시계열 데이터의 양자화된 문자열 변환을 통한 새로운 패턴 분석 기법

김형준, 윤태진, 조환규
부산대학교 정보컴퓨터공학과
e-mail:{hjkim83, ytj, hgcho}@pusan.ac.kr

A New Pattern Analysis Methodology for Time-Series Data using Symbol String Quantization

Hyong-Jun Kim, Taijin Yoon, Hwan-Gue Cho
Dept of Computer Science, Pusan National University

요 약

시계열 데이터에서 패턴을 분석하는 기법은 많은 발전이 이루어져 오고 있으나 주식시장의 경우 패턴 분석 및 예측에 관련되어 많은 연구가 이루어져 있지 않고 있다. 이는 주가의 등락 자체가 본질적으로 무작위하다고 생각되어지고 있기 때문이다. 본 연구에서는 주가의 등락이 보여주는 무작위성의 정도를 Kolmogorov Complexity로 측정, 그 무작위성의 정도와 본 논문에서 제시한 반전역정렬로 예측하는 주가의 예측 간의 상관관계를 보인다. 이를 위하여 KOSPI 주식 데이터 28년 690개의 데이터를 수집하여 이들 주식 데이터의 등락을 양자화된 문자열로 변환하여 본 논문에서 제시한 방법의 의미를 평가하였다. 그 결과 Kolmogorov Complexity가 높은 경우에는 주가 변동 예측이 어려우며, Kolmogorov Complexity가 낮은 경우에는 주식 변동 예측은 가능하나 등락 예측 율은 단기 예측은 12%이상의 예측 율을 보일 수 없으며, 장기 예측의 경우 54%의 예측 율로 수렴함을 확인하였다.

1. 연구동기

최근 재테크에 대한 열풍에 따라 주식에 대한 사람들의 관심이 높아지고 있다. 특히 어떻게 하면 주식 시장을 예측하여 좀 더 많은 이윤을 남길 수 있을지에 사람들의 관심이 몰리고 있다. 이런 가운데 컴퓨터를 이용한 주식 시장 예측은 사람들의 큰 관심을 가져오고 있다.

데이터의 속성에 시간적 요소가 차지하는 비중이 크다는 점에서 주식 예측과 일기 예보는 많은 공통점을 가지고 있다. 하지만 일기예보와는 달리 주식 예측은 아직 낮은 예측 율을 보인다. 가장 큰 이유는 예측 결과가 실제 미래에 미치는 영향력이라고 볼 수 있다. 일기 예보의 경우 예측 결과는 실제 미래에 영향을 주지 않지만 주식 예측은 실제 투자자들에게 영향력을 행사하여 예측 결과를 다르게 낼 수 있다. 이런 주식 시장의 특수성을 분석하여 실제 주가는 랜덤하게 움직인다는 이론이 발표되기도 하였다[1].

이런 주식 예측에 대한 부정적인 시각에도 불구하고 최근까지 많은 연구가 이루어 졌으며 컴퓨터 사이언스 분야에서는 크게 두 가지 접근 방법이 연구 중에 있다. 그 중 하나는 인동지능의 한 분야인 신경회로망이다. 신경회로망은 주어지는 데이터를 통해 학습을 하여 예측을 하는 방식으로 많은 연구가 이루어지고 있다. 또 다른 분야로는 패턴 검색으로 최근 Eugene의 이론[1]에 반하는 연구결과들이 발표됨에 따라 다시 활발하게 연구되기 시작하고 있다. 패턴 검색 방식은 주식 데이터를 시간의 흐름에 따른 순차적인 데이터들의 집합으로 보고 이를 시간 공간에 매핑하여 패턴을 검색한다.

이 논문에서는 주식 데이터의 패턴 존재 유무를 판별하기 위하여 Kolmogorov Complexity를 도입하였다. Kolmogorov Complexity를 이용하여 주식 데이터에 패턴이 존재함을 밝힌 다음 이를 생물학에서 유전자의 패턴을 찾는 방법 중 하나인 반전역정렬을 이용하여 예측한다.

그 결과 Kolmogorov Complexity가 높은 경우에는 변동 예측이 어려우며, 낮은 경우에는 주식 변동 예측은 가능하나 등락 예측 율은 단기 예측은 12%이상의 예측 율을 보일 수 없으며, 장기 예측의 경우 54%의 예측 율에 수렴함을 확인하였다.

2. 관련 연구

주식 시장 예측은 매우 어려운 작업이다. Eugene Fama가 제안한 The Efficient Market Theory에 따르면 주식은 그 시점의 모든 알려진 정보를 반영하고 있으며 과거의 주식 패턴을 이용하여 주식을 예측하는 것은 불가능하다고 하였다. 왜냐하면 어떠한 주식도 과거의 시점과 똑같은 정보를 가지고 있지 않으며 주식 시장 또한 같은 상태로 반영되지 않기 때문이라고 한다. 그러나 Technical Analysis라고도 불리는 이런 주식 예측 시스템 개발은 계속 이루어지고 있었다. 특히 컴퓨터 시스템을 이용한 주가

예측 시스템은 이런 기술적 분석을 바탕으로 하고 있는데 그중 가장 유명한 방법 중 하나는 신경회로망을 이용하는 것이다[2]. 다른 방법으로는 기술적 분석에서 사람이 판단하는 방식을 컴퓨터로 시뮬레이션 하는 방법이다. 실제로 사람이 판단할 때에도 몇 가지 지표를 이용하여 그 지표에 따라 결정을 하기 때문에 전문가 시스템에 응용하여 좀 더 빠르고 정확하게 주식 시장을 예측할 수 있을 것으로 기대되고 있다[3].

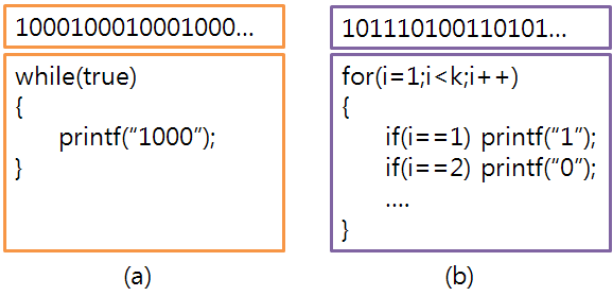
이 논문에서 제시한 방법론의 원형은(Prototype) 생물학에서 사용되어 큰 주목을 받고 있다. 생물학에서는 어떤 유전자의 특성(발현정도)을 시간 축에 따라 어떻게 변화하는지 살펴보는 것을 매우 중요한 주제로 간주하고 있다. 예를 들어서 어떤 암 유전자가 높은 온도에서 또는 낮은 온도에서 산소가 많은 상태, 적은상태 등 각각의 독립된 상태에서 어떻게 변화가 있는지 알아내고 그런 변화가 이미 잘 알려진 다른 암 유전자와 얼마나 유사한지 규명하는 것은 암 치료법이나 신약개발에 매우 중요한 역할을 수행한다. 이 실험은 보통 microarray라고 하는 특별한 실험 장치를 사용하여 측정하는데 이 실험결과 한 유전자에 대하여 대략 30개 내외의 시계열 데이터가 생성된다. [4] 등은 이러한 짧은 시간의 시계열 데이터를 이용하여 어떤 유전자의 내제된 특성을 규명하는 방법을 제시하였다. 즉, 이미 유전자의 특성이 잘 알려진 데이터가 확보된 상황에서 어떤 질의 유전자의 데이터를 비교하여 비슷한 패턴을 보인다면 질의 유전자도 유사한 특성을 가진다는 것이다. 우리는 이런 유전자의 발현 정도가 주식의 등락과 매우 유사함에 주목하여 응용하였다.

3. 주식 예측 기법

주식을 예측하기 위하여 우리는 윈도우 개념을 도입하였다. 주식데이터를 일별로 사용하는 것이 아닌 윈도우의 크기에 따라 변화시켜 예측함으로서 단기예측과 장기예측을 모두 수행할 수 있었다. 또한 주식 예측을 위해 크게 두 가지 방법을 이용하였다. 먼저 주식 데이터에 실제로 패턴이 존재하는지 유무를 판별한 다음 생물학에서 사용되는 반전역정렬을 이용하여 실제 주식을 예측하여 연관성을 찾아보았다.

3.1. 주식 데이터 내 패턴 검색

주식 데이터에 패턴이 존재하는지를 알아보기 위해서는 반대로 주식 데이터가 무작위성을 가지는지를 조사하면 된다. 특정 문자열이 무작위성을 가지는지를 판별하기 가장 좋은 방법은 Kolmogorov randomness를 이용하는 것이다. Kolmogorov randomness(algorithmic randomness)에 의하면 어떤 문자열의 크기가 그 문자열의 Kolmogorov complexity보다 작으면 그 문자열은 랜덤하다고 한다. Kolmogorov complexity에 입각하면 위의 무작위성을 만족하는 문자열은 압축이 불가능하며 그 문자열보다 짧은 길이를 가지는 프로그램으로는 절대 나타낼 수 없다고 정의되어 진다. 즉, 주식 데이터를 특정한 문자열로 나타낼 경우 만일 주식 데이터 내에 어떠한 패턴도 존재하지 않으면 Kolmogorov complexity에 의하여 압축을 할 수 없다는 결론이 나온다.



(그림 1) Kolmogorov Complexity의 예. (a)는 Kolmogorov Complexity가 낮은, 즉 패턴이 반복되는 경우로 문자열을 표현하는 프로그램의 길이가 문자열보다 짧게 나타날 수 있다. 그에 반해 (b)는 Kolmogorov Complexity가 높은 예이며 문자열보다 프로그램의 길이가 짧아지기 힘들다. 즉, 패턴이 존재하지 않는 경우이다.

그림-1은 Kolmogorov Complexity의 예이다. (a)는 Kolmogorov Complexity가 낮은, 즉 패턴이 반복되는 경우로 문자열을 표현하는 프로그램의 길이가 문자열보다 짧게 나타날 수 있다. 그에 반해 (b)는 Kolmogorov Complexity가 높은 예이며 문자열보다 프로그램의 길이가 짧아지기 힘들다. 즉, 패턴이 존재하지 않는 경우이다.

그러나 Kolmogorov complexity는 계산이 불가능하기 때문에 본 논문에서는 [5]와 [7]에서 사용한 바 있는 압축 알고리즘을 이용한 Kolmogorov complexity 측정을 이용하였다. 특정 텍스트를 압축을 하였을 때 압축이 가능하다는 것은 그 텍스트를 서술할 수 있는 좀 더 작은 텍스트가 존재한다는 것이며 이를 통하여 우리는 Kolmogorov complexity를 유추할 수 있다.

3.2. 반전역정렬을 이용한 주식 예측

주가 예측을 위한 패턴 매칭을 위해서 우리는 Semi-global 정렬을 사용하였다. 전역 정렬의 경우 쿼리 시퀀스와 비교 시퀀스 모두를 사용하며, 지역 정렬의 경우 쿼리 시퀀스의 일부와 비교 시퀀스의 일부만 사용한다. 하지만 반전역정렬의 경우 쿼리 시퀀스의 모두와 비교 시퀀스의 일부만 이용하여 정렬을 수행하기 때문에 주식 패턴 검색에 매우 유리하다.

반전역정렬은 상대적으로 크기가 작은 시퀀스의 앞, 뒤에 가상의 공간이 존재한다고 가정한다. 즉, 두 시퀀스의 크기는 동일하게 처리하면서 크기가 작은 시퀀스의 앞, 뒤에 유사도 값을 계산하지 않음으로서 크기가 다른 두 시퀀스에서 전역정렬을 수행하는 것이다. 주식 데이터에서 패턴

을 찾을 때에는 기준이 되는 주식 패턴은 크기가 비교 대상에 비해 매우 짧다. 그러므로 전역 정렬을 수행한다면 찾고자 하는 패턴을 대상 주식의 모든 영역에서 찾음으로서 패턴을 찾는 것이 아니라 그 패턴과 유사한 주식을 찾는 결과가 되어 버린다. 하지만 반전역정렬을 이용하면 대상 주식에서 필요한 패턴을 빠르게 찾을 수 있다.

3.2. 양자화 문자열을 이용한 주가 변동 예측 방법

주가 예측을 위해서 많은 방법들이 개발되었다. 앞서 설명한 바와 같이 크게 패턴 매칭과 신경회로망을 이용하는 방법들로 나뉠 수 있다. 본 보고서에서는 주식 데이터를 일정한 가공을 통하여 불필요한 정보들을 제거한 뒤 반전역정렬을 이용하여 유사영역을 탐색하는 방식을 채택하였다. 불필요한 정보의 제거를 위하여 본 논문에서는 앞선 연구결과[6]를 바탕으로 주식 데이터를 양자화 하여 실험하였다. 반전역정렬을 이용하여 주가를 예측하기 위해서는 몇 가지 최적화해야할 변수들이 존재한다. 특히 예측을 하고자하는 주식 데이터의 범위는 매우 중요한 변수이다.

정의1. *주식 예측을 위한 기준 주식 K번째 일자의 데이터는 $S_o(k)$로 나타내어지며 다음과 같이 정의된다.*

$$S_o(k) = S_o(k-a) \odot S_o(k-a+1) \odot ... S_o(k-2) \odot S_o(k-1)$$

정의1에 따르면 예측을 하고자 하는 주식 데이터는 예측을 하고자 하는 일자의 a번째 이전 데이터부터 예측일의 바로 앞 데이터까지의 모음이라고 정의된다. 여기서 a의 크기에 따라 유사한 주식 패턴의 개수가 다르게 나오므로 정확도를 위해서 매우 중요한 변수가 된다. 이 변수에 대해서는 차후 실험을 통하여 최적화 과정이 필요하지만 본 논문에서는 a=15로 고정하고 실험하였다.

정의2. *주식 예측을 위한 기준 주식 K번째 일자의 데이터에 대해 대상 주식데이터는 $S^i_t(k)$로 나타내어지며 다음과 같이 정의된다.*

$$S^i_t(k) = S^i_t(0) \odot S^i_t(1) \odot ... S^i_t(k-3) \odot S^i_t(k-2)$$

예측 대상 데이터는 모든 주식들이 해당이 되며 자기 자신의 데이터도 포함되어 있다. 이는 자기 자신의 데이터에서도 반복되는 패턴이 존재할 수도 있기 때문이며, 차후 전역 정렬을 이용하여 유사한 종목으로 한정하여 좀 더 나은 예측도를 구할 수 있을 것으로 기대된다. 위의 정의에서 재미있는 점은 데이터 내에서 기준 데이터는 예측 바로 전 데이터까지를 사용하지만 대상 주식 데이터는 예측 2단계 전 데이터를 사용한다는 점이다. 이는 패턴이 먼저 선행되어 나타나야지만 예측이 가능하다는 것을 나타내며 만일 똑같이 움직이는 주식이 존재할 경우에는 기준 주식과 동일하게 움직이는 주식을 이용하여 예측하기는 매우 힘들다는 것을 나타낸다.

정의3. *주식 예측을 위한 기준 주식 K번째 일자의 데이터에 대한 주식 예측 값은 SP(k)로 나타내며 다음과 같이 정의된다.*

$$SP(k) = S_t^{Msga(k)}(AlignLast(S_o(k), S_t(k)) + 1)$$

$$Msga(k) = \max_i \{\mathrm{SemiGlobalAlignment}(\mathrm{S_o(k)}, \mathrm{S_t^i(k)})\}$$

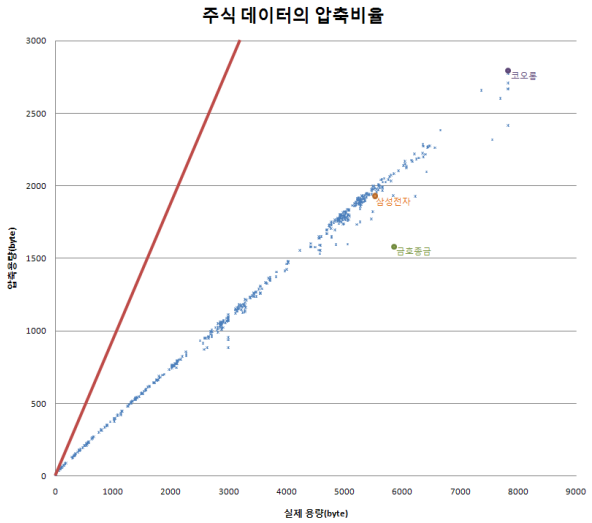
정의3은 주식 예측값 $SP(k)$를 나타내고 있다. 위의 정의에 따르면 가장 높은 지역정렬 유사도를 가지는 주식 데이터의 반전역정렬의 k+1번째 값으로 정의 된다. 즉, 비슷한 패턴을 찾아서 그 패턴의 다음번 값을 주식 예측 값으로 간주한다.

4. 실험

실험을 위하여 본 논문에서는 KOSPI 주식 데이터 28년 690개의 데이터를 수집하여 이들 주식 데이터의 등락을 양자화된 문자열로 변환하여 실험을 진행하였다.

4.1. Kolmogorov Complexity 실험

주식 데이터에 패턴이 존재하는지 알아보기 위한 실험은 다음과 같이 진행하였다. 우선 주식 데이터들을 종목별로 분류하여 각 주식별로 주가 데이터를 저장한 후 압축 알고리즘을 이용하여 얼마나 압축되는지를 확인하였다.



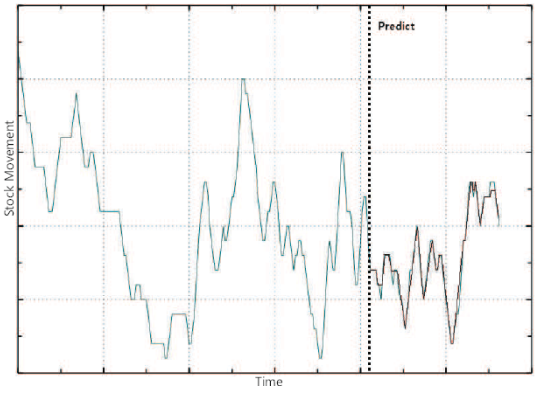
(그림 2) 주식 데이터의 압축 비율. 직선은 압축이 되지 않았을 경우에 분포해야할 위치를 나타내며 파란 점들은 실제로 압축된 주식 데이터들의 압축률을 나타내고 있다.

그림-2와 같이 주식데이터를 압축하였을 때 실제로 압축이 되는 것을 확인할 수 있다. 이는 주식 데이터들이 실제로 패턴을 가지고 있음을 나타내고 있다.

4.2. 주가 변동 예측 실험

주가 예측 모델을 실제로 예측하기 위해서는 매일의 종가로 되어있는 주식 데이터를 이용하면 실제 예측 율은

매우 떨어진다. 이는 예측에 큰 도움이 되지 않는 정보들이 많이 포함되어 있기 때문이며 이런 정보들을 제거하면 예측 율을 증가시킬 수 있다. 하지만 너무 많은 정보를 제거하면 필요한 정보가 제거됨에 따라 예측의 효용성이 떨어지게 된다.



(그림 3) 삼성 SDI 주식 변동 예측 윈도우 크기가 3일 경우에 대해 2월 한 달 동안의 예측으로 비교적 높은 예측 율을 보인다.

그림-3은 삼성 SDI 주식에 대해 변동 예측을 수행한 그래프이다. 2008년 2월 한 달에 대해 예측을 수행하였으며 비교적 높은 예측 율을 보이는 것을 확인할 수 있다 그러나 위의 케이스는 예측 율이 높은 케이스이며 모든 경우에서 위와 같은 결과를 얻을 수는 없었다.

<표 1> KOSPI 데이터의 윈도우크기별 등락예측율. 최소 12.57%에서 최대 54.53%까지 예측이 가능하다.

|w|	등락예측율	|w|	등락예측율
1	12.57 %	6	45.23 %
2	21.53 %	7	47.28 %
3	27.44 %	8	50.94 %
4	38.54 %	9	53.21 %
5	42.63 %	10	54.53 %

표 1은 KOSPI 데이터의 윈도우 크기별 등락 예측 율을 나타내고 있다. 윈도우 크기가 작으면 등락 예측 율은 매우 낮게 나오며 윈도우 크기가 커지면 최대 54.53%의 예측 율이 나오는 것을 확인할 수 있다. 하지만 너무 큰 윈도우 크기의 예측은 효용성이 떨어진다.

5. 결론

재테크에 대한 열풍과 불안한 금융 시장의 현 상황은 많은 사람들의 관심을 주식 예측에 몰리게 했다. 하지만 주식 시장의 예측은 매우 어려우며 데이터의 방대함과 복잡성으로 좋은 결과를 얻지 못하고 있다.

본 논문에서는 생물학에서 사용되는 기법을 주식시장에 적용하여 주식시장을 예측해보았다. 이를 위하여 먼저 주식데이터에 패턴이 존재하는지를 확인하기 위하여 Kolmogorov complexity를 이용하여 주식이 완전히 랜덤하지 않다는 것을 확인하였다. 그러나 주식에 패턴이 존재함에도 불구하고 예측 율은 단기 예측의 경우 12%정도이며 장기 예측의 경우 54%의 예측 율로 수렴함을 확인하였다. 이는 몇 가지 변수들이 최적화 되지 않은 수치이며 최적화를 통해 더 높은 예측 율을 얻을 수 있을 것으로 기대되지만 실제 주식 투자에 응용가능성은 미지수이다.

참고문헌

[1] Eugene F Fama. Efficient capital markets: A review of theroy and empirical work. *Journal of Finance,* 25(2):383-417, May 1970

[2] 김환용 김성곤, 균등다층연산 신경망을 이용한 금융지표지수 예측에 관한 연구. *한국컴퓨터정보학회 논문지*, 8(3):113-123, 2003

[3] 조근식 이강희, 양인실. 캔들스틱 분석을 이용한 주식매매 타이밍 예측을 위한 전문가 시스템. *한국전문가시스템학회지*, 3(2):57-70, 12 1997

[4] Andrew T. Kwon, Holger H. Hoos, and Raymond Ng. Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics*, 19(8): 905-912, 2003.

[5] Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P. 2003. The similarity metric. *In Proceedings of the Fourteenth Annual ACM-SIAM SODA 2003,* 863-872.

[6] H.-J. Kim, C.-K. Ryu, and H.-G. Cho. A phylogenetic analysis for stock market indices using time-series string alignments. *ICCIT '08. Third International Conference on,* 1:487-492, Nov. 2008.

[7] Mehmet M. Dalkilic, Wyatt T. Clark, James C. Costello, and Predrag Radivojac. Using compression to identify classes of inauthentic texts. *In SIAM '06,* 604-608