ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 March: 27(1): 65~82 http://dx.doi.org/10.13088/jiis.2021.27.1.065

투자자별 거래정보와 머신러닝을 활용한 투자전략의 성과

김경목
국민대학교 비즈니스IT전문대학원
(seobi197805@naver.com)

김선웅
국민대학교 비즈니스IT전문대학원
(swkim@kookmin.ac.kr)

최흥식
국민대학교 비즈니스IT전문대학원
(hschoi@kookmin.ac.kr)

주식시장에 참여하는 투자자들은 크게 외국인투자자, 기관투자자, 그리고 개인투자자로 구분된다. 외국인투자자 같은 전문투자자 집단은 개인투자자 집단과 비교하여 정보력과 자금력에서 우위를 보이고 있으며, 그 결과 시장 참여자들 사이에는 외국인투자자들이 좋은 투자 성과를 보이는 것으로 알려져 있다. 외국인 투자자들은 근래에는 인공지능을 이용한 투자를 많이 하고 있다.

본 연구의 목적은 투자자별 거래량 정보와 머신러닝을 결합하는 투자전략을 제안하고, 실제 주가와 투자자별 거래량 데이터를 이용하여 제안 모형의 포트폴리오 투자 성과를 분석하는 것이다. 일별 투자자별 매수 수량과 매도 수량 정보는 한국거래소에서 공개하고 있는 자료를 활용하였으며, 여기에 인공신경망을 결합하여 최적의 포트폴리오 전략을 도출하고자 하였다.

본 연구에서는 자기 조직화 지도 모형 인공신경망을 이용하여 투자자별 거래량 데이터를 그룹화하고 그룹화한 데이터를 변환하여 오류역전파 모형을 학습하였다. 학습 후 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 투자자별 거래정보 분석을 활용한 포트폴리오 구성 프로그램 개발의 방향을 제시하였고 실제 주식 투자를 위한 프로그램 개발에 활용할 수 있음을 보였다.

주제어 : 투자자별 거래량 정보, 기계학습, 로보어드바이저, 트레이딩시스템

논문접수일 : 2020년 12월 29일 논문수정일 : 2021년 2월 19일 게재확정일 : 2021년 3월 8일
원고유형 : 일반논문 교신저자 : 최흥식

1. 서론

1.1. 연구의 배경 및 목적

2016년 이세돌과 알파고의 딥마인드 챌린지 매치(Google Deepmind Challenge Match)는 사람들의 인식에 인공지능(Artificial Intelligence)이라는 뜨거운 이슈를 만들었다. 금융 분야에서도 로보어드바이저(Robo-advisor)라는 인공지능형 자산관리 프로그램이 급속도로 도입되기 시작하였다. 양적 성장에 비해서 투자 성적은 좋지 않은 편으로 시류에 편승한 나머지 본질보다는 인공지능이라는 외형 포장에 치중한 결과이다. 로보어드바이저와 같은 인공지능 기술을 활용한 투

자전략의 성과는 주가에 영향을 미치는 무수히 많은 변수들 중 어떤 입력 변수를 선택할 것인가에 달려 있다.

본 연구에서는 주가와 더불어 투자 주체들의 거래량 정보를 입력변수로 선택하였다. Daigler and Wiley(1999)는 대규모 기관투자자와 같은 차별화된 정보력을 가진 투자자들의 거래량이 주가의 변동성에 큰 영향을 미치고 있음을 실증분석으로 입증하였다. 이러한 투자자별 거래량의 정보효과(information effect)는 자본시장이 완전 개방된 우리나라와 같은 경우, 외국인 투자자의 거래량 영향력은 더 막대할 것으로 판단된다. 많은 연구들은 선진 투자기법으로 무장한 외국인 투자자들이 뛰어난 정보력을 바탕으로 한국 주식시장에서 큰 영향력을 행사하고 있음을 실증적으로 보여주고 있다(Cho and Lee, 2001; Ko and Lee, 2003; Kim and Ahn, 2010; Kim, 2013; Kim and Choi, 2015).

본 연구는 투자자별 거래정보와 머신러닝 기법을 활용하는 로보어드바이저 모형을 제안하고, 국내 상장 주식에 분산 투자하는 포트폴리오를 산출하고 그 수익성을 실증 분석하고자 한다.

1.2. 연구의 내용

머신러닝 학습을 위한 자기 조직화 지도 모형 인공신경망을 이용하여 외국인투자자, 기관투자자, 그리고 개인투자자별 거래량 데이터와 일별 주가 수익률 자료를 20가지 경우로 군집화 한다. 투자자별 거래량 데이터와 주식의 일별 수익률은 정규화 작업을 진행한다. 그룹화한 데이터는 오류역전파 모형 인공신경망에 사용할 수 있게 5일 간격으로 60일 연속 데이터의 변환 작업을 진행한다. 변환 작업으로 만들어진 데이터는 오류역전파 모형 인공신경망에 적용하기 위해 정규화 작업을 진행한다. 오류역전파 모형 인공신경망을 학습하고 검증데이터를 사용하여 월별 포트폴리오 구성을 진행한다.

오류역전파 모형 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, Maximum Drawdown(MDD), 표준편차, 샤프지수(Sharpe Ratio) 등을 사용한다. 샤프지수의 계산에서 무위험자산 수익률은 1년 만기 국채 수익률을 이용한다. 벤치마크의 Buy And Hold 수익률도 분석한다. 실제 투자에서 발생하는 매수, 매도 시의 수수료와 세금도 고려한다.

2. 이론적 배경

2.1. 자기 조직화 지도 모형

자기 조직화 지도(Self Organizing Map) 모형은 인공신경망(Artificial Neural Network)으로서 무감독(Unsupervised Learning) 학습에 의한 클러스터링(Clustering)을 수행하는 모형으로 1984년 이후 코호넨(Teuvo Kohonen)에 의해 소개된 모형이다(Kohonen, 1990; Dreiseitl and Machado, 2002).

면 개념을 사용하여 면내의 인공신경세포(Artificial

Neuron)들 간의 경쟁을 구현하며 모든 연결은 아래서 위로 가는 비회귀 인공신경망이다. 다층으로의 확장도 가능하지만 일반적으로 단층을 많이 사용한다. 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.

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

한편, Kim and Kim(2020)은 모의실험을 통해 자기조직화지도의 노드의 개수가 최적의 군집의 개수를 찾는데 있어 중요한 요인 중에 하나라는 사실을 밝혔다.

2.2. 오류역전파 모형

인공신경망으로서 감독학습(Supervised Learning)에 의한 분류(Classification)를 수행하는 모형이다. 오류역전파(Back-propagation) 모형의 핵심은 다음과 같다. 인공신경망으로서 감독학습(Supervised Learning)에 의한 분류(Classification)를 수행하는 모형이다(Olson and Delen, 2008; Schmidhuber, 2015).

'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 인공신경세포에 있는 것만 아니라 그것에 달린 더 아래층 인공신경세포에도 있기 때문에 그들 모두에 책임을 물어 연결 가중치를 수정한다. 이러한 잘못에 대해 연대 책임을 묻기 위해서 출력 층에서 발생하는 에러를 아래층으로 역전파 시키는 것을 오류역전파 모형이라고 부른다.

Jeong and Yun(1998)은 인공신경망모델인 오류역전파 모형으로 선형모형이 인식하지 못하는 패턴이 주가의 움직임에 나타나며 패턴의 인식이 주가예측에서 의미를 가질 수 있는 가능성을 보여주었다.

2.3. 투자자별 거래량 정보

투자자별 거래량 분석은 주식시장의 주가 흐름에 영향을 미치는 투자 주체들의 거래량 정보를 분석하여 향후 주가의 방향을 예측하려는 기술적 분석이다. 특히 우리나라 주식 시장 참여자들 사이에서는 외국인투자자 같은 큰 손들의 영향력이 막대한 것으로 알려지고 있으며, 대부분의 실증 분석 연구 결과들도 외국인투자자들의 정보력 우위를 지지하고 있다. Oh and Hahn(2008)은 외국인과 내국인 투자자 사이의 상대적 정보우위성을 조사한 결과, 6년간의 외국인투자자의 총 수익률은 82.6%로 내국인투자자의 21.9%보다 60.7%포인트나 높은 것으로 나타나 외국인투자자들의 자산배분전략의 우위성을 보여주었다. Kwark and Jun(2013)은 1998년부터 2010년까지의 투자자별 투자 손익 비중을 분석한 결과 외국인투자자의 손익 비중이 1.10으로 개인투자자 0.83, 기관투자자 1.08보다 높았다. 특히, 주가 상승기에서 뿐만 아니라 주가 하락 기간에서도 외국인투자자의 투자성과가 상대적으로 우월함을 보여주었다. Park(2014)는 외국인투자자들의 투자행태를 분석한 결과 1999년부터 2013년까지의

〈Table 1〉 Sample Data

Code	Date	Price data		Supply-Demand Data		
		Open	Close	Individual	Foreign	Institutional
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

분석 기간 동안 전반적으로 유의한 양의 투자성과가 나타났고, 암묵적 거래비용에서도 낮은 거래비용이 나타남으로써 외국인투자자의 양의 투자성과에 대한 연구 결과를 지지하였다.

국내 증권사들은 투자자들에게 매일매일 상장주식의 투자자별 순매수 데이터를 제공한다. 본 논문에서는 대신증권 API를 이용하여 투자자별 순매수 데이터를 사용하였다. 투자 주체는 개인, 외국인, 기관, 금융, 보험, 투신, 은행, 기타금융, 연기금, 기타법인, 기타외인, 사모펀드, 국가지자체 순매수 데이터를 제공한다. 크게 개인, 외국인, 기관으로 분류할 수 있고 그 외 기타법인, 기타외인은 이에 속하지 않는다. 기관은 증권, 보험, 투신, 은행, 기타금융, 연기금, 사모펀드, 국가지자체가 이에 속한다(Yi and Lee, 2004; Moon et al., 2016).

3. 주식 포트폴리오 구성

3.1. 데이터 수집

대신증권에서 사이보스플러스로 제공하는 API을 사용하여 C# 프로그램언어로 데이터 수집 프로그램을 개발하였고 KOSPI 200종목 중 2007년 1월 2일부터 2017년 7월 31일까지의 일별 시가, 종가 및 개인, 외국인, 기관의 순매수 데이터가 존재하는 151종목을 수집하여 오픈소스 DBMS의 일종인 SQLite에 담아 연구하였다. 오픈소스 프로그램이라 가벼운 데이터 작업을 하기에는 편리하다.

<Table 1>은 삼성전자 수집 데이터 샘플로 시가와 종가, 개인, 외국인, 기관 수급 데이터이다.

3.2. 자기 조직화 지도 모형

SOM(Self Organizing Map) 모델을 통해 작업을 구현하기 위해 Ohm사에서 발행한 'Turbo C로 길들이는 학습하는 기계 신경망' 책에서 Turbo C 언어로 제공하는 머신러닝 프로그램을 C# 프로그램언어로 변환하여 프로그램을 구현하였다(Lee, 1993). 저자는 어렵게 변환하여 클러스터링 작업을 구현하였지만 이후 클러스터링 연구를 하면서 Accord.NET Framework에서 제공하는 AI 프로그램이 있어 독자는 클러스터링 작업을 쉽게 접근하여 구현할 수 있을 것이다.

3.2.1. 데이터 정규화 방법

인공신경망 인공신경세포에서 연결 가중치는 학습을 통해 입력 패턴과 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 반

〈Table 2〉 Normalized Data

Code	Date	Normalized state			
		Returns	Individual	Foreign	Institutional
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

응해 버리는 경우가 생긴다. 머신러닝의 학습을 위해서 입력 패턴의 정규화(Normalization) 과정을 진행하였다(Heaton, 2012).

n기간 일별 가격데이터는 당일 종가에서 당일 시가를 빼고 그 값을 당일 시가로 나눠서 당일 수익률(PL_i)로 정의하였고, 당일 정규화 데이터(N_i)는 다음과 같이 계산하였다.

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

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

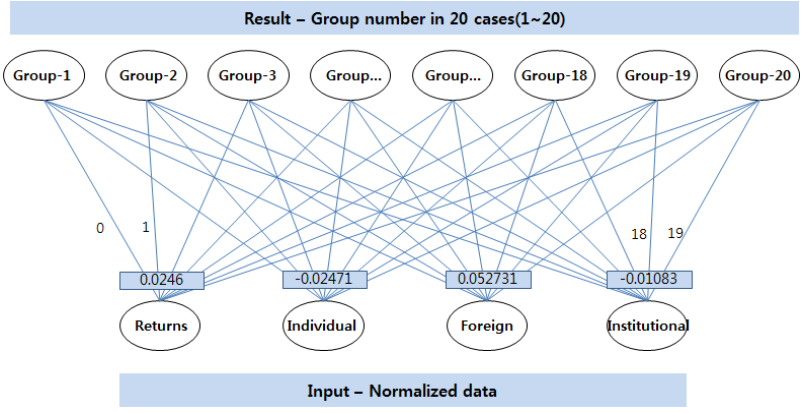
$T = MAX(S_i) - MIN(S_i)$, 일 i= 1, 2, …, n, $S_i = \sum_{i=1}^{n} d_i$, 일 i= 1, 2, …, n

<Table 2>는 머신러닝 학습을 위한 정규화 데이터를 보여주고 있다.

3.2.2. 학습 방법

정규화된 수익률, 개인, 외국인, 기관 거래량 데이터를 인공신경망의 자기조직 지도 모형에 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 역전파 알고리즘 모형에서 학습 검증 테스트에 미래 데이터가 포함되지 않기 위한 것이다. 학습 데이터 일수는 1986일(75.71%)이며 클



〈Figure 1〉 Clustering Learning Model

러스터링 데이터 일수는 2623일(100.0%)이다.

<Figure 1>은 클러스터링 학습을 도식화를 한 내용을 보여준다.

Input Data는 4개이고 Output Data는 20개이며 Layer는 2개로 Hidden Layer는 없다. 출력은 승자 인공신경세포의 결정으로 <Figure 1>처럼 그룹1부터 그룹20까지의 승자 인공신경세포가 출력된다.

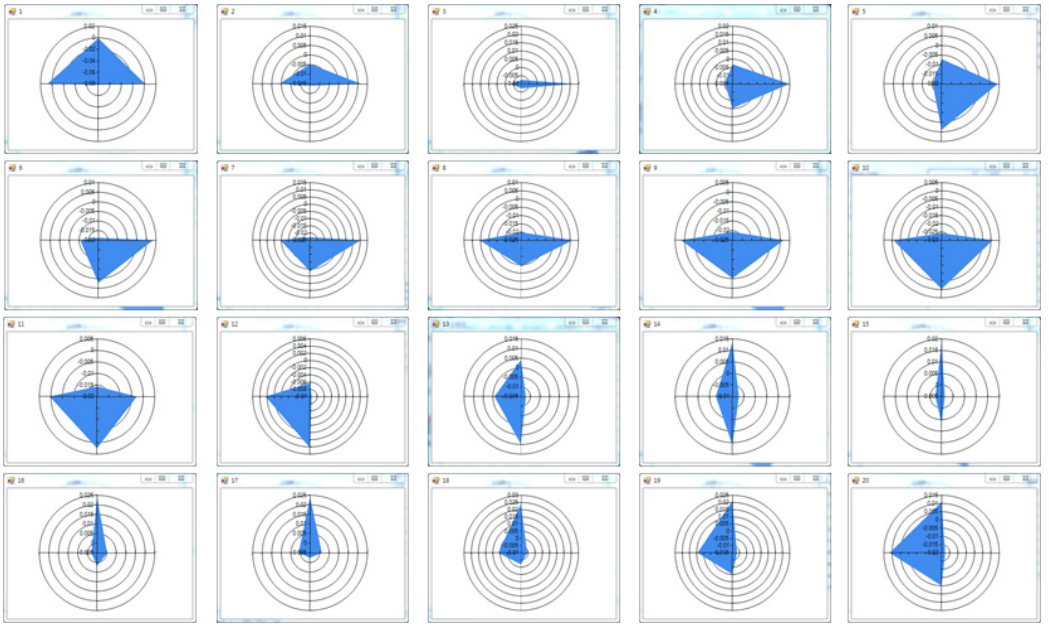
<Table 3>은 클러스터링을 진행하고 나온 결과로 Number는 승자 인공신경세포의 번호이다.

<Figure 2>는 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 거래량 데이터 순이다. 방사형 차트를 보면 일별 수익률과 거래량 데이터의 비율로 의미 있는 패턴을 보여주고 있다. 1번을 보면 수익률이 높고 개인, 기관

〈Table 3〉 Clustering Result

Code	Date	Result	Individual	Foreign	Institutional	Number
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



〈Figure 2〉 Average Radar Chart of Group Patterns in 20 Cases

〈Table 4〉 Clustering Data

TYPE	A005930	A000660	A005380	A015760	A012330	A035420	A005490	A055550	A090430	A051910
1	29	43	40	124	54	20	91	46	62	63
2	40	11	42	22	45	27	44	33	7	24
3	336	12	90	20	97	243	40	78	10	69
4	182	35	78	31	81	118	43	84	19	72
5	40	217	78	201	64	21	92	131	54	102
6	97	253	127	223	135	69	132	158	116	142
7	85	84	178	83	159	182	95	108	89	153
8	18	17	46	63	47	29	70	36	27	60
9	19	42	57	82	48	22	86	48	54	68
10	31	217	134	128	104	76	162	141	483	106
11	71	133	108	105	124	97	136	129	72	104
12	210	94	134	62	137	192	73	121	24	89
13	239	54	109	50	136	155	71	89	8	95
14	25	37	57	90	65	31	68	47	20	67
15	31	65	58	129	68	48	100	83	51	125
16	46	121	109	125	103	52	136	123	281	144
17	54	62	102	54	92	73	145	86	227	94
18	49	171	137	183	125	90	227	156	282	182
19	64	136	131	119	116	125	99	116	65	99
20	319	181	170	91	185	315	75	172	34	127

의 거래량 비율이 높다. 반면 10번은 수익률은 낮고 개인, 외국인, 기관의 거래량이 높은 것을 볼 수 있다. 16번은 수익률은 높으나 개인, 외국인, 기관의 거래량이 적음을 알 수 있다. 이러한 클러스터링을 통해 데이터들의 정보가 의미를 갖도록 그룹화할 수 있다.

<Table 4>는 각 종목별 학습데이터 1985일을 클러스터링 한 클러스터 분포도이다.

3.3. 오류역전파 모형

3.3.1. 데이터 변환 방법

오류역전파 모형에서는 Input Data와 Output Data 패턴이 자기 조직화 지도 모형 출력 데이터 패턴을 그대로 가져다 쓰는 것이 아니다. 오류역전파 모형에서는 Input Data는 60일 데이터 패턴을 입력하고 Target Data 패턴은 2가지의 패턴을 이용하여 상승과 하락을 입력해준다. 따라서 클러스터링에서 나온 데이터를 변환해야 한다.

<Figure 3>은 데이터 변환 방법을 보여주고 있다.

<Figure 3>에서 왼쪽 데이터는 자기 조직화 지도 모형에서 나온 출력 정보이다. 오른쪽 데이터는 자기 조직화 지도 모형의 데이터를 5일 간격으로 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 수

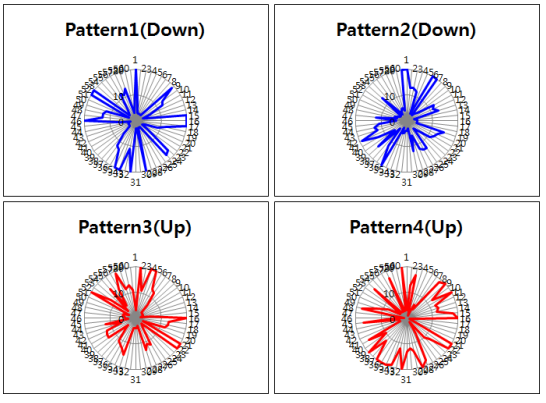
Code	Date	Group
A005930	2007-01-02	20
A005930	2007-01-03	7
A005930	2007-01-04	3
A005930	2007-01-05	3
A005930	2007-01-08	3
A005930	2007-01-09	3
A005930	2007-01-10	3
A005930	2007-01-11	3
A005930	2007-01-12	19
A005930	2007-01-15	13
A005930	2007-01-16	12
A005930	2007-01-17	6
A005930	2007-01-18	3
A005930	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

Group information for 60 days every 5 days																	Type
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...	60	
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

〈Figure 3〉 Data Conversion Method

익률을 구하여 다음 수식을 사용하여 데이터를 만들 수 있다. <Figure 3>에서 오른쪽 데이터의 Type에 들어갈 데이터는 1의 경우는 상승을 의미하며 즉 수익률이 0보다 큼을 의미하며, 2의 경우는 하락을 의미하며 수익률이 0보다 작거나 같은 경우를 말한다.

<Figure 4>는 Input Data의 상승, 하락의 샘플 데이터를 방사형 차트로 그린 것이다.



〈Figure 4〉 Radar Chart of Input Data

원 중심은 1부터 바깥쪽으로 갈수록 20에 가깝다. 상승 패턴을 보면 20에 가까운 데이터가 많은 것을 볼 수 있다. <Figure 2>에서 번호가 클수록 가격이 상승률이 높은 것을 알 수 있는데 <Figure 4>에서도 상승 패턴에 가격 상승률이 반영되는 것을 볼 수 있다.

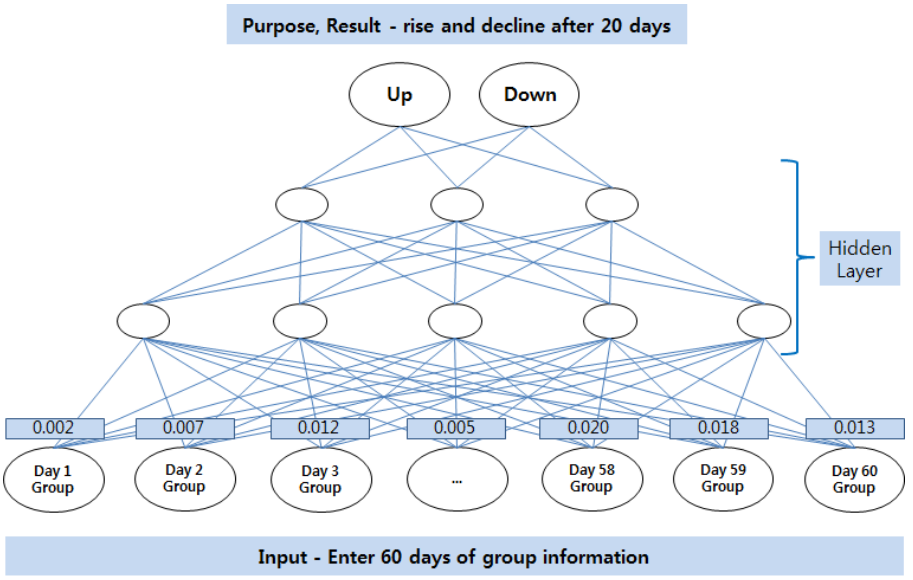
3.3.2. 데이터 정규화 방법

머신러닝의 학습을 위해서 입력 패턴의 정규화 과정을 진행하였다.

모든 일별 그룹화 데이터를 정규화 한다.

$G_i = g_i * 0.01$, 일 $i = 1, 2, \cdots, n$, g_i : 일별 그룹화 데이터, G_i : 정규화된 일별 그룹화 데이터

다음으로 목표구분 데이터를 정규화 한다.



〈Figure 5〉 Classification Learning Model

1 : [1, 0]

2 : [0, 1]

3.3.3. 학습 방법

기계학습에 적용하기 위해 2007년 1월 2일부터 2014년 12월 29일까지를 학습데이터로 사용하였고, 2014년 12월 30일부터 2017년 7월 31일까지를 검증데이터로 사용하였다. 학습 데이터 일수는 1985일(75.68%) 이며, 검증 데이터 일수는 638일(24.32%) 이다.

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

<Figure 5>는 분류 학습 모형을 보여주고 있다.

Input Data는 60개이고 Output Data는 2개이며 Layer는 6개로 Hidden Layer는 4개이다. 출력은 [1, 0] 또는 [0, 1] 예측으로 <Figure 5>처럼 상승, 하락 결과가 출력된다.

3.4. 포트폴리오 구성 방법

포트폴리오 구성은 시가총액 상위 10종목을 각각 오류역전파 모형에 학습시키고 그 중 매월 첫 거래일에 상승 예측 결과가 출력되는 종목들을 매수 진입하고 20일 후 청산하는 방식이다. 포트폴리오 구성을 매월 재구성하고 자산 비중을 동일하게 리밸런싱(Rebalancing)하였다(Park et al., 2014).

시가총액 상위 10종목은 2016년 12월 29일 기준 종가와 현재 발행수량으로 시가총액을 계산하였다. 시가총액 상위 10종목을 벤치마크(Benchmark)로 정하였다. 시가총액 계산은 중간

〈Table 5〉 Prediction Results of Error Propagation Model

Date	Return on Portfolio	Entry number	Success rate(%)	Date	Return on Portfolio	Entry number	Success rate(%)
2015-01-02	8.63	6	100	2016-05-02	-1.32	6	50
2015-02-02	5.11	6	100	2016-06-01	1.16	7	57
2015-03-02	3.92	5	67	2016-07-01	-0.14	6	50
2015-04-01	4.95	3	80	2016-08-01	-1.17	6	50
2015-05-04	-4.42	4	0	2016-09-01	-0.41	4	60
2015-06-01	-4.96	8	29	2016-10-04	0.17	6	75
2015-07-01	-0.21	6	33	2016-11-01	-2.32	2	33
2015-08-03	-4.89	6	0	2016-12-01	3.23	6	71
2015-09-01	8.65	7	100	2017-01-02	0.79	6	50
2015-10-01	2.28	4	75	2017-02-01	-0.56	4	50
2015-11-02	3.31	6	75	2017-03-02	0.14	5	40
2015-12-01	0.04	4	50	2017-04-03	-1.53	6	50
2016-01-04	-2.22	6	33	2017-05-02	9.89	4	100
2016-02-01	9.23	5	100	2017-06-01	7.04	3	100
2016-03-02	2.66	5	67	2017-07-03	4.47	4	75
2016-04-01	4.36	4	100	Total	55.88	160	62.57

변동내용을 적용하기 어려워 2017년 최초 개발을 시작할 때 2016년 마지막일로 정했다. 이후 검증 데이터가 적어 개발 진행 도중 2017년에 7개월간의 데이터를 추가하였다.

포트폴리오는 2015년 1월부터 2017년 7월까지 매월 첫 거래일에 오류역전파 모형 인공신경망의 예측 결과가 상승으로 분류한 종목들로 구성된다. 매수 가격은 첫 거래일 시가이고 매도가격은 20일 후 종가이다.

수익률 PL 계산은 다음과 같다. 매수, 매도 수수료는 0.015%, 제세금은 0.3%로 수익률을 가정하였다.

$$PL = \frac{(C-O)-O*0.00015-C*0.00015-C*0.003}{O} * 100,$$

O : 매수가, C : 매도가

4. 머신러닝 성능 평가 및 포트폴리오 성과 분석

<Table 5>는 포트폴리오의 월별 성과, 진입 종목 수량과 예측 성공률의 결과이다. 검증 데이터의 포트폴리오 연도별 예측 성공률의 결과는 2015년 59.05%, 2016년 62.24%, 2017년 66.43%로 총 평균 성공률은 62.57%로 검증 데이터의 포트폴리오 예측 성공률은 상당히 좋은 결과를 보여주었다. 예측 성공률이 보여주듯이 월별 성과도 비례하여 상당히 높은 것을 알 수 있다. 이는 머신러닝의 학습 데이터의 학습이 잘 되었음을 보여주는 성과로 볼 수 있을 것이다.

〈Table 6〉 Return on Portfolio

Date	Same allocation rate(%)				Compound interest rate(%)			
	Portfolio	Benchmark	KOSPI200	KOSPI	Portfolio	Benchmark	KOSPI200	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
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
Average	18.02	7.19	8.76	7.71				
Total	55.88	22.3	27.18	23.9	69.83	22.94	29.76	25.88

<Table 6>에서 개별종목의 월별 수익률을 이용하여 포트폴리오의 월별 수익률과 복리 수익률을 구하였다.

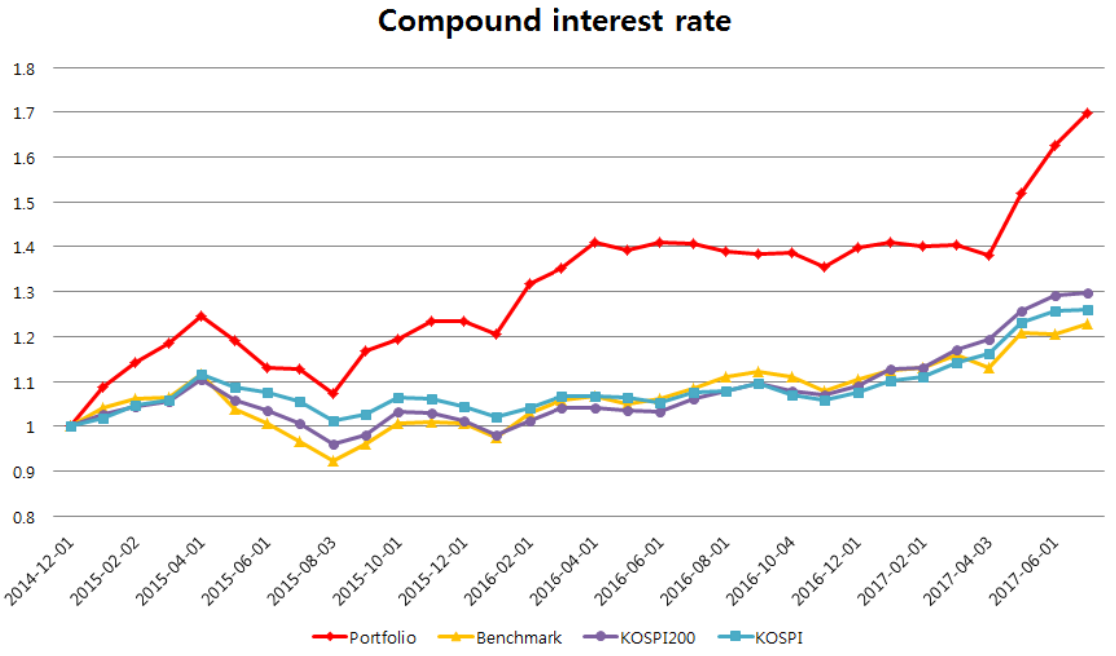
동일배분 수익률(PL_i)은 다음과 같이 계산하였다.

$PL_i = P_i * 1/n$, 종목 $i = 1, 2, \ldots, n$, P_i : 개별 종목 수익률

SPL = $\sum_i^n PL_i$, : 개별 종목 동일배분 수익률의 합

복리 수익률(CI)은 다음과 같이 계산하였다.

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



〈Figure 5〉 Compound Interest Rate

<Table 6>에서 동일배분 수익률의 KOSPI200 지수 연평균 수익률 8.76%, KOSPI 지수 연평균 7.71%로 2015년, 2016년, 2017년도는 1년 만기 국채 수익률의 3년 평균 1.531%보다 수익률이 높았던 연도였다. 하지만 포트폴리오 연평균 수익률은 18.02%로 KOSPI200, KOSPI 지수보다 더 좋은 성과를 보였다. 벤치마크의 연평균 수익률은 7.19%로 포트폴리오 수익률이 훨씬 뛰어난 성과를 보여주었다.

복리 수익률도 벤치마크 수익률은 22.94%, KOSPI200 지수는 29.76%, KOSPI 지수는 25.88%로 비슷한 수익률을 보여준 반면, 포트폴리오 수익률은 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 투자라고 볼 수 있다.

<Figure 5>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.

<Table 7>은 포트폴리오와 벤치마크, KOSPI200, KOSPI 지수의 성과를 분석한 결과이다.

벤치마크와 KOSPI200, KOSPI 지수의 결과는 비슷하였다. 연평균 수익률은 <Table 6>의 포트폴리오의 동일배분 수익률 평균을 연율화한 것이다. 포트폴리오는 연평균 수익률에서도 벤치마크보다 우수한 성적을 보이고 있다. MDD는

〈Table 7〉 Performance Analysis

	Portfolio	Benchmark	KOSPI200	KOSPI
Same allocation rate(%)	55.88	22.30	27.18	23.90
Compound interest rate(%)	69.83	22.94	29.76	25.88
Average annual return(%)	21.63	8.63	10.52	9.25
MDD(%)	14.85	13.96	9.82	10.10
Standard deviation	0.140	0.110	0.089	0.079
Sharpe ratio	1.436	0.645	1.009	0.974

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. 샤프지수는 1.436으로 벤치마크, KOSPI200, KOSPI 지수보다 높게 나타나고 있다.

5. 결론 및 향후 연구

본 연구는 머신러닝과 투자자별 거래량 데이터를 활용한 주식 포트폴리오의 투자 전략과 그 성과를 분석하였다.

기존 연구들은 주가 및 수익률 데이터를 머신러닝에 적용한 사례 연구들이 많다. 투자자별 거래량의 데이터 연구도 데이터가 주는 의미의 연구만을 찾을 수 있었다. 그래서 이 연구는 머신러닝에 투자자별 거래량과 같은 의미 있는 데이터가 학습에도 어떠한 영향을 미치고 있는지 한 예를 선보인 연구라고 할 수 있다.

오류역전파 모형 인공신경망을 이용한 포트폴리오의 총 평균 예측률은 62.57%이다. 과거 연구 사례들 보다도 상당히 좋은 결과를 확인할 수 있었다. 포트폴리오 성과 분석은 시가총액 상위 10종목인 벤치마크보다도 수익률이 2배 이상 좋았으며 KOSPI200 지수와 KOSPI 지수로 시장 수익률 비교를 하여도 포트폴리오 수익률이 상당히 좋았다. 이것은 클러스터링의 그룹화와 학습데이터의 학습이 잘 되었음을 더욱 확실하게 보여주고 있다. 이는 투자자별 거래량이 주가에 미치는 영향이 상당히 높음을 알 수 있는 근거가 될 것이다.

또한 이 연구 이전에 주가 및 거래량 데이터로 먼저 연구를 하였지만 의미 있는 결과를 찾을 수가 없었다. 여러 연구 중 나온 아이디어가 만들어 낸 연구 결과이며 이전 연구 결과들을 저장하지 않아 비교할 수는 없지만 추후 머신러닝 성능 평가 자료로 사용할 수 있도록 다시 한번 시도하여 데이터를 저장한다면 좋은 비교 자료가 될 수 있을 것이다.

수익률은 매월 포트폴리오 구성을 하고 자산을 동일 비중으로 리밸런싱한 결과이다. 매월 포트폴리오를 구성할 때 동일 종목은 매도를 하지 않고 리밸런싱 하는 방법으로 시스템을 구현한다면 더욱 좋은 성과가 예상된다.

주식시장에는 주가에 영향을 줄 많은 데이터가 있다. 이런 데이터를 찾아 머신러닝에 활용한다면 이보다 좋은 결과를 찾을 수도 있을 것이다. 보다 주가에 영향을 주는 데이터를 찾아 실제 거래가 가능한 상태의 연구를 구현한다면 좋

은 성과를 낼 수 있는 실 거래도 가능해 보인다.

이 연구는 Turbo C언어를 C#언어 프로그램으로 변환해서 구현하여 다소 시간이 걸렸지만 R, Python, MATLAB과 같은 프로그래밍 언어에서는 Clustering 및 Classification을 다양하게 지원하고 있으며 쉽게 배우고 사용할 수 있다. 이를 이용한다면 좀 더 수월한 연구가 될 것이다.

참고문헌(References)

Cho, H. Y. and P. S. Lee, "A study on the relationship between price volatility and trading volume for trader type," *Korean Journal of Financial Studies*, Vol.29(2001), 373~405.

Daigler, R. and M. Wiley, "The impact of trader type on the futures volatility-volume relation," *The Journal of Finance*, Vol.54(1999), 2297~2316.

Dreiseitl, S., Machado, L. O., "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, Vol.35(2002), 352~359.

Heaton, J., *Introduction to the Math of Neural Networks*, Heaton Research, Inc., 2012.

Kim, J-H, "An analysis on the impact of investor's information superiority and negative feedback trading on stock return," *Korean Journal of Financial Studies*, Vol.42(2013), 667~698.

Kim, S. and H. Choi, "Performance analysis on trading system using foreign investors' trading information," *Korean Management Science Review*, Vol.32(2015), 57~67.

Kim, S. M. and J. J. Kim, "A new cluster validity index based on connectivity in self-organizing map.", The Korean Journal of Applied Statistics, Vol.33, No.5(2020), 591~601.

Kim, S. W. and H. C. Ahn, "Development of an intelligent trading system using Support Vector Machines and Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.16(2010), 71~92.

Ko, K. and J. Lee, "Foreigner's trading information and stock market: 10 years' experience of stock market liberalization," *The Korean Journal of Finance*, Vol.16(2003), 159~192.

Kohonen, T., "The self-organizing map," *Proceedings of IEEE*, 78(1990), 1464~1480.

Kwark, N. K. and S. G. Jun, "Performance and impact of foreign investment," *The Korean Journal of Financial Management*, Vol.15, No.2(1998), 369~399.

Jeong, Y. G. and Y. S. Yun, "A Study on the Predictability of Stock Price Using Artificial Neural Network Model.", The Korean Journal of Financial Management, Vol.15(1998), No.2, p.369-399.

Lee, S. W., *Turbo-C Learning Machine Neural Network*, Book publishing Ohm, 1993.

Moon, J., S. Kang, and J. Kim, "A study on the performance and investment behavior classified by the type of investors," *Korean International Accounting Review*, Vol.65(2016), 155~178.

Oh, S. H. and S. B. Hahn, "Analyzing the cumulative returns on investments of domestic and foreign investors in Korean stock market," *Korean Journal of Financial Studies*, Vol. 37(2008), 537~567.

Olson, D. and D. Delen, *Advanced Data Mining Techniques*, Springer, 2008.

Park, K. I., "Trading performance of foreign investors and exchange rate," *International Business Review*, Vol. 18(2014), 119~135.

Park, S. C., S. W. Kim, and H. S. Choi, "Selection model of system trading strategies using SVM," *Journal of Intelligence and Information Systems,* Vol.20(2014), 59~71.

Schmidhuber, J., "Deep learning in neural networks: An overview," *Neural Networks*, 61(2015), 85~117.

Yi, K. Y. and Y. G. Lee, "The differences in investment behavior and performance by investor types," *Journal of Industrial Economics and Business*, Vol.17(2004), 1233~1253.

Abstract

Performance of Investment Strategy using Investor-specific Transaction Information and Machine Learning

Kyung Mock Kim** · Sun Woong Kim** · Heung Sik Choi*

Stock market investors are generally split into foreign investors, institutional investors, and individual investors. Compared to individual investor groups, professional investor groups such as foreign investors have an advantage in information and financial power and, as a result, foreign investors are known to show good investment performance among market participants.

The purpose of this study is to propose an investment strategy that combines investor-specific transaction information and machine learning, and to analyze the portfolio investment performance of the proposed model using actual stock price and investor-specific transaction data. The Korea Exchange offers daily information on the volume of purchase and sale of each investor to securities firms. We developed a data collection program in C# programming language using an API provided by Daishin Securities Cybosplus, and collected 151 out of 200 KOSPI stocks with daily opening price, closing price and investor-specific net purchase data from January 2, 2007 to July 31, 2017.

The self-organizing map model is an artificial neural network that performs clustering by unsupervised learning and has been introduced by Teuvo Kohonen since 1984. We implement competition among intra-surface artificial neurons, and all connections are non-recursive artificial neural networks that go from bottom to top. It can also be expanded to multiple layers, although many fault layers are commonly used. Linear functions are used by active functions of artificial nerve cells, and learning rules use Instar rules as well as general competitive learning. The core of the backpropagation model is the model that performs classification by supervised learning as an artificial neural network.

We grouped and transformed investor-specific transaction volume data to learn backpropagation models through the self-organizing map model of artificial neural networks. As a result of the estimation

* Corresponding author: Heung Sik Choi
Graduate School of Business IT, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 136-702, Korea
Tel: +82-2-910-4567, Fax: +82-2-910-4017, E-mail: hschoi@kookmin.ac.kr
** Graduate School of BIT, Kookmin University

of verification data through training, the portfolios were rebalanced monthly. For performance analysis, a passive portfolio was designated and the KOSPI 200 and KOSPI index returns for proxies on market returns were also obtained.

Performance analysis was conducted using the equally-weighted portfolio return, compound interest rate, annual return, Maximum Draw Down, standard deviation, and Sharpe Ratio. Buy and hold returns of the top 10 market capitalization stocks are designated as a benchmark. Buy and hold strategy is the best strategy under the efficient market hypothesis. The prediction rate of learning data using backpropagation model was significantly high at 96.61%, while the prediction rate of verification data was also relatively high in the results of the 57.1% verification data. The performance evaluation of self-organizing map grouping can be determined as a result of a backpropagation model. This is because if the grouping results of the self-organizing map model had been poor, the learning results of the backpropagation model would have been poor. In this way, the performance assessment of machine learning is judged to be better learned than previous studies.

Our portfolio doubled the return on the benchmark and performed better than the market returns on the KOSPI and KOSPI 200 indexes. In contrast to the benchmark, the MDD and standard deviation for portfolio risk indicators also showed better results. The Sharpe Ratio performed higher than benchmarks and stock market indexes. Through this, we presented the direction of portfolio composition program using machine learning and investor-specific transaction information and showed that it can be used to develop programs for real stock investment.

The return is the result of monthly portfolio composition and asset rebalancing to the same proportion. Better outcomes are predicted when forming a monthly portfolio if the system is enforced by rebalancing the suggested stocks continuously without selling and re-buying it. Therefore, real transactions appear to be relevant.

Key Words : Investor-specific Transaction Information, Machine Learning, Robo-advisor, Trading System

Received : December 29, 2020 Revised : February 19, 2021 Accepted : March 8, 2021
Corresponding Author : Heung Sik Choi

저 자 소 개

김 경 목

현재 주식회사 엠알에스에서 재직 중이다. 숭실대학교 전산원을 졸업하였고 국민대학교 비즈니스IT전문대학원에서 트레이딩시스템 전공 석사학위를 취득하였다. 주요 관심분야로는 인공지능, 주식, 파생상품, 시스템트레이딩, 자산배분 전략, 금융공학, 로보어드바이저 등이다.

김 선 웅

현재 국민대학교 비즈니스IT전문대학원 교수로 재직 중이다. 서울대학교 경영학과에서 경영학사를 취득하고, KAIST 경영과학과에서 투자론을 전공하여 공학석사와 공학박사 학위를 취득하였다. 주요 관심분야는 트레이딩시스템, 투자공학, 헤지펀드와 자산운용이다.

최 흥 식

현재 국민대학교 경영대학 경영정보학부 및 동 대학 비즈니스IT전문대학원 교수로 재직 중이다. KAIST에서 경영과학 석사학위를 취득하였으며 미국 로체스터 대학에서 경영학 석사 및 박사학위를 취득하였다. 관심분야로는 파생상품 시스템트레이딩, 트레이딩계량분석, 옵션 변동성매매 등이다.