POINT SET의 계산기하학적 속성을 응용한 주식 클러스터링

김영준
부산대학교 정보컴퓨터공학부
Kim Hyong-jun
Pusan National University
hjkim83@pearl.cs.pusan.ac.kr

요 약

주어 예측 모델에는 많은 변수들이 존재한다. 특히 예측 결과가 주어 변동에 영향이 미친다는 점에서 매우 복잡한 모델이다. 이런 주식 시스템에서 주가 예측을 하기 위해서는 변수를 최대한 줄이고 비교적 간단한 모델을 사용하는 작업이 필요하다. 그러나 주식 시장에서 제공하고 있는 분류 시스템은 상장되는 기업의 명칭만을 이용한 분류로서 실제 주식시장의 상호작용을 전혀 고려되지 않는다. 그러므로 실제 주식시장에서 일어나는 움직임에 상호작용까지 모두 고려되는 주식 분류 시스템이 필요하다. 본 보고서에서는 각 주식 종류를 기하 공간에 배치하여 두 주식들의 관계를 살펴보고 이를 이용하여 주식 클러스터링을 한다. 그 결과를 살펴보고 정단점을 살펴본다.

주제어: 주가예측, 주식 클러스터링, Computational Geometry

1 연구 동기

주가 예측은 많은 변동들이 존재하기 때문에 매우 어렵다. 이런 변수들을 제한하기 위해서는 비교적 간단한 속성을 가진 주식들을 묶은 다음 주기를 예측하는 방법이 필요하다. 이를 위해서 주가들을 분류하는 작업이 필요하다. 대부분의 주식시스템에서는 기준적인 주식 분류를 제공하고 있다. 하지만 이 주식 분류는 주식을 상장하는 기업의 특성으로 분류한 것으로 실제 주기가 서로 영향을 주고 받는 것들에 대해서는 고려하지 않는다. 예를 들면 기업 A가 식용유 관련 종목으로 분류되어 있고 기업 B가 운송으로 분류되어 있다면 하면 대한민국의 경우 식유를 외국에서 전량 수입하기 때문에 B회사의 영향을 크게 받게 된다. 하지만 기준의 주식 분류 시스템에서는 이런 맥락이 전혀 드러나지 않기 때문에 주식 클러스터링에 큰 어려움을 겪게 된다. 그러므로 이런 주가들의 속거짐 상호작용까지도 고려되어야 하는 주식 분류 시스템이 필요하다.

주식시스템의 많은 변수로 인하여 기존의 주가 분석은 복잡한 특성을 주안점을 두고 연구가 진행되었다. 기본적으로 주식 시스템은 그 주가 시스템이 속해있는 국가의 영향을 많이 받는다. 몇 가지 주가 분석 연구는 그 나라의 주가 특성을 이용하여 예측하거나 분류하는 방식으로 이루어졌다[4, 5, 7]. 몇몇 다른 연구들은 주식시스템을 다른 연구 분야로 접목하여 이루어졌다. 현재 가장 많이 진행된 연구 로는 Wavelet 변환을 주식 시스템에 응용한 것이다[10]. 이 방식은 현재도 계속해서 연구가 진행되어지고 있다[8, 9]. 주식 시스템은 자신이 예상한 결과가 실제로 주식 시스템에 영향을 주기 때문에 예측이 더욱 난해하다. 이런 점을 감안하여 주식과 관련된 뉴스 기사들을 이용하여 주식을 단기 예측 하려는 방법도 제시되었다[6].

본 보고서에서는 주식을 계산기하학 분야에 접목하여 빠르게 유사한 주식추보들을 검출해 내는 방법을 제안하고 그 결과를 바탕으로 장단점을 분석하여 본다.

2 사전 연구

주가를 예측하는 문제는 금융 분야에서 매우 중요한 관심사이 되어왔다. 주식 시장은 시장 환경의 변화에 매우 민감하게 변화하며 이런 주식 시세를 예측을 통해 주식 투자자로부터 이익을 창출하기 위해서는 주식을 사고 파는 시점을 결정하는 문제가 매우 중요하다. 이런 주식을 예측하기 위해서 이루어진 연구들 중 컴퓨터 사이언스 분야에서는 크게 인공지능을 이용한 방법이 주를 이루고 있다.
일반적으로, 주식 투자에서 필요한 중요한 결정사항은 종목 선정과 매매 타이밍 결정이다. 종목 선정 은 앞으로 좋은 수익을 가져다 줄 수 있는 종목을 선택하는 것이고, 매매 타이밍 결정은 주가가 동락을 반복할 때 가장 높은 수익을 얻기 위한 매매 시점을 결정하는 문제이다.

이런 여러가지 투자 결정 사항 중에서 컴퓨터 사이언스 분야에서는 매매 타이밍 결정에 주로 초점을 맞추어서 주가 예측을 수행하고 있다.

주식시장 예측을 위해 적용된 인공지능 방법들은 크게 신경망(neural network), 퍼지 이론(fuzzy logic), 전문가 시스템(expert systems)로 분류될 수 있다[1].

특히 전문가 시스템을 이용한 방식은 75퍼센트 정도의 정확도를 보여 주었다. 하지만 이런 전문가 시스템의 가장 큰 단점은 자동 학습 능력이 없기 때문에 스스로 패턴과 규칙을 만들지 못하여 새로운 변화에 대해서는 취약하다는 단점이 있다.

주가는 일정 펌프가 없이 동락을 거듭하는 것을 처리하기 보다는 실제로 특정한 패턴을 가지고 동락이 이루어진다고 가정하여 주식을 예측하는 방법도 있다[3]. 실제로 주식시장의 예측에는 일정한 패턴이 있다고 가정하고 이루어지고 있다. 그런 패턴을 찾아내어 예측에 적용을 한다면 매우 높은 수익률을 낼 수 있을 것으로 기대된다.

3 계산기하학을 응용한 주식 클러스터링

이 섹션에서는 우선 주식 정보를 양자화 하여 일정 기초로 나타내는 과정에 대해서 설명한다. 그리고 나서 양자화된 주식 정보를 이용하여 계산 기하학 영역으로 매핑하여 두 주식의 유사성을 판단하는 방법에 대해 설명한다.

3.1 주식정보의 양자화

주식을 클러스터링 하기 위해서는 주가의 양자화와 두 유사한 주식의 유사도를 수치화 하는 작업이 필요하다. 이 섹션에서는 우선 생물 정보학을 이용한 주식 분류 방법에 대해 설명한다.

주식은 시간별로 주가를 가지고 있다. 그러므로 하나의 주식에서 하나의 주가를 다음과 같이 정의 한다.

**Definition 3.1** 하나의 주식 $S$에 대해, $i$번째 날의 주가는 $S_i$로 나타낸다.

매일의 주가는 매우 큰 변동폭을 가지고 있기 때문에 이를 이용해서 분석하기가 매우 어렵다. 이런 문제로 우리는 주가 계산을 위해 슬라이딩 윈도우를 사용한다.

**Definition 3.2** 주식 $S$에 대해, $i$번째 날의 원도우크기가 $w$인 평균주가 $SW_i(w)$는 다음과 같이 정의된다:

$$ SW_i(w) = \sum_{k=1-w}^{k} S_k $$

정의 3.2를 이용하여 우리는 주가를 단기간 또는 장기적으로 분석이 가능하다. 또한 두 주식의 유사도를 판별할 수라도 매우 유용하다. 이에 대해서는 나중에 설명한다.

**Definition 3.3** 주가 원도우 $SW_i(w)$에 대해 $i$번째 날의 양자화된 주가 원도우 $Q_i(w)$는 다음과 같이 정의된다:

$$ Q_i(w) = 100 \frac{SW_i(w) - SW_{i-1}(w)}{SW_{i-1}(w)} $$
주가를 문자열로 변환하기 위해서는 우선 주가를 양자화 할 필요가 있다. 정의 3.3를 이용하여 우리가 이전 완도우의 차이의 비율을 나타내는 $Q_i^{(w)}$를 얻을 수 있다. $Q_i^{(w)}$를 이용하여 양자화 과정을 적용하여 문자열로 변환할 수 있다. 

테이블 1는 각 비율에 상응하는 문자를 나타낸다.

<table>
<thead>
<tr>
<th>$QC_i^{(w)}$</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>$UB &lt; Q_i^{(w)}$</td>
</tr>
<tr>
<td>U</td>
<td>$LB &lt; Q_i^{(w)} \leq UB$</td>
</tr>
<tr>
<td>C</td>
<td>$-LB &lt; Q_i^{(w)} \leq LB$</td>
</tr>
<tr>
<td>D</td>
<td>$-UB &lt; Q_i^{(w)} \leq -LB$</td>
</tr>
<tr>
<td>B</td>
<td>$Q_i^{(w)} \leq UB$</td>
</tr>
<tr>
<td>Z</td>
<td>$S_i가 존재하지 않음</td>
</tr>
</tbody>
</table>

$UB$는 양자화 값의 상위 제한값을 나타내며 $LB$는 하위 제한값을 나타낸다. 이 두 변수를 조정하여 우리는 결과를 최적화할 수 있다.

주가는 몇가지 이유로 몇몇 날에는 존재하지 않는다. 예를 들면 특정 일자에는 아직 주식이 상장되지 않았다고 해당 기업이 부도 또는 다른 이유로 인하여 상장을 중지한경우가 있다. 이런 날들의 $QC_i^{(w)}$는 \(Z\)로 나타내고 아무런 계산도 하지 않는다.

**Definition 3.4** 주식 $a$에 대하여, 주식 $a$의 완도우 크기가 $w$인 양자화 집합 $QS_a^{(w)}$는 다음과 같이 정의된다.

\[
QS_a^{(w)} = \{QC_i^{(w)} | i \in N\}
\]

$QC_i^{(w)}$는 특정 일자에 대한 문자만 가지고 있기 때문에 시간순의 문자열을 얻기 위해서는 모든 시간대를 다 모아야한다. 이렇게 시간순으로 모든 문자열을 이용하여 우리는 Global Alignment 및 주가 분석을 할 수 있다. 또한 주식 클러스터링도 이를 이용하여 할 수 있다.

**Definition 3.5** 두개의 주가 양자화 집합 $QS_a^{(w)}$와 $QS_b^{(w)}$에 대하여 Global Alignment 값 $GA_w(a, b)$는 다음과 같이 정의된다.

\[
GA_w(a, b) = GAS(QS_a^{(w)}, QS_b^{(w)})
\]

Global Alignment 값 $GA_w(a, b)$는 하나의 파라미터만 가지고 있다. 이 파라미터를 이용하여 주식의 완도우 크기를 조정할 수 있다. 이는 주식의 유사도를 단기간 즉, 짧은 시간내에 주가의 변화량을 가지고 분석할 수도 있으며 장기적 즉, 긴 시간동안 변한 주가의 변화량을 가지고 분석할 수도 있다.

### 3.2 주식정보의 기여학적 요소 변화

임의의 두 주식 $a, b$가 있을때, 두 주식의 $QS_a^{(1)}, QS_b^{(1)}$값을 기여학적 요소로 변환한다.

기본적으로 $QS_a^{(1)}$는 문자열로 이루어져 있다. 각 문자를 숫자로 변환하여 좌표를 계산한다.

표 2는 $QS_a^{(1)}$의 각 문자에 따른 좌표 보정값을 나타내고 있다. 좌표값은 0에서 시작하여 문자에 따라 값을 보정하여 새로운 좌표를 만드는데 사용한다.

임의의 두 주식 $a, b$에 대해 $a$는 $X$좌표값으로 매핑하고, $b$는 $Y$좌표값으로 매핑한다. 이를 통해 두 $QS_a^{(1)}$의 모든 문자를 탐색하면 2차원 공간에 매핑되는 포인트셋이 만들어진다. 이 기하학 오브젝트를 이용하여 두 주식의 연관성을 조사한다.
표 2: 각 문자에 해당하는 좌표 보정값. 상용하는 문자에 해당하는 숫자를 좌표값에 더하여 새로운 좌표를 구한다.

<table>
<thead>
<tr>
<th>Character</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>+2</td>
</tr>
<tr>
<td>U</td>
<td>+1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>-1</td>
</tr>
<tr>
<td>B</td>
<td>-2</td>
</tr>
</tbody>
</table>

기본적으로 \( a, b \)가 완전히 일치하면 포인트셋은 \( y = x \) 그래프에 모두 존재한다. 만일 두 주식의 문자가 불일치하면 \( y = x \)그래프에서 벗어나게 된다. 이 속성을 이용하여 \( y = x \) line에서 가장 멀리 벗어져 있는 포인트까지의 거리를 두 주식의 유사도 정도로 나타낸다.

그런데 \( y = x \)에서 가장 멀리 벗어진 포인트를 계산할 때 모든 포인트셋에 대해 검사를 할 경우 계산량이 매우 많아지게 된다. 그러므로 포인트셋의 특성 중 하나인 입의의 점에 대해 가장 멀은 포인트셋의 컨텍스 홀에 존재한다는 것을 이용하여 포인트셋의 컨텍스 포인트들에 대해서만 검사를 수행한다.

그런 식으로 검사를 하였을 경우에 특이한점은 두 주식이 모두 증가하지만 증가량이 풀려서 불일치할 경우의 차이값이 두 주식 중 하나는 증가하고 하나는 감소하는 경우 보다 작다는 점이다. 이는 두 차 증가나 감소의 경우에는 \( y = x \) 그래프에서 조금씩 바꾸는데 비해 두 주식 중 하나는 증가하고 하나는 감소하는 경우는 \( y = -x \) 그래프에 가깝게 좌표가 움직이기 때문이다.

4 실험결과

그림 1: 계산 기하 변환에서 가장 높은 점수를 받은 주식 쌍의 포인트 셋의 분포 빌간 선은 \( y = x \)그래프를 나타낸다.

그림 1은 계산 기하 변환에서 가장 높은 점수를 받은 주식양의 포인트 셋 분포이다. 그래프가 확대 되었으나, 비교적 \( y = x \)에 가깝게 분포되어 있음을 확인할 수 있다.

그림 2는 산 기하 변환에서 가장 높은 점수를 받은 주식양의 실제 주식 변동률을 나타낸다. 큰 변동률은 민감하지 못하나 대부분의 변동률에서 비슷한 양상을 보이는 것을 확인할 수 있다.
그림 2: 계산 기하 변화에서 가장 높은 점수를 받은 주식 채도의 실제 주식 변동률. x축은 시간의 변화를 나타내며, y축은 전날 주가 대비 변화율을 나타낸다.

그림 3: 계산 기하 변화에서 가장 낮은 점수를 받은 주식 채도의 포인트 셋의 분포 범위 선은 y = x 그래프를 나타낸다.

그림 3은 계산 기하 변화에서 가장 낮은 점수를 받은 주식 채도의 포인트 셋 분포이다. 원쪽 주식이 오름에 따라 다른 주식은 계속해서 하락하는 y = −x에 가까운 분포를 보이고 있다.

그림 4는 계산 기하 변화에서 가장 낮은 점수를 받은 주식 채도의 실제 주식 변동률이다. 원쪽 주식이 오름에 따라 반대쪽 주식은 완전 반대의 움직임을 보이고 있다.

위의 그림들과 같이 계산 기하 변화를 통한 주식 유사도 분석은 매우 일치함을 확인할 수 있다. 이는 y = x 그래프에서 가장 면 점이 되기 위해서는 계속해서 두 주식이 일치하지 않아야 하기 때문이다.

테이블 3은 상위 10개의 계산 기하 변화 쌍이다. 거리는 y = x 그래프에서 가장 면 점의 거리를 나타내는데 이 거리가 가까울 수록 두 주식 쌍이 풀더 유사함을 나타낸다. 실제 분류와 비교하면 제조업의 경우가 많이 존재하며 대부분이 잘 일치함을 확인할 수 있다. 한가지 제외있는점은 비슷한 거리를 가지는 쌍들이 존재하는데 이는 문자열에 따른 최표의 이동이 1단위로 일어나기 때문이다. 이를 이용하여, 일부 거리를 벗어나는 주식 쌍들은 서로 유사도가 없는 것으로 판단하여 Global Alignment를 수행하지 않음
그림 4: 계산 기하 변환에서 가장 낮은 점수를 받은 주식 쌍의 실제 주식 변동률. x축은 시간의 변화를 나타내며, y축은 전날 주가 대비 변화율을 나타낸다.

표 3: 계산 기하 변환 쌍에서 가장 높은 10개의 주식 쌍. 대부분의 쌍이 같은 대분류를 가지는 것을 확인할 수 있다.

<table>
<thead>
<tr>
<th>주식A</th>
<th>주식B</th>
<th>거리</th>
<th>대분류A</th>
<th>대분류B</th>
</tr>
</thead>
<tbody>
<tr>
<td>전왕홀딩스</td>
<td>S&amp;TC</td>
<td>3.54</td>
<td>서비스</td>
<td>제조업</td>
</tr>
<tr>
<td>다이아씨</td>
<td>CJ</td>
<td>4.24</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>신성ENG</td>
<td>S&amp;TC</td>
<td>4.24</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>한국화장</td>
<td>피시스</td>
<td>4.95</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>CJ</td>
<td>S&amp;TC</td>
<td>5.66</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>대상홀딩스</td>
<td>엠파인</td>
<td>5.66</td>
<td>서비스</td>
<td>전철업</td>
</tr>
<tr>
<td>백산</td>
<td>S&amp;TC</td>
<td>5.66</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>삼성카우프로</td>
<td>한국프랜</td>
<td>5.66</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>삼화전기</td>
<td>우선</td>
<td>5.66</td>
<td>제조업</td>
<td>제조업</td>
</tr>
<tr>
<td>국보</td>
<td>우선</td>
<td>5.66</td>
<td>제조업</td>
<td>금융</td>
</tr>
</tbody>
</table>

면 주식의 유사도를 판별하는 시간을 대폭 축소할 수 있을것으로 기대된다.

한편 이 방법의 문제점으로는 위의 리스트에서 여러번 등장하는 'S&TC' 주식에서 나타난다. 실제 주식 데이터에서는 'S&TC'의 주식은 실제 상장되지 않아データ가 별로 존재하지 않는다. 이로 인하여 'S&TC'는 포인트셋이 별로 존재하지 않아서 전체의 이동이 많이 일어나지 않는 바람에 매우 높은 순위를 차지하고 있다. 차후에는 이를 개선하여 포인트셋의 크기에 따른 스케일 변화가 필요할 것으로 판단되어 진다.

5 결론
주식의 변동은 많은 변수로 인하여 예측하기가 매우 어렵다. 그래서 주식의 변동을 예측하기 위해서는 변수를 제한하는 작업이 선행되어야 한다. 우리는 이 보고서에서 변수를 제한하기 위하여 주식의 클러스터링을 하였다. 비슷한 속성의 주식들을 같은 클러스터로 묶음으로서 우리는 특성을 이용하여 주식의 변동을 예측할 수 있을것으로 기대된다. 지금까지 주식 클러스터링을 위해서 생물 정보학을 접목하여 주
References


