韓國컴퓨터情報學會 論文誌
第14卷 第5號, 2009. 5.

2009-14-5-1-3

CRM을 위한 은닉 마코프 모델과 유사도 검색을 사용한 시계열 데이터 예측

조 영 희*, 전 진 호*, 이 계 성**

Time-Series Data Prediction using Hidden Markov Model and Similarity Search for CRM

Cho Young Hee*, Jeon Jin Ho*, Lee Gye Sung**

요 약

시계열의 예측에 대한 문제는 오랫동안 많은 연구자들의 연구의 대상이었으며, 예측을 위한 많은 방법이 제안되었다. 본 논문에서는 은닉 마코프 모델(Hidden Markov Model)과 우도(likelihood)를 사용한 유사도 검색을 통하여 향후 시계열 데이터의 운행 방향을 예측하는 방법을 제안한다. 이전에 기록된 시계열 데이터에서 질의 시퀀스(sequence)와 유사한 부분을 검색하고 유사 부분의 서브 시퀀스를 사용하여 시계열을 예측하는 방법이다. 먼저 주어진 질의 시퀀스에 대한 은닉 마코프 모델을 작성한다. 그리고 시계열 데이터에서 순차적으로 일정 길이의 서브 시퀀스를 추출하고 추출된 서브 시퀀스와 작성된 은닉 마코프 모델과의 우도를 계산한다. 시계열 데이터로부터 추출된 서브 시퀀스 중에서 우도가 가장 높은 시퀀스를 유사 시퀀스로 결정하고 결정된 부분 이후의 값을 추출하여 질의 시퀀스 이후의 예측 값을 추정한다. 실험 결과 예측 값과 실제 값이 상당한 유사성을 나타내었다. 제안된 방법의 유효성은 코스피(KOSPI) 종합주가지수를 대상으로 실험하여 검증한다.

Abstract

Prediction problem of the time-series data has been a research issue for a long time among many researchers and a number of methods have been proposed in the literatures. In this paper, a method is proposed that similarities among time-series data are examined by use of Hidden Markov Model and Likelihood and future direction of the data movement is determined. Query sequence is modeled by Hidden Markov Modeling and then the model is examined over the pre-recorded time-series to find the subsequence which has the greatest similarity between the model and the extracted subsequence. The similarity is evaluated by likelihood. When the best subsequence is chosen, the next portion of the subsequence is used to predict the next

• 제1저자 : 조영희 교신저자 : 조영희
• 투고일 : 2009. 04. 07, 심사일 : 2009. 04. 27, 게재확정일 : 2009. 05. 08.
* 단국대학교 전자계산학과 ** 단국대학교 컴퓨터학부

phase of the data movement. A number of experiments with different parameters have been conducted to confirm the validity of the method. We used KOSPI to verify suggested method.

▸ Keyword : 은닉 마코프 모델(Hidden Markov Model), 예측(Prediction), 시계열(Time Series)

I. 서 론

시간의 흐름에 따라 발생한 데이터를 수집하여 기록한 시계열은 과학뿐만 아니라 경제, 의료 등 다양한 분야에서 대량으로 생산되고 있다. 이렇게 발생한 시계열 데이터를 분석하여 그 시계열 데이터가 내포하고 있는 특징을 찾아낸다면 그 특징은 시계열 데이터를 이해하고 분석하는데 도움을 줄 뿐만 아니라, 현재 또는 미래의 방향성을 제시하는데도 유용하게 활용될 수 있을 것이다. 특히 과거의 데이터에서 찾아낸 특징을 사용하여 미래의 데이터 변화를 예측하는 문제는 오랜 동안 관심의 대상이 되었다. 그런데 대부분의 시계열은 다음에 어떠한 값을 나타낼 것인가를 알 수 없는 비 규칙적인 변화를 갖는 동적 특성을 가지고 있다. 이러한 동적인 변환을 갖는 시계열이 미래에 어떠한 값을 나타낼 것인가를 정확히 예측해 낸다는 것은 매우 어려운 일이다. 그 중에서도 주식 시계열 데이터는 대표적인 동적 특성을 가진 시계열이라 할 수 있다. 그것은 주가의 결정에 영향을 미치는 요인이 하나가 아닌 많은 변인에 의해 영향을 받기 때문이다. 그 변인에는 정치, 경제, 사회적인 이슈에 영향을 받는다. 그 밖에 개별적인 산업 동향, 개별 기업체의 경영상황, 전 세계적으로 변화해가는 경향 등 여러 요소가 복합적으로 가미되어 단일 수치로 나타나는 동적 운행을 이해하는 것이 매우 힘든 과제임에 틀림없다. 이처럼 과거의 시계열 데이터에 투영된 여러 가지 요소에 의한 종합적인 현상을 모델링할 수 있다면 이를 현재의 상황에 투영하여 앞으로의 시계열 데이터 운행 방향을 예측하는 것이 가능할 것이다. 본 논문은 이러한 복잡하고 동적인 특성을 갖는 주식 시계열의 과거 데이터를 통해서 미래 변동을 예측하는 하나의 방법을 제안한다.

본 논문에서 제안하는 시계열의 예측 방법은 시계열의 모델링과 유사도 검색을 사용하는 방법이다. 먼저 현재의 시계열을 확률적 모델을 기반으로 하는 은닉 마코프 모델(Hidden Markov Model)[11, 12, 13]로 작성한다. 그리고 이전 시계열 데이터들을 일정 길이로 분할하여 서브 시퀀스를 생성하고 생성된 모델과의 우도(likelihood)를 계산한다. 서브 시퀀스들 중에서 은닉 마코프 모델과 가장 우도가 높은 서브 시퀀스를 유사 시퀀스로 결정한다. 우도가 가장 높다는 것은, 작성된 모델로부터 생성 가능한 시계열들 중에는 유사 시퀀스로 결정된 서브 시퀀스와 같은 흐름을 갖는 시계열이 존재할 가능성이 가장 높다는 것을 의미한다. 그러므로 유사 시퀀스로 결정된 서브 시퀀스의 변화를 현재 시계열에 적용할 수 있을 것이다. 즉, 유사 시퀀스로 선택된 부분 이후의 데이터 흐름을 현재 이후의 값의 변화를 예측하는데 활용하는 것이다. 이 방법은 주식 데이터의 값을 정확히 예측하기보다는 현재 이후의 시계열이 어떤 추세나 흐름을 나타낼 것인가를 예측하기에 적당한 방법이다. 본 논문에서는 1991년부터 2001년 사이의 한국종합주가지수(KOSPI) 데이터를 사용하여 실험하였다.

II. 관련 연구

1. 기존 연구

관측된 시계열이 미래에 어떠한 변동 형태를 나타낼 것인가에 대한 관심은 오랜 동안 계속되어 왔으며 많은 연구자들에 의해서 미래 시계열의 예측을 위한 다양한 방법이 제안되었다. 그리고 이 방법들의 대부분은 과거에 발생한 시계열에서 미래를 추측할 수 있는 정보를 추출하거나 특정 의미를 나타내는 특징을 추출하여 현재에 적용하는 방식을 사용하고 있다. 그러므로 미래의 값이나 상태를 예측하기 위해서는 과거에 관측된 데이터들 중에서 몇 개나 사용할 것인가 하는 문제와 어떤 데이터를 선택할 것인가 하는 문제를 생각해 볼 수 있다.

먼저, K-NN(K-Nearest Neighbors)과 상호정보(Mutual Information)[1, 2] 값을 사용하여 이런 문제를 해결하고 있다. 상호정보 값은 두 변수 사이의 상호 의존도로부터 정보량을 계산하여 얻은 값이다. 현재 지점과 다른 지점들 사이의 상호정보 값을 계산하고 K-NN 방법을 사용하여 상호 의존도가 높은 지점들을 선택한다. 이렇게 선택된 K개의 데이터들은 다음 시기의 예측을 위한 입력 값이 되는 것이다. 이때 입력 개수를 나타내는 K는 1부터 하나씩 변경하면서 테스트를 수행하여 오류 값이 최소가 되는 것으로 결정한다.

다음은 주어진 시계열의 데이터를 변환시킨 후 그 값을 예측에 활용하고 있다. 먼저 주어진 시계열을 증가나 감소 등의 변화에 대한 방향을 나타내는 벡터로 변환하여 그 값을 사용하는 방법[3]이다. 이렇게 변환된 벡터 값을 패턴 모델링에 이용하게 된다. 여기서 패턴 모델링이란 것은 패턴들의 나열을 사용하여 시계열을 설명하는 과정을 말한다. 만약 변환 벡

터의 길이가 n이라면 n+1번째의 증가와 감소를 결정하는 것이다. 먼저 변환 데이터에서 가장 최근의 데이터들 k개를 사용하여 패턴을 만들고 변환 벡터에서 생성된 패턴과 가장 가깝게 일치하는 부분을 패턴을 나타내는 부분을 검색하여 찾는다. 검색된 패턴의 벡터 값에 따라 n+1번째의 증가와 감소를 결정하게 된다.

다음은 시계열을 로그 비[4]로 변환하여 사용하는 방법이다. 시계열은 로그 비 값의 범위에 따라 상태 값을 할당하게 된다. 그러면 주어진 시계열의 상태가 어떻게 변화되었는가를 계산하여 상태전이 확률로 나타낼 수 있게 된다. 이것은 마코프 체인(Markov Chain)의 개념을 기본으로 하여 마코프 모델을 생성하는 것이다. 만약 시계열이 마코프 체인의 특성을 갖는다면, $t-1$은 $t-2$의 영향에 의해 결정되고 t는 $t-1$에 의해 결정된다고 할 수 있다. 그렇다면 현재시간 t의 상태는 시간 $t-1$뿐만 아니라 $t-2$의 상태 값과도 연관이 있을 것이다. 그래서 시간 $t+1$의 값은 시간 t와 $t-1$이 갖는 상태 값과 상태 전이 확률에 따라서 증가와 감소를 결정하게 된다.

이외에도 인공신경망[5, 6, 7, 8]을 사용한 예측 방법들이 많이 제안되고 있다. 먼저 피드포워드(feedforward) 신경망을 사용하는 방법이다. 이 신경망의 입력의 개수는 PCA[7] 등과 같은 통계적인 방법을 사용하여 결정하고 출력의 개수는 1개로 한다. 이때 출력 값은 예측 값의 부호를 나타내게 된다. 입력 데이터는 본래의 데이터를 자동상관계수로 변환한 값을 사용하며, 시계열의 규칙성과 불규칙성에 따라서 출력 값을 할당한다. 이렇게 구성된 신경망에 변환된 데이터를 입력하고 학습을 수행한다. 학습이 종료되었을 때 최종적으로 갖게 되는 입력 값은 주어진 데이터에 대한 예측 값이 된다. 신경망을 하나만 생성하는 것이 아니라 클러스터 별로 작성하여 주어진 시계열이 해당되는 클러스터에 작성된 신경망을 사용하여 시계열의 값을 예측하는데 사용하기도 한다. 그러나 이러한 신경망은 파라미터 값을 어떻게 할당하느냐에 따라서 또는 데이터의 특성에 따라서 그 예측 정확도가 다소 달라질 수 있다. 그리고 다중 계층의 피드포워드 알고리즘을 사용하여 구성한 신경망에 진화 절차를 추가하여 시계열 모델링을 작성하는 FNT(Flexible Neural Tree)[8] 모델을 생성하거나 RBF(Radial Basis Fucntion)[9]를 갖는 신경망을 구성하여 시계열 예측에 사용하는 방법들이 제안되었다.

2. 은닉 마코프 모델

은닉 마코프 모델은 주어진 시계열을 가장 잘 설명하는 모델을 생성하는 방법이다. 모델이 결정되면 그 모델이 갖는 상태의 수와 파라미터의 값은 주어진 시계열에 가장 적합한 것으로 생각한다. 관측 열이 주어졌을 때 이 관측 열을 위한 은닉 마코프 모델을 추정 방법은 여러 가지가 있지만 여기서는 베이지안 정보 기준(Bayesian Information Criterion : BIC) 방법과 Baum-Welch 방법을 사용한다[10].

먼저 베이지안 정보기준에 대해서 알아보자. 데이터 X와 모델 M의 확률이 각각 $P(X)$, $P(M)$이고, 데이터의 한계우도는 $P(X|M)$이라고 할 때, 모델 M의 파라미터 구성 θ가 주어지면 한계 우도는 식(1)과 같다.

$$P(X|M) = \int_{\theta} P(X|\theta, M)\, P(\theta|M) d\theta \quad \cdots\cdots\cdots\cdots (1)$$

데이터의 개수 N이 대규모일 때, 식(1)에 로그를 취하고 라플라스 근사법을 적용하면 식(2)와 같은 식이 된다. 여기서 $\hat{\theta}$ 는 로그를 취한 한계 우도 값을 최대로 하는 파마미터 구성이다.

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

N 데이터 개수 , d 파라미터 개수

식(2)에서 첫 번째 항 $\log P(X|M, \hat{\theta})$은 데이터를 자세히 잘 설명할수록, 즉 파라미터가 많을수록 큰 값을 갖게 되는 우도 값이다. 두 번째 항 $\frac{d}{2} log N$은 모델 복잡도에 대한 페널티(penalty) 항으로 파라미터의 수가 작을수록 한계우도 값을 크게 한다. 그러므로 식(2)는 두 항목이 서로 배타적인 특성을 갖도록 구성되어 있다. 즉, 파라미터 수가 많으면 모델은 데이터를 잘 설명하게 되고 첫 번째 항의 값도 커지지만 계산 복잡도가 높아지며 전체적인 한계우도 값은 그리 커지지 않게 된다. 반대의 경우, 계산 복잡도는 낮아지지만 모델이 데이터를 잘 설명하지 못하게 된다. 그러므로 두 항목이 적절한 값을 가질 때 한계우도 값은 최고값을 갖게 되며, 이때가 모델을 위한 최적의 상태 수가 되는 것이다. 이 상태의 수를 기반으로 반복 실행을 통해서 최적의 파라미터의 값들을 추정하게 된다.

Baum-Welch방법은 파라미터를 추정(Estimation)하는 단계와 생성된 모델과 주어진 시퀀스와의 우도(likelihood)인 $P(O|\lambda)$가 최대(Maximization)가 되도록 하는 단계를 반복하면서 최적의 파라미터 값을 획득하게 된다. $P(O|\lambda)$가 최대가 되도록 하는 파라미터의 추정은 전향(forward)-후향(backward) 절차를 사용한다. 전향 변수와 후향 변수를 사용하여 방출확률과 전이 확률을 계산한다. 새로 구해진 파라미터 값을 사용하여 모델과 시퀀스 사이의 우도를 계산하고

이전의 우도 값 보다 현재의 우도 값이 더 작아질 때까지 파라미터 추정과 우도 계산을 반복한다.

III. 본론

1. 시계열의 검색과 예측 방법

본 논문에서는 은닉 마코프 모델과 우도를 사용하여 유사 시퀀스를 찾고 그 시퀀스를 활용하여 시계열을 예측하는 방법을 제안한다.

1.1 정규화

동적으로 변화하는 시계열, 특히 주식 시계열의 경우에는 업종별, 종목별로 지수의 편차가 크다. 그런데 유사 패턴을 검색하는 것은 지수 값이 일치하는 것이 아니라 값의 흐름이나 추세 즉, 시계열이 나타내는 모양의 유사성을 찾아내는 것이다. 그러므로 비교에 사용되는 데이터는 일정 크기에 분포되도록 변환시키는 전처리작업을 수행하는 것이 필요하다. 본 논문에서는 이러한 전처리를 위해서 아래의 식(3)과 같이 시계열 값의 평균과 표준편차를 사용하여 정규화 한다.

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

v_t : 시간 t에서의 주가, μ_s : 시계열 s의 평균 값

σ_s : 시계열 s의 표준편차,

v_t' 시간 t에서의 전처리된 결과 값

정규화는 질의 시퀀스와 시계열 데이터에서 유사도 검색을 위해 추출된 서브시퀀스에 각각 적용된다.

1.2 로그 비(log ratio) 상태 시계열

주어진 시계열 데이터를 식(4)와 같이 로그 비 형태로 변환하여 그 값을 두 가지 용도로 활용한다.

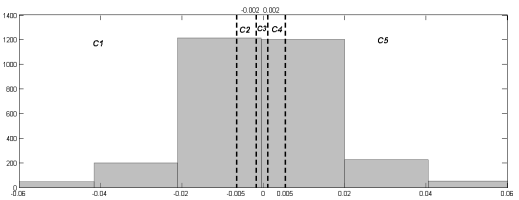


그림 1. 로그 비 상태 그래프

Fig 1. State Graph of log ratio

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

y_t : 시간 t에서의 주가, y_{t-1}에서의 주가

로그 비의 첫 번째 용도는 유사 시퀀스 검색에 사용하는 것이다. 시계열 데이터에서 질의 시퀀스와 유사한 부분을 검색하는 가장 간단한 방법은 시계열을 하나씩 이동시키면서 일정 길이로 데이터를 추출하고 그 데이터를 비교하는 방법이다. 그러나 이것은 시계열의 길이가 매우 길 경우에 시간을 많이 소비하게 된다. 따라서 검색 시간을 줄이기 위해서 시계열을 로그 비 상태 시계열로 변환한 후 그 상태 값을 검색에 사용한다. 우선 유사 시퀀스를 찾으려는 시계열 데이터에서 질의 시퀀스의 로그 비 상태 값이 두 개 이상 일치하는 부분을 찾아내고 그 시점에서부터 일정 길이의 시퀀스를 추출하여 유사도를 계산하게 된다. 만약 검색의 시작에서 로그 비 상태 값의 비교를 실행하지 않고 검색할 경우에는 서로 다른 질의 시퀀스에 대해서 동일한 시퀀스를 유사 시퀀스로 검색하는 경우가 종종 발생한다. 그러나 2개의 로그 비 상태 값이 비교하여 일치하는 것들을 대상으로 검색하는 경우에는 이와 같은 결과가 거의 발생하지 않는다.

로그 비의 두 번째 용도는 최종적인 유사 시퀀스 결정에 사용된다. 하나의 질의 시퀀스에 대한 유사 시퀀스 검색을 여러 번 수행하는 경우에 검색 결과가 동일하지 않고 서로 다른 시퀀스가 결정되는 경우가 발생한다. 이렇게 검색된 시퀀스들은 최종적인 유사 시퀀스의 후보들이 된다. 이 후보 유사 시퀀스들의 로그 비 상태 값을 질의 시퀀스의 로그 비 상태 값과 비교하여 값이 같으면 '1'을 더하고 다르면 '1'을 뺀다. 여러 후보 유사 시퀀스들 중에서 최종적으로 가장 높은 값을 갖는 시퀀스를 유사 시퀀스로 결정하게 된다. 그런데 여러 번의 실행에 서로 다른 시퀀스가 유사 시퀀스로 결정되는 것은 은닉 마코프 모델이 생성될 때마다 파라미터의 초기 값이 임의로 설정되기 때문이다. 은닉 마코프 모델이 생성될 때 주어진 초기 값이 달라지면 모델의 파라미터 값이 조금씩 달라 질수 있고 그 결과 은닉 마코프 모델에 대한 서브 시퀀스의 우도 값도 달라질 수 있게 되는 것이다. 로그 비 상태 시계열 변환방법은 아래와 같은 순서로 이루어진다.

① 시계열을 식(4)를 사용하여 로그 비 r_t로 변환한다.

② 그림 1과 같이 변환된 시계열을 로그 값의 분포에 따라 5개의 그룹으로 나누어 각각에 레이블을 붙인다.

그림 1은 시계열을 로그 비로 변환한 그래프이다. 그림 1에서와 같이 로그 값에 따라 그룹으로 나누어 상태 값을 할당

하였다. 상태 할당은 변화가 거의 없는 경우 즉, 로그 비 값이 0에 가까운 부분은 C3으로 이름을 붙였다. 이 부분은 로그 비의 값이 0인 부분을 선택하려 했지만 로그 비 값이 0이 되는 데이터의 개수가 너무 작아서 레이블 할당의 의미가 없어지기 때문에 범위를 조금 확장하여 적용하였다. C2와 C4는 로그비의 값이 0보다는 크지만 그리 큰 값을 갖지 않는 것들이 포함되도록 선택하였다. C1과 C5는 로그 비 값의 변화의 폭이 큰 것들에 할당한 것이다. 여기서 C2와 C4의 범위를 좁게 한 것은 C1과 C5와의 구분을 명확히 하기 위해서이다. 즉, 상승과 하락의 차이가 많이 나는 부분과 거의 없는 부분을 좀 더 확연히 구분할 수 있도록 하기 위해서이다.

1.3 은닉 마코프 모델과 우도

질의 시퀀스는 정규화 후에 은닉 마코프 모델 λ_Q를 생성한다. 이 모델과 유사한 특성을 나타내는 부분을 시계열 데이터에서 검색하는 것이다. 우선 로그 비 상태 시계열에서 질의 시퀀스와 상태 값이 2개 이상 일치하는 부분에서 부터 $h+\alpha\,(0<\alpha\le\beta,\ 5\le\beta<h)$개의 데이터를 추출한다. 여기서 α와 β값을 이와 같이 설정한 이유는 3장의 1.5절에 기술한다. 그리고 추출된 서브시퀀스 s_i와 작성된 모델 λ_Q와의 우도를 계산한다. 아래 식 5와 같이 추출된 서브 시퀀스들 중에서 가장 우도 L_i가 높은 서브시퀀스 s_i를 주어진 질의 시퀀스의 유사 시퀀스로 결정한다.

$$L_i = \max_i P(s_i|\lambda_Q) \quad \cdots\cdots (5)$$

s_i 서브시퀀스, i : 1, 2, … k

본 논문에서 유사 시퀀스 검색을 위해 사용된 모든 시계열 데이터에 정규화를 적용하는 이유는 다음과 같다. 만약 정규화를 적용하지 않는다면, 후보 시퀀스들 중에서 시퀀스가 갖는 값의 범위와 크기가 질의 시퀀스가 갖는 것과 거의 같은 분포를 나타내는 것만 유사 시퀀스로 결정되어 버린다. 다시 말해, 질의 시퀀스의 값들이 600 ~ 700 사이의 분포이면 결정된 유사 시퀀스가 갖는 값의 분포도 600 ~ 700 사이의 값이다. 그것은 은닉 마코프 모델과 시퀀스 사이의 우도 값은 두 시퀀스의 값의 분포가 유사할수록 커지기 때문이다. 그러나 여기서 찾는 유사 시퀀스는 값의 크기는 다르더라도 모양이 유사한 것을 검색하고자 하는 것이다. 그러므로 시퀀스의 값을 일정 값의 범위 안에 분포하도록 변환하는 정규화를 적용하는 것이 필요하게 된다. 그리고 정규화하지 않았을 경우에는 값의 분포는 다르더라도 모양이 더 유사한 시퀀스를 검색해 내지 못하게 될 것이다.

1.4 제안된 방법

다음 그림 2의 순서에 따라서 마코프 모델을 생성하고 우도를 계산한 후에 유사 시퀀스로 결정하여 예측에 사용한다.

먼저, 그림 2의 단계1과 2에서는 각 시계열을 로그 비 상태 값을 가진 시계열로 변환한다. 그리고 단계 3에서는 유사 시퀀스들을 검색하고 예측 시퀀스 후보들을 추출하는 작업을 수행한다. 우선 질의 시퀀스의 은닉 마코프 모델을 생성한다. 이제부터 시계열 데이터에서 질의 시퀀스의 1, 2 번째의 로그 비 상태 값과 같은 부분을 찾고 상태 값이 같은 부분부터 비교를 위한 시퀀스를 추출한다. 이렇게 추출된 시퀀스와 질의 시퀀스에 대한 은닉 마코프 모델과의 우도를 계산한다. 그리고 우도가 가장 높은 것을 유사 시퀀스 후보로 하고 유사 시퀀스 이후의 값을 예측 값 후보로 결정한다.

하나의 질의 시퀀스에 대해서 m번 반복 실행해서 유사 시퀀스 후보와 예측 값 후보를 m개 생성한다. 그것은 은닉 마코프 모델 작성 시에 주어진 초기화 값에 따라서 모델의 파라미터의 값에 변경이 생길 수 있고 검색된 유사 시퀀스도 달라질 수 있기 때문이다. 그러나 여러 번 반복 실행 결과 유사 시퀀스로 검색된 시퀀스는 대부분 몇 개의 시퀀스로 압축되는 것을 알 수 있다. 단계 4에서는 이렇게 선택된 유사 시퀀스 후보들 중에서 가장 적당한 시퀀스를 구하는 것이다. 최종적인 유사 시퀀스를 결정하기 위해서는 질의 시퀀스와 유사 시퀀스 후보들의 로그 비 상태 값의 합계를 계산한다. 즉 두 시퀀스의 같은 위치의 로그 비 상태 값이 같으면 "1"을 더하고 다르면 "1"을 빼는 것이다. 마지막 단계 5에서는 로그 비 상태의 합계 값을 비교하여 가장 큰 값을 갖는 시퀀스를 질의 시퀀스의 유사 시퀀스로 결정하고 이 시퀀스 이후의 값을 여러 예측 시퀀스 후보들 중에서 최종적인 예측 시퀀스로 결정한다.

1.5 오차분산의 추정치를 사용한 α와 β값 결정

질의 시퀀스 모델과 우도를 계산할 시퀀스의 길이는 $h+\alpha$

이고 유사 시퀀스를 결정한 후에 예측에 사용할 길이는 β이

1. 길이 h인 질의 시퀀스$Q = \{q_1, q_2, \ldots, q_h\}$을 로그 비(Log ratio) 상태 시계열 v_Q로 변환한다.
2. 길이 n인 시계열 데이터 $Y = \{y_1, y_2, \ldots, y_n\}$를 로그 비 상태 시계열 v_Y로 변환한다.
3. for i = 1 to m
   3.1 길이 h인 질의 시퀀스$Q = \{q_1, q_2, \ldots, q_h\}$의 은닉 마코프 모델 λ_Q 작성한다.
   3.2 for j = 1 to n
   if $v_Q(1)$ == $v_S(j)$ and $v_Q(2)$ == $v_S(j+1)$
   y_j에서부터 길이 $l(l = h + \alpha,\ 0 < \alpha \le \beta,\ 5 \le \beta < h)$인 서브시퀀스 s_k와 v_{s_k} 추출한다.
   s_k와 λ_Q의 우도 L_{s_k} 계산한다.
   end if
   end for
   3.3 $L_{s_k} = \max_k P(s_k|\lambda_Q),\ 1 \le k \le n$,
   서브 시퀀스 s_k 선택한다.
   3.4 서브 시퀀스 시작에서부터 $h + \beta$ 데이터 추출한다.
   3.5 β 부분을 예측 값 후보 p_i로 결정한다.
   end for
4. for i = 1 to m
   $qsum_i = 0$
   for j = 1 to h
   if $v_{s_k}(j)$ == $v_Q(j)$
   $qsum_i$의 값을 1 증가 시킨다.
   else $qsum_i$의 값을 1 감소시킨다.
   end for
   end for
5. $qsum_i$중에서 최고 큰 값을 갖는 i번째의 p_i를 예측 시계열로 결정한다.

그림 2. 시퀀스 예측 알고리즘
Fig 2. Sequence prediction algorithm

다. 여기서 h는 질의 시퀀스의 길이이므로 미리 주어질 것이다. 그러나 α의 값은 β보다는 적고 0보다는 큰 수중에서 가장 작은 오차 분산을 나타내는 값을 사용한다. 모든 오차 항은 서로 독립이며, 식(6)과 같은 오차항의 정규분포를 나타낸다. 여기에서 오차 분산의 추정치는 식(7)과 같다.

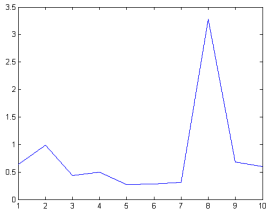
$$\epsilon \sim (\mu, \sigma^2) \rightarrow \epsilon \sim (0,1) \quad \cdots\cdots (6)$$

ϵ : 오차항, μ : 오차항의 평균, σ^2 : 오차항의 분산

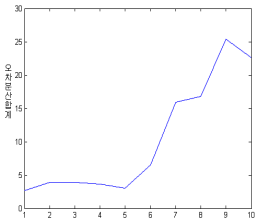
$$\sigma^2 = \frac{SSE}{n-2},\ SSE = \sum (y_i - \hat{y_i})^2 \quad \cdots\cdots (7)$$

y_i : 관찰 값, $\hat{y_i}$: 관찰 값의 평균, n: 관찰 값의 개수

β값은 예측 시퀀스의 길이 값이므로 시퀀스의 추세를 알 수 있어야 한다. 그래서 β값은 예측 시퀀스가 5일 이동 평균선이나 10일 이동 평균선을 작성할 수 있도록 5, 10, 15, 20... 중에서 선택한다. 그러나 이때 β값은 질의 시퀀스의 길이보다는 작은 값을 사용할 것이다. 그림 3의 a는 β값이 10일 때 α의 값을 1에서 10까지 변환시키면서 실제 값과 예측 값 사이의 오차 분산 값을 계산하여 그래프로 나타낸 것이다. 이 그래프는 α값이 5일 때 최소의 오차 분산을 나타내고 있다. 7개의 테스트 시퀀스에 대한 예측 시퀀스를 검색하고 7개의 실제 값과 예측 값의 오차를 모두 더하여 오차 분산을 계산하였다. 그림 3의 b는 같은 테스트 데이터를 10번 반복하여 α값의 변화에 따른 오차분산의 합계를 나타낸 것이다. 그림 3의 b에서 보면 α값이 1과 5일 때 오차 분산의 합계가 작게 나타나는 것을 알 수 있다. 본 논문에서는 α값을 5로 사용하여 실험하였다.



a. β 값이 10일 때, α값의 변화에 따른 오차분산 (단위 : 10-27)



b. α 값의 변화에 따른 오차분산합계(단위 : 10-27)

그림 3. β와 α값에 의한 오차분산
Fig 3. Error Variance by β and α

2. 실험 결과

본 논문에서는 1991년부터 2001년 사이의 코스피(KOSPI) 종합주가지수 3000개를 실험 데이터로 사용했다. 우선 1991년부터 3000개를 추출하여 검색을 위한 데이터베이스 시계열로 사용하고 3000개 이후에 500개의 데이터를 추출하여 질의 시퀀스를 위한 시계열로 사용하여 테스트하였다.

먼저, 시계열 데이터와 질의 시퀀스는 검색과 예측 값

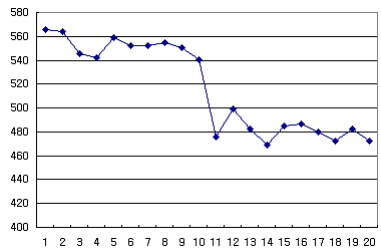
의 선택을 위해서 로그 비 상태 시계열로 변환하고 그림 1과 같이 로그 비 값의 크기에 따라 상태 값을 할당한다. 그리고 테스트에서 α와 β값은 각각 5와 10으로 설정하고 질의 시퀀스의 길이는 20으로 설정하였다. 즉, 우도 계산을 위해서는 질의 시퀀스보다 5개의 데이터를 더 사용하고 10개의 값을 예측하는 것이다. 테스트를 위해 사용된 500개 데이터 중에서 20개씩을 선택하여 질의 시퀀스로 사용하고 20개 이후에 나타난 10개의 데이터를 예측을 위한 비교 값으로 사용하였다.

우선, 질의 시퀀스를 은닉 마코프 모델로 만들었다. 그리고 시계열 데이터를 순차적으로 이동하면서 주식 시계열 데이터와 질의 시퀀스의 로그 비 상태 값이 2개 이상 같은 부분부터 25개의 데이터를 선택하여 서브시퀀스로 결정하였다. 은닉 마코프 모델과 서브시퀀스와 우도를 계산하고 가장 높은 우도 값을 나타내는 서브시퀀스를 질의 시퀀스와 가장 유사한 시퀀스로 결정한다. 유사 시퀀스로 결정된 시퀀스 부분의 끝에서 앞으로 5개 뒤로 5개를 선택하여 β값인 10개의 예측 값을 결정한다. 그런데 은닉 마코프 모델의 경우 초기 확률 값의 설정에 따라서 모델의 파라미터 값이 조금씩 달라질 수 있다. 그러므로 유사 시퀀스의 검색 결과가 다르게 나올 수 있다. 이러한 문제점을 해결하기 위해서 유사 시퀀스 검색을 10회 이상 실시하고 로그 비 상태 값을 사용한다. 질의 시퀀스와 후보 유사 시퀀스의 로그 비 상태 값을 비교하여 같은 값이면 '1'을 더하고 다른 값이면 '1'을 뺀 값을 계산한다. 그리고 후보 유사 시퀀스들 중에서 로그 비 상태의 합계 값이 가장 큰 시퀀스를 최종적인 유사 시퀀스로 결정하였다. 만약 로그 비 상태 값이 같은 경우에는 우도 값이 가장 큰 것을 선택하였다. 표 1에 첫 번째 질의 시퀀스에 대한 테스트 결과 검색된 유사 시퀀스들의 실험 횟수별 로그 비 상태 값과 우도 값을 나타내었다.

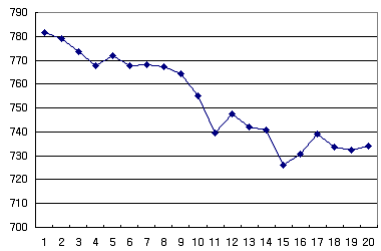
표 1. 실험 횟수별 로그 비 상태 값과 우도 값
Table 1. Log Ratio State Value and Likelihood

횟수	로그 비 상태 값	우도 값
1	0	-63.146
2	0	-43.426
3	4	-59.411
4	0	-49.844
5	4	-59.189
6	0	-45.769
7	0	-48.072
8	0	-69.994
9	4	-60.731
10	0	-51.715

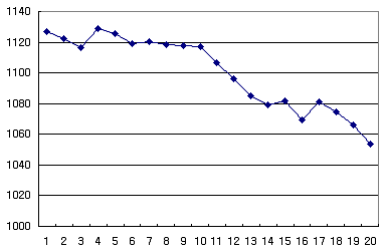
표 1에서 열 번의 테스트 중에서 3, 5, 9번째는 동일한 시퀀스가 유사 시퀀스로 검색되었다. 이것은 그림 4의 b에 그래프로 나타내었다. 그리고 1, 2, 4, 5, 7, 9, 10 번째 실행 결과로 검색된 시퀀스가 동일하다. 이것은 그림 4의 c에 그래프로 나타내었다. 그림 3을 보면 a에 있는 질의 시퀀스와 b의 시퀀스가 더 유사하게 보인다. 그리고 b 시퀀스의 로그 비 상태 값도 c 시퀀스보다 큰 것을 알 수 있다.



a. 질의 시퀀스



b. 3,5,9 회에서 검색된 시퀀스



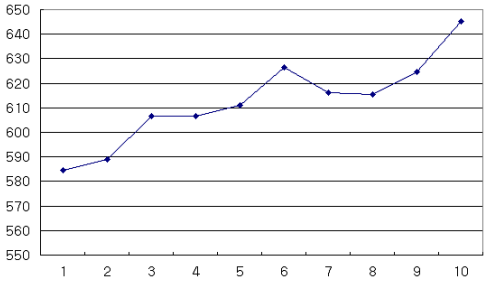
c. 1,2,4,6,7,8,10 회에서 검색된 시퀀스

그림 4. 질의 시퀀스와 검색된 유사 시퀀스
Fig 4. Query Sequence and Similarity Sequence

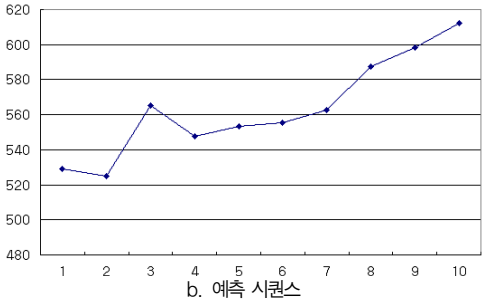
그림 5에는 실제 시퀀스와 제안된 방법으로 예측한 예측 시퀀스의 그래프를 나타내었다. 지면 관계상 실험 결과

중에서 유사한 흐름을 보이는 것과 정 반대의 흐름을 보이는 것을 몇 개만 대표적으로 나타내었다. 그림 5에서 a는 실제 시퀀스이고 b는 제안된 방법을 사용하여 예측한 예측 시퀀스이다. 그림 5의 (1)은 4번 째 테스트한 시퀀스 이후의 실제 값(a)와

(1) 시퀀스 4번 이후의 예측

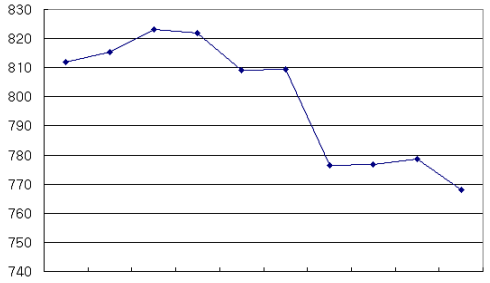


a. 실제 시퀀스

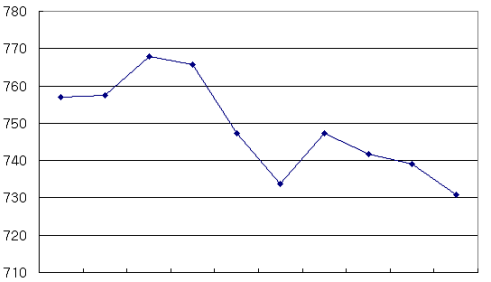


b. 예측 시퀀스

(2) 시퀀스 18번 이후의 예측

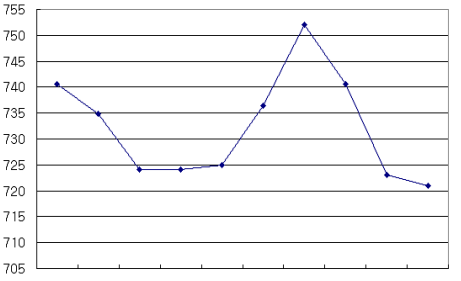


a. 실제 시퀀스

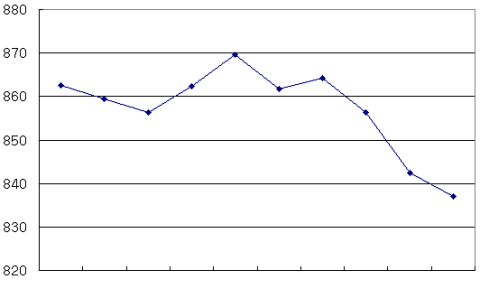


b. 예측 시퀀스

(3) 시퀀스 23번 이후의 예측

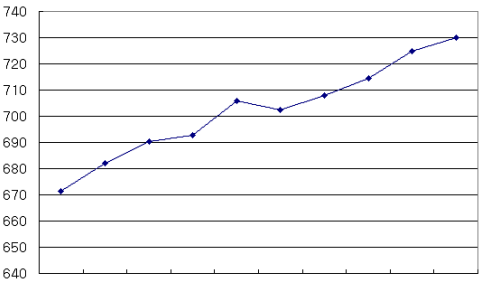


a. 실제 시퀀스

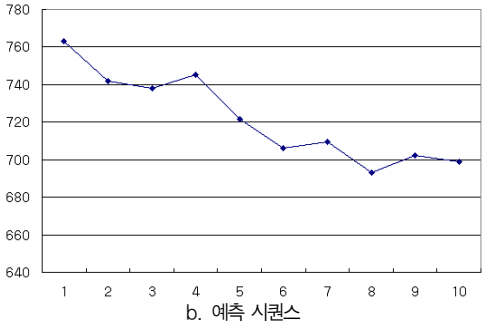


b. 예측 시퀀스

(4) 시퀀스 29번 이후의 예측



a. 실제 시퀀스



b. 예측 시퀀스

그림 5. 시계열의 실제 값과 예측 값
Fig 5. Real value and Prediction value of Time series

유사 시퀀스를 통해서 추출해 낸 예측 시퀀스의 값(b)를 그래프로 표시한 것이고 (2)는 시퀀스 번호 18 번째에 대한 실제 시퀀스와 예측 시퀀스를 나타낸 것이다. 그림 5에서 보듯이 (1)은 실제와 예측의 시퀀스들이 상승의 추세를 나타내는 것을 확인할 수 있고 (2)도 역시 둘 다 하락의 추세를 나타내는 것을 볼 수 있다. 그러나 29번 시퀀스 이후의 실제와 예측을 나타낸 (3)의 경우에는 정 반대의 흐름을 보이고 있다. 그림 5의 (4)를 보면 시각적으로 봤을 때는 그 유사 여부를 확인하기 어렵다. 그러나 실제와 예측 시퀀스의 각 시기별 값의 변화를 보면 유사성을 찾을 수 있다. (4)의 a와 b 시퀀스에서 5번째와 6번째 사이의 값의 변화가 실제 값의 경우는 앞의 값에 비해 상승하지만 예측 값의 경우는 하락한다. 그 외에 3, 4번째의 변화가 다르게 나타난다. 그러나 나머지 1, 2번째, 6, 7번째 등에서 값의 상승과 하락의 변화는 a 시퀀스와 b 시퀀스가 같은 형태로 변화하는 것을 알 수 있다. 그리고 값의 차이는 있지만 모양의 유사성을 찾을 수 있다. 그러므로 실제와 예측 시퀀스가 유사하다고 할 수 있을 것이다.

다음의 표 2에는 그림 5에 나타낸 실제 시퀀스와 예측 시퀀스 사이의 평균제곱오차를 계산한 값을 나타낸 것이다. 예측 값과 실제 값은 모두 정규화를 적용한 후 계산하였다.

표 2. 평균제곱오차
Table2. Mean Square Error

시퀀스 번호	평균제곱오차(MSE)
4	0.238015
18	0.444669
23	0.950340
29	3.443078

표 2의 평균제곱오차 값을 보면, 그림 5에서 정 반대의 추세를 나타내고 있는 29번 시퀀스 이후의 예측 값에 대한 평균제곱오차의 값이 크게 나타나는 것을 볼 수 있다. 그리고 흐름이 유사하게 나타났던 시퀀스 4번과 18번 이후의 예측에 대한 평균제곱오차 값은 크지 않은 것을 알 수 있다. 또한 뚜렷하게 상승이나 하락의 추세를 나타내지 않았던 시퀀스 23번 이후의 예측 값에 대한 평균제곱오차 값도 1을 넘지 않는다. 이것은 추세를 알 수는 없지만 상승과 하락의 시점이 유사하게 변하므로 시퀀스 23번의 경우 실제와 예측 시퀀스가 유사 흐름을 나타낸다고 할 수 있을 것이다. 이와 같이 추세를 명확히 알 수 있는 시퀀스들과 추세를 결정할 수는 없지만 변화 시점이 60% 이상 동일한 시퀀스들을 모두 포함하여 정확도를 계산하였다. 이렇게 하여 계산된 정확도는 아래 표 3에 나타내었다.

표 3. 예측 시퀀스의 정확도
Table 3. Accuracy of Prediction Sequence

테스트 시퀀스 개수	유사 예측 시퀀스 개수	정확도
35	22	62.8%

IV. 결 론

본 논문은 확률적 모델을 기반으로 하는 은닉 마코프 모델과 우도를 사용하여 과거의 데이터 중에서 현재 상황과 가장 유사한 유사 시퀀스를 찾아내고 유사 시퀀스 이후의 값을 사용하여 주식 시계열의 값을 예측하는 방법을 제안했다. 그것은 유사 시퀀스 이후에 나타날 데이터의 운행 방향이 현재 시퀀스의 은닉 마코프 모델로부터 생성될 데이터의 운행 방향과 유사할 것으로 생각한 것이다. 실험 결과 예측된 데이터들은 앞으로의 주가 흐름이 어떻게 변화할 것인가에 대한 추세를 실제와 유사하게 예측하는 것을 볼 수 있었다. 정확한 값을 예측하는 것이 아닌 그 흐름을 예측하는데 있어서는 본 논문의 제안 방법이 유효할 수 있을 것으로 생각된다. 그러나 일부분의 예측 열은 실제 값과 정 반대의 추세를 나타내는 것을 볼 수 있다. 이러한 정 반대 방향을 나타내는 예측열의 경우에는 최종적으로 결정된 유사 시퀀스가 아닌 후보 유사 시퀀스로 검색된 것들 중에서 실제 추세와 유사한 흐름을 나타내는 것이 있었다. 이것은 유사 시퀀스 결정 방법에 또 다른 방법을 추가할 필요성을 느끼게 하는 부분이다. 앞으로 좀 더 연구하여 정확성을 더 높인 유사 시퀀스 결정 방법을 찾아낸다면 지금보다 명확한 예측 시퀀스를 찾아낼 수 있을 것으로 생각된다.

참고문헌

〔1〕 A. Sorjamaa, et al., "Methodology for long-term prediction of time series," Neurocomputing, Vol. 70, No. 16-18, pp.2861-2869, Oct, 2007.

〔2〕 A. Sorjamaa, J. Hao, A. lendasse, "Mutual Information and k-Nearest Neighbors Approximator for Time Series Prediction", International Conference on Artificial Neural Networks, Vol. 3697, pp.553-558, 2005 Sep.

〔3〕 S. Singh, "Pattern Modelling in time-series forecasting," Cybernetics and Systems-An International Journal, Vol. 31, No. 1, pp.49-66, 2000.

〔4〕 C. P. Papageorgiou, "High Frequency Time Series Analysis and Prediction using Markov Models," inProdeedings of the conference on Computational Intlligence for Financial, pp.182-185, Mar. 1997.

〔5〕 N. G. Pavlidis, D. K. Tasoulis, M. N. Vrahatis, "Time Series Forecasting Methodology for Multiple-Step-Ahead Prediction," The IASTED International Conference on Computational Intelligence, pp.456-461, 2005.

〔6〕 C. Chatfield, "Time Series Forecasting with Neural Networks," Neural Networks for signal Processing VIII, pp.419-427, 31 Aug -2 Sept. 1998

〔7〕 P. Cortez, M. Rocha, J. Machado, J. Neves, "A Neural Network Based Time Series Forecasting System," , IEEE International Conference on Neural Networks, Proceedings Vol. 5, pp.2689-2693, Nov. 1995

〔8〕 Y. Chen, B. Yang, J. Dong, A. Abraham, "Time-series forecasting using flexible neural tree model", Information Sciences : an International Journal, Vol. 174, No. 3-4, pp.219-235, Aug, 2005.

〔9〕 D. Zhang, X. Ning, X. Liu, Y.Han, "NonLinear Time Series Forecasting with Dynamic RBF Neural Network," Proceeding of the 7th World COngress on Intelligent Control And Automation, pp.6988-6993, Chongqing, China, Jun. 2008.

〔10〕 J. Hamaker and J. Zhao, "Bayesian Information criterion for automatic model selection," Technical Report, Mississippi State University, May 1999.

〔11〕 M. Azzouzi, I. T. Nabney, "Analysing time series structure with Hidden Markov Models," in Proceeding of Neural Network for Signal Processing VIII, pp.402-408, 31 Aug - 2 Sept. 1998

〔12〕 A. Panuccio, M. Bicego and V. Murino, "A Hidden Markov Model-based approach to sequential data clustering", In Caelli, T., Amin, A., Duin, R., Kamel, M., de Ridder, D., eds.: Structural, Syntactic and Statistical Pattern Recognition. LNCS 2396, Springer pp.734 -742, 2002.

〔13〕 C. Bahlmann, H. Burkhardt, "Measuring Hmm Similarity with the Bayes Probability of Error and its Application to Online Handwriting Recognition," In Proc. of the 6th ICDAR, pp.406-411, 2001.

저 자 소 개

조 영 희

1995년 2월 단국대학교 전자계산학과 (이학사)

2000년 2월 단국대학교 전자계산학과 (이학석사)

2005년 8월 단국대학교 전자계산학과(박사과정 수료)

〈관심분야〉 데이터마이닝, 기계학습, 에이전트

전 진 호

1998년 명지대학교 경영정보학과 (경영학석사)

2007년 단국대학교 전자계산학과 (이학박사)

2003년 3월 ~ 현재 관동대학교 경영정보학부 겸임교수

〈관심분야〉 기계학습, 데이터마이닝

이 계 성

1980년 서강대학교 전자공학과(학사)

1982년 한국과학기술원 전자계산학과(석사)

1994년 Vanderbilt University 전자계산학과(공학박사)

1994년 ~ 1995년 대구대학교 전산정보학과 전임강사

1996년 ~ 현재 단국대학교 컴퓨터과학과 교수