논문 2010-6-20

다중 유사 시계열 모델링 방법을 통한 예측정확도 개선에 관한 연구

A Study on Improving Prediction Accuracy by Modeling Multiple Similar Time Series

조영희*, 이계성*

Young-Hee Cho, Gye-Sung Lee

요 약 본 연구에서는 시계열 자료처리를 통해 예측정확도를 개선시키는 방안에 대해 연구하였다. 단일 예측 모형의 단점을 개선하기 위해 유사한 시계열 자료를 선정하여 이들로부터 모델을 유도하였다. 이 모델로부터 유효 규칙을 생성해내 향후 자료의 변화를 예측하였다. 실험을 통해 예측정확도에 있어 유의한 수준의 개선효과가 있었음을 확인하였다. 예측모델 구성을 위해 고정구간과 가변구간을 두고 모델링하여 고정구간, 창이동, 누적구간 방식으로 구분하여 예측정확도를 측정하였다. 이중 누적구간 방식이 가장 정확도가 높게 나왔다.

Abstract A method for improving prediction accuracy through processing time series data has been studied in this research. We have designed techniques to model multiple similar time series data and avoided the shortcomings of single prediction model. We predicted the future changes by effective rules derived from these models. The methods for testing prediction accuracy consists of three types: fixed interval, sliding, and cumulative method. Among the three, cumulative method produced the highest accuracy.

Key Words : 시계열(time series), 마코프모델(Markov model), 예측모델(Prediction model).

Ⅰ. 서 론

우리는 시간 속에 살아가기 때문에 실세계에서 생산되는 자료는 실질적으로 모두 시계열 자료에 해당된다고 말할 수 있다. 특히 주가나 사회 경제 지표와 같은 자료는 속성상 재화적 가치 창출과 연관되어 있어 많은 연구의 대상이 되고 왔다. 이들을 분석하는 방법론에 있어서 다양한 접근방법이 시도되고 있으나 주요 처리과정을 살펴보면 자료의 형태나 패턴을 분석하면서 자료를 이해하거나 해석하는 방법과 이를 활용하여 향후 추세를 판단하는 방법으로 구분할 수 있다. 이 과정에서 자주 사용되는 방법으로 모델링을 들 수 있다. 복잡하면서 다양한 현상을 좀 더 체계적으로 이해하기 위한 수단이다. 이를 모델 기반 해석이라 부른다. 모델을 토대로 시간적 흐름에 대해 어떻게 변화할 것인지를 파악하려는 시도가 이어지는데 이를 모델 기반 예측 방법이라 부른다. 본 연구에서는 모델에 기반을 둔 예측방법에 대하여 연구하였다.

본 논문에서는 마코프 체인 모델을 기본 모델링 방법으로 선정하여 예측에 활용한다. 모델링을 통해 유용한 패턴을 찾아 그 패턴을 이용하여 새로운 상황에 적용하여 향후 추세를 결정할 수 있는 방법을 제안한다. 모델링의 결과로 유효한 규칙을 유도해 내어 이를 통해 예측 정확도를 개선시킨다. 본 연구가 제안하는 방법은 단일 계열을 통한 모델링이 아니라 복수개의 시계열 자료로부터 유도된 모델을 통해 예측하는 방법을 통해 정확도를 개

*정회원, 단국대학교 컴퓨터과학과
접수일자 : 2010.11.18, 수정완료일자 : 2010.12.12
게재확정일자 2010.12.15

선시켰다. 기간별 예측을 통해 예측정확도를 비교 분석하였고 유사 계열을 찾아 유사 계열 전체를 통합한 모델을 통해 예측 정확도를 개선시키는 방법을 제안하였다.

2장에서는 관련연구를 기술하고, 3장에서는 마코프 체인 모델을 예측에 활용하는 방안을 제시한다. 4장에서는 이를 실험하여 산출된 결과에 대해 분석하고 그 방법에 대한 타당성을 검토한다. 마지막으로 5장에서는 결론을 기술한다.

Ⅱ. 관련연구

시계열 자료에 대한 모델링 기법에는 여러 가지가 있다. 그중 마코프 프로세스를 모델링하는 방법으로 마코프 체인모델과 은닉마코프 모델이 있다. 모델링하는 대상에 미지의 파라미터가 존재할 것이고 이들은 마코프 프로세스일 것으로 가정하여 그 가정에 기초하여 관측된 파라미터로부터 숨겨진 파라미터를 찾아 모델로 완성하는 방법이다[1]. 마코프 체인 모델은 마코프 모델의 가장 간단한 모델로 어떤 계열의 시간적 운행 모양이 그 계열의 현재 상태와 과거의 일련의 상태가 미래의 상태를 결정짓는 확률과정으로 정의된다. 상태간의 이동을 확률적 표현으로 나타내며 이는 전이확률로 표현된다.

은닉 마코프 모델은 시계열 자료 처리 및 예측에 활발히 응용되고 있는 모델이다[2,4,8,9]. 은닉 마코프 모델만으로 주가를 예측하는 경우 연구 성과가 제한적일 수밖에 없다. 그 이유는 통상 은닉 상태에 대한 내용을 파악하기 어려워 모델 구성에서 어려움을 겪는다. 통상 일별 주가는 시가, 종가, 최고가, 최저가를 가지므로 4개의 은닉상태를 설정하여 은닉 마코프 모델을 구성한다[4]. 이 4개의 은닉 상태는 상태로서의 의미가 전혀 없다. 이들은 관측값을 형성하는 것으로 은닉 상태로 취급하는데 근본적인 오류가 있는 것이다. 이런 문제를 안고 모델을 구성하여 시계열 자료를 분석한다면 그 분석결과도 당연히 정당화될 수 없을 것이다.

시계열 자료에 대한 모델링에 예전부터 사용하여 오던 자기회귀 모델과 이동 평균 모델을 비롯하여 베이지안 모델 등이 있다[5]. 자료의 시점 사이의 관계를 추정하는 방법으로 자기회귀 모델과 이동평균 모델이 사용되어 왔다. 이들은 과거의 값이나 특정 요소가 현재의 값을 결정한다는 가정에서 비롯한 모델이다. 현재의 값은 과거의 값에 의해 결정되기 때문에 이론적으로 잡음 요소와 같은 특정한 방해요소가 없다고 가정한다면 장기적인 예측도 가능하다고 본다. 그러나 이와 같은 단순 모델로 복잡하게 변화하는 경제지표까지 설명하거나 예측하는 일에는 한계가 있다고 본다.

보다 복잡한 모델로 ARMA (ARIMA), 인공신경망, 상호정보[3] 등이 있다. 전통적인 통계 기법으로 앞서 언급한 자기회귀모델과 이동평균 모델을 결합한 ARMA 모델은 계절성, 비 정지성(non-stationary) 등 다른 요소에 의해 예측에 한정된 역할을 하게 되는 단점이 있다. 그리고 통계적 방법은 항상 고도의 기술적 제한 조건 및 적용 환경을 동반하지 않는 경우 모델링은 제한적일 수밖에 없다[6]. 인공신경망도 시계열 자료 처리 및 예측에 자주 사용되고 있다. 문제에 한정적인 신경망 구조 때문에 매우 문제 지향적인 해결방법이라 할 수 있다[3,4]. 예측의 정확도 여부를 떠나 신경망이 갖고 있는 전형적인 한계인 설명능력 부족으로 인해 실험결과 활용에 제한적일 수 있다는 단점도 갖고 있다[6]. 본 연구에서는 다수의 시계열 자료를 활용하는 자료기반 분석을 통해 단순 모델을 구성하여 예측에 활용하고자 한다. 이를 실험을 통해 확인해 보인다. 단순 모델로는 마코프 모델을 기본 모델로 선정한다.

1. 마코프 체인 모델

마코프 모델[7]이 N개의 상태를 갖고 상태를 Q라 하자. 현재 시각 t에서의 상태는 t이전의 지난 n개의 상태에 영향을 받는다고 할 수 있다. 이것을 바탕으로 전이확률 $A(t)$를 식(1)과 같이 나타낼 수 있다. 현재 $(n-1)$번째의 상태와 전이확률을 안다면 n 번째의 상태는 전이확률을 통해서 알 수 있다

$$A(t) = \left\{ a_{i_n \cdots i_1 j}(t) \right\}_{i_1, \ldots, i_n, j = 1}^{N} \quad (1)$$

$$a_{i_n \cdots i_1 j}(t) = P(Q(t) = j \mid Q(t-1) = i_1, \; Q(t-2) = i_2, \ldots, Q(t-n) = i_n)$$

주어진 시계열 자료를 상태와 상태 전이확률을 갖는 마코프 모델로 만들기 위해서는 시계열 자료에 대한 상태 값을 할당하는 작업을 수행해야 한다. 먼저 시계열 자료를 식 (2)와 같은 로그 비(log ratio) r_t 형태로 변환한 후 그 로그 비 값의 크기에 따라 구간을 나누고 각 구간에 상태 값을 할당한다.

$$r_t = \ln\left(\frac{y_t}{y_{t-1}}\right) \quad (2)$$

y_t: 시각 t에서의 관측 값, y_{t-1}: 시각 $t-1$에서의 관측 값

현재의 상태가 과거 r 개의 관측값에 의해 결정된다고 가정하면 r 차 마코프 체인 모델이 된다. 일반적으로 1차 마코프 체인 모델이 사용된다. 본 연구에서는 r 값을, 1에서 5까지 시도하였다. r이 3 이상에서는 주요한 패턴을 찾기가 힘들었고, 예측결과 또한 랜덤한 예측보다 더 개선되지 않았다. 그 이유는 전이확률의 개수가 기하급수적으로 늘어나 확률밀도 값이 너무 작아 주도적인 규칙을 산출하기가 어려웠기 때문이다. r을 2로 정하여 방법간의 예측 정확도를 측정하여 비교 분석하기로 한다. 즉, 과거 2개의 정보와 현재의 값으로 이뤄진 패턴 찾기를 통해 다음 값을 예측하는 시스템이 된다.

2. 자료의 정규화

시계열 자료를 모델링하기 위해서는 자료에 대한 정규화 작업이 필수적이다. 동적으로 변화하는 시계열 자료의 경우에 업종별, 종목별로 지수 값의 편차가 크다. 이들의 변화 패턴을 조사 분석하는 것은 지수 값이 일치하는 것이 아니라 값의 흐름이나 추세 즉, 시계열 자료가 나타내는 모양의 유사성을 찾아내는 것이다. 여기서 사용한 전처리 방법은 정규화 과정으로 식(3)과 같다.

$$v_t^{'} = \frac{v_t - \mu_s}{\sigma_s} \quad (3)$$

v_t: 시간 t에서의 주가, μ_s:시계열 자료 s의 평균 값

σ_s: 시계열 자료 s의 표준편차,

$v_t^{'}$: 시간 t에서의 정규화된 값

Ⅲ. 본론

본 연구에서는 마코프 체인 모델을 이용하여 패턴을 구분하고 각 패턴이 출현할 가능성을 확률분포로 찾는다. 그 후 각 패턴의 발생가능성이 가장 큰 패턴을 찾아 그 패턴 이후의 값을 가지고 향후 변화를 예측하는 방법을 취한다. 시계열 자료는 업종별 주가지수를 선택하였다. 각 자료는 2006년도 기간의 자료로 총 247개로 이뤄진다. 자료 선정에는 상승, 하락, 보합이 적절히 포함되어 있는 자료를 선정하도록 노력하였다. 상승이나 하락으로 치우쳐진 년도의 경우에 실험결과가 왜곡될 가능성이 있기 때문에 자료 선정에 주의를 기울였다.

예측정확도와 예측율을 조사하는 방식에는 크게 세 가지로 구분한다. 모델 구성에 사용되는 데이터 구간의 변동에 따른 구분과 예측모델 구성에 사용되는 데이터의 길이에 관한 종류에 따라 3가지로 구분한다. 모델 형성에 사용되는 데이터 구간이 고정되는 방식을 고정방식으로 부르기로 한다. 모델형성에 사용되는 데이터의 길이는 고정되어 있으나 그 구간이 이동하면서 모델이 형성되는 방법은 창이동 방식이라 부르기로 한다. 모델형성 구간이 점진적으로 증가하면서 누적자료를 이용하여 모델이 갱신되는 방식을 누적구간 방식이라 정의한다.

예측하는 데이터 구간의 길이도 세 가지로 나눠 실험한다. 246개(로그비를 사용하므로 247개보다 1이 적음)에 대한 자료를 대상으로 모델 구성 데이터 구간을 처음 200개, 220개, 230개로 나눠서 모델을 구성한다. 각 경우에 대해 테스트 자료는 40개, 26개, 16개로 나눠서 테스트한다. 40개의 경우를 장기 예측, 26개의 경우를 중기 예측, 16개를 단기 예측으로 구분하여 실험하기로 한다. 만일 200개의 업종별 지수자료를 모델형성 구간으로 정하면 201번째 자료부터 40개의 테스트 자료가 예측 정확도를 측정하기 위한 자료로 사용된다. 이와 같이 최소의 테스트 자료 개수를 16으로 정하는 이유는 만일 테스트 자료의 개수가 10이하로 작아지면 예측정확도의 평균수치가 급변하여 안정적인 예측정확도를 측정하기가 힘들기 때문이었다.

대부분의 시계열 예측에는 단일 모델을 중심으로 예측하는 방법을 적용한다. 즉, 하나의 시계열의 변화추세를 감안하여 그 후의 미래 상황을 예측하는 방법이다. 본 연구에서는 단일 시계열 자료대신에 복수 개의 자료를 대상으로 모델을 구성한 후 예상 대상이 되는 시계열 자료의 향후 운행 방향을 예측하는 방법을 제안한다. 먼저 예측을 위한 시계열 자료를 중심으로 이와 유사한 시계열 자료를 선정한다. 개별 업종별 지수자료에 대해 가장 유사한 시계열을 일정 개수로 수집하는 방법을 적용한다. 유사도 측정은 모델기반 방식을 적용하기로 한다. 주어진 시계열 자료를 잘 설명하는 모델로 은니마코프 모델이 있다. 은닉마코프 모델 방식으로 모델을 생성된 후 다른 시계열 자료와 그 모델과 유사한 정도를 나타내는 것으로 우도(likelihood)를 사용한다[8].

시계열 자료 X와 모델 λ가 주어질 때 자료와 모델의 확률을 각각 $p(X)$와 $p(\lambda)$로 표시된다. 모델 λ의 파라

미터 구성, θ가 주어지면 한계우도는 식(4)과 같이 정해진다.

$$p(X|\lambda) = \int_{\theta} p(X|\theta, M)p(\theta|M)d\theta \quad (4)$$

바움 웰치 알고리즘을 적용하여 생성된 모델, λ와 주어진 시계열 자료인 X와의 우도인 $p(X|\lambda)$를 최대값이 되도록 최적의 파라미터가 결정될 때까지 추정단계와 최대화 단계를 반복한다. 전체 알고리즘은 그림 1에 기술되어 있다.

알고리즘
input: 시계열 자료 v_q, 전체 자료집합 $D = \{v_1, v_2, ... v_n\}$, c: 유사시계열 자료수 output: 유사시계열 자료 $SD = \{v_{s1}, v_{s2}, ...\}$ v_q에 대한 은닉마코프 모델 λ_q을 구한다. for i=1 to n ($i \neq q$) LL_i = 우도계산(v_i, v_q) end sort(LL_i's) 최상위 c개 시계열 자료 $\rightarrow SD$ $SD \leftarrow SD + v_q$ SD에 대한 마코프 모델 생성 마코프 모델 생성

그림 1. 유사시계열 모델링
Fig.1. Modeling Similar Series

유사시계열 자료의 선택기준은 모델과 개별 시계열 자료간의 유사도 측도인 로그 우도를 계산한다. 로그 우도는 식(4)에 log를 취한 값으로 결정된다. 상위 c개의 시계열은 우도관점에서 가장 유사한 시계열이 될 것이다. 다중 시계열 자료가 선택되었다면 이로부터 마코프 규칙을 추출해야 한다. 마코프 전이 확률 값을 조사하여 가장 큰 값을 갖는 전이 규칙을 유효규칙으로 결정한다. 단, 만일 가장 큰 전이 확률 값과 차상위의 전이확률 값을 비교하여 그 차이가 특정 값 이상일 때 유효규칙으로 선정된다. 본 실험에서는 10% 이상의 차가 있을 경우에 유효규칙으로 정하기로 한다. 이 차이를 너무 크게 설정하면 유효규칙의 수가 줄어드는 문제가 있으며 너무 작게 설정할 경우에는 예측 오류의 가능성을 높이는 결과를 초래할 것이다.

이와 같은 규칙을 추출하여 적용하게 되므로 일부 상태값 조합에서는 예측을 하지 못하는 경우가 발생한다. 이를 예측율로 조사하여 분석하기로 한다. 예측율은 적용가능한 규칙이 있어 이를 적용하여 예측하는 사례와 예측적용 가능한 규칙이 없어 예측을 유보하는 경우의 비를 가지고 결정한다.

예측율 = 예측된 케이스 수 / 총 케이스 수 x 100%

Ⅳ. 실험 및 결과

1. 실험자료

실험에 사용된 자료는 코스피 지수 중 2006년도 업종별 자료를 중심으로 자료를 선정하였다. 업종에 속한 여러 개별 종목을 통합하여 지수형태로 종합하였기 때문에 개별 자료 보다는 다양한 변수에 민감하게 반응하여 자료를 왜곡할 수 있는 가능성을 제한할 수 있어 이를 선택하였다. 업종별 자료에는 총 22가지 종류의 자료가 있다. 이들 자료는 표 1에 기술하였다.

표 1. 업종별 항목
Table 1. Items by Industrial Category

1.음식료	2.섬유의복	3.종이목재	4.화학
5.의약품	6.비금속광물	7.철강및금속	8.기계
9.전기전자	10.의료정밀	11.운수장비	12.유통업
13.전기가스	14.건설	15.운수창고	16.통신
17.금융	18.은행	19.증권	20.보험
21.서비스	22.제조업		

실험에 사용할 2006년도 업종별 지수 22개의 자료를 대상으로 한 로그 비 자료의 예가 그림 2에 나와 있다. 우측 하단에 있는 축에는 로그비 값을 나타내도록 되어 있다. 대부분의 값이 -0.03에서 0.03까지 이르는 구간에 분포되어 있음을 알 수 있다. 좌측 하단에는 시계열 일련번호 값을 갖는 축이다. 일련번호는 1번에서 22번까지이다. 수직 축은 각 계열의 로그비에 해당하는 빈도수를 나타낸다. 이 그림을 보면 모든 계열의 자료분포가 0을 중심으로 분포된 형태를 취한 것으로 볼 수 있다. 로그비가 음수의 경우는 하락을 의미하고 양수의 경우는 상승을 의미한다.

이 자료는 전처리 과정으로 이전 일과 비교한 로그비를 사용하였고 적정한 임계점을 활용하여 상승, 하락 또

는 보합의 상태를 가지도록 값을 준비하였다. 임계값은 -0.005와 +0.005를 기준으로 +0.005 이상은 상승 상태 값으로, -0.005 이하에는 하락 상태 값으로 설정하고, 그 사이는 보합의 상태로 설정하여 구분하기로 하였다.

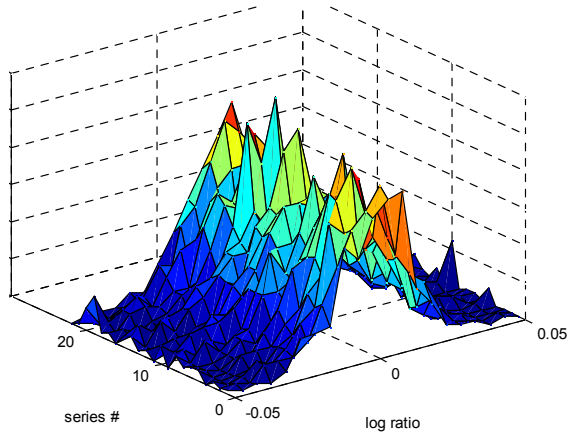

그림 2. 로그비 자료
Fig. 2. Log Ration Data

2. 실험결과

실험은 각 22개 업종별 자료를 개별로 테스트한다. 먼저 개별 업종별 자료에 대해 예측정확도를 살펴보자. 개별 업종에 대해 예측기간에 따라 장기(long), 중기(mid), 단기(short)로 구분하여 정확도를 측정하였다. 단기 기간에 대한 예측정확도가 평균적으로 가장 높을 것으로 예상할 수 있었으며 그것을 수치로 확인할 수 있다(표 2). 개별 업종별 예측 정확도를 표시하면 그림 3 좌측과 같다.

표 2. 전체 예측정확도와 예측율(%)
Table 2. Overall Prediction Accuracy and Prediction Rate(%)

방법	단기	중기	장기
고정구간 (예측율)	49.99 (56.25)	43.18 (58.57)	41.54 (59.66)
창이동 (예측율)	48.40 (58.24)	43.01 (59.09)	43.59 (58.98)
누적구간 (예측율)	47.79 (56.53)	41.30 (58.57)	42.29 (58.07)

중기와 장기는 거의 일정한 변동폭으로 움직인다. 그 변화폭도 상대적으로 크지 않음을 그림을 보고 확인할 수 있다. 반면, 단기예측에 대한 정확도는 전체적으로 높게 형성되었으나 그 편차는 매우 크다. 13번 계열에 대한 예측정확도는 10%대로 하강함을 볼 수 있으며 17번 계열에 있어서는 80%대의 높은 정확도를 산출한다. 전체적으로 정확도를 높이는 것도 중요하지만 일관성이 유지되어야 그 효과가 클 것이다. 따라서 이 그림은 중장기 예측에 대한 정확도는 다소 떨어지나 일관성면에 있어서는 단기예측의 경우보다 좀더 안정성 있는 예측을 한다는 이점이 있음을 보여주고 있다.

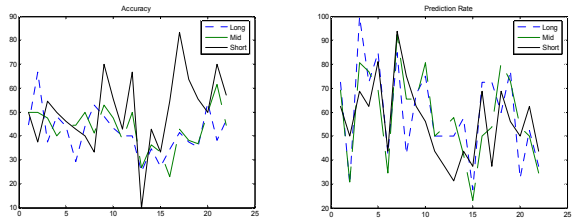

그림 3. 예측정확도와 예측율
Fig. 3. Prediction Accuracy and Prediction Rate

그림 3 우측에 예측율을 표시하였다. 단기, 중기, 장기로 구분하여 각 업종별 예측율을 조사하였는데 전체적인 평균값은 각각, 56.25%, 58.57%, 59.66%로 조사되었다. 다시 말하면 기간에 관계없이 예측을 결정할 수 있는 정도는 거의 비슷해 보인다. 장기 기간에 대한 예측이 다른 2가지 경우보다 약간 상회하지만 그 차이는 미미한 수준으로 나타났다.

예측정확도와 예측율을 요약한 그래프가 그림 4에 표시되었다. 대체로 예측구간이 길어질수록 예측 정확도와 예측율에는 서로 반대되는 역비례 관계가 있음을 확인할 수 있다. 그러나 장기 예측으로 갈수록 예측정확도가 일률적으로 하락하는다. 또 예측율도 일률적으로 상승하지는 않는다.

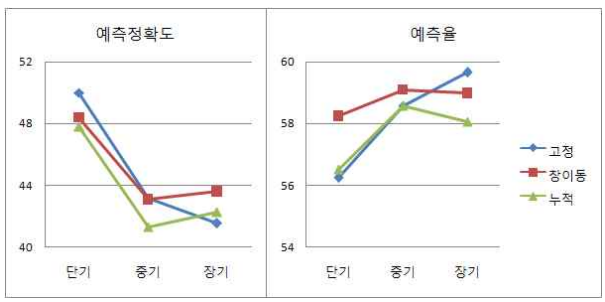

그림 4. 예측정확도와 예측율
Fig. 4. Prediction Accuracy and Prediction Rate

다중 시계열에 대한 모델링을 통해 유효규칙을 유도하여 예측정확도를 개선할 수 있는지 실험해 보았다. 우선 대상이 되는 업종에 대해 다른 시계열과 비교하기 위해 그림 1의 알고리즘을 적용하여 유사계열을 찾았다. 유사도가 높을수록 우도가 높게 산출되므로 우도가 가장 높게 나온 상위 4개의 시계열 자료를 확보하였다. 먼저 시계열 자체를 이용하여 예측정확도를 산출하였고 다음에는 가장 유사한 시계열 자료와 함께 마코프 모델을 만들고 유효 규칙을 산출한 후 이를 이용하여 예측정확도를 측정하였다. 유사한 시계열을 하나씩 추가하면서 예측정확도의 추이를 살펴보았다. 가장 예측 정확도가 높은 누적구간 방식에 다중 유사 시계열 모델 방법을 적용하면 표 3과 같은 결과를 얻는다. 이 표에 대한 그래프가 그림 5에 나타나있다.

표 3. 유사계열 모델의 예측정확도 (%)
Table 3. Prediction Accuracy for Similar Series(%)

유사계열수	단기	중기	장기	평균값
1	48.40	43.07	43.59	45.02
2	41.88	38.39	42.64	40.97
3	43.07	42.22	43.63	42.97
4	45.88	46.51	47.49	46.63
5	41.73	36.27	42.39	40.13

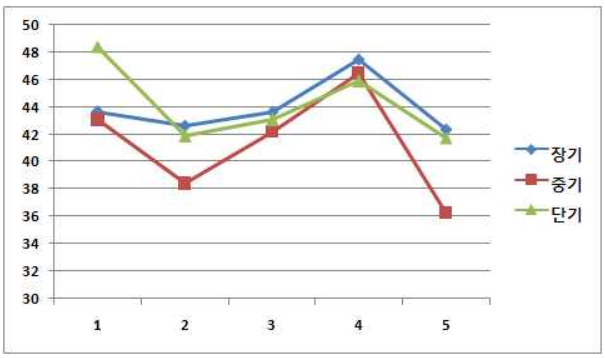

그림 5. 누적구간방식의 예측정확도
Fig. 5. Prediction Accuracy for Cumulative Method

그림을 보면 유사계열이 추가되면서 예측정확도가 변하는 것을 알 수 있다. 그래프의 하단에 있는 일련번호는 유사계열의 수를 의미한다. 그림을 보면 정확도가 일률적으로 상승하거나 하강하는 것이 아니라 유사시계열의 개수가 증가할수록 하강하다가 상승하는 구간이 나타나는 것을 볼 수 있다. 이는 유사계열과 함께 모델링하였을 때 예측정확도가 개선될 여지가 있음을 보여주는 것이라 사료된다. 일반적으로 개별 단기 예측정확도가 가장 높게 나왔으나 중기나 장기에서는 항상 그렇지는 않다. 그리고 유사계열이 여러개로 조합되어 모델링될 때 그 모델은 평균적으로 예측정확도가 가장 높아질 수도 있다. 누적구간 방법에서는 유사 시계열의 총 수가 4일 때 평균적으로 가장 높은 예측 정확도를 나타내었다. 그때 기간 전체에 대한 평균값은 46.63%에 이르는 최대값을 나타내었다.

V. 결 론

본 연구에서는 주가자료를 중심으로 과거의 변화 패턴을 분석하여 향후 상황을 예측하는 방안에 대해 연구하였다. 변화를 상태로 구분한 후 그 상태의 전이확률에 따라 특정 패턴에 따른 다음 상태를 예측하는 방안을 제안하였다. 예측정확도에 있어 기간별에서는 큰 차이를 발견할 수 있었는데 단기예측의 경우는 다른 2가지 기간의 예측에 비해 월등히 정확한 예측이 가능함을 보였다.

개별 시계열자료로부터 예측하는 통상적인 방법에서 발전하여 다중 시계열 자료를 통해 예측하는 방법을 제안하였다. 다중 시계열 선정을 위해 모델과 시계열을 비교하는 측도인 우도를 이용하였고 우도의 크기에 따라 유사도를 결정하여 유사시계열을 선정한 후 다중 시계열 마코프체인 모델을 구성하였다. 다중 시계열 모델로부터 유효규칙을 유도하여 예측에 활용한 결과 3개 또는 4개의 시계열이 통합되어 모델로 생성되었을 때 평균적으로 높은 정확도를 산출하였고, 특히, 중기나 장기 예측에 있어서는 단일 시계열로부터 예측하는 예측정확도보다 더 높은 정확도를 생성해 내었다.

본 논문에서는 개별 유사도에 의한 모델을 구축하려고 시도하였으나 향후 보다 체계적인 집단 구성방법을 연구할 예정이다. 아울러 업종별 지수 분석에서 세부 종목에 적용하는 방법을 연구할 예정이다.

참 고 문 헌

[1] http://en.wikipedia.org/wiki/Hidden_Markov_model

[2] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. of the IEEE, vol.77, no.2, pp.557-286, 1989

[3] M.R. Hassan, B. Nath, M. Kirley, "A fusion model of HMM, ANN, and GA for stock market forecasting," Expert Systems with Applications 33, pp. 171-180, 2007.

[4] Hassan, M.R., & Nath, B., "Stock market forecasting using HMM: a new appreoach," Proc. of 5th International conference on intelligent system design and application, pp.286-291, 2005.

[5] 오유진, 한규숙, 김유섭, "하이브리드 주가예측모델," 한국금융학회 학술대회 논문집, 31-41쪽, 2007.

[6] Duan, J. et al., "A prediction algorithm for time series based on adaptive model selection," Expert Systems with Applications 36, pp. 1308-1314, 2009.

[7] Papageorgiou, C. P., "High frequency time series analysis and prediction using Markov models," in Proc. of the conf. on Comp. Intelligence for Finance, pp.182-185, Mar. 1997.

[8] 전진호, 이계성, "시계열 데이터의 모델기반 클러스터 결정에 관한 연구", 한국콘텐츠학회 논문지 제 7권 6호, 22-30쪽, 2007년 6월.

[9] C. Li, and G. Biswas, "Building models of ecological dynamics using HMM based temporal data clustering," IDA 2001, pp. 53-62. 2001

[10] A. Sorjamaa, et al., "Methodology for long-term prediction of time series," Neurocomputing, pp. 178-186. Elsevier, 2007.

※ 본 연구는 2009학년도 단국대학교 대학연구비 지원으로 연구되었음.

저자 소개

조 영 희(준회원)

• 2000년: 단국대학교 이학석사.
• 2008년: 단국대학교 이학박사.
• 2008년 ~ 현재: 단국대학교 컴퓨터과학과 강사

<주관심분야 : 데이터마이닝, 지능형시스템, 시계열 자료분석>

이 계 성(정회원)

• 1982년: 한국과학기술원 이학석사.
• 1994년: Vanderbilt 대학 공학박사.
• 1994년 ~ 현재: 단국대학교 컴퓨터과학과 교수

<주관심분야 : 데이터마이닝, 지능형시스템, 시계열 자료분석, 바이오 인포마틱스>