주가 운동양태 예측을 위한 예측 모델결정에 관한 연구

A Study on Determining the Prediction Models for Predicting Stock Price Movement

전진호, 조영희, 이계성
단국대학교 전자계산학과

Jin-Ho Jeon(jhjy@dankook.ac.kr), Young-Hee Cho(zerowh@daum.net),
Gye-Sung Lee(gslee@dku.edu)

요약

주식투자의 대중화, 관심의 증가에 따라 주가예측의 중요성이 증대되고 있다. 주가의 변화는 어떤 경향이나 패턴에 의해 움직인다고 가정할 때, 과거의 주가분석을 통해 이들의 변화를 잘 설명할 수 있는 모델의 구성이 가능할 것이다. 동적인 현상을 반영하는 최적의 모델이 구성된다면 이를 통해 향후의 일정기간의 주가의 운동양태의 예측이 가능할 것이다. 본 연구에서는 주가와 같은 템포랄(temporal) 데이터를 잘 설명할 수 있는 모델결정에 대한 방법론으로서 오토마타 기반의 모델을 가정한다. 모델의 최적 상태 수를 결정하기 위한 기준으로서 베이지안정보기준(BIC : Bayesian Information Criterion) 근사법을 사용한다. 베이지안정보기준의 유효성을 살펴보고 베이지안정보기준을 실제 주가데이터 모델의 상태 수 결정과정에 적용하여 모델을 생성한 후 결정된 모델을 통하여 일정 기간의 일별주가곡선의 운동양태를 예측한다. 실제의 주가곡선에 적용하여 모델의 유효성을 확인하였고 예측 주가곡선의 운동양태가 실제 주가 곡선과 유사함을 확인하였다.

■ 중심어 : | 템포랄 데이터 | 모델기반 | 베이지안정보기준 | 주가지수 |

Abstract

Predictions on stock prices have been a hot issue in stock market as people get more interested in stock investments. Assuming that the stock price is moving by a trend in a specific pattern, we believe that a model can be derived from past data to describe the change of the price. The best model can help predict the future stock price. In this paper, our model derivation is based on automata over temporal data to which the model is explicable. We use Bayesian Information Criterion(BIC) to determine the best number of states of the model. We confirm the validity of Bayesian Information Criterion and apply it to building models over stock price indices. The model derived for predicting daily stock price are compared with real price. The comparisons show the predictions have been found to be successful over the data sets we chose.

■ Keyword : | Temporal Data | Model-based | Bayesian Information Criterion | Stock Price Index |

* 본 연구는 2005학년도 단국대학교 대학연구비의 지원으로 연구되었습니다.

접수번호 : #060406-001
접수일자 : 2006년 04월 06일
심사완료일 : 2006년 05월 16일
교신저자 : 전진호, e-mail : jhjy@dankook.ac.kr

I. 서 론

현대는 주식투자가 대중화되면서 그에 대한 관심도 증가되고 있다. 최근의 주식시장의 불안정한 상태와 급속한 금융환경변화에 따른 변동성 증가에 따른 주가예측의 중요성이 증가되고 있다. 주가의 형성과정은 단순하지 않다. 이러한 형성과정을 통한 주가의 예측가능성에 대하여 상반된 두 입장이 존재하고 있다. 미래 주가에 영향을 미치게 될 요소가 매우 많고 다양하기 때문에 주가의 미래를 예측한다는 것은 매우 어렵거나 불가능하다는 입장이 있다. 또 다른 한편에서는 주가의 변화는 과거 주가의 분석을 통해 즉, 어떤 경향이나 패턴에 의해 미러 주가의 움직임을 어느 정도 예측할 수 있다고 주장한다. 14세기 영국의 논리학자가 주창한 Occam's Razor 원리처럼 아무리 복잡한 현상이라도 최소의 이론으로 설명할 수 있다는 논리는 후자의 관점을 지원하는 것이라 할 수 있을 것이다. 본 연구에서는 후자의 입장을 통하여 시간적 특징들로 묘사되는 주가데이터의 모델결정 방법론을 살펴보고자 한다. 이는 주어진 주가데이터에 대하여 오토마타기반의 모델을 가정하는 것인데, 데이터에 대하여 가장 적합한 모델을 생성하는 것이다. 즉 궁극적 목적은 동적시스템을 표현하는 템포랄(temporal) 데이터를 잘 설명할 수 있는 정확한 모델을 개발하는 것이다.

본 연구에서는 최적의 모델을 결정하는 과정에 대한 방법론으로 오토마타기반의 모델에서 최적의 상태 수를 결정할 베이지안정보기준(Bayesian Information Criterion) 근사법에 대해서 고찰과 실험을 통해 유효성을 살펴보고 베이지안정보기준 근사법을 실제 주가데이터의 모델 결정과정에 적용하여 모델을 생성한다. 향후 10일과 1개월의 일별주가곡선의 운동양태를 예측하여 실제의 주가곡선과 비교함으로서 모델결정의 유효성을 검증하고 보다 정확한 주가의 운동양태의 향후 변화를 예측하는데 본 연구의 목적이 있다.

II. 배경 연구

과거에서 현재까지 템포랄(temporal) 데이터의 모델기반의 예측기법들에 대하여 많은 연구들이 있었다. 모델기반 방법론은 주어진 데이터에 대하여 가장 적합한 모델의 집합을 찾는 방법론이다. 모델기반 방법은 각 데이터에 대하여 분석적인 함수 또는 오토마타 기반 모델들을 가정한다. 위의 기법들은 가장 적합한 모델들에 객체들을 반복적으로 할당하며 새로운 객체 분배와 함께 모델 파라미터를 갱신한다. 이러한 과정은 수렴 시까지 반복되어진다. 모델기반의 방법들은 회귀모델[1], 시계열분석[2], 신경망, 그리고 비결정적 유한상태 오토마타인 마코프체인(Markov Chain)[3], 은닉마코프모델(Hidden Markov Model)[4] 등이 있다.

모델기반의 방법들의 각 특징을 살펴보면, 회귀모델은 길지 않은 데이터를 다루므로 동적현상의 특성묘사를 나타내기 어렵다. 신경망은 많은 부분에서 템포랄(temporal) 현상을 예측하는 작업에 성공적으로 적용되어 왔으나, 일반적 템포랄(temporal) 데이터의 모델링에는 적합하지 않다. 그 이유는 첫째, 모델의 구조가 알려져 있다는 것이다. 즉, 모델에서 은닉층 수, 노드들에서 사용되는 기준함수뿐만 아니라 각 층에서 노드들의 수가 정해져 있다는 것이다. 둘째, 모델의 해석을 지원하지 않는 것이다. 이는 훈련과정동안, 모델 파라미터 값들의 조정의 목적은 객관적 기준함수에 따라 산출층에 값들을 최적화하는 것이다. 그러므로 신경망에서 노드들 사이의 연결들과 노드들과 관련된 모델적 의미가 없다는 것이다.

마코프체인 모델은 모델의 단순성 때문에 하나의 이산값을 갖는 템포랄(temporal) 특징으로 묘사되는 템포랄(temporal) 데이터의 표현 모델링에 유용하다[5]. 그러므로 일반적인 템포랄(temporal) 데이터의 모델링에 사용될 때 다음과 같은 제한점이 있다 첫째, 연속적인 값을 갖는 템포랄(temporal) 데이터 특징을 묘사하는 데이터 모델에 적합하지 않으며 둘째, 다수의 템포랄(temporal) 특징에 의하여 묘사되는 데이터를 표현에 어렵다. 이러한 문제를 해결하기 위하여 각 상태에서 특징들에 대한 적합한 확률함수를 사용하여 연속적인 값을 갖는 템포랄(temporal) 시퀀스를 쉽게 다루며, 다수의 템포랄(temporal) 특징들을 가진 데이터의 묘사가 쉬운 은닉마코프 모델을 사용하는 것이 일반적 템포랄(temporal) 데이터의 모델링에서는 효과적이라고 할 수 있다.

III. Temporal 데이터의 모델결정방법론

본 연구에서의 모델결정 방법론으로서 템포랄(temporal) 데이터의 모델결정과정은 크게 두 단계로 나누어 볼 수 있다. 첫 번째 단계는 최적의 상태의 수를 결정하는 것과 두 번째 단계는 최대 우도값을 주는 모델 파라미터의 추정이다.

3.1. 모델의 상태 수 결정을 위한 기준(BIC)

은닉마코프모델의 모델 사이즈 선택의 목적은 주어진 데이터로부터 최적의 상태들의 수를 가진 은닉마코프모델을 선택하는 것이다. 은닉마코프모델 λ에 대하여 시간적 데이터 X가 주어지면, 베이지안 모델 선택구조에서, 한계우도의 파라미터 구성과 결부되어질 때, 최고의 사후확률 모델을 주는, 최선의 모델 사이즈는 하나이다. 우리는 베이즈 이론으로부터 다음을 알 수 있으며

$$P(M\ X) \propto P(X|M) \tag{1}$$

즉, 식 (1)처럼 모델의 사후확률은 데이터의 한계우도에 비례한다. 그러므로 베이지안 모델 선택의 목적은 가장 큰 한계우도를 주는 모델을 선택하는 것이다. 모델 M의 파라미터 구성 θ가 주어지면, 데이터의 한계우도는 식 (2)와 같이 계산되어진다.

$$P(X|M) = \int_{\Theta} P(X|\Theta, M)P(\Theta|M)d\Theta \tag{2}$$

파라미터들이 연속적인 값들을 가질 때 적분계산은 폐형해(closed form solution)를 획득하는 것이 어렵다. 한계우도를 구하기 위한 근사기법은 다양하다. 몬테-카를로 방법 그리고 라플라스 근사[6] 등이 있는데 이 둘은 매우 정확하기는 하나 계산적으로 비용이 많이 드는 것으로 알려져 있다. 따라서 본 연구에서는 더 효율적인 근사기법인 베이지안정보기준을 살펴볼 것이다. 베이지안정보기준은 다량의 데이터가 있을 때 우도함수나 사전확률이 다변량 가우시안 분포에 근사된다는 점에서 유도된다[6]. 식 (2)의 내부의 항에 로그를 취한 것을 $g(\theta)$로 정의하고 $g(\theta)$를 최대화 시키는 파라미터 구성을 $\hat{\theta}$라 할 때 이는 사후확률을 최대화하게 된다. 이를 θ의 최대 사후확률(MAP)이라 부른다. 여기에 2차 테일러 다항 근사법을 적용한 후 e의 지수를 취하고 다시 원식에 대입하여 다음 식을 산출한다.

$$\log P(X|M) \approx \log P(X|\hat{\Theta}, M) + \log P(\hat{\Theta}|M) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A| \tag{3}$$

여기서 d는 모델에서 파라미터의 수이고, A는 $\hat{\Theta}$에서 계산되는 $g(\Theta)$ 음의 Hessian이다. 식 (3)는 다시 데이터의 수인 N에 비례하는 항만 남기고 나머지를 제거함으로 더욱 근사 시켜 식 (4)를 유도된다.

$$\log P(X|M) \approx \log P(X|M, \hat{\Theta}) - \frac{d}{2}\log N \tag{4}$$

위의 식에서 첫 번째 항인 자료의 우도 $\log P(X|M, \hat{\Theta})$는 데이터를 가장 잘 설명할 수 있는 상세한 데이터의 모델을 찾도록 유도하는 성분이다. 두 번째 항인 $-\frac{d}{2}logN$은 모델 내의 파라미터 개수에 대한 패널티 항으로 볼 수 있다. 베이지안정보기준[7]은 이러한 두 항에 상호 배타적인 특성이 서로 조화되는 타협점에서 최선의 모델 상태의 수가 결정되는 것을 제시한다. 베이지안정보기준 곡선의 특징은 상태의 수가 증가할수록 우도의 값이 점진적으로 증가하지만 상태의 수가 증가함에 따른 패널티 항에 의해 상쇄되므로 최적의 상태 수를 나타내는 점에서 정점을 이루고 점차 하강하는 곡선으로 표현된다. 베이지안정보기준 근사기법의 장점은 수식의 의미가 직관적이면서 동시에 사전확률을 구할 필요가 없다는 점이다.

3.2. 모델 파라미터 추정단계

모델의 상태의 수가 결정되어 모델의 구조가 주어지면, 은닉마코프모델의 파라미터의 추정은 모델 파라미터들 $\bar{\pi}$, $A(a_{ij})$ 그리고 $B(\bar{\mu}, \Sigma)$의 최적화를 유도한다.

파라미터의 추정방법으로서 E-M 알고리즘의 한 형태인 관측열에 대하여 최대확률을 주는 최대우도(ML)기법인 바움-웰치(Baum-Welch) 파라미터 추정기법을 사용한다[8]. 최대우도(ML) 은닉마코프모델 파라미터 추정절차는 전향-후향 동적프로그래밍 절차에 의해 구현되어진다.

모델 파라미터의 재추정 단계에서 필요한 변수인 전향변수 α와 후향변수 β는 전향, 후향 절차를 통해 계산되어지며 다음의 식 (5), 식 (6)과 같다.

$$\alpha_t(i) = \left(\sum_{j=1}^{N} \alpha_{t-1}(j) a_{ij} \right) . P_i(O_t) \quad (5)$$

$$\beta_t(i) = \sum_{j=1}^{M} a_{ij} . P_i(O_t) . \beta_{t+1}(j) \quad (6)$$

위의 식을 이용하여 시간이 t에서 $t+1$로 흐를 때, 상태 i에서 상태 j로 전이한 횟수의 기대값과 관측열이 시간 t에서 i 상태에 방문될 횟수의 기대값은 다음의 식 (7), 식 (8)과 같다.

$$E(A_{ij}) = \sum_{t=1}^{L} \alpha_t(i) a_{ij} P_j(O_{t+1}) \beta_{t+1}(j) \quad (7)$$

$$E(A_i) = \sum_{t=1}^{L} \sum_{j=1}^{M} \alpha_t(i) a_{ij} P_j(O_{t+1}) \beta_{t+1}(j) \quad (8)$$

위의 계산된 변수 값들을 이용하여 모델의 매개변수를 재추정(reestimate)하는데 모델 파라미터들의 갱신규칙은 다음의 식 (9), 식 (10), 식 (11)과 같다.

$$a'_{ij} = \frac{\sum_{t=1}^{L} \alpha_t(i) a_{ij} P_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{L} \sum_{j=1}^{M} \alpha_t(i) a_{ij} P_j(O_{t+1}) \beta_{t+1}(j)} \quad (9)$$

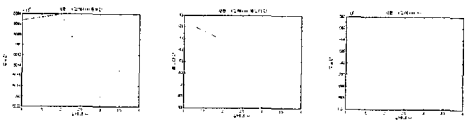
$$\mu'_{ik} = \frac{\sum_{t=1}^{L} \alpha_t(i) . \beta_t(i) . O_t^k}{\sum_{t=1}^{L} \alpha_t(i) \beta_t(i)} \quad (10)$$

$$\sigma'_{ik} = \sqrt{\frac{\sum_{t=1}^{L} \alpha_t(i) . \beta_t(i) . (O_t^k - \mu'_{ik})}{\sum_{t=1}^{L} \alpha_t(i) \beta_t(i)}} \quad (11)$$

IV. 실험 및 결과

실험을 통하여, 모델의 결정시에 제일 먼저 파악되어야 할 상태 수를 결정짓는 판단기준으로 사용될 베이지안정보기준의 효용성을 살펴보고자 데이터의 시퀀스의 수와 시퀀스의 길이에 따라 정확한 상태의 수를 찾는지를 살펴본다. 그리고 실질적인 주가데이터에 적용하여 상태 수를 결정하고 이에 따른 모델을 생성하여 주가곡선을 예측하는 실험을 통해 모델이 최적으로 생성되는지를 확인한다.

먼저 베이지안정보기준의 효용성을 실험하기 위해 데이터를 생성해야 한다. 모델의 상태는 둘이며 각 상태의 전이확률과 방출확률의 파라미터는 다음과 같은 조건으로 정의하였다.[1] 첫 번째 실험은 실험집단을 두 집합으로서 각 시퀀스의 길이는 3000으로 같고 시퀀스의 수를 3과 6으로 하였을 경우를 살펴보았으며 두 번째 실험은 실험집단을 세 집합으로서 각 집합의 시퀀스의 수를 6개로 하며 각 시퀀스의 길이를 2000, 3000, 4000으로 하였을 경우를 살펴보았다.

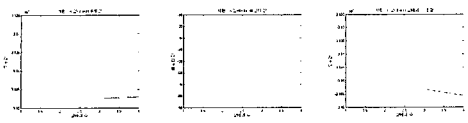


(a) likelihood (b) penalty (c) BIC

그림 1. 각 시퀀스 길이 3000, 시퀀스의 수가 3개인 데이터 집합의 liklihood, penalty, BIC 곡선

첫 번째 실험에서, 주어진 두 데이터의 집합에 대한 우도, 패널티, 베이지안정보기준 곡선은 [그림 1][그림 2]와 같다. [그림 1(c)]에서 보는 것처럼 시퀀스의 수가 3개인 실험군은 모델의 상태 수를 하나의 상태에서 가장 큰 베이지안정보기준 값을 갖는 것으로 나타났으며, 시퀀스의 수가 6개인 실험군에서는 [그림 2(c)]에서처럼 상태의 수를 2개인 것에서 가장 큰 베이지안정보기준값을 갖는 것으로 정확히 추정하는 것을 볼 수 있다.

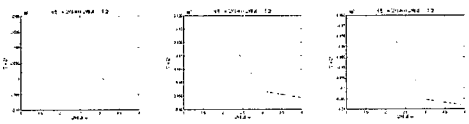
1) 상태초기=[0.7 ; 0.3]
전이확률=[0.95 0.05 ; 0.1 0.9]
방출확률=[0.16667 0.16667 0.16667 0.16667 0.16667 0.16667 ; 0.1 0.1 0.1 0.1 0.1 0.5]



(a) liklihood (b) penalty (c) BIC

그림 2. 각 시퀀스 길이 3000, 시퀀스의 수가 6개인 데이터 집합의 liklihood, penalty, BIC 곡선

[그림 1]에서 상태가 1인 경우는 큰 의미를 갖지 못한 것이기 때문에 그 다음 크기를 갖는 2에서 최대값을 갖는 것으로 해석할 수 있다. 데이터의 시퀀스 수가 충분히 많은 경우에서는 바로 정확한 상태의 수를 추정한다는 것을 알 수 있다.

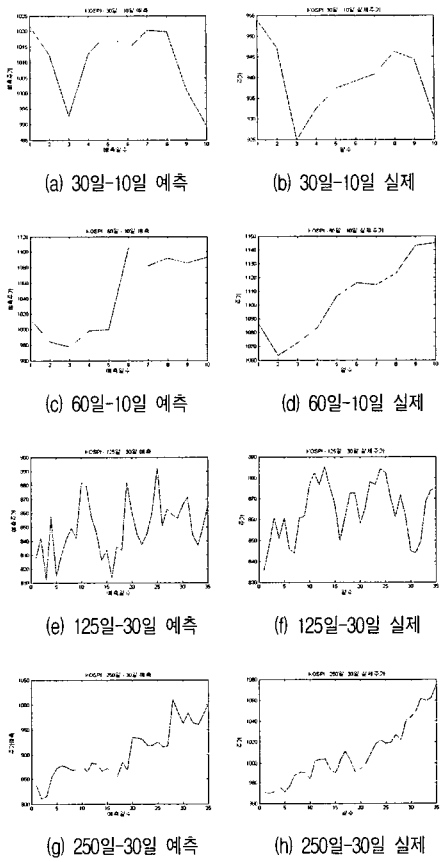


(a) 길이-2000 (b) 길이-3000 (c) 길이-4000

그림 3. 시퀀스의 수는 6, 각 시퀀스의 길이가 2000, 3000, 4000 일 때의 BIC 곡선

두 번째 실험에서는, 주어진 세 데이터의 집합에 대한 베이지안정보기준 곡선은 [그림 3]과 같다. [그림 3]에서 보는 것과 같이 모두 두 개의 상태를 정확히 추정하는 것을 볼 수 있다. 즉, 데이터의 시퀀스의 길이도 상대적으로 충분하다면 정확한 상태의 수를 보여준다. 위에서 베이지안정보기준 근사법의 유효성이 증명됨에 따라 실질적인 주가데이터에 적용하여 주가데이터를 잘 설명하는 모델을 생성해보고 이를 통해 예측되는 주가곡선의 패턴을 실질의 주가곡선과 비교하여 확인해 봄으로서 모델이 최선으로 생성된 것인지를 확인해 보고자 한다.

실험에 사용된 주가데이터는 크게 세 분류로 나누어 보았다. KOSPI(옛 종합주가지수)지수, 산업별지수에서 전기전자업종 그리고 개별종목으로 나누었으며 개별종목에서는 삼성전자와 하이닉스의 데이터를 선정하였다.



(a) 30일-10일 예측 (b) 30일-10일 실제

(c) 60일-10일 예측 (d) 60일-10일 실제

(e) 125일-30일 예측 (f) 125일-30일 실제

(g) 250일-30일 예측 (h) 250일-30일 실제

그림 4. KOSPI 지수 - 30, 60, 125, 250일의 실험데이터를 통한 10, 30일의 주가곡선의 비교[2]

개별종목으로서 삼성전자는 대표적인 우량종목으로서 주가의 형성에 있어 외적환경요소에 덜 민감할 것으로 생각되어 좀 더 안정적인 예측모델이 가능할 것으로 생각되어 선택하였으며 하이닉스는 중형주로서 예측모델의 예측률이 우량종목의 예측률과 차이가 있는지를 확인함으로서 모델결정의 안정성을 확인하기 위하여 선정하였다. 모든 실험데이터의 표본기간은 2004년 3월 2일부

2) 그림 4의 각 그래프의 데이터 길이는 다음과 같다
(a)실험데이터:05.03.02-05.04.13(b)실제데이터:05.04.14-05.04.27
(c)실험데이터:05.04.01-05.06.28(d)실제데이터:05.06.29-05.07.12
(e)실험데이터:04.05.03-04.10.29(f)실제데이터:04.11.01-04.12.17
(g)실험데이터:04.05.03-05.05.02(h)실제데이터:05.05.03-05.06.22

터 2005년 12월 29일까지 시퀀스의 길이는 460이다.

위의 [그림 4]는 KOSPI 지수의 30일, 60일의 실험데이터를 통해 10일(단기)을 그리고 125일, 250일의 실험데이터를 통해 30일(장기)의 예측한 곡선과 실제곡선을 보여준다.[3] (a)와 (b)는 3일경까지 하락 후 상승을 하다 8일경부터 하락을 하는 매우 유사한 양태의 곡선을 보여주고 있으며 (c)와 (d)는 전체적으로 상승하는 모양을 보여주고 있으나 (c)에서 초, 후반부에 보합을 보여주는 것이 약간 차이가 있다. 즉 10일(단기)의 예측에서는 실험데이터가 60일보다는 30일의 데이터가 매우 유사함을 보여주는 것으로 확인할 수 있다. 그리고 (e)와 (f)는 5일, 10일, 30일 부근에서 최저 하락점을 보여주는 매우 유사한 운동양태를 보여주며 (g)와 (h)는 20일경 후는 둘 다 상승하지만 (g)는 초중반에 보합을 (h)는 초중반에 상승을 보여준다. 30일(장기)에서도 마찬가지로 기간이 긴 시퀀스보다는 짧은 시퀀스(125일)가 좀 더 유사한 운동양태를 보여준다. 긴 기간의 훈련데이터보다 짧은 기간의 훈련데이터에서 예측된 곡선들이 실제 곡선의 운동양태와 좀 더 유사한 곡선을 보여주는 것은 긴 기간의 데이터일수록 과거의 수많은 불규칙적인 외적변수가 과거 데이터에 많이 내재되어 있기 때문이다.

위의 결과를 통해 예측곡선과 실제곡선의 운동양태 유사성을 확인하는 과정으로 본 연구에서는 두 곡선을 정규화[4]시킨 후, 두 곡선의 각 일별변화의 차이를 통해 두 곡선의 일별변화의 차이가 적을수록 에러율(예측곡선과 실제곡선의 차이)이 적은 것으로서 유사한 곡선임을 확인하는 방법으로 적용하였다. 정규화 시킨 에러율 평균치는 0.3에서 0.9사이에 존재함을 실험을 통해 확인하였다. 중앙값에 해당하는 평균치 임계값을 0.6으로 정하였을 때 에러율값이 임계값보다 낮을수록(0.3에 근접) 두 곡선의 운동양태가 유사함을 보여주었으며 높을수록 (0.9에 근접) 비유사성을 보여주었다. 위의 임계값을 기준으로 낮은 수치 사례에서는 두 곡선 즉, 실제곡선과 예측곡선의 일별 상승, 하락의 패턴이 불일치의 날짜가 10일 중 평균 2일로서 80%의 예측정확도를 확인할 수 있었다.

표 1. 실험데이터의 기간별 시행횟수와 유사패턴 수

	30일	60일	125일	250일
KOSPI	20회 (16회)	20회 (8회)	20회 (20회)	20회 (12회)
전기전자 (업종별)	20회 (16회)	20회 (8회)	20회 (16회)	20회 (8회)
삼성전자 (개별종목)	20회 (16회)	20회 (8회)	20회 (16회)	20회 (8회)
하이닉스 (개별종목)	20회 (16회)	20회 (7회)	20회 (16회)	20회 (9회)

[표 1]은 KOSPI, 산업별지수에서 전기전자업종, 개별종목에서 삼성전자와 하이닉스의 주가데이터를 대상으로 시행한 기간별 실험 횟수와 유사패턴이 발생된 횟수를 보여주는 표이다. 모두 60일, 250일보다는 30일과 125일의 실험데이터에서 유사패턴이 잘 예측되는 것으로 확인할 수 있었으며 개별종목에서 모델결정에 따른 유사패턴의 예측정확도가 대형우량주(삼성전자)와 중소형주(하이닉스)에 있어서 별 차이가 없음을 보여준다.

V. 결 론

본 연구에서는 각 템포랄(temporal) 데이터가 주어진 후 이 데이터를 잘 설명할 수 있는 모델을 결정짓는 과정으로서 오토마타 기반의 모델결정 방법을 적용하였다. 각 템포랄(temporal) 데이터에 대한 모델을 결정하는 첫 단계는 모델의 상태 수를 찾는 것이다. 베이지안정보기준 근사법을 이용하여 상태수를 구했다. 베이지안정보기준 근사법이 일반적으로 모델의 상태 수를 정확하게 추정하는 실험 결과를 보여주고 있으나 데이터 시퀀스의 수와 시퀀스의 길이에 영향을 받는 것을 확인하였다. 이 점은 베이지안정보기준 근사법의 유도가 자료의 개수가 많은 경우에 다변량 가우시안 분포로 근사할 수 있다는 점에서 볼 때 당연한 결과로 예측된 것이다. 이렇게 베이지안정보기준 근사법을 이용하여 모델의 상태 수를 결정하여 모델이 결정되는 과정을 실질적인 주가데이터 즉,

3) 지면상, 본 연구에서 실험결과들 중 KOSPI 지수의 실험결과 그래프만 그림4에서 제시하였다.

4) 정규화변환을 통해 시퀀스가 갖는 요소값의 절대크기를 무시할 있으며, 이는 요소값의 크기는 다르지만 변화하는 패턴이 유사한 시퀀스들을 파악하는데 유용하다. 정규화는 다음의 식에 따른다.

$$s'_i = \frac{s[i] - \frac{Max(S)+Min(S)}{2}}{\frac{Max(S)-Min(S)}{2}}$$

KOSPI 지수, 산업별 종목에서 전기전자업종 지수, 개별 종목에서 우량주인 삼성전자와 중소형주인 하이닉스의 주가에 적용하여 모델을 생성 후 향 후 10일과 30일의 일별 주가곡선의 운동양태를 예측하는 실험을 한 결과 단기(10일), 장기(30일) 예측 모두 실제와 유사한 운동양태를 예측하는 것을 확인할 수 있었다. 그러나 향후 예측정확도를 보다 개선시키는 노력이 필요하고 유사성을 측정하는 측도에 대한 연구가 진행되어야 한다. 다양한 업종 및 종목으로 확대 연구하고 측정 장비나 공정과정에서 발생하는 템포랄(temporal) 데이터에 이 모델링 방법을 적용하여 복잡하고 동적인 시스템들과 프로세스들을 가진 현상을 설명하는 모델 구성 연구에 노력할 것이다.

참 고 문 헌

[1] S. Gaffney and P. Smyth, "Curve Clustering with Random Effects Regression Mixtures," Proc. Ninth Inter. Workshop on Artificial Intelligence and Staristics, 2003.

[2] W. S. Wei, Time Series Analysis: Univariate and Multivariate Methods, Addison-Wesley Publishing Co. Inc., 1990.

[3] M. Kijima, Markov Processes for Stochastic Modeling. The University Press, Cambridge, 1997.

[4] J. Frederick, Statistical Methods for Speech Recognition, The MIT Press, 2001.

[5] P. Sebastiani, M. Ramoni, P. Cohen, J. Warwick, and J. Davis, "Discovering dynamic using bayesian clustering," Advances in Intelligent Data Analysis, Springer-Verlag, D. J. Hand, J. N. Kok, and M. R. Berthold, Eds. Berlin, Springer-Verlag, pp.199-210, 1999(8).

[6] D. Heckerman, D. Geiger, and D. M. Chikering, "A tutorial on learning with Bayesian Network," Machine Learning, Vol.20, pp.197-243, 1995.

[7] S. S. Chen and P. S. Gopalkrishana, "Speaker, enviroment, and channel change detection and clustering via the Bayesian information criterion," Proceedings of the IEEE Interational Conference on Vol.2, pp.645-648, 1998(5).

[8] T. Koski, Hidden Markov Models for Bioinformatiocs, Kluwer Academic Publishers, 2001.

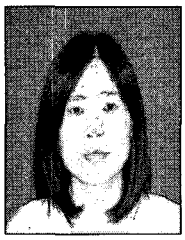
저 자 소 개

전 진 호(Jin-Ho Jeon) 정회원

- 1998년 : 명지대학교 경영정보학과(경영학석사)
- 2003년 2월 : 단국대학교 전자계산학과(박사과정 수료)
- 2003년 3월~현재 : 관동대학교 경영정보학부 겸임조교수

<관심분야> : 기계학습, 데이터마이닝

조 영 희(Young-Hee Cho) 정회원



- 2000년 : 단국대학교 전자계산학과(석사)
- 2005년 : 단국대학교 전자계산학과(박사과정)

<관심분야> : 기계학습, 데이터마이닝, Ontology

이 계 성(Gye-Sung Lee) 정회원

- 1980년 : 서강대학교 전자공학과(학사)
- 1982년 : 한국과학기술원 전자계산학과(석사)
- 1994년 : Vanderbilt University 전자계산학과(공학박사)
- 1994년~1996년 : 대구대학교 전산정보학과 전임강사
- 1996년~현재 : 단국대학교 컴퓨터과학전공 부교수

<관심분야> : 기계학습, 데이터마이닝, 바이오인포메틱스, 비디오마이닝