韓國컴퓨터情報學會 論文誌
第16卷 第2號, 2011. 2.

2011-16-2-4-5

마코프 모델에 기반한 시계열 자료의 모델링 및 예측

조 영 희*, 이 계 성*

Modeling and Prediction of Time Series Data based on Markov Model

Young-Hee Cho*, Gye-Sung Lee*

요 약

주식 가격이나 경제 지표, 사회적 현상의 추세나 변화 등은 통상 시간에 따라 변화하기 때문에 시계열 자료로 구분된다. 시계열 자료는 시간 축에 대해 변화하는 자료의 표현 가치뿐 아니라 그 변화 추세나 향후 방향성까지 제시할 수 있다는 점에서 이에 대한 방법론에 대해 많은 연구와 노력이 지속되어 왔다. 본 논문에서는 전통적으로 예측 모형을 구축하여 예측하는 방법을 취하되 그 모형이 복잡하고 정교한 모델을 활용하여 예측 정확도를 높이려는 시도와는 달리 자료 클러스터링 방법과 자료 구간 선정을 통해 예측정확도를 높이려 시도하였다. 기본 모델은 마코프 모델이다. 구간별 유사 구간을 추출하여 모델링하는 구간별 모델링 방법과 클러스터링을 통한 그룹별 모델링을 통해 모델의 예측정확도를 개선하려 시도하였다. 실험을 통해 클러스터링을 거친 그룹별 마코프 모델이 정확도를 개선 시켰으나 예측율은 현저히 떨어지는 결과를 낳았다.

▸ Keyword : 시계열, 마코프 모델, 예측모델

Abstract

Stock market prices, economic indices, trends and changes of social phenomena, etc. are categorized as time series data. Research on time series data has been prevalent for a while as it could not only lead to valuable representation of data but also provide future trends as well as changes in direction. We take a conventional model based approach, known as Markov chain modeling for the prediction on stock market prices. To improve prediction accuracy, we apply Markov modeling over carefully selected intervals of training data to fit the trend under consideration to the model. Another method we take is to apply clustering to data and build models of the resultant clusters. We confirmed that clustered models are better off in predicting, however, with the loss of prediction rate.

▸ Keyword : time series, Makov model, Prediction model

• 제1저자 : 조영희 교신저자 : 이계성
• 투고일 : 2010. 11. 12, 심사일 : 2010. 11. 24, 게재확정일 : 2010. 11. 25.
* 단국대학교 컴퓨터과학과(Dept. of Computer Science, Dankook University)
※ 이 연구는 2010학년도 단국대학교 대학연구비 지원으로 연구되었음.

Ⅰ. 서 론

많은 실세계 시스템은 그 특성상 동적인 성질을 지니는 경우가 많다. 동적 특성을 지닌 시스템이나 프로세스의 운행 행태나 그들로부터 산출 되는 자료의 형태를 분석하여 이들의 특성을 이해하는 것이 매우 중요한 과제이다. 자료를 분석하여 모델로 구성하여 시스템이나 프로세스를 설명한다. 이렇게 유도된 모델은 다양한 용도로 활용된다. 예로, 공정제어 관리와 같은 경우, 공정의 성능 평가를 실시간으로 모니터하면서 모델을 사용하여 설정된 목표에 부합하면서 시스템이 제어되는지 여부를 확인할 수 있다. 기계적인 작동에 오류 여부를 확인하는데도 사용될 수 있다. 이들 동적인 특성은 시계열 자료로 표현되는 경우가 대부분이다. 그 외에도, 우리가 흔히 접하는 주가나 경제지표와 같은 자료들은 전형적인 시계열 자료이다.

주식 가격이나 경제 지표 외에도 의학분야의 진료정보, 과학 공학 분야의 실험자료, 사회적 현상의 추세나 변화 등은 통상 시간에 따라 변화하기 때문에 시계열 자료로 구분된다. 시계열 자료는 시간 축에 대해 변화하는 자료의 표현 가치뿐 아니라 그 변화 추세나 향후 방향성까지 제시할 수 있다는 점에서 이에 대한 방법론에 대해 많은 연구와 노력이 지속되어 왔다. 금융시장에서의 예측 문제는 특별히 집중적으로 연구되어 오는 분야 중 하나이다. 전통적으로 예측 모형을 구축하여 예측하는 방법을 취한다. 그 모형은 과거의 변화 형태에 기초하는 경우가 많다. 여러 가지 기술적 변수, 미시적, 거시적 관점의 경제 지표 등, 다양한 자료나 정보에 대해 복잡한 모델을 형성해 나가는 시도가 계속되어 왔다. 목표는 복잡하고 정교한 모델을 활용하여 예측 정확도를 높이려는 시도이다. 더욱이, 인공신경망, 유전자 알고리즘, Support Vector Machine (SVM) 등, 다양한 인공지능 기술과 베이지안 통계 및 은닉 마코프 모델 등을 활용하여 다양하고 복잡한 방법들이 제안되고 이들을 통해 개선된 결과들이 계속 출현하고 있는 것이 현재의 상황이다[2,3,6,10,11]. 그러나 많은 경우 이 방법들의 강력한 모델링 성능과 설명 가능한 표현력은 뛰어나지만 금융시장의 운행과 같은 실세계 문제를 모델링하는 데에는 여러 가지 어려움을 극복해야 하는 경우가 많다. 금융지표나 경제 지표 등은 경제뿐만 아니라 정치, 사회, 및 국제적 요소들이 복합적으로 작용하고 시장 참여자들의 구성 및 내외적 상세 요소들에 의해 민감하게 반응하며 통합 작용하여 단순 수치로 결정되는 만큼 경제나 금융 재무시장의 변화를 예측한다는 것은 대단히 어려울 수밖에 없기 때문이다. 결국 모델은 대체로 복잡해지고 복합적이며 다양해 질수밖에 없을 것이다[13].

모델이 복잡해짐으로 예측 정확도가 비례적으로 개선된다면 모델의 복잡도는 타당하게 여겨질 수 있을 것이다. 그러나 실제로 모델의 복잡도와 예측 정확도는 원하는 것처럼 비례하지 않는 것이 일반적인 결론이다. 오캄의 면도날[1]이라는 이론에서는 모델의 구성에 관한 방향을 제시해 준다. 즉, 단순한 모델과 복잡한 모델이 대등하게 어떤 현상을 잘 설명할 수 있다면 단순한 모델이 선호된다는 사실이 있다. 단순한 모델의 유연성과 일반성을 겸비하여 예측을 위한 활용에 있어 복잡한 모델에 비해 월등한 성능을 발휘하는 경우가 많기 때문이다.

본 논문에서는 시계열 자료 분석이나 예측에 자주 활용되고 있는 단순한 모델의 하나인 마코프 체인 모델을 활용하는 방안에 대해 연구한 결과를 기술한다. 제안하는 방법은 마코프 모델을 선택하여 그 모델의 유효성을 조사하고, 모델의 부족한 부분을 보완하는 과정을 보인다. 본 연구에서는 2가지의 중요한 방법을 마코프 체인 모델에 부수적으로 함께 활용하여 예측 정확도를 증진시키는 방안에 대해 기술한다. 그 방법은 구간별 모델링과 클러스터 별 모델링 방법이다. 2장에서는 배경이론을 설명하고, 3장에서는 마코프 체인 모델을 예측에 활용하는 방안을 제시하고, 4장에서는 이를 실험하여 산출된 결과에 대해 분석하고 그 방법에 대한 타당성을 검토한다. 마지막으로 5장에서는 결론을 기술한다.

Ⅱ. 관련 연구

시계열 자료에 대한 모델링 기법에는 여러 가지가 있다. 그 중 시계열 자료의 특징을 가장 손쉽게 표현할 수 있는 방법의 하나로 마코프 체인 모델이 있다. 이는 마코프 모델의 가장 간단한 모델로 n 개의 상태로 이루어진 시간적 처리 모델이다. 현 상태는 바로 직전의 과거 k개의 상태에 의해 시간적으로 연결되어 있다. 연결 상황은 n^k개의 상태전이 확률에 의해 표현된다. 최근에는 시계열 자료의 모델기반 분석이 좀 더 큰 관심을 끌고 있다. 예전부터 사용하여 오던 자기회귀 모델과 이동 평균 모델을 비롯하여 베이지안 모델, 은닉 마코프 모델 [2,3,4,6,10]등이 그것이다. 자료의 시점 사이의 관계를 추정하는 방법에 자기회귀 모델과 이동평균 모델이 있다. 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 단순한 모델로 복잡하게 변화하는 경제지표까지 설명하거나 예측하는 일에는 한계가 있다고 본다.

보다 복잡한 모델로 은닉 마코프 모델 (HMM, Hidden Markov Model)[7,11], ARMA (ARIMA), 인공신경망[6], 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[4]. 인공신경망도 시계열 자료 처리 및 예측에 자주 사용되고 있다. 문제에 한정적인 신경망 구조 때문에 매우 문제 지향적인 해결방법이라 할 수 있다[2, 4]. 예측의 정확도 여부를 떠나 신경망이 갖고 있는 전형적인 한계인 설명능력 부족으로 인해 실험결과 활용에 제한적일 수 있다는 단점도 갖고 있다[5].

은닉 마코프 모델은 시계열 자료 처리 및 예측에 활발히 응용되고 있는 모델이다[2,10]. 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[6,7]. 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 분석 결과도 당연히 정당화 될 수 없을 것이다.

시계열 자료의 예측 문제를 접근할 때 하나의 시계열 자료에 대한 예측문제에 한정하는 경우가 많다[5,6]. 하나의 시계열이 갖고 있는 자료의 한정성도 문제지만, 특히 주가와 같은 시계열은 무작위성 (랜덤워크)에 의한 움직임이 큰 경향을 가지므로 이런 단일 계열의 자료로 단순히 그것의 기술적 분석을 통해 미래 주가를 예측한다는 것은 매우 어려운 일로 알려져 있다. 무작위성에 의한 예측의 한계성을 조금이나마 해결하기 위해 본 연구에서는 다수의 시계열 자료에 의거한 예측 방법과 구간별 유사 구간에 한정하여 모델링하는 구간별 모델링 방법을 제안하려 한다.

또한 위에서 언급한 여러 방법들이 시계열 분석에서 제한적 분야나 문제에 한정하여 능력을 발휘하거나 또는 문제의 특성을 무시하고 방법론에 문제를 맞춰 해결하려는 시도는 예측이라는 어려운 문제를 더욱 어렵게 할 수 있다. 또한 복잡한 문제를 복잡한 모델로 접근하려는 시도에서는 범용성 적용 문제가 있다고 볼 수 있다. 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.

2.1 마코프 체인 모델

마코프 체인 모델에 대해 먼저 설명하고 시계열 자료의 준비에 대해 정규화와 로그비 (log ratio)로 변환시키는 과정을 설명한다. 자료의 정규화는 여러 시계열 자료를 처리해야 하는 과정에서 각 계열간 절대 수치의 편차를 고려해야 하기 때문에 이를 정규화 시켜야 한다. 시계열 자료는 마코프 체인 모델에 적합하게 적용될 수 있도록 로그비 값으로 전환한다.

마코프 모델[8]은 시간에 따른 상태 추이의 변화를 포착하는 통계적 모델이다. 마코프 모델이 N개의 상태를 갖는다면, N^2의 전이확률로 구성된다. 상태를 Q라 하면 $Q = \{q_i\}_{i=1}^{N}$와 같이 나타낼 수 있고 $Q(t)$는 시간 t에서의 상태를 나타낸다. 시간 t에서의 상태 변화를 나타내는 상태전이확률 $A(t)$는 식 (1)과 같이 나타낼 수 있다.

$$A(t) = \{a_{ij}(t)\}_{i,j=1}^{N} \quad \cdots\cdots (1)$$

$$a_{ij}(t) = P(Q(t) = j | Q(t-1) = i)$$

여기서 $a_{ij}(t)$는 시간 $t-1$에서 상태가 i일 때 시간 t에서 상태가 j로 전이될 확률을 나타낸다. 그리고 상태 i에서 다른 상태로 전이할 모든 확률의 합은 식 (2)와 같이 1이 된다.

$$\sum_{j=1}^{N} a_{ij}(t) = 1 \quad \cdots\cdots (2)$$

마코프 모델의 상태 변화는 바로 이전 상태에 의존적인 특성을 갖는다. 즉, 시간 t에서의 상태는 오직 시간 $t-1$에서의 상태에만 영향을 받고 시간 $t-1$에서의 상태는 시간 $t-2$에서의 상태에만 영향을 받는다. 이것을 좀 더 확장하면 현재 시각 t에서의 상태는 t이전의 지난 n개의 상태에 영향을 받는다고 할 수 있다. 이것을 바탕으로 전이확률을 식(3)과 같이 나타낼 수 있다. 식(3)을 통해서 현재 $(n-1)$ 번째의 상태와 전이확률을 안다면 n 번째의 상태는 전이확률을 통해서 알 수 있다

$$A(t) = \{a_{i_n \cdots i_1 j}(t)\}_{i_1, \ldots, i_n, j=1}^{N} \quad \cdots\cdots (3)$$

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

주어진 시계열 자료를 상태와 상태 전이확률을 갖는 마코프 모델로 만들기 위해서는 시계열 자료에 대한 상태 값을 할당하는 작업을 수행해야 한다. 먼저 시계열 자료를 식 (4)와 같은 로그 비(log ratio) r_t 형태로 변환한 후 그 로그 비 값의 크기에 따라 구간을 나누고 각 구간에 상태 값을 할당한다.

$$r_t = \ln\left(\frac{y_t}{y_{t-1}}\right) \quad \cdots\cdots (4)$$

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

로그 비 상태로 변환된 시계열 자료에서 상태 전이확률을 작성한다. 먼저 시각 $t-2, t-1, t$에서 각 대응하는 상태 값을 i, j, k라고 하면 C_{ijk}는 시각 $t-2$에서는 상태 i, 시각 $t-1$에서는 상태 j, 시각 t에서는 상태 k와 같은 전이가 발생하는 횟수를 모두 더한 것이다. 이제 시각 t에서 상태 값이 i, j, k의 순서로 전이될 확률 a_{ijk}는 식 (5)와 같이 나타낼 수 있다.

$$a_{ijk} = \frac{C_{ijk}}{\sum_{x=1}^{N} C_{ijx}} \quad \cdots\cdots (5)$$

이제 이렇게 작성된 상태 전이확률을 사용하여 시계열 자료의 다음 상황을 예측하게 된다. 만약 $Q(t-1) = i$ 이고 $Q(t) = j$ 라면 $Q(t+1)$은 상태를 결정하기 전에 먼저 로그 비 값을 적용하여 확률 값을 계산한다.

마코프 모델은 가장 최근 r 개의 관측값에 의존한다는 가정하에 t 시간에 관측될 특정 값의 확률을 추론하는데, r 값에 따라 마코프 모델의 차수가 결정된다. 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 시도하였으다. 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 방법 간의 예측 정확도를 측정하여 비교 분석하기로 한다. 즉, 과거 2개의 정보와 현재의 값으로 이뤄진 패턴 찾기를 통해 최선의 값을 예측하는 시스템이 된다.

2.2 자료의 정규화

시계열 자료를 클러스터링하기 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. 그러므로 비교에 사용되는 자료는 일정 크기에 분포되도록 변환시키는 전처리작업을 수행하는 것이 필요하다. 본 논문에서는 이러한 전처리를 위해서 시계열 자료 값의 평균과 표준편차를 사용하여 정규화 한다. 여기서 사용한 전처리 방법은 정규화 과정으로 다음과 같다.

$$v_t' = \frac{v_t - \mu_s}{\sigma_s} \quad \cdots\cdots (6)$$

v_t: 시간 t에서의 주가, μ_s: 시계열 자료 s의 평균 값

σ_s: 시계열 자료 s의 표준편차,

v_t': 시간 t에서의 정규화된 값

이러한 전 처리 작업 절차를 걸쳐 구해진 시계열 자료는 종목이나 업종에 따라 커다란 편차를 나타내지 않고 값들이 일정한 범위 안에 분포되도록 변환된다.

III. 본론

3.1 제안 방법

본 연구에서는 마코프 체인 모델을 이용하여 패턴을 구분하고 각 패턴의 가능성을 확률분포로 찾는다. 그 후 각 패턴의 발생가능성이 가장 큰 패턴을 찾아 그 패턴 이후의 값을 가지고 향후 변화를 예측하는 방법을 취한다. 각 자료는 1년 기간의 자료로 총 247개로 이뤄진다. 자료 선정에는 상승, 하락, 보합이 적절히 포함되어 있는 자료를 선정하도록 노력하였다. 총 22개 자료 중 3개의 자료(철강 및 금속, 의료정밀, 건설)를 그림 1에 표시하였다. 그림에서 보듯이 각 자료는 상승, 하락, 보합 부분이 뒤섞여 있음을 보여주고 있다. 상승이나 하락으로 치우쳐진 년도의 경우에 실험결과가 왜곡될 가능성이 있어 자료 선정에 주의를 기울였다.

실험에 사용되는 방법은 3가지이다. 첫 번째는 마코프 체인 모델을 직접 할용 하는 방안이다. 각 자료 중 첫 200개 자료는 과거 자료로 설정하여 마코프 모델 구성에 학습 자료로

활용된다. 이후 일정기간의 자료(40일간)를 선정하여 모델을 테스트하는 자료로 활용할 것이다. 본 논문에서는 이 방법을 MC 모델링 방법이라 명명한다.

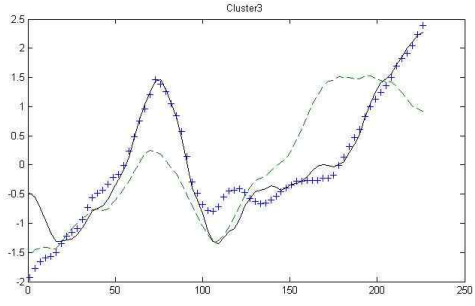

그림 1. 업종지수: 클러스터 3(철강금속, 의료정밀, 건설)
Fig. 1. Industrial indices for Cluster 3

두 번째 방법은 학습자료의 끝에 위치한 모드를 기준으로 학습자료를 인위적으로 선별하여 모델을 만든다. 만일 학습자료의 끝이 상승모드에 있는 경우 상승 구간에 있는 자료들로 이뤄진 모델을 구성하여 테스트 자료를 갖고 테스트한다. 현재 예측하기 직전의 구간이 상승모드에 있다면 과거 자료 중에서 하락장세의 구간을 되도록 피하고 상승장에 있는 구간을 선별하여 이를 모아 마코프 모델을 구성한다면 보다 정확한 예측이 이뤄질 것이라는 예상을 하게 되고 이를 확인하고자 실험을 시행할 것이다. 예측할 구간과 유사한 자료로 모델링 하였으므로 예측정확도가 더 개선될 것으로 기대되는 것은 당연할 것이다. 이 방법은 구간별 모델링이라 부르기로 한다.

세 번째 방법은 클러스터링을 이용한다. 시계열 자료들은 보이는 변화하는 모양의 유사성에 따라 분류하기 위해 클러스터링을 적용한다. 클러스터링은 주어진 자료들 중에서 유사성이 높은 자료들을 같은 클러스터에 할당하고 서로 다른 특성을 달리하여 특성이 상이한 자료는 서로 다른 클러스터에 할당하는 방식으로 진행된다. 이것은 같은 클러스터에 속한 자료들은 유사 특성을 갖는다는 것을 의미하고 (내적 유사성), 클러스터 간에는 자료 들 간의 차이성이 큰 특성을 갖는다는 것 (외적 상이성)을 의미한다. 시계열 자료들에 대한 클러스터링 작업을 수행하면 모양이 유사한 시계열 자료들이 같은 클러스터에 포함될 것이다. 결국 모양이 유사한 시계열 자료들이 같은 클러스터에 포함되어 모양의 유사성에 따라 시계열 자료들을 분류할 수 있게 된다.

클러스터링 방법은 베이지안 클러스터링 방법을 사용한다[9]. N개의 자료에 대해 최선의 분할을 표현하는 최적 클러스터의 수는 1부터 N까지 매우 다양할 수 있다. 최악의 경우 이것은 N개의 클러스터가 될 수 있다. 이것은 계산적으로 매우 큰 비용이 소요된다. 이를 피하기 위해서 미리 선택되어 정의된 베이지안 정보기준 함수[11]에 의해 클러스터의 수를 결정한다. 최선의 분할사이즈 k는 일반적으로 N보다 훨씬 작다. $k=2$로부터 시작하여 클러스터의 수를 하나씩 증가하여 계속 반복해 나가다가 가장 높은 베이지안 기준함수의 값을 가질 때 그 때의 클러스터의 수가 최적의 클러스터 수가 된다. 이와 같은 특성은 베이지안정보기준 측도의 특성[11]에 근거하여 활용한 것이다.

모델기반 클러스터링에서, 자료는 확률분포의 혼합(Mixture)에 의해 생성되어지는 것을 가정한다. 혼합모델 M은 K개의 컴포넌트 모델들에 의해 표현되고 각 컴포넌트는 C_k로 표현된다. 자료 $X=(x_1,\ldots,x_N)$이 주어지면, k번째 컴포넌트(k번째 클러스터) 모델 λ_k에 속하는 개체 x_i인 확률을 $f(x_i \mid \theta_k, \lambda_k)$으로 표현한다. 파라미터들은 θ_k로서 표현되어지며, 혼합모델이 주어졌을 때, 자료의 우도(likelihood)는 식(8)과 같이 표현되어진다.

$$P(X \mid \theta, M) = P(X \mid \theta_1,\ldots,\theta_K,\lambda_1,\ldots,\lambda_K)$$

$$= \prod_{i=1}^{N} P(x_i \mid \theta_1,\ldots,\theta_K,\lambda_1,\ldots,\lambda_K)$$

$$= \prod_{i=1}^{N}\sum_{k=1}^{K} P_k \cdot f(x_i \mid \theta_k,\lambda_k) \quad \cdots\cdots (8)$$

위에서 P_k는 컴포넌트모델 λ_k의 사전확률이다. $P_k = P(x_i \in \lambda_k),\ i=1,\ldots,N,\ k=1,\ldots,K.$ 이다.

베이지안 클러스터링은 모델기반 클러스터링 문제를 베이지안 모델 선택의 문제 형태로 바꾼다. 서로 다른 컴포넌트 클러스터들을 갖는 분할들이 주어 졌을 때, 목적은 가장 큰 사후확률을 갖는 가장 좋은 모델 M을 선택하는 것이다. 다시 말하면 자료에 대한 최적의 혼합모델 M을 찾는 것이다. 최적의 클러스터링 혼합모델 M은 가장 높은 분할사후확률(PPP), $P(M \mid X)$를 갖는다. 혼합모델의 한계우도 $P(X \mid M)$를 분할사후확률에 근사시킨다. 여기에서 클러스터 분할선택을 위한 한계우도의 계산에 베이지안정보기준[11]을 적용한다. $\lambda_1,\ldots,\lambda_K$로서 모델된 K클러스터를 갖는 분할에 대하여, 식(9)처럼 분할사후확률이 정의되고 베이지안정보기준 근사법을 사용하여 그 값이 계산된다.

$$\log P(X \mid \hat{\theta}, M) = \prod_{i=1}^{N}\sum_{k=1}^{K} P_k f(x_i \mid x_i \in \lambda_k, \hat{\theta}, \lambda_k) \quad \cdots\cdots (9)$$

위에서 $\hat{\Theta}$는 K 클러스터의 한계우도 모델 파라미터의 구성을 나타낸다. P_k는 클러스터 k의 사전확률이 되고 $f(x_i|x_i \in \lambda_k, \hat{\Theta}, \lambda_k)$은 클러스터 k에 대한 모델이 주어졌을 때 자료 x_i의 확률을 나타낸 것이며 바움-웰치의 전향절차[12]에 의하여 계산된다. 베이지안정보기준을 적용하여 K 클러스터들을 갖는 분할에 대한 분할사후확률은 식(10)과 같다.

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

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} logP_k + \sum_{i=1}^{N}\sum_{k=1}^{K} f(x_i|\hat{\Theta}_k, \lambda_k) - \frac{K + \sum_{k=1}^{K} d_k}{2}\log N \quad \cdots\cdots (10)$$

식(10)에서 첫 번째 항 $\log P(X|\hat{\Theta}, M)$는 자료에 대한 모델의 우도값을 나타내며, 두 번째 항 $\frac{d}{2}\log N$는 모델 복잡도에 따른 패널티 항이다. d_k는 클러스터내의 의미 있는 파라미터의 수를 나타낸다. 각 개체는 분할에서 주어진 하나의 클러스터에 할당된다. 최선의 모델은 전체 클러스터 분할의 복잡도와 전체 자료의 우도의 조화를 이루는 가운데 결정되는 것이다.

일단 3가지 모델이 완성된 후에는 테스트 자료를 각 모델에 적용하게 된다. 테스트할 때 우리는 전이확률 값에서 가장 높은 확률에 해당하는 상태를 예측하는 것이 아니라 특정 조건을 만족해야 한다. 즉, 예측 상태의 확률 값이 가장 큰 것의 상태가 예측될 것이다. 단, 차상위에 있는 확률보다 일정 수준 이상일 때 상태를 예측할 수 있게 된다. 본 실험에서는 최상위와 차상위의 예측확률이 10%이상의 차이를 가질 때에 예측상태를 결정하게 된다. 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% (11)

IV. 실험결과

4.1 실험자료

실험에 사용된 자료는 2006년도 업종별 자료를 선택하였다. 개별 종목별 자료 보다는 업종별 자료가 외적 환경 요소의 변동요인에 의한 민감도에 둔감하기 때문에 22가지 업종별 자료를 선택하였다. 22개의 업종에 대해서는 표 1에 기술하였다.

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

1.음식료	2.섬유의복	3.종이목재	4.화학
5.의약품	6비금속광물	7철강및금속	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		

2006년도 개별지수의 추세곡선이 상승이나 하강으로 편중되어 있지 않아 실험의 공정성을 평가하기에 상대적으로 적정할 것으로 판단하여 이 지수 자료를 사용하기로 하였다.

실험에 사용할 2006년도 업종별 지수 22개의 자료를 대상으로 한 로그 비 자료의 예가 그림 2에 나와 있다. 이 그림을 보면 분포가 0을 중심으로 분포된 형태를 취한 것으로 볼 수 있다. 음수의 경우는 하락을 의미하고 양수의 경우는 상승을 의미한다. 하나의 계열에서 양수 쪽에 자료의 분포가 많다면 대상 기간 동안 전체적으로 상승국면임을 나타낸다고 할 수 있다. 그러나 22개의 지수자료 전체를 대상으로 한 자료에 대한 히스토그램은 전체적으로 섞여 있어 개별적인 대세를 구분하기 쉽지 않다. 실제로 상승한 일수가 하락할 일수가 약 4%정도가 많으므로 전체적으로는 미세한 상승 추세로 봐야 할 것이다.

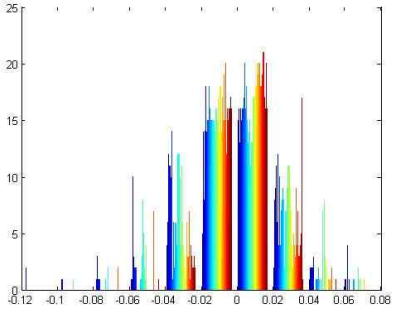

그림 2. 자료의 분포도
Fig. 2. Distribution of Data

이 자료는 전처리 과정으로 이전 일과 비교한 로그비를 사용하였고 적정한 임계점을 활용하여 상승, 하락 또는 보합의 상태를 가지도록 값을 준비하였다. 그림2에 나타난바와 같이 대부분의 빈도수가 -0.03에서 +0.03 사이에 밀집되어 있음을 보여준다. 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.

4.2 실험 결과

실험은 각 22개 업종별 자료를 개별로 테스트한다. 각 자료에는 247개의 지수 자료로 이뤄져 있다. 처음 200개의 자료를 통해 마코프 모델을 구성한다. 200개 이후 40개의 자료를 테스트 자료로 설정하여 생성된 모델에 적용하여 예측 정확도와 예측율을 산출해 낸다. 여기에 3가지 방법이 실험에 사용되었다. 첫 번째 방법은 학습자료인 200개 자료에 대하여 마코프 모델을 구성하여 테스트 자료를 적용하는 방법이다. 두 번째 방법은 200개 직전의 자료가 전체적으로 상승 국면을 이루고 있으므로 상승구간을 식별하여 수동으로 이 구간을 추출하여 이를 바탕으로 마코프 모델을 형성한다. 이 방법은 전체구간에 걸쳐 마코프 모델을 구축하는 것에 비해 패턴이 유사한 구간을 비교해 봄으로 예측 정확도를 인위적으로 높이려는 시도에서 이 방법을 고안하게 되었다. 이후 결과에서도 다시 논의되겠지만 인위적으로 예측정확도를 높이려는 의도와 반대되는 실험 결과가 산출되어 주가 지수의 랜덤워크의 특성을 재확인하는 계기가 되었다. 세 번째 방법은 22개 업종에 대해 클러스터링을 적용한 후 클러스터별 마코프 모델을 구성한다. 이 방법은 유사한 추세로 움직이는 시계열 자료를 묶어 유사한 패턴들을 강화시켜 예측에 적용하여 예측 정확도를 높이려는 시도이다. 클러스터링 방법에 의해 3개의 클러스터가 구성되었다. 클러스터 별 소속 업종이 표 2에 표시되었다.

표 2 클러스터
Table 2. Clusters

클러스터 1	1 3 4 5 9 11 12 15 17 18 19 21 22
클러스터 2	2 6 8 13 16 20
클러스터 3	7 10 14

시각적으로 구분하기 쉽도록 가장 작은 규모의 클러스터 3에 속하는 업종의 지수에 대한 그래프가 그림 1에 표시되었다. 이 그래프는 식별하기 쉽도록 20일 이동평균선을 그린 것이다. 좌측 축은 로그비의 값을 표시하고 x 축은 일별 자료를 보여둔다. 그림 1에서 보듯이 유사한 곡선들이 그려짐을 확인할 수 있다. 이들 클러스터들이 마코프 체인의 전이 확률을 구하기 때문에 유사한 패턴에 대한 것을 강조할 수 있어 예측 정확도 개선에 좀 더 기여하지 않을까 예상하게 된다.

3가지 방법에 대한 실험 결과가 표 3에 표시되었다. 표 3은 전체 시계열 자료에 대한 예측 정확도를 측정한 평균 예측 정확도이다. 표에서 보듯이 클러스터링 방식에 의한 모델이 예측 정확도가 가장 높게 산출되었다. 그 다음은 구간별 모델의 예측 정확도가 MC 모델링 방법 보다 낮게 산출된 점이 매우 이례적이라고 할 수 있다. 유사한 자료를 통해 모델링한 후 모델과 유사한 상황에 적용시켰는데도 불구하고 정확도는 도리어 반대되는 결과를 나타내 보였다는 것은 주식과 같은 시계열은 패턴대로 움직이는 것이 아니라는 것을 다시 한번 확인하게 된다. MC 모델에 의한 예측 정확도는 대상 자료의 분포에 근거하여 비교하여 보면 그것을 약간 상회하는 정도의 예측 정확도를 보인 것으로 판단된다.

표 3. 전체 예측정확도와 예측율
Table 3. Overall Prediction Accuracy and Prediction Rate

방법	예측 정확도(%)	예측율(%)
MC	44.6	64.7
구간별	42.8	70.6
클러스터	53.6	38.4

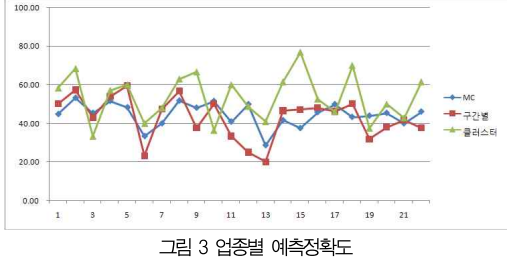

그림 3 업종별 예측정확도
Fig. 3. Prediction Accuracy by Category

그림 3은 개별 자료에 대한 예측 정확도를 나타낸 도표이다. 그래프의 y 축은 예측 정확도를 나타내고 x축은 업종 번호를 나타내고 있다. 클러스터링에 의한 예측 정확도가 거의 대부분의 경우에서 다른 두 모델의 결과를 상회하고 있음을 확인할 수 있다. 구간별 특정 상황을 추출하여 예측정확도를 측정하였지만 예상과 달리 예측 정확도는 가장 나쁜 것으로 나왔음을 보여주고 있다. 전체에 대한 평균 예측 정확도와 비교해도 같은 결과를 이 그래프를 통해 확인할 수 있다. 예상을 훨씬 뛰어 넘는 53%의 예측 정확도를 나타내고 있는 클러스터링에 의한 모델에서는 개별에서 소폭으로 다른 모델보다 상회하거나 소폭 하회하는 경우도 있으나 여러 종목에 걸쳐 월등히 상회하는 결과에 의해 전체 수치가 크게 개선된 것으로 나타났다. 그러나 이와 같이 크게 개선된 이면에는 다른 희생을 감수한 결과로 판명되었다. 즉, 클러스터링 모델에서는 추출된 규칙이 적어 예측을 하지 못하는 경우가 빈번히 발생하기 때문이다. 예측 가능성을 정량화한 예측율에 있어 클러스터링 모델의 예측율이 다른 모델에 비해 현저히 낮음을 그림 4를 통해 확인 할 수 있다. 다시 말하면 예측정확도는 높으나 예측을 유보한 경우가 다른 모델에 비해 상당히 많다는 사실을 말하는 것이다. 이를 해석한다면, 매우 보수적인 예측을 통해 예측정확도를 높인 효과라고 해석할 수 있겠다.

표 4. 클러스터별 정확도와 예측율
Table 4. Accuracy and Prediction Rate by Cluster

방법	클러스터1 (%)		클러스터 2 (%)		클러스터 3 (%)	
	정확도	예측율	정확도	예측율	정확도	예측율
MC	45.4	59.8	43.06	65.8	44.4	83.3
구간	42.8	69.8	40.5	65.8	47.9	83.3
클러스터	55.3	28.5	52.5	47.1	48.7	64.2

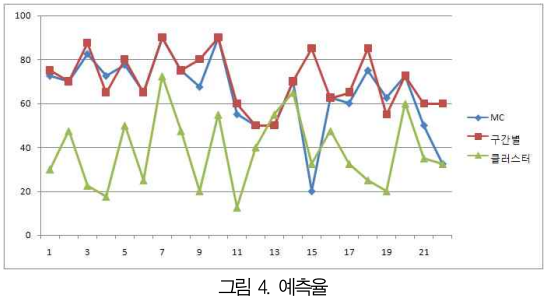

그림 4. 예측율
Fig. 4. Prediction Rate

이를 클러스터별로 구분하여 예측정확도와 예측율을 비교하여 보자(표 4). 전체적으로는 클러스터링을 이용한 예측 정확도가 크게 개선된 것으로 나왔으나 개별 클러스터 각각에 대해 분석하면 클러스터별로 차이가 있음을 확인하게 된다.

클러스터 1이나 2에 속해 있는 업종 들 중에서 예측정확도는 전체의 평균과 약간의 차이가 있는 것을 확인할 수 있다. 클러스터 1에서는 클러스터링의 전체 평균 예측정확도보다 증가되었음을 보여준다(55.3%). 클러스터 3에서는 구간별 예측 정확도가 높아져있고 반면 클러스터링에 의한 예측 정확도는 반대로 낮아져 있다. 클러스터 3은 소수개로 이뤄진 클러스터이다. 이는 소수 개로 이뤄진 업종들의 유사성에 의해서 구간별 예측 방법이 좀 더 정확도를 높이는 것이라고 해석할 수 있을 것이다. 클러스터링 방법은 53.6%에서 48.7%로 감소되는 결과가 나왔는데 이는 소규모의 클러스터는 클러스터링의 효과를 크게 기대하기는 어렵다고 볼 수 있을 것으로 해석된다. 그러나 클러스터 3에서의 예측율은 전체적으로 크게 개선된 것을 보여주는데, 특히 클러스터링에 의한 방법의 예측율이 크게 증대되었다.

V. 결론

주가 지수와 같은 변동성이 심하고 외적 환경요소에 영향을 받아 모델링이 매우 어려운 분야는 모델의 복잡성으로 그들을 해석하고 분석하여 예측하려는 시도가 많았다. 본 연구에서는 변화를 예측하기 힘든 시계열 자료를 처리하는데 있어 복잡한 모델로 접근하기 보다는 간략한 모델로 접근하고 그 특성에 맞게 자료를 가공하고 부수적인 방법을 동원하여 예측정확도를 높이는데 그 목적이 있다고 본다. 그 부수적인 방법에는 구간별 모델링와 클러스터링을 통한 그룹별 모델링을 통해 예측 정확도를 개선하고자 시도하였다. 개별 종목지수가 가질 수 있는 변동성을 축소시키기 위해 업종별 주가지수 시계열 자료를 중심으로 실험하였고 마코프 체인 모델과 같이 간략한 패턴을 중심으로 한 예측방법을 활용하였다. 마코프 모델을 이용한 예측 정확도를 기준으로 두 방법에 대한 결과를 분석해 보았다. 시각적으로 현저한 상승구간을 택하여 상승장에 연속인 구간을 예측하여 예측 정확도를 높이려 하였으나 주가지수의 랜덤워크의 한계를 벗어나지 못하였을 뿐만아니라 도리어 예측정확도가 떨어진 것을 확인할 수 있었다. 반면 클러스터링을 이용하여 그룹별 모델을 구성한 후 예측정확도를 계산한 결과 현저한 개선을 가져올 수 있었다. 클러스터별로 각각, 9%, 9.44%, 4.3%의 개선효과를 가져올 수 있었다. 그러나 예측율을 보면 83%에서 64%로 저하됨을 확인하였다. 예측이 가능한 케이스가 현저히 줄어든 현상을 통해 얻어진 결과에 기인한 것이라 할 수 있다. 클러스터링을 통한

예측은 보수적인 접근으로 활용될 수 있음을 보여주고 있다. 예측을 정확히 하기 위해서는 유사 시계열로 이뤄진 집단에 대한 모델을 통해 비록 예측 비율이 낮을지라도 예측을 하는 경우에는 정확도가 현저히 높은 결과에 이른 것이다.

본 논문에서는 고정된 기간 동안 예측 확률을 수집하여 분석하였다. 향후, 단기, 중기, 장기 예측을 위한 실험방법을 연구하여 이들간의 차이를 비교분석해 볼 것이다. 아울러 업종별 지수 분석에서 세부 종목에 적용하는 방법을 연구할 예정이다.

참고문헌

[1] P.H. Winston, “Artificial Intelligence,” Addison Wesley, 3rd ed., 1992.

[2] H.J. Park, "Use of HMM for stock market Prediction," MA thesis, Sungkyunkwan Univ., 2008.

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

[4] Y, Oh, et al., "Hybrid stock price prediction," Proc. of Korea Financial Society, pp. 31-41, 2007.

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

[6] 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.

[7] 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.

[8] Papageorgiou, C. P., "High Frequency Time Series Analysis and Prediction using Markov Models," in Proc.of the conf. on Computational Intelligence for Finance, pp. 182-185, Mar. 1997.

[9] J. Jeon, G. Lee, “A study on determination of model based clusters of time series data,", J. of the Korea Contents Society, vol 7, no 6, pp. 22-30, Jun. 2007.

[10] Y. Cho, J. Jeon, G. Lee, “Prediction of time-series data using HMM and similarity search for CRM," J. of the Korea Society of Computer and Infomation, Vol. 14, No. 5, pp. 19-28, May, 2009.

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

[12] I.L. MacDonald and W. Zucchini, Hidden Markov and other models for discrete valued time series, Chapman and Hall/CRC, 1997.

[13] J. Lee, D. Park, “Earning rate analysis using Knowledge base HTS for individual, institute, foreigners during different periods of time," Journal of the Korea Society of Computer and Information , Vol. 15, No. 1, pp. 207-217, Jan. 2010.

저 자 소 개

조 영 희

2000: 단국대학교 이학석사.

2008: 단국대학교 이학박사.

2008- 현재: 단국대학교 컴퓨터과학과 강사

관심분야: 데이터마이닝, 지능형시스템, 시계열 자료분석

E-mail : zeroch@dankook.ac.kr

이 계 성

1982: 한국과학기술원 이학석사.

1994: Vanderbilt 대학 공학박사.

1994~: 현재: 단국대학교 컴퓨터과학과 교수

관심분야: 데이터마이닝, 지능형 시스템, 시계열 자료분석

E-mail : gslee@dku.edu