Earth Mover's Distance に基づくクラスタリングによる株価の変動点の検出

2018SS050 大部智也
指導教員：小市俊悟

1　はじめに

新型コロナウイルスの流行を原因とする将来の不安のために資産を形成したいと考える人が多くなってきているように感じる．特に以前にも増して，株式投資やビットコインといった仮想通貨への投資，積み立て NISA のような投資信託など，「投資」に対して人々の関心が高まっている．しかし，実際に投資を行うにあたっては，十分な知識や情報が必要であるため，投資に興味はあるが手を出すことができない人が多いのもまた事実である．本研究では，「投資」の中でも，株式投資，特に，株価の変動によるキャピタルゲイン獲得のために，どのタイミングで株式投資を行うべきかに関心があり，株価が上昇から下降，下降から上昇へと変化するような，いわゆる変動点を検出することを目的とする．

2　使用するデータ

株式投資メモ [1] というサイトからダウンロードした．大手家電量販店の 2016 年から 2020 年までの過去 5 年分のデータを使用する．株価は，すべて各取引日における終値とする．株価の値動きには小幅なところもあれば，大きく変動しているところもある．また，上昇傾向にあるところや下降傾向のところもある．本研究では，このような短期間のデータの質の変化を投資のタイミングを計るための情報と捉える．

3　研究手法

本研究では，離散確率分布間の距離の一種である Earth Mover's Distance(EMD) を用いる [2]．特に離散的なヒストグラムを対象にし，次のように定義される EMD を利用する．ヒストグラムの横軸は，n 個の離散値であり，$[n]$ は $[n] = \{1, 2, \ldots, n\}$ を表す．各 $i \in [n]$ における値がそれぞれ a_i, b_i であるようなヒストグラム $a = \{a_i\}_{i=1}^n$ と $b = \{b_i\}_{i=1}^n$ を考え，a と b の距離 $D(a, b)$ を次の最適化問題の最適値として定める．

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i,j \in [n]} |a_i - b_j| x_{ij} \\
\text{subject to} \quad & \sum_{j \in [n]} x_{ij} = 1 \quad (i \in [n]) \\
& \sum_{i \in [n]} x_{ij} = 1 \quad (j \in [n]) \\
& 0 \leq x_{ij} \leq 1 \quad (i, j \in [n])
\end{aligned}
$$

定義からわかるように，EMD では数値の発生順に意味はなく，数値の集合としての違いを定量化する．

4　株価差分データの作成とクラスタリング

EMD の特性も踏まえて, 次のような差分データを作成する．時刻を $[T] = \{1, \ldots, T\}$ とし, 時刻 $i \in [T]$ の株価を p_i と表すとき, $q_i = p_{i+1} - p_i$ を各 $i \in [T-1]$ について求める．このような差分データ q_i について, 連続する m 個を集めた $S_i = \{q_i, q_{i+1}, \ldots, q_{i+m-1}\}$ $(i \in [T-m])$ を対象に EMD の計算とそれに基づくクラスタリングを行う．ただし，各 $i \in [T-m]$ に対して得られる S_i の中には S_i と S_{i+1} のように重複が大きいものも含まれるので，クラスタリングがうまく行えない．そこで，n 日間隔で得られる S_i，すなわち，$\mathcal{S}_n = \{S_i \mid i = kn+1, k = 0, 1, \ldots, \lfloor (T-m-1)/n \rfloor\}$ をクラスタリングの対象とする．$\mathcal{S}_n$ についてクラスタリングすれば，各 $S_i \in \mathcal{S}_n$ がどのクラスターに属しているかに応じて，時刻 $i = kn+1$ の方も分類できる．異なるクラスターに属す S_i は，傾向が異なるデータであると考えられ，分類先が切り替わる時刻 i が変動点となる．上昇，下降，停滞のみならず，上昇の中でも急激な上昇などを表すようなクラスターに分類されることを期待する．年度や m と n を変えて分析を行う．

5　クラスタリングの結果

5.1　2016 年の $m = 5, n = 3$ のとき

コンプリート法で得られたデンドログラムは図 1 で，図 2 は得られたクラスターを色で株価のグラフに示したものである．閾値を 90 としたとき，7 つのクラスターに分かれ，左から順に，

- 日々の変動が大きく全体では急な下降
- 日々の変動が大きく全体では急な上昇
- 日々の変動は穏やかであり全体では緩やかな上昇
- 日々の変動は大きいが全体では緩やかな上昇
- 日々の変動は穏やかであり全体では緩やかな下降
- 日々の変動は大きいが全体としては停滞気味のもの，
- 日々の変動に特異なものが含まれており全体では下降

というようなクラスターとなった．各クラスターが上昇，下降または停滞のいずれかのみに占有されることをもって，分類が成功していると判断するのであれば，得られた 7 つのクラスターはいずれかで占有されており，概ね狙い通りの結果を得ることができたと言える．

5.2　2016 年の $m = 7, n = 4$ のとき

ウォード法で得られたデンドログラムは図 3 で，得られたクラスターを色で示したものが図 4 である．閾値を 90

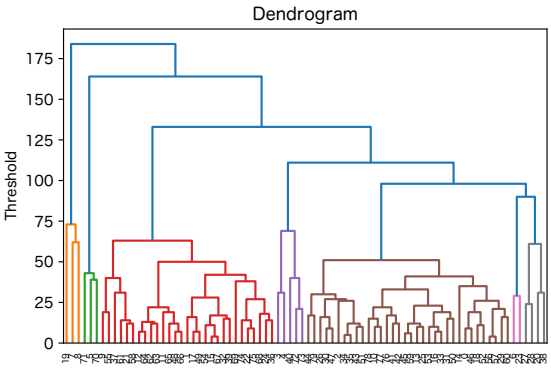

図 1　コンプリート法を適用した 2016 年の株価データに対して $m = 5$, $n = 3$ のときに得られるデンドログラム

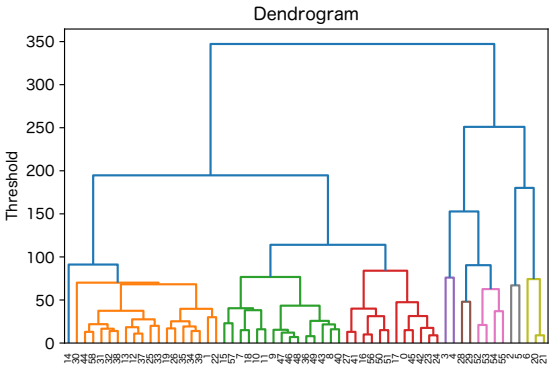

図 3　ウォード法を適用した 2016 年の株価データに対して $m = 7$, $n = 4$ のときに得られるデンドログラム

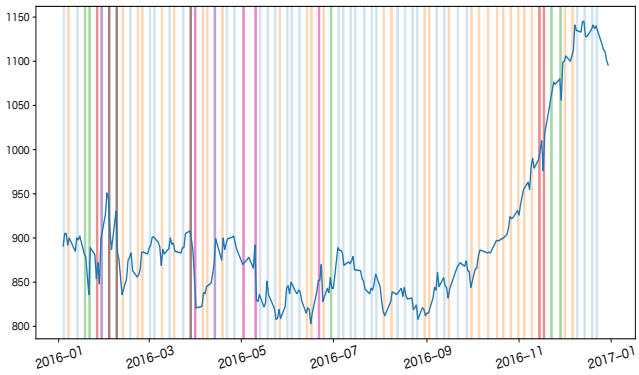

図 2　コンプリート法を適用した 2016 年の株価データに対して $m = 5$, $n = 3$ のときに得られるグラフ

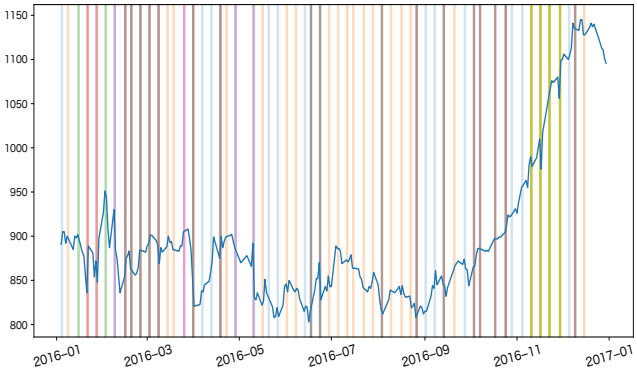

図 4　ウォード法を適用した 2016 年の株価データに対して $m = 7$, $n = 4$ のときに得られるグラフ

としたとき，9 つのクラスターに分かれ，左から順に，

- 日々の変動は穏やかであり全体として急な下降
- 日々の変動は穏やかで全体として緩やかな下降
- 日々の変動は穏やかで全体として緩やかな上昇
- 日々の変動は穏やかで全体として急な上昇
- 日々の上下変動が大きく全体として上昇
- 日々の変動に特異なものが含まれており全体として上昇
- 日々の変動が大きく全体として急な上昇
- 日々の変動が大きく全体として急な下降
- 日々の変動に特異なものが含まれており全体として急な下降

というようなクラスターとなった．同じデータに対して，コンプリート法で得たクラスターには上昇と下降が混在するものがあったが，それもないので $m = 7, n = 4$ のときはコンプリート法よりウォード法の方が結果が良い．

6　おわりに

EMD による距離に基づくクラスタリングによって，株価のトレンドを一定程度捉えることができた．得られたクラスターの中には意味付けが難しいものもあったが，株価の変動には突発的，例外的変動も含まれることを考慮すれば，それなりの規模のクラスターに対して，上昇や下降といった傾向をあてはめることができたことは，本研究の研究手法に期待されることを，ある程度実現することができたと考える．

本研究では，EMD とクラスタリングを利用したアプローチで株価の推移を分析したが，企業の財務諸表や新聞インターネットからの情報を利用したりと，複数のアプローチを組み合わせることでより精度を上げることができると考える．

参考文献

[1] 株式投資メモ https://kabuoji3.com (アクセス日:2021 年 7 月 26 日)

[2] 小池めぐみ：『特殊な Earth Mover's Distance を用いた通信異常検知』．2019 年度南山大学理工学部システム数理学科卒業研究, 2020.