은닉 마코프 모델을 이용한 시계열 데이터의 의미기반 패턴 매칭

Conceptual Pattern Matching of Time Series Data using Hidden Markov Model

조영희*, 전진호*, 이계성**
단국대학교 전자계산학과*, 단국대학교 컴퓨터과학부**

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

요약

시계열 데이터에서 패턴을 찾고 검색하는 문제는 여러 분야에서 오랫동안 관심을 가지고 연구되어 왔다. 본 논문은 시간의 흐름에 따라 값의 변화를 나타내는 시계열 형태의 주식 데이터에 적용할 수 있는 새로운 패턴 매칭 방법을 제안한다. 우선, 의미를 기반으로 패턴을 정의하고 정의된 패턴에 일치하는 데이터들을 추출하여 학습모델을 작성한다. 그리고 새로운 질의 시퀀스가 어떤 종류의 패턴과 일치하는가는 각 학습 모델과의 유사도를 측정하여 결정하게 된다. 학습 모델은 시계열을 잘 설명하는 것으로 알려진 은닉 마코프 모델을 사용하여 작성하였다. 실험 결과 은닉 마코프 모델의 특성을 사용하여 생성된 각 학습 모델은 주어진 의미를 잘 나타내는 패턴을 생성하였으며, 새로운 시퀀스가 주어졌을 때 일치하는 패턴에 따라서 시퀀스가 가진 의미를 파악할 수 있었다.

■ 중심어 : | 패턴매칭 | 은닉마코프모델 | 시계열 |

Abstract

Pattern matching and pattern searching in time series data have been active issues in a number of disciplines. This paper suggests a novel pattern matching technology which can be used in the field of stock market analysis as well as in forecasting stock market trend. First, we define conceptual patterns, and extract data forming each pattern from given time series, and then generate learning model using Hidden Markov Model. The results show that the context-based pattern matching makes the matching more accountable and the method would be effectively used in real world applications. This is because the pattern for new data sequence carries not only the matching itself but also a given context in which the data implies.

■ keyword : | Pattern Matching | Hidden Markov Model | Time Series |

I. 서 론

시계열은 의학, 과학, 경제 등의 다양한 분야에서 발생한다. 예를 들어 주간 또는 일간 별로 작성된 주가 데이터, 반도체 공정 중에 발생하는 플라즈마의 변화, 온라인 모니터링 시스템에서 실시간으로 저장되는 모니터링 정보 등이 있다. 이렇게 발생하는 수많은 데이터들에서 의미 있는 정보를 찾아내는 것은 중요한 문제이며, 이러한 정보를 구하는 방법으로 일정한 형태로 변화하거나 규칙적으로 변화하는 패턴을 찾는 것은 아주

접수번호 : #080123-003
접수일자 : 2008년 01월 23일
심사완료일 : 2008년 05월 19일
교신저자 : 조영희, e-mail : zeroch@dankook.ac.kr

유용한 방법이라 할 수 있다. 패턴을 찾는 방법으로 시계열에서 시각적으로 보이는 모양을 기반으로 해서 유사 모양을 찾는 패턴 매칭 방법이 많이 연구되었다. 그리고 유사 모양을 찾기 위해서 어떠한 유사도 척도를 사용할 것인가 또한 중요 관심사이다.

본 논문은 여러 분야의 시계열 데이터 중에서 주식데이터에서 발생하는 패턴을 찾는 문제를 다룬다. 주가의 변동은 하나의 변수가 아닌 수많은 변수의 영향을 받아서 결정되기 때문에 주식 시계열의 변화를 예상하기는 어렵다. 그러나 과거의 주식 데이터에 나타난 패턴이 현재에도 유사하게 반복적으로 발생할 것이라는 가정하에 유사 패턴을 찾으려는 연구가 진행되어 왔다. 그동안의 다른 연구들에서는 모양을 기반으로 하여 유사 패턴을 찾는 문제를 논의했다. 그러나 본 논문은 주식 데이터가 가지고 있는 의미를 기반으로 하는 패턴 매칭 방법을 제안한다. 이 방법은 시계열 데이터에서 의미를 기반으로 하여 패턴을 정의하고 패턴에 일치하는 데이터를 추출하여 은닉 마코프 모델 형태의 학습 모델을 생성한다. 생성된 은닉 마코프 학습 모델과 새로운 질의 시퀀스와의 유사도를 측정하여 패턴을 결정하는 방법이다.

II. 관련 연구

1. 기존 연구

일반적으로 패턴 정의는 시계열이 나타내는 모양을 보고 그 모양이 갖는 특징에 따라 패턴으로 표현하는 방법[1][3][4]을 사용한다. 그러나 시계열을 일정간격으로 나누어서 세그먼트로 만들고 각 세그먼트에 명칭적 표현[5]을 적용하여 나타내는 방법을 사용하기도 한다. [2][9]의 경우에는 패턴을 확률적 모델로 표현하여 사용한다. 그리고 확률적 모델의 경우 시퀀스에 대한 은닉 마코프 모델을 생성[7][10]하여 사용하기도 한다. 또한 패턴을 k-상태 세그먼트를 갖는 semi-Hidden Markov Model[[2] 방법을 사용하여 확률적 모델을 표현하기도 한다. [8]에서는 현재 지점과 다음 지점 사이의 차이를 코드화 하고 그것들을 하나의 세그먼트로 생각한다. 이렇게 생성된 세그먼트들을 두 개 이상 사용하여 코드로 표현된 패턴으로 정의 한다.

주어진 시계열에 대해 정의된 패턴을 새로운 시계열에서 찾아내는 시계열 검색의 문제는 시퀀스들 사이의 유사도 측정 방법을 고려하게 된다. 시퀀스들 사이의 유사도 측정에 대해서는 많은 연구자들에 의해 다양한 방법들이 제안되었다. 우선, 시퀀스의 유사도를 측정하는 방법으로 가장 일반적인 방법은 두 시퀀스 사이의 유클리드 거리(Euclidean distance)를 계산하여 사용하는 것이다. 단순한 방법이지만 비교적 정확한 유사도 측정 효과를 나타낸다. 그러나 이것은 시각적인 모양은 유사하지만 시간의 차이가 발생하는 경우에는 다른 시퀀스로 인식되어 유사 시퀀스로 검색해 내지 못하는 경우가 많이 발생한다. 이러한 문제점을 해결하고자 Dynamic Time Warping 방법[9]을 적용하여 시간의 차에 의한 문제를 어느 정보 보완하기도 한다. 그리고 시계열에 대한 전처리 작업으로 시계열을 구간 선형 표현법(linear piecewise representation)으로 나타낸 세그먼트로 분할 후 그 세그먼트들을 사용하여 비교하는 방법들을 제안하고 있다. 우선 [3]에서는 세그먼트 된 시퀀스의 길이에 대한 편차를 계산하여 유사도의 측도로 사용하기 한다. 그리고 [1]에서는 상대적 위치를 나타내는 데이터 포인트로 표현된 세그먼트에서 상대적 위치의 값의 일치여부를 사용하여 유사성을 결정한다. [4]에서는 확률적 거리 모델을 유사도 측정 방법으로 제안하고 있다. 또한 확률적 모델로 생성된 패턴의 경우에는 패턴 모델과 질의 시퀀스 사이의 우도(likelihood) 값을 사용하여 유사도를 측정하기도 한다. 이때 세그먼트들 사이의 거리는 시간 변형에 대한 유연성을 허용하도록 하고 있다. [7][10]에서는 시퀀스에 대한 은닉 마코프 모델과 시퀀스들 사이의 로그 우도(log-likelihood) 값을 계산하여 유사 정도를 확인하는 방법을 사용한다. 코드로 표현된 패턴을 사용하는 경우에는 정의된 패턴과 코드 값을 비교하여 패턴을 찾아내는 방법을 사용하고 있다[8].

2. 은닉마코프 모델(Hidden Markov Model)

은닉 마코프 모델은 주어진 시계열을 가장 잘 설명하

는 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 모델의 파라미터 추정 방법은 여러 가지가 있지만 여기서는 Bauam-Welch 방법을 사용하고 상태의 수를 추정하기 위해서는 베이지안 정보 기준(Bayesian Information Criterion : BIC) 방법을 사용한다[6].

먼저 주어진 관측 열에 대한 최적의 모델 크기, 즉 최적의 상태 수를 결정 한다. 상태의 수를 결정하는 방법은 여러 가지가 연구되었지만 여기서는 베이즈 이론을 기반으로 하는 베이지안 정보 기준(Bayesian Information Criterion, BIC) 방법을 사용한다. 이 방법은 정확도는 약간 떨어지지만 계산 복잡도를 많이 줄여주기 때문에 효율성을 높인 근사기법이다.

데이터 X와 모델 M의 확률이 각각 $P(M)$, $P(X)$라 하고 데이터의 한계우도 $P(X|M)$하자. 이때, 모든 모델의 사전 확률이 같다면 베이즈 정리에 의해 $P(M|X) \propto P(M)(X|M)$로 나타낼 수 있다. 즉, 어떤 모델의 사후 확률 $P(M|X)$는 그 데이터의 한계 우도에 비례($P(M|X) \propto (X|M)$)하게 된다. 여기에 모델 M의 파라미터 구성 θ가 주어지면 한계 우도는 아래와 같은 식이 된다.

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

이제, 위의 (1)식에 로그를 취해보자. 그리고 데이터의 개수 N이 대규모일 때, 로그를 취한 한계 우도 값을 최대로 하는 파마미터 구성을 $\hat{\theta}$이라 하고 (1)식에 라플라스 근사법을 적용하여 보면 아래와 같은 식이 된다.

$$logP(X|M) = logP(X|M, \hat{\theta}) - \frac{d}{2}logN$$

N: 데이터 개수 , d : 파라미터 개수 (2)

위의 식에서 첫 번째 항 $logP(X|M, \hat{\theta})$은 데이터를 자세히 잘 설명 할수록, 즉 파라미터가 많을수록 큰 값을 갖게 되는 우도 값이다. 두 번째 항 $\frac{d}{2}logN$은 모델 복잡도에 대한 페널티 항으로 파라미터의 수가 작을수록 한계우도 값을 크게 한다. 그러므로 (2)번 식은 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. 우향 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. 의미기반 학습 모델을 이용한 패턴 매칭

본 논문에서는 시계열에서 특징에 따라 추출된 데이터가 갖는 의미를 기반으로 하는 패턴 매칭 방법을 제안한다.

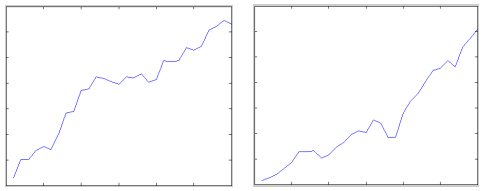
1. 제안된 방법

패턴 매칭을 위한 실험 방법은 우선, 주어진 주식 시계열에서 의미 기반으로 패턴을 정의하고 은닉 마코프 형태의 학습 모델을 생성하여 패턴별 학습 모델을 작성한다. 그리고 생성된 학습 모델과 새로운 질의 시퀀스

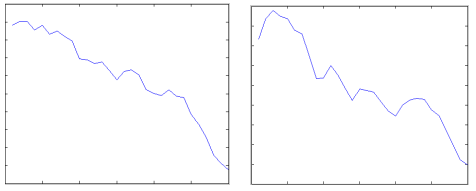
사이의 유사도를 측정하여 새로운 질의 시퀀스가 어떤 패턴과 유사한 특성을 나타내는 가를 결정한다. 이때 유사도 측정 방법은 학습 모델과 시퀀스 사이의 우도(likelihood)를 계산하여 사용한다. 그리고 학습 데이터와 테스트 데이터들의 값의 편차가 일정 범위 안에 분포되도록 하기 위해서 각 학습 데이터와 테스트 데이터에 전처리 작업을 수행한다.

시계열에서 패턴의 정의는 시계열이 나타내는 시각적 특성이 얼마나 유사한가를 생각하여 반복적으로 발생하는 모양을 하나의 패턴으로 정의하여 사용한다. 그런데 주식 데이터에서 가장 관심을 갖는 사항은 주식의 시세가 상승하느냐 또는 하락 하느냐의 문제이다. 그리고 현재의 상승 추세가 일정 기간 중에서 최고로 오른 상태인가 또는 현재의 하락 추세가 일정 기간 중에서 최저로 내려간 상태인가에 관심을 갖게 된다. 본 논문에서는 이러한 주식이 나타내는 추세에 상승, 하락, 최고, 최저의 의미를 부여하고 이러한 의미를 기반으로 해서 각각을 패턴으로 정의한다. 이렇게 정의된 패턴들을 의미기반 패턴 이라고 한다. 여기에 가격 변동의 폭이 거의 없는 보합 상황을 추가하여 상승선, 하락선, 최고점, 최저점, 보합선의 5가지로 패턴으로 정의한다.

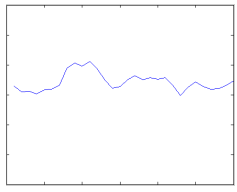
아래 [그림 1]은 각 패턴의 예를 나타낸 것이다. [그림 1]에서 보면 a 와 b가 시각적으로 유사하고 c와 d가 거의 유사한 형태를 나타내고 e는 다른 그래프와 시각적으로도 완전히 다른 형태를 나타낸다. 여기서 그림 a의 상승선은 계속적으로 주가가 상승하는 부분을 그래프로 나타낸 것이고 b의 최고점은 일정 기간에서 최고의 주가를 나타낸 부분을 t라 했을 때 그 시점부터 이전 데이터를 그래프로 그린 것이다. 즉, $t, t-1, t-2, ..., t-n$ 까지 데이터를 그래프로 그린 것이다. c에 나타난 하락선의 경우는 계속적으로 주가가 하락하는 부분을 그래프로 그린 것이고 d의 최저점은 최고점과 유사하게 결정된다. 즉, 일정 기간에서 최저의 주가를 나타낸 부분을 t라 했을 때, $t, t-1, t-2, ..., t-n$ 의 데이터를 그래프로 그린 것이다.



a. 상승선 b. 최고점



c. 하락선 d. 최저점



e. 보합선

그림 1. 패턴의 예

[그림 1]에서 보듯이 그래프들이 나타내는 모양은 유사하지만 각 시퀀스가 내포하고 있는 의미적인 측면에서 보면 서로 다른 의미를 나타내므로 각각 다른 패턴으로 정의하여 사용하게 된다.

주식 데이터에서 정의된 패턴을 위한 학습 모델을 생성한다. 그런데 학습 모델 생성을 위해 사용되는 주식 시계열 데이터는 각 업종별, 종목별로 주가 지수가 커다란 차이를 나타낸다. 이런 데이터를 그대로 사용하게 되면 같은 패턴을 위한 데이터도 편차가 아주 커져서 일관된 값을 갖는 패턴 모델을 생성하기 어렵게 된다. 그러므로 값의 편차가 일정 범위 안의 값으로 나타나도록 전 처리 작업을 수행한다. 여기서 사용한 전처리 방법은 정규화 과정으로 다음과 같다. i) 주어진 주가 데이터의 평균과 표준 편차를 구한다. ii) 현재 주가에서 평균을 뺀다. iii) ii)의 값을 표준 편차로 나눈다. 이것을 식으로 나타내면 아래와 같다.

$$s'_t = \frac{s_t - \mu_s}{\sigma_s}$$

s_t : 시간 t에서의 주가,

μ_s : 시계열 s의 평균 값,

σ_s : 시계열 s의 표준편차,

$s_i^{'}$: 시간 t에서의 전 처리 된 결과 값

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

이제 고른 분포를 나타내는 주식 시계열 데이터에서 각 패턴의 의미에 맞는 데이터를 추출한다. 추출된 데이터의 길이는 m으로 한다. 우선 최고점 데이터는 주어진 주식 시계열에서 가장 큰 값을 갖는 지점에서부터 앞쪽으로 m 개를 사용한다. 즉, 최고지점 시간 t에서 시작해서 $t-1, t-2, \ldots, t-m$ 까지 데이터를 추출한다. 최저점의 경우도 주어진 시계열에서 가장 작은 값을 갖는 최저지점에서 시작해서 이전 값들을 m 개 추출하고 상승선과 하락선을 위한 데이터도 역시 각각 주가가 계속적으로 상승하는 부분과 계속적으로 하락하는 부분을 m개 선택한다. 보합선은 계속적으로 같은 값을 유지하는 부분이 거의 없기 때문에 주가의 변화가 거의 크지 않고 변동의 크기가 작은 부분을 선택하여 사용한다. 이때 주어진 주식 시계열이 n개라면, 최고점과 최저점의 데이터는 최대 n개가 추출되고 상승선, 하락선, 보합선을 위한 데이터는 n개 이상이거나 그보다 작은 수가 추출될 수도 있다.

각 패턴에 맞게 추출된 데이터를 사용하여 패턴별 학습 모델을 생성한다. 학습 모델은 은닉 마코프 모델을 사용한다. 은닉 마코프 모델은 시계열에서 최적의 상태 수를 찾고 그 상태 수에서 최적의 파라미터를 구해서 주어진 데이터들을 가장 잘 설명하는 모델을 생성하게 된다. 생성된 모델들은 상승선모델(Increase Model), 하락선모델(Decrease Model), 최고점모델(Highest Model), 최저점모델(Lowest Model), 보합선모델(Steady Model)이며, 각각의 모델들은 영문 첫 글자를 따서 λ_I, λ_D, λ_H, λ_L, λ_S 로 표기한다.

질의 시퀀스 Q가 주어졌을 때, 이 시퀀스가 어떤 패턴과 일치하는 가를 결정하기 위한 유사도 측정 방법은 각 패턴들의 학습 모델과 질의 시퀀스 사이의 우도(likelihood) 값을 사용한다. 질의 시퀀스 Q와 각 패턴 모델 별 우도 값은 5개의 모델에 대한 우도 값 중 가장 큰 값의 모델을 Q에 대한 패턴으로 결정한다.

$$\lambda_k = \max P(Q|\lambda_j),\ j \in (I, D, H, L, S)$$

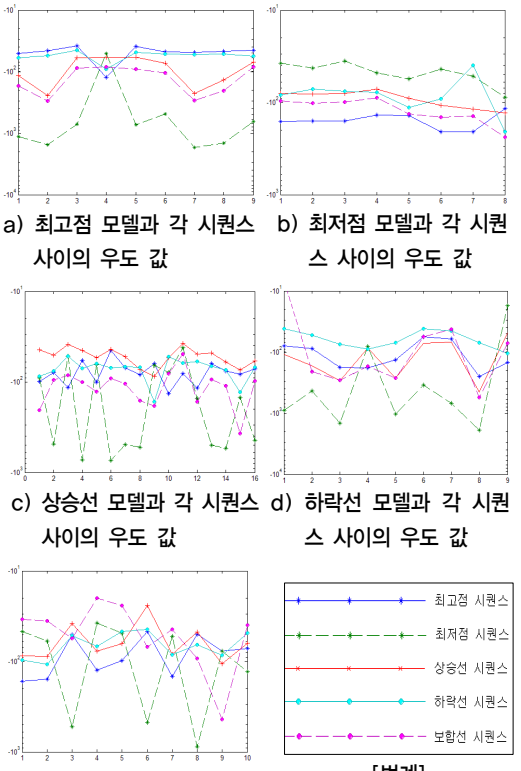
질의 시퀀스에 대한 패턴을 찾은 후에 그 패턴의 값이 최고점 모델이면 현재의 질의 시퀀스는 주가가 최고인 시점임을 나타내게 되어 단순 상승과는 분리되는 의미를 찾을 수 있게 된다. 즉, 앞으로 주가가 상승하게 된다 하더라도 현재 시점의 주가보다 더 높지 않을 가능성이 크다는 것을 의미하게 된다. 그리고 최저점 모델로 결정되면 주가가 최저의 시점임을 알 수 있게 된다. 이것은 앞으로 주가가 현재보다 더 낮아지는 상황이 발생하지 않을 가능성이 크다는 것을 의미하게 된다.

Ⅳ. 실험결과

실험에 사용한 데이터는 2000년에서 2006년 사이의 업종별 KOSPI 지수이다. 이 데이터에서 년도나 종목에 구애받지 않고 시퀀스를 추출하기 위해서 데이터 추출 전에 앞에서 설명한 전처리 과정을 수행했다. 전처리 과정을 통해 변환된 데이터에서 상승선, 하락선, 보합선, 최고점, 최저점의 특징을 잘 나타내는 시퀀스들을 추출했다. 이때 데이터의 길이는 30일로 하였다. 30일 데이터를 사용한 이유는 HMM 학습 모델을 이용하여 주식의 양태를 예측[12]하는 경우에 30일 데이터를 사용하였을 경우 가장 유사한 패턴의 예측 데이터를 생성하였기 때이다. 최고점의 경우에 1년 동안의 주식데이터에서 가장 높은 주가를 나타내는 부분을 기점으로 이전 데이터 30개를 선택했다. 최저점이 경우도 1년 동안의 주가에서 가장 낮은 주가를 나타내는 부분을 기점으로 이전 데이터 30개를 선택했다. 이렇게 추출된 시퀀스들은 학습 모델 생성과 테스트를 위해 사용 했다. 학습모델은 추출된 데이터 들 중에서 각 패턴 별로 5개씩의 시퀀스를 사용하여 은닉 마코프 모델 형태의 학습

모델을 생성하였다.

각 패턴별 학습모델들과 테스트 데이터를 사용하여 우도 값을 측정하였다. 다음의 [그림 2]는 각 학습 모델과 상승, 하락, 보합, 최고, 최저의 특징을 갖는 시퀀스들과의 우도 값에 로그를 취한 결과를 그래프로 그린 것이다. x축은 시퀀스 번호를 나타내고 y축은 로그를 취한 우도 값이다. 여기서 우도 값에 로그값을 취한 것은 각 모델별 우도 값의 편차를 줄이기 위한 것이다.



a) 최고점 모델과 각 시퀀스 사이의 우도 값

b) 최저점 모델과 각 시퀀스 사이의 우도 값

c) 상승선 모델과 각 시퀀스 사이의 우도 값

d) 하락선 모델과 각 시퀀스 사이의 우도 값

e) 보합선 모델과 각 시퀀스 사이의 우도 값

[범례]

그림 2. 각 모델과 각 시퀀스 사이의 우도값

[그림 2]에서 a)를 보면 최고점 시퀀스들 중에서 하나만 제외하고 나머지는 모두 다른 어떤 시퀀스 보다 최고점 모델과의 우도 값이 가장 높게 나타나고 있는 것을 볼 수 있다. 이것은 최고점의 시퀀스들이 상승선과 유사한 모양임에도 불구하고 상승선 모델로 분류되지 않고 최고점 모델로 분류 된 것을 의미한다. b)에서도 최저점 시퀀스들은 거의 대부분 최저점 모델과 가장 높은 우도 값을 나타내는 것이 볼 수 있다. 이것 역시 하락선과 유사한 모양을 나타내는 최저점 시퀀스가 대부분 하락선 모델이 아닌 최저점 모델로 분류된 것을 알 수 있다. c), d), e)의 경우도 상승선 시퀀스들은 상승선 모델과, 하락선 시퀀스들은 하락선 모델과, 보합선 시퀀스들은 보합선 모델과의 우도 값이 높게 나타나는 것의 수가 많은 것을 알 수 있다.

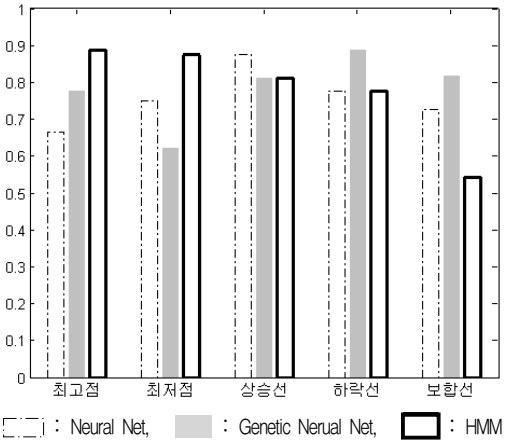


그림 3. 알고리즘 별시퀀스 분류 정확도

[그림 3]은 은닉 마코프 학습 모델 이외에 신경망과 신경망을 사용한 유전자 알고리즘으로 학습시킨 후 테스트한 결과를 나타내는 그림이다. 이 그림에서 x축은 시퀀스의 종류를 나타내고 y축은 분류해 낸 정확도를 나타낸다. 그림에서 보면, 최고점과 최저점 시퀀스의 경우 HMM 학습모델을 사용한 것의 정확도가 높고 하락선과 보합선 시퀀스의 경우에는 신경망을 사용한 유전자 알고리즘을 적용한 테스트 결과가 정확도가 높은 것을 볼 수 있다. 그리고 신경망을 사용한 테스트의 경우 상승선 시퀀스의 분류 정확도가 높게 나타난다.

다음으로 상승선과 최고점 시퀀스들이 각 모델과의 우도 값에 의해서 어떤 모양의 시퀀스로 분류되는가를 예측하고 실제 값을 비교하였다. 그리고 하락선과 최저점에 대해서도 같은 방법을 적용하여 실험하였다. 상승선과 최고점을 분류하는 정확도와 하락선과 최저점의 분류 정확도를 각 [표 1]과 [표 2]에 나타내었다. [표 1]

은 최저점의 시퀀스가 최저점 모델, 하락선 모델과 비교해서 최저점 시퀀스로 분류 되는 정확도와 하락선 시퀀스가 하락선 모델, 최저점 모델과 비교하여 하락선 시퀀스로 분류되는 정확도를 나타낸다. 유사 모양을 나타내는 두 개의 시퀀스가 각각 75 ~ 80% 정도의 정확도를 나타내며 분류되는 것을 볼 수 있다. [표 2]에서도 상승선과 최고점 시퀀스가 모양은 유사하지만 서로 다른 패턴으로 분류되어 대략 70%에서 80% 정도 분류해 내는 것을 볼 수 있다.

표 1. 최저점과 하락선 분류 정확도

시퀀스 종류	정확도	에러율
최저점	76.81%	23.19%
하락선	81%	19%

표 2. 최고점과 상승선 분류 정확도

시퀀스 종류	정확도	에러율
최고점	79.5%	20.5%
상승선	70.6%	29.4%

이러한 결과를 통해서 모양만이 아닌 의미에 따라서 분류될 수 있음을 알 수 있다. 이 결과 값은 또한 하락의 상황이더라도 최저점의 시퀀스로 분류되면 1년 중 가장 낮은 주가를 나타내는 시기임을 알 수 있게 된다. 이것은 앞으로 더 이상 주가가 낮아지는 상황은 발생하지 않을 가능성을 예측할 수 있고 앞으로 상승의 가능성을 생각할 수 있게 된다. 그리고 상승의 상황에서도 최고점의 시퀀스로 분류되면 1년 중 가장 주가가 높은 시점이므로 앞으로 더 이상 높은 주가가 발생하지 않을 가능성을 예측할 수 있게 된다.

V. 결론

본 논문은 주식 시계열이 갖는 모양에 의해서만 패턴을 정의하지 않고 그 시계열의 특징이 나타내는 의미를 기반으로 해서 패턴을 정의하고 은닉 마코프 모델 형태의 학습 모델을 생성하였다. 각 패턴별 질의 시퀀스는 60% ~ 80% 정도의 분류 정확도를 나타내었다. 그러므로 은닉 마코프 모델을 사용하여 시계열의 특징에 대한 패턴별 모델을 생성할 수 있음을 알 수 있다. 특히 최고점의 경우 상승선과 유사한 추세의 모양을 나타내고 있기 때문에 모양만을 생각했다면 같은 패턴으로 분류되었을 것이다. 그러나 의미를 기반으로 하는 학습 모델링에서의 최고점 시퀀스는 상승선과는 다른 패턴으로 분류됨을 볼 수 있었다. 그리고 상당히 높은 분류 정확도를 나타내는 것을 알 수 있다. 이것은 동일한 패턴으로 분류 될 것을 다른 패턴으로 분류한 것이다. 그리고 질의 시퀀스에 그 패턴의 의미를 적용할 수 있게 된다. 그러나 본 논문에서는 단순화 된 의미만을 부여하여 패턴을 정의하고 적용하였기 때문에 의미의 적용 범위가 너무 좁고 정의된 패턴의 수도 너무 작아서 좀더 정확한 분류에는 문제가 발생하게 된다. 앞으로 확장된 의미의 패턴을 정의하고 학습 모델로 작성하여 세분화되고 정확도가 더 높아진 분류와 예측이 가능하도록 연구를 진행할 것이다.

참 고 문 헌

[1] G. Xianping, "Pattern Matching in Financial Time series Data," 1998.

[2] X. Ge and P. Smyth, “Deformable Markov model templates for time-series pattern matching,” In proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, MA, Vol.20, No,23, pp.81-90, 2000(8).

[3] E. Keogh, "A fast and robust method for pattern matching in time-series databases," In Proc. of the 9th Int. Conf. on Tools with Artificial Intelligence, pp.578-584, 1997.

[4] E. Keogh and P. Smyth, "A probabilistic Approach to fast pattern matching in time series databases," In the third conference on Knowledge discovery in Database and Data mining, pp.24-30, 1997.

[5] W. Wang and J. Yang, P. S. Yu, "Mining Patterns in Long Sequential Data with noise," ACM SIGKDD Exploration, pp.28-33, 2001.

[6] J. Hamaker and J. Zhao, "Bayesian Information criterion for automatic model selection," Technical Report, Mississippi State University, 1999(5).

[7] 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, 1998.

[8] S. Singh, "Pattern Modelling in time-series forecasting," Cybernetics and Systems-An International Journal, Vol.31, issue 1, 2000.

[9] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, "Unsupervised Speaker Change Detection Using Probabilistic Pattern Matching," IEEE Signal Processing Letters, Vol.13, No.8, 2006(8).

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

[11] 전호상, 남궁재찬, "혼합된 GA-BP 알고리즘을 이용한 얼굴 인식 연구", 한국정보처리학회 논문지, 제7권, 제2호, 2000.

[12] 전진호, 조영희, 이계성, "주가 운동양태 예측을 위한 모델 결정에 과한 연구", 한국컨텐츠학회논문지, 제6권, 제6호, 2006.

저 자 소 개

조 영 희(Young-Hee Cho) 정회원

- 1995년 2월 : 단국대학교 전자계산학과(이학사)
- 2000년 2월 : 단국대학교 전자계산학과(이학석사)
- 2005년 8월 : 단국대학교 전자계산학과(박사과정 수료)

<관심분야> 데이터마이닝, 기계학습, 에이전트

전 진 호(Jin-Ho Jeon) 정회원

- 1998년 명지대학교 경영정보학과(경영학석사)
- 2007년 단국대학교 전자계산학과(이학박사)
- 2003년 3월 ~ 현재 : 관동대학교 경영정보학부 겸임교수

<관심분야> 기계학습, 데이터마이닝

이 계 성(Gye-Sung Lee) 정회원

- 1980년 : 서강대학교 전자공학과(학사)
- 1982년 : 한국과학기술원 전자계산학과(석사)
- 1994년 : Vanderbilt University 전자계산학과(공학박사)
- 1994년 ~ 1995년 : 대구대학교 전산정보학과 전임강사
- 1996년 ~ 현재 : 단국대학교 컴퓨터과학과 교수

<관심분야> 기계학습, 데이터마이닝, 비디오마이닝, 바이오인포메틱스