업들 중에서 시가총액 순으로 우량한 기업들 위주로 포트폴리오를 구성하고 있으며, 우량한 기업들 순으로 자산 배분도 이뤄지고 있다. 때문에 ETF를 매입하는 것은 투자자가 여러 기업들을 분산 투자 하면서 발생하는 불편함을 해소한다는 장점이 가장 크게 부각되며 과학적인 방법에 의해서 포트폴리오를 구성하지 않기 때문에 시장 수익률 대비 초과 수익이 그리 크지 않다.

물론 ETF가 안전 자산이라는 성향을 갖고 있지만 ETF 성격 상 약 20개에서 30개 정도의 자산을 포함해야 한다는 가정이 있다면 보다 더 과학적인 접근을 통해서 투자 가치가 높은 기업으로 구성할 수 있다. 예를 들어 KODEX IT는 ETF는 상장된 기업들 중에서 IT와 관련된 기업들을 구성하였고 약 25개의 기업을 포함하고 있다. 그러나 시가총액이 가장 높은 기업들 위주로 구성되어 있으며, 포트폴리오 자산 배분 비율의 상위 1위와 4위의 기업이 전체 포트폴리오의 70%이상의 비율로 자산 배분이 편중되어 있다.

때문에 본 연구에서는 기업들의 재무 자료를 이용하여 ETF 상품을 개발하는 방법론을 제안하고자 DEA 기법을 이용하여 포트폴리오를 구성하고 현재 상장된 ETF 상품과 비교·분석하고자 한다. 이를 위해서 본 연구에서는 국내 상장된 IT 업종을 대상으로 실험하고 유사한 ETF 상품인 TIGER 200 IT와 TIGER 소프트웨어의 수익률 측면으로 비교하고자 한다. 또한 상품의 포트폴리오의 성과 측정 분석을 위해서 샤프 지수(sharp ratio), 젠센의 알파(Jensen's alpha) 그리고 정보비율(Information Ratio; 이하 IR)을 산출하여 실효성을 분석하고 그 우수성을 보이하고자 한다.

제 2 절 포트폴리오 구성

2.1 마코위츠 모형의 이론적 고찰

투자전략에는 크게 적극적인 투자 전략(active investment management)과 소극적인 투자 전략(passive management strategy)으로 구분한다. 적극적인 투자 전략은 평균적으로 기대하는 수익률보다 높은 수익률을 추구하는 것을 목표로 포트폴리오를 구성하는 전략으로써 투자자는 지속적으로 시장의 수익률보다 높은 수익률을 창출할 수 있다는 판단 하에 투자를 진행한다. 이는 주식 시장에 대한 분석이 불완전하다는 가정 하에 특정 종목의 기업에 대한 성장 가능성과 안정성 등을 모두 반영한 내재적 가치와 시장에서 형성되는 시장 가격(market price) 간의 괴리가 발생하는 종목들에 투자하는 방법이기도 하다. 반면에 소극적인 투자 전략은 체계적인 분석 능력을 필요로 하는 적극적인 투자 전략과는 반대로 기계적이고 단순하게 투자하는 방식이다. 소극적 투자 전략에는 단순 매입 전략과 평균 투자 전략 그리고 시장지수 펀드 전략(index fund)이 있다.

포트폴리오 구성에 있어서 정량적이며 수리적인 포트폴리오 관리를 위해서 가장 대표적이고 일반적인 마코위츠 포트폴리오 선정 모형을 이용한다. Markowitz, H. M.(1952)에 의하면 본 모형은 포트폴리오의 위험률을 최소화하면서 평균 수익률을 보장해 주는 단일기간의 정적 모형이다.

마코위츠 포트폴리오 선정 모델을 이용하여 포트폴리오를 구성하기 위해서는 같은 위험 하에서는 기대 수익률이 높은 것이 그리고 같은 기대 수익률 하에서는 위험이 작은 것이 그렇지 않은 투자 대상을 지배해야 한다는 지배원리(dominance principle)를 고려해야 한다. 이러한 이론적 배경에 입각한 것이 효율적 투자대상(efficient investment)이다.

마코위츠 모형은 위험의 정도를 나타내는 포트폴리오의 수익에 대한 분산을 최소화(minimize)하는 것을 목적함수로 정의한다. 더불어 기본적으로 공매도가 존재하지 않는 가정 하에 투자자가 요구하는 최소기대수익률을 달성해야 하고, 투자가 가능한 금액을 모두 포트폴리오에 투자해야 한다는 제약조건이 따른다.

마코위츠 모형의 목적 함수와 제약 식은 다음과 같다.

$$\text{Minimize} \quad \sum_{i=1}^N \sum_{j=1}^N w_i w_j \sigma_{ij} \quad (1)$$

$$\text{Subject to} \quad \sum_{j=1}^N \mu_j w_j \geq K \quad (2)$$

$$\text{Subject to} \quad \sum_{j=1}^N w_j = 1 \quad (3)$$

여기서,

N 포트폴리오에 포함시킬 수 있는 투자 대상의 종목 수

μ_j 주식 j 의 평균수익률($j=1,2,\dots, N$)

σ_{ij} $i \neq j$ 인 경우 주식 i 와 주식 j 의 수익률에 대한 공분산

K 포트폴리오에 요구되는 최소 기대수익률

w_j 주식 j 에 투자하는 비율, 결정 변수($j=1,2,\dots,N$)

식 (1)은 마코위츠 포트폴리오 선정 모형의 목적함수로서, 포트폴리오를 구성할 수 있는 종목들의 공분산을 최소화하기 위한 식이다. 즉 포트폴리오에 투자 위험을 최소로 할 수 있는 종목들을 구성함으로써 제약식인 식 (2)에서의 K 만큼의 포트폴리오에 대한 최소 기대수익률을 기대할 수 있다. 식 (3)은 투자 원금의 100%를 포트폴리오에 투자한다는 제약식이다.

2.2 마코위츠 모형의 한계점

마코위츠 모형은 다음과 같은 몇 가지의 조건을 가정하고 있으며 이는 다음과 같다.

- (1) 모든 투자자는 위험을 기피하려는 성향이 있기 때문에 같은 수익률이라면 위험이 작은 투자를 선호한다.
- (2) 모든 자산에 대한 기대 수익률이 알려져 있다.
- (3) 모든 자산의 분산과 공분산이 알려져 있다.
- (4) 평균과 분산만으로 모든 자산의 수익률 분포를 알 수 있다. 즉 첨도나 외도는 고려하지 않는다.
- (5) 거래 수수료나 세금과 같은 거래 비용은 고려하지 않는다.
- (6) 공매도가 존재하지 않는다.

이러한 한계점은 실무적으로 다양한 제약을 갖게 되기 때문에 현실적으로 마코위츠 모형을 이용해서 포트폴리오를 관리하는 방법론은 현실적인 보완작업 없이는 활용하기 어렵다. 또한 본 모형은 기업의 과거 주가에만 의존하고 기업의 재무 자료 등을 고려할 수 없다는 한계점도 내포하고 있다.

더불어 마코위츠 모형은 기대 수익률보다 과거에 상승한 종목들을 기준으로 선별해주기 때문에 기대 수익률을 정의하기가 애매하고, 과거에 크게 하락하다가 기술적 반등 또는 재무구조 개선 등에 의해서 상승을 하는 종목들에 대한 투자 기회를 가질 수 없다는 한계점도 있다. 따라서 본 장에서는 기업의 재무 자료를 바탕으로 경영 효율성을 고려하여 포트폴리오를 구성하기 위해 경영과학 분야의 DEA 기법을 통해서 포트폴리오 구성 방법론을 제시하고자 한다.

2.3 경영 효율성 평가의 이론적 배경

기업의 경영 효율성을 산출하는 방법론인 DEA 모형은 효율적인 DMU의 개별적인 관찰에 초점을 두어 개선 가능성에 대한 유용한 정보를 제공하기 때문에 각 DMU를 상대적으로 평가하여 효율성을 측정함과 동시에 개선안을 제시할 수 있다는 장점이 있다(이철기, 2016).

DEA 모형을 통한 효율성 평가를 실시함에 있어서 가장 중요한 요소 중 하나는 투입 및 산출요소의 선정에 있다. 따라서 본 연구에서는

선행연구를 통한 투입 및 산출요소 선정 과정을 거쳐 요인을 선정하기로 한다.

다음 <표 2-1>은 다양한 산업에서 DEA 모형을 통해 효율성 평가를 실시하였을 때 투입 및 산출 요소를 선정한 요소들을 선행 연구를 통해서 정리한 표이다(최다영 2011). DEA 모형을 통한 효율성 평가를 위해서 평가 하고자 하는 업종들의 대상은 같은 산업 군으로 제한을 둔다.

본 연구에서의 투입 요소와 산출 요소는 선행연구를 참고하고 업종의 특성을 고려하여 투입 요소는 판매비와 관리비, 자본총계, 자산총계 그리고 산출 요소는 매출액, 반기순이익으로 정의한다.

경영 효율성을 분석할 수 있는 DEA 모형에는 CCR(Charnes-Cooper-Rhodes), BCC(Banker- Charnes-Cooper), 가변형 모형(additive model), 슬랙 중심 측정모형(slacks-based measure) 등 많은 종류가 개발되었다(김종기, 2008). 본 논문에서는 규모수익성(Return to Scale; RTS)의 가변을 가정한 BCC 모형을 기반으로 실험을 실시하고자 하며 투입중심 BCC 모형을 선형계획모형으로 정리하면 다음과 같다(유재필, 2013).

$$\text{Minimize} \quad \eta \quad (4)$$

$$\text{Subject to} \quad \eta x_0 - X\lambda \geq 0 \quad (5)$$

$$y_0 - Y\lambda \leq 0 \quad (6)$$

$$e\lambda = 1 \quad (7)$$

여기서,

η DMU₀의 투입물 승수

x_0, y_0 DMU₀의 투입물과 산출물 벡터

X, Y 전체 DMU들의 투입물과 산출물 행렬

λ 가중치 벡터

e 1로만 이루어진 벡터

<표 2-1> 선행연구에서의 투입 및 산출 요소

구분	투입요소	산출요소	업종
박철수 (2003)	월전수표, 인당관리비, 직원 수, 사무실면적, 유가증권, 투자액	대출금 총액, 예수금 총액	은행업
이형록 (2010)	총자본, 판매비와, 관리비, 종업원수	매출액, 당기순이익	건설업
박홍균 (2010)	직원수, IT시스템 수, 창고 수	매출액	운송업
하헌구 (2007)	종업원 수, 고정자산, 자본총계, 운영비용	매출액, 당기순이익	물류업
이형석 (2007)	종업원 수, 고정자산, 총자본	매출액, 당기순이익	철강업
김주백 (2004)	자본금, 고정자산	당기순이익, 매출액	해운업
신정훈 (2016)	총자산, 종업원수, CAPEX 증감액	당기순이익, 매출액	자동차업

위 식 (4)~(7)를 바탕으로 DEA 경영 효율성 분석을 통해 산출된 효

효율성 값은 기업 별로 100에서 0까지의 점수로 산출되며, 이렇게 산출된 효율성 점수를 바탕으로 포트폴리오 자산 배분의 비율에 적용되는데 이는 다음 절에서 설명한다.

2.4 경영 효율성을 고려한 포트폴리오 선택

DEA 모형을 이용하여 기업들의 경영 효율성을 평가하면 <그림 2-2>와 같이 기업 별로 경영 효율성 점수를 산출할 수 있다. 본 연구에서는 국내 상장된 약 70개에 해당하는 IT 기업을 대상으로 실험을 하였고 매년 편입 또는 상장폐지 등으로 인해서 실험 기간 동안 실험 대상의 종목 수는 다를 수 있다. <그림 2-2>는 DEA 모형을 계산할 수 있는 Frontier Analyst 소프트웨어를 이용하여 기업들의 효율성 점수를 산출한 결과를 예시로 보여주고 있다.

Units	
Unit name	Score
(주)케이티스	100.0%
(주)큐로컴	100.0%
(주)시공테크	100.0%
(주)케이티뮤직	100.0%
한솔인티큐브(주)	100.0%
유엔젤(주)	95.7%
(주)정원엔시스	92.1%
(주)텍셀네트컴	89.5%
아이크래프트(주)	88.5%
(주)링네트	68.8%
네이버(주)	68.3%
(주)경봉	66.7%
(주)디지털조선일보	64.3%
(주)카카오	62.9%
(주)사람인에이치알	61.7%
에스넷시스템(주)	59.3%
엔에이치엔한국사이버결제(주)	58.6%

<그림 2-2> 효율성 점수 산출 사례

<그림 2-2>는 2016년도 1분기에 2015년도 4분기의 재무 자료를 바탕으로 효율성 점수를 산출한 사례인데 실제 효율성 점수를 산출하면 네이버와 카카오는 상대적으로 점수가 높지 않은 것을 알 수 있다. 하지만 본 연구의 비교 대상인 TIGER 200 IT는 동일 기간에 네이버는 약 20% 그리고 카카오는 약 13%의 포트폴리오 비중을 보이고 있다. 반면 DEA 모형을 이용하여 기업의 경영 효율성 점수를 산출하면 케이티스, 큐로컴, 시공테크와 같은 기업이 더욱 효율성 점수가 높게 나오는 것을 알 수 있다. 본 연구에서는 효율성 점수를 기반으로 자산 비율을 선정하는데 한 자산 당 5%의 비율을 초과할 수 없도록 설정한다. 일반적으로 약 70개의 기업을 대상으로 DEA 모형을 통해서 효율성 점수를 산출하면 대략 5개에서 10개 기업의 효율성 점수가 100이 나온다. 즉 효율성 점수를 바탕으로 다음과 같은 공식을 이용하여 투자 비율을 산정한다. 효율성 점수로 내림차순하고 식(8)을 적용하면 약 70개의 종목에서 대략 20개에서 30개 사이로 포트폴리오가 구성된다.

$$100\% \asymp \sum_{i=1}^N E_i * 0.05 \quad (8)$$

여기서,

E 기업들의 효율성 점수

본 연구에서는 이렇게 구성된 포트폴리오와 실제 ETF와 비교 분석하고자 하며, 이는 다음 절에서 설명한다.

제 3 절 실험결과 및 분석

3.1 실험 계획

본 장에서는 본 연구의 실험 계획 및 DEA 기법을 통해서 구성된 포트폴리오와 실제 시장에서 거래되고 있는 ETF와 수익률 비교와 포트폴리오 성과 분석을 통한 결과를 설명하고자 한다.

포트폴리오 구성을 위해 2013년을 기준으로 KOSPI에 상장된 IT 업종 약 70개 기업을 선정하였고 <표 2-2>는 실험 데이터와 실험 계획에 대한 설명이다.

<표 2-2> 실험 계획	
대상	상장된 약 70개의 국내 IT 기업
기간	2013년 1월 1일 ~ 2016년 12월 30일
평가 주기	4회/년 (분기마다 포트폴리오 재구성)
성과 측정	벤치마크 대비 포트폴리오 수익률, 샤프지수, 켄센의 알파, IR 등
자료 수집	이데일리 MARKETPOINT, FN 가이드

본 연구에서 제시하는 DEA를 이용한 포트폴리오 구성은 선행 실험 데이터를 활용한다. 예컨대 2012년도 4분기의 재무 자료를 이용하여 2013년도 초에 포트폴리오를 구성하고 2013년도 1분기의 재무 자료를 바탕으로 2013년도 2분기 초에 포트폴리오를 구성한다. 수익률 비교의

형평성을 위해서 벤치마크인 TIGER 200 IT와 TIGER 소프트웨어의 수익률도 포트폴리오와 동일한 기간만큼의 수익률을 산출하여 비교 분석하였다. 편의를 위해 DEA를 이용한 포트폴리오 구성 전략을 PM_{DEA} 로 TIGER 200 IT를 PM_{ETF1} 로 표현하고 TIGER 소프트웨어는 PM_{ETF2} 로 표현하였다.

실험 기간은 2013년도부터 2016년도 말까지이며, 분기 별 재무 자료를 바탕으로 분기마다 포트폴리오를 구성하기 때문에 총 16번의 포트폴리오를 구성하고 이에 해당하는 분기 수익률을 각각 산출한다.

벤치마크와는 분기 별 수익률 분석과 함께 포트폴리오 성과 측정 지표인 샤프지수, 켄센의 알파 그리고 IR 등 다양한 측면으로 비교 분석하고자 한다. 기업들의 재무 자료는 2012년도 4분기부터 2016년도 4분기까지 FN 가이드에서 제공하는 재무 자료를 활용하고, 주가 자료는 이테일리의 MARKETPOINT를 이용하여 수집하였다.

3.2 포트폴리오 수익률 분석

<표 2-3>은 PM_{DEA} 와 PM_{ETF1} 그리고 PM_{ETF2} 의 분기 별 수익률과 년도 별 누적 수익률 그리고 4개년도 총 누적 수익률을 나타내고 있다. 2013년도는 PM_{ETF2} 가 39.7% 누적 수익률을 보이면서 PM_{DEA} 와 PM_{ETF1} 에 비해서 수익률 측면에서 우수하다. 특히 PM_{DEA} 는 4분기 수익률이 약 -18%를 기록하면서 2013년 누적 수익률에 큰 영향을 끼쳤다. 2014년에는 PM_{DEA} 의 누적 수익률이 약 48%로 가장 월등하게 나타나는데 특

히 1분기와 3분기에 각 각 약 16%, 약 31%의 높은 수익률을 기록했다.

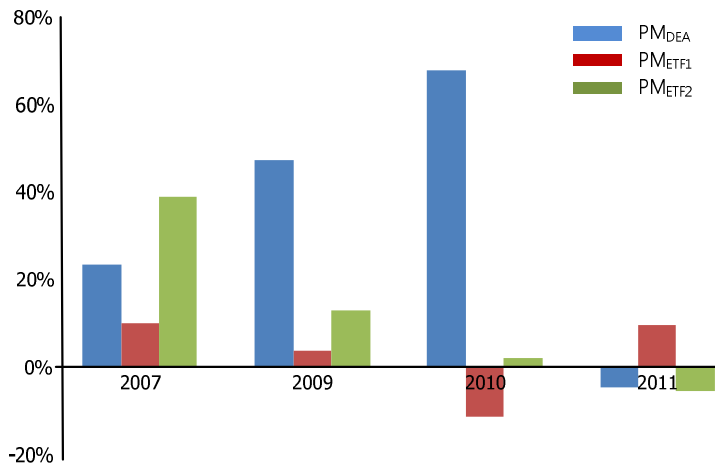
2015년도는 PM_{ETF1} 는 -11.6%를 보이면서 저조한 실적을 보인 반면에 PM_{DEA} 는 약 70%의 누적 수익률을 기록했다. 그러나 2016년도에는 PM_{ETF1} 가 약 10%로 가장 우수하게 나타났고 PM_{DEA} 는 약 -5%로 저조한 수익률을 보였다. 4년 동안의 총 누적 수익률은 PM_{DEA} 가 약 197%로 PM_{ETF1} 와 PM_{ETF2} 에 비해서 매우 우수한 결과를 보인다. 이는 분기 별로 4년 동안 누적된 수익률인데 2013년도부터 2016년도까지 분기 별로 일정한 수익을 기록했기 때문이다.

<그림 2-3>은 이해를 돕기 위해 PM_{DEA} 와 PM_{ETF1} 그리고 PM_{ETF2} 의 년도 별 수익률을 그림으로 보여주고 있다. <그림 2-4>는 실험 기간 동안의 누적 수익률을 나타내고 있는데 2014년도에 PM_{ETF1} 와 PM_{ETF2} 에 비해서 PM_{DEA} 의 누적 수익률이 높아지는 것을 확인 할 수 있다.

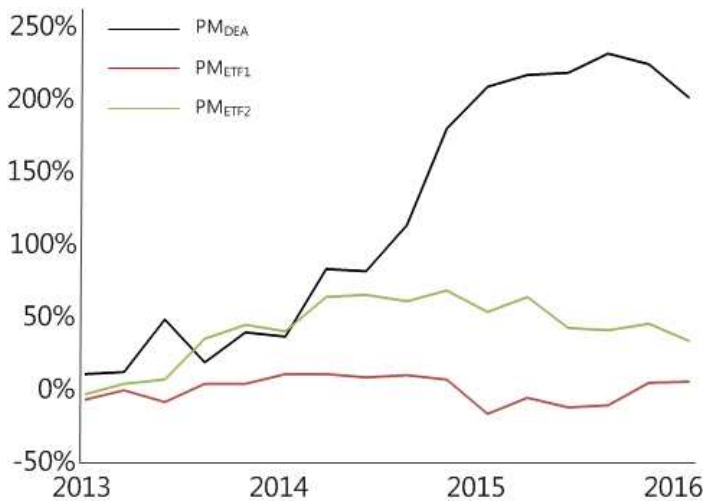
이처럼 PM_{DEA} 와 PM_{ETF1} 그리고 PM_{ETF2} 의 수익률 측면에서 분석해보면 본 연구에서 제안하는 PM_{DEA} 가 실제 시장에서 거래되고 있는 PM_{ETF1} 와 PM_{ETF2} 에 비해서 월등하게 우수하다는 것을 확인할 수 있다. 다음 절에서는 포트폴리오 운용의 성과 측정을 정량적으로 분석해보고자 한다.

<표 2-3> 분기 별 수익률 결과(단위, %)

년도	분기	PM_{DEA}	PM_{ETF1}	PM_{ETF2}	시장 수익
2013	1	16.6	-0.7	3.3	-0.2
	2	1.1	6.5	6.8	-1.3
	3	28.7	-7.4	2.4	-3.7
	4	-18.4	12.6	23.6	5.4
	누적	23.8	10.3	39.7	0.0
2014	1	16.0	0.0	6.5	-3.3
	2	-1.8	5.6	-3.1	2.7
	3	31.4	-0.4	15.8	2.6
	4	-0.8	-1.3	0.8	-4.8
	누적	48.4	3.9	13.2	-3.0
2015	1	16.2	0.7	-2.7	1.4
	2	29.5	-2.4	4.4	4.1
	3	9.9	-19.3	-8.1	-7.9
	4	2.6	11.5	6.2	4.1
	누적	69.6	-11.6	1.9	1.1
2016	1	0.5	-6.3	-12.5	-2.2
	2	4.0	1.4	-0.9	1.0
	3	-2.1	15.0	3.2	3.9
	4	-6.9	0.6	-7.8	-0.9
	누적	-4.8	9.9	-5.8	1.7
4년 총 누적		196.6	11.3	37.8	-0.2



<그림 2-3> PM_{DEA} , PM_{ETF1} , PM_{ETF2} 의 연도별 수익률



<그림 2-4> PM_{DEA} , PM_{ETF1} , PM_{ETF2} 의 누적 수익률

3.3 성과 측정

일반적으로 포트폴리오 성과 측정을 위한 도구로 수익률의 변동성, 샤프지수와 켈센의 알파 그리고 IR이 있다. 샤프지수는 포트폴리오의 위험 1 단위에 대한 초과 수익의 정도를 나타내는 지표, 즉 초과 수익이 얼마인가를 측정하는 지표이고 모형은 식(9)과 같다.

$$Sharpe\ Ratio = \frac{R_i - R_f}{\sigma_i} \quad (9)$$

여기서,

R_i 포트폴리오 i 의 수익률

R_f 무위험 수익률(국고채 3년 만기)

σ_f 포트폴리오 i 의 표준편차

젠센의 알파는 포트폴리오의 수익률이 균형 상태에서의 수익률보다 얼마나 높은지를 나타내는 지표, 즉 포트폴리오 수익률에서 기대 수익률을 뺀 값을 의미하며 모형은 식(10)과 같다.

$$\text{젠센의 알파} = (R_i - R_f) - b_p^*(K_i - R_f) \quad (10)$$

여기서,

b_p 포트폴리오의 베타

K_i 시장 수익률

그리고 IR은 포트폴리오 관리자의 능력을 측정할 수 있는 지표로 포트폴리오의 초과 수익률을 추적 오차로 나눈 값을 말하며 RVR(Reward-to-Variability Ratio)라고도 부른다. 세 가지 모두 측정된 결과 값이 높을수록 투자 성과가 우수하다고 할 수 있으며 IR의 경우에는 실무적으로 미국에서는 50% 이내인 경우에 ‘우수’한 것으로 평가한다. IR의 산출 모형은 식(11)과 같다.

$$IR = \frac{(R_i - K_i)}{Te} \quad (11)$$

여기서,

Te 추적 오차의 표준편차

앞서 설명한 산출 공식을 이용하여 본 연구에서 제안하는 PM_{DEA} 와 실제 시장에서 거래되고 있는 PM_{ETF1} 와 PM_{ETF2} 를 성과 측정한 결과는 <표 2-4>와 같다. 변동성을 제외한 나머지 성과 지표는 측정된 결과 값이 클수록 포트폴리오 운용에 대한 성과가 좋다고 할 수 있는데 PM_{DEA} 가 PM_{ETF1} 와 PM_{ETF2} 에 비해서 운용 성과가 월등하게 좋다는 것을 알 수 있다. 또한 PM_{ETF1} 비해서 PM_{ETF2} 가 켄센의 알파를 제외한 나머지 지표에서 운용 성과가 좋게 나오는 것을 알 수 있다.

변동성은 년도 별 누적 수익률에 대한 변동성인데 PM_{ETF1} 와 PM_{ETF2} 에 비해서 PM_{DEA} 가 높게 나타난다. 이는 우량주로만 편성된 PM_{ETF1} 와 PM_{ETF2} 에 비해서는 PM_{DEA} 가 다소 변동성에 대한 위험에 노출된 것이라 해석할 수 있다.

<표 2-4> 성과 측정(단위, %)

	변동성	샤프지수	켄센의 알파	IR
PM_{DEA}	32.0	93.0	11.9	122.9
PM_{ETF1}	10.2	-13.1	-3.2	44.8
PM_{DEA}	32.0	93.0	11.9	122.9
PM_{ETF2}	19.9	39.2	-3.9	84.0

제 3 장. 섹터 투자 전략

제 1 절 문제의 정의

주식을 투자하는 사람들은 대부분 포털을 통해서 다양한 투자 정보를 수집한다. 그리고 주가와 관련이 깊은 다양한 경제적 용어 및 투자 주체에 대한 연관된 단어들을 포털을 통해서 검색을 한다.

이미 야후(Yahoo)에서 제공하는 트렌드를 통해서 주식 시장의 거래량과 이와 관련된 단어들의 검색량 간의 연관성이 있다는 것과 구글 트렌드 데이터가 자동차 판매량, 실업수당 신청률, 소비자신뢰지수 등 여러 경제적 지표와 연계될 수 있다는 것이 연구를 통해서 입증되었다(Choi and Varian, 2012). 또한 투자자는 시장에 대한 불안감이 증폭되는 기간 동안에는 주식 매입 또는 매도에 대한 의사결정을 위해서 시장에 대한 다양한 정보 검색을 평소보다 더 많이 한다는 것을 구글 트렌드 분석을 통해서 밝혀진바 있다(Preis T, Moat H S and Stanley H E, 2013). 이러한 현상은 투자자의 심리와 밀접한 관련이 있다. 실제로 일반 개인 투자자들은 정량적인 트레이딩 시스템(trading system)과는 거리가 멀고 대부분 포털 검색을 통해서 정보를 수집한 후 정성적인 판단을 통해서 매매하는 경향이 있기 때문이다.

이처럼 투자자의 투자 심리는 주식 시장과의 밀접한 연관이 있는데 이는 투자자의 심리적 요인을 분석하여 투자 전략을 수립하는 방식과 함께 빅데이터 등의 트렌드를 정량적으로 분석하여 투자 전략을 수립

하는 배경이 되고 있다. 연구를 통해서도 미국의 주식 시장에서 개인 투자자는 집단적 거래행태가 존재하고 개인 투자자의 투자 심리와 주식 수익률 간의 유의한 관계가 있음을 보였다(Kumar and Lee, 2006).

그 외 호주와 독일의 주식 시장에서도 브로커를 통한 개인 투자자들의 거래를 각각 분석함으로써 개인 투자자들 간의 거래에 있어서 유의한 상관관계가 있다는 것이 입증되었다(Jackson A, 2003; Dorn D, Huberman G and Sengmueller P, 2008). 이와 같이 투자 심리에 대한 연구는 빅데이터나 매체를 통해서 투자자의 심리를 분석하는 연구에 비해서 많이 연구되어 왔다. 그리고 주식 투자자들이 의견을 공유할 수 있는 포털과 SNS(Social Networking Service) 등이 대중화되면서 주식 투자를 하는 투자자들의 약 93%가 포털을 통해서 투자 주체 및 그와 연관된 단어 검색을 하고 있으며, 최근 빅데이터의 관심도가 증가하면서 이와 관련된 연구도 함께 진행되고 있다.

Barber B M, Odean T and Zhu N(2009)은 감정이 개개인의 행동과 의사결정에 큰 영향을 미친다는 행동경제학(behavioral economics) 이론을 바탕으로 대규모 트위터 피드(twitter feeds)로부터 추출한 집단적 감정 상태의 변화가 일정한 기간 동안에 다우존스산업지수(DJIA)의 가치 변화와 상관관계가 있음을 밝혔다. 또한 2004년부터 2010년까지 구글 검색 엔진을 통해서 질의된 다양한 검색어들의 주별(weekly) 검색량과 주식 시장의 변동성간에 상관관계가 있다는 점과 특히 S&P500에 포함된 개별 종목들의 검색량이 개별주식의 거래량과도 유의한 상관관계가 있다는 것이 연구결과 입증되었다(Preis T, Reith D

and Stanley H E, 2010).

Bordino et al(2012)은 2010년 5월부터 2011년 4월까지 야후의 검색 엔진을 통해서 질의된 검색어의 검색량과 NADAQ100에 상장된 종목들의 거래량 간의 상관관계를 보였고, 이들 두 시계열 간의 인과관계를 분석하였다. 김유신(2012)은 오피니언 마이닝을 통해서 지능형 투자자의사결정모형을 제시하였고 이를 통해서 주가지수 변동성을 예측하는 가능성을 제시하였다.

박원준(2012)은 통계나 실험을 통해서 얻은 정형화된 데이터뿐만 아니라 인간의 정서나 심리 정보에 해당하는 기분이나 감정이 내재된 비정형화된 데이터를 분석함으로써 소비자 중심의 정보를 산출할 수 있으며, 기업은 물론 공공 영역에서도 광범위하게 사용될 수 있다고 판단하였다. 이득환(2014)은 빅데이터에 나타난 투자자별 감성이나 정보가 KOSPI200 선물 지수 수익률에 미치는 영향을 실증 분석하여 빅데이터가 KOSPI200 선물 지수 수익률을 예측하는 정보를 포함한다는 것과 빅데이터를 사용한 투자 전략이 높은 수익을 가져옴을 증명하였다.

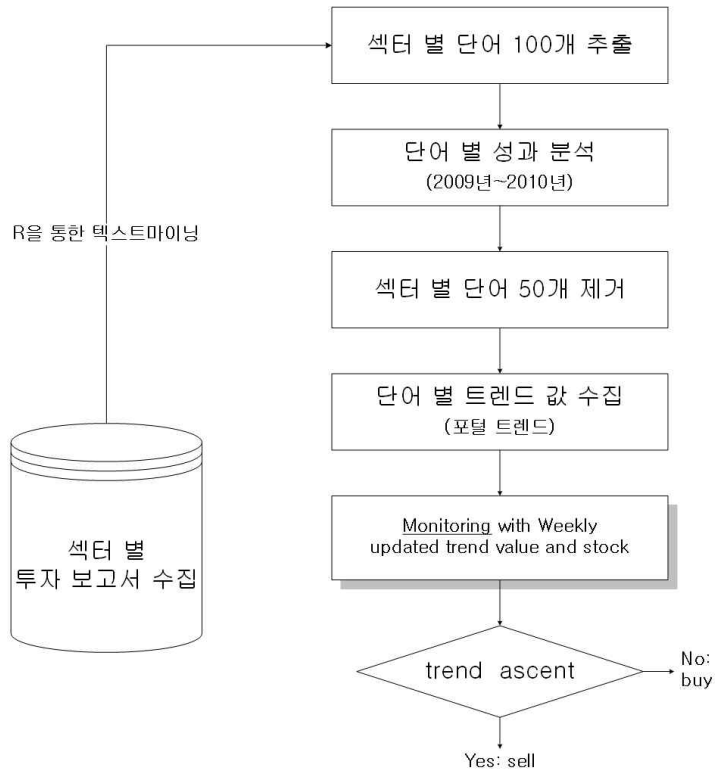
앞서 설명한 기존의 연구를 살펴보면 우리나라 인터넷 포털 검색엔진에서 형성되는 섹터별 빅데이터 트렌드를 국내 주식시장 섹터투자 전략에 적용한 연구는 찾아보기 어렵다. 따라서 본 연구는 국내 대표적인 검색엔진을 통해 형성되는 사용자의 경제관심의 변화 트렌드를 섹터별 ETF에 반영하는 투자전략을 제안하고 그 성과를 분석하고자

한다.

일반적으로 시장에 대한 불안감이 높아질수록 시장에 대한 관심도는 증가한다(Preis T, Moat H S and Stanley H E, 2013). 예컨대 투자자가 투자한 자산에 대한 불안감이 높아지면 포털에 투자 자산과 관련된 단어를 검색하거나 SNS를 통해서 본인의 불안한 심리를 표현한다. 또한 증권관련 포털을 보면 종목 상담의 대부분은 현재의 주가가 매입가보다 높을 경우에 비해서 낮을 경우가 월등하게 많다.

신현준(2015)은 국내 빅데이터 트렌드를 이용하여 KOSPI 주가지수 투자 전략을 수립하고 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 것을 입증하였다.

따라서 본 연구에서는 주식시장을 세분화하여 분할된 섹터 내에서 수집된 트렌드를 해당 섹터에 반영하여 매매하는 일종의 섹터투자 전략을 제시하고자 한다. 이를 위해서 주식시장을 11개의 섹터로 구분하고 각 섹터별 35~50개의 트렌드 검색어들을 텍스트마이닝과 브레인스토밍 기법을 이용하여 추출하여 섹터투자 전략 포트폴리오를 구성한다. 매매는 11개 섹터별로 거래되고 있는 ETF 상품들을 대상으로 한다. 4년간(2011~2014)의 섹터별 투자 성과를 누적수익률 관점에서 비교 분석함으로써 유의미한 결과를 도출하고자 한다. <그림 3-1>은 본 연구에서 제안하는 섹터투자 전략에 대한 과정을 보여주는 그림이다.



<그림 3-1> 텍스트마이닝을 이용한 ETF 투자 전략

제 2 절 연구의 자료

2.1 섹터 선정

일반적으로 개별 종목에 대한 포트폴리오를 구성하는 것은 너무 많은 수의 종목들이 투자 대상에 포함되기 때문에 종목을 선정하는 것이 힘들며, 몇 몇의 투자자로 인해서 명확한 근거가 없이 주가가 급등 또는 급락하는 종목들이 있기 때문에 효과적인 포트폴리오를 구성하는 전략을 수립하는 것이 매우 어렵다. 반면 상장지수펀드인 ETF는 KOSPI200 또는 KOSPI50과 같이 다소 우량한 기업들을 구성하고 인덱스 펀드와는 달리 거래소에 상장되어 일반 주식처럼 자유롭게 사고 팔 수 있기 때문에 본 연구와 같이 새로운 투자 전략을 제안하고자 실험하는 연구에 있어서 적합하다고 사료된다.

따라서 본 연구는 한국 거래소에 상장된 총 11개의 섹터인 TIGER 은행, KODEX 에너지화학, KODEX 운송, KODEX 조선, KODEX 반도체, KODEX 철강, TIGER IT, KODEX 소비재, KODEX 건설, KODEX 자동차, TIGER 미디어 통신을 선정한다. 또한 포털에서 제공하는 트렌드의 변화량 자료가 주 별로 제공된다는 점을 감안하여 각 섹터의 주가를 증권사에서 제공하는 HTS(Home Trading System)를 통해서 본 연구의 실험 기간인 2009년부터 2014년까지 주 별로 수집한다.

ETF는 일반적으로 KOSPI 지수와 양의 상관관계를 보이는데 본 연

구의 투자 대상인 11개의 섹터와 KOSPI 지수와의 상관계수는 <표 3-1>과 같다. KOSPI 지수와 섹터간의 상관계수를 보면 TIGER 은행이 KOSPI 지수와 가장 높은 상관관계를 보인다. 이는 은행들의 주가는 큰 변동성이 없이 KOSPI 지수와 유사한 방향으로 움직이기 때문이다. 다음으로 반도체가 0.60으로 높는데 KODEX 반도체에 KOSPI 지수의 움직임에 높은 영향을 주는 삼성전자가 구성되어 있기 때문이다. 다음 절에서는 각 섹터 별로 키워드를 선정하는 과정과 키워드 선정을 위해 사용하는 텍스트마이닝에 대해서 설명하고자 한다.

<표 3-1> KOSPI와 섹터간의 상관계수

섹터종류	IT	건설	미디어	반도체	소비재	에너지	운송	은행	자동차	조선	철강
상관계수	0.50	0.20	0.22	0.60	0.32	0.40	0.42	0.77	0.37	0.44	0.43

2.2 텍스트마이닝

텍스트마이닝은 비정형화된 대규모 텍스트 자료로부터 일반화된 패턴 및 관계를 발견하고 추출하는 과정이다(강병욱, 2015). 텍스트마이닝은 크게 데이터를 수집하는 과정, 용어를 추출하는 과정, 정보를 추출하는 과정 그리고 정보를 분석하는 과정의 총 4단계의 과정을 거친다.

먼저 텍스트마이닝의 첫 번째 단계인 데이터를 수집하는 과정은 비정형 대규모 텍스트 자료를 수집하는 단계이다. 두 번째 용어를 추출하는 과정은 구문의 패턴이나 단어들의 연관성을 고려하여 연관성 분

석에 의해 단어들을 추출하여 분석 대상 용어들의 후보를 선정하고, 선정된 단어에 대해서 여러 가지 통계적인 방법을 통해서 전체를 대표하는 단어들을 추출하는 과정이다. 즉 첫 번째 단계에서 얻어진 분석대상 데이터를 중심으로 정보를 추출할 수 있도록 데이터를 가공하는 단계로 수집된 문서를 기본으로 관련 키워드를 추출한다. 세 번째 정보를 추출하는 과정은 문서 자체를 찾는 것이 아니라 문서 내에서 유용한 정보를 찾는 과정이다. 예컨대 제품을 소개하는 문서의 경우에 제품명, 제품의 기능 그리고 주의 사항 등과 같이 상세 정보를 포함하는 용어를 추출하기 위한 과정이다. 네 번째 정보를 분석하는 과정은 세 번째 과정에서 얻어진 최종 단어에 대해서 빈도, 분류, 군집화(Clustering) 그리고 컨셉 도출 방법 등을 이용하여 유용한 정보를 도출해 내는 과정이다. 빈도는 가장 많이 얻어지는 단어가 무엇인지에 대한 정보를 도출해 내고, 분류는 앞서 추출된 단어의 내용에 따라 문서들을 범주화 시켜주는 과정이다. 즉 주어진 신문 텍스트 문서가 금융 분야인지 또는 정치 분야인지 등을 단어에 따라 분류하는 것을 의미한다. 군집화는 문서에 포함되어 있는 추출된 단어들을 유사도에 따라 여러 개의 텍스트 집단으로 군집화 시켜주는 과정이다. 컨셉 도출은 어떤 특정한 키워드를 중심으로 또 다른 키워드들 간의 관계를 파악하는 기법이다.

본 연구에서는 섹터별로 25개의 리포트를 선정하는데 선정하는 기준은 2014년도에 매출이 가장 높았던 3개의 증권사에서 발행하는 리포트를 대상으로 한다. 또한 3개의 증권사에서 발행하는 리포트가 방대하기 때문에 해당 섹터와 동일한 주제의 리포트 중에서 조회 수가 가

장 많은 순으로 리포트를 선정한다. 선정된 리포트를 통해서 섹터 별로 해당 키워드를 선정해야 하는데 이는 다양한 분야에서 사용되는 프로그램 R을 이용한다. 텍스트마이닝을 구현하는 과정은 tm 패키지를 이용하여 문서별 Corpus 생성, 명사 추출, 불 용어 및 기타 의미 없는 기호를 제거, 두 자리 이상의 명사 추출 그리고 KoNLP 패키지를 사용하여 Corpus 내의 한글 형태소 단위를 인식하는 절차로 구성된다. 그리고 명사들의 빈도를 이용하여 Matrix 구조의 유사도를 판단 및 구현하고 유사한 문서의 그룹화를 거쳐서 키워드와 복합 명사들을 추출한다(이종화, 2015).

<그림 3-2>는 KODEX 에너지화학에 해당하는 키워드를 R을 통해서 추출한 결과를 사례로 보여주는 그림이며, 해당 섹터에 대한 대표성이 강한 단어가 더 크게 보인다. 25개의 리포트에서 R을 통해서 100개의 단어를 추출하고 2009년부터 2010년까지 각 단어의 트렌드 변화량을 바탕으로 섹터 투자를 수행한 후, 이 중에서 성과가 높았던 50개를 최종적으로 해당 섹터의 연관 단어로 선정한다. <표 3-2>은 KODEX 에너지화학에 해당하는 최종 키워드들이다.



<그림 3-2> R을 이용한 키워드 선정

<표 3-2> 선정된 에너지화학 섹터 키워드

최종적인 에너지화학 섹터 키워드				
S-OIL	두바이유	신재생에너지	연료 전지	한국전금
LG화학	SK가스	프로필렌	합성수지	부타디엔
Opex	bbl	LNG	배럴	브랜드유
한샘	현대리마트	라이온캠텍	KCC	호남석유화학
국제유가	SK이노베이션	한화케미칼	동남아	친환경
천연가스	원자력	휘발유	석유	폴리우레탄
톨루엔	HDPE	MEG	PVC	BPA
스프레드	부동산 가격	플라스틱	나프타	에틸렌
원유	정유	석유화학	유가	석탄
ABS	LPG	BENZENE	디젤	태양전지
제거된 에너지화학 섹터 키워드				
경기	코스닥	테마주	삼성	예금
조세정책	코스피	시황	헤지펀드	스톡옵션
간접세	기업어음	급등주	출구전략	중국
직접세	콜금리	ELW	경제	국민은행
환헤지	우선주	금융	모기지론	신용등급
나스닥	주식	유가	공매도	정책
지급준비율	채테크	거래소	경제성장률	부양
GDP	증권	증시	인플레이션	경매
GNP	부동산	ELS	물가	서울은행
환율	펀드	탄소세	세금	증장비

제 3 절 섹터 투자 전략

앞서 설명한 섹터 선정과 해당 섹터의 키워드를 텍스트마이닝을 통해서 선정하면, 다음으로는 투자 전략을 수립해야 한다. 텍스트마이닝을 통해서 선정된 키워드의 트렌드는 네이버 트렌드를 통해서 자료를 수집하는데 이는 주 별 자료를 수집한다. 따라서 전 주 대비하여 해당 키워드의 검색량이 증가하면 해당 섹터의 ETF는 매도하는 전략을 취한다. 예컨대 KODEX 에너지화학의 키워드로 선정된 석유화학 키워드가 전 주 대비하여 검색량이 떨어지면 KODEX 에너지화학을 매수한다.

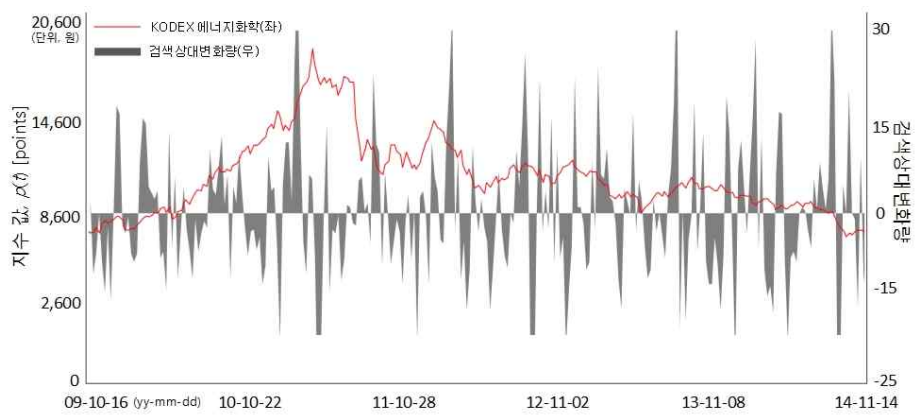
분석의 단순화를 위해 투자 대상인 ETF는 공매도(short selling)가 가능하다고 가정한다. 또한 키워드의 변화량을 전 주 대비가 아닌 과거 2주, 3주 등의 이동평균 값의 변화량을 바탕으로 매매 실험을 진행한다. 예컨대 $t-1$ 주에 MMF 용어가 네이버에서 검색된 검색량, $n(t-1)$ 을 네이버 트렌드를 통해 산출할 수 있다. 시장 참여자들의 정보검색 움직임을 정량화하기 위해 검색 상대변화량(relative search volume change) $\Delta n(t, \Delta t) = n(t) - N(t-1, \Delta t)$ 을 이용한다. 여기서 $N(t-1, \Delta t)$ 는 $\{n(t-1) + n(t-2) + \dots + n(t-\Delta t)\} / \Delta t$ 로 정의되며 $t-1$ 주부터 최근 Δt 주 동안의 검색량의 이동평균을 뜻한다(Preis T, Moat H S and Stanley H E, 2013).

<그림 3-3>은 KODEX 에너지화학의 주별 종가와 검색용어 석유화학에 대한 검색 상대변화량의 시간에 따른 변화를 보여주며, 각각은 KODEX 에너지화학 t 주 첫 거래일의 종가, 즉 $p(t)$ 의 시계열과 검색용

어 석유화학의 $\Delta t = 3$ 주로 산출한 검색 상대변화량을 뜻한다.

실험하는 과정에서 매매 전략은 신현준(2015)의 연구에서 제안한 NT-OS전략을 바탕으로 실험한다. 일대일청산(one-to-one settlement; 이하 OS)을 의미하는 NT-OS전략은 매매신호 발생 시에 취한 포지션을 바로 청산하지 않고 보유해 나가는 방식으로써 청산은 기존의 유지하고 있는 포지션과 반대의 매매신호가 발생하면 보유 중인 가장 오래된 반대 포지션과 일대일로 시행한다. 예컨대 만약 $t-1$ 주에 $\Delta n(t-1, \Delta t) > 0$ 이라면 t 주의 첫 거래일 종가인 $p(t)$ 에 매도하고 만약, $t+1$ 주에 $\Delta n(t+1, \Delta t) > 0$ 이라면 $t+2$ 주의 첫 거래일 종가인 $p(t+2)$ 에 다시 한 번 매도포지션을 취한다. 그리고 만약 $t+2$ 주에 $\Delta n(t+2, \Delta t) < 0$ 이라면 $t+3$ 주의 첫 거래일 종가인 $p(t+3)$ 로 매수함으로써 기존에 누적된 매도 포지션들 중 t 주의 매도포지션과 일대일로 청산된다.

이 경우 누적수익률은 $\log p(t) - \log p(t+3)$ 이다. NT-OS전략은 일종의 레버리지 전략으로써 동일한 방향의 매매 신호가 발생하면 동일한 포지션을 누적해가기 때문에 방향성 예측이 정확하다면 큰 수익을 얻을 수 있지만 반대의 경우 큰 손실의 위험도 존재한다.



<그림 3-3> 검색상대변화량과 ETF 가격과의 관계

제 4 절 실험 결과 및 분석

4.1 실험 계획

본 연구에서 제안하는 섹터 투자 전략의 성능을 분석하기 위해서 <표 3-3>의 실험 계획을 수립한다. 앞서 설명했듯이 투자 대상은 총 11개의 ETF이며, 매매 전략은 앞 장에서 설명한 NT- OS전략을 이용한다. 매매 시 발행하는 거래 수수료는 0.004%로 정의하며 본 연구의 실험 기간은 2011년부터 2014년까지 설정한다.

또한 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매한 결과와 비교하기 위해서 추가적으로 50개의 키워드를 선정하고 이를 이용하여 매매한 결과와 비교한다. 이는 금융업에 종사하는 5명을 통해서 각 섹터 별로 의미 있는 키워드를 선정하였다. 본 방법론은 편의상 브레인스토밍(Brain Storming; 이하 BS)기법이라고 정의한다. 더불어 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매(Trading)하는 방식을 TM-T으로 BS 기법을 통해서 선정된 키워드를 바탕으로 매매하는 방식을 BS-T라고 정의한다.

<표 3-3> 실험 계획	
실험 요인	값
투자 대상	11개의 ETF
키워드 선정 방법론	텍스트마이닝
거래 수수료	0.004%
비교대상	BS-T 기법
실험 기간	4 년(2011.01.02~2014.12.26)

4.2 실험결과 및 분석

<표 3-4>는 TM-T와 BS-T를 바탕으로 섹터 별 매매한 연도별 로그수익률과 누적 로그수익률에 대한 결과를 정리한 표이다. 표에서 $\Delta 1$, $\Delta 2$, $\Delta 3$ 은 이동평균기간(Δt , 주)을 의미한다. 대체적으로 텍스트마이닝을 통해서 추출한 단어를 바탕으로 매매한 TM-T가 수익률 측면에서 BS-T를 바탕으로 매매한 수익률에 비해서 우수하다. 또한 소비재와 철강의 섹터를 제외하고는 이동평균선기간이 클수록 누적 로그수익률이 높는데, 이는 일반적인 트렌드가 특정한 모멘텀을 발생하기 때문에 단기적인 트렌드 변화에 대응하여 매매하는 것 보다는 일정하게 유지되는 트렌드를 바탕으로 매매하는 것이 더 높은 수익률을 보인다는 것을 알 수 있다.

더불어 단기적이고 민감한 트렌드 변화로 인한 빈번한 매매 타이밍은 매매 비용이 많이 발생한다. 그리고 동일한 섹터와 동일한 년도의 경우에는 $\Delta 1 \sim \Delta 3$ 간에 로그수익률에 대한 편차가 낮다. 예컨대 KODEX 건설의 경우, 2011년을 보면 Δt 별로 TM-T의 로그수익률이 각각 약 -0.10, -0.12, -0.13이고, BS-T의 경우에는 Δt 별로 각각 0.01, 0.44, 0.41이다. 반면 2012년도에는 TM-T의 경우에 Δt 별 로그수익률이 모두 0.40 대 이다. 이처럼 동일한 섹터와 동일한 년도의 경우에 Δt 별로 로그수익률 간의 편차가 작은 것을 알 수 있는데, 이는 주 별로 수집한 단어 별 포털 트렌드 값의 경우에 Δt 간의 편차가 크지 않기 때문이다.

앞서 <표 3-1>의 KOSPI 지수와 섹터 별 상관관계와 섹터 별 수익률 간의 유의한 상관관계는 없는 것으로 사료된다. 많은 경우의 수를 실험함으로써 각각에 대한 개별적 성과를 로그수익률을 통해서 확인할 수 있지만 종합적인 매매 성과를 정량적인 평가 척도로 이해하기는 쉽지 않다(신현준, 2013). 따라서 다음 절에서 일반적으로 포트폴리오의 운용 성과를 측정하는 샤프지수와 쟈센의 알파 그리고 IR을 통해서 $TM-T$ 를 바탕으로 Δt 별로 매매한 경우에 대해서 성과 평가를 진행한다.

<표 3-4> 11개 섹터의 투자 수익률

		2011			2012			2013			2014			누적수익률		
		Δ1	Δ2	Δ3	Δ1	Δ2	Δ3	Δ1	Δ2	Δ3	Δ1	Δ2	Δ3	Δ1	Δ2	Δ3
건설	TM-T	-0.10	-0.12	-0.13	0.40	0.48	0.46	0.68	0.77	0.82	1.26	1.28	1.41	3.81	4.24	4.59
	BS-T	0.01	0.44	0.41	-0.07	0.30	0.54	0.09	0.22	0.27	0.22	0.29	0.44	0.23	1.97	2.98
소비재	TM-T	-0.07	-0.27	-0.43	0.09	0.14	0.07	0.22	0.30	0.36	0.88	0.96	0.84	1.31	1.09	0.52
	BS-T	-0.10	-0.30	-0.40	0.07	0.13	0.11	0.19	0.18	0.21	0.75	0.92	1.05	0.99	0.80	0.67
에너지	TM-T	0.32	0.34	0.23	0.31	0.38	0.41	0.58	0.61	0.76	0.21	0.33	0.33	2.30	2.97	3.04
	BS-T	0.17	0.10	0.07	0.51	0.75	0.51	0.47	0.58	0.18	0.77	0.67	0.81	3.62	4.08	2.46
자동차	TM-T	0.55	0.71	0.48	0.19	0.28	0.60	-0.17	0.08	0.09	0.20	-0.07	-0.07	0.84	1.21	1.42
	BS-T	-0.04	-0.28	-0.61	0.18	0.20	0.27	-0.12	0.15	0.15	0.14	-0.17	-0.14	0.12	-0.18	-0.51
조선	TM-T	0.24	0.38	0.18	0.29	0.27	0.28	-0.04	0.55	0.07	0.24	0.11	0.68	0.90	2.01	1.72
	BS-T	-0.92	-0.39	0.39	-0.88	-1.81	-2.84	0.13	-0.31	-0.62	-2.85	-2.98	-3.27	-1.02	-0.32	1.23
철강	TM-T	0.11	0.26	0.17	-0.04	0.12	-0.08	0.33	0.02	0.13	0.26	0.12	0.03	0.78	0.61	0.25
	BS-T	-0.17	-0.19	-0.59	-0.04	-0.18	0.24	0.08	-0.38	-0.31	0.00	0.19	0.11	-0.13	-0.51	-0.61
반도체	TM-T	-0.33	0.12	0.09	0.15	0.02	-0.19	0.34	0.46	0.58	0.40	0.56	0.66	0.45	1.60	1.29
	BS-T	-0.97	-1.22	-1.08	0.06	0.27	0.19	1.20	1.22	1.24	0.77	0.78	0.85	-0.88	-2.10	-1.38
IT	TM-T	0.06	0.09	0.24	0.13	0.15	0.22	0.44	0.47	0.60	0.73	0.65	0.74	1.98	2.03	3.22
	BS-T	0.31	0.32	0.41	-0.38	-0.54	-0.60	-0.28	-0.35	-0.49	-0.28	-0.40	-0.57	-0.58	-0.76	-0.88
미디어	TM-T	0.24	0.45	0.61	0.02	0.06	0.14	0.55	0.12	0.06	0.08	0.32	0.74	1.11	1.27	2.36
	BS-T	-0.33	-0.41	-0.45	-0.13	-0.42	-0.36	-1.16	-1.03	-0.89	0.05	-0.05	-0.31	-1.10	-1.01	-0.97
운송	TM-T	0.05	0.11	0.16	0.26	0.27	0.50	0.48	0.43	0.81	0.38	0.42	0.65	1.70	1.83	4.16
	BS-T	-0.03	-0.03	0.33	-0.12	-0.02	0.08	0.04	0.14	0.23	-0.57	-0.11	0.25	-0.62	-0.04	1.21
은행	TM-T	0.12	0.16	0.23	0.33	0.69	0.91	0.35	0.35	0.10	0.46	0.84	0.88	1.92	3.83	3.85
	BS-T	0.28	0.38	0.42	0.32	0.53	0.56	-0.47	-0.14	-0.02	-0.11	-0.23	-0.18	-0.21	0.41	0.78

4.3 성과 측정

일반적으로 포트폴리오 성과 측정을 위한 방법론인 샤프지수와 켄센의 알파 그리고 IR의 이론적 설명은 2장에서 설명하고 있기 때문에 본 장에서는 생략하고자 한다.

<표 3-5>는 TM-T를 바탕으로 매매한 결과에 대한 성과를 측정한 결과를 Δt 별로 보여준다. 특히 수익률이 높다고 성과 평가의 결과가 반드시 좋다는 것은 아니다. 예컨대 KODEX 건설의 경우에 $\Delta 3$ 의 경우가 가장 높은 수익률을 보이지만 샤프지수는 $\Delta 1$, 켄센의 알파는 $\Delta 2$ 그리고 IR은 $\Delta 1$ 이 높게 나타난다. 이는 수익률을 통한 평가와는 다르게 성과 평가는 연도별 수익률의 표준편차와 벤치마크 대비 초과수익 등을 고려하기 때문에 반드시 누적수익률이 높다고 성과 평가 결과가 우수한 것은 아니다. 즉 우수한 운용 능력은 수익률과 함께 운용 위험(Risk)을 항상 고려해야 한다. 여기서 벤치마크는 BS-T로 설정하였다.

샤프지수 중에서는 약 84%의 누적수익률을 기록한 KODEX 자동차가 4.51($\Delta 1$)로 가장 우수했고, 켄센의 알파는 약 400%의 누적수익률을 보인 KODEX 건설이 1.59($\Delta 2$)로 가장 운용 성과가 높았다. 마지막으로 IR은 약 230%의 누적수익률을 보인 TIGER 미디어가 1.61($\Delta 3$)로 가장 높았다.

<표 3-5> TM-T의 성과 측정

	샤프 지수			젠센의 알파			IR		
	$\Delta 1$	$\Delta 2$	$\Delta 3$	$\Delta 1$	$\Delta 2$	$\Delta 3$	$\Delta 1$	$\Delta 2$	$\Delta 3$
건설	2.70	1.84	1.49	0.04	1.59	0.84	1.08	0.69	0.51
소비재	1.13	1.09	0.76	0.08	0.10	0.04	0.14	0.09	-0.07
에너지	0.69	0.75	1.04	0.40	0.36	0.45	-0.63	-0.50	0.14
자동차	4.51	-8.63	-3.11	0.16	0.21	0.25	0.67	0.90	0.94
조선	-0.14	-0.22	-0.18	0.10	0.45	0.14	1.21	1.37	1.23
철강	-4.65	-0.76	-0.29	0.18	0.12	0.01	1.19	1.22	0.71
반도체	0.44	1.02	0.86	0.06	0.24	0.20	-0.19	0.04	-0.03
IT	-1.98	-1.31	-1.37	0.25	0.25	0.36	1.27	1.33	1.40
미디어	-0.51	-0.45	-0.72	0.07	0.30	0.69	1.22	1.49	1.61
운송	-1.57	-5.74	2.27	0.25	0.29	0.64	1.41	1.51	1.20
은행	2.91	3.53	2.62	0.30	0.50	0.49	1.01	0.98	0.84

제 5 절 응용 사례

5.1 사례 적용 배경

금융 시장에 대한 위험 변수(risk variable)가 다양화되면서 투자자들에게 금융 상품의 가치를 예측하는 관심은 높아지고 있다. 최근 위험을 최소화하는 상장지수펀드와 같은 패시브펀드(passive fund)의 관심도가 높아지고 있지만 2017년 4월 기준으로 국내 ETF 순자산은 약 17조이며, 주식형 공모형 펀드는 약 40조의 순자산을 기록하고 있다. 그만큼 아직까지는 고수익과 고위험을 동반하는 주식형 펀드의 투자 비중이 월등하게 높다고 볼 수 있다. 또한 최근 KOSPI 지수가 상승하면서 2017년 5월 KOSPI와 KOSDAQ 시장의 하루 평균 거래대금은 약 10조원을 기록했다.

주식 시장의 투자 비중이 타 금융 상품에 비해서 월등하게 높아지고 금융 시장에 대한 변동성과 위험이 증가하면서 주가에 대한 예측은 다양한 분야에서 관심의 대상이 되고 있으며, 주가 예측이 가능할 경우에 시장 수익률 대비 어느 정도의 초과 수익이 가능한지에 대한 연구가 많이 진행되고 있다. 특히 최근에 인공 지능(artificial intelligence)의 분야의 발달로 인해서 기계 학습(machine learning) 기법을 이용한 주가의 방향을 예측하는 연구가 나오고 있으며, 이는 상당한 수준의 신뢰도를 보이고 있다(Saeed & Ying, 2014, Park et al., 2016).

더불어 인공 신경망(artificial neural network)을 이용하여 주가 자료

를 클러스터링하는 것은 주가를 예측하는 과정에서 매우 중요한 과정이다. 실제로 주가의 패턴을 분석하여 주가를 예측하는 실효성은 과거에 입증된 바 있다(Jung & Yoon, 1998). 기술적 분석(technical analysis)을 이용하여 주가를 예측하는 연구들은 일반적으로 주가의 변동성의 추세를 효율적으로 표현할 수 있는 이동 평균(moving average)에 기초한 입력 특성(input feature) 기법을 사용하고 있다(Lee, 2013). 예를 들어 MACD(moving average convergence divergence)의 경우는 이동 평균 자료들의 발산과 수렴을 표현하기 위해서 장기 이동 평균과 단기 이동 평균의 차이를 이용한다. 그러나 이러한 기술적 자료들을 이용하면 대부분 주가를 선행하지 못한 예측 결과를 보인다는 한계점을 내포하고 있다.

주가 자료를 이용하여 패턴을 분석한 국내 연구를 살펴보면, Park(2007)은 시계열 데이터인 주식 데이터를 대상으로 신경망을 적용하여 주가 예측의 정확도를 향상시키는 방법에 대하여 연구하였고, Min & Jeong(2007)은 유전알고리즘을 통해서 기업들의 집단분류를 통해서 특정 기업들의 주가의 움직임을 집단화 한 연구를 보고하였다. 그리고 Park & Shin(2011)는 시계열 네트워크를 통해서 패턴을 분석하고 이를 이용하여 주가를 예측할 수 있다는 주장을 하였다. 더불어 해외 연구 자료를 살펴보면 Lee and Jo(1999)는 candle 차트를 이용하여 기업들의 주가의 움직임의 공통점을 찾아내어 분류를 통한 패턴 분석을 하였다.

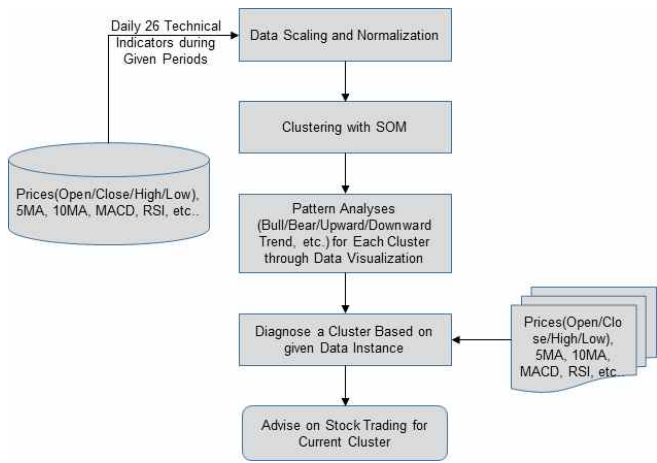
Lo et al.(2000)은 비모수 kernel 회귀분석을 활용하여 주식의 변동성

을 도식화하여 패턴을 발견했고, Fu et al.(2007)은 과거 주가 시계열 자료를 이용하여 패턴을 선별하고 각 패턴들의 정의를 하였다. 또한 Liu and Kwong(2007)은 Radial basis function net work 기법을 이용해서 기술적 자료를 이용하여 계량적인 패턴을 분석하였다. 이처럼 다양한 국내·외 연구들은 주가의 패턴 분석의 발전에 기여하고 있다.

반면 미래 주가를 예측하는 연구와는 다르게 효율적 시장 가설(efficient market hypothesis)에 근거하여 미래 주가를 예측하는 것은 어렵다는 주장을 하고 있는 연구도 다양하다. 효율적 시장 가설에 의하면 주식 시장이 가장 완벽한 중재자이며 주가 움직임의 랜덤워크로 인해서 과거의 주가 자료를 이용하여 미래의 주가를 예측하는 것은 불가능하다고 주장한다(Lippens, 1987). 또한 주가의 움직임은 객관적인 시장 데이터뿐만 아니라 정치적인 요인 등으로 인해서 주가 움직임을 예측하는 것은 한계가 있다는 연구도 있었다(Cootner, 1964). 특히 패턴 분석의 이론적인 반론과 함께 다양한 연구에서 전문 인력과 정보력을 갖고 있는 대형 기관 투자자들과는 다르게 개인 투자자들은 기업의 내부 정보 등과 같이 고급 정보에 습득 능력이 낮기 때문에 개인 투자자들은 객관적인 기술적 자료를 이용하여 투자할 수밖에 없는 한계점을 내포하고 있다.

본 연구에서는 개인 투자자들도 쉽게 얻을 수 있는 기술적 자료를 이용하여 자기조직화지도(Self Organizing Map; 이하 SOM) 통한 주가 패턴 분석 기법을 제안하고자 한다. 이는 과거 자료를 기반으로 유사한 패턴끼리 분류를 하고 daily로 어느 패턴을 내포하고 있는지 찾

아낼 수 있으며, 본 연구를 통해서 향후 패턴들이 내포하고 있는 의미와 미래 움직임의 예측 등에 관한 연구를 진행하는데 도움이 되기를 기대한다. 또한 본 연구는 각 패턴들을 방사형 차트를 이용하여 쉽게 분별할 수 있는 방법론을 제안하고 일반적으로 패턴 분석에 사용되는 이동평균 자료뿐만 아니라 약 26개의 기술적 자료를 이용하여 정확도 있는 패턴을 분류하는 방법론을 제안한다(<그림 3-4> 참고).



<그림 3-4> 패턴 분석 과정

5.2 클러스터링을 이용한 패턴 분석

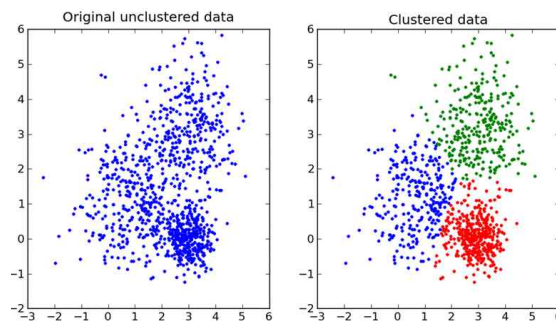
본 연구는 상장된 기업의 기술적 자료를 바탕으로 자기조직화 맵을 통해 클러스터링을 하고 클러스터링의 결과를 방사형 차트를 이용하여 검증해보고자 한다. 또한 그 결과를 바탕으로 급등 또는 급락 이전의 패턴을 알아보고자 한다. 특히 본 연구에서 사용되는 26개의 기술적 자료는 장중 실시간으로 변화하는 빅데이터로써, 일별로는 1년에

대략 200개 이상이 축적된다.

이러한 빅데이터는 산출물을 이해하기 쉽게 시각화하는 것이 어렵고 해석에 있어서도 난해한 부분이 있다. 따라서 본 연구에서 제안하는 시각화 기법에 대해서 설명하기에 앞서서 본 장에서는 클러스터링의 이론적 배경과 함께 본 연구에서 제안하는 SOM 기법에 대해서 기술한다.

5.2.1 클러스터링의 이론적 배경

빅데이터 마이닝(big data mining) 기법 중에서 하나인 클러스터링은 데이터들 간의 유사성(similarity)에 의해서 비슷한 개체들을 집단화하는 통계적인 방법으로 비지도학습(unsupervised learning)에 속한다. 특히 패턴인식, 인공지능, 기계학습 그리고 마케팅 등 다양한 연구분야와 산업에서 활용되고 있다.



<그림 3-5> 클러스터링의 개념

<그림 3-5>는 클러스터링의 기본적인 개념을 설명하고 있는 그림이다. 왼쪽은 클러스터링을 하지 않았을 때 데이터들의 성격을 나타내고 있으며 오른쪽은 클러스터링을 했을 때 유사한 집단은 동일한 색으로 표현되고 있다. 클러스터링은 1957년 k-mean을 근간으로 발전하였는데 k-mean 알고리즘은 클러스터링의 방법론 중에서 분할법에 속한다. 분할법이란 주어진 데이터를 여러 집단으로 나누는 방법이다. 예컨대 n개의 데이터가 있을 때 입력 데이터를 n개 보다 작거나 유사한 K개의 클러스터로 나누는데, 이 때 각 군집은 클러스터를 형성하게 된다. 즉 데이터를 한 개 이상의 데이터 오브젝트로 구성된 K개의 클러스터로 나누는 것이다. 이 때 클러스터를 나누는 과정은 거리 기반의 클러스터 간의 비유사도(dissimilarity)와 같은 비용 함수(cost function)을 최소화하는 방법으로 진행하며 이 과정에서 같은 클러스터 내 데이터 오브젝트 사이의 유사도는 증가하고 다른 클러스터에 있는 데이터 오브젝트와의 유사도는 감소하게 된다. k-mean 알고리즘은 각 클러스터의 중심(centroid)과 클러스터 내의 데이터 오브젝트와의 거리의 제곱합을 비용 함수로 정하고, 이 함수 값을 최소화하는 방향으로 각 데이터 오브젝트의 소속 클러스터를 업데이트 해 줌으로써 수행된다(Lee et al., 2010).

그러나 본 알고리즘은 몇 가지 한계점을 내포하고 있다. 우선 k-mean 알고리즘은 k값에 따라 결과 값이 완전히 달라진다. 예컨대 비용 함수의 함수 공간에서 최적화를 시행할 때, 에러(error)가 줄어드는 방향으로 최적의 값을 찾아가게 되는데, 전역이 아닌 지역 최소한의 값에 도달해도 알고리즘의 수렴 조건을 만족하게 되므로 더 이상

최적화를 진행하지 않게 된다. 이를 방지하기 위해 시뮬레이티드 어닐링(simulated annealing) 기법 등과 병행하기도 한다. 이러한 하이브리드 방법을 통해 의도적으로 에러가 줄어드는 방향을 피하는 방식을 이용하거나 서로 다른 초기 값으로 여러 번 시도해본 뒤 가장 에러가 낮은 결과를 사용하는 기법 등을 이용함으로써 이 문제를 완화할 수 있다.

또한 k-mean 알고리즘은 초기값이 결과 값에 영향을 미칠 뿐만 아니라 큰 이상 값(outlier)에 민감하게 반응한다. 이상 값이란 다른 대부분의 데이터와 비교했을 때 멀리 떨어져 있는 데이터를 의미하는데 이러한 이상 값은 알고리즘 내에서 중심점을 갱신하는 과정에서 클러스터 내의 전체 평균값을 크게 왜곡시킬 수 있다. 따라서 클러스터의 중심점이 클러스터의 실제 중심에 있지 않고 이상 값 방향으로 치우치게 위치할 수 있다. 이를 방지하기 위해 k-mean 알고리즘을 실시하기 전에 이상 값을 제거하는 프로세스를 먼저 실행하거나 분할법의 일종인 k-medoids을 이용하면 이상 값의 영향을 줄일 수 있다.

앞서 설명한 k-mean 알고리즘의 한계점에 대응하기 위하여 본 연구에서는 SOM 기법을 이용하여 클러스터링을 수행한다.

5.2.2 SOM의 이론적 배경

Kohonen(1982)에 의해서 고안된 SOM은 신경망 모델 중에서 하나이며, 자율학습을 사용한 신경망의 하나로 경쟁 학습 모형을 기반으로 하고 있다. 기대되는 산출물이 없이 입력 자료가 주어지는 자율 학습

에서는 비슷한 입력 값들이 들어오는 경우에 그것들을 같은 집단으로 묶기 위해서 스스로 연결 강도를 조절한다. 따라서 입력 자료 집합의 통계적인 특성들을 산출하고 이를 유사한 데이터를 같은 집단으로 묶어 가는 학습 알고리즘이다. 또한 SOM은 입력 층과 출력 층으로 나뉘지는데 관측되어지는 데이터 수만큼 뉴런을 가지고 있는 입력 층은 각각의 뉴런을 통해서 실제 데이터로부터 정보를 받는 역할을 한다. 그리고 입력 층에 존재하는 뉴런들은 출력 층에 있는 뉴런 전체와 연결되어 있고 각각의 연결선은 연결 강도(weight)를 나타낸다.

학습을 통해서 임의로 산정되어 있던 연결 강도는 조정되어갈 수 있는데 우선 입력된 데이터들이 출력 층에 투영되고 출력 층에 존재하는 뉴런들은 학습 알고리즘의 규칙인 유사성의 정도를 고려해서 서로 비교를 한다. 이 때 유사성의 척도로 유클리디안 거리(squared euclidian distances)가 활용되고 여기서 결정된 승자 뉴런과 연결된 연결 강도들은 입력 데이터에 상응하도록 조절된다. 이러한 과정을 반복하면 상대적으로 큰 유사성을 갖고 있는 입력 데이터들끼리 클러스터를 형성하게 된다. 이 과정에서 Kohonen(1982)은 뉴런들의 연결 강도를 조정하기 이전의 신경망과는 다르게 정해진 승자 뉴런만이 유일한 출력 신호(signal)를 보낼 수 있는 승자 독점(winner takes all)의 이론을 제안하였다.

또한 이웃(neighborhood) 개념을 적용하여 승자 뉴런의 주변에 있는 뉴런들이 지닌 연결 강도도 승자 뉴런의 연결 강도와 함께 수정될 수 있도록 하여 입력 패턴의 유사성을 좀 더 잘 반영할 수 있는 특성 지

도를 형성한다. 이와 관련된 내용은 SOM을 이용하여 기업의 부도 예측에 적용한 Kiviluoto(1998)의 연구 그리고 SOM을 이용하여 제조업체의 경쟁적 위치를 연구한 Back et al.(1998)의 연구에서 자세하게 설명하고 있다.

5.3 실험 결과 및 분석

5.3.1 실험 계획

본 연구에서는 상장된 기업의 일별 주가 자료와 함께 일별로 업데이트되는 기술적 자료들을 바탕으로 클러스터링을 시행한다. 상장된 기업들을 대상으로 클러스터링을 진행하여 유사한 기업들을 집단화하는 것이 보편적으로 진행되어온 방법이나, 본 연구에서는 하나의 기업을 대상으로 주가의 패턴을 클러스터화하여 각 클러스터의 특성을 분석한다.

본 연구에서는 특정 기업 하나를 어떻게 선택하느냐에 대한 의사결정 문제는 크게 중요하지 않지만 거래량과 과거 주가의 변동성 등을 고려해서 KOSPI 시장에 상장된 현대모비스로 선정하였다.

클러스터링 대상에 해당하는 기간은 2015년 4월 28일부터 2017년 4월 7일까지 약 2년의 기간이며, 수집된 자료는 <표 3-6>과 같다.

수집된 기술적 자료들은 우리가 일반적으로 사용하는 HTS에서 쉽게 수집할 수 있는 데이터이며 본 연구에서는 이들이 갖는 의미는 크

게 중요하지 않기 때문에 각 기술적 자료들에 해당하는 설명은 따로 하지 않는다. 또한 이렇게 수집된 자료들은 두 차례에 걸쳐서 사전 데이터 정제작업(data scaling)을 실시한다. 각각의 기술적 자료들의 의미하는 바가 다르기 때문에 수집된 원시 자료들은 측정 단위 등이 다를 수밖에 없다. 따라서 본 연구에서는 수집된 자료들을 Z-score Normalization와 Min-max Normalization을 이용해서 표준화 하였다.

<표 3-6> 수집 자료			
기술적 지표			
1	시가	14	Slow Stc. D%
2	종가	15	Fast Stc. k%
3	저가	16	Fast Stc. D%
4	고가	17	RSI
5	매수량	18	DMI
6	매도량	19	ADX
7	Envelop	20	TRIX
8	5일 이동평균	21	SONAR
9	10일 이동평균	22	CCI
10	20일 이동평균	23	MACD
11	60일 이동평균	24	OBV
12	120일 이동평균	25	Parabolic SAR
13	Slow Stc. k%	26	Bollinger Band

<표 3-7>는 앞서 설명한 실험 계획을 표로 보여주고 있다. 본 연구에서는 481일의 수집된 일별 데이터들을 SOM 기법을 이용하여 클러스터링을 시행한다.

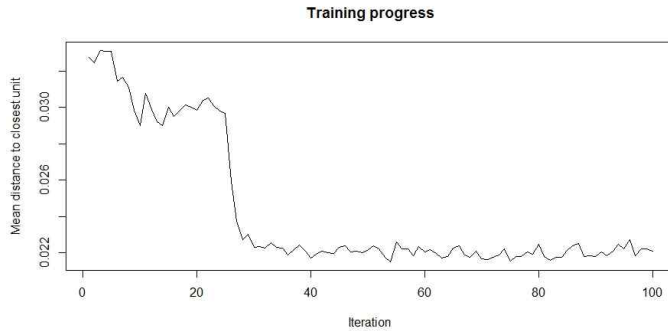
클러스터링을 시행하기 이전에 몇 개의 클러스터로 나눌 것인지 설정해야 하는데 본 연구에서는 선행실험을 통하여 총 10개의 클러스터로 결정하였다. 클러스터의 수가 너무 작으면 군집화하는 과정에서 정확도가 다소 떨어질 수 있으며, 클러스터의 수가 너무 많으면 패턴의 구분이 불명확해지기 때문에 각 클러스터에 대한 정의가 어려울 수 있다.

<표 3-7> 실험 계획

대상	현대모비스 26개 기술적 지표
기간	2015년 4월 28일 ~ 2017년 4월 7일
수집 주기	일별, 총 481일 영업일
분류 기법	Self Organizing Map(10개 클러스터)
자료 수집	이대일리 MARKETPOINT
실험 환경	R (kohonen package ver. 3.0.2)

5.3.2 클러스터링 결과 및 분석

본 연구에서 클러스터링을 통해서 패턴을 분류하였고, 데이터 시각화를 위해 방사형 차트를 활용하여 그 결과를 분석하였다. kohonen package의 SOM 모델은 xdim과 ydim의 값을 각각 5와 2로 총 10개의 클러스터로 구성하였다. 시행횟수는 100, 학습율(learning rate)은 기본 값인 c(0.05, 0.01)로 설정하였고 학습과정은 <그림 3-6>과 같다.

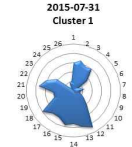
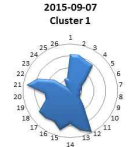


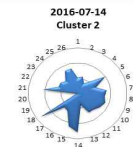

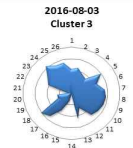

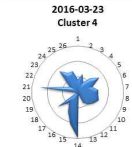
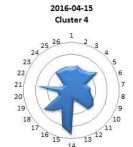

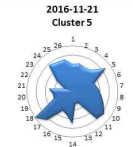



<그림 3-6> 학습 과정

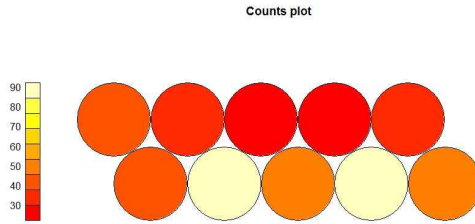
<그림 3-7>는 총 481일간의 각 일별로 어느 클러스터에 속하는지 결과를 확인할 수 있으며 클러스터 1부터 클러스터 10까지 각 클러스터에 속한 날의 수를 나타내는 그림이다.

평균적으로 한 클러스터에 약 48일 정도가 속해있는 것을 알 수 있다. 또한 각 클러스터 간의 표준편차는 약 21일 정도인데 주가의 움직임은 불규칙적으로 움직이기 때문에 반드시 모든 클러스터에 적절하게 분산되어 클러스터링이 되었다고 좋은 것은 아니다. 즉 주가의 움직임에 따라서 특정 클러스터에 많은 날이 속할 수 있고 반대로 모든 클러스터에 적절하게 속할 수도 있다.

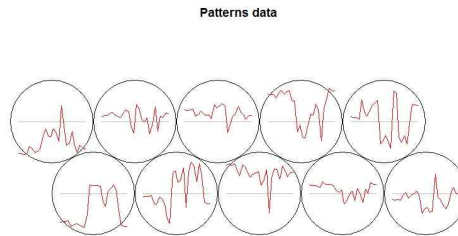
<표 3-8> 클러스터 별 주가 움직임

클러스터	대표 패턴 시각화			패턴 특징
1	 <p>2015-07-31 Cluster 1</p>	 <p>2015-09-07 Cluster 1</p>	 <p>2015-09-08 Cluster 1</p>	주가 상승추세, 분할 매수 추천
2	 <p>2016-06-21 Cluster 2</p>	 <p>2016-06-23 Cluster 2</p>	 <p>2016-07-14 Cluster 2</p>	주가 상승 추세 끝자락, 이후 하락추세로 전환될 가능성이 크다. 매도신호
3	 <p>2016-08-02 Cluster 3</p>	 <p>2016-08-03 Cluster 3</p>	 <p>2016-08-19 Cluster 3</p>	상승 중이거나, 주가 상승 구간의 정점부근에 가깝다. 이후 하락 추세로 전환될 가능성이 크다. 매도신호
4	 <p>2016-03-23 Cluster 4</p>	 <p>2016-04-15 Cluster 4</p>	 <p>2016-04-18 Cluster 4</p>	전체적으로 상승하고 있는 추세이다. 보유하고 있는 주식은 매도하지 말고 좀 더 지켜보거나, 분할매수 추천
5	 <p>2016-11-11 Cluster 5</p>	 <p>2016-11-21 Cluster 5</p>	 <p>2016-11-22 Cluster 5</p>	하락하던 추세를 멈추고 곧 상승으로 추세가 전환될 가능성이 높다. 지켜보다가 적극 매수하는 것을 추천

6	<p>2015-07-13 Cluster 6</p>	<p>2015-07-14 Cluster 6</p>	<p>2015-07-15 Cluster 6</p>	<p>하락하던 추세를 멈추고 곧 상승으로 추세가 전환될 가능성이 높다. 지켜보다가 적극 매수하는 것을 추천</p>
7	<p>2015-10-20 Cluster 7</p>	<p>2015-10-29 Cluster 7</p>	<p>2015-11-10 Cluster 7</p>	<p>주가의 지속적인 상승구간, 분할 매수 추천, 매도 보류</p>
8	<p>2016-09-06 Cluster 8</p>	<p>2016-09-26 Cluster 8</p>	<p>2016-09-28 Cluster 8</p>	<p>기존의 상승구간이 끝나감. 곧 하락추세로 전환될 가능성이 높으므로 분할 매도 추천</p>
9	<p>2016-09-29 Cluster 9</p>	<p>2016-11-04 Cluster 9</p>	<p>2017-01-25 Cluster 9</p>	<p>주가 급락 구간, 매수금지, 분할 매도 추천</p>
10	<p>2017-03-31 Cluster 10</p>	<p>2017-03-28 Cluster 10</p>	<p>2017-04-07 Cluster 10</p>	<p>주가 하락추세 구간, 곧 상승으로 전환될 수 있으므로 매도와 매수는 보류하고 관망</p>



<그림 3-7> 클러스터 별 포함된 일 수



<그림 3-8> 클러스터 별 대표 패턴

<그림 3-8>는 클러스터 별 대표 패턴을 보여주고 있으며, 각 클러스터별로 구분되는 고유의 특징을 보여주는 것을 확인할 수 있다. 본 연구에서 클러스터링을 한 결과, 각 클러스터 별로 속한 날들의 방사형 모양은 <표 3-8>에 표현된 대표 3개의 방사형 모양과 모두 유사한 모양을 보이는 것을 알 수 있었다. 또한 그림을 보면 알 수 있듯이 각 클러스터 별로 방사형 모양의 특징을 쉽게 파악할 수 있고 이러한 방식의 시각적 표현 방식은 향후 기계 학습과 같은 연구 과정에서 도움이 될 것으로 사료된다.

향후 연구계획으로 각 클러스터에 속한 패턴들이 실제 주가의 움직임에 어떠한 영향을 주는지 분석하고 그 영향들을 계량화하여 기계학습을 통한 매매 전략을 진행하고자 한다. 그 전에 본 연구에서는 각 패턴들이 실제 주가흐름을 어떻게 대표하는지 분석한 결과는 <표 3-8>에 정리되어 있다. 이는 각 각의 패턴들이 발생한 후 영업일 기준으로 전후 약 5일 이내의 주가 움직임을 분석한 결과이며, 추후 연구에서 좀 더 정량적인 방법을 통해서 각 패턴이 갖는 의미에 대해서 설명하고자 한다.

제 4 장. 결 론

본 논문은 최근 다양한 분야에서 관심이 높은 빅데이터를 계량적인 방법론으로 금융 분야에 적용하여 포트폴리오 구성과 섹터 투자 전략 그리고 인공지능화가 가능한 방사능 차트를 이용한 개별 종목에 대한 매매 전략 등을 제안한다. 특히 본 논문의 2장에서는 동일한 섹터의 개별 종목들을 주가가 아닌 재무 자료를 이용해서 투자 가치가 높은 종목들로 포트폴리오를 구성하는 방법론을 제안하고 이에 대한 성과를 측정하기 위해서 최근 금융 시장에서 자산 규모가 커지고 있는 ETF와 비교 분석한다. 또한 3장에서는 이러한 ETF의 투자 전략을 빅데이터와 텍스트마이닝을 통해서 금융상품에 기본이 되는 주식과 파생상품의 하나인 선물을 정량적이고 과학적인 접근 방식으로 매매할 수 있는 방법론을 제안하였다. 다음으로 설명하는 결론은 3장의 응용 사례에 관한 본문 내용은 본 장에서 설명하지 않고 향후 연구 계획으로 대체하고자 한다.

본 논문의 2장에서는 투자자들의 수요가 증가하고 있는 ETF 상품의 포트폴리오 구성을 보다 더 정량적인 방법론으로 구성하고자 경영 과학 분야의 DEA 모형을 포트폴리오 구성 전략에 적용하는 방법론을 제안하였다. DEA 모형은 기업의 경영 효율성을 평가하는 방법론으로써 기업의 재무 자료를 바탕으로 경영 효율성 점수를 산출한다. 이를 포트폴리오 구성 전략에 적용하고 그에 대한 성과를 분석하기 위해서 실제 시장에서 거래되고 있는 ETF 상품과 비교 및 분석하였다. ETF 상품의 특성 상, 특정 업종을 선택해야하는데 본 연구에서는 IT 업종

을 선택하였고 본 연구에서 제안하는 방법론과 실제 시장에서 거래되는 TIGER 200 IT와 TIGER 소프트웨어와 수익률 측면 그리고 포트폴리오 성과 측면으로 분석하였다.

그 결과 실제 시장에서 거래되는 ETF 상품에 비해서 DEA 모형을 이용하여 ETF를 구성했을 때 월등하게 높은 수익률을 보였다. 본 연구에서 제안하는 DEA 모형을 통해서 ETF를 구성하였을 경우에는 2013년부터 2016년도까지 약 200%의 누적 수익률을 기록한 반면에 TIGER 200 IT와 TIGER 소프트웨어는 각각 약 11%, 38%의 누적 수익률을 보였다. 또한 포트폴리오의 성과 측정 도구인 샤프지수, 켄센의 알파 그리고 IR을 산출해본 결과, 모두 DEA 모형을 이용하였을 때의 ETF가 우수한 결과를 보였다.

다음으로 3장에서는 대중들의 감정이 개인행동과 의사결정에 큰 영향을 미칠 수 있다는 행동경제학의 이론을 토대로 국내 빅데이터 트렌드를 이용한 ETF 투자전략을 제안하였다. 시장 참여자가 경제에 갖는 관심은 인터넷 사용자의 경제 관련 검색어와 연결될 수 있으며 특정 기간의 해당 검색량은 경제 분야 빅데이터의 트렌드로 이해할 수 있다.

따라서 본 연구는 시장 참여자의 경제에 대한 관심도가 증가하는 시점이 주식시장 주가의 하락 시점을 선행하고, 반대로 관심도의 하락은 주식시장 주가의 상승을 선행한다는 이론을 수립하였고, 국내 주식시장의 주가 및 네이버 트렌드(Naver trends) 검색량 데이터 변화량의

연관성을 이용한 ETF 투자전략을 제안하고 결과를 분석하였다. 또한 텍스트마이닝 기법을 이용하여 섹터 별로 의미 있는 키워드(Keyword)를 선정하였다.

이는 브레인스토밍 기법(Brain Storming Method)인 BS 기법을 통해서 선정된 키워드를 바탕으로 2011년부터 2014년까지 ETF를 매매한 수익률과 비교하였고, 그 결과 대체적으로 본 연구에서 제안한 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매했을 때 보다 더 높은 수익률을 보였다. 또한 시장 참여자들의 정보검색 움직임을 정량화하기 위해 검색 상대변화량 $\Delta_1 \sim \Delta_3$ 로 나눠 실험한 결과, 누적 로그수익률 측면에서는 Δ_3 가 다소 높은 성과를 보였다. 그러나 매매 운용에 따른 수익률만을 갖고 운용 성과를 정량적으로 판단하기는 한계가 있기 때문에 샵프 지수, 켄센의 알파, IR을 산출하고 Δ 간의 성과를 분석하였다.

그 결과 수익률이 다소 높았던 Δ_3 보다는 Δ_1 와 Δ_2 가 우수한 성과를 보였다. 섹터 별로 선정된 키워드에 따라서 운용 성과가 다르게 나타난 본 연구의 결과를 통해서 투자자의 심리가 반영된 포털 트렌드를 바탕으로 매매 전략을 수립할 시, 투자 자산과 연관된 키워드를 효과적으로 선정하는 것이 무엇보다 중요하다는 것을 알 수 있었다.

향후 연구 계획으로는 본 연구에서 제안하는 다양한 계량적인 방법론과 3장의 응용사례에서 설명한 클러스터링을 이용한 금융 상품 투자 방법론을 보완하여 인공지능망과 같은 기계 학습 기법을 이용해서

파생상품 매매에 있어서 인공지능 매매 시스템의 개발 및 실효성 분석을 하고자 한다. 특히 응용 사례에서 제안하는 26개의 기술적 자료를 대체할 수 있는 요인들을 정량적인 기법을 통해서 정의하고 각 패턴들이 갖는 의미를 기계 학습에 적용하는 향후 연구에 본 논문이 큰 기여를 할 것으로 사료된다.

참 고 문 헌

- [1] 김유신, 김남규, 정승렬, “뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형”, 지능정보연구, 제18권, 제2호, pp.143-156, 2012.
- [2] 김종기, 강다연 “DEA 모형을 이용한 국내 아파트 건설기업(상장기업)의 효율성 분석”, 한국콘텐츠학회논문지, 제8권, 제7호 pp. 201-207, 2008.
- [3] 박재연, 유재필, 신현준 “로보어드바이저를 이용한 포트폴리오 관리”, 한국정보기술아키텍처논문지, 제13권, 제3호, pp. 467-476, 2016.
- [4] 박원준, “‘빅 데이터(Big Data)’ 활용에 대한 기대와 우려”, Journal of Communications & Radio Spectrum, 제51권, pp.28-47, 2012.
- [5] 신현준, 라현우, “금융시장의 빅데이터 트렌드를 이용한 주가지수 투자 전략”, 한국경영과학회지, 제32권, pp.91-103, 2015.
- [6] 신현준, 유재필, “DEA-마코위츠 결합 모형을 이용한 건설기업의 효율적 포트폴리오 구성 방안”, 한국산학기술학회, 제14권, pp.899-904, 2013.
- [7] 이철기, 이우기 “DEA에 의한 할리우드 영화 효율성 분석 및 이를 응용한 영화의 분류 시스템”, 한국정보기술아키텍처논문지, 제13권, 제3호, pp. 487-495, 2016.
- [8] 유재필, 신현준 “DEA-마코위츠 결합 모형을 이용한 건설업종 투자 전략”, 한국산학기술학회논문지, 제14권, 제2호, pp. 899-904, 2013.
- [9] 이득환, 김수현, 강형구, “빅데이터를 사용한 시스템 트레이딩 : KOSPI200 선물을 대상으로”, 2014년 5개 학회 공동학술연구발표회: 한국재무학회, 2014.

- [10] 옥기율, 김지수, “소비자 심리지수가 KOSPI 수익률에 미치는 비대칭적 영향에 대한 연구”, 금융공학연구, 제11권, 제1호, pp.17-37, 2012.
- [11] 최다영, 안범준, 신현준 “기업의 성장가능성을 고려한 포트폴리오 선택 전략”, 한국산학기술학회논문지, 제12권, 제9호, pp. 3849-3855, 2011.
- [12] Back B, Sere K, Vanharanta H, “Managing complexity in large data bases using self-organizing maps”, Management and Information Technologies, Vol.8, No.4, pp. 191 - 210, 1998.
- [13] Barber, B. M., Odean, T. and Zhu, N., “Systematic Noise”, Journal of Financial Markets, Vol.12, pp.547-569, 2009.
- [14] Bordino, I., Battiston, S., Caldarelli, G., Cristell, M. and Ukkonen, A., “Web Search Queries Can Predict Stock Market Volumes”, PloS ONE, Vol.7, No.7, pp.1-17, 2012.
- [15] Choi, H. and Varian, H., “Predicting the Present with Google Trends”. The Economic Record, Vol.88, pp.2-9, 2012.
- [16] Cootner, P.H., “The random character of stock market prices”, MIT Press, 1964.
- [17] Dorn, D., Huberman, G. and Sengmueller, P., “Correlated Trading and Returns”, Journal of Finance, Vol.63, pp.885-920, 2008.
- [18] Fu, T.C., F.L. Chung, R. Luk, and C.M. Ng, "Stock time series pattern matching : Template-based vs. rule-based approaches", Engineering Applications of Artificial Intelligence, Vol.20, No.3, pp. 347-364, 2007.
- [19] Jackson, A., “The Aggregate Behaviour of Individual Investors”, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.536942>, 2003.
- [20] Jung, Y.K. and Y.S. Yoon, “An Study on Predictability of Stock Prices

- using Artificial Neural Networks”, The Korean Journal of Financial Management, Vol.15, No.2, pp.369-399, 1998.
- [21] Kiviluto, K. and E. Oja, "S-map : A network with a simple self-organization algorithm for generative topographic mappings", MIT Press, 1998.
- [22] Kumar, A. and Lee, C. M. C., "Retail Investor Sentiment and Return Comovements", Journal of Finance, Vol.61, No.5, pp.2451-2486, 2006.
- [23] Lee, J.W. "A Stock Trading System based on Supervised Learning of Highly Volatile Stock Price Patterns", KIISE Transactions on Computing Practices, Vol.19, No.1, pp. 23-29, 2013.
- [24] Lee, K.H. and G.S. Jo, "Expert system for predicting stock market timing using a candlestick chart", Expert Systems with Applications, Vol.16, No.4, pp. 357-364, 1999.
- [25] Lee, W.K., W.H. Lee, and H.K. Lee, "Web Clustering Simulation Using Genetic Algorithm", Journal of Information Technology and Architecture, Vol.7, No.1, pp. 111-121, 2010.
- [26] Lippens, R.E., "Multimaturity efficient market hypotheses : Sorting out rejections in international interest and exchange rate markets", International Journal of Forecasting, Vol.3, No.1, pp. 149-158, 1987.
- [27] Liu, J.N.K. and R.W.M. Kwong, "Automatic extraction and identification of chart patterns towards financial forecast", Applied Soft Computing, Vol.7, No.4, pp. 1197-1208, 2007.
- [28] Lo, A.W., H.M. Mamaysky, and J. Wang, "Foundations of technical analysis : computational algorithms, statistical inference, and

empirical implementation”, *Journal of Finance*, Vol.55, pp. 1705–1770, 2000.

- [29] Min J.H. and C.W. Jeong, “Design and Performance Measurement of a Genetic Algorithm–based Group Classification Method : The Case of Bond Rating”, *Journal of the Korean Operations Research and Management Science Society* , Vol.32, No.1, pp. 61–74, 2007.
- [30] Park, J.Y., J.P. Ryu, and H.J. Shin, “Portfolio Management Using Robo–Advisors”, *Journal of Information Technology and Architecture*, Vol.13, No.3, pp. 467–476, 2016.
- [31] Park, S.J, “Stock Price Data Mining Using Neural Network”, *Journal of Knowledge Information Technology and Systems*, Vol.2, pp. 25–32, 2007.
- [32] Preis, T., Reith, D. and Stanley, H. E., “Complex Dynamics of Our Economic Life on Different Scales: Insights from Search Engine Query Data”, *Philosophical Transaction of the Royal Society A*, Vol.368, pp.5707–5719, 2010.
- [33] Preis, T., Moat, H. S. and Stanley, H. E., “Quantifying Trading Behavior in Financial Markets Using Google Trends“, *Scientific Reports*, 3, doi:10.1038/srep01684, 2013.
- [34] Saeed A. and W.T. Ying, “Stock market co–movement assessment using a three–phase clustering method”, *Expert Systems with Applications*, Vol.41, pp.1301–1314, 2014.

ABSTRACT

Portfolio Management Using Big Data and Artificial Intelligence

Chang Hoon Han

Dept. of Management Engineering

The Graduate School

Sangmyung University

This paper studies portfolio management and sector investment strategies through quantitative methodologies using various big data existing in the financial market. Chapters 2 and 3 describe the technical procedures for portfolio construction and sector investment via big data trends, respectively.

Chapter 2 deals with portfolio construction strategies using data envelopment analysis (DEA) method based on a variety of financial data of corporations. Recently, as concerns about the global economic crisis have increased, many investors have become

increasingly distrustful of the active market, while passive markets such as ETF products are becoming more popular. However, the ETF, which is the representative of the passive market in Korea, has many ETFs listed in a short period of time. However, most of the ETFs constitute portfolios with high market capitalization, and the composition ratio also has a limit of being concentrated in a specific company. As such, it does not include stocks that are highly likely to grow or are undervalued relative to the value of the company. This study constructs a portfolio using DEA method that evaluates the management efficiency of a firm and compare it with ETF products that are actually listed on exchange. It is necessary to select a specific sector based on the characteristics of the ETF product. This study selects the IT industry and constructs a portfolio using data envelope analysis for IT companies listed on the KOSPI and KOSDAQ, and analyzed ETF products, yield analysis, and portfolio performance, called TIGER 200 IT and TIGER software, which constitute the most similar sectors . As a result, ETF product management using the data envelope analysis method proposed in this study is superior to the benchmark ETF products listed on the actual market in terms of profitability and performance measurement.

Chapter 3 mentions quantitative trading strategies for ETF introduced in chapter 2 using big data, which are various keywords existing in the financial market. Researches on applying big data trends to financial market have been actively conducted, while a lot of attempts using big data for various industries are increasing. In addition, researches show that there is a correlation between the movement of the financial market and the sentimental changes of the public participating directly or indirectly in the market and applies the relationship to investment strategies for stock market. Unlike previous studies, this study breaks down the stock market into 11 sectors in order to closely capture the trends from each markets. Keywords for each sector are selected by text mining and brainstorming methods, and trends data of these keywords are collected for recent five years. The computational results illustrate that the invest strategy based on text mining shows better performance than one based on brain storming in terms of accumulated rate of returns.

Since the various methodologies proposed in this paper rely on a non-qualitative, scientific and quantitative methodology, it will contribute to the development of AI based portfolio management and trading system that apply machine learning in the future.