2007년도 대한전기학회 하계학술대회 논문집 2007. 7. 18 - 20

주식 예측을 위한 은닉 마코프 모델의 이용

박형준, 홍다혜, 김문현
성균관대학교 정보통신공학부 인공지능 연구실

Using Hidden Markov Model for Stock Flow Forecasting

Hyoung-Joon Park, Da-Hye Hong, Moon-Hyun Kim
School of Information and Communication Engineering, SungKyunKwan University

Abstract - 주식 예측은 주식 시장이 생긴 이래로 투자자들이나, 금융 전문가들 사이에서 매우 중요한 일이 되어 왔다. 그러한 중요성으로 인해 엘리오트 파동이론과 같은 많은 주식 예측 기법이 제시되었고, 또한 이러한 예측의 자동화를 위해 인공지능분야에서도 많은 연구가 있어왔다. 주가 예측에 패턴인식 방법을 적용한 기존의 연구로는 주로 ANN(Artificial Neural Network)방식과 은닉 마코프 모델(HMM, Hidden Markov Model)이 있었고, 본 논문에서는 HMM을 이용한 방법을 제안한다. HMM은 시간 순차적인 패턴을 가지는 모델의 인식에 좋은 성능을 보여 주로 음성인식 분야에서 많이 이용되고 있다. 주식 변화 역시 시간 순차적 흐름에 따라 기울기의 변화가 어느 정도 일정한 패턴을 가지는 성질이 있고, 이것은 HMM을 이용한 패턴인식으로 주식의 앞으로의 변화를 예측하기에 적합한 요인이 된다. 본 논문에서는 이를 위해 다음과 같은 과정을 걸쳤다. 첫 번째로 실존 회사의 장기간의 주식 데이터를 기반으로 여러 개의 HMM모델을 학습 하였다. 두 번째로 예측하고자 하는 기간 이전의 주식 변화 데이터를 입력으로 하여, 이전에 이와 유사한 패턴이 있었는지를 HMM을 통해 알아냈다. 마지막으로 이렇게 알아낸 패턴을 이용하여 앞으로의 주식 변화를 예측하였다. 실험은 실제 주식 변화와 예측값의 비교를 통해 정확도를 검증하였다.

1. 서 론

주식 투자는 주식 시장이 생긴 이래 현재까지, 그 높은 수익률을 이유로 가장 인기 있는 투자 방법 중 하나다. 하지만, 그 주식의 예측하기 힘든 변화율 때문에 투자에 있어 항상 높은 위험이 있었던 것 또한 사실이다[1]. 이러한 이유로 주식을 예측하는 여러 가지 기법들은 많은 관심을 받아왔는데, 기존의 연구로는 ARMA(Auto-Regression Moving Average)기법이나, ANN(Artificial Neural Network)등을 이용한 방법들이 있다. 하지만 그러한 대부분의 연구들은 각각의 한계가 있는데, 예를 들어 ANN의 경우 정형화된 구조를 따라야만 한다는 단점이 있다[2]. 그러한 단점 때문에 몇몇 연구자들에 의해 퍼지 시스템을 도입한 예측 모델이 제시되기도 했다. 또한 HMM을 이용한 주식 예측 방법도 소개된 바 있으나, 그 예측범위가 하루단위로 제한되었다. 즉, 기존의 방식은 다음날의 주식 예측을 위해 오늘의 주식 데이터(시가, 종가, 고가, 저가)를 입력하여, 주식 모델에 대한 확률을 구하고, 이전의 데이터 중에서 그와 가장 유사한 확률을 나타내는 날짜를 찾아내, 그 다음날의 가격 차이를 이용하여 예측하는 방식으로 이용되었다. 하지만 이런 기법은 요즘과 같이 주식 가격의 변화폭이 큰 시기에는 그렇게 좋은 예측 방법이 될 수 없다. 예를 들어 1년 전의 주식 가격이 5000원일 때의 주가 변화와 현재 10000원일 때의 주가 변화의 크기는 다를 수밖에 없다. 또한 전체적인 시스템에서 이용된 HMM모델의 개수가 하나였는데, 이 역시 오차를 야기할 수 있다. 많은 데이터열을 기반으로 학습을 하였지만, 하나의 모델을 만든 탓에 전혀 다른 패턴이 비슷한 확률로서 나타나는 경우가 생길 수 있기 때문이다. 이러한 방법은 순차적인 흐름의 패턴을 잘 설명하는 HMM모델의 특징을 제대로 이용하지 못한 부분이 있다[3]. 이에 본 논문에서는 이산형 HMM을 이용하여 학습과 예측에 이용될 데이터열을 주식의 가격이 아닌 그 등락폭으로 하고, 그 길이 또한 하루단위가 아닌 학습에는 2주일간의 데이터열, 검증에는 1주일간의 데이터열을 이용한다. 즉 가장 최근의 1주일간의 등락률의 데이터열을 입력하면, 그와 가장 흡사한 지난 2주간의 모델을 찾아내어 남은 1주간의 주식을 예측한다. 이 기법의 정확도는 (주)사라콤, (주)삼성전자, (주)sk증권, (주)교보증권의 4개의 실제 1년 4개월간의 데이터를 이용하여 평가한다.

2. 은닉 마코프 모델 및 주식 예측으로의 적용

2.1 은닉 마코프 모델(HMM, Hidden Markov Model)

HMM은 시간적(temporal)패턴인식분야에 적합한 이론이다. 시간적 패턴이란 인식 대상 패턴이 어느 시간적인 간격 동안 부분별로 나뉘어져서 순차적(sequential)인 입력으로 제공되는 것을 말하며, 음성 인식시스템에서 하나의 음절이 얼마의 시간 동안 마이크의 감지 신호로 주어지는 것, 문자 인식시스템에서 하나의 문자영상이 구성 원소 단위로 분할되어 순차적인 입력으로 제공되는 것들이 이의 예들이다[3]. HMM은 상태와 특징과의 관계를 특징방출행렬로 표현하는 이산형HMM 모델과, 확률밀도함수로 표현하는 연속형 HMM 모델로 나뉘어진다.

2.1.1 이산형 은닉 마코프 모델의 정의

HMM은 두 개의 통계적 프로세스를 갖는 유한 상태 머신(finite state machine)으로, 상태(state)와 상태간을 연결하는 천이확률(transition probability)로 구성되어 있다. 이산형 HMM 모델의 정의는 다음과 같다. 이산형 HMM은 한 상태에서 출력 가능한 심벌의 개수가 유한개이며, 이들의 출력확률이 이산 분포특성을 갖는 모델로, 다음과 같이 5개의 파라미터로 표현된다.

$$\lambda = (A, B, \pi, N, M)$$

여기서, λ는 HMM 자체를 나타내며, N은 모델에서 사용되는 상태수를 그리고 M은 모델에서 사용되는 심벌수를 의미한다. HMM의 정의를 위해서 사용되는 파라미터의 의미는 아래와 같다.

i) N: N은 HMM을 구성하는 상태의 수이다. 비록 상태가 은닉되어 있으나, 일부 응용 프로그램의 경우 상태가 특정한 의미를 가지는 경우도 있다. 각 상태를 1,2,3,.., N으로 명명하고, 시점 t에서의 상태를 q_t로 표시한다.

ii) M: M은 상태 당 관측 심벌의 수를 나타내는 것으로, 심벌의 크기를 나타낸다. 관측심벌들은 모델화된 시스템의 물리적인 출력에 대응된다. 각각의 관측심벌들을 V = {v1, v2}

iii) A=a_{ij}; A는 상태 천이 확률이다. 즉 a_{ij} 라 하면 이는 i라는 상태에서 j라는 상태로 천이될 확률을 의미한다.

iV) B=b_j(k); B는 상태에서의 관측 확률이다. 즉 b_j(k)라 하면 이는 j라는 상태에서 k라는 관측값이 나올 확률이다.

V) π = π_i; π는 초기상태 확률로서 π_i라 하면 처음에 i라는 상태가 선택될 확률이다. 일단 π_i가 결정되고 나면 그다음부터는 상태 천이 확률인 A에 의존하여 다음의 상태가 결정된다.

2.1.2 은닉 마코프 모델의 3가지 문제와 해법

i) 확률 평가 문제

관측열 O={$O_1, O_2, O_3, O_4, \ldots$}과 모델 $\lambda = (A, B, \pi, N, M)$이 주어져 있을 때, 이 관측열이 어떠한 모델로부터 발생할 확률이 가장 큰가? 즉, 우도 P($O|\lambda$)를 어떻게 효과적으로 계산할 것인가? 에 대한 문제로 전향(forward)과 후향(Backward)알고리즘으로 해결할 수 있다.

ii) 최적 상태열을 찾는 문제

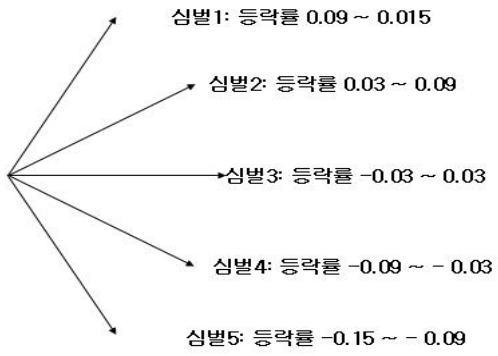
관측열 O={$O_1, O_2, O_3, O_4, \ldots$}과 모델 $\lambda = (A, B, \pi, N, M)$이 주어져 있을 때, 주어진 관측열을 가장 잘 생성해낼 수 있는 최적의 상태열 q=($q_1, q_2, q_3 \ldots$)을 어떻게 찾을 것인가? 에 대한 문제로 이 문제는 비터비(Viterbi)알고리즘으로 해결할 수 있다.

iii) 파라미터 추정의 문제

하나의 관측열 $O=\{O_1, O_2, O_3, O_4, \ldots\}$을 가지고 그 관측열의 우도 $P(O|\lambda)$를 최대화하는 모델 $\lambda = (A, B, \pi, N, M)$을 어떻게 추정하여 구할 것인가에 대한 문제로 이는 바움-웰치(Baum-Welch) 재추정 알고리즘을 이용하여 해결할 수 있다[4].

2.2 주식 예측을 위한 은닉 마코프 모델의 이용

어떠한 목적으로 HMM을 이용하던지 간에 상태와 각 상태에서 방출될 심벌의 종류와 개수를 정하는 것은 매우 중요한 작업이다. 방출될 심벌을 결정하는 것은 곧 주식예측에 이용될 관측열을 결정하는 것과 같으므로 본 논문에서는 실제 주식의 가격의 매일 매일의 등락률((오늘종가-어제종가)/어제종가)을 그 심벌로 정한다. 등락률을 관측열, 즉 심벌로 정한 이유는 만약 단순히 주식의 가격을 심벌로 정할 경우, 각각의 주식마다 가격의 범위가 틀리기 때문에, 매번 다른 회사의 주식 가격을 예측할 때 마다 그 심벌 또한 변화시켜 줘야 한다. 하지만 등락률 또한 그 범위가 -0.15 ~ 0.15까지의 무한대의 경우가 나올 수 있으므로 이산형 HMM의 모델로의 적용을 위해 이러한 등락률의 범위를 정해 이산적으로 나누어 줄 필요성이 있다. 본 논문에서는 다섯 개의 심벌을 가지는 HMM으로 표현하기 위해 즉, 하나의 등락률이 가지는 수치를 주식 차트에서의 기울기와 같은 개념으로 보고 이를 급한 증가 기울기, 완만한 증가 기울기, 평행 기울기, 완만한 감소 기울기, 급한 감소 기울기의 5가지 특징으로 나누어 표현하였다. 이는 <그림 1>과 같다.



<그림 1> 등락률의 범위와 심벌의 설정

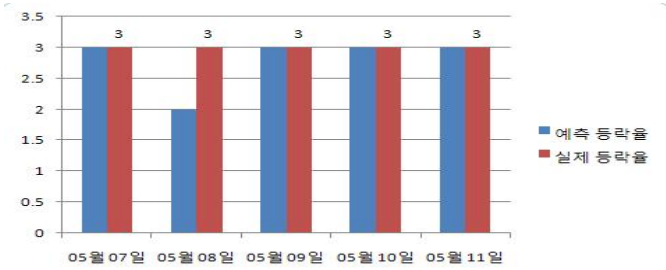
또한 상태의 개수는 초기값을 정하기 쉽도록 하기 위해 심벌의 수와 일치하도록 5개로 정하였다. 그리고 상태 당 관측 심벌의 수인 M을 구하기 위해 각 상태 1, 2, 3, 4, 5를 심벌 1, 2, 3, 4, 5와 일치하도록 설정하였고 그래서 초기 상태 천이행렬은 실제 각 심벌의 변화율을 직접 계산하여 구성하였다. 예를 들어 심벌 1에서 2로 변화의 개수가 30번이 나왔고 총 변화 개수가 300번이라면 상태 1에서 2의 천이 확률은 1/10이 되도록 설정하였다. 그리고 초기 상태확률도 마찬가지로 실제 심벌의 관측 확률로서 계산하였다.

2.2.1 주식의 HMM모델 학습

주식의 HMM모델의 학습을 위해 여기서 Baum-Welch 알고리즘을 적용하였다. 한 가지 주식을 임의로 선정한 후 2006년 1월 1일부터 2007년 4월 30일까지의 데이터를 이용하여 이를 66개의 모델로 나누었다. 현재 목표하고자 하는 예측 방법은 각각의 모델을 2주단위로 학습을 하고 최근 1주일간의 매일 매일의 등락율을 관측열로 하여 확률 평가 문제 방법으로 최고의 확률을 나타내는 모델을 알아내는 것이다. 그럼 그 모델의 실제 등락율의 변화가 예측값이 되는 것이다. 하지만 그냥 단순히 1년 4개월간의 실제 데이터를 등분하게 되면 실제 관측열이 모델의 중간이나 끝부분부터 일치하게 될 경우 예측 정확도가 무척 떨어지게 된다. 그래서 2주 단위로 학습을 하지만 각각의 학습 모델에서 1주간씩은 서로 겹치도록 설정하였다. 즉, 1일부터 14일까지가 첫 번째 학습 모델이 된다면, 두 번째 학습 모델은 7일부터 21일까지가 된다.

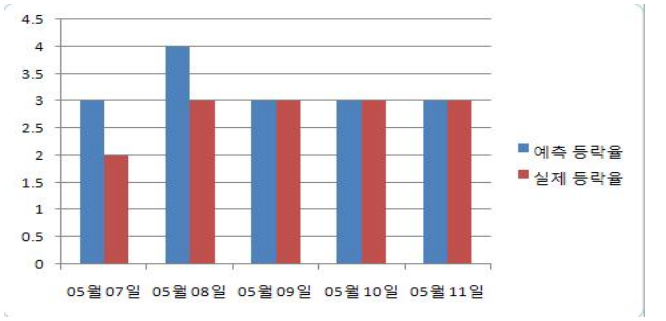
3. 실험결과

정확도 평가를 위해 실제 몇 가지 주식을 선택 한 후에 2006년 1월 1일부터 2007년 4월 30일까지의 데이터를 이용하여 학습을 하였다. 학습과 예측을 위한 프로그램은 Matlab7.0을 이용하였다. 사용된 주식 데이터는 (주)사라콤, (주)삼성전자, (주)sk증권, (주)교보증권 의 4가지 종목을 선택하였고 각각 66개의 모델로 나누었다. 그리고 2006년 4월 27일부터 5월 4일까지의 5일간의 데이터열을 입력값으로 하여 5월 7일부터 11일까지의 5일간의 결과값을 예측하였고, 실제 데이터와 비교하였다. 결과는 다음과 같다.

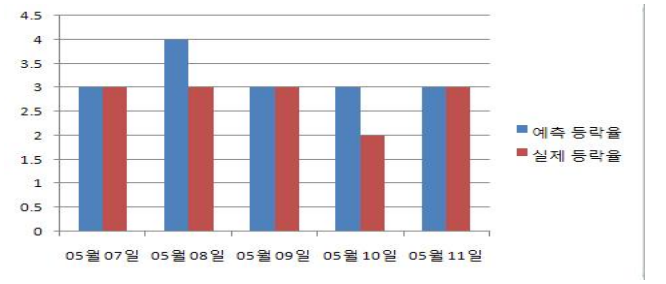


<그림 2> 삼성전자 결과

<그림 3> 사라콤 결과



<그림 4> sk증권 결과



<그림 5> 교보증권 결과

이상의 각각의 예측에 대해 80%, 100%, 60%, 60%의 정확도가 나왔다. 본 논문에서는 단지 심벌과 상태를 5가지로 분류를 하여서 평가를 하였지만 만약 더 많은 심벌과 상태로 나눌 경우 정확도와 그 실제 주식예측에 있어서 실용성이 증가할 것으로 보인다.

[참 고 문 헌]

[1] Ramon Laurence, "Using Neural Networks to Forecast Stock Market Prices.", 1997. 12. 12.
[2] Md. Rafiul Hassan and Baikunth Nath, "Stock Market Forecasting Using Hidden Markov Model", IEEE, 2005.
[3] 김문현, "인공지능", 생능출판사, 2001.7.10.
[4] 김창주, "은닉 마르코프 모델을 이용한 한국어 숫자 음성의 인식", 1999.