하이브리드 주가예측 모델

오유진*, 한규숙**, 김유섭***

초 록

본 논문에서는 주가예측의 정확도를 향상시키기 위하여 공적분 (Co-Integration)과 인공 신경망 (Artificial Neural Networks)을 사용한 2단계 하이브리드 주가예측 모델을 제시한다. 과거 주가의 예측을 위하여 많은 계량적인 모델이 수립되었으나 지나치게 복잡한 모델의 특성상 광범위한 사용으로 이어지기 어려운 문제가 있었다. 또한 이를 극복하기 위하여 인공 신경망에 기반한 모델을 사용하여 주가 예측을 시도하였으나 성격이 판이하게 다른 종목들을 하나의 모델에 통합하고자 하여 모델이 지나치게 일반화되어 신뢰성 있는 정확도를 보일 수 없었다. 따라서 본 논문에서는 이러한 종목의 다양성에서 유발하는 문제를 최소화하기 위하여 장기적 관련성을 보유한 종목들을 집단화하고, 이들을 대상으로 신경망을 학습하여 예측 성능을 향상시키고자 하였다. 실증분석은 KOSPI 및 KOSDAQ 시장의 시가 총액 상위 종목들을 대상으로 이루어졌으며, 이들 실험을 통하여 인공 신경망 모델이 일반 선형 회귀분석에 비하여 매우 좋은 예측 성능을 가지고 있으며, 공적분 관계가 있는 것으로 판명된 종목군의 학습이 임의로 구성된 종목군의 학습에 비하여 높은 예측 정확도를 가지는 것을 확인할 수 있었다.

- 주제어: 공적분, 인공 신경망, 선형 회귀분석, 주가예측 모델, KOSPI, KOSDAQ

I. 서 론

금융 시장에서의 예측과 관련한 전산학적 입장에서의 많은 연구는 예측 모형의 구축에 집중되어 왔는데, 이 모형은 과거 가격 시퀀스에 기반한 기술 변수, 미시적 관점의 주식 관련 변수, 그리고 거시적 관점의 경제 변수와 같은 여러 설명 변수를 사용하였다 (Ghosn and Bengio (1997)). 또한 Kendall and Ord (1997) 역시 시계열의 입장에서 많은 기법들을 활용하였으며, Fama (1988)는 시장을 완벽하게 예측하는 것은 효율적 시장 가설 (Efficient Market Hypothesis)의 입장에서는 불가능하나 어느 정도의 이익은 유도가 가능하다고 하였다. 특히 인공지능 분야의 알고리즘들이 강력한 표현력과 모델링 능력을 지니고 있기 때문에, 이들을 이용한 예측의 사례들이 보고되고 있는데 Malkiel (1996), 인공 신경망(Artificial Neural Network), 결정트리(Decision Tree), SVM(Support Vector Machine) 등이 이에 적용되어 왔다 (Dempster et. al. (2001), Fan and Palaniswami (2001) Kim et. al. (2002) Saad et. al. (1998)).

주가 예측에 있어서 인공 지능 측면에서의 이러한 시도들은 수학적인 모델에 비하여 매우 단순하지만 주목

* 서울시립대학교 경제학과, e-mail: ohyj@ewhain.net
** 서울시립대학교 경제학과, e-mail: schiz33@uos.ac.kr
*** 한림대학교 컴퓨터공학과, e-mail: yskim01@hallym.ac.kr

할 만한 성능을 보여주었다. 하지만, 실제 매매에 있어서는 성능의 한계를 보여주었기 때문에, 이를 극복하기 위하여 다양한 방법론이 시도되었는데, 김유섭 (2004)은 매매 행위를 모델링하여 예측 행위를 보완하였고, Armano (2005)는 다양한 예측 모델을 통합한 앙상블 모델을 수립하였다. 하지만 이러한 방법론은 기본적인 예측 모델의 성능에 따라 그 성능이 좌우되기에, 기본 예측 모델의 성능 향상 문제는 꾸준히 제기되고 개선되어야 할 사항이다. 본 논문에서는 인공 신경망에 기반한 주가예측 모델의 성능 향상을 위하여 공적분 검정을 이용하였다. 기존 인공 신경망의 학습 데이터 구축에는 여러 개별 종목들의 일자별 데이터가 사용되는데, 이러한 과정에서 전혀 성향이 다른 종목들의 데이터가 모두 학습에 사용되게 된다. 그 결과 각 종목별 특성은 학습에서 반영되지 못하여 해당 종목의 미래 가격 예측이 불안정해진다. 반면에 개별 종목만을 가지고 학습을 하게 되면, 학습 데이터의 양이 부족하게 되는 현상을 빚게 되어 재현에 한계가 있기 때문에 보다 강력한 예측 엔진의 구축이 어렵다. 따라서 본 논문에서는 이러한 문제를 2단계 하이브리드 예측 모형을 제시함으로써 해결하고자 하였다. 먼저 1단계에서는 예측을 시도하고자 하는 종목과 Johansen의 공적분 검정을 하여 통과한 종목들을 수집하고 유사 종목군을 형성한다. 이 단계에서는 매매를 원하는 종목과 유사한 시계열 패턴을 보이지만 보다 다양한 움직임을 보이는 종목군을 추출하여 향후 학습될 모델이 보다 강력할 수 있도록 한다. 그리고 2단계에서는 이 종목군의 일자별 데이터로 인공 신경망을 학습하여 최종 예측 모델을 수립하는데, 데이터의 시계열성을 감안하여 학습에 필요한 속성을 설계한다.

본 연구에서는 실험을 위하여 KOSPI와 KOSDAQ 시장의 시가총액 상위 종목들의 데이터를 수집하였다. 인공 신경망은 뚜렷한 특성을 주로 학습하는 경향이 있어, 이를 이용한 모델은 주로 가격 변동폭이 큰 소형주 위주로 추천이 이루어진다. 하지만 실제 매매에 있어서 이러한 소형주는 여러 제한이 있기 때문에 본 실험에서는 대형주를 주요 대상으로 삼았다. 실험 결과 KOSPI와 KOSDAQ을 막론하고 공적분 검정을 통하여 구성된 유사 유형의 종목군을 대상으로 학습한 모델이 랜덤하게 추출하여 구성된 종목군을 학습한 모델보다 더 높은 예측 정확도를 보여주었다. 또한 인공 신경망 모델은 선형 회귀분석 모델에 비하여 월등히 높은 예측 정확도를 보여주었다.

본 논문의 2절에서는 본 연구에 사용된 자료의 성격에 대하여 설명하였고, 3절에서는 예측 모델의 구성에 대하여 설명하였다. 예측 모델에 대한 설명에는 공적분에 대한 간략한 설명과 더불어 인공 신경망에 대한 설명이 포함된다. 4절에서는 실험 결과 및 평가를 논의한 후에 5절에서 결론 및 향후 연구에 대하여 논하였다.

Ⅱ. 자 료

본 연구에서는 KOSPI 및 KOSDAQ의 시가 총액 상위 종목 중에서 공적분 검정 및 예측을 위한 학습에 충분한 시계열 데이터가 있는 30 종목을 각각 선택하여 실험에 활용하였다. <표 1>은 KOSPI 및 KOSDAQ 시장에서 본 실험을 위하여 추출된 종목들의 리스트로써 이 종목들의 2007년 1월 26일 현재 시가 총액을 보여주고 있다 (KOSCOM 2007).

<표 1> 실험에 사용된 KOSPI 및 KOSDAQ 시장에서의 시가 총액 상위 30 종목들

연 번	KOSPI		KOSDAQ	
	종 목 명	시가 총액 (백만원)	종 목 명	시가 총액 (백만원)

1	삼 성 전 자	86,464,711	NHN	6,011,386
2	한 국 전 력	28,549,763	LG 텔레콤	2,587,008
3	POSCO	27,769,007	하나로 텔레콤	1,722,617
4	SK 텔레콤	15,995,161	아시아나 항공	1,196,416
5	하 이 닉 스	15,124,775	하 나 투 어	780,608
6	현 대 차	14,572,559	동 서	688,380
7	KT	12,890,823	CJ 홈쇼핑	687,060
8	현 대 중 공 업	10,488,000	다 음	671,234
9	신 세 계	10,392,136	휴 맥 스	638,695
10	삼 성 전 자 우	10,389,209	포 스 데 이 터	487,679
11	SK	9,007,742	태 웅	481,032
12	KT&G	8,463,199	GS 홈쇼핑	475,781
13	외 환 은 행	7,771,127	쌍 용 건 설	456,968
14	삼 성 화 재	7,575,600	CJ 인터넷	445,482
15	SK 네트웍스	7,288,065	네 오 위 즈	410,560
16	S-Oil	7,261,590	지 엔 텍	405,840
17	기 업 은 행	6,968,607	서 울 반 도 체	401,840
18	현 대 모 비 스	6,573,197	파 라 다 이 스	360,133
19	현 대 건 설	5,343,681	플래닛 82	336,528
20	LG	4,952,390	매 일 유 업	335,000
21	KTF	4,941,453	엠 넷 미 디 어	324,748
22	삼 성 중 공 업	4,836,780	인 터 파 크	282,773
23	삼 성 물 산	4,803,696	에스에프에이	250,525
24	두 산 중 공 업	4,572,685	심 텍	244,800
25	기 아 차	4,079,598	LG 마이크론	228,000
26	GS 건설	3,972,900	화 인 텍	216,000
27	현 대 산 업	3,806,901	KTH	205,620
28	CJ	3,232,427	유 진 기 업	202,621
29	삼 성 증 권	3,064,392	HK 저축은행	178,943
30	대 우 증 권	3,022,604	오 디 코 프	148,566

본 논문의 실험을 위해서는 최소한 2003년도부터 2006년도까지의 4년간의 데이터가 있는 종목이 필요하다. 따라서 최근에 설립 또는 인수/합병으로 변동이 심하고, 꾸준한 데이터가 없는 종목들은 시가총액의 크기에 불문하고 실험 대상에서 제외하였다. 예를 들어 KOSPI 시장의 신한지주, 우리금융, 롯데쇼핑과 KOSDAQ 시장의 메가스터디, SSCP, 헬리아텍과 같은 종목들은 2003년부터의 데이터가 없어 실험 대상에서 제외하였다. 또한 일부 종목들은 거래 정지, 감자 또는 증자로 인하여 가격 및 거래량에 있어서 비정상적인 움직임을 보인 경우들도 있었으나, 전체 실험 데이터의 크기에 비하여 무시할 수 있는 정도라 가정하여 이를 일반적인 움직임으로 여기고 실험에 임하였다.

모든 종목들은 매매에 따라서 가격 및 거래량과 관련한 여러 데이터를 매일 생성해낸다. 본 논문에서는 시중의 가정 매매 시스템(HTS: Home Trading System)을 통하여 쉽게 구할 수 있는 데이터를 토대로 하여 이를 변환 및 가공하여 예측에 활용한다. 예를 들어 특정 종목이 특정 일자에 보여준 시가, 고가, 저가, 종가와 거래량을 나타내는 수치를 비롯하여 5일, 10일, 20일, 60일, 120일의 가격 및 거래량의 이동 평균값을 기초 데이터로 활용한다. 공적분 검정에서는 이들 데이터 중에서 종가의 시계열만을 그 대상으로 선택하였다. 한편, 인공 신경망을 활용한 예측 시스템에서는 이들 기초 데이터를 가공하여 새로운 데이터를 생성 및 추출하여 활용한다. 예측 시스템은 기초 데이터로부터 이동 평균선의 기울기, 종가와 이동 평균선의 간격 등 19개의 속성을 새로이 설계하여 사용하는데, 특히 데이터의 시계열성을 반영하기 위하여 전일과 전전일 종가

의 현재가 대비 비율을 속성으로 사용한다. 이들 19개의 속성들은 모두 비율값으로 되어 있는데, 이는 종목별로 값의 범위의 차이가 크기 때문에 이를 정규화하기 위함이다.

Ⅲ. 주가예측 모델 - 공적분과 인공 신경망

본 논문에서 제시하고 있는 주가예측 모델에서는 예측하고자 하는 종목과 유사성이 있다고 판단되는 종목들을 Johansen의 공적분 검정을 통하여 추출하는 추출 단계와, 추출된 종목들의 데이터를 가지고 인공 신경망으로 예측 모델을 구축하는 학습 단계로 나누어진다. 본 절에서는 공적분의 간단한 설명과 예를 보임으로써 추출 단계에 대하여 논하고, 인공 신경망의 기본 원리와 적용 사례를 보임으로써 학습 단계에 대하여 논한다.

1. 추출 단계 - Johansen의 공적분 검정

공적분은 확률보행과정(random walk)을 따르는 변수들 간에 장기적 관계가 존재하는지를 살펴보는 것이며, 기존의 공적분 검정 중에서는 Johansen의 공적분 검정(1988)이 가장 일반적으로 사용되고 있다. Johansen의 최우추정법은 다변수 분석틀에서 진단할 수 있으며, 모수 추정치에 대해 명백히 가설 진단을 수행할 수 있는 장점이 있다. 다음은 벡터자기회귀모형(VAR, Vector Autoregression Model)이다.

$$z_t = D(t) + \sum_{i=1}^{p} \phi_i z_{t-i} + a_t \qquad (1)$$

여기서 z_t는 k개의 확률보행을 따르는 변수들을 포함하고 있으며, $D(t)$는 추세(trend)를 의미하며, a_t는 오차항이다. 여기서 p는 VAR모형의 차수를 의미하며, 본 연구에서는 적정한 시차를 정하기 위하여 Akaike 기준을 사용하여 모형을 선택하였다. 실험결과 대부분의 경우 2시점의 과거시차까지를 포함하는 것이 최적인 것으로 나타났다.

식(1)을 오차수정모형(ECM, Error Correction Model)으로 전환하면 다음 식(2)와 같으며,

$$\Delta z_t = D(t) + \Pi z_{t-1} + \sum_{i=1}^{p} \phi_i^* z_{t-i} + a_t \qquad (2)$$

여기서 $\Pi = -\phi(1)$, $\phi^* = -\sum_{i=i+1}^{p} \phi_j$, $i = 1,...,p-1$, 그리고 $\Delta z_t = z_t - z_{t-1}$이다. 여기서 관심사는 행렬 Π에 있다. 만약 $Rank(\Pi) = 0$이면, 변수들 간에 공적분 관계가 존재하지 않는다는 것을 의미하며, $Rank(\Pi) = m > 0$이면 z_t는 m개의 공적분 관계를 가진다. 여기서 Johansen의 공적분 검정은 다음의 Trace 검정과 Max검정의 2 가지로 구축된다.

우선 Trace 검정의 가설들은$H_0 : m = m_0 \quad vs. \quad H_a : m > m_0$과 같이 설정되며, 여기에서 m_0는 0과 $k-1$사이의 수이다. Johansen의 Trace 검정통계량 $L_{tr}(m_o)$은

$$L_{tr}(m_0) = -(T - kp) \sum_{i=m_o+1}^{k} \ln(1 - \lambda_i) \qquad (3)$$

으로 정의되며, 여기서는 고유값(eigenvalues)들을 의미한다. 검정통계량의 값이 주어진 임계치보다 크면, 귀무가설을 기각하게 된다.

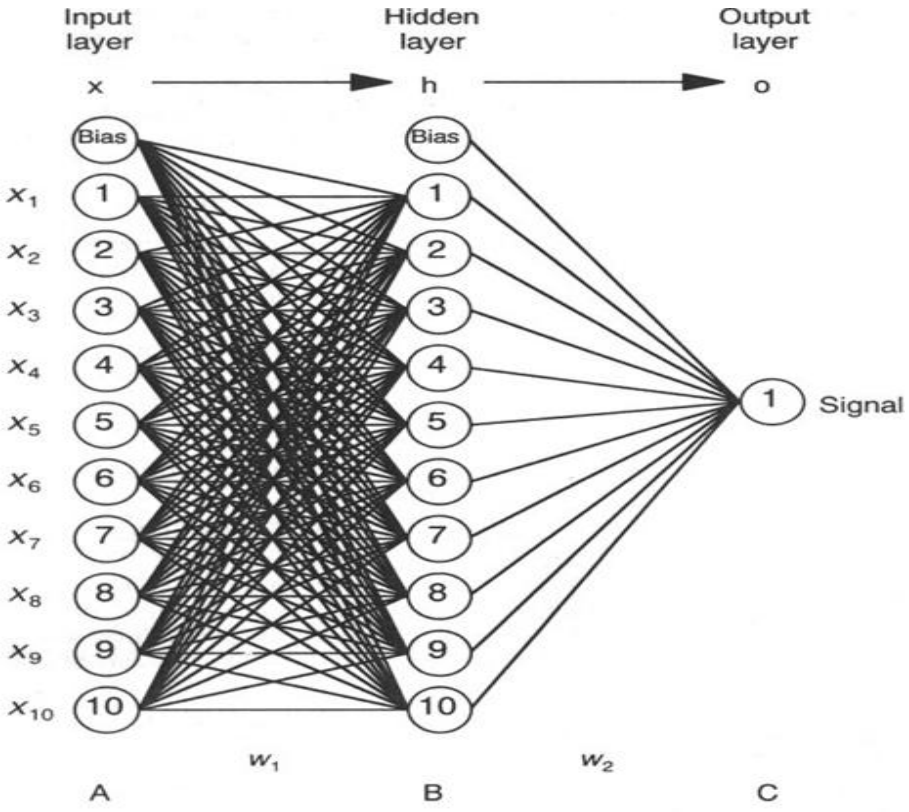
Max검정의 가설들은 $H_0 : m = m_0 \quad vs. \quad H_a : m = m_0 + 1$로 설정되며, Johansen의 Max 검정통계량

$L_{\max}(m_o)$은

$$L_{\max}(m_0) = -(T-kp)\sum_{i=m_o+1}^{k}\ln(1-\lambda_{m_0+1}) \quad (4)$$

으로 정의된다. 마찬가지로 검정통계량의 값이 주어진 임계치보다 크면, 귀무가설을 기각하여 공적분 관계의 유무뿐만 아니라, 공적분 관계의 개수를 검정할 수 있다.

2. 학습 단계 - 인공 신경망

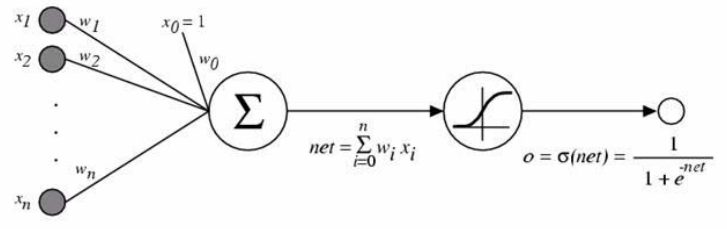


(그림 1) 인공 신경망의 전체 구조

(그림 1)은 학습 단계에서 사용되는 인공 신경망의 대략적인 모습을 보여준다. 인공 신경망은 크게 3개의 층으로 이루어진다. 그리고 각 층의 모든 노드들은 다른 층의 모든 노드들과 간선으로 연결되어 있으며, 각 간선에는 가중치 값이 부여되어 있다. 첫째, 입력층(그림 1의 A층)은 실제 데이터의 입력이 이루어지는 곳으로써 하나의 데이터 레코드는 수십 또는 수백개의 속성으로 이루어지고 각각의 개별 속성값들이 적절한 정규화 과정을 거쳐서 입력층의 개별 노드에 입력된다. 둘째, 출력층(그림 1의 C층)은 입력된 개별 레코드의 인공 신경망에서의 최종 결과값이 출력되는 층으로써, 본 연구에서는 종가의 전일대비 증가율이 그 출력값이 된다. 그리고 마지막으로 은닉층(그림 1에서 B층)은 입력층과 출력층과는 달리 외부에 그 값이 노출되지 않는 특성을 가지고 있으며 인공 신경망의 성능에 가장 큰 영향을 미치는 요소이다. 본 논문에서는 이 은닉층의 노드수를 10개로 제한하였다.

은닉층과 출력층의 노드들은 각각 입력되는 값들을 토대로 하여 하나의 값을 계산하는데 이 때 시그모이드 함수(Sigmoid Function)을 사용하여 계산하기 때문에 시그모이드 단위(Sigmoid Unit)이라고 하며 다음 (그

림 2)와 같은 구조를 가진다 (Mitchell (1997)).



(그림 2) 시그모이드 단위(Sigmoid Unit)의 구조

개별 노드들은 자신의 하위층에서 출력된 값(x_i)과 연결 간선의 가중치 값(w_i)을 이용하여 $y = \sum_{i=0}^{n} w_i x_i$를 계산하고, 이 값을 데이터의 분포를 고려하여 시그모이드 함수 $\sigma(y) = \frac{1}{1+e^{-y}}$를 계산하여 자신의 출력 값으로 결정한다. 이러한 과정이 최종적으로 출력층까지 이루어지면서 주어진 데이터에 대한 출력값이 결정된다. 이 때 가중치 w_i의 값이 수십 또는 수백회의 학습을 통하여 최적화되는데, 이와 관련한 자세한 알고리즘은 Mitchell (1997)에서 찾아볼 수 있다.

Ⅳ. 실험 및 평가

1. Johansen 공적분 검정을 통한 유사 종목 추출

본 연구에서는 실험을 위하여 KOSPI 및 KOSDAQ 시장의 지수를 기준 시계열로 보고 <표 1>의 종목 중에서 이와 유사한 시계열을 추출하였는데, 검정을 위하여 2003-2004년 데이터가 사용되었다. 여기서 지수를 기준 시계열로 삼은 이유는 실험 결과가 특정 종목에 편향되는 것을 방지하고자 하는 것에 있다. <표 2>는 <표 1>의 지수와 개별 종목 간의 2변량 Johansen 공적분 검정 중에서 Trace 검정통계량의 값을 보여준다. 여기서 귀무가설은 지수와 해당 종목간에 공적분 관계가 없다는 것이며, 이것이 기각되면, 1개 이상의 공적분 관계가 존재하여 장기적 연관이 있는 것으로 해석할 수 있다. 식 (1)에서 모형은 KOSPI와 KOSDAQ 모두 Akaike 기준을 사용하여 과거 2시차까지 포함하는 모형으로 선택하였다. <표 2>에 제시된 Trace 검정통계량들은 지수와 해당 종목간의 공적분 관계를 검정한 것으로 검정통계량의 값이 클수록 유의하게 장기적 관계에 있음을 시사한다.

<표 2> 공적분 검정에서의 Trace 검정통계량

연 번	KOSPI		KOSDAQ	
	종 목 명	Trace 검정통계량	종 목 명	Trace 검정통계량
1	삼 성 전 자	2.69	NHN	12.04
2	한 국 전 력	4.46	LG 텔레콤	14.37*
3	POSCO	11.74	하나로 텔레콤	9.18
4	SK 텔레콤	15.14**	아시아나 항공	6.14
5	하 이 닉 스	10.07	하 나 투 어	17.73**
6	현 대 차	10.78	동 서	10.83

7	KT	9.47	CJ 홈쇼핑	15.76**
8	현 대 중 공 업	8.79	다 음	11.84
9	신 세 계	9.88	휴 맥 스	9.75
10	삼 성 전 자 우	7.06	포 스 데 이 터	7.87
11	SK	5.90	태 웅	5.04
12	KT&G	7.05	GS 홈쇼핑	10.47
13	외 환 은 행	6.61	쌍 용 건 설	9.84
14	삼 성 화 재	19.98**	CJ 인터넷	13.84*
15	SK 네트웍스	11.66	네 오 위 즈	18.07**
16	S-Oil	3.37	지 엔 텍	11.81
17	기 업 은 행	13.97*	서 울 반 도 체	6.60
18	현 대 모 비 스	12.61	파 라 다 이 스	9.95
19	현 대 건 설	4.67	플래닛 82	8.11
20	LG	7.53	매 일 유 업	8.11
21	KTF	17.86**	엠 넷 미 디 어	10.04
22	삼 성 중 공 업	10.93	인 터 파 크	8.64
23	삼 성 물 산	4.34	에스에프에이	5.57
24	두 산 중 공 업	3.02	심 텍	9.82
25	기 아 차	14.63*	LG 마이크론	6.33
26	GS 건설	5.64	화 인 텍	9.22
27	현 대 산 업	12.16	KTH	14.78*
28	CJ	19.87**	유 진 기 업	18.10**
29	삼 성 증 권	8.67	HK 저축은행	6.91
30	대 우 증 권	5.56	오 디 코 프	13.41*

주: **는 유의수준 0.05에서 유의함을 의미하며, *는 유의수준 0.10에서 유의함을 의미함.

<표 3>은 Trace 검정통계량의 기초통계를 정리하였다. 이에 따르면, KOSPI 시장보다는 KOSDAQ 시장의 시가총액 상위 종목들이 지수와 보다 더 강력하게 연동되어 있음을 알 수 있다. 특히 시가 총액의 10% 이상을 차지하며 KOSPI 시장을 좌우하는 것으로 알려진 삼성전자의 낮은 유사성은 매우 주목할 만하다. 이 종목은 지수와의 연동성이 극히 떨어지는 특성상 시장의 분위기에 따른 매매는 매우 신중할 필요가 있으며, 실제로 해당 종목의 거래 주체들은 시장 분위기 보다는 해당 업체의 실적 및 향후 전망에 의하여 투자를 해왔다고 볼 수 있다.

<표 3> Trace 검정통계량의 기초 통계

KOSPI			KOSDAQ		
평 균	표준 편차	중간값	평 균	표준 편차	중간값
9.54	4.79	9.47	10.67	3.71	11.81

전반적으로 본다면, KOSDAQ 시장은 높은 평균 및 중간값과 작은 표준 편차를 보여주었다. 이를 통하여 KOSPI 시장의 개별 종목들이 KOSDAQ 시장의 개별 종목 보다 시장의 분위기에 비교적 독립적으로 매매가 이루어진다고 판단된다. 즉 KOSDAQ 시장은 개별 종목의 실적 및 특성 보다는 외부 변수 및 시장 분위기에 의하여 가격이 결정되는 정도가 더 심하기 때문에, 투자에 있어서 KOSPI 종목은 종목별 미시적인 관점에서, 그리고 KOSDAQ 종목은 시장의 거시적인 관점에서 접근을 하는 것이 보다 타당하다고 본다.

다음 <표 4>는 각 시장에서 지수와 유사성이 있다고 설명되는 종목들과 이들의 Trace 검정통계량의 값을 보여주고 있다.

<표 4> 유사성 검정으로 추출된 종목과 Trace 검정통계량

연 번	KOSPI 시장		KOSDAQ 시장	
	종 목	Trace 검정통계량	종 목	Trace 검정통계량
1	삼 성 화 재	19.98	유 진 기 업	18.10
2	CJ	19.87	네 오 위 즈	18.07
3	기 아 차	18.07	하 나 투 어	17.73
4	KTF	17.86	CJ홈쇼핑	15.76
5	SK텔레콤	15.14	KTH	14.78
6	기 업 은 행	13.97	LG텔레콤	14.37
7			CJ인터넷	13.84
8			오 디 코 프	13.41

2. 인공 신경망을 이용한 예측 모델의 구축

인공 신경망의 학습을 위하여 <표 4>의 종목들을 하나의 종목군으로 결정하여 해당 종목들의 데이터들을 수집하고 학습하였다. 학습은 공적분 검정에 사용한 데이터들과 기간이 일치하도록 2003년-2004년 데이터를 그 대상으로 하였다. KOSPI 시장에서는 6개의 종목이 추출 단계에서 선정되었으며 이들은 총 2,900여개의 개별 데이터 레코드를 포함한다. 또한 KOSDAQ 시장에서는 8개의 종목이 선정되었고, 3,900여개의 개별 데이터 레코드를 포함한다. 또한 구축된 모델의 성능을 평가하는 테스트 데이터로는 KOSPI 시장은 2,700여개 데이터를, KOSDAQ 시장은 3,600여개의 데이터를 사용하였다. 이들 테스트 데이터는 학습 데이터와 상호 배제된 데이터로써 해당 종목들의 2005년-2006년 데이터로 구성하였다.

추출 단계가 예측 성능에 기여하는 정도를 평가하기 위하여 추출 단계에서 구성된 종목군과 동일한 개수의 종목을 랜덤하게 선택하여 인공 신경망 모델을 학습하였다. 그리고 인공 신경망 모델의 성능을 비교분석하기 위하여 선형 회귀분석(Linear Regression) 모델을 구축하여 그 성능을 살펴보았다. 모델의 예측 정확도를 측정하는 방법으로는 예측값(predicted value)과 목표값(target value)의 상관 계수 (Correlation Coefficient)와 예측값과 목표값의 차를 이용한 제곱근-상대-제곱 오차 (Root Relative Squared Error: RRSE)를 사용하였다. 이 중 RRSE는 다음과 같은 방식으로 계산된다 (Gepsoft 2007).

$$E = \sqrt{\frac{\sum_{j=1}^{n}(P_j - T_j)^2}{\sum_{j=1}^{n}(T_j - \overline{T})^2}} \qquad (5)$$

여기서 P_j는 예측 시스템이 테스트 데이터 j에 대하여 예측한 값이고, T_j는 데이터 j의 목표값이다. 그리고 $\overline{T}$는 $\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$ 를 통하여 계산되며 전체 데이터의 개수는 n이다. 이 평가 방식은 기본적으로 예측값과 목표값의 차를 오차 E라 하여 실제 예측의 정확도를 평가하는 것인데, 오차의 정도를 가상 예측 시스

템이 모든 예측을 목표값의 평균치로 시도하였을 때와 비교하여 그 상대적인 정확도를 계산하는 것이다.

<표 5> 선형 회귀분석 모델의 예측 정확도

KOSPI				KOSDAQ			
상관 계수		RRSE(%)		상관 계수		RRSE(%)	
공적분	랜덤	공적분	랜덤	공적분	랜덤	공적분	랜덤
0.9992	0.7188	4.04	120.12	0.9902	0.9452	15.03	32.64

<표 6> 인공 신경망 모델의 예측 정확도

KOSPI				KOSDAQ			
상관 계수		RRSE(%)		상관 계수		RRSE(%)	
공적분	랜덤	공적분	랜덤	공적분	랜덤	공적분	랜덤
0.9999	0.9965	0.72	8.75	0.9999	0.9973	2.32	7.62

<표 5>와 <표 6>은 예측 모델로써 선형 회귀분석 및 인공 신경망을 선택하였을 때의 예측 정확도에 대하여 보여준다. 전체적인 모델의 정확도를 비교하여 본다면, 모든 경우에 있어서 인공 신경망은 선형 회귀분석에 비하여 매우 높은 성능을 보여주었다. 따라서 주가 예측과 같은 문제에 있어서는 선형적인 모델보다는 비선형적인 모델이 월등히 좋은 결과를 보여줄 수 있다는 점을 확인할 수 있었다.

추출 과정을 거친 신경망 모델에서는 KOSPI의 경우가 KOSDAQ의 경우보다 더 높은 성능을 보여주었다. 반면 랜덤하게 추출한 종목들로 학습을 한 모델에서는 KOSDAQ의 경우가 미세하지만 더 좋은 성능을 보여주었다. 이는 추출 과정을 통하여 선택된 종목들의 유사도의 차이에서 비롯되는 것으로 보인다. 즉, KOSPI의 경우는 해당 종목들의 Trace 검정통계량의 중간값이 17.86이지만 KOSDAQ은 중간값이 14.78에 불과할 정도로 지수와의 유사성이 떨어진다. 이러한 차이가 학습의 성능에 영향을 미친 것으로 보인다. 반면에 랜덤 추출의 경우는 KOSDAQ이 더 많은 학습 데이터를 사용하였기 때문에 약간이나마 더 좋은 성능을 보인 것으로 추정된다. 최적화해야 하는 계수가 선형 회귀분석에 비하여 월등히 많은 인공 신경망은 훨씬 더 많은 학습 데이터가 필요하다.

Ⅴ. 결 론

본 연구에서는 주가의 예측을 위하여 2단계 하이브리드 예측모델을 제시하였다. 첫째, 추출 단계에서는 Johansen의 공적분 검정을 통하여 매매를 원하는 종목과 종가 시계열의 유사성이 높은 종목들을 추출한다. 둘째, 추출된 종목들을 대상으로 학습 데이터를 구축하고 이 데이터로 인공 신경망을 학습하여 예측 모델을 수립한다. 이러한 2단계 모델을 통하여 인공 신경망의 예측 성능을 향상시키고, 유사 종목군을 구성하여 종목의 심층 분석을 가능하게 할 것이다.

향후에 보다 향상된 예측 성능으로, 실제 매매에 있어 현실적으로 적용할 수 있는 예측 시스템을 개발하기 위하여 다음과 같은 후속 연구가 필요하다고 판단된다. 첫째, 시스템의 구현을 통한 예측의 자동화이다. 현재로서는 모델의 구축에 초점이 맞추어져 있기 때문에 추출 단계와 학습 단계가 서로 별도의 단계로서 구성되어 있다. 이 들 두 단계를 하나의 시스템으로 통합하여 매매를 원하는 종목이 결정되면 바로 예측치를 생성하여 투자자의 의사 결정을 도울 수 있도록 하여야 한다. 둘째, 매매와 연동된 시스템의 설계이다. 이 모

델은 예측만을 위한 모델이다. 그러나 투자 수익의 향상을 위해서는 예측뿐만 아니라 매매 정책의 최적화 역시 필요하다. 이러한 최적화에 적합한 모델을 구축하고 해당 모델을 실제 시스템으로 구현하여야 한다. 마지막으로 다양한 예측 모델의 구축이다. 본 연구는 인공 신경망에 기반한 모델에 관한 것이다. 그러나 동일한 인공 신경망이라 하더라도, 사용되는 데이터의 속성이 틀리면 전혀 다른 예측 결과가 나올 수 있다. 이처럼 이종의 예측 모델을 다수 구축하고, 이들 모델로부터 통합된 예측 결과를 얻게 된다면 보다 보완된 예측 모델을 구축할 수 있을 것이다.

참 고 문 헌

김유섭·이재원·이종우(2004), 「다중 에이전트 Q-학습 구조에 기반한 주식 매매 시스템의 최적화」, 『정보처리학회논문지 B』, 제11-B권 제2호,

Armano, G., Marchesi, M., and Murru, A., "A Hybrid Genetic-Neural Architecture for Stock Indexes Forecasting," *Information Sciences*, 170, 2005.

Dempster, M. A. H., Payne, T. W., Romahi, Y., and Thompson, G. W. P., "Computational Learning Techniques for Intraday FX Trading Using Popular Technical Indicators," *IEEE Transactions on Neural Networks*, 12(4), 2001.

Fan, A. and Palaniswami, M., "Stock Selection Using Support Vector Machines," *In Proceedings of International Joint Conference on Neural Networks*, 2001.

Fama, E. F., "Multiperiod Consumption Investment Decisions," *American Economic Review*, 60, 1988.

Gepsoft, http://www.gepsoft.com/GXPT4KB/Chapter10/section1/SS07.htm, 2007.

Ghosn, J. and Bengio, Y. "Multi-Task Learning for Stock Selection," *Advances in Neural Information Processing Systems*, 9, M. C. Mozer, M. I. Jordan and T. Petsche editor, The MIT Press, 1997.

Johansen, S., "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control* 12, pp.231-254, 1988.

Kendall, S. M. and Ord, K. *Time Series.* Oxford, 1997.

Kim, S. D, Lee, J. W., Lee, J. and Chae, J.-S., "A Two-Phase Stock Trading System Using Distributional Differences," *Proceedings of International Conference on Database and Expert Systems Applications*, 2002.

Koscom, http://www.koscom.co.kr, 2007.

Malkiel, B. G., *A Random Walk Down Wall Street*, Norton, 1996.

Mitchell, T., *Machine Learning*, McGraw Hill, 1997.

Saad, E. W., Prokhorov, D. V. and Wunsch II, D. C., "Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks," *IEEE Transactions on Neural Networks*, 9(6), 1998.