碩士學位論文

머신러닝과 수급분석을 활용한 주식 포트폴리오 구성 연구

A Study on Stock Portfolio Modeling
using Machine Learning and Supply-Demand Analysis

國民大學校 비즈니스IT專門大學院
트레이딩시스템 專攻

김 경 목

2017

머신러닝과 수급분석을 활용한 주식 포트폴리오 구성 연구

A Study on Stock Portfolio Modeling
using Machine Learning and Supply-Demand Analysis

指導敎授 金 善 雄

이 論文을 碩士學位 請求論文으로 提出함

2017 年 12 月

國民大學校 비즈니스IT 專門大學院

트레이딩시스템 專攻

김 경 목

2017

金耿睦의

碩士學位 請求論文을 認准함

2017年 12月

審査委員長 최 홍 식 ㊞

審査委員 김 선 웅 ㊞

審査委員 김 남 규 ㊞

國民大學校 비즈니스IT 專門大學院

차 례

표 차례

그림 차례

국문 요약

머신러닝과 수급분석을 활용한 주식 포트폴리오 구성 연구

김 경 목

트레이딩시스템 전공

본 논문은 머신러닝과 수급분석을 활용한 주식 포트폴리오 구성 방법을 개발하고 그 성과를 분석하는 연구이다. 금융 알파고라 불리는 주식 로보-어드바이저 프로그램의 개발 방향을 제시하고 누구나 접할 수 있도록 하여 경제적 가치 향상에 기여 하는데 의의가 있다.

자기 조직화 지도 모델 인공신경망을 이용하여 수급분석 데이터를 그룹화하고 그룹화한 데이터를 변환하여 오류역전파 모델 인공신경망에 학습을 시켜서 검증 데이터 예측결과로 매월 포트폴리오 구성을 하도록 개발하였다. 성과 분석을 위해 포트폴리오의 벤치마크를 지정하였고 시장 수익률 비교를 위해 KOSPI200, KOSPI 지수 수익률도 구하였다. 포트폴리오의 동일배분 수익률, 복리 수익률, 연평균 수익률, MDD, 표준편차, 샤프지수, 벤치마크로 지정한 시가총액 상위 10종목의 Buy and Hold 수익률 등을 사용하여 성과 분석을 진행하였다. 분석 결과 포트폴리오가 벤치마크 대비 2배 수익률을 올렸으며 시장 수익률보다 좋은 성과를 보였다. MDD와 표준편차는 포트폴리오와 벤치마크가 비슷한 결과로 성과 대비 비교한다면 포트폴리오가 좋은 성과라고 할 수 있다. 샤프지수도 포트폴리오가 벤치마크와 시장 결과보다 좋은 성과를 내었다.

이를 통해 머신러닝과 수급분석을 활용한 포트폴리오 구성 프로그램 개발 방법의 방향을 제시하였고 우수한 성과로 실제 주식 투자에 프로그램 개발 적용에 활용할 수 있음을 보였다.

키워드 : 인공지능, 머신러닝, 수급분석, 로보-어드바이저, 트레이딩시스템, 주가 수익률, KOSPI200, KOSPI

제1장 서론

1.1 연구의 배경 및 목적

이세돌과 알파고와의 딥마인드 챌린지 매치(Google Deepmind Challenge Match)[1]는 2016년 3월 9일부터 10일, 12일, 13일, 15일까지 총 5회에 걸쳐 이루어진 최고의 바둑 인공지능 프로그램과 바둑 최고 인간 실력자의 대결로 알파고가 4승 1패로 승리하였다. 사람들의 인식에 인공지능(Artificial Intelligence)[2]이라는 뜨거운 이슈를 만들었으며 이때부터 인공지능에 대한 관심이 더욱 뜨거워지기 시작했다. 이후 금융 쪽에서도 자산관리 발전을 위해 금융의 알파고라 할 수 있는 로보-어드바이저(Robo-advisor)[3] 프로그램 개발에 열을 올렸다. 로보-어드바이저라는 자산관리 프로그램들이 하나둘 모습을 나타내면서 투자자들의 관심은 날이 갈수록 커지고 있지만 인공지능이라는 트렌드에만 집중한 나머지 좋지 않은 실적들이 보고되고 있다. 로보-어드바이저 속내를 들여다보면 아쉽게도 여러 펀드를 출시했지만 상당부분 기존의 알고리즘 트레이딩에서 크게 벗어나지 않았다. 국내의 로보-어드바이저 서비스가 알파고와 같은 기술로 투자자들에게 보다나은 서비스를 보여주기 위해서는 더욱 기술 발전에 대한 노력이 필요하다. 투자에서 인공지능 기술을 활용하여 좋은 결과를 얻기란 쉬운 일이 아니다. 따라서 구체적으로 머신러닝을 이용한 한 방법을 제시 하고자 한다.

본 연구에서는 실질적인 머신러닝 기술과 주식 수급데이터를 이용한 포트폴리오 구성 방법으로 로보-어드바이저 인공지능 기술에 활용할 수 있는 기초를 다지며 경제적 가치 향상에 기여할 수 있는 방법을 구현해 보고자 한다.

---

1) 딥마인드 챌린지 매치(Google Deepmind Challenge Match)란 2016년 3월에 있었던 구글 딥마인드사의 바둑 인공지능 프로그램인 알파고와 한국의 프로 기사인 이세돌 9단과의 바둑 대국이다.

2) 인공지능(Artificial Intelligence)이란 인간의 생각하는 능력을 일부 또는 전체를 인공적으로 구현한 것이다.

3) 로보-어드바이저(Robo-advisor)란 '로봇(robot)'과 '투자전문가(advisor)'를 결합한 단어로 인공지능 자산관리 프로그램을 말한다.

1.2 연구의 내용

머신러닝 학습을 위한 자기 조직화 지도 모델 인공신경망을 이용하여 투자주체별 수급분석 데이터를 군집화 한다. 군집화 방법은 투자주체별 수급데이터의 개인, 외국인, 기관과 일별 수익률을 20가지 경우로 그룹화 한다. 군집화하기 전에 투자주체별 수급데이터와 일별 수익률은 정규화 작업을 진행한다. 그룹화한 데이터는 오류역전파 모델 인공신경망에 사용할 수 있게 5일 간격으로 60일 연속 데이터를 변환 작업을 진행한다. 변환 작업으로 만들어진 데이터는 오류역전파 모델 인공신경망에 적용하기 위해 정규화 작업을 진행한다. 오류역전파 모델 인공신경망을 학습하고 검증데이터를 사용하여 월별 포트폴리오 구성을 진행한다.

오류역전파 모델 인공신경망에 학습된 데이터의 예측률과 검증 데이터의 예측률을 분석한다. 검증 데이터 개별종목의 월별 수익률을 사용하여 포트폴리오의 동일배분 수익률과 복리 수익률을 구한다. 검증 데이터의 포트폴리오와 비교할 시가총액 상위 10종목을 벤치마크로 정하고 수익률을 비교 분석한다. 시장 수익률 비교를 위해 KOSPI200 지수와 KOSPI 지수를 비교 분석한다. 분석 방법은 포트폴리오의 동일배분 수익률, 복리 수익률, 연평균, MDD, 표준편차, 샤프지수 등을 사용한다. 샤프지수의 계산에서 무위험자산 수익률은 1년 만기국채 수익률을 이용한다. 벤치마크의 Buy And Hold 수익률도 분석한다. 실제 같은 운용처럼 분석하기 위해 진입, 청산 수수료와 세금도 차감해서 분석한다.

제2장 머신러닝

머신러닝(Machine Learning)은 컴퓨터과학의 인공지능의 한 부분이며 패턴인식과 컴퓨터 학습이 진행되는 원인이 무엇인지 설명하는 이론의 연구에서 발전한 분야이다. 과거 데이터를 바탕으로 학습하고 예측하며 스스로의 성능을 높여가는 시스템과 이를 위한 모델을 연구하고 구현하는 기술이다. 모델들은 정해진 프로그램 명령들을 수행하기보다는 입력 데이터를 바탕으로 예측하거나 결과를 도출하기 위한 모델을 구현하는 방식이다.

2.1 자기 조직화 지도 모델

자기 조직화 지도(Self Organizing Map) 모델은 인공신경망(Artificial Neural Network)[4]으로써 무감독(Unsupervised Learning) 학습에 의한 클러스터링(Clustering)[5]을 수행하는 모델로 1984년 이후 코호넨(Teuvo Kohonen)[6]에 의해 소개된 모델이다.

면 개념을 사용하여 면내의 인공신경세포(Artificial Neuron)[7]들 간의 경쟁을 구현하며 모든 연결은 아래서 위로 가는 비회귀 인공신경망이다. 다층으로의 확장도 가능하지만 일반적으로 단층을 많이 사용된다. 인공신경세포들의 활성 함수로는 선형 함수를 사용하며 학습 규칙은 일반적인 경쟁 학습과 마찬가지로 인스타 규칙(Instar rule)을 사용한다. 인스타 규칙은 연결 가중치 조절의 기본이 된다.

그 핵심은 다음과 같다.

---

4) 인공신경망(Artificial Neural Network)은 생물학적 신경망을 인공적으로 구현하고자한 통계학적 학습 모델이다.
5) 클러스터링(Clustering)이란 데이터의 유사성에 기초하여 데이터를 여러 그룹으로 분류하는 방법이다.
6) 코호넨(Teuvo Kohonen) : 1934년 7월 11일 출생 하였고 핀라드 아카데미 학술 연구원이자 명예 교수이다.
7) 인공신경세포(Artificial Neuron)란 생물학적 신경계를 이루는 기본적인 단위세포를 인공신경망에서 수학적으로 모델링한 신경세포를 말한다.

'어떤 인공신경세포가 특정 연결을 자극하면 그것의 연결 가중치를 그 자극과 같아지도록 조절한다.'

인스타 규칙을 식으로 쓰면 다음과 같다.

$$w(N)_{ij} = w(O)_{ij} + \alpha(a_i - w(O)_{ij}) \qquad \text{식(1)}$$

$w(N)_{ij}$: 인공신경세포 i, j 사이의 수정된 후의 연결 가중치
$w(O)_{ij}$: 인공신경세포 i, j 사이의 수정되기 전의 연결 가중치
α : 학습률($0 < \alpha \leq 1$)
a_i : 인공신경세포 i의 활성 값

식(1)에서 $w(N)_{ij}$는 새롭게 변경된 인공신경세포 i, j 간의 연결 가중치이다. $w(O)_{ij}$는 수정되기 전의 연결 가중치이다. α는 학습률, a_i는 인공신경세포 i의 활성 값이다.

경쟁 관계 구현에 있어 주목할 만한 것은 이웃(neighborhood)이라는 개념을 사용하고 있다. 이것은 멕시코 모자(Maxican Hat) 형태[8]의 경쟁 형태를 단순화시켜 구현하기 위해 도입된 것으로 승자 인공신경세포로부터 일정거리에 있는 인공신경세포들을 말하며 이웃들은 승자 인공신경세포와 같은 면에 있는 인공신경세포들로 구성된다. 승자 인공신경세포의 연결 가중치가 조절 되며 이웃들의 연결 가중치도 함께 수정한다. 승자 인공신경세포뿐만 아니라 그것의 이웃들도 해당 입력 패턴에 반응하게 되는 것이다. 이러한 이웃이라는 개념을 사용하여 멕시코 모자 경쟁 형태를 구현하는 이유는 학습이 진행됨에 따라서 이웃의 범위를 줄이는데 있다.

즉 학습 초기에는 같은 면내의 모든 인공신경세포들을 이웃으로 하고 그렇게 하면 모든 인공신경세포들의 연결 가중치가 조절되므로 처음에는 임의로 분산되어져 있던 인공신경세포들의 연결 가중치들은 거의 같은 점에서 모이게 되어 결국 모든 인공신경세포들이 거의 같은 패턴에 반응하게

8) 멕시코 모자(Maxican Hat) 형태란 인공신경망에서 경쟁 관계를 나타내기 위한 인공신경세포들의 형태이다. 이 형태의 경쟁은 실제 동물 신경계와 유사한 형태를 가지고 있다.

된다. 학습이 진행되면서 이웃의 범위는 점차 줄어들고 그때부터 인공신경세포들의 연결 가중치는 조금씩 자신의 입력된 패턴을 쫓아 다시 분산된다.

이때도 역시 범위는 좁아졌지만 이웃의 연결 가중치를 함께 조절함으로써 이웃과 함께 이동하게 된다. 결국 주어진 입력 패턴에 대한 승자 인공신경세포의 이웃 인공신경세포들은 언제나 승자가 반응한 패턴과 유사한 패턴에 반응하게 되는 것이다.

자기 조직 지도 모델에서 승자 인공신경세포의 선정은 연결 가중치와 입력 패턴과의 거리로 구한다.

다음은 거리를 구하는 식이다.

$$distance_{pj} = \sqrt{\sum_{j}(a_{pi} - w_{ij})^2}$$

$distance_{pj}$: 인공신경세포 j의 연결 가중치와 입력된 패턴 p 간의 거리
a_{pi} : 입력층에서 인공신경세포 i의 출력
w_{ij} : 인공신경세포 i와 인공신경세포 j 사이의 연결 가중치

자기 조직 지도 모델에서 인공신경망을 학습시키는 과정이다.

① 입력층에 입력 패턴을 제시한다.
② 모든 층에 대해서 입력층을 제외한 아래쪽 층부터 동작시킨다. 해당하는 층 내의 모든 면에서 승자 인공신경세포를 구한다. 승자 인공신경세포와 이웃의 연결 가중치를 조절한다.

$$w(N)_{ij} = w(O)_{ij} + \alpha(a_i - w(O)_{ij})$$

③ 출력층 끝까지 ②과정을 반복한다.
④ ① ~ ③과정을 입력 패턴 전부에 대해 반복한다.
⑤ 이웃의 범위와 학습률을 감소시킨다. ④과정을 반복한다.
⑥ ⑤과정을 이웃 범위가 자기의 자신이 되기까지를 반복한다.

2.2 오류역전파 모델

인공신경망으로써 감독(Supervised Learning) 학습에 의한 분류(Classification)를 수행하는 모델이다.

오류역전파(Back-propagation) 모델의 핵심은 다음과 같다.

'만일 어떤 인공신경세포의 활성이 다른 인공신경세포가 잘못된 출력에 공헌을 하였다면 두 인공신경세포 간의 연결 가중치를 그것에 비례하여 수정해 주어야 한다. 그리고 그러한 과정은 그 아래에 있는 인공신경세포들까지 계속된다.'

출력층 인공신경세포의 잘못된 출력에 대한 책임이 바로 아래층 인공신경세포에게만 있는 것만 아니라 그것에 달린 더 아래층 인공신경세포에게도 있기 때문에 그것들에게도 책임을 물어 연결 가중치를 수정한다. 이러한 잘못의 대해 연대 책임을 묻기 위해서 출력 층에서 발생하는 에러를 아래층으로 역 전파 시키는 것을 오류역전파 모델이라고 부른다.

인공신경망이 주어진 입력에 대해 동작하고 나면 출력층 인공신경세포들의 에러가 구해지는데 목적 패턴에서 출력 인공신경세포의 활성 값을 뺀 값이 해당하는 출력 인공신경세포의 에러이다.
이 에러를 한 번 더 가공하여 각각의 출력층 인공신경세포에 대하여 델타를 구한다.

$$\delta_j = f'(n_j)e_j = a_j(1-a_j)e_j \qquad \text{식(2)}$$

출력층 인공신경세포의 경우

$$e_j = t_j - a_j$$

시그모이드 함수(Sigmoid Function)[9]의 경우

$$f'(n_j) = \frac{\partial f(n_j)}{\partial n_j} = a_j(1 - a_j)$$

δ_j : 출력층 인공신경세포 j의 델타
$f'(n_j)$: 출력층 인공신경세포 j의 활성 함수의 미분 값
e_j : 출력층 인공신경세포 j의 에러
t_j : 출력층 인공신경세포 j에 대응하는 목적 패턴의 성분
a_j : 출력층 인공신경세포 j의 활성 값

이렇게 출력층 인공신경세포들의 델타가 구해지면 이 델타를 아래층 인공신경세포들로 역 전파 한다. 이 델타의 역 전파는 마치 인공신경망의 정상적인 동작을 거꾸로 완전히 뒤집는 것과 같으며 위층 인공신경세포에서 발생한 델타에서 그것에 연결된 연결 가중치의 값이 곱해져 아래층 인공신경세포로 전달되고 그렇게 전달되는 델타들은 거기서 합쳐진다. 그렇게 합쳐진 값은 은닉층 해당 인공신경세포의 에러가 된다. 이렇게 은닉층 인공신경세포들의 에러가 구해지면 식(2)에서 은닉층 인공신경세포의 델타를 값을 구할 수 있다.

이것을 식으로 정리하면 다음과 같다.

$$\delta_i = f'(n_i)e_i = a_i(1 - a_i)e_i$$

은닉층 인공신경세포의 경우

$$e_i = \sum_j w_{ij}\delta_j \qquad \text{식(3)}$$

9) 시그모이드 함수(Sigmoid Function)는 선형 함수(Linear Function)와 역치 함수(Threshold Function)의 특징을 가지고 있는 비선형 함수(Non-Linear Function)이며 미분 가능하고 S와 같은 형태를 가지는 함수이다. 인공신경망에서 학습 곡선으로 나타낸다.

시그모이드 함수의 경우

$$f'(n_i) = \frac{\partial f(n_i)}{\partial n_i} = a_i(1 - a_i)$$

δ_i : 은닉층 인공신경세포 i의 델타
$f'(n_i)$: 은닉층 인공신경세포 i의 활성 함수의 미분 값
e_j : 은닉층 인공신경세포 i의 에러
w_{ij} : 인공신경세포 i에서 인공신경세포 j로 가는 연결 가중치
δ_j : 출력층 인공신경세포 j의 델타
a_i : 은닉층 인공신경세포 i의 활성 값

위 식에서 은닉층 인공신경세포의 경우에는 에러를 구하는 식(3)이 다를 뿐이다. 다른 식들은 위에서 이야기했던 출력층 인공신경세포의 경우와 같은 것을 볼 수 있다. 은닉층 인공신경세포의 경우는 출력층 인공신경세포와 달리 목적 패턴으로부터 직접 에러를 구할 수 없다. 따라서 출력층 인공신경세포에서 구해진 에러를 델타로 변화시킨 값을 역 전파 받아서 자신의 에러로 하는 것이다.

다음은 오류역전파 모델에 의해 연결 가중치를 조절하는 식이다.

$$w(N)_{ij} = w(O)_{ij} + \alpha \delta_j a_i \qquad \text{식(4)}$$

$$\delta_j = a_j(1 - a_j)e_j$$

출력층 인공신경세포의 경우

$$e_j = t_j - a_j$$

은닉층 인공신경세포의 경우

$$e_j = \sum_k w_{jk}\delta_k$$

식(4)에서 δ_j는 연결 가중치를 고치고자 하는 인공신경세포 j의 델타 값이다. 그리고 a_i는 인공신경세포 j에 연결되어 있는 아래층 인공신경세포이다. 연결 가중치는 그것에 연결되어 있는 인공신경세포의 활성 값에 비례해서 수정된다.

δ_j는 인공신경세포 j의 에러로부터 구하고 있는데 인공신경세포 j가 출력층 인공신경세포인지 은닉층 인공신경세포인지에 따라서 에러를 계산하는 방법이 달라진다.

인공신경세포 j가 은닉층 인공신경세포인 경우에는 자신의 위층에 있는 인공신경세포들로부터 연결 가중치가 곱해져서 전달되는 델타들의 합에 의해서 에러가 계산되어진다.

다음은 오류역전파 모델 인공신경망 학습 과정이다.

① 입력층 인공신경세포에 입력 패턴을 제시한다.

② 인공신경망을 동작시킨다. 바이어스(bias)[10]를 사용할 경우 이를 포함시킨다.

$$a_j = \frac{1}{(1 + \exp(n_j + bias_j))}$$

③ 출력층 인공신경세포들의 에러와 델타를 구해 이를 은닉층으로 역전파한다.

$$e_j = t_j - a_j$$

10) 바이어스(bias)는 모든 인공신경세포에서 입력층 인공신경세포를 제외한 인공신경세포가 가지는 특성이며 시그모이드 함수에서 좌우 이동을 결정하는 상수의 역할에 해당한다. 인공신경망에서 역치 함수의 역치 기능을 한다.

$$\delta_j = a_j(1 - a_j)e_j$$

④ 역 전파되어진 델타로부터 은닉층 인공신경세포들의 에러와 델타 값을 구하여 이를 역 전파한다.

$$e_j = \sum_k w_{jk}\delta_k$$
$$\delta_j = a_j(1 - a_j)e_j$$

⑤ ④과정을 입력층의 바로 위층까지 반복한다.
⑥ 출력층과 모든 은닉층의 에러와 델타 값이 구해졌으면 일반화된 델타 규칙에 의하여 연결 가중치를 수정한다. 모멘텀(Momentum)[11]을 사용할 경우에는 이를 포함시키며 바이어스를 사용할 경우에는 바이어스도 수정한다.

$$w(N)_{ij} = w(O)_{ij} + \alpha\,\delta_j a_i + \beta\triangle w_{ij}(O)$$
$$bias(N)_{ij} = bias(O)_{ij} + \alpha\,\delta_j \bullet 1 + \beta\triangle bias_{ij}(O)$$

⑦ ① ~ ⑥과정을 입력 패턴 모두에 대하여 반복한다.
⑧ ④과정을 인공신경망이 학습이 완전하게 될 때까지 반복한다.

11) 모멘텀(Momentum)이란 물리학 용어이다. 인공신경망의 연결 가중치를 수정하는 식에서 관성을 줌으로 학습하는 시간을 단축하고 학습 성능 향상을 위해 사용한다.

제3장 수급분석

3.1 기본 개념

수급분석은 주식시장의 자금이 유입되는 흐름을 중심으로 하여 시장상황의 동향이나 주가의 방향을 예측 하려는 기술적 분석[12]이라 볼 수 있다. 주가는 주식시장의 수요와 공급에 의하여 결정된다는 기본인식에서 출발하여 수요와 공급에 영향을 미치는 요인들을 정리하고 계량화함으로써 주식가격의 동향을 파악하고자 하는 것이다.

국내 증권사에서는 이런 추세에 맞혀서 투자주체별 순매수 데이터를 제공한다. 여기서는 대신증권 API를 이용하여 투자주체별 순매수 데이터를 사용하였다. 투자주체는 개인, 외국인, 기관, 금융, 보험, 투신, 은행, 기타금융, 연기금, 기타법인, 기타외인, 사모펀드, 국가지자체 순매수 데이터를 제공한다. 크게 개인, 외국인, 기관으로 분류 할 수 있고 그 외 기타법인, 기타외인은 이에 속하지 않는다. 기관은 증권, 보험, 투신, 은행, 기타금융, 연기금, 사모펀드, 국가지자체가 이에 속한다.

<그림 1>은 투자주체별 순매수 데이터로 만든 투자주체별 매매동향 프로그램 차트이다. 임의의 주식종목 2015년 10월 23일부터 2017년 10월 22일간의 수급분석 차트이다. ①번은 개인으로 주가와 반대방향으로 움직이고 있는 것을 볼 수 있다. ②, ③번은 외국인, 기관으로 주가와 비슷하게 움직이고 있다. 수급분석을 통해 개인, 외국인, 기관의 투자주체별 매매동향으로 주가와의 연관성이 있음을 알 수 있다.

12) 기술적 분석은 증권분석으로 주가와 거래량의 데이터를 근거로 과거 흐름의 패턴을 분석하여 미래의 변화 추세를 예측하는 방법이다.

<그림 1> 투자주체별 매매동향 프로그램 차트

3.2 투자주체 정의

이번 연구에서는 투자주체 개인, 기관, 외국인의 3가지 데이터만을 이용하기로 하였다.

개인 투자자(Individual Investors)는 적은규모의 자본을 갖고 주식을 투자하는 불특정한 다수를 의미한다. 개인 투자자들은 자금력 및 정보력 또는 투자기법에 기관 투자자에 비해서 부족함을 많이 가지고 있다. 따라서 주식시장의 움직임을 적극적으로 대응하지 못하는 투자자 집단이라고 할 수 있다. 이에 비해서 기관 투자자들은 많은 자금을 동원하여 전문적인 지식과 우수한 투자기법을 이용하여 주식시장에서 거대한 힘을 발휘하는 투자자 집단이라고 할 수 있다.

기관 투자자(Institutional Investors)는 개인 투자자에 대하여 상대적인

개념으로 볼 수 있는데 이들은 불특정한 다수인 개인으로부터 금융상품을 통하여 유입되는 자금에 자기자금을 더하여 전문적으로 운영하는 운영의 주체라 할 수 있다. 하지만 이러한 기관 투자자에 대하여 명확하게 정의가 내려지진 않은 상태이다.

외국인 투자자(Foreign Investors)는 외국인 투자 촉진법 제2조에 따르면 다음 중 어느 하나에 해당하는 것을 말한다.

첫째, 외국인이 이 법에 따라 대한민국법인(설립 중인 법인을 포함한다) 또는 대한민국 국민이 경영하는 기업의 경영활동에 참여하는 등 그 법인 또는 기업과 지속적인 경제관계를 수립할 목적으로 대통령령으로 정하는 바에 따라 그 법인이나 기업의 주식 또는 지분(이하 주식 등 이라한다)을 소유하거나

둘째, 다음의 어느 하나에 해당하는 자가 해당 외국인투자기업에 대부하는 5년 이상의 차관(최초의 대부계약 시에 정해진 대부기간을 기준으로 한다)

1) 외국인투자기업의 해외 모기업(母企業)
2) 1)의 기업과 대통령령으로 정하는 자본출자관계가 있는 기업
3) 외국투자자
4) 3)의 투자자와 대통령령으로 정하는 자본출자관계가 있는 기업

셋째, 외국인이 이 법에 따라 과학기술 분야의 대한민국법인(설립 중인 법인을 포함한다)으로서 연구인력 · 시설 등에 관하여 대통령령으로 정하는 기준에 해당하는 비영리법인과 지속적인 협력관계를 수립할 목적으로 그 법인에 출연하거나

넷째, 그 밖에 외국인의 비영리법인에 대한 출연으로서 비영리법인의 사업내용 등에 관하여 대통령령으로 정하는 기준에 따라 제27조에 따른 외국인투자위원회(이하 “외국인투자위원회” 라 한다))가 외국인 투자로 인정하는 것으로 정의 하고 있다.

제4장 주식 포트폴리오 구성

4.1 데이터 수집

대신증권에서 사이보스플러스로 제공하는 API을 사용하여 C# 프로그램 언어로 데이터 수집 프로그램을 개발하였고 KOSPI 200종목 중 2007년 1월 2일부터 2017년 7월 31일까지의 일별 시가, 종가 및 개인, 외국인, 기관의 수급 데이터가 존재하는 151종목을 수집하였다.

<표 1> 수집데이터 샘플

코드	일자	가격 데이터		수급 데이터		
		시가	종가	개인	외국인	기관
A005930	2007-01-02	183000	187500	-19379	31942	-7544
A005930	2007-01-03	187500	185500	44250	-18255	-28570
A005930	2007-01-04	187000	180000	82101	-43094	-50596
A005930	2007-01-05	182000	175500	220424	-128968	-89324
A005930	2007-01-08	175500	176500	192465	-219604	11867
A005930	2007-01-09	175500	169500	25109	-26588	3977
A005930	2007-01-10	171000	165500	122340	-53726	-72355
A005930	2007-01-11	167500	170000	97542	-51509	-37104
A005930	2007-01-12	170000	169500	-33996	46332	-12003
A005930	2007-01-15	169000	170500	-82074	218224	-132585
A005930	2007-01-16	169500	171000	-51364	-23528	8282
A005930	2007-01-17	171000	169000	3587	7113	-79207
A005930	2007-01-18	169500	166500	23888	-107289	6815

<표 1>은 삼성전자 수집데이터 샘플로 시가와 종가, 개인, 외국인, 기관 수급 데이터이다.

4.2 자기 조직화 지도 모델

4.2.1 데이터 정규화 방법

인공신경망 인공신경세포에서 연결 가중치(Weight)는 학습을 통해 입력 패턴(Pattern)과 유사해진다. 입력패턴들 중 어느 하나가 현저하게 클 경우 인공신경세포의 연결 가중치는 그 큰 값을 닮아 갈 것이고, 결국 그 인공신경세포의 연결 가중치만 커질 것이다. 따라서 모든 입력 패턴이 가장 큰 값에 반응해 버리는 경우가 생긴다. 머신러닝 학습을 하려면 <표 1>의 데이터 형식으로는 위와 같은 문제로 머신러닝 학습을 진행 할 수 없다. 머신러닝의 학습을 위해서 입력 패턴의 정규화(Normalization)[13] 과정을 진행하였다.

n기간 일별 가격데이터는 당일 종가에서 당일 시가를 빼고 그 값을 당일 시가로 나눠서 당일 수익률로 정규화 하였다.

$PL_i = (C_i - O_i) / O_i$, 일 i = 1, 2, …, n

PL_i : 당일 수익률, C_i : 당일 종가, O_i : 당일 시가

n기간 일별 수급데이터는 전체 기간 동안의 전체유통물량[14]을 구하고 전체유통물량으로 일별 정규화 데이터로 변환한다. 전체유통물량은 n기간 동안의 누적수급데이터의 MAX, MIN 데이터를 구해서 MAX에서 MIN을 뺀다. 당일 정규화 데이터는 당일 수급데이터로 전체유통물량을 나눠서 구한다.

$$S_i = \sum_{i=1}^{n} d_i \text{ , 일 } i = 1, 2, \cdots, n$$

13) 입력 패턴의 정규화(Normalization)란 패턴 정규화 과정으로 일부 패턴이 다른 것들 에 비해 현저하게 큰 경우를 데이터 변환을 거쳐 미연에 방지하는 것이다. 정규화 과정을 통해 입력 패턴 모두 공평한 경쟁에 참가할 수 있다.

14) 전체유통물량은 주식시장에서 유통되는 물량을 파악하기 위해 수급데이터로만 구한 전체 수급유통물량이다.

d_i : 당일 수급데이터

S_i : 당일까지 누적수급데이터

$T = MAX(S_i) - MIN(S_i)$, 일 i = 1, 2, …, n

T : 전체유통물량

$MAX(S_i)$: 당일까지 누적수급데이터들의 최대값

$MIN(S_i)$: 당일까지 누적수급데이터들의 최소값

$N_i = \frac{d_i}{T}$, 일 i = 1, 2, …, n

N_i : 당일 정규화 데이터

<표 2> 정규화 데이터

코드	일자	정규화 상태			
		수익률	개인	외국인	기관
A005930	2007-01-02	0.0246	-0.02471	0.052731	-0.01083
A005930	2007-01-03	-0.0107	0.056425	-0.03014	-0.04101
A005930	2007-01-04	-0.0374	0.10469	-0.07114	-0.07263
A005930	2007-01-05	-0.0357	0.28107	-0.21291	-0.12823
A005930	2007-01-08	0.0057	0.245419	-0.36253	0.017035
A005930	2007-01-09	-0.0342	0.032017	-0.04389	0.005709
A005930	2007-01-10	-0.0322	0.156	-0.08869	-0.10387
A005930	2007-01-11	0.0149	0.124379	-0.08503	-0.05326
A005930	2007-01-12	-0.0029	-0.04335	0.076487	-0.01723
A005930	2007-01-15	0.0089	-0.10466	0.360255	-0.19033
A005930	2007-01-16	0.0088	-0.0655	-0.03884	0.011889
A005930	2007-01-17	-0.0117	0.004574	0.011742	-0.1137
A005930	2007-01-18	-0.0177	0.03046	-0.17712	0.009783

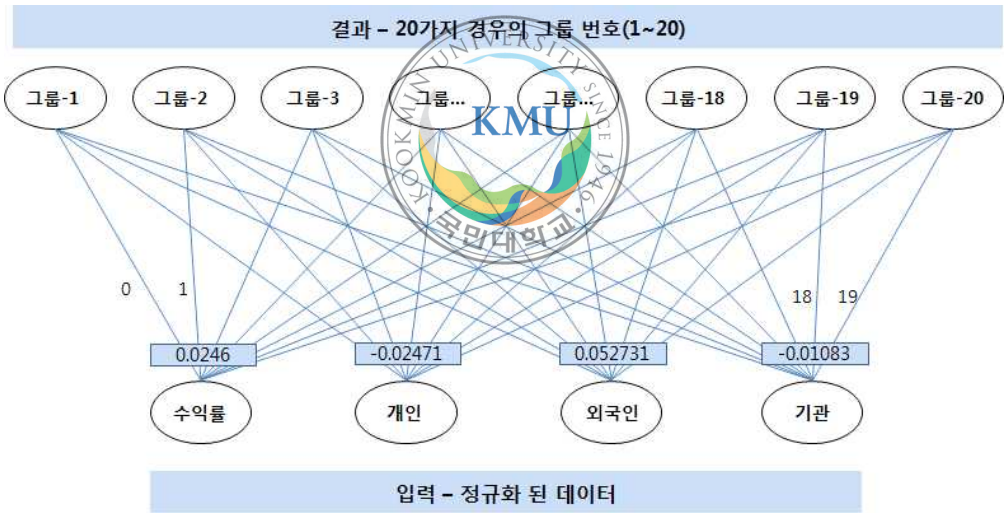
<표 2>는 수집된 데이터를 머신러닝 학습을 위해 정규화한 상태이다.

4.3.2 학습 방법

정규화 된 수익률, 개인, 외국인, 기관 수급 데이터를 인공신경망의 자기조직 지도 모델에 학습을 시키고 학습정보로 클러스터링 한다.

클러스터링 학습을 시키기 위해 2007년 1월 2일부터 2014년 12월 31일까지를 학습데이터로 사용하였고, 2015년 1월 1일부터 2017년 7월 31일까지를 클러스터링 하였다. 학습 데이터를 2007년 1월 2일부터 2014년 12월 31일까지 사용한 것은 역 전파 알고리즘 모델에서 학습 검증 테스트에 미래 데이터가 포함되지 않기 위한 것이다.

학습데이터 일수 : 1986일(75.71%)
클러스터링데이터 일수 : 2623일(100.0%)



<그림 2> 클러스터링 학습 모형

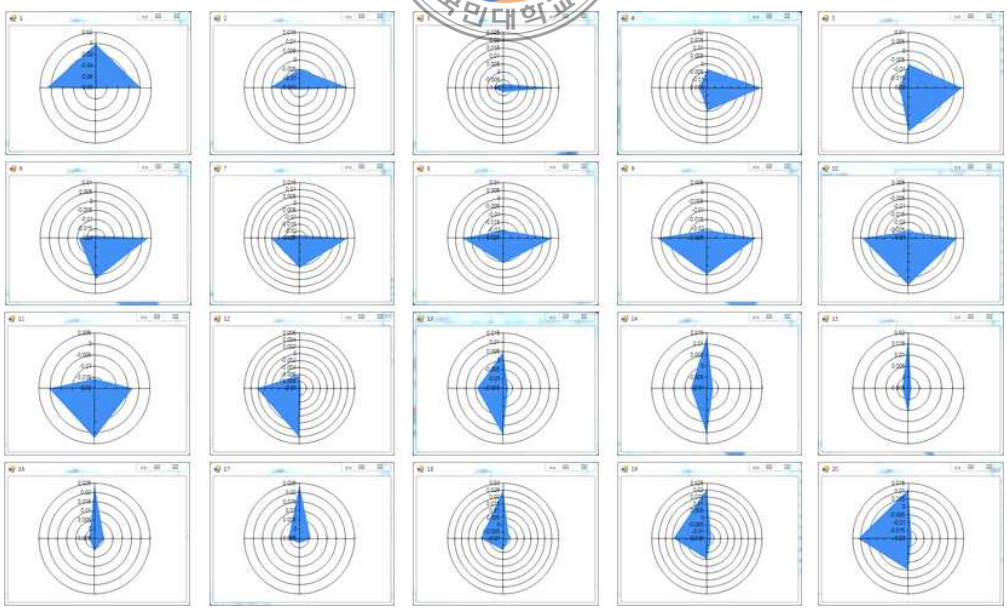
<그림 2>는 클러스터링 학습을 도식화 한 것이다.

Input Data는 4개이고 Output Data는 20개이며 Layer는 2개로 Hidden Layer는 없다. 출력은 승자 인공신경세포의 결정으로 <그림 2>처럼 그룹1부터 그룹20까지의 승자 인공신경세포가 출력된다.

<표 3> 클러스터링 결과 샘플

종목코드	날짜	수익률	개인	외국인	기관	그룹번호
A005930	2007-01-02	0.0081	-0.0083	0.002	-0.0007	20
A005930	2007-01-03	-0.0255	0.0189	-0.0012	-0.0027	7
A005930	2007-01-04	-0.0065	0.035	-0.0028	-0.0048	3
A005930	2007-01-05	-0.0214	0.094	-0.0083	-0.0084	3
A005930	2007-01-08	-0.0169	0.0821	-0.0141	0.0011	3
A005930	2007-01-09	-0.0017	0.0107	-0.0017	0.0004	3
A005930	2007-01-10	-0.0069	0.0522	-0.0034	-0.0068	3
A005930	2007-01-11	0.0052	0.0416	-0.0033	-0.0035	3
A005930	2007-01-12	0.022	-0.0145	0.003	-0.0011	19
A005930	2007-01-15	0	-0.035	0.014	-0.0125	13
A005930	2007-01-16	-0.0033	-0.0219	-0.0015	0.0008	12
A005930	2007-01-17	-0.0131	0.0015	0.0005	-0.0075	6
A005930	2007-01-18	-0.0067	0.0102	-0.0069	0.0006	3
A005930	2007-01-19	-0.0051	0.0204	-0.0007	-0.007	4

<표 3>은 클러스터링을 진행하고 나온 결과로 그룹번호는 승자 인공신경세포의 번호이다.



<그림 3> 20가지 경우 그룹 패턴의 평균 방사형 차트

<그림 3>은 클러스터링을 통해 그룹화한 20가지 결과들의 평균을 계산하여 그린 방사형 차트이다. 왼쪽 상단부터 오른쪽에서 아래로 이동하며 오른쪽 하단까지 1부터 20의 결과이다. 차트의 모양은 4방위로 위쪽은 수익률이고 오른쪽부터 시계 방향으로 개인, 외국인, 기관의 수급 데이터 순이다. 방사형 차트를 보면 일별 수익률과 수급 데이터의 비율로 의미 있는 패턴을 보여주고 있다. 클러스터링을 통해 데이터들의 정보가 의미 있도록 그룹화 할 수 있는 것을 알 수 있다.

4.3 오류역전파 모델

4.3.1 데이터 변환 방법

오류역전파 모델에서는 Input Data와 Output Data 패턴이 자기 조직화지도 모델 출력 데이터 패턴을 그대로 가져다 쓰는 것이 아니다. 오류역전파 모델에서는 Input Data는 60일 데이터 패턴을 입력하고 Target Data 패턴은 2가지의 패턴을 이용하여 상승과 하락을 입력해준다. 따라서 클러스터링에서 나온 데이터를 변환 해줘야한다.

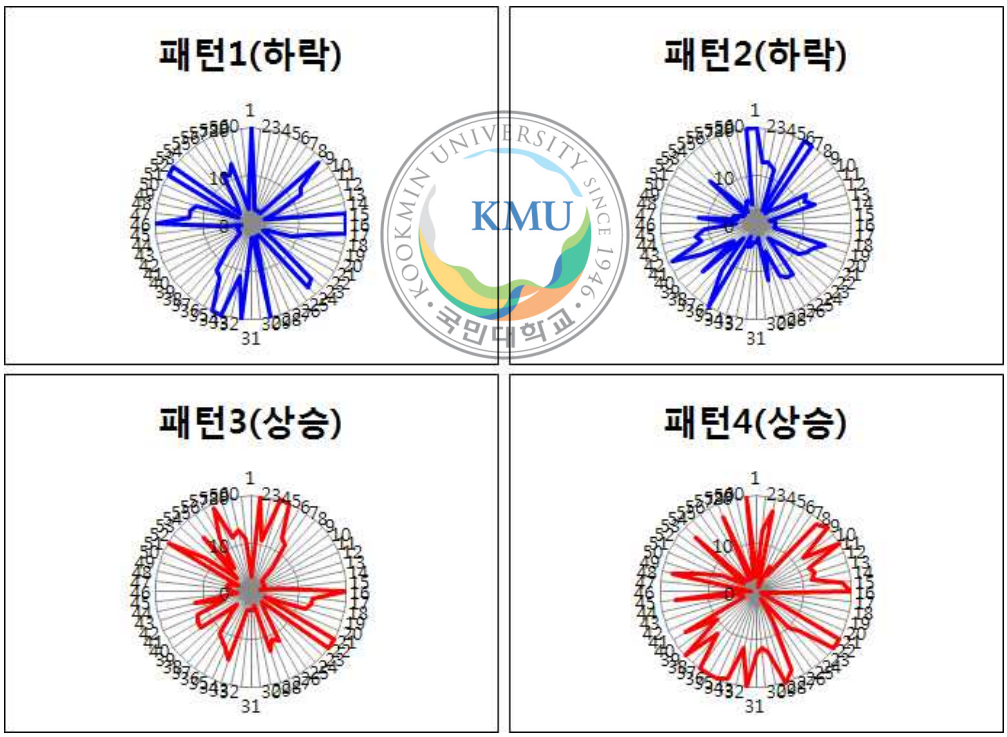
종목코드	날짜	그룹
A005930	2007-01-02	20
A005930	2007-01-03	7
A005930	2007-01-04	3
A005930	2007-01-05	3
A005930	2007-01-08	3
A005930	2007-01-09	3
A005930	2007-01-10	3
A005930	2007-01-11	3
A005930	2007-01-12	19
A005930	2007-01-15	13
A005930	2007-01-16	12
A005930	2007-01-17	6
A005930	2007-01-18	3
A005930	2007-01-19	4
A005930	2007-01-22	20
A005930	2007-01-23	20
A005930	2007-01-24	20
A005930	2007-01-25	12
A005930	2007-01-26	9
A005930	2007-01-29	4
A005930	2007-01-30	3
A005930	2007-01-31	6
A005930	2007-02-01	17
A005930	2007-02-02	18
A005930	2007-02-05	3
A005930	2007-02-06	3

5일 간격의 60일간 그룹 정보																	목표 구분
1일	2일	3일	4일	5일	6일	7일	8일	9일	10일	11일	12일	13일	14일	15일	...	60일	
20	7	3	3	3	3	3	3	19	13	12	6	3	4	20	...	4	2
3	3	3	19	13	12	6	3	4	20	20	20	12	9	4	...	13	2
12	6	3	4	20	20	20	12	9	4	3	6	17	18	3	...	13	2
20	20	12	9	4	3	6	17	18	3	3	3	3	20	3	...	3	2
3	6	17	18	3	3	3	3	20	3	3	20	11	20	20	...	3	2
3	3	3	20	3	3	20	11	20	20	13	12	7	3	3	...	6	2
3	20	11	20	20	13	12	7	3	3	3	3	3	2	6	...	11	1
13	12	7	3	3	3	3	3	2	6	20	13	13	12	3	...	4	1
3	3	3	2	6	20	13	13	12	3	20	20	3	4	5	...	3	1
20	13	13	12	3	20	20	3	4	5	12	11	13	3	4	...	20	2
20	20	3	4	5	12	11	13	3	4	4	4	4	15	13	...	20	1
12	11	13	3	4	4	4	4	15	13	12	11	6	4	13	...	20	1
4	4	4	15	13	12	11	6	4	13	13	12	6	12	3	...	12	1
12	11	6	4	13	13	12	6	12	3	4	4	5	5	3	...	16	1
13	12	6	12	3	4	4	5	5	3	20	12	4	15	6	...	13	2
4	4	5	5	3	20	12	4	15	6	13	19	12	13	11	...	8	2
20	12	4	15	6	13	19	12	13	11	4	12	4	5	4	...	12	2
13	19	12	13	11	4	12	4	5	4	3	3	13	6	3	...	5	2
4	12	4	5	4	3	3	13	6	3	4	5	5	4	20	...	7	2

<그림 4> 데이터 변환 방법

<그림 4>에서 왼쪽 데이터는 자기 조직화 지도 모델에서 나온 출력 정보이다. 오른쪽 데이터는 자기 조직화 지도 모델의 데이터를 5일 간격씩 60일 연속 데이터를 세로에서 가로 데이터로 변환하여 Input Data를 만든다. 이때 Target Data도 준비하는데 Target Data는 연속 데이터의 마지막 날짜의 다음날 시가와 20일 후 종가로 수익률을 구하여 다음 수식을 사용하여 데이터를 만들 수 있다. <그림 4>에서 오른쪽 데이터의 목표구분에 들어갈 데이터이다.

1(상승) : 수익률 > 0
2(하락) : 수익률 <= 0



<그림 5> Input Data 방사형 차트

<그림 5>는 Input Data의 상승, 하락의 샘플 데이터를 방사형 차트로 그린 것이다. 원 중심은 1부터 바깥쪽으로 갈수록 20에 가깝다. 상승 패턴을 보면 20에 가까운 데이터가 많은 것을 볼 수 있다. <그림 3>에서 번호가

클수록 가격이 상승률이 높은 것을 알 수 있는데 <그림 5>에서도 상승 패턴에 가격 상승률이 반영 되는 것을 볼 수 있다.

4.3.2 데이터 정규화 방법

머신러닝의 학습을 위해서 입력 패턴의 정규화 과정을 진행하였다.

모든 일별 그룹화 데이터를 정규화 한다.

$G_i = g_i * 0.01$, i 일 = 1, 2, …, n

g_i : 일별 그룹화 데이터
G_i : 정규화 된 일별 그룹화 데이터

목표구분 데이터를 정규화 한다.

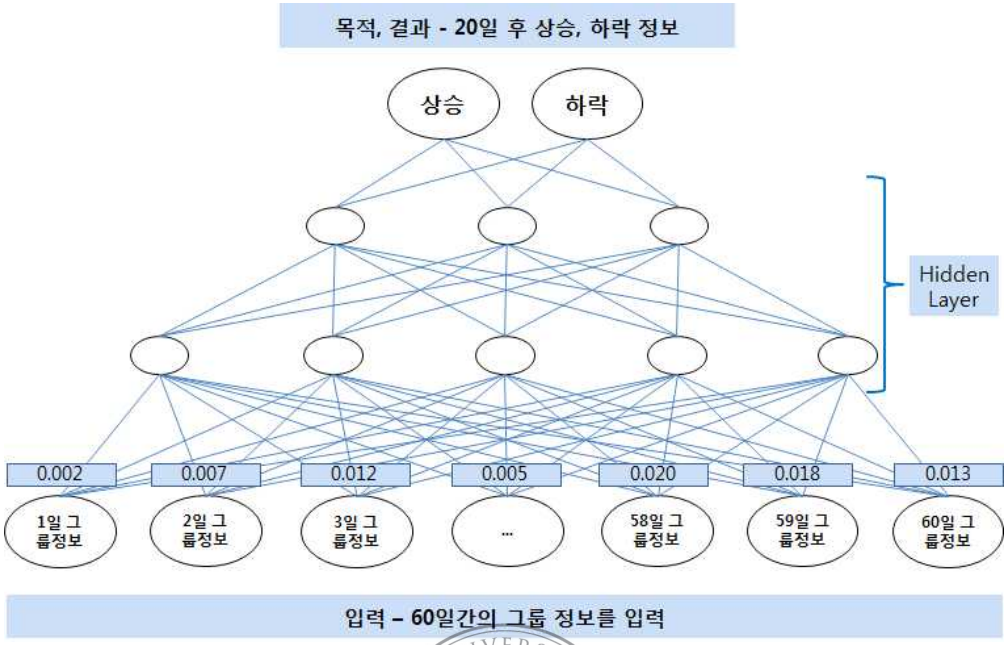
1 : [1, 0]
2 : [0, 1]

4.3.3 학습 방법

기계학습에 적용하기 위해 2007년 1월 2일부터 2014년 12월 29일까지를 학습데이터로 사용하였고, 2014년 12월 30일부터 2017년 7월 31일까지를 검증데이터로 사용하였다.

학습데이터 일수 : 1985일(75.68%)
검증데이터 일수 : 638일(24.32%)

각 종목별 5일 간격으로 연속되는 60일 데이터를 학습데이터로 만들면 386개의 학습데이터가 나온다. 검증데이터는 2년 7개월로 31개의 데이터를 사용하였다.



<그림 6> Classification 학습 모형

Input Data는 60개이고 Output Data는 2개이며 Layer는 6개로 Hidden Layer는 4개이다. 출력은 [1, 0] 또는 [0, 1] 예측으로 <그림 6>처럼 상승, 하락 결과가 출력된다.

4.4 포트폴리오 구성 방법

포트폴리오 구성은 시가총액 상위 10종목을 각각 오류역전파 모델에 학습 시키고 그 중에 매월 첫 거래일에 상승 예측 결과가 출력되는 종목들을 구성하여 진입하고 20일 후 청산하는 방식이다. 포트폴리오 구성을 매월 재구성 하고 자산 비중을 동일하게 리밸런싱(Rebalancing)[15] 하였다.

<표 4> 벤치마크(시가총액 상위 10종목)

코드	종목명	발행수량	2016년12월29일 종가	시가총액 (단위 : 백만)
A005930	삼성전자	129,768,494	1,802,000	233,842,826
A000660	SK하이닉스	728,002,365	44,700	32,541,706
A005380	현대차	220,276,479	146,000	32,160,366
A015760	한국전력	641,964,077	44,050	28,278,518
A012330	현대모비스	97,343,863	264,000	25,698,780
A035420	NAVER	32,962,679	775,000	25,546,076
A005490	POSCO	87,186,835	257,500	22,450,610
A055550	신한지주	474,199,587	45,250	21,457,531
A090430	아모레퍼시픽	58,458,490	321,500	18,794,405
A051910	LG화학	70,592,343	261,000	18,424,602

<표 4>와 같이 2016년 12월 29일 기준 종가와 현재 발행수량으로 시가총액을 계산하였다. 시가총액 상위 10종목을 벤치마크(Banchmark)[16]로 정하였다.

<표 5> 포트폴리오 구성

연월	종목코드	종목명	진입가	청산가	수익률	목표 구분	예측 구분
2015년1월	A015760	한국전력	42350	42600	0.26	1	1
	A051910	LG화학	179500	199000	10.5	1	1
	A055550	신한지주	44150	45600	2.94	1	1
	A090430	아모레퍼시픽	221100	268000	20.82	1	1

15) 리밸런싱(Rebalancing)이란 운용하는 자산의 처음 편입비중에서 어느 시점에서 변경된 편입비중을 재설정하는 것을 말한다.

16) 벤치마크(Banchmark)란 어떤 비교할 수 있는 것의 기준이 되는 것으로 그 값을 측정할 수 있는 기준을 말한다.

연월	종목코드	종목명	진입가	청산가	수익률	목표 구분	예측 구분
2015년2월	A005490	POSCO	255500	276500	7.86	1	1
	A005930	삼성전자	1365000	1437000	4.93	1	1
	A012330	현대모비스	248500	254500	2.08	1	1
	A015760	한국전력	42900	45250	5.13	1	1
	A090430	아모레퍼시픽	262500	278000	5.56	1	1
2015년3월	A035420	NAVER	662000	672000	1.18	1	1
	A051910	LG화학	233000	226500	-3.11	2	1
	A090430	아모레퍼시픽	285800	326000	13.69	1	1
2015년4월	A000660	SK하이닉스	45950	46950	1.84	1	1
	A005380	현대차	167500	172500	2.65	1	1
	A005490	POSCO	242000	260500	7.29	1	1
	A005930	삼성전자	1437000	1366000	-5.26	2	1
	A012330	현대모비스	247000	240500	-2.95	2	1
	A015760	한국전력	46000	47950	3.9	1	1
	A035420	NAVER	672000	686000	1.75	1	1
	A051910	LG화학	228000	274500	20	1	1
	A055550	신한지주	42100	44050	4.29	1	1
	A090430	아모레퍼시픽	333800	388400	15.98	1	1
2015년5월	A005490	POSCO	255500	240000	-6.38	2	1
	A015760	한국전력	46850	45850	-2.46	2	1
2015년6월	A000660	SK하이닉스	50600	42000	-17.27	2	1
	A005380	현대차	157500	135000	-14.57	2	1
	A005490	POSCO	244000	227000	-7.28	2	1
	A012330	현대모비스	221500	209000	-5.96	2	1
	A015760	한국전력	46300	45200	-2.7	2	1
	A051910	LG화학	253000	276500	8.93	1	1
	A090430	아모레퍼시픽	401000	419000	4.14	1	1
2015년7월	A005930	삼성전자	1268000	1230000	-3.32	2	1
	A015760	한국전력	46200	48200	3.99	1	1
	A090430	아모레퍼시픽	415500	411500	-1.29	2	1
2015년8월	A000660	SK하이닉스	36500	35800	-2.24	2	1
	A005490	POSCO	194000	190000	-2.39	2	1
	A055550	신한지주	42150	39550	-6.48	2	1
	A090430	아모레퍼시픽	412000	378500	-8.44	2	1
2015년9월	A005380	현대차	147000	164000	11.2	1	1
	A012330	현대모비스	204500	231500	12.83	1	1
	A015760	한국전력	47950	49000	1.85	1	1
	A051910	LG화학	240000	285500	18.57	1	1
	A055550	신한지주	39350	41400	4.86	1	1
	A090430	아모레퍼시픽	374000	385000	2.6	1	1
2015년10월	A012330	현대모비스	229000	238500	3.81	1	1
	A015760	한국전력	49050	51600	4.85	1	1
	A055550	신한지주	41700	42600	1.82	1	1
	A090430	아모레퍼시픽	385500	381500	-1.36	2	1

연월	종목코드	종목명	진입가	청산가	수익률	목표 구분	예측 구분
2015년11월	A005490	POSCO	182500	171000	-6.61	2	1
	A012330	현대모비스	238000	252000	5.53	1	1
	A051910	LG화학	303500	321000	5.42	1	1
	A090430	아모레퍼시픽	377500	412500	8.91	1	1
2015년12월	A005490	POSCO	170000	169500	-0.62	2	1
	A012330	현대모비스	246500	247500	0.07	1	1
	A051910	LG화학	318000	332000	4.06	1	1
	A055550	신한지주	41400	40150	-3.34	2	1
2016년1월	A000660	SK하이닉스	30550	27300	-10.93	2	1
	A005490	POSCO	167000	178500	6.53	1	1
	A012330	현대모비스	242000	258500	6.47	1	1
	A051910	LG화학	332500	295500	-11.42	2	1
	A055550	신한지주	39000	38600	-1.35	2	1
	A090430	아모레퍼시픽	415000	405500	-2.61	2	1
2016년2월	A000660	SK하이닉스	28100	31950	13.33	1	1
	A005380	현대차	132000	147500	11.38	1	1
	A005490	POSCO	178500	213500	19.22	1	1
	A005930	삼성전자	1152000	1220000	5.55	1	1
	A015760	한국전력	53700	58500	8.58	1	1
	A051910	LG화학	290000	308000	5.86	1	1
	A055550	신한지주	38250	38650	0.71	1	1
2016년3월	A015760	한국전력	59800	59800	-0.33	2	1
	A051910	LG화학	305000	323500	5.72	1	1
	A090430	아모레퍼시픽	375000	386000	2.59	1	1
2016년4월	A035420	NAVER	645000	677000	4.62	1	1
	A055550	신한지주	40300	41800	3.38	1	1
	A090430	아모레퍼시픽	386500	407500	5.09	1	1
2016년5월	A000660	SK하이닉스	27950	28700	2.34	1	1
	A005380	현대차	142000	139500	-2.09	2	1
	A005490	POSCO	237500	208000	-12.71	2	1
	A012330	현대모비스	263500	255000	-3.55	2	1
	A015760	한국전력	61300	62900	2.27	1	1
	A035420	NAVER	679000	720000	5.69	1	1
	A055550	신한지주	41400	39550	-4.78	2	1
	A090430	아모레퍼시픽	407000	417500	2.24	1	1
2016년6월	A000660	SK하이닉스	28850	31800	9.86	1	1
	A005380	현대차	137500	139000	0.76	1	1
	A005490	POSCO	206500	202000	-2.5	2	1
	A005930	삼성전자	1298000	1396000	7.2	1	1
	A012330	현대모비스	251500	258500	2.44	1	1
	A051910	LG화학	270000	257000	-5.13	2	1
	A055550	신한지주	39100	37450	-4.54	2	1
2016년7월	A005930	삼성전자	1427000	1507000	5.26	1	1
	A015760	한국전력	60300	61600	1.82	1	1
	A035420	NAVER	713000	710000	-0.75	2	1

연월	종목코드	종목명	진입가	청산가	수익률	목표 구분	예측 구분
	A051910	LG화학	259000	242500	-6.68	2	1
	A055550	신한지주	37150	40200	7.85	1	1
	A090430	아모레퍼시픽	435000	400000	-8.35	2	1
2016년8월	A012330	현대모비스	256000	259000	0.84	1	1
	A015760	한국전력	61200	58100	-5.38	2	1
	A055550	신한지주	40600	40550	-0.45	2	1
	A090430	아모레퍼시픽	384500	387000	0.32	1	1
2016년9월	A005380	현대차	133000	139000	4.17	1	1
	A005490	POSCO	230500	231500	0.1	1	1
	A012330	현대모비스	258000	280000	8.17	1	1
	A051910	LG화학	271500	237500	-12.81	2	1
	A055550	신한지주	41250	40700	-1.66	2	1
2016년10월	A005380	현대차	138500	140000	0.75	1	1
	A005930	삼성전자	1610000	1639000	1.47	1	1
	A012330	현대모비스	280500	274000	-2.64	2	1
	A051910	LG화학	243000	246500	1.11	1	1
2016년11월	A005380	현대차	141000	135000	-4.57	2	1
	A005490	POSCO	236000	258500	9.17	1	1
	A005930	삼성전자	1630000	1677000	2.54	1	1
	A012330	현대모비스	273000	251500	-8.18	2	1
	A015760	한국전력	49950	47600	-5.02	2	1
	A090430	아모레퍼시픽	350000	323500	-7.88	2	1
2016년12월	A000660	SK하이닉스	43000	45350	5.12	1	1
	A005380	현대차	133000	143500	7.54	1	1
	A005490	POSCO	249500	255500	2.07	1	1
	A035420	NAVER	790000	763000	-3.74	2	1
	A051910	LG화학	230000	261000	13.11	1	1
	A055550	신한지주	44550	45750	2.36	1	1
	A090430	아모레퍼시픽	328000	316500	-3.83	2	1
2017년1월	A000660	SK하이닉스	44750	53700	19.61	1	1
	A012330	현대모비스	264000	242000	-8.64	2	1
	A090430	아모레퍼시픽	315500	317000	-4.28	1	1
	A015760	한국전력	44200	42450	-2.89	2	1
	A035420	NAVER	778000	758000	0.77	2	1
	A055550	신한지주	45400	45900	0.14	1	1
2017년2월	A005490	POSCO	269000	283500	5.04	1	1
	A005930	삼성전자	1977000	1922000	-3.1	2	1
	A035420	NAVER	761000	776000	1.63	1	1
	A090430	아모레퍼시픽	318500	301000	-5.81	2	1
2017년3월	A005490	POSCO	291000	286500	-1.87	2	1
	A012330	현대모비스	258000	236000	-8.83	2	1
	A035420	NAVER	799000	864000	7.78	1	1
	A051910	LG화학	286000	299500	4.38	1	1
	A055550	신한지주	47500	47300	-0.75	2	1
2017년4월	A000660	SK하이닉스	50800	54000	5.95	1	1

연월	종목코드	종목명	진입가	청산가	수익률	목표 구분	예측 구분
	A005380	현대차	156500	144000	-8.29	2	1
	A005930	삼성전자	2070000	2231000	7.42	1	1
	A012330	현대모비스	242000	222000	-8.57	2	1
	A051910	LG화학	296000	274000	-7.74	2	1
	A055550	신한지주	46450	47550	2.03	1	1
2017년5월	A000660	SK하이닉스	55300	56400	1.65	1	1
	A012330	현대모비스	222000	273000	22.57	1	1
	A035420	NAVER	800000	842000	4.9	1	1
	A051910	LG화학	273500	303000	10.42	1	1
2017년6월	A000660	SK하이닉스	56700	68500	20.42	1	1
	A035420	NAVER	850000	853000	0.02	1	1
	A055550	신한지주	49400	49900	0.68	1	1
2017년7월	A000660	SK하이닉스	66500	64600	-3.18	2	1
	A012330	현대모비스	251000	255500	1.46	1	1
	A015760	한국전력	40700	44950	10.08	1	1
	A051910	LG화학	294000	323000	9.5	1	1

<표 5>는 2015년 1월부터 2017년 7월까지 매월 첫 거래일에 오류역전파 모델 인공신경망에서 예측 결과가 상승 분류한 종목들이다. 진입가격은 첫 거래일 시가이고 청산가격은 20일 후 종가이다.

수익률 계산은 다음과 같다.

$$PL = ((C - O) - O * 0.00015 - C * 0.00015 - C * 0.003) / O * 100 \quad \text{식(5)}$$

PL : 수익률
O : 진입가
C : 청산가

식(5)에서 진입, 청산 수수료는 0.015%, 제세금은 0.3%로 수익률을 계산하였다.

제5장 머신러닝 성능 평가 및 성과 분석

<표 6>은 오류역전파 모델 인공신경망으로 시가총액 상위 10종목을 각 종목별로 학습을 시키고 그 학습 데이터의 예측 결과이다. 학습 예측률 평균은 96.61%로 학습 결과는 상당히 높았다.

<표 6> 오류역전파 모델 학습 예측 결과

코드	종목 명	학습데이터수량	예측성공수	학습 예측률
A005930	삼성전자	386	326	84.46
A000660	SK하이닉스	386	377	97.67
A005380	현대차	386	366	94.82
A015760	한국전력	386	378	97.93
A012330	현대모비스	386	378	97.93
A035420	NAVER	386	380	98.45
A005490	POSCO	386	384	99.48
A055550	신한지주	386	377	97.67
A090430	아모레퍼시픽	386	379	98.19
A051910	LG화학	386	384	99.48

<표 7>은 검증 데이터의 기간별 예측 결과이다. 상승, 하락 종목의 종목 수는 시가총액 상위 10종목의 실제 상승, 하락한 종목 수이고 예측성공 수는 오류역전파 모델 예측 결과에서 실제 상승, 하락한 종목 중에 상승 또는 하락 예측 결과가 일치하는 종목 수이다. 연별로 본 예측률은 2015년 64.17%, 2016년 53.33%, 2017년 51.43%로 총 평균 예측률은 57.1%로 좋은 성과를 내고 있다.

<표 7> 검증 데이터 기간별 예측 결과

진입연월	상승 종목		하락 종목		전체 종목
	종목 수	예측성공수	종목 수	예측성공수	예측률(%)
2015년 1월	7	4	3	3	70
2015년 2월	6	5	4	4	90
2015년 3월	4	2	6	5	70
2015년 4월	8	8	2	0	80
2015년 5월	1	0	9	7	70

진입연월	상승 종목		하락 종목		전체 종목
	종목 수	예측성공수	종목 수	예측성공수	예측률(%)
2015년 6월	3	2	7	2	40
2015년 7월	3	1	7	5	60
2015년 8월	1	0	9	5	50
2015년 9월	8	6	2	2	80
2015년 10월	7	3	3	2	50
2015년 11월	5	3	5	4	70
2015년 12월	6	2	4	2	40
연별합계(평균)	59	36	61	41	(64.17)
2016년 1월	3	2	7	3	50
2016년 2월	7	7	3	3	100
2016년 3월	7	2	3	2	40
2016년 4월	6	3	4	4	70
2016년 5월	5	4	5	1	50
2016년 6월	6	4	4	1	50
2016년 7월	6	3	4	1	40
2016년 8월	8	2	2	0	20
2016년 9월	7	3	3	1	40
2016년 10월	5	3	5	4	70
2016년 11월	3	2	7	3	50
2016년 12월	7	5	3	1	60
연별합계(평균)	70	40	50	24	(53.33)
2017년 1월	6	3	4	1	40
2017년 2월	7	2	3	1	30
2017년 3월	6	2	4	1	30
2017년 4월	4	3	6	3	60
2017년 5월	8	4	2	2	60
2017년 6월	5	3	5	5	80
2017년 7월	6	3	4	3	60
연별합계(평균)	42	20	28	16	(51.43)
총합(평균)	171	96	139	81	(57.10)

<표 8>에서 개별종목의 월별 수익률을 이용하여 포트폴리오의 월별 수익률과 복리 수익률을 구하였다.

동일배분 수익률은 다음과 같이 계산하였다.

$PL_i = P_i * 1/n$, 종목 i = 1, 2, …, n

$SPL = \sum_{i}^{n} PL$

P_i : 개별 종목 수익률
PL_i : 개별 종목 동일배분 수익률
SPL : 개별 종목 동일배분 수익률의 합

복리 수익률은 다음과 같이 계산하였다.

$CI = (1+PL_1) * (1+PL_2) * \cdots * (1+PL_n)$

PL_i : 개별 종목 동일배분 수익률, 월 i = 1, 2, …, n
CI : 복리 수익률

<표 8> 포트폴리오 기간별 수익률

진입 연월	동일배분 수익률(%)				복리 수익률(%)			
	Port folio	Bench mark	KOSPI 200	KOSPI	Port folio	Bench mark	KOSPI 200	KOSPI
2015-01	8.63	4.03	2.66	1.92	8.63	4.03	2.66	1.92
2015-02	5.11	1.99	1.79	2.59	14.18	6.10	4.51	4.56
2015-03	3.92	0.21	0.95	1.16	18.66	6.32	5.50	5.77
2015-04	4.95	4.95	4.72	5.52	24.53	11.58	10.48	11.60
2015-05	-4.42	-7.06	-4.22	-2.62	19.03	3.70	5.82	8.68
2015-06	-4.96	-3.02	-2.06	-0.94	13.13	0.57	3.64	7.66
2015-07	-0.21	-3.88	-2.88	-1.84	12.89	-3.33	0.65	5.68
2015-08	-4.89	-4.63	-4.48	-4.18	7.38	-7.81	-3.85	1.26
2015-09	8.65	4.34	2.09	1.47	16.67	-3.81	-1.84	2.74
2015-10	2.28	4.68	5.21	3.6	19.32	0.69	3.27	6.44
2015-11	3.31	0.4	-0.32	-0.2	23.28	1.09	2.95	6.22
2015-12	0.04	-0.31	-1.73	-1.76	23.33	0.77	1.17	4.35
2016-01	-2.22	-3.15	-3.01	-2.17	20.59	-2.40	-1.88	2.09
2016-02	9.23	5.43	3.1	2.01	31.73	2.89	1.16	4.14
2016-03	2.66	2.82	2.97	2.6	35.23	5.80	4.17	6.85
2016-04	4.36	0.85	-0.16	-0.03	41.13	6.69	4.00	6.82

진입 연월	동일배분 수익률(%)				복리 수익률(%)			
	Port folio	Bench mark	KOSPI 200	KOSPI	Port folio	Bench mark	KOSPI 200	KOSPI
2016-05	-1.32	-1.58	-0.5	-0.42	39.26	5.01	3.48	6.36
2016-06	1.16	0.95	-0.16	-1.04	40.87	6.01	3.32	5.26
2016-07	-0.14	2.19	2.82	2.21	40.67	8.33	6.23	7.59
2016-08	-1.17	2.57	1.42	0.38	39.03	11.12	7.73	7.99
2016-09	-0.41	0.98	1.66	1.58	38.46	12.20	9.52	9.70
2016-10	0.17	-1.01	-1.38	-2.37	38.70	11.06	8.01	7.10
2016-11	-2.32	-2.96	-0.94	-1.26	35.48	7.78	7.00	5.75
2016-12	3.23	2.53	1.91	1.86	39.86	10.51	9.04	7.72
2017-01	0.79	1.9	3.34	2.24	40.95	12.61	12.68	10.13
2017-02	-0.56	0.36	0.33	0.79	40.17	13.02	13.06	11.00
2017-03	0.14	2.55	3.57	2.94	40.36	15.90	17.10	14.25
2017-04	-1.53	-2.4	2.02	1.82	38.21	13.12	19.47	16.33
2017-05	9.89	6.9	5.35	5.92	51.88	20.92	25.85	23.22
2017-06	7.04	-0.37	2.57	2.02	62.57	20.48	29.08	25.70
2017-07	4.47	2.04	0.52	0.14	69.83	22.94	29.76	25.88
평균	18.02	7.19	8.76	7.71				
총합	55.88	22.3	27.18	23.9	69.83	22.94	29.76	25.88

<표 8>에서 동일배분 수익률의 KOSPI200[17] 지수 연평균 수익률 8.76%, KOSPI[18] 지수 연평균 7.71%로 2015년, 2016년, 2017년도는 1년 만기국채 수익률의 3년 평균 1.531%보다 수익률이 높았던 연도였다. 하지만 포트폴리오 연평균 수익률은 18.02%로 KOSPI200, KOSPI 지수보다 더 좋은 성과를 내었다. 벤치마크의 연평균 수익률은 7.19%로 벤치마크 보다 포트폴리오 수익률이 훨씬 뛰어난 성과를 보여주었다.

복리 수익률도 벤치마크 수익률은 22.94%, KOSPI200 지수는 29.76%, KOSPI 지수는 25.88% 비슷한 수익률을 보여준 반면에 포트폴리오 수익률이 69.83%로 복리 수익률에서도 뛰어난 성과를 내었다.

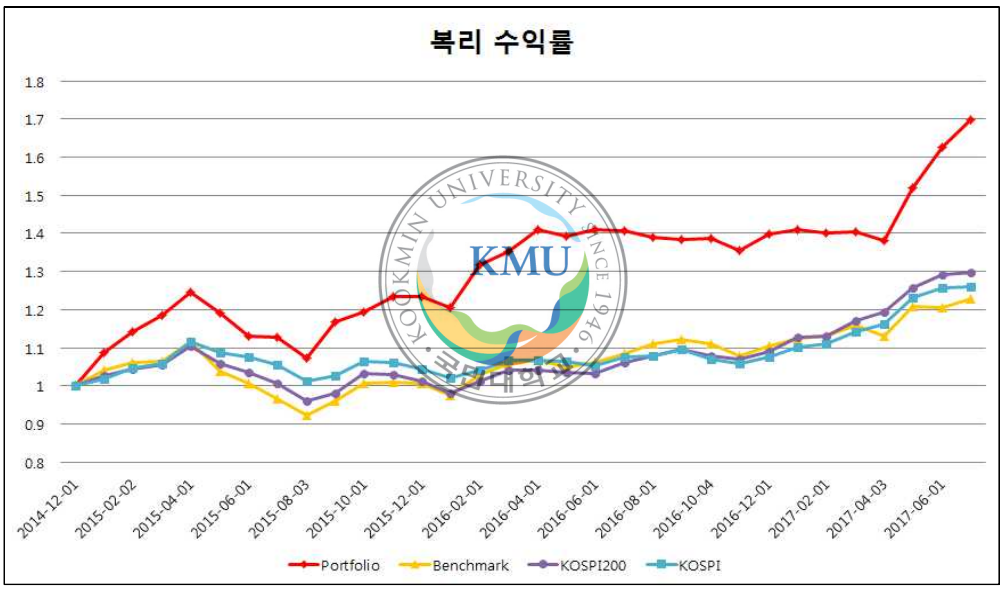
포트폴리오의 동일배분 수익률과 복리 수익률의 차이는 13.95%로 복리 수익률이 높다. 이것은 손실이 수익 보다 크지 않다는 것이다. 안정적인

17) KOSPI200란 한국 주식 종목 중 대표하는 주식 200개 종목을 특정시점 1990년 1월 3일 시가총액을 비교해 지수화한 것이다.

18) KOSPI란 한국증권거래소에 상장된 모든 주식을 전체 시장의 흐름을 파악하기 위해 전체의 주가를 산출해 나타내는 종합주가지수이다.

투자라고 볼 수 있다.

<그림 7>은 복리 수익률 차트이다. 차트를 보면 벤치마크, KOSPI200 지수, KOSPI 지수의 수익률은 비슷한 움직임을 보이고 있다. 반면 포트폴리오 수익률 성과는 큰 차이가 나는 것을 한눈에 보여주고 있다. 하지만 2015년 5월부터 8월까지 하락장일 때 포트폴리오 수익률도 같이 내려간 것은 체계적 위험(Systematic Risk)[19]에 대한 대비는 못하고 있다는 것을 알 수 있다. 이것은 이 포트폴리오 연구의 단점이라 할 수 있고 앞으로 보완이 더 필요한 부분이다.



<그림 7> 복리 수익률 차트

<표 9>는 포트폴리오와 벤치마크, KOSPI200, KOSPI 지수의 성과를 분석한 결과이다. 벤치마크와 KOSPI200, KOSPI 지수의 결과는 비슷하였다. 연평균 수익률은 <표 8>의 포트폴리오의 동일배분 수익률 평균을 연율화[20] 작업을 한 것이다. 포트폴리오는 연평균 수익률에서도 벤치마크보다

19) 체계적 위험(Systematic Risk)이란 증권시장 전반에 미치는 영향으로 분산투자로도 위험을 줄일 수 없는 투자 위험이다. 원인으로는 사회적, 정치적, 경제적 조건으로 들 수 있다.

20) 연율화는 기간별 기준으로 본 수익률을 연별 기준으로 수익률을 변환하는 것이다.

우수한 성적을 보이고 있다. MDD(Max Draw Down)는 최고점 대비 최대 손실 폭으로 포트폴리오는 14.85%이고 벤치마크는 13.96%로 비슷한 결과이지만 수익률 대비 비교를 하면 포트폴리오가 좋은 성적이라고 할 수 있다. 표준편차는 수익률 평가할 때 위험(변동성) 기준으로 보며 샤프지수(Sharpe Ratio)[21] 계산할 때 이용하고 있다. 샤프지수를 산출 할 때 무위험채권이자율은 1년 만기국채 수익률을 사용 했다. 1년 만기국채 수익률은 2015년 1.698%, 2016년 1.433%, 2017년 7월 기준 1.464%로 평균 1.531%이다. 포트폴리오 샤프지수는 1.436이며 벤치마크 샤프지수는 0.645로 벤치마크의 2배 이상이며 KOSPI200, KOSPI 지수 보다는 샤프지수가 크게 웃돌았다. 샤프지수가 1.436이라는 것은 해당 위험자산에 1단위 투자를 늘릴수록 이자 대비 1.436만큼의 초과 수익률을 낸다는 의미다. 초과 수익률이 발생하기 때문에 포트폴리오 복리 수익률이 69.83%로 높게 나타난 것이다.

<표 9> 성과 분석

	Portfolio	Benchmark	KOSPI200	KOSPI
동일배분 수익률(%)	55.88	22.30	27.18	23.90
복리 수익률(%)	69.83	22.94	29.76	25.88
연평균 수익률(%)	21.63	8.63	10.52	9.25
MDD(%)	14.85	13.96	9.82	10.10
표준편차	0.140	0.110	0.089	0.079
샤프지수	1.436	0.645	1.009	0.974

벤치마크의 Buy and Hold[22] 수익률은 27.37%로 벤치마크 자신보다는 수익률이 높았지만 포트폴리오 수익률보다는 낮았다.

21) 샤프지수(Sharpe Ratio)는 성과를 평가하는 지표로 표준편차를 사용한다. 위험자산에 투자하여 얻은 초과 수익률이라 볼 수 있다. 계산은 수익률에서 무위험 채권 이자율을 빼고 그 값을 수익률의 표준편차로 나누어 구한다.
22) Buy and Hold는 매수 후 보유로 최종 성과를 비교하기 위한 한 방법이다.

제6장 결론 및 향후 연구

6.1 결론

본 연구에서는 머신러닝과 수급분석을 활용한 주식 포트폴리오 구성 연구를 진행하였다.

오류역전파 모델 인공신경망을 이용한 학습 데이터의 예측률은 96.61%로 상당히 높았고 검증 데이터의 예측률은 57.1% 생각보다 낮았지만 과거 인공신경망 연구들과 비교하면 괜찮은 결과이다. 자기 조직화 지도 모델 인공신경망의 그룹화의 성능 평가는 오류역전파 모델 인공신경망의 결과로 판단할 수 있다. 자기 조직화 지도 모델의 그룹화 결과가 안 좋았다면 오류역전파 모델의 학습 결과도 좋지 않았을 것이기 때문이다. 이렇게 머신러닝의 학습 성능평가는 이전 연구들에 비해 학습이 잘 되어진 걸로 판단된다. 이유는 포트폴리오 성과에서도 볼 수 있다. 포트폴리오 성과 분석은 시가총액 상위 10종목인 벤치마크보다도 수익률이 2배 이상 좋았으며 KOSPI200 지수와 KOSPI 지수로 시장 수익률 비교를 하여도 포트폴리오 수익률이 상당히 좋았다.

수익률은 매월 포트폴리오 구성을 하고 자산을 동일 비중으로 리밸런싱한 결과이다. 매월 포트폴리오 구성할 때 동일 종목은 청산을 하지 않고 리밸런싱 하는 방법으로 시스템을 구현한다면 더욱 좋은 성과가 예상된다. 따라서 실제 거래 적용도 가능해보인다.

결과적으로 본 연구의 의의는 머신러닝과 수급분석 데이터를 활용해서 포트폴리오 구성하여 실전에서 성과를 내는 주식 시스템 트레이딩을 할 수 있는 주식 로보-어드바이저 프로그램 개발의 첫 걸음을 내딛었다는 것에 있다.

6.2 향후 연구

논문을 준비하며 아쉬웠던 부분이 준비 했던 151종목 전체를 학습 시키고 예측 하지 못한 부분이다. 전체 학습을 시도하지 않은 것은 아니다. 하지만 학습과정에서 데이터의 양이 많은 부분과 양이 많으니 비정상적인 데이터도 많이 포함 되어 있었을 것이다. 입력 데이터 정규화 방법과 데이터의 필터 방법을 연구 한다면 더 좋은 연구의 성과를 얻을 수 있을 것으로 보인다.

이번 연구를 통해 여러 가지 연구 방법들이 생각났다.

① 오류역전파 모델 인공신경망 학습의 60일 입력 데이터를 90, 120일 등으로 늘려 포트폴리오 구성기간을 늘리는 방법.
② 주식 재무제표 데이터를 이용하여 3개월 마다 포트폴리오 구성하는 방법.
③ KOSPI200 선물 분봉 수급 데이터에 60분 학습시키고 61분 진입하여 종가 청산 하는 방법.
④ 자기 조직화 지도 모델 학습의 그룹화 과정 데이터에 기술적 분석 데이터를 넣는 방법.

연구를 하면서 다양한 방법들이 생각나지만 실력과 시간이 부족하여 다음 연구 과제로 진행 하여야할 것 같다.

참 고 문 헌

김선웅, 안현철, “Support Vector Machines와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발”, 지능정보연구, 16권, 제1호, 2010, pp.71-92.

문준철, 강성수, 김준호. “투자주체별 투자행태 및 투자성과에 관한 연구”, 국제회계연구, 65, 2016, 155-178.

박성철, 김선웅, 최홍식. “SVM을 이용한 시스템트레이딩전략의 선택모형”, 지능정보연구, Vol.20(2), 2014, pp.59-71.

이상원, 『Turbo C로 길들이는 학습하는 기계 신경망』, 도서출판 Ohm사, 1993, pp.267-282, pp.340, pp.375-381.

Barber, S. (2007), AI: neural network for beginners(part 1 of 3), [http://www.codeproject.com/Articles/16419/AI-Neural-Network-for-beginners-Part-of]

Dreiseitl, S., Machado, L.O. (2002), “Logistic regression and artificial neural network classification models: a methodology review”, Journal of Biomedical Informatics, Vol. 35, Issues 5-6, pp. 352-359.

Heaton, J. (2012), 『Introduction to the math of neural networks』, Heaton Research, Inc., Kindle Edition

Singh, A. (2010), Neural networks, [http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/NeuralNetworks.pdf]

Abstract

A Study on Stock Portfolio Modeling using Machine Learning and Supply-Demand Analysis

by Kim, Kyung-Mock

Department of Trading System,
Graduate School of Business IT,
Kookmin University, Seoul, Korea

Using Self-organizing map, which is one of the most popular neural network models, this study first groups the demand-and-supply analysis data. With this grouping data the study trains an artificial neural network via back-propagation to develop a monthly portfolio.

For the analysis of portfolio performance, the study calculates the KOSPI200 index and KOSPI index returns, and analyzes the model performance using portfolio returns, compound returns, average annual returns, MDD, standard deviation, Sharpe ratio, and buy and hold strategy of the benchmark.

The results of the performance analysis show that the study's model portfolio doubles its benchmark and surpasses the market returns. The model portfolio's MDD and standard deviation are similar to those of the benchmark. The Sharpe ratio is better than that of benchmark and market.

The study proposes a method to develop a program for portfolio construction using machine learning techniques and demand-supply analysis. The results of the study show that it can be applied to developing an actual stock investment program.

Keywords : Artificial Intelligence, Supply-Demand Analysis, Robo-Advisor, Trading System, Stock Returns, KOSPI200, KOSPI