2단계 하이브리드 주가 예측 모델 : 공적분 검정과 인공 신경망

오 유 진† · 김 유 섭††

요 약

본 논문에서는 주가예측의 정확도를 향상시키기 위하여 공적분 검정(Cointegration Tests)과 인공 신경망(Artificial Neural Networks)을 사용한 2단계 하이브리드 예측 모델을 제시한다. 기존의 연구에서는 예측을 시도하고자 하는 종목의 일자별 개별 레코드를 인공 신경망과 같은 방법으로 학습함으로써 주식 데이터가 가지는 시계열적 특성을 충분히 반영하지 못하였는데, 새로 제안한 모형에서는 주식자료의 과거시차들의 값들도 인공 신경망의 속성(feature)으로 사용하여 기존 연구의 한계를 보완하였다. 또한, 예측대상종목의 정보들 외에도 장기적으로 높은 시계열 유사성을 보유한 종목들을 선발한 후 속성으로 사용하여 모형의 예측성능을 향상 시켰다. 구체적으로 1단계는 Johansen의 공적분 검정을 통하여 예측대상종목과 장기적 관계(long-term relationship)에 있는 종목을 추출하고, 2단계는 이 선발된 종목들과 예측대상종목의 시계열 정보 특성을 속성으로 구축한 인공 신경망으로 학습하여 관심 종목을 예측한다. 제안된 모델의 성능을 확인하기 위하여 KOSPI 지수의 방향성을 예측하는 시스템을 구현하였으며, 시가총액 상위 종목군을 대상으로 지수와의 공적분 검정을 하였다. 성능을 살펴보기 위하여 본 연구에서는 시계열 정보가 속성으로 반영된 단순 인공 신경망 모델, 공적분 검정을 통과한 종목들의 시계열 속성이 포함된 모델, 그리고 그 모델과 속성의 개수를 동일하게 하기 위하여 임의로 종목을 선택하여 이들의 시계열 속성이 포함된 모델을 구축하였다. 실험 결과 공적분 검정을 통과한 종목군의 속성이 결합된 모델은 단순 인공 신경망만으로 학습된 기존 모델에 비하여 평균적으로는 11.29% (최대 29.98%) 정확도가 향상되었고, 임의로 선택된 종목군의 속성이 결합된 모델에 비해서는 평균적으로는 10.59% (최대 25.78%) 가 향상된 예측 정확도를 보여주었다.

키워드 : 공적분 검정, 인공 신경망, 주가예측 모델, KOSPI

A Two-Phase Hybrid Stock Price Forecasting Model : Cointegration Tests and Artificial Neural Networks

Oh, Yujin† · Kim, Yu-Seop††

ABSTRACT

In this research, we proposed a two-phase hybrid stock price forecasting model with cointegration tests and artificial neural networks. Using not only the related stocks to the target stock but also the past information as input features in neural networks, the new model showed an improved performance in forecasting than that of the usual neural networks. Firstly in order to extract stocks which have long run relationships with the target stock, we made use of Johansen's cointegration test. In stock market, some stocks are apt to vary similarly and these phenomenon can be very informative to forecast the target stock. Johansen's cointegration test provides whether variables are related and whether the relationship is statistically significant. Secondly, we learned the model which includes lagged variables of the target and related stocks in addition to other characteristics of them. Although former research usually did not incorporate those variables, it is well known that most economic time series data are depend on its past value. Also, it is common in econometric literatures to consider lagged values as dependent variables. We implemented a price direction forecasting system for KOSPI index to examine the performance of the proposed model. As the result, our model had 11.29% higher forecasting accuracy on average than the model learned without cointegration test and also showed 10.59% higher on average than the model which randomly selected stocks to make the size of the feature set same as that of the proposed model.

Key Words : Co-Integration Tests, Artificial Neural Networks, Stock Price Forcasting Model, KOSPI

1. 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 [1]. 또한 [2] 역시 시계열의 입장에서 많은 기법들을 활용하였으며, [3]에서는 시장을 완벽하게 예측하는 것은 효율적 시장 가설 (Efficient Market Hypothesis)의 입장에서는 불가능하나 어느 정도의 이익은

※ 이 논문은 2007년도 한림대학교 교비 학술연구비(HRF-2007-041)에 의하여 연구되었음.
† 정 회 원 : 고려대학교 경영전문대학원 연구교수
†† 종신회원 : 한림대학교 정보통신공학부 부교수 (교신저자)
논문접수 : 2007년 4월 23일, 심사완료 : 2007년 10월 1일

유도가 가능하다고 하였다. 특히 인공지능 분야의 알고리즘들이 강력한 표현력과 모델링 능력을 지니고 있기 때문에, 이들을 이용한 예측의 사례들이 보고되고 있는데 [4], 인공신경망(Artificial Neural Network), 결정 트리(Decision Tree), SVM(Support Vector Machine) 등이 이에 적용되어 왔다[5-8].

주가 예측에 있어서 인공 지능 측면에서의 이러한 시도들은 수학적인 모델에 비하여 매우 단순하지만 주목할 만한 성능을 보여주었다. 하지만, 실제 매매에 있어서는 성능의 한계를 보여주었기 때문에, 이를 극복하기 위하여 다양한 방법론이 시도되었는데, [9]는 매매 행위를 모델링하여 예측 행위를 보완하였고, [10]은 다양한 예측 모델을 통합한 앙상블 (Ensemble) 모델을 수립하였다. 하지만 이러한 방법론은 기본적인 예측 모델의 성능에 따라 그 성능이 좌우되기에, 기본 예측 모델의 성능 향상 문제는 꾸준히 제기되고 있으며 개선되어야 할 사항이다.

다양한 인공 지능 알고리즘 중에서 인공 신경망은 높은 성능과 구현의 용이성으로 가장 널리 사용되고 있다 [5, 7, 8, 9, 10, 11]. 그러나 기존 인공 신경망의 학습 데이터 구축에는 주식 데이터가 가지는 시계열적 특성을 제대로 반영할 수 없었다. 이는 신경망 학습이 횡단면 (cross-sectional) 자료를 가지고 학습을 진행하기 때문인데, 이는 주식 데이터의 시계열성을 간과한 것이다. 즉 신경망에서는 데이터가 상호 독립적이라는 가정 아래에서 일자별 개별 데이터 레코드들을 단순히 수집하여 이들을 학습 데이터로 활용하였다. 시계열 변동성이 높은 주식의 특성을 감안하여, 시계열적 속성(feature)을 학습 모델 구축 시 반영할 필요가 있다고 판단된다. 또한 단일 종목의 특성만이 반영된 데이터의 학습은 학습된 모델이 해당 종목의 과거 움직임에 지나치게 의존하게 되어 향후 발생하게 될 보다 다양한 움직임을 예측하는데 한계를 보이게 된다.

따라서 본 논문에서는 가장 널리 사용되고 있는 인공 신경망에 기반한 예측 모델의 기본적인 성능 향상을 위하여 공적분 검정을 이용한 2단계 하이브리드 예측 모형을 제시하였다. 1단계에서는 예측을 시도하고자 하는 종목과 Johansen의 공적분 검정[12]을 하여 통과한 종목들을 추출하여 유사 종목군을 구성한다. 이 단계에서는 매매를 위하여 미래 가치를 예측하고자 하는 종목과 장기적으로 유사한 시계열 패턴을 보였던 종목군을 추출하는데, 이는 기본적으로 유사한 추세를 보이지만 보다 다양한 움직임을 학습에 반영할 수 있도록 하여 향후 학습될 모델이 주식 데이터의 시계열 특성을 보다 강력히 반영할 수 있도록 하기 위함이다. 또한 통계학과 경제학에서 주식 데이터를 분석할 때에는, 한 종목만을 가지고 분석하는 것은 매우 어려운 것으로 알려져 있기 때문에, 일반적으로 유사 시계열간의 관계를 포함하여 분석한다. 이처럼 주식 데이터의 분석에 있어서 유사 종목군의 추출은 필수적인데, 본 논문에서 유사 종목군을 구성하기 위하여 공적분 검정을 사용한 이유는 다음과 같다. Granger and Newbold [13]는 회귀분석으로 확률보행(random walk)인 불안정(nonstationary) 시계열 자료들 간의 관계를 분석할 때, 실제로는 서로 관련이 없는데도 분석 결과는 매우 유의하다고 판정되는 이른바 가성적 회귀(spurious regression)현상이 발생할 소지가 있다고 하였다. 또한 Phillips [14]는 가성적 회귀현상 하에서는 추정된 계수의 분포가 기존의 t-분포로 수렴하지 않고, 발산함을 보여, 가성적 회귀현상이 초래하는 통계적 문제점을 수리적으로 증명하였다. 이를 방지하기 위하여 Engle and Granger [15] 그리고 Johansen [12]은 공적분이라는 통계적 기법을 제안하였으며, 그 이후부터는 이들 기법인 공적분 검정과 오차수정모형(Vector error correction model)이 주로 이용되고 있는 추세이다. 주식은 대표적인 확률보행과정으로 알려져 있기 때문에, 본 연구에서는 유사 종목군을 선별하기 위하여 공적분 검정을 사용하였다. 2단계에서는 추출된 종목군들의 시계열 특성을 반영하여 인공 신경망 모델을 구축한다. 즉 인공 신경망의 입력 속성을 설계할 때, 해당 종목의 과거 시차 정보뿐만 아니라, 시계열이 유사한 여러 종목들의 과거 시차 정보를 통합한다.

본 연구에서 제안한 예측 모델의 실증 분석을 위하여 KOSPI 지수의 방향성을 예측하는 시스템을 구현하였고, 지수와 유사한 종목은 KOSPI 시장의 시가총액 상위 종목 중에서 선택하도록 하였다. 인공 신경망은 뚜렷한 특성을 주로 학습하는 경향이 있어, 이를 이용한 [7]과 같은 모델은 주로 가격 변동폭이 크고 가격에 영향을 미치는 요소가 비교적 적은 소형주 위주로 추천이 이루어졌다. 본 연구에서는 이러한 문제를 상쇄하기 위하여 대형주들에 적용하기 용이하도록 지수를 대상으로 하는 예측 모델을 구축하였고 이 모델은 지수가 아닌 일반 대형주에 대해서도 쉽게 적용될 수 있다. 실험 결과 공적분 검정을 통과한 종목군의 속성이 결합된 모델은 단순 인공 신경망만으로 학습된 모델에 비하여 평균적으로는 11.29%(최대 29.98%)의 정확도가 향상되었다. 또한 실험에서는 입력 벡터의 크기의 차이로 인한 성능의 차이가 생기지 않게 하기 위하여, 임의로 선택된 종목군의 속성을 입력 벡터에 추가한 모델을 수립하였는데, 공적분을 활용한 모델은 이 모델에 비해서도 평균적으로는 10.59% (최대 25.78%) 가 향상된 예측 정확도를 보여주었다. 또한 일반적인 예상과는 달리 학습 데이터의 크기는 학습 모델의 예측 정확도와 큰 상관관계는 보여주지 못하였다.

본 논문은 2절에서는 본 연구에 사용된 자료의 성격에 대하여 설명하고 학습 모델에 사용된 속성들에 대하여 설명한다. 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 논한 후에 5절에서 결론 및 향후 연구에 대하여 논하였다.

2. 자료 및 속성

2.1 KOSPI 종목 자료

〈표 1〉 실험에 사용된 KOSPI 시장에서의 시가 총액 상위 30 종목들

연 번	종 목 명	시가 총액 (백만원)	연 번	종 목 명	시가 총액 (백만원)
1	삼 성 전 자	86,464,711	16	S-Oil	7,261,590
2	한 국 전 력	28,549,763	17	기 업 은 행	6,968,607
3	POSCO	27,769,007	18	현 대 모 비 스	6,573,197
4	SK 텔레콤	15,995,161	19	현 대 건 설	5,343,681
5	하 이 닉 스	15,124,775	20	LG	4,952,390
6	현 대 차	14,572,559	21	KTF	4,941,453
7	KT	12,890,823	22	삼 성 중 공 업	4,836,780
8	현 대 중 공 업	10,488,000	23	삼 성 물 산	4,803,696
9	신 세 계	10,392,136	24	두 산 중 공 업	4,572,685
10	삼 성 전 자 우	10,389,209	25	기 아 차	4,079,598
11	SK	9,007,742	26	GS 건설	3,972,900
12	KT&G	8,463,199	27	현 대 산 업	3,806,901
13	외 환 은 행	7,771,127	28	CJ	3,232,427
14	삼 성 화 재	7,575,600	29	삼 성 증 권	3,064,392
15	SK 네트웍스	7,288,065	30	대 우 증 권	3,022,604

본 연구에서는 지수의 방향성을 예측하기 위하여 KOSPI 시장의 시가 총액 상위 30 종목을 대상으로 지수와의 공적분 검정을 실시하였다. <표 1>은 KOSPI 시장에서 본 실험을 위하여 추출된 종목들의 리스트로써 이 종목들의 2007년 1월 현재 시가 총액을 보여주고 있다 [16].

본 논문의 실험을 위해서는 최소한 2003년도부터 2006년도까지의 4년간의 데이터가 축적된 종목이 필요하다. 따라서 최근에 설립 또는 인수/합병으로 변동이 심하고, 꾸준한 데이터가 없는 종목들은 시가총액의 크기에 불문하고 실험 대상에서 제외하였다. 예를 들어 신한지주, 우리금융, 롯데쇼핑과 같은 종목들은 2003년부터의 데이터가 없어 상당히 높은 시가총액 순위에도 불구하고 실험 대상에서 제외하였다. 신한지주와 우리금융의 경우에는 일반 은행으로써 오래전부터의 데이터가 축적되어 있으나, 2006년도 연말을 기준으로 하였을 때, 최근에 금융 지주 회사로 변경되면서 관련 회사들과 통합되었기 때문에, 과거와 현재가 연속성을 가졌다고 인정하기 어렵다. 그리고 롯데쇼핑은 상장된지 얼마 되지 않았기 때문에 충분한 데이터를 얻을 수 없었다. 그런데, 이외의 일부 종목들에서도 거래 정지, 감자 또는 증자로 인하여 가격 및 거래량에 있어서 비정상적인 움직임을 보인 기간들이 있었으나, 전체 실험 데이터의 크기에 비하여 무시할 수 있는 정도라 보고 이를 일반적인 움직임으로 여기고 특별하게 처리하지 않았다.

2.2 입력 속성의 설계

모든 종목들은 매매에 따라서 가격 및 거래량과 관련한 여러 데이터를 매일 생성해낸다. 본 논문에서는 시중의 가정 매매 시스템(HTS: Home Trading System)[17]을 통하여 쉽게 구할 수 있는 데이터를 토대로 하여 이를 변환 및 가공하여 예측에 활용한다. 일반적인 HTS는 지수를 포함하여 모든 종목이 특정 일자에 보여준 시가(*SP*), 고가(*HP*), 저가(*LP*), 종가(*CP*)와 거래량(*VOL*)을 나타내는 수치를 비롯하여 5일, 10일, 20일, 60일, 120일 동안의 가격 이동평균(*MA_P*)와 거래량 이동평균(*MA_V*)을 기초 데이터로 제공한다. 여기서 t 시점에서의 과거 d일 동안의 가격 이동평균 $MA_P_d(t)$와 거래량 이동평균 $MA_V_d(t)$는 다음과 같은 방식으로 계산된다.

$$MA\_P_d(t) = \frac{\sum_{i=0}^{d-1} CP(t-i)}{d}, \qquad MA\_V_d(t) = \frac{\sum_{i=0}^{d-1} Vol(t-i)}{d} \quad (1)$$

본 연구에서는 이들 데이터 중에서 종가의 시계열(*CP(t)*)만을 그 대상으로 하여 공적분 검정을 적용하였다. 공적분 검정으로 지수와 개별 종목들의 유사성을 분석하여 인공 신경망의 속성에 반영할 유사 종목들을 추출한다. 한편, 인공 신경망을 활용한 예측 시스템에서는 이들 기초 데이터를 가공하여 새로운 데이터를 생성 및 추출하여 활용한다. 인공 신경망에서 사용한 입력 속성 변수들은 <표 2>에 정리되어 있다.

〈표 2〉 인공 신경망 입력 속성 변수

No.	변 수 명	변 수 설 명	비고
1	SL_P5(t)	5일 가격 이동평균선의 기울기	관심 종목의 횡단면적 속성
2	SL_P10(t)	10일 가격 이동평균선의 기울기	
3	SL_P20(t)	20일 가격 이동평균선의 기울기	
4	SL_P60(t)	60일 가격 이동평균선의 기울기	
5	SL_P120(t)	120일 가격 이동평균선의 기울기	
6	SL_V5(t)	5일 거래량 이동평균선의 기울기	
7	SL_V10(t)	10일 거래량 이동평균선의 기울기	
8	SL_V20(t)	20일 거래량 이동평균선의 기울기	
9	GA_P5(t)	5일 가격 이동평균선과 t시점의 종가의 차이	
10	GA_P10(t)	10일 가격 이동평균선과 t시점의 종가의 차이	
11	GA_P20(t)	20일 가격 이동평균선과 t시점의 종가의 차이	
12	GA_P60(t)	60일 가격 이동평균선과 t시점의 종가의 차이	
13	GA_P120(t)	120일 가격 이동평균선과 t시점의 종가의 차이	
14	GA_V5(t)	5일 거래량 이동평균선과 t시점의 거래량과의 차이	
15	GA_V10(t)	10일 거래량 이동평균선과 t시점의 거래량과의 차이	
16	GA_V20(t)	20일 거래량 이동평균선과 t시점의 거래량과의 차이	
17	Body(t)	t 시점에서의 시가대비 종가	
18	RP(t-1)	t 시점 종가대비 t-1 시점 종가	관심 종목의 시계열 속성
19	RP(t-2)	t 시점 종가대비 t-2 시점 종가	
20	RPc(t-1)	공적분 검정을 통과한 종목의 t 시점 종가대비 t-1 시점 종가	유사 종목들의 시계열 속성
21	RPc(t-2)	공적분 검정을 통과한 종목의 t 시점 종가대비 t-2 시점 종가	

1-19번은 HTS에서 구한 기초 데이터로부터 가격 및 거래량 이동 평균선의 기울기, 가격 및 거래량의 종가와 이동평균선의 간격, 시가대비 종가의 변동 비율, 당일 종가대비

〈표 3〉 인공 신경망 입력 속성 계산 방법

$$SL\_P_d(t) = \frac{MA\_P_d(t) - MA\_P_d(t-1)}{MA\_P_d(t-1)} \times 100, \quad SL\_V_d(t) = \frac{MA\_V_d(t) - MA\_V_d(t-1)}{MA\_V_d(t-1)} \times 100$$

$$GA\_P_d(t) = \frac{CP(t) - MA\_P_d(t)}{MA\_P_d(t)} \times 100, \quad GA\_V_d(t) = \frac{Vol(t) - MA\_V_d(t)}{MA\_V_d(t)} \times 100$$

$$Body(t) = \frac{CP(t) - SP(t)}{SP(t)} \times 100$$

$$RP(t-i) = \frac{CP(t-i) - CP(t)}{CP(t)} \times 100$$

$$RP_c(t-i) = \frac{CP_c(t-i) - CP_c(t)}{CP_c(t)} \times 100$$

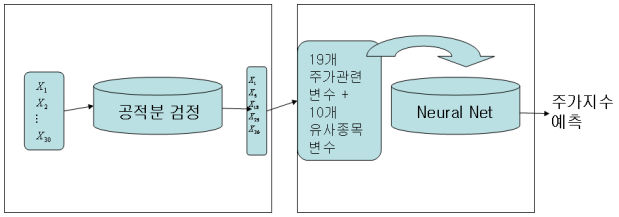
전일 및 전전일 종가의 변동 비율인 관심 종목 속성이며, 20번과 21번은 공적분 검정을 통과한 유사 종목들의 전일 및 전전일 종가의 변동 비율이다.

<표 3>은 위에서 설명한 속성을 계산하는 방법을 구체적으로 보여주고 있다. 여기서 $RP_c(t-1)$, $RP_c(t-2)$ 은 공적분 검정을 통과한 종목의 수에 따라서 그 개수가 정해진다.

$RP(t-1)$과 $RP(t-2)$은 각각 $t-1$ 및 $t-2$ 시점의 종가를 의미하는데, 이를 통하여 학습될 모델이 종목들의 시계열 특성을 반영할 수 있도록 하였다. 또한 $RP_c(t-1)$과 $RP_c(t-2)$는 공적분 검정을 통과한 유사 종목들의 시계열 특성을 입력 속성에 반영함으로써 유사한 시계열을 가지고 있으나 보다 다양한 움직임을 보이는 특성을 모델에서 학습할 수 있도록 하였다. 공적분 검정 결과 유사 시계열이 n개인 경우 n쌍의 $RP_c(t-1)$과 $RP_c(t-2)$ 값을 입력 속성으로 사용한다. 여기서 부분 시계열 정보를 나타내기 위하여 과거 2시차의 정보만을 입력 속성에 포함시켰는데, 이는 본 연구의 경우 3장에서 설명하고 있는 VAR 모델에서 과거 2시점까지를 포함할 경우에 최적의 결과를 얻을 수 있다는 것을 실험적으로 발견하였기에 그렇게 한 것이다.

3. 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. 전체 과정은 (그림 1)에서 간단하게 확인할 수 있다. (그림 1)을 간단히 설명하면 다음과 같다. 먼저 KOSPI 시가 총액 상위 30개 종목을 선택하고 이들 개별 종목과 지수와의 공적분 검정을 하고, 5개의 종목이 유사성이 있는 것으로 판별되었다고 하자. 그러면 5개의 종목들에 대하여 각각 이들의 t-1, t-2 시점의 종가 비율을 계산하고 이들을 인공 신경망의 입력 속성에 포함시킨다. 이 때, 지수와 관련한 19개의 입력 속성(t-1, t-2 시점의 종가 비율을 포함한)은 이미 인공 신경망의 입력 속성으로 설계되어 있다고 가정할 수 있다. 지수의 기존 19개 속성과 유사 종목의 10개의 속성, 총 29개의 속성이 인공 신경망의 입력 속성이 되고 이를 토대로 학습이 이루어진 후에 예측 모델이 구축된다.



(그림 1) 주가 예측 모델 구축 개요

3.1 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[12]이 가장 일반적으로 사용되고 있다. Johansen의 최우추정법은 다변수 분석틀에서 진단할 수 있으며, 모수 추정치에 대해 명백히 가설 진단을 수행할 수 있다는 장점이 있다. 다음은 벡터자기회귀모형(VAR, Vector Autoregressive Model)이다.

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

여기서 z_t는 k개의 확률보행을 따르는 변수들을 포함하고 있는 $(k \times 1)$ 확률변수벡터, $D(t)$는 추세(trend), ϕ_i는 $(k \times k)$ 모수행렬, a_t는 $(k \times 1)$ 오차벡터, 그리고 p는

VAR모형의 차수를 각각 의미한다.

식(2)를 오차수정모형(ECM, Error Correction Model)으로 전환하면 다음 식(3)과 같으며,

$$\Delta z_t = \alpha + D(t) + \Pi z_{t-1} + \sum_{i=1}^{p-1} \phi_i^* \Delta z_{t-i} + a_t \quad (3)$$

여기서 $\Pi = -I + \phi_1 + ... + \phi_p$, $\phi_i^* = -\sum_{i=i+1}^{p} \phi_j$, $i = 1,...,p-1$, 그리고 $\Delta z_t = z_t - z_{t-1}$이다. 여기서 관심사는 행렬 Π에 있다. 공적분의 개수는 행렬 Π의 rank만큼 존재하며, 만약 $m(= rank(\Pi))$개의 공적분이 존재할 경우, $(k \times k)$ 행렬 Π는 $\Pi = BA'$와 같이 표현되어 지며, 여기서 $(m \times k)$ 행렬 A'의 각 행은 z_t에 포함된 변수들 간의 공적분 관계를 의미하는 선형관계이며, $(k \times m)$ 행렬 B는 계수행렬이 된다. 즉, $Rank(\Pi) = 0$이면, 변수들 간에 공적분 관계가 존재하지 않는다는 것을 의미하며, $Rank(\Pi) = m > 0$이면 z_t는 m개의 공적분 관계를 가진다.

Johansen의 공적분 검정은 Trace 검정과 Max검정의 2가지로 구축된다. 우선 Trace 검정의 귀무가설과 대립가설은 각각 $H_0 : m = m_0 \quad vs. \quad H_a : m > m_0$으로 설정되며, 여기서 k를 Π의 행의 개수라고 하였을 때, m_0는 0과 $k-1$사이의 수이다. Johansen의 Trace 검정통계량 $L_{tr}(m_o)$은

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

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

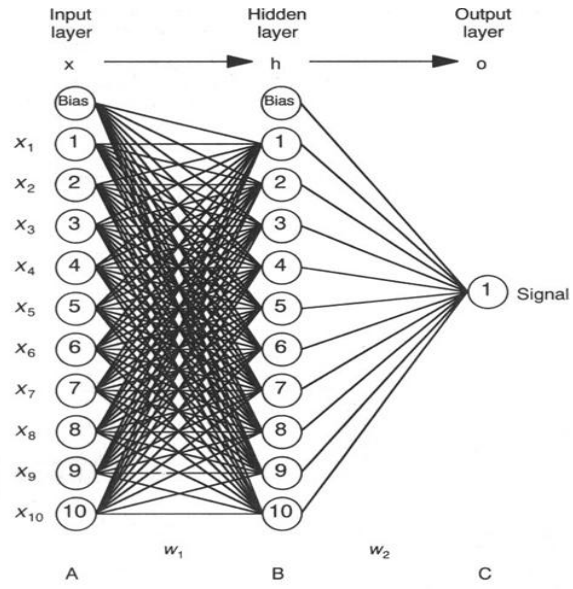
한편, Max검정의 가설들은 각각 $H_0 : m = m_0 \quad vs. \quad H_a : m = m_0 + 1$이며, Johansen의 Max 검정통계량 $L_{\max}(m_o)$은

$$L_{\max}(m_0) = -(T - kp) \sum_{i=m_o+1}^{k} \ln(1 - \lambda_{m_0+1}) \quad (5)$$

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

3.2 학습 단계 - 인공 신경망

(그림 2)은 학습 단계에서 사용되는 인공 신경망[18]의 대략적인 모습을 보여준다. 본 연구에서 사용된 인공 신경망은 3개의 층으로 이루어진다. 그리고 각 층의 모든 노드들은 다른 층의 모든 노드들과 간선으로 연결되어 있으며, 각 간선에는 가중치 값이 부여되어 있다. 첫째, 입력층(그림 2의 A층)은 실제 데이터의 입력이 이루어지는 곳으로써 하나



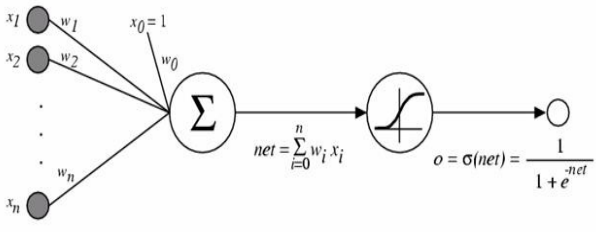
(그림 2) 인공 신경망의 전체 구조

의 데이터 레코드는 수십 또는 수백개의 속성으로 이루어지고 각각의 개별 속성값들이 적절한 정규화 과정을 거쳐서 입력층의 개별 노드에 입력된다. 본 연구에서는 2.2에서 계산된 속성값들이 x_i에 각각 입력된다. 둘째, 출력층(그림 2의 C층)은 입력된 개별 레코드의 인공 신경망에서의 최종 결과값이 출력되는 층으로써, 본 연구에서는 다음과 같은 값이 출력되도록 하였다.

$$C = \begin{cases} +1 \text{ if } CP(t+1) - CP(t) \geq 0 \\ -1 \quad otherwise \end{cases} \quad (6)$$

그리고 마지막으로 은닉층(그림 2에서 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 정하였으며 단일 계층으로 설계하였다.

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 (그림 3)와 같은 구조를 가진다[14].



(그림 3) 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, 본 연구에서는 인공 신경망 알고리즘에서 가장 널리 사용되는 역전파 (back propagation) 알고리즘을 사용하여 w_i 값을 학습하였으며, 이를 간략히 요약하면 다음과 같다.

먼저 모든 가중치 값은 랜덤 값을 부여한다. 본 연구에서는 0을 초기값으로 설정하였다. 그리고 가지고 있는 학습 예제 모두에 대하여 해당 예제를 네트워크의 입력으로 하여 네트워크 출력값을 계산한다. 그리고, 모든 출력 단위 (output unit) k에 대하여

$$\delta_k \leftarrow o_k(1-o_k)(t_k-o_k) \tag{7}$$

을 계산한다. 여기서 o_k는 실제 계산된 출력값을 그리고 t_k는 목적값을 의미한다. 또한 은닉 단위 (hidden unit) h에 대하여

$$\delta_h \leftarrow o_h(1-o_h)\sum_{k\in outputs} w_{h,k}\delta_k \tag{8}$$

를 계산한다. 마지막으로 각각의 네트워크 가중치 $w_{i,j}$를

$$w_{i,j} \leftarrow w_{i,j} + \triangle w_{i,j} \tag{9}$$

의 방식으로 갱신한다. 이 때,

$$\triangle w_{i,j} = \eta \delta_j x_{i,j} \tag{10}$$

이다. 여기서 η는 학습 비율 (learning rate)를 말하는데, 본 연구에서는 이 값을 0.3으로 하였다. 그리고 $x_{i,j}$는 노드 j에서 i로의 입력값이다. 본 연구에서는 모멘텀 (momentum)을 사용하여 국소 최소화 (local minima) 의 문제를 해결하려고 하였으며

$$\triangle w_{i,j}(n) = \eta \delta_j x_{i,j} + \alpha \triangle w_{i,j}(n-1) \tag{11}$$

의 방식으로 적용하였다. 여기서 $0 \le \alpha < 1$ 가 모멘텀인데, 본 연구에서는 이 값을 0.2로 하였으며, $\triangle w_{i,j}(n)$은 n번째 순환에서 적용된 가중치 변경치를 말한다. 본 연구에서는 실험적으로 학습 비율과 모멘텀의 값을 구하여 사용하였다.

4. 실험 및 평가

4.1 Johansen 공적분 검정을 통한 유사 종목 추출

본 연구에서는 실험을 위하여 KOSPI 시장의 지수를 기준 시계열로 보고 <표 1>의 종목 중에서 이와 유사한 시계열을 추출하였는데, 공적분 검정 기간을 기준으로 3개의 공적분 검정을 실시하였다. 이를 각각 KOSPI1, KOSPI2, KOSPI3라 하였는데, 이들의 검정 기간은 KOSPI1이 2003년-2004년, KOSPI2가 2003년-2005년, 그리고 KOSPI3가 2003년-2006년 6월로 정하였다. 이와 관련하여 공적분 검정에 사용된 시계열의 길이는 KOSPI1이 493, KOSPI2가 743, KOSPI3가 867이다. 본 연구에서 지수를 기준으로 삼은 이유는 실험 결과가 특정 종목의 특성에 편향되는 것을 방지하여 보다 일반적인 경우를 반영하기 위함이다. <표 3>은 KOSPI 지수와 <표 1>의 개별 종목 간의 2변량 Johansen 공적분 검정 중에서 Trace 검정통계량의 값을 보여준다. 여기서 귀무가설은 지수와 해당 종목 간에 공적분 관계가 없다는 것이며, 이것이 기각되면, 1개 이상의 공적분 관계가 존재하여 장기적 연관이 있는 것으로 해석할 수 있다. 식 (2)에서 모형은 KOSPI1, KOSPI2, KOSPI3 모두 Akaike 기준[1]을 사용하여 과거 2시차까지 포함하는 모형으로 선택하였다. <표 4>에 제시된 Trace 검정통계량들은 지수와 해당 종목간의 공적분 관계를 검정한 것으로 검정통계량의 값이 클수록 유의하게 장기적 관계에 있음을 시사한다.

〈표 4〉 공적분 검정에서의 Trace 검정통계량

연 번	종 목 명	KOSPI1	KOSPI2	KOSPI3
1	삼 성 전 자	2.69	7.79	6.32
2	한 국 전 력	4.46	9.61	10.50
3	POSCO	11.74	8.32	9.74
4	SK 텔레콤	15.14**	17.14**	20.02**
5	하 이 닉 스	10.07	13.22*	18.44**
6	현 대 차	10.78	17.96**	20.27**
7	KT	9.47	9.40	13.89*
8	현 대 중 공 업	8.79	9.84	4.22
9	신 세 계	9.88	13.63*	20.97**
10	삼 성 전 자 우	7.06	8.50	10.24
11	SK	5.90	4.50	3.28
12	KT&G	7.05	11.90	12.82
13	외 환 은 행	6.61	16.42**	12.13
14	삼 성 화 재	19.98**	17.17**	17.05**
15	SK 네트웍스	11.66	7.60	6.01
16	S-Oil	3.37	3.99	4.98
17	기 업 은 행	13.97*	9.23	12.56
18	현 대 모 비 스	12.61	9.57	13.59*
19	현 대 건 설	4.67	11.24	10.64
20	LG	7.53	5.63	6.05
21	KTF	17.86**	15.56**	15.44**
22	삼 성 중 공 업	10.93	8.25	4.63
23	삼 성 물 산	4.34	5.97	8.72
24	두 산 중 공 업	3.02	17.13**	7.39
25	기 아 차	14.63*	11.89	11.84
26	GS 건설	5.64	11.49	6.39
27	현 대 산 업	12.16	16.22**	12.56
28	CJ	19.87**	11.87	12.36
29	삼 성 증 권	8.67	7.67	5.34
30	대 우 증 권	5.56	10.91	4.54

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

1) 모형 선택은 Akaike 기준(AIC, Akaike Information Criterion)을 사용하였다. Akaie 기준은 $AIC=2k-2\ln(L)$ 이며, 여기서 k는 모수의 개수, L은 우도함수(likelihood function)를 의미한다. 특히, 변수가 정규분포를 따를 경우, $AIC=2k+n\ln\left(\frac{RSS}{n}\right)$으로 표현이 되며, 여기서 RSS는 잔차최소자승합이며, n은 자료의 개수를 의미한다. AIC에서 2k는 변수의 개수 증가함에 따라 RSS값이 작아지는 것을 방지하는 penalty다. 최종적으로 모형은 AIC값이 가장 작은 후보모형으로 선택한다.

<표 5>는 각 실험별 Trace 검정통계량의 기초통계를 정리하였다. 이에 따르면 공적분 검정 기간이 길어질 수록 대체적인 검정 통계량은 증가하는 것으로 나타났다. 이러한 양태는 개별 종목들의 주가 움직임들은 단기적으로는 개별 주가의 특성에 따라서 다양한 모습을 보여주지만, 장기적으로 본다면 지수의 움직임에 점차로 일치해가는 특성을 보여주는 것이다. 또한 지수의 움직임을 선도하는 것으로 알려진 삼성전자의 경우에는 의외로 검정통계량이 매우 낮은 것을 볼 수 있는데, 삼성전자의 단일 종목의 투자가 결국 시장 평균에 근접한 수익률을 가능하게 할 것이라는 일반인들의 생각은 논란의 여지가 충분하다고 할 수 있다.

〈표 5〉 Trace 검정통계량의 기초 통계

KOSPI1			KOSPI2			KOSPI3		
평 균	표준 편차	중간값	평 균	표준 편차	중간값	평 균	표준 편차	중간값
9.50	4.80	9.13	10.99	4.01	10.38	10.76	5.16	10.57

〈표 6〉 유사성 검정으로 추출된 종목과 Trace 검정통계량

연번	KOSPI1		KOSPI2		KOSPI3	
	종 목	Trace 검정통계량	종 목	Trace 검정통계량	종 목	Trace 검정통계량
1	삼성화재	19.98	현 대 차	17.96	신 세 계	20.97
2	CJ	19.87	삼성화재	17.17	현 대 차	20.27
3	기 아 차	18.07	SK텔레콤	17.14	SK텔레콤	20.02
4	KTF	17.86	두산중공업	17.13	하이닉스	18.44
5	SK텔레콤	15.14	외환은행	16.42	삼성화재	17.05
6			현대산업	16.22	KTF	15.44
7			KTF	15.56		

<표 4>의 Trace 검정통계량을 기반으로 한 유사종목의 추출 단계에서 유의수준 0.05에서의 임계값(Critical Value)은 15.34로 계산되었다. 본 실험에서는 모든 실험에 있어서 유의수준 0.05를 기준으로 하여 임계값이 15.34를 넘기는 종목들이 지수와 장기 유사성이 있다고 판단하여 추출하고, 이들 종목들의 부분 시계열 정보를 인공 신경망의 입력 속성에 추가하였다. 다음 <표 6>은 각 실험에서 지수와 유사성이 있다고 설명되는 종목들과 이들의 Trace 검정통계량의 값을 보여주고 있다.

4.2 인공 신경망을 이용한 예측 모델의 구축

본 실험에서는 인공 신경망의 학습을 위하여 개별 종목이 가지는 19개의 기본 속성에 <표 6>의 종목들이 가지는 부분 시계열 속성을 추가하였다. 예를 들어 KOSPI1의 경우에는 5개의 종목이 공적분 검정을 통하여 추출되었는데, 각각의 종목 c가 가지는 $RP_c(t\text{-}1)$, $RP_c(t\text{-}2)$ 속성값 10개가 추가되어 총 29개의 속성이 인공 신경망의 입력값이 되었다. 학습 데이터로는 공적분 검정에 사용한 데이터들과 기간이 일치하도록 KOSPI1은 2003년-2004년 데이터를, KOSPI2는 2003년-2005년 데이터를, 그리고 KOSPI3는 2003년-2006년 6월 데이터를 그 대상으로 하였다. 또한 구축된 모델의 성능을 평가하는 테스트 데이터로는 학습 데이터와 상호 배제된 데이터를 사용하였다.

추출 단계가 예측 성능에 기여하는 정도를 평가하기 위하여 각 실험에서는 관심 종목(본 실험에서는 KOSPI 지수)의 19개 속성만으로 단순 인공 신경망을 학습한 모델(M_s), 추출 종목의 개수만큼 랜덤하게 종목을 선택하여 부분 시계열 정보를 속성에 추가한 모델(M_r), 그리고 공적분 검정 단계를 통하여 추출된 종목의 부분 시계열 정보가 입력 속성에 추가된 모델(M_c)로 예측하여 각각의 성능을 비교 분석하였다. 예측은 $t+1$ 시점의 주가가 t 시점에 비하여 상승하였는지, 아니면 하락하였는지를 예측하도록 하였다. 또한 각 실험에서는 학습 예제의 크기가 성능에 어떠한 영향을 미치는지를 파악하기 위하여 학습 예제의 크기에 따른 성능도 확인하였다.

<표 7>은 KOSPI1 실험에서의 예측 성능을 보여준다. 예측 정확도(Acc)은 N개의 테스트 예제 중에서 예제 $\mathbf{x}_i$가 가지는 목표값 y_i와 $\mathbf{x}_i$를 모델의 입력으로 주어졌을 때 출력되는 예측 결과값 $\hat{f}(x_i)$에 대하여 다음과 같이 계산되는데, 여기서 y_i 및 $\hat{f}(x_i)$는 상승일 경우에는 1, 하락일 경우에는 -1의 값을 가진다.

$$Acc = \frac{\sum_{i=0}^{N-1}(2-|y_i-\hat{f}(x_i)|)}{2N} \times 100 \quad (12)$$

또한 향상도(Imp_c)는 각 모델 M_a의 예측정확도 Acc_a를 기준으로 하였을 때, 공적분 검정을 반영한 모델 M_c의 예측정확도 Acc_c가 향상된 정도를 보여주며 다음과 같이 계산된다.

$$Imp_\alpha = \frac{Acc_c - Acc_\alpha}{Acc_\alpha} \times 100 \quad (13)$$

<표 7>에서의 첫 번째 행은 학습에 사용된 학습 예제의 크기를 말하는데, 여기서 예제의 크기가 100이라 함은 y_i 의 값이 1인 양성 예제 및 -1인 음성 예제의 수가 각각 100개씩 학습에 사용되었음을 의미한다. 본 실험에서는 학습 예제의 목표값에 의하여 학습이 영향을 받지 않도록 각각의 목표값에 대하여 동수의 예제를 선택하여 학습하였다. 또한 학습은 각 예제의 발생 날짜에 따라서 순차적으로 이루어졌다.

〈표 7〉 KOSPI1 실험에서의 예측 정확도 및 성능 향상도

	100		150		200		평 균	
	정확도	향상도	정확도	향상도	정확도	향상도	정확도	향상도
M_s	50.56	7.14	50.83	8.2	50.56	-3.3	50.65	4.68
M_r	51.94	4.29	48.61	13.15	47.78	2.32	49.44	7.24
M_c	54.17	0	55	0	49.89	0	53.02	0

〈표 8〉 KOSPI2 실험에서의 예측 정확도 및 성능 향상도

	100		150		200	
	정확도	향상도	정확도	향상도	정확도	향상도
M_s	46.67	23.81	47.78	11.62	50.56	0
M_r	49.44	16.87	46.11	15.66	48.33	4.61
M_c	57.78	0	53.33	0	50.56	0
	250		300		평 균	
	정확도	향상도	정확도	향상도	정확도	향상도
M_s	48.89	12.5	47.22	12.94	48.22	11.99
M_r	46.67	17.85	48.33	10.35	47.78	13.02
M_c	55	0	53.33	0	54	0

〈표 9〉 KOSPI3 실험에서의 예측 정확도 및 성능 향상도

	100		150		200		250	
	정확도	향상도	정확도	향상도	정확도	향상도	정확도	향상도
M_s	44.29	25.78	50	5.72	42.86	29.98	50	-5.72
M_r	44.29	25.78	48.57	8.83	50	11.42	50	-5.72
M_c	55.71	0	52.86	0	55.71	0	47.14	0
	300		350		평 균			
	정확도	향상도	정확도	향상도	정확도	향상도		
M_s	48.57	17.64	48.57	11.78	47.38	13.57		
M_r	50	14.28	50	8.58	48.81	10.24		
M_c	57.14	0	54.29	0	53.81	0		

<표 7>을 보면 모델 M_c가 학습 예제의 수와 상관없이 가장 좋은 예측 성능을 보여준다는 것을 알 수 있다. 평균적으로 M_s는 50.65%, M_r은 49.44%, 그리고 M_c는 53.02%의 예측 정확도를 보여주었으며, 평균 향상도는 M_s에 대해서는 4.68%, M_r에 대해서는 7.24%를 보여주었다. 또한 대부분의 경우에 있어서 모델 M_c는 다른 모델들에 비하여 최저 -3.3%에서 최고 13.15%의 성능 향상을 보여주었는데, 이 정도의 성능 향상이라면 공적분이 기존의 인공 신경망 학습 모델의 성능에 매우 긍정적인 영향을 미친다고 볼 수 있다. 다음 <표 8>과 <표 9>는 KOSPI2 및 KOSPI3 실험 결과를 순차적으로 보여준다.

각 실험에서 각 모델별 평균 정확도를 보면, KOSPI2의 경우에는 M_s는 48.22%, M_r은 47.78%, M_c는 54%의 예측 정확도를, KOSPI3의 경우에는 M_s는 47.38%, M_r은 48.81%, M_c는 53.81%의 예측 정확도를 보여주었다. 또한 KOSPI2에서는 M_s에 비해서는 11.99%, M_r에 비해서는 13.02%의 향상도를, 그리고 KOSPI3에서는 M_s에 비해서는 13.57%, M_r에 비해서는 10.24%의 향상도를 보여주었다. 또한 모든 경우에 있어서 M_c는 M_s에 비해서 최대 29.98%, 그리고 M_r에 비하여 25.78%의 향상도를 보여주었다. 또한 M_c는 최대 57.78%의 예측 정확도를 보여줌으로써 등락에 대한 예측이 어느 정도 가능함을 보여주었다.

이상의 세 가지 실험 결과를 토대로 다음과 같은 추론이 가능하다. 첫째, 인공 신경망의 학습에 사용된 예제의 크기와 예측 성능과는 별다른 상관관계가 보이지 않는다. 이는 모든 모델의 실험 결과에서 학습 예제의 크기가 커진다고 예측 정확도가 향상된다는 점을 발견할 수 없다는 점에서 알 수 있다. 따라서 단순히 학습 예제의 크기를 크게 함으로써 학습 모델의 예측 성능을 향상시킬 수 있다는 것은 주식 데이터와 같이 불안정성(non-stationary) 시계열 데이터의 경우에는 성립하기 어렵다고 볼 수 있다.

둘째, 공적분 검정에 사용된 시계열 데이터의 크기는 예측 성능과 별다른 상관관계가 없다. KOSPI1, KOSPI2, KOSPI3에서의 M_c의 평균 정확도 보면 거의 차이가 보이지 않는다. 또한 학습 예제의 크기가 동일한 환경에서의 평균 정확도 역시 별다른 상관관계를 보이지 않는다. 분명히 공적분 검정을 통하여 성능 향상이 이루어졌음에도 불구하고 이러한 현상이 벌어진 것은 일정 크기 이상의 시계열 데이터에서 공적분 검정이 이루어졌다면 그 이상의 크기는 큰 의미가 없음을 의미한다고 볼 수 있다. 이는 주가 데이터의 불안정성에 기인하는 것으로 예상된다.

셋째, 공적분 검정을 통하여 유사성이 인정되는 시계열이 아닌 여타의 시계열 정보는 시계열 데이터의 예측에 도움이 되지 못한다. 이는 공적분 검정을 통과한 종목의 수와 동일한 수의 종목을 임의로 선택하여 임의 종목의 시계열 정보를 학습의 입력 속성에 반영한 모델 M_r의 예측 정확도가 지수의 속성만으로 이루어진 모델 M_s과 별다른 성능 차이를 보이지 못하는 점에서 알 수 있다. 반면에 공적분 검정을 통하여 추출된 종목들의 시계열 정보를 입력 속성에 반영한 모델 M_c의 경우 예측 정확도가 매우 향상되었다. 따라서 주가 데이터와 같이 불안정성이 심각한 시계열 데이터의 학습 모델을 구축하기 위해서는 학습에 필요한 입력 속성을 설계할 때 데이터의 불안정성을 보완할 수 있는 속성을 발굴하여 이를 활용해야 한다.

4.3 타 연구와의 비교

기존에 진행되어 온 인공 신경망 기반의 주식 예측 모델의 성능 향상과 관련한 연구는 크게 세 가지로 나누어 볼 수 있다. 첫째, [8, 15]의 연구와 같이 신경망 모델에 통계 모델을 접목시키거나 신경망 구조의 최적화를 통하여 신경망 자체를 최적화하고자 한 방향이다. 본 연구에서 제안된 방법은 이들 연구와 접목이 가능하며 자체적으로 성능 향상된 신경망을 2단계에서 사용된 단순 신경망 대신으로 활용할 수 있다. 이러한 방식으로 공적분 검정을 통하여 보다 시계열성이 강화된 입력 속성을 설계 및 적용하여 기존 연구보다 더 향상된 신경망 모델을 구축할 수 있다. 둘째, [5]의 연구와 같이 인공 신경망의 입력 속성을 원시 데이터가 아닌 기술 지표 (technical indicators)와 같은 가공된 데이터로 구성하고 이와 같이 입력 속성들을 최적화함으로써 전체 예측력을 향상시키는 경우이다. 이 경우도 첫째 경우와 마찬가지로 기술 지표 시계열 데이터를 기반으로 공적분 검정을 하여 유사 시계열을 추출한 후, 인공 신경망의 입력 속성을 [5]에서 사용한 기술 지표로 설계하면 역시 인공 신경망만으로 이루어진 예측 모델보다 더욱 향상된 모델을 수립할 수 있다. 마지막으로 [7]의 연구는 본 연구와 같이 인공 신경망을 실행하기에 앞서 일종의 전처리 단계를 두어 입력을 제어하는 방식이다. 그러나 이 방식은 근본적으로 본 연구에서 제안한 방식과 성능 비교를 하기 매우 어려운데, 그 이유는 [7]에서는 원하는 패턴을 가지고 있는 모든 종목의 데이터를 한꺼번에 학습하고 원하는 시점에서 해당 패턴을 가진 모든 종목을 일단 추출한 후 이들을 대상으로 예측을 하기 때문이다. 이러한 방식의 예측에서는 원하는 시계열 패턴을 보이는 종목들에 대한 예측만 가능하기 때문에, 궁극적으로 주가 지수나 관심 종목들의 연속적인 예측은 불가능하게 된다.

5. 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 종목들의 시계열 특성을 학습 모델의 입력 속성으로 구축하고 이 데이터로 인공 신경망을 학습하여 예측 모델을 수립한다. 이러한 2단계 모델을 통하여 인공 신경망의 예측 성능을 최고 30%까지 향상시키고, 유사 종목군을 구성하여 종목의 심층 분석을 가능하게 하였다.

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. 이들 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 모델은 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 연구에서는 인공 신경망에 기반한 예측 모델을 구축하였으나, 기계 학습 분야에서 널리 사용되고 있는 많은 학습 모델들 중에서 분류(classification) 문제에 적합한 것으로 입증된 모델들을 이 문제에 적용하여 성능 향상 여부를 판단하고자 한다.

참 고 문 헌

[1] 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.

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

[3] Fama, E. F., "Multiperiod Consumption Investment Decisions," *American Economic Review*, Vol.60, pp.163-174, 1988.

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

[5] 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*, Vol.12, No.4, pp.744-754, 2001.

[6] Fan, A. and Palaniswami, M., "Stock Selection Using Support Vector Machines," *In Proceedings of International Joint Conference on Neural Networks*, pp.1973-1983, 2001.

[7] 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*, pp.143-152, 2002.

[8] 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*, Vol.9, No,6, pp.1456, 1998.

[9] 김유섭, 이재원, 이종우, "다중 에이전트 Q-학습 구조에 기반한 주식 매매 시스템의 최적화, *정보처리학회논문지 B*, 제11-B권 제2호, pp.207-212, 2004.

[10] Armano, G., Marchesi, M., and Murru, A., "A Hybrid Genetic-Neural Architecture for Stock Indexes Forecasting," *Information Sciences*, Vol.170, Issue 1,

pp.3-33, 2005.
[11] 이상원, “주가 예측을 위한 최적 인공신경망 모형 선택에 관한 연구,” 인제대학교 석사학위논문, 2002.
[12] Johansen, S., “Statistical Analysis of Cointegration Vectors,” *Journal of Economic Dynamics and Control* Vol.12, pp.231-254, 1988.
[13] Granger, C.W.J. and Newbold, P., “Spurious regressions in econometrics,” *Journal of Econometrics*, 2, pp.111-120, 1974.
[14] Phillips, P.C.B., “Understanding spurious regressions in Econometrics,” *Journal of Econometrics*, 33, pp.311-340, 1986.
[15] Engle, R.F. and Granger, C.W.J., “Co-integration and error-correction: representation, estimation and testing,” *Econometrica*, 55, pp.251-276, 1987.
[16] Koscom, http://www.koscom.co.kr, 2007.
[17] 한국투자증권, http://www.truefruend.com, 2007.
[18] Mitchell, T., *Machine Learning*, McGraw Hill, 1997.

오 유 진

e-mail : ouj92@hotmail.com
1996년 이화여대 수학과 졸업(학사)
1998년 이화여대 대학원 통계학과 (석사)
1985년 이화여대 대학원 통계학과(박사)
현 재 고려대학교 경영전문대학원 연구교수
관심분야 : 시계열분석, 계량경제 등

김 유 섭

e-mail : yskim01@hallym.ac.kr
1992년 서강대학교 전자계산학과 (학사)
1994년 서울대학교 대학원 컴퓨터공학과 (공학석사)
2000년 서울대학교 대학원 컴퓨터공학과 (공학박사)
현 재 한림대학교 정보통신공학부 부교수
관심분야 : 전산금융, 자연언어처리, 기계학습, e-learning 등