■ 2010년도 학생논문 경진대회 수상작

서열 정렬 알고리즘을 이용한 주가 패턴 탐색 시스템 개발

(Developing Stock Pattern Searching System using Sequence Alignment Algorithm)

김 형 준 † 조 환 규 ††
(Hyong-Jun Kim) (Hwan-Gue Cho)

요 약 시계열 데이터에서 패턴을 분석하는 기법은 많은 발전이 이루어져 오고 있다. 그러나 주식시장의 경우 시계열 데이터임에도 불구하고 패턴 분석 및 예측은 많은 연구가 이루어지지 않고 있으며 예측도가 매우 낮다. 그 이유는 주가의 등락 자체가 본질적으로 무작위하다고 하면 어떠한 과학적 방법으로도 그 예측은 불가능하다. 본 연구에서는 주가의 등락이 보여주는 무작위성의 정도를 Kolmogorov 복잡도를 이용해 측정하여 그 무작위의 정도와 본 논문에서 제시한 반 전역정렬(semi-global alignment)로 예측할 수 있는 주가의 예측의 정확간의 깊은 상관관계가 있음을 보인다. 이를 위해서 주가지수의 등락을 양자화된 문자열로 변환하고 그 문자열의 Kolmogorov 복잡도를 이용해 주가 변동의 무작위성을 측정하였다. 우리는 KOSPI 주식 데이터 28년 690개의 데이터를 수집하여 이를 실험용 데이터로 사용하여 본 논문에서 제시한 방법의 의미를 평가하였다. 그 결과 Kolmogorov 복잡도가 높은 경우에는 변동 예측이 어려우며, Kolmogorov 복잡도가 낮은 경우에는 주식 변동 예측은 가능하나 3종류의 예측율에 대해서 투자자들이 관심이 많은 등락 예측율은 단기 예측은 12% 이상의 예측율을 보일 수 없으며, 장기 예측의 경우 54%의 예측율로 수렴함을 확인하였다.

키워드 : 주식, 주가, 패턴 분석, 반 전역정렬, Kolmogorov 복잡도

Abstract There are many methods for analyzing patterns in time series data. Although stock data represents a time series, there are few studies on stock pattern analysis and prediction. Since people believe that stock price changes randomly we cannot predict stock prices using a scientific method.

In this paper, we measured the degree of the randomness of stock prices using Kolmogorov complexity, and we showed that there is a strong correlation between the degree and the accuracy of stock price prediction using our semi-global alignment method. We transformed the stock price data to quantized string sequences. Then we measured randomness of stock prices using Kolmogorov complexity of the string sequences. We use KOSPI 690 stock data during 28 years for our experiments and to evaluate our methodology. When a high Kolmogorov complexity, the stock price cannot be predicted, when a low complexity, the stock price can be predicted, but the prediction ratio of stock price changes of interest to investors, is 12% prediction ratio for short-term predictions and a 54% prediction ratio for long-term predictions.

Key words : Stock, Stock Prices, Searching Pattern, Semi-global Alignment, Kolmogorov Complexity

† 비 회 원 : 부산대학교 컴퓨터공학과
hjkim83@pusan.ac.kr
†† 정 회 원 : 부산대학교 컴퓨터공학과 교수
hgcho@pusan.ac.kr
논문접수 : 2010년 6월 1일
심사완료 : 2010년 10월 7일

정보과학회논문지: 시스템 및 이론 제37권 제6호(2010.12)

1. 서 론

주식 시장에 대한 관심이 높아지면서 주식 시장의 흐름을 예측하는 방법에 대한 연구가 활발하게 진행되고 있다. 주식시장의 예측에 대해서는 여러 가지 방법이 있을 수 있겠지만 주식시장 특유의 복잡성과 자료의 방대함으로 인하여 주식시장 예측에 있어서 컴퓨터의 도움에 많이 의존하고 있다. 하지만 아직까지는 컴퓨터를 이용하여 최종 의사 결정을 내리기보다는 단순한 수학적 계산을 통하여 최종 의사 결정에 도움이 될 수 있는 정보만 제공하는 수준이다. 하지만 단순히 제공되는 정보만을 이용하여 주식을 예측하기에는 주식시장이 매우 복잡하기 때문에 단순히 정보 제공만이 아니라 어느 정도의 예측이 가능한 시스템에 대한 사람들의 필요성이 높아지고 있다.

데이터의 속성에 시간적 요소가 차지하는 비중이 크다는 점에서 주식 예측과 일기 예보는 많은 공통점을 가지고 있다. 두 분야 모두 컴퓨터의 발달에 따라 빠르게 발달하고 있는 분야이다. 하지만 일기예보가 80%이상의 정확도를 보이고 있는 데 비하여 주식 예측은 이러한 정확도를 보여주지 못하고 있다. 즉 컴퓨터를 이용하여 최종 의사 결정을 하는 것이 아니라 컴퓨터는 단순히 최종 의사 결정에 도움을 주는 정보를 제공하는 도구로서만 사용되고 있다. 이는 주식시장이 일기 예보에 비해 데이터가 부족하다고 할 수도 있겠지만 실제 일기 예보와 주식시장 모두 데이터가 구축되기 시작한 것은 100년이 되지 않는다. 주식 예측과 일기 예보의 가장 큰 차이점은 예측 결과가 실제 미래에 영향력을 얼마나 가지고 있느냐이다. 일기 예보의 경우 예측 결과는 실제 미래에 영향을 전혀 주지 않는다. 하지만 주식 예측의 경우 예측한 결과가 발표되면 실제 투자자들에게 영향력을 행사하여 예측 결과 자체가 틀려지게 된다. 이런 주식시장의 특수성을 분석하여 실제주가는 랜덤하게 움직인다는 이론이 발표되기도 하였다[1].

이런 주식 예측에 대한 부정적인 시각에도 불구하고 최근까지 많은 연구가 이루어졌으며 컴퓨터 사이언스 분야에서는 크게 두 가지 접근 방법이 연구 중에 있다. 그 중 하나는 인공지능의 한 분야인 신경회로망이다. 신경회로망은 주어지는 데이터를 통해 학습을 하게 되어 이를 바탕으로 예측을 가능하게 하며 많은 연구가 활발하게 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 매핑한다. 이런 주식 데이터를 이용하여 주식시장을 예측하는 일은 '시간 데이터 마이닝(temporal data mining)'으로 볼 수 있다. 이를 위해서 Lin, Orgun과 Williams은 시간 데이터 마이닝을 두 가지로 분류하였다[2]. 첫 번째는 '유사한 패턴 매칭'이며, 두 번째는 '시간 데이터베이스에서 주기적인 패턴 찾기'이다. 이를 주식시장 예측에 응용하면 첫 번째는 유사한 주식 분류로 볼 수 있으며 두 번째는 과거의 주식 데이터를 응용하여 주식시장의 미래 예측이라고 볼 수 있다.

주식시장을 예측을 하기 위해서는 변수를 줄이기 위해서 비슷한 속성의 주식들을 분류하는 것 이 선행되어야 한다. 이런 주식 분류의 방법에는 여러 가지가 존재하지만 주식시장 예측에 사용하기 위해서는 과거의 주식 데이터들을 이용하여 실제 주식의 특성을 파악하는 방법이 필요하다. 주식이 상장될 때 등록된 주식 분류는 실제로 그 주식의 위치를 알려주기는 하지만 다른 주식들과의 상호 연동에 의한 주식의 고유한 특성은 나타내지 못하는 단점이 존재한다. 그러므로 과거의 주식 데이터를 이용하는 방법이 필요하게 되는데 이런 과거의 주식 데이터를 이용할 때에도 약간의 데이터 가공이 필요하게 된다. 주식은 실제 경제 흐름과 다른 주식의 등락에 매우 민감하기 때문에 이런 정보들이 모두 주식에 반영되어 매우 복잡하게 보이게 된다. 이런 불필요한 정보를 제거하게 되면 주식의 고유한 특성을 파악할 수 있게 되고 이를 이용하여 주식 분류에 활용할 수 있다.

주식을 예측하기 위하여 패턴 매칭을 이용할 수 있다. 과거의 주식 데이터들을 이용하여 주식의 변화를 파악하여 현재 주식 변화와 가장 유사한 변화를 보이는 패턴을 찾은 다음 그 패턴을 이용하여 주식을 예측하는 방식이다. 이를 위하여 DNA 염기서열에서 유사한 영역을 찾는 기법인 정렬(alignment) 기법을 응용가능하며 실제로 많은 주식 데이터들 가운데서 빠르게 패턴을 검색할 수 있어서 유용한 방법으로 사용되고 있다.

그림 1은 본 논문에서 구현한 시스템을 이용하여 예측한 결과이다. S모 기업의 2007년 1월부터 2008년 2월까지의 주식 데이터에 대해 유사한 주식 패턴을 보이는 다른 주식들을 분류하여 이들 정보를 바탕으로 유사한 패턴을 찾아내는 방식으로 주식을 예측하였다. 위의 결과를 얻기 위하여 우선 예측 단위를 일별로 하지 않고 특정한 구간만의 데이터를 살펴보는 범위(이하 '윈도우'로 표기)를 도입하여 몇 일간의 주식 데이터의 합의 평균을 이용하였다. 이를 통하여 불필요한 정보를 제거하였으며 유사한 주식 패턴을 보이는 다른 주식들을 모두 패턴 예측 데이터로 사용함으로서 빠르게 패턴을 찾으며 또한 결과의 정확도를 높일 수 있었다. 예측은 예측 데이터들 중 유사한 패턴들을 찾아 적용 하는 방식을 사용하였다. 본 논문에서 주식 예측 시스템을 구현함으

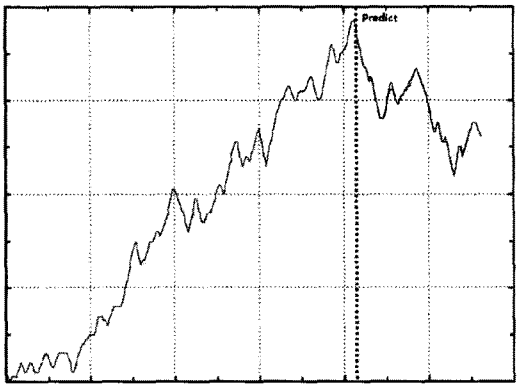

그림 1 S기업 주가의 예측 그래프. 가로축은 윈도우 크기 3의 시간 변화, 세로축은 시간 변화에 따른 주식의 변화율

로서 본 논문에서 밝힌 점은 다음과 같다.

1. 주가에 정렬 기법을 적용하여 특정 시간대에 어느 주식과 유사한지를 검색하였으며 패턴을 통하여 주식 예측을 가능하도록 하였다.
2. 주식들에 전역 정렬을 적용하여 실제 주식시장에서 제공하는 분류법으로 주식들의 상관관계를 모두 나타낼 수 없음을 보였다.
3. 주가 데이터를 가공한 뒤, 정렬 기법을 이용하여 주가의 변동을 예측가능 하며 투자자들이 관심이 많은 등락 예측율은 단기 예측은 12% 이상의 예측율을 보일 수 없으며, 장기 예측의 경우 54%의 예측율로 수렴함을 확인하였다.

위와 같은 점을 밝히기 위해서 본 논문은 우선 주식 데이터를 정렬 기법을 적용하기 적합한 형태로 변형한다. 그리고 전역 정렬을 이용하여 유사한 주식들을 클러스터링 하여 실제 주식 분류와는 다르게 유사한 주식들이 존재함을 보인다. 마지막으로 유사한 주식들의 과거 주식 데이터를 윈도우를 이용하여 특정주식의 미래를 거시적, 미시적으로 예측하여 예측도의 변화를 밝혀내어 예측 모델을 제안한다.

2. 관련 연구

2.1 이전의 주가 예측 방법론

주식시장 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 한다. 왜냐하면 어떠한 주식도 과거의 시점과 똑같은 정보를 가지고 있지 않으며 주식시장 또한 같은 상태로 반영되지 않기 때문이라고 한다. 이 논문에서 Fama는 주식은 완전히 랜덤하게 움직인다고 주장했다[3].

그럼에도 불구하고 주식의 과거 정보를 이용한 주식 예측 시스템 개발은 계속 진행되어져 왔다. '의사 과학(pseudo science)'이라고도 취급 받지만 이런 시스템들을 주로 '기술적 분석'이라고 부른다. 기술적 분석의 방법으로는 '캔들차트 투자기법(candlestick charting)'과 '다우 이론' 등이 있다. 캔들차트 투자기법은 18세기 일본에서 쌀을 매매하던 호마 무네히사가 개발하였다. 바 그래프와 라인 그래프를 혼합한 이 차트는 실제 추세가 어떻게 변하고 있는지까지 나타내기 때문에 예측에 많은 도움을 주었다.

다우 이론은 미국 통신사인 다우존스사의 창설자 찰스 다우, 에드워드 존스, 찰스 버그스트레서 등에 의해 고안돼 발전된 이론으로서 다우존스 평균주가라고도 한다. 다우존스 평균주가는 다우 공업주 30종 평균과 철도주 20종 평균의 상호 움직임에서 주식시세의 대세를 살피는 것으로 공업주 평균과 철도주 평균이라는 두 가지 지표만으로 장세의 대세, 나아가 경기 동향의 향방까지도 판단하고 하는데 특징이 있다. 이는 주가라는 것이 모든 경제활동 정보를 반영한다는 이론에 바탕을 두고 있으며 공업주 평균에서 주요 기업의 생산 활동을 파악하며 철도주 평균은 국내 산업운수의 활동을 각각 반영하고 있어 이 두 개의 지표만으로도 경기 활동의 전반을 파악할 수 있다는 전제로 이루어진 것이다. 하지만 다우 이론은 1929년의 '뉴욕 주가 대폭락'을 예견한 이후 성과는 별로 좋지는 않다.

컴퓨터 시스템을 이용하여 주가를 예측하는 시스템은 이런 기술적 분석을 바탕으로 하고 있다. 기술적 분석 많은 정보를 가지고 있는 주식시장에서 필요한 정보만을 추출하여 그 정보를 지표로 삼아 주식을 예측한다는 기본 틀을 가지고 있는데 이를 컴퓨터 시스템에 접목하여 좀 더 많은 정보를 빠르게 판단하여 좀 더 정확하게 예측을 하고자 하는데 기본 의미를 가지고 있다. 그리고 이렇게 주가를 예측하는 시스템은 크게 인공지능을 이용한 방법과 패턴 분석을 이용한 방법으로 구분지을 수 있다.

컴퓨터 인공지능을 이용한 주가 예측 방법 중 가장 유명한 방법 중 하나는 신경 회로망을 이용하는 것이다[4]. 신경 회로망은 복잡한 패턴들을 빠르게 학습할 수 있는데 이런 점을 응용하여 테스트 데이터로 패턴들을 분석한 다음 이를 이용하여 미래를 예측하는 방법이다. 하지만 테스트 데이터에 따라 다른 결과를 낼뿐 아니라 테스트 데이터와 실제 데이터가 다른 패턴을 가지면 제대로 결과를 내놓지 못하는 점이 약점이다. 그러므로 신경 회로망을 이용한 주식 예측은 테스트 데이터에 독립적으로 모든 패턴에 적용할 수 있는 쪽으로 개발이 진행되고 있다.

다른 방법으로는 기술적 분석에서 사람이 판단하는 방식을 컴퓨터로 시뮬레이션 하는 방법이 있다. 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 기대되어 개발되고 있다[5]. 하지만 이 시스템들 또한 특정 조건에서만 제대로 작동하는 단점이 존재한다. 주식 데이터를 시간의 흐름에 따른 순차적인 데이터들의 집합으로 간주하고 나열된 데이터들 사이에서 패턴을 찾아내어 미래를 예측하거나 주식 분류를 하는 방법도 연구되고 있다[6-10]. 이 방법은 실제주식에 영향을 주는 사건들을 모두 분석하는 것이 아닌 결과를 통하여 원인을 분석하는 방식으로 변수를 제한함으로서 빠른 분석은 가능하나 결과를 해석하는데 시간이 걸린다는 단점도 존재한다.

위에서 나열한 방법들은 모두 주식 데이터에 일정한 패턴이 존재한다고 가정하고 개발된 방법들이다. 하지만 지금까지 주식 데이터에 일정한 패턴이 존재하는가에 대해서는 연구가 이루어지지 않았다. 이 때문에 주식 예측이 의사 과학이라는 평가를 받고 있다.

2.2 생물학에서의 서열 간 유사성 탐색

특정 데이터들 간의 유사도를 측정하는 방법은 여러 방향에서 다양한 연구가 이루어져 왔다. 특히 유전자 DNA를 비교하여 DNA 간의 유사도를 비교하는 생물정보학과 유사한 문서를 찾아내어 유사도 여부를 가리는 유사도 탐색 분야에서 많은 유사도 측정 기법들이 개발되었다[11]. 생물정보학의 경우 데이터의 특성상 유사도 측정 기법의 초점이 대용량의 데이터에서 필요한 영역을 찾는 검색 속도에 초점이 맞추어져서 연구가 진행되어져 왔다. 그 결과 다양한 정렬기법이 나왔고 속도 면에서도 빠른 속도를 자랑하고 있다.

특히 이 논문에서 제시한 방법론의 원형은 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 마이크로 배열이라고 하는 특별한 실험 장치를 사용하여 측정하는데 이 실험결과 한 유전자에 대하여 대략 30개 내외의 시계열 데이터가 생성된다. Kwon[12] 등은 이러한 짧은 시간의 시계열 데이터를 이용하여 어떤 유전자의 내재된 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.

2.3 데이터 간의 유사도 측정법

특정 데이터들 간의 유사도를 측정하는 방법은 여러 방향에서 다양한 연구가 이루어져왔다. 특히 유전자 DNA를 비교하여 DNA 간의 유사도를 비교하는 생물정보학과 유사한 문서를 찾아내어 유사 도 여부를 가리는 유사도 탐색 분야에서 많은 유사도 측정 기법들이 개발되었다. 생물정보학의 경우 데이터의 특성상 유사도 측정 기법의 초점이 대용량의 데이터에서 필요한 영역을 찾는 검색 속도에 초점이 맞추어져서 연구가 진행되어져 왔다. 그 결과 다양한 정렬기법이 나왔고 속도 면에서도 빠른 속도를 자랑하고 있다.

그에 비해 유사도 탐색영역은 아직까지 많은 연구가 진행되고 있다. 현재유사도 탐색은 크게 두 가지 영역으로 나뉘는데 프로그램 소스코드 유사도 탐색과 자연어 문서 유사도 탐색이다. 특히 자연어 문서 유사도 탐색의 경우에는 유사도 데이터를 이루고 있는 자료가 방대하기 때문에 많은 어려움이 존재한다. 그 때문에 문서 유사도는 현재 특정 언어 도메인에 대해서만 유사도 검사 연 구가 진행 중에 있다. 문서 유사도 검사 방식에는 구문(syntax)을 이용한 방식과 문맥(semantics)을 이용한 방식이 있다. 문맥을 이용한 방식은 문장의 구조나 단어의 앞뒤 순서에 상관없이 문장의 뜻을 이용 하여 유사도를 탐색하기 때문에 유사도 여부를 가장 정확히 판별할 수 있는 방식이다. 하지만, 자연어의 의미 자체를 파악하는 작업 자체가 비용이 클 뿐 아니라 현재로서는 올바르게 의미를 파악하는 시스템도 존재하지 않기 때문에 문맥을 이용한 유사도 탐색은 실제 그 실용성에 대해서는 의문 이 제기되고 있다[13].

현재 유사도 탐색에서 가장 활발하게 연구되고 있는 영역인 구문을 이용한 유사도 탐색에는 특성변수 계산 방식(attribute counting)과 구조 통계적 방식(structure metric)이 존재한다. CloneChecker[14], SCAM[15]은 특성변수 계산법을 이용하며, Plague[16], YAP[17], SIM[18]는 구조 통계적 방식을 사용하고 있다. 프로그램 소스 코드의 유사도 탐색도 문서 유사도 탐색과 유사한 방법을 이용하여 연구되어지고 있다. 그러나 프로그램 소스 코드는 언어 자체의 특성상 구조화되어 있기 때문에 자연어 탐색보다는 쉽게 이루어지고 있다. 특히 소스 코드의 세부적인 내용을 이용한 탐색이 아닌 Kolmogrov 복잡도를 응용하여 빠르게 소스코드간의 유사도를 분석하는 방법이 제안되기도 하였다[19]. 그리고 계산이 불가능한 Kolmogorv 복잡도의 한계를 극복하기 위하여 압축율을 이용하여 프로그램 간의 유사도를 분

석하는 기법도 제안되었다[20]. 이들 방법은 유사도 프로그램들 사이에 유사한 패턴이 있다고 가정하고 그 패턴을 나타내는 방법이 동일하기 때문에 유사도를 판별 가능할 것이라고 가정하고 있다.

3. 서열 정렬을 이용한 주가 예측 시스템 개발

3.1 주가 데이터 변환

주식 데이터를 나타낼 수 있는 척도는 여러 가지가 연구되었고 현재 가장 많이 사용되는 방법은 4가지 정도가 있다[21]. 우선 시간 t에서의 주가를 Y(t)라고 한다면 다음과 같은 주가 척도가 있다.

1. $Z(t) = Y(t+\Delta t) - Y(t)$
2. $Z_D(t) = [Y(t+\Delta t) - Y(t)]D(t)$
3. $R(t) = \frac{Y(t+\Delta t) - Y(t)}{Y(t)} = \frac{Z(t)}{Y(t)}$
4. $S(t) = \ln Y(t+\Delta t) - \ln Y(t)$

1번 방법은 주가 변동만을 조사하는 방법으로 가장 간단한 방법이기는 하지만, 주가간의 크기에 따른 변화율을 나타내지 못한다는 단점이 있다. 즉, 주가가 100인 경우에 10의 변화를 가지는 경우와 주가가 1000인 경우에 10의 변화를 가지는 경우를 구분하지 못한다는 단점이 존재한다. 2번 방법은 이런 단점을 해결하고자 D(t)라는 해당 시점의 '정적' 인자를 추가하는 것이다. 이를 통하여 일정한 변화율을 얻을 수 있는 장점은 있지만 특정 시점의 D(t)를 정확하게 명시할 수 없기 때문에 실제로는 거의 사용되지 않는다. 3 번째 방법은 t동안의 주가 변화율을 조사하는 방식으로 퍼센테이지를 돌려주기 때문에 1번 방법에서 나타나는 문제를 해결할 수 있다. 하지만 이 방법은 장기적인 시간변화에 따른 주가척도의 변화에 민감하여 잘못된 결과를 나타낼 수 있다는 단점이 있다. 4 번째 방법은 2번 방법과 3번 방법의 장점을 따서 만든 방법으로 선형성을 유지하기 위해서 ln을 이용하였다. 4번 째 방법은 경제 성장률이 일정할 경우에는 제대로 척도가 반영이 되지만 보통의 경우에 경제 성장은 일정하지 않기 때문에 약간의 단점은 존재한다. 하지만 S(t)는 주가 변동과 경제 성장 지표의 방향성을 모두 가지고 있기 때문에 주식 데이터의 척도로 매우 유용하다. 특히 주식 데이터는 보통의 경우 매일의 데이터를 이용하기 때문에 4번의 방법은 매우 유용하게 사용된다.

3.1.1 k-level 양자화

이 장에서는 두 주식간의 유사도를 구하는 방법에 대해 설명한다. 한 주식의 주가는 시간순의 주식 값들로 이루어져 있다. 그러므로 우리는 한 주식의 하나의 주식 값을 정의할 수 있다.

정의 3.1 어떤 주식 i의 주가S_i에 대하여, k번째 날의 주식 값은 $S_i(k)$로 나타낸다.

매일의 주식 값은 많은 변수와 노이즈를 포함하고 있기 때문에 이 주식 값만 이용하면 좋은 결과를 얻을 수 없다. 그렇기 때문에 우리는 슬라이딩 윈도우 개념을 주가 계산에 도입하였다.

정의 3.2 어떤 주식의 주가S_i에 대하여, $|w|$가 w인 k번째 *timestamp* 주식 윈도우는 $SW_i^w(k)$로 표기하고 다음과 같이 정의된다.

$$SW_i^w(k) = \sum_{j=k-w}^{k} S_i(j)$$

정의 3.2를 이용하여 우리는 주식을 미시적, 거시적으로 분석이 가능하다. 윈도우를 이용하여 어떻게 두 주식을 분석하는데 도움이 되는지는 나중에 이야기한다.

정의 3.3 주식 윈도우 SW_i^w에 대하여, k번째 *timestamp*의 주식윈도우의 양자화는 $Q_i^w(k)$로 표기하고 다음과 같이 정의된다.

$$Q_i^w(k) = 100 \cdot \frac{SW_i^w(k) - SW_i^w(k-1)}{SW_i^w(k-1)}$$

주식 값을 문자열로 변환하기 위해서 우리는 주식 값을 양자화 해야 한다. 정의 3.3을 통해 우리는 바로 전의 주식윈도우와의 차 비율을 나타내는 $Q_i^w(k)$를 구할 수 있다. $Q_i^w(k)$를 이용하여 우리는 양자화 룰을 적용하여 주식 값을 문자열로 변환할 수 있다. 표 1을 이용하여 우리는 정수 값을 문자 값으로 변환 가능하다.

UB는 양자화 값의 상위 제한 값을 나타내며 LB는 하위 제한 값을 나타낸다. 이 값들을 조절함으로서 우리는 좀 더 좋은 정렬 값을 얻을 수 있다. 또한 어떤 주식들은 어떤 날들에는 몇 가지 이유로 인하여 값들이 존재하지 않는다. 이런 값이 존재하지 않는 주식이 존재하

표 1 i번째 날의 $|w|=w$인 $QC_i^w(k)$. 총 5개 단계 'T', 'U', 'C', 'D', 'B'와 'Z(존재하지 않음)'로 나타내어진다. UB는 양자화 값의 상위 제한 값, LB는 하위 제한 값을 나타낸다.

$QC_i^w(k)$	범위
T	$UB < Q_i^{(w)}$
U	$LB < Q_i^{(w)} \le UB$
C	$-LB < Q_i^{(w)} \le LB$
D	$-UB < Q_i^{(w)} \le -LB$
B	$Q_i^{(w)} \le -UB$
Z	S_i does not exist

면 우리는 $QC_i^w(k)$를 'Z'로 설정하고 계산에 어떠한 영향도 주지 않도록 했다.

정의 3.4 주식 S_i에 대하여, $|w|=w$인 S_i의 양자화 집합을 QS_s^w로 표기하고 다음과 같이 정의한다:

$$QS_i^w = \{QC_i^w(k) | i \in N\}$$

QS_s^w는 특정한 한 주식 값을 나타내는 문자이기 때문에 우리는 시간 순 문자열을 만들기 위해서 하나의 주식에 대해 모든 문자들을 모을 필요가 있다. 이 시간 순 문자열을 이용하여 우리는 주식들 간의 정렬 값을 구하고 두 주식간의 유사도를 구할 수 있다. 그리고 이를 바탕으로 유사한 주식들을 클러스터링을 할 수 있다.

표 2는 주식 데이터를 이용하여 양자화 된 문자열로 변환된 예이다. 숫자로 이루어진 주식 데이터들이 전후 데이터간의 차이를 이용하여 양자화 된 문자열로 변환되는 것을 확인할 수 있다.

표 2 2007년 1월 4일부터 2007년 2월 27일까지의 7개 기업의 실제 주식 데이터들의 양자화 된 문자열 변환

주식명	양자화된 문자열 (20070104~20070227)
S1기업	BTBBBTCBBBTCBTBUUBUBBCUBBUUUUTBBTTBTBC
S2기업	BBBUBUTUCBBBUUUUBBUBCUCCCBUBBUUTUUBBBB
L1기업	CBBUBCUBUBBBUBBUBBUBUUUUUUUCBTUUUBBBUU
S3기업	CBCBBBBUUBUBUCUUBBBBUUTBBCUBBUCBUTUBBB
K1기업	BUBBBUBBTBUCBBUCCUBBUUTCUBTBCCUUCBCCCC
L1기업	BUCTBUBBCBCCCTUUCTBBBUBBUTBCUCBBBCTUCU
H1기업	BUBCUCBCUUBBUUCBBUUBCCUBBBBTBCBUBUTBU

3.1.2 이동 평균선에 의한 추세 예측

이동 평균선은 주로 시계열 데이터에서 단기간의 변화를 억제하고 보다 장기간의 변화율을 부각하기 위해서 사용된다. 주로 주식이나 금융 데이터의 테크니컬 분석에 사용되며 각종 거시경제학에서의 시계열 데이터 분석에 사용된다. 이동 평균선은 신호분석에서의 저역통과 필터(low pass filter)와 비슷한 역할을 하는데 이를 통하여 단기간의 변화를 뺀 장기간의 변화율을 파악할 수 있다.

가장 간단한 이동 평균으로는 SMA(Simple Moving Average)가 있다. SMA는 과거 n 포인트에 대한 비 가중 평균을 나타낸다. 수식으로 나타내면 다음과 같다.

$$SMA = \frac{p_M + p_{M-1} + \ldots + p_M - n}{n}$$

SMA는 가장 기초적이고 많이 사용되는 이동 평균이기는 하지만 주기적인 방향성만을 제거할 수 있다는 단점이 있다. 즉, 완벽하게 주기적인 사이클이 존재할 수 없는 경제학이나 주식 데이터에 대해서는 제대로 작용할 수 없다는 단점이 존재한다.

SMA에 대비되는 이동 평균으로 가중 이동 평균이 있다. 이는 각각의 다른 데이터 포인트에 대하여 다른 가중치를 적용하여 이동 평균을 내는 방법이다. 테크니컬 분석에서 WMA는 가중치가 산술적으로 줄어든다는 점에서 SMA와 비교된다. WMA의 수식은 다음과 같다.

$$WMA = \frac{np_M + (n-1)p_{M-1} + \ldots + 2p_{M-n+2} + p_{M-n+1}}{n + (n-1) + \ldots 2 + 1}$$

WMA는 최근의 데이터에 좀 더 민감하게 반응하게 되어 방향 예시 기능은 강화되나 이동평균이 가지는 고유의 장점인 '신뢰도'를 없애버릴 수 있다는 단점이 존재한다. 신뢰도란 지표를 어느 정도 신뢰할 수 있는 정도를 나타내며 어느 지표든지 주가의 변화를 지나치게 빠르게 따라잡을 경우 신뢰도가 떨어진다. 이는 정보가 주가의 변화에 영향을 느리게 미치기 때문이다.

그림 2는 가중 이동 평균의 가중치를 나타내는 그래프이다. 최근의 Data일수록 높은 가중치를 받기 때문에 최근의 추세를 좀 더 많이 반영 할 수 있다는 장점이 있다.

WMA의 산술적인 가중치를 지수 승으로 변경하여 EMA(Exponential Moving Average)가 만들어졌다. 지수이동평균의 수식은 다음과 같다.

$$EMA = \alpha \times (p_1 + (1-\alpha)p_2 + (1-\alpha)^2 p_3 + (1-\alpha)^3 p_4 + \ldots)$$

지수이동평균의 파라미터로는 smoothing factor라고 불리는 α가 있다. 보통 0에서 1사이의 값을 가지며 다음과 같이 결정된다.

$$\alpha = \frac{2}{N+1}$$

제시된 수식들을 이용하여 지수이동평균의 가중치를 그래프로 나타내면 그림 3과 같다.

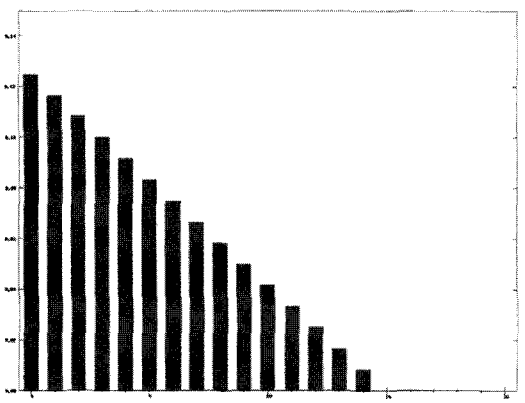

그림 2 n = 15일 때의 Weighted Moving Average의 가중치. 최근의 Data일수록 높은 가중치를 받으나 이동평균이 가지는 고유의 장점인 '신뢰도'를 없애버릴 수 있다는 단점이 존재한다.

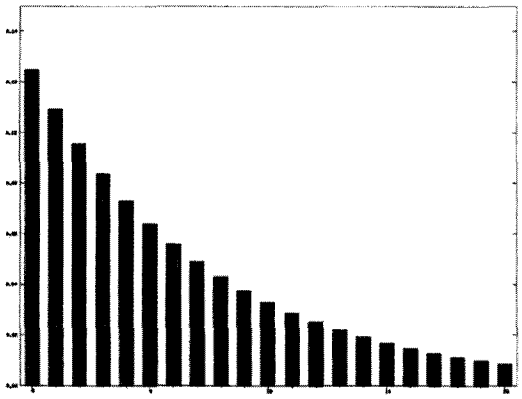

그림 3 n = 15일 때의 Exponential Moving Average의 가중치. 최근의 Data일수록 높은 가중치를 받는다. WMA보다 더욱 최근의 Data에 민감하게 반응한다. α값을 조절하여 WMA와 SMA의 중간지표로 활용가능하다.

지수이동 평균은 WMA보다 더욱 최근의 Data에 민감하게 반응한다. 하지만 α값을 잘 변경한다면 WMA와 SMA의 중간적인 지표로 사용이 가능하다.

3.2 주가들 간의 유사 패턴 검출

3.2.1 반 전역정렬 방법

주가 예측을 위한 패턴 매칭을 위해서 우리는 반 전역 정렬(semi-global alignment)을 사용하였다. 전역 정렬의 경우 쿼리 시퀀스와 비교 시퀀스 모두를 사용하며, 지역 정렬의 경우 쿼리 시퀀스의 일부와 비교 시퀀스의 일부만 사용한다. 하지만 반 전역정렬의 경우 쿼리 시퀀스의 모두와 비교 시퀀스의 일부만 이용하여 정렬을 수행하기 때문에 주식 패턴 검색에 매우 유리하다. 기본적으로 전역 정렬의 경우 두 시퀀스의 크기가 다를 경우에 두 시퀀스의 크기가 동일하다고 가정하고 정렬을 수행하기 때문에 다음과 같은 결과가 나온다.

C A G C A C T T G G A T T C T C G G
C A G C - - - - - G - T - - - - G G

그러나 우리는 찾고자 하는 패턴 내에는 최대한 gap이 존재하지 않는 상태로 나오길 바라고 있다. 즉, 두 시퀀스의 크기가 틀리더라도 gap을 최대한 적게 하고 정렬을 수행하기를 바란다.

반 전역 정렬은 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 전역 정렬을 수행하는 것이다.

C A G C A - C T T G G A T T C T C G G
- - - C A C G T G G - - - - - - - - -

위의 정렬은 반 전역 정렬을 수행하였을 때의 결과 값이다. 아래쪽의 시퀀스의 크기가 작기 때문에 앞, 뒤에 gap을 추가하여 정렬을 수행하였다. 그리고 정렬을 수행하여 유사도 값을 계산할 때 begin gap과 trailing gap은 계산에서 제외한다. 그 결과 가장 유사도 값이 높은 정렬 값이 찾아진다.

주식 데이터에서 패턴을 찾을 때에는 기준이 되는 주식 패턴은 크기가 비교 대상에 비해 매우 짧다. 그러므로 전역 정렬을 수행한다면 결과 값은 처음 정렬 결과와 비슷하게 나올 것이다. 즉, 찾고자 하는 패턴을 대상 주식의 모든 영역에서 찾음으로서 패턴을 찾는 것이 아니라 그 패턴과 유사한 주식을 찾는 결과가 되어 버린다. 하지만 반 전역 정렬을 이용하면 대상 주식에서 필요한 패턴을 빠르게 찾을 수 있다.

3.2.2 유사 주가 사전조사를 통한 패턴 검색

비록 정렬 기법으로 빠르게 두 시퀀스 간의 유사도를 검출할 수 있지만, 모든 다른 주식들에 대하여 유사도 검사를 하기에는 대상 주식들의 수가 많다. 그렇기 때문에 사전에 유사한 패턴을 보이는 주식들을 조사함으로서 실제주식 패턴 검색 시 비교할 대상을 줄이고, 예측 결과의 정확도를 높일 수 있다. 물론 주식시장에서 기본적으로 제공하는 주식의 분류가 존재하기는 하지만 이를 통해서는 실제 주식간의 숨은 영향관계를 파악할 수 없기 때문에 실제주식 데이터를 이용하여 유사 주가들을 분류한다.

유사 주가들을 분류하기 위해서 우리는 전역 정렬 기법을 이용하였다. 전역 정렬 기법은 비슷한 길이의 두 시퀀스의 전체 유사도를 파악하는 방식으로 이를 이용하면 우리는 각 주식들이 얼마나 서로 유사한지 클러스터링을 할 수 있다. 본 논문에서 제시한 양자화된 주식 데이터에 대한 전역 정렬 기법의 적용은 다음과 같이 정의된다.

정의 3.5 윈도우 크기가 w 인 두 개의 서로 다른 양자화 집합 QS_i^w와 QS_j^w에 대하여 전역 정렬 값은 $GA_w(a,b)$로 표기하고 다음과 같이 정의된다:

$$GA_w(i,j) = GlobalAlignmentScore(QS_i^w, QS_j^w)$$

전역 정렬값 $GA_w(i,j)$는 주식의 윈도우 크기만을 파라미터로 사용한다. 이 주식 윈도우 크기를 이용하여 우리는 두 주식의 유사도를 미시적, 거시적으로 분석 가능하다.

3.3 패턴 유사도를 이용한 주식 예측

주가 예측을 위해서 많은 방법들이 개발되었다. 앞서

설명한 바와 같이 크게 패턴 매칭과 신경 회로망을 이용하는 방법들로 나뉠 수 있다. 본 보고서에서는 주식 데이터를 일정한 가공을 통하여 필요 없는 정보들을 제거한 뒤 반 전역 정렬을 이용하여 유사 영역을 탐색하는 방식을 채택하였다. 반 전역 정렬을 이용하여 주가를 예측하기 위해서는 몇 가지 최적화해야 할 변수들이 존재한다. 특히 예측을 하고자 하는 주식 데이터의 범위는 매우 중요한 변수이다.

정의 3.6 주식 예측을 위한 기준 주식 K 번째 일자의 데이터는 $S_o(k)$로 나타내어지며 다음과 같이 정의된다.

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

정의 3.6은 예측을 하고자 하는 주식 데이터의 범위를 나타내는 정의이다. 위의 정의를 따르면 예측을 하고자 하는 주식 데이터는 예측을 하고자 하는 일자의 a번째 이전 데이터부터 예측일의 바로 앞 데이터까지의 모음이라고 되어있다. 여기서 a의 크기에 따라 유사한 주식 패턴의 개수가 다르게 나오므로 정확도를 위해서 매우 중요한 변수가 될 수 있다. 예를 들면 a의 값이 작다면 유사한 패턴이 매우 많이 나오게 될 것이고 이에 따라 정확도가 떨어질 수도 있다. 반면 a값이 너무 크다면 매치 되는 패턴의 수가 하나도 없어서 예측이 불가능하게 될 수도 있다.

정의 3.7 주식 예측을 위한 기준 주식 K번째 일자의 데이터에 대해 대상 주식 i 데이터는 $S_t^i(k)$로 나타내어지며 다음과 같이 정의된다.

$$S_t^i(k) = S_t^i(0) \odot S_t^i(1) \odot \ldots S_t^i(k-3) \odot S_t^i(k-2)$$

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 유사 구한 종목으로 한정하여 좀 더 나은 예측을 할 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.

정의 3.8 주식 예측을 위한 기준 주식 K번째 일자의 데이터에 대한 주식 예측 값은 $SP(k)$로 나타내며 다음과 같이 정의된다.

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

$$Msga(k) = \max\{SemiGlobalAlignment(S_{o(k)}, S_t^i(k))\}$$

정의 3.8은 주식 예측 값 $SP(k)$를 나타내고 있다. 위의 정의에 따르면 가장 높은 지역정렬 유사도를 가지는 주식 데이터의 반 전역 정렬의 $k+1$번째 값으로 정의된다. 즉, 비슷한 패턴을 찾아서 그 패턴의 다음번 값을 주식 예측 값으로 간주한다.

정의 3.9 어떤 주식 S_i에 대하여 주식 예측율 P_i^+, P_i^-, P_i^0로 정의된다.

$$P_i^+ = \frac{\text{실제 상승한 주식 데이터 갯수}}{\text{상승할 것으로 예상한 주식 데이터 갯수}} = \frac{|SR(K) = 'T' \text{ or } 'U'|}{|SP(K) = 'T' \text{ or } 'U'|}$$

$$P_i^- = \frac{\text{실제 하강한 주식 데이터 갯수}}{\text{하강할 것으로 예상한 주식 데이터 갯수}} = \frac{|SR(K) = 'B' \text{ or } 'D'|}{|SP(K) = 'B' \text{ or } 'D'|}$$

$$P_i^0 = \frac{\text{변동이 없는 주식 데이터 개수}}{\text{변동이 없을 것으로 예상한 주식 데이터 개수}} = \frac{|SR(K) = 'C'|}{|SP(K) = 'C'|}$$

양자화된 문자열로 설명하면 P_i^+는 'T'와 'U'로 이루어져 있으며, P_i^-의 경우는 'B'와 'D'이며 P_i^0는 'C'를 나타낸다. 이 중 실제 투자자들이 관심을 가지는 것은 P_i^+, P_i^-에 관심이 있기 때문에 우리는 주식의 등락 예측율을 분리하여서 설명한다. 즉, 실제 주식의 등락 예측율은 주식의 상승 하강에 대한 예측율로 주식투자에 중요한 지표가 된다.

정의 3.10 어떤 주식 Si에 대하여 주식 등락 예측율은 다음과 같이 정의된다.

$$P_i^* = P_i^+ + P_i^-$$

$$P_i^{\sim} = P_i^+ + P_i^- + P_i^0$$

예를 들어 실제 KOSPI 주식 중 2008년 2월 15일 예측의 경우 $a=15, |w|=3$일 때의 양자화 된 문자열 예측은 다음과 같다.

표 3 실제 KOSPI 주식 중 $a=15$이고 $|w|=3$일 때의 2008년 2월 15일의 양자화된 문자열 예측 결과

종목 명	패턴	예측값
S1	BTBBBTCBBBTCBTBBUBUBUCUB BUUUTBBTTBTBCC	B
S2	----------BBUBUTUCBBBUU UT----------	B
D1	CCCTUUCTBBBUBUBUBBUUBUBU UUUUUCUTUCBUB	U
S3	----------BU-UBUBBCUB-BUUU U----------	U
C1	CUBUBBBUBBUBBU-BUUUUUBBC BBUUUBBUBCUCCC	C
H1	-------U-UCBBUUBCCUBBBB---- ----------	B

위와 같이 최대유사패턴의 값에 대상주식의 최근 패턴을 반 전역 정렬을 이용하여 패턴을 찾은 다음 최대유사패턴에서 대상 주식과 일치하는 패턴의 최종 값을 다음 일자의 예측치로 사용한다. 최대 유사패턴은 과거의 주식 데이터 중 어디서든 나타날 수 있다. 예를 들어 S2의 경우 예측 값은 'B'로 나타나며 실제주식의 값도 'B'로 나타난다. S3도 'U'로 일치하지만 H1의 경우 예측 값은 'C'이고 실제 값은 'B'로 서로 다르게 나타나는 것을 확인할 수 있다.

4. 실험 및 결과 분석

4.1 실험데이터

주가 예측 모델을 실제로 예측하기 위해서는 매일 매일 종가로 되어 있는 주식 데이터를 이용해서는 실제 예측율은 매우 떨어진다. 그 원인으로는 실제 하루하루의 종가로 이루어진 데이터는 효용성 없는 정보로 인하여 하루하루 등락을 거듭하기 때문으로 볼 수 있으며 만약 이런 쓸모없는 정보를 제거할 수 있다면 주식 예측의 예측도는 매우 증가할 수 있다. 반면 이런 정보를 너무 제거하게 된다면 필요한 정보까지 제거되어 예측도는 올라가더라도 별로 효용성이 없는 예측이 될 것이다. 이런 정보 제거의 방법으로 이 보고서에서는 주식이 시간 데이터라는 점을 이용하여 $|w|$를 이용하여 정보를 제거하였다. $|w|$의 주식 데이터를 하나의 단위로 두고 이들의 평균을 이용하여 주식의 등락 정보를 모으면 흥미로운 사실을 발견할 수 있다. 만일 $|w|$가 증가함에 따라 당일의 주식 변화율은 전체 데이터에 크게 영향을 미치지는 못하기 때문에 쓸모없는 정보가 제거된다는 점과 너무 큰 $|w|$로 주식 데이터를 가공하였을 경우에는 효용성이 떨어진다는 점이다. 즉, $|w|$ 조정으로 통하여 우리는 적절한 예측 범위를 찾을 수 있을 것이다. 예측을 위한 $|w|$ 결정을 위한 실험에 사용된 데이터는 표 4와 같다.

데이터는 1980년부터 2008년까지 총 28년의 데이터를 이용하여 패턴을 검색하였다. 패턴을 찾기 위해서 만일 주식 데이터가 존재하지 않는다면 그 패턴은 사용하지 않았다. 즉 반 전역 정렬을 통하여 유사하다고 판별된 패턴내부에 데이터가 존재하지 않는 문자열이 나타날 경우 그 패턴은 순위 에서 제거하였다.

표 4 주가 예측 실험에 사용된 1980년부터 2008년까지 총 28년간의 KOSPI 데이터의 속성

Date	KOSPI
StartDate	1980/01/02
EndDate	2008/02/29
Year	28
Num	690

예측도를 검사하기 위해서 실험은 다음과 같이 진행되었다. 우선 2008년 2월 한 달 동안의 각각 주식 690개에 대하여 2008년 2월 1일부터 하루하루 각각 예측을 실행하여 실제주식 데이터와 얼마나 일치하는지를 확인하였다. 예측도는 2008년 2월 한 달 동안의 전체 주가의 개수에 대해 일치한 주가 데이터의 비율로 측정하였다.

4.2 시스템 파라미터의 최적 값 설정

표 5는 $|w|$별로 예측 실험을 수행한 경우의 예측율을 나타낸 결과이다. $|w|$가 커짐에 따라 예측율은 급격하게 증가하는 것을 확인할 수 있지만 $|w|$가 10이상이 되면 변화율이 상수로 수렴되기 때문에 예측율이 올라가는 것으로 확인할 수 있다. 그 결과 그림 4는 $|w|$에 따른 예측율을 나타내고 있다.

$|w|$가 50일까지 조사하였으며 30일 이후로부터는 거의 100퍼센트가 나오는 것을 확인할 수 있다. 이는 $|w|$가 증가함에 따라 변화율이 거의 상수에 수렴하기 때문

표 5 이동 평균선에 따른 주가 예측율. 이동 평균선의 일 $|w|$이 커질수록 주가 예측율도 올라가는 것을 확인할 수 있다.

$\|w\|$	$P_i^{\sim}$	$\|w\|$	$P_i^{\sim}$
1	23.77%	6	75.51%
2	43.62%	7	70.58%
3	57.68%	8	87.97%
4	67.83%	9	88.41%
5	72.32%	10	89.86%

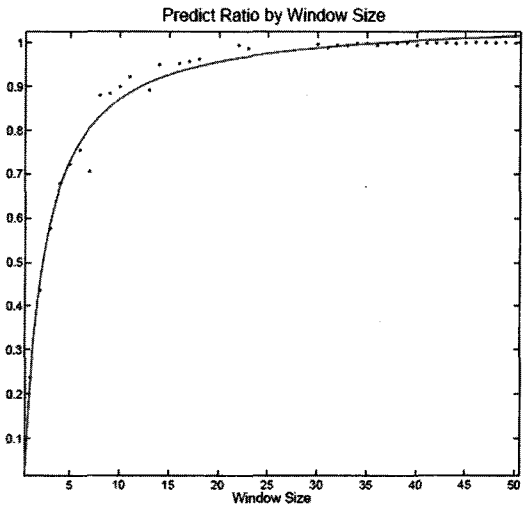

그림 4 Window Size 에 따른 예측율 그래프. f(x)=(p1 · x + p2)/(x + q1)그래프로 수렴되는 것을 확인할 수 있다.

이다. 위의 값들은 다음과 같은 선형 다차방정식으로 나타내어진다.

$$f(x) = (p1 \cdot x + p2)/(x + q1)$$

위의 그래프는 p1 =1.055, p2 = -.485, q1 =1.562로 나타내어지며 $|w|$ 변화에 따른 예측도는 일정한 식에 의해 나타내어지며 예측 가능하다는 것을 나타낸다. 위의 $|w|$에 따른 예측율을 $|w|$에 따른 압축 비율과 비교해 보면 유사 패턴의 개수와 그에 따른 예측율을 확인해 볼 수 있다. 즉 높은 유사도를 보이더라도 압축율이 높다는 것은 그만큼 많은 패턴이 존재하지 않으며 거의 한 종류의 값으로 수렴하고 있다는 것을 나타낸다.

4.2.1 패턴 검색 대상의 최적화

논문에서 제안하는 알고리즘을 데이터 집합에 적용하면 우리는 모든 쌍의 주식들에 대해 $GA_w(a,b)$값 을 구할 수 있다. 이를 통하여 우리는 주식을 분류하고 주식들 간의 숨겨진 영향력을 확인할 수 있다. 표 6은 최상위 10쌍의 $GA_w(a,b)$값을 가지는 주식 쌍을 나타내고 있다. 5개의 쌍은 같은 분류를 가지지만, 나머지 5개는 서로 다른 카테고리를 가지고 있다.

그림 5는 H기업과 L기업의 주식 그래프를 보여주고 있다. 붉은 그래프는 H기업을 나타내며 녹색 그래프는 L기업을 나타내고 있다. 비록 주식 값의 크기는 다르지만 둘 다 같은 시간에 비슷한 움직임을 보이고 있다. 이것은 H기업과 L기업이 서로에게 영향력을 미치고 있다는 것을 나타내고 있다. 그리고 그림 6은 H기업과 L기업의 양자화 그래프를 보여주고 있다. $|w|$를 조절함으로서 우리는 각 주식의 고유한 속성을 밝혀낼 수 있다. $|w|$가 증가하면 그래프의 변화율은 감소한다. 이것은 $|w|$가 증가하면 매일의 주식 변동율이 미치는 영향이 줄어듦을 나타낸다. 이를 통하여 우리는 각 주식의 고유 속성을 밝혀낼 수 있다.

$GA_w(a,b)$를 이용하여 우리는 주식 시장의 진화 계통

표 6 상위 10쌍의 $GA_w(a,b)$. 대부분의 쌍들이 KOSPI에서 기본적으로 제공하는 분류에서 서로 같은 분류를 가진다.

주식A	주식B	$GA_w(a,b)$	A분류	B분류
H1	L1	5358	제조업	제조업
E1	C1	5144	제조업	제조업
S1	S2	5090	제조업	제조업
S3	K1	4979	제조업	금융업
B1	C2	4968	제조업	판매업
D1	T1	4904	제조업	제조업
I1	I2	4887	제조업	제조업
L2	D2	4884	운송업	제조업
L3	S4	4883	제조업	건설업
S5	H2	4861	제조업	금융업

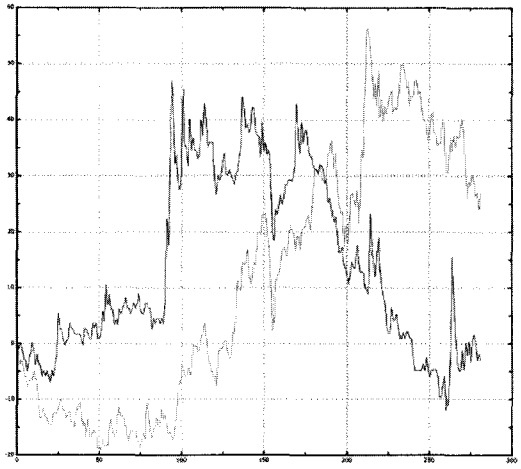

그림 5 H기업(붉은색)과 L기업(녹색)의 주식 그래프. 증감의 폭은 다르지만 증가, 감소 자체는 비슷하게 이루어졌다.

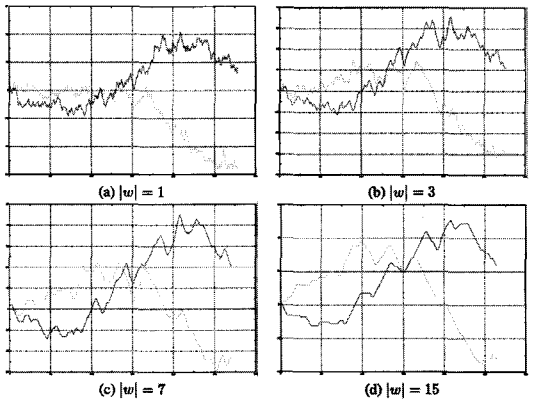

그림 6 H기업(붉은색)과 L기업(녹색)의 주식 그래프. 주식 그래프가 다른 모양을 그리는 것으로 보이지만 증가, 감소는 비슷하게 이루어졌다. 이는 $GA_w(a,b)$과 주식쌍의 등락이 관계가 있음을 의미한다.

그래프를 구할 수 있다. $GA_w(a,b)$의 쌍 들을 연결함으로서 우리는 몇 개의 분리된 그래프를 구할 수 있다. 즉 연결된 그래프들은 유사한 주식이므로 우리는 이 그래프를 통하여 주식을 분류할 수 있다.

그림 7은 양자화 클러스터 그래프의 일부분을 보여주고 있다. 각 노드는 주식을 나타내며 에지는 두 노드 간에 유사한 속성이 있음을 나타내고 있다. 그림 7의 그래프를 보면 L1과 L2, L3와 D1, T1은 비슷한 주식 변화를 보이는 클러스터에 포함되어 있음을 확인할 수 있다.

그림 8은 양자화 클러스터 노드들의 주식 변동률을

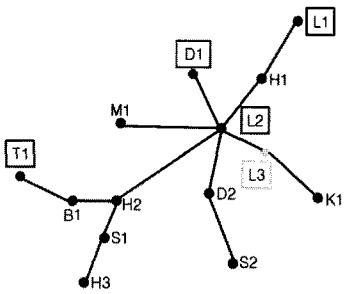

그림 7 양자화 클러스터 그래프의 일부분, 각 노드는 주식을 나타내며 에지는 두 노드 간에 유사한 속성이 있음을 나타내고 있다.

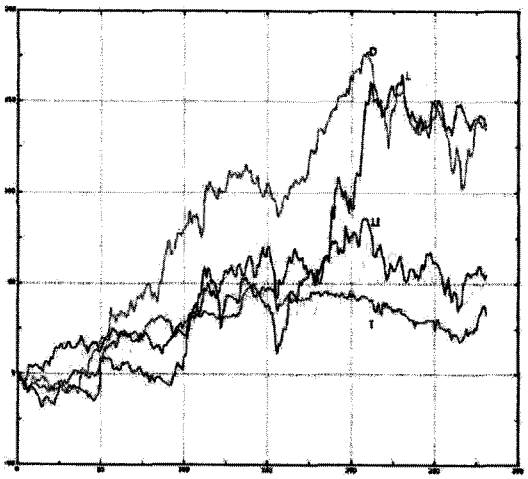

그림 8 양자화 클러스터 노드들의 주식 변동률. 같은 그래프 내에 존재하는 주식들은 모두 비슷한 움직임을 보인다.

보여주고 있다. 각 주식들이 다른 변동률을 보이기는 하지만 비슷한 주식 변동 움직임을 보여주고 있다. 위의 결과를 통하여 비록 주식의 기본 분류가 존재하기는 하지만 주식들을 분류하기 위해서 기본 분류만을 사용한다면 주식 간의 숨겨진 영향력을 파악할 수 없다는 것을 알 수 있다. 즉, 다른 분류로 되어 있더라도 상호 협력이나 기타 분류로는 드러나지 않는 영향력이 주식들 간에 영향을 미치며 이는 주식을 상장할 때 분류되는 기본 분류로는 알 수 없다는 것을 나타낸다.

표 7은 실제 주식의 $|w|$에 따른 예측율이다. KOSPI 데이터 18년의 데이터를 사용하였으며 실제 예측은 2008년 2월 한 달 동안의 예측율을 측정한 것이다. $|w|$이 증가함에 따라 예측율 또한 증가하는 것을 확인할 수 있다. 이는 예측에 불필요한 정보들이 제거됨에 따라 예측율이 증가하는 것으로도 볼 수 있으나, 너무 많은

표 7 실제 주식의 $|w|$에 따른 $P_i^{\sim}$. KOSPI 데이터 18년의 데이터를 사용하였으며 실제 예측은 2008년 한 달 동안의 $P_i^{\sim}$를 측정하였다. $|w|$이 증가함에 따라 예측율 또한 증가하는 것을 확인할 수 있다.

주식명	1	2	3	4
	5	6	7	8
S1	33.77	52.12	64.24	75.21
	81.37	85.45	79.23	80.15
H1	31.53	51.11	61.23	74.23
	80.21	84.22	76.22	80.13
S2	23.76	43.54	57.66	67.81
	72.29	75.48	70.48	87.99
L1	23.76	43.53	57.65	67.82
	72.33	75.50	70.58	88.00
S3	23.77	43.60	57.68	67.80
	72.33	75.53	70.59	88.01
K1	5.90	20.12	35.65	45.22
	53.34	57.57	55.27	68.45
K2	5.10	22.21	36.24	44.21
	52.54	56.51	52.54	67.21
L2	3.77	23.62	37.68	47.83
	52.32	55.51	50.58	67.97

정보들이 제거된 것일 수도 있다.

4.3 시스템의 예측 성능 평가

실제 예측도를 검사하기 위해서 실험은 다음과 같이 진행되었다. 우선 주식 690개에 대하여 2008년 2월 1일부터 한 달 간의 하루하루 각각 예측을 실행하여 실제 주식 데이터와 얼마나 일치하는지를 확인하였다. 예측도는 2008년 2월 한 달 동안의 전체 주가의 개수에 대해 일치한 주가 데이터의 비율로 측정하였다.

그림 9는 실제 주식의 예측율 그래프이다. 전반적으로

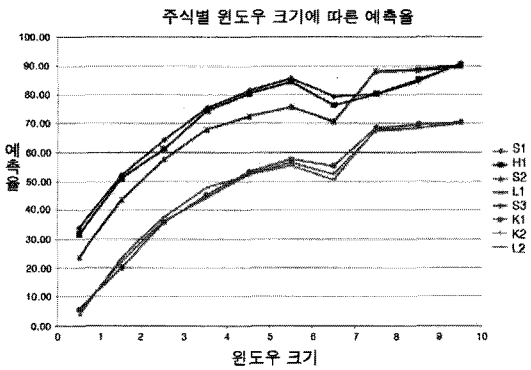

그림 9 $|w|$에 따른 실제 주식의 예측율 그래프. 7일째에 예측율이 떨어지는 것은 주식 시장에서 거래가 일주일에 5일 단위로 이루어지기 때문에 $|w|$에 대한 오차가 발생하기 때문이다.

$|w|$이 증가함에 따라 예측율 또한 증가하는 것을 확인할 수 있으나, 주식에 따라 차이가 나는 것을 확인할 수 있다. 이는 주식에 따라 패턴의 존재 확률이 다르기 때문이며 비슷한 패턴이 적을 경우 주식을 예측하기 힘들어진다. 또한 7일째에 예측율이 전반적으로 감소하는 것은 주식시장의 거래가 일주일에 5일 단위로 이루어지기 때문에 $|w|$가 7일로 되면서 패턴이 많이 사라지기 때문으로 보인다. 그림 10은 실제 주식 변동에 따른 예측 그래프이다. 붉은색 그래프는 실제 주식 변동그래프이며 파란색은 해당 주식의 예측 그래프이다.

모든 그래프의 $|w|$는 3이며 2008년 2월 한 달에 대해 주가 예측을 수행하였다. (a)와 (b)의 경우는 주식 예측이 잘 수행된 경우로 주식이 반등되는 경우는 잘 찾지 못하지만 전체적으로 높은 예측율을 보였다. (c)와 (d)의 경우는 주식 예측이 잘 이루어지지 않은 경우로 예측값과 실제 주식 값이 반대로 나오는 경우도 보였다. 전반적으로 본 시스템은 높은 예측율을 보였다. 하지만 $|w|$값이 증가함에 따라 'C'값이 증가하기 때문에 예측율이 높아졌다고도 볼 수 있다.

표 8은 UB = 2.0이고 LB = 0.3일 때 특정 $|w|$의 양자화 문자 비율을 보여주고 있다. 표에 따르면 $|w|$가 증가하면 주식의 변동율은 감소하는 것을 확인할 수 있다. 즉, $|w|$가 증가하면 대부분의 양자화된 주식 데이터들이 'C'의 값을 가지기 때문에 높은 예측율을 가진다고 볼 수 있다. 이는 다른 말로 하면 예측율이 높다고 해서 실제로 투자의 용도로 사용하기가 어렵다는 말이 된다. 투자를 통하여 이득을 보기 위해서는 전체 예측율 보다는

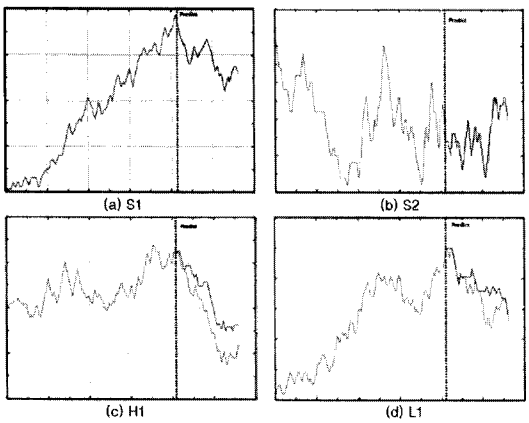

그림 10 예측과 실제주식 변동 그래프. 파란색 그래프는 실제주식 변동 그래프이며 검은색 그래프는 해당 주식의 예측 그래프이다. 모든 그래프의 윈도우 크기는 3이며 2008년 2월 한 달에 대해 주가 예측을 수행하였다. (a)와 (b)는 예측율이 높고 (c)와 (d)는 예측율이 낮다.

표 8 UB = 2.0이고 LB = 0.3일 때 특정 $|w|$의 양자화 문자 비율. 열은 양자화 문자들을 나타내며 행은 윈도우 크기를 나타낸다. $|w|$가 증가함에 따라 C의 비중이 높아지며 이는 $|w|$가 증가하면 하루의 등락이 전체 데이터에 미치는 영향이 작아짐을 의미한다.

	1	3	7	10	15	30
T	19.64	11.76	6.09	4.38	2.99	1.85
U	21.33	27.78	30.34	30.80	30.58	28.37
C	14.49	18.89	27.01	30.39	34.98	43.55
D	23.49	30.21	30.68	30.11	28.25	23.83
B	19.19	9.52	4.05	2.50	1.40	0.65
Z	1.85	1.84	1.83	1.82	1.80	1.75

표 9 KOSPI 데이터의 $|w|$별 P_i^*. 실제 투자자들은 등락 예측율인 P_i^*에 더욱 관심을 가진다.

$\|w\|$	P_i^+	P_i^-	P_i^*	P_i^0
1	7.82%	4.75%	12.57%	11.20%
2	10.24%	11.29%	21.53%	22.09%
3	12.63%	14.81%	27.44%	30.24%
4	21.03%	17.51%	38.54%	29.29%
5	20.32%	22.31%	42.63%	29.69%
6	21.23%	24.00%	45.23%	30.28%
7	22.23%	25.05%	47.28%	23.30%
8	25.54%	25.40%	50.94%	37.03%
9	32.23%	20.98%	53.21%	35.20%
10	26.21%	28.32%	54.53%	35.33%

실제 주가의 변동을 예측하는 등락 예측율이 더욱 중요하기 때문이다.

표 9는 실제 투자자들이 관심을 가질 수 있는 등락 예측율을 나타낸다. 등락 예측율은 예측율에서 변동이 없는 예측인 'C'의 예측율을 뺀 것으로 주식이 등락을 하는 경우의 예측율이다. 즉, 변동이 없는 경우를 제외하여 이익을 볼 수 있는 등락인 경우만 예측한 결과로 이는 $|w|$가 증가함에 따라 양자화 된 문자열에서 'C'의 개수가 증가함에 따른 예측율의 상승을 제외할 수 있는 장점이 있다.

4.4 다른 주식 예측 시스템과의 비교

본 논문에서 제시한 주식 예측 시스템이 얼마나 효율적인지는 다른 시스템들과의 비교를 통해서 알 수 있다. 그러나 많은 시스템들이 서로 다른 데이터들을 사용하기 때문에 단순히 비교하기에는 무리가 있다. 본 논문에서 제시한 주식 예측 시스템과 비교하기 위해서 우리는 신경 회로망을 이용한 예측 시스템과 비교를 하였다. 패턴 분석의 경우에는 대부분의 시스템이 특정 패턴이 주식 내에 발견될 때만 예측을 하는 방식을 채택하기 때문에 비교하기가 어렵기 때문이다.

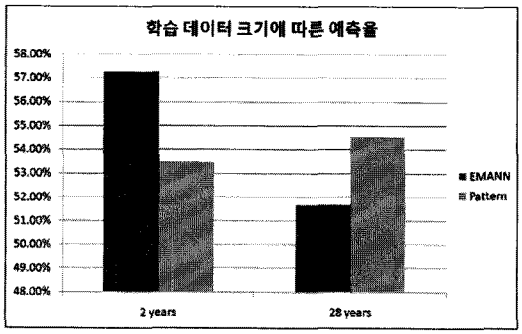

그림 11 학습 데이터의 크기에 따른 예측율 변화. 학습 데이터의 크기가 작을 때에는 신경회로망을 이용한 방식이 유리하고, 학습 데이터의 크기가 크면 패턴을 검색하는 시스템의 정확도가 높다.

본 논문에서 비교한 시스템은 [4]에서 제시한 알고리즘을 이용하였다. 이 시스템은 주식의 종가, 거래량, MACD (Moving Average Convergence Divergence) 값과 투자 심리선 값을 이용하기 때문에 단순히 본 논문에서 제시한 데이터를 사용하기에는 무리가 있었다. 그러므로 같은 시기에 대하여 각각의 방법에 대해서 특정 시기의 데이터를 이용하여 학습한 뒤 등락 예측율을 비교하였다.

결과는 그림 11과 같이 학습 데이터의 크기에 따라 두 시스템의 예측율이 달라지는 것을 확인할 수 있다. 즉, 충분히 큰 학습 데이터가 존재할 경우에는 본 시스템의 예측정확도가 더욱 증가하는 것을 확인할 수 있다.

5. 결론 및 향후 연구 과제

주식시장의 예측은 매우 어려우며 데이터의 방대함과 복잡성으로 인하여 주식시장의 예측은 거의 불가능한 것으로 여겨지고 있다. 주식시장을 예측을 하기 위해서는 변수를 줄이기 위해서 비슷한 속성의 주식들을 분류하는 것이 선행되어야 한다. 이런 주식 분류의 방법에는 여러 가지가 존재하지만 주식시장 예측에 사용하기 위해서는 과거의 주식 데이터들을 이용하여 실제 주식의 특성을 파악하는 방법이 필요하다. 주식이 상장될 때 등록된 주식 분류는 실제로 그 주식의 위치를 알려주기는 하지만 다른 주식들과의 상호 연동에 의한 주식의 고유한 특성은 나타내지 못하는 단점이 존재한다. 그러므로 과거의 주식 데이터를 이용하는 방법이 필요하게 되는데 이런 과거의 주식 데이터를 이용할 때에도 약간의 데이터 가공이 필요하게 된다. 주식은 실제 경제 흐름과 다른 주식의 등락에 매우 민감하기 때문에 이런 정보들이 모두 주식에 반영되어 매우 복잡하게 보이게 된다. 이런 불필요한 정보를 제거하게 되면 주식의 고유한 특성을 파악할 수 있게 되고 이를 이용하여 주식 분류에 활용 할 수 있다.

주식을 예측하기 위하여 패턴 패칭을 이용할 수 있다. 과거의 주식 데이터들을 이용하여 주식의 변화를 파악하여 현재 주식 변화와 가장 유사한 변화를 보이는 패턴을 찾은 다음 그 패턴을 이용하여 주식을 예측하는 방식이다. 이를 위하여 DNA 염기서열에서 유사한 영역을 찾는 기법인 정렬 기법을 응용 가능하며 실제로 많은 주식 데이터들 가운데서 빠르게 패턴을 검색할 수 있어서 유용한 방법으로 사용되고 있다.

주식 예측 시스템을 구현함으로서 본 논문에서 밝힌 점은 다음과 같다.

1. 주가에 정렬 기법을 적용하여 특정 시간대에 어느 주식과 유사한지를 검색하였으며 패턴을 통하여 주식 예측을 가능하도록 하였다.
2. 주식들에 전역 정렬을 적용하여 실제주식시장에서 제공하는 분류법으로 주식들의 상관관계를 모두 나타낼 수 없음을 보였다.
3. 주가 데이터를 가공한 뒤, 정렬 기법을 이용하여 주가의 변동을 예측가능 하며 투자자들이 관심이 많은 등락 예측율은 단기 예측은 12% 이상의 예측율을 보일 수 없으며, 장기 예측의 경우 54%의 예측율로 수렴함을 확인하였다.

위와 같은 점을 밝히기 위해서 본 논문은 우선 주식 데이터를 정렬 기법을 적용하기 적합한 형태로 변형하였다. 그리고 전역 정렬을 이용하여 유사한 주식들을 클러스터링 하여 실제주식 분류와는 다르게 유사한 주식들이 존재함을 보였으며 유사한 주식들의 과거 주식 데이터를 $|w|$를 이 용하여 특정주식의 미래를 거시적, 미시적으로 예측하여 예측도의 변화를 밝혀내어 예측 모델을 제안하였다. 이를 위하여 우리는 KOSPI 주식 데이터 28년 690개의 데이터를 수집하여 2008년 2월의 주식 데이터에 대하여 예측 실험을 수행하였다. 그 결과 S1, S2의 경우 각각 $|w|$ =5일 때 81.37%, 80.37% 의 예측율을 보임을 밝혀내었다. 하지만 예측율이 특정 $|w|$에서 현저히 떨어지는 현상을 보이는 것은 추후 실험을 통하여 좀 더 보강해야 할 것이다. 또한 장기 예측의 경우 높은 예측율을 보이지만 이는 실제 수익률과는 무관하다. 즉, 한 주식에 대한 예측율은 그 주식의 등락을 예측 하는 수치일 뿐, 그 예측에 따라 얻는 수익과는 별개의 문제이다. 왜냐하면 만일 주가가 하락한다고 예상하였으나 실제로 주가가 올라간다면 주식 투자에 따른 손실이 클 것이기 때문이다. 또한 반대로 주가가 상승한다고 예측하였으나 주가가 하락한다면 그 손실 또한 매우 크다. 즉, 이 논문에서 예측한 주가 예측율은 주가 예측에 따른 이익율과는 관계가 없다.

현재 주식시장은 등락을 거듭하여 예측이 매우 어렵

다. 하지만 주식은 장기적으로 보면 비슷한 상황이 반복되기 때문에 이전에 비슷한 상황이었던 IMF와 비교한다면 지금의 주식시장을 예측 가능할 것으로 기대된다. 즉, 본 시스템은 최근의 주식시장에 대한 충분한 데이터 즉 2008년 2월 이후의 국내 주식 데이터 등이 제공된다면 유사한 상황이었던 IMF때와의 패턴비교를 통하여 불안한 주식시장의 미래를 조금이나마 예측할 수 있을 것으로 기대된다. 또한 외국 나스닥과 같은 외국 데이터가 추가된다면 외국 주식과 국내 주식과의 상관관계를 통하여 좀 더 정확한 주식 예측이 가능할 것이다.

참 고 문 헌

[1] Eugene F Fama, "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, vol.25, no.2, pp.383-417, May 1970.

[2] Lin W., Orgun M., and Williams G, "An overview of temporal data mining," *ADM 02*, pp.83-83, 2002.

[3] Eugene F. Fama, "The behavior of stock-market prices," *The Journal of Business*, vol.38, no.1, pp.34-105, 1965.

[4] H. Y. Kim and S. G. Kim, "The Study of the Financial Index Prediction Using the Equalized Multi-layer Arithmetic Neural Network," *Journal of KSCI*, vol.8, no.3, pp.113-123, 1 2003. (in Korean)

[5] K. S. Cho, K. H. Lee and I. S. Yang. Expert System for Predicting the Stock Market Timing Using Candlesticks Chart. *Journal of KIISS*, vol.3, no.2, pp.57-70, Dec. 1997. (in Korean)

[6] Richi Nayak and Paul te Braak, "Temporal pattern matching for the prediction of stock prices," AIDM 2007, pp.95-103, 2007.

[7] Bartolozzi M., Leinweber D.B., and Thomas A.W. "Self-organized criticality and stock market dynamics: an empirical study," *Physica A: Statistical Mechanics and its Applications*, vol.350, no.2-4, pp.451-465, 2005.

[8] Gilmore Claire G., Lucey Brian M., and McManus Ginette M, "The dynamics of central european equity market comovements," *The Quarterly Review of Economics and Finance*, vol.48, no.3, pp.605-622, 2008.

[9] Z. R. Struzik, "Wavelet Methods in (Financial) Time-series Processing," *Physica A: Statistical Mechanics and its Applications*, vol.296 no.1-2, pp.307-319, June 2001.

[10] Gyozo Gidofalvi, "Using news articles to predict stock price movements," 2001.

[11] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences." *J Mol Biol*, vol.147, no.1, pp195-197, March 1981.

[12] Andrew T. Kwon, Holger H. Hoos, and Raymond Ng, "Inference of transcriptional regulation relationships from gene expression data," *Bioinformatics*, vol.19, no.8, pp.905-912, 2003.

[13] Sven Meyer zu Eissen and Benno Stein, "Intrinsic plagiarism detection," *Lecture Notes in Computer Science*, vol.3936, pp.565-569. Springer, 2006.

[14] CloneChecker : A Software Plagiarism Detector. http://ropas.snu.ac.kr/n/clonechecker/.

[15] Narayanan Shivakumar and H´ector Garc´ıa-Molina, "SCAM: A copy detection mechanism for digital documents," 1995.

[16] Geoff Whale, "Plague user manual(release1.2)," Department of Computer Science, University of New South Wales, 1989.

[17] Wise, "YAP3: Improved detection of similarities in computer program and other texts," SIGCSEB: SIGCSE Bulletin (ACM Special Interest Group on Computer Science Education), vol.28, 1996.

[18] David Gitchell and Nicholas Tran. "Sim: a utility for detecting similarity in computer programs," *In SIGCSE '99: The proceedings of the thirtieth SIGCSE technical symposium on Computer science education*, pp.266-270, 1999.

[19] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul Vit´anyi, "The similarity metric," *In SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp.863-872, Philadelphia, PA, USA, 2003.

[20] Mehmet M. Dalkilic, Wyatt T. Clark, James C. Costello, and Predrag Radivojac, "Using compression to identify classes of inauthentic texts," *In SIAM '06: Proceedings of the 2006 SIAM International Conference on Data Mining*, pp.604-608, 2006.

[21] Harry Eugene Stanley Rosario Nunzio Mantegna, "An introduction to econophysics," Cambridge University Press, 2000.

김 형 준

2008년 부산대학교 정보컴퓨터공학부(학사)
2010년 부산대학교 컴퓨터공학과(석사)
2010년~현재 Daum Communications.
관심분야는 그래프 이론, 언어 처리 등

조 환 규

정보과학회논문지 : 시스템 및 이론
제 37 권 제 2 호 참조