平成 30 年度 公立はこだて未来大学卒業論文

個別銘柄の連動類似度を用いた株価予測

鈴木　雄士

複雑系知能学科　　1015100

指導教員　　竹之内　高志

提出日 平成 31 年 1 月 29 日

Stock Prices Prediction Using Similarity of Stock Prices

by

Yuji Suzuki

BA Thesis at Future University Hakodate, 2019

Advisor:　Takashi Takenouchi

Department of Complex and Intelligent Systems
Future University Hakodate
January 29, 2019

Abstract– A lot of researches have been proposed to predict stock prices using mathematic technique, the computer simulation and machine learning. When a brand in a sector of stock market fluctuates, brands in the same sector are affected and behave in a similar manner. In this research, I propose a method to improve stock prices prediction by incorporating information of brands being similar with the targeted brand.

Keywords: clustering, prediction, stock price, brand

概 要: 現在までに数理的手法やコンピュータ・シミュレーションをはじめとする様々な手法で株価を予測する研究がされてきた. その中で, 機械学習を用いることで株価の予測精度を向上させている手法がある. 株式市場ではある銘柄が変動した際に, その影響を受けて同じセクターやテーマの銘柄が似たような値動きをする事がある. 本研究では, 対象銘柄と類似度の高いデータを予測モデルに組み込むことで, 対象銘柄の株価予測精度を向上させることを目指す.

キーワード: クラスタリング, 予測, 株価, 銘柄

目 次

第1章　序論

1.1　背景

株式市場の動きを予測する研究は, 古くから行われてきた. 数理的手法やコンピュータ・シミュレーションを金融データに適用し, トレーディングや投資決定などに活用する試みが盛んである [1]. これらの研究では, 過去の時系列データから将来の値を予測するテクニカル分析と, 財務諸表や経済指標などから値を予測するファンダメンタル分析が主流となっている. しかし, 株価などの金融時系列データはほとんどランダムウォークの動きをするため予測は難しい [2]. また, 株式取引を行う際には専門的な知識を必要とする上, 利益を上げにくいという問題がある.

1.2　研究目標

株価を予測する際には現在の相場と似た過去の相場を参考にすることがある. また, ある銘柄が変動した際に, その影響を受けて同じセクターやテーマの銘柄が似たような値動きをする事がある. これを取引に利用する手法はトレーダーにはよく知られているがそれぞれのトレーダーの主観によって銘柄の類似度の判断が異なる問題がある. そこで, 機械学習を用いて対象銘柄と類似度が高いデータを特徴量として予測モデルに組み込むことで, 対象銘柄の株価予測の精度を上げ, 株式取引に役立たせることを目的とする.

1.3　構成

本論文の構成は次のようになっている.

- 第 2 章：関連研究
- 第 3 章：本研究で使用する要素技術
- 第 4 章：提案手法
- 第 5 章：実験
- 第 6 章：結論

第2章　関連研究

本章では, 本研究に関連する研究について述べる.

2.1　株価変動パターンの類似性を用いた株価予測 [3]

この研究では現在の相場と似た過去の相場を参考にする予測プロセスを定量的に再現し, 予測に活用する手法の提案を行なっている. 現在の株価変動が過去のどの時点の株価変動に近いかを定量的に抽出する. 抽出には, 系列データ間の類似度を測るために日次の株価変動を月初の値との比で表現した月間の株価変動に対して DTW を適用する Indexation DTW(IDTW) による手法を使用した. 結果として IDTW を用いて抽出した現在の株価変動に近い, 過去の株価変動パターンが予測に有効な特徴量であることを実証した.

2.2　関連銘柄同定のための時系列データ類似度尺度の提案 [4]

この研究ではある銘柄の株価が変動したときに, その影響を受けて遅れて株価が変動する銘柄を発見する事が目的である. そこで, 株価等の時系列データの類似度を測定するために Derivative Dynamic Time Warping(DDTW) による手法を提案している. 結果として同じような業種であれば高い類似度が得られることが確認できた. 一方, 異なる業種でも類似度が高いと判断される銘柄も見つかった.

2.3　クラスタ分析におけるクラスタ数自動決定法の比較 [5]

この研究ではクラスタ分析における様々なクラスタ数自動決定法を概観し, その一部についてモンテカルロシミュレーションで比較し, 評価を行なっている. 結果として幾つかのクラスタ数自動決定法を色々な方法で試してみたがシミュレーションでは良い結果を残せてはいても, それ単独でクラスタ数を決定するには不安が残った. 様々なクラスタ数自動決定法があり, それぞれに特徴があるので 1 つの方法で決めようとするのではなくそれらを組み合わせることで, かなりの確率で正しいクラスタ数を決定できるようになると考えられる.

2.4 テクニカル分析に基づくペアトレードの有効性と日本の株式市場の効率性 [9]

この研究ではテクニカル指標の 1 つである移動平均乖離率とペアトレードを組み合わせるテクニカル分析に基づくペアトレードを提案している. 結果として良好な成績かつ運用リスクが低いことが示され, ペアトレードもテクニカル分析も有効性があることが示唆された. 今後の課題としては本研究では移動平均乖離率とペアトレードと組み合わせて分析を行ったが, 他のテクニカル指標とペアトレードとの組み合わせを行うことで新たな運用手法の発見が考えられる.

2.5 機械学習を用いたポートフォリオの最適化 [11]

この研究では, 2 つの手法により市場から超過収益を得るための手法を提案している. 第一の手法は、市場におけるモメンタムの抽出フィルタをモデル化する手法である. 第二の手法は、AntiCor という個別株式の平均値回帰の性質・時間的なずれを検出することにより超過収益を得るアルゴリズムの改良版を実装している. 上記の手法を米国株式の過去１２年間の５００程度の銘柄を対象とし、その有効性を検証している. 結果として AntiCor フィルターはナスダック 100 銘柄から選択する場合, 取引コストが 0.1%以内で強いダウントレンドの市場でなければ超過収益が挙げられる可能性が高いと述べている. また, モーメンタムフィルターは選択候補が 50 銘柄以上になれば有効性が高くなり, 市場が緩やかな下降トレンドであっても超過収益を得ることが可能であると述べている.

第3章　要素技術

本章では, 実験で使用する要素技術について述べる.

3.1　Dynamic Time Warping(DTW) 距離

現在の株価変動に近い過去の株価変動を抽出する際に, 時系列データ間の類似度 (距離) を計測する必要がある. 時系列データ間の類似度を測るため, これまで多くの尺度が提案されてきた. よく使用される単純な尺度としてユークリッド距離が挙げられる. ただし単純なユークリッド距離では 2 つの時系列の長さが異なる場合に計算ができないという問題点がある. DTW 距離はユークリッド距離における当該問題点を克服する尺度である. すなわち長さの異なる時系列データの比較を可能とし, 2 つの時系列に対して時間軸を伸縮させながら最適な対応付けを行う距離尺度である.

DTW 距離とは, 二つの時系列データの各点の距離を総当りで比較し, 最短となるパスでかかる総コストのことである. 例として, 長さ m のデータ $X = (x_1, \cdots, x_m)$ と長さ n のデータ $Y = (y_1, \cdots, y_n)$ を考える. この 2 つのデータの DTW 距離は以下の手順で求められる.

1. $m \times n$ の大きさの類似度行列 D を作る.
2. $D_{0,0} = 0$，それ以外の類似度行列の値に ∞ を代入し初期化する.
3. 以下の式により $i = 1, 2, \cdots, m$ ； $j = 1, 2, \cdots, n$ の全ての組み合わせの $D_{i,j}$ を計算する.
   $D_{i,j} = d(x_i, y_i) + \min(D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1})$
   d は距離測度であり，マンハッタン距離 $d(x_i, y_i) = |x_i - y_i|$ を使用する

これによって得られた類似度行列 $D_{m-1,n-1}$ の値がデータ X とデータ Y の DTW 距離となる.

3.2　Derivative Dynamic Time Warping(DDTW) 距離

DTW では, 一方の時系列のある時点と, もう一方の時系列の複数の時点を対応付けて距離を算出することができる. ただし, データの個々の値を基に時系列データの距離を定義しており, データの個々の値を基に時系列データ間の距離を定義しており, 距離が近い時系列データ同士でも, その形状は必ずしも類似しているわけではない. 例えば, 時系列データのある区間で値が上昇し, もう一方の時系列データでは下降しているような場合でも, その区間の値そのものに差がなければ, 類似していると判断されてしまう. よって, DTW などの

値に基づく距離尺度は, 形状に着目して時系列データの類似性を測定したい場合は, 適切な距離/類似度の尺度とはならない. そこで小沢ら [4] は, 時系列の形状に着目した Derivative Dynamic Time Warping(DDTW) を提案している. 形状の類似性を考慮するため, 以下の式を用いて時系列データを変換したのちに, DTW により距離の算出を行う.

$$x'_i = x_i - x_{i-1}.$$

3.3 x-means 法

x-means 法は, k-means 法の拡張である. あらかじめクラスタ数を決めておかなければならない k-means 法とは違い, 最適なクラスタ数 m を推測することができる. x-means 法の考え方は, m = 2 で再帰的に k-means 法を実行していくというもので, クラスタの分割前と分割後で BIC 値 (ベイズ情報量規準) を比較し, 値が改善しなくなるまで分割を続ける. つまり, 分割前のベイズ情報量を BIC, 分割後のベイズ情報量を BIC' とし,

- BIC $>$ BIC' ならば 2 分割する
- BIC $\leq$ BIC' ならば 2 分割しない

を全ての場合で 2 分割出来なくなるまで繰り返すことにより最適なクラスタ数を決定する. p 変量正規分布を

$$f(x;\theta_p) = (2\pi)^{-\frac{p}{2}}|V_p|^{-\frac{1}{2}}\exp[-\frac{1}{2}(x-\mu_p)^t V_p^{-1}(x-\mu_p)]$$

とする μ_p は p 次の平均値ベクトル, V_p は $p \times p$ の分散共分散行列である. データセット $x_1, \cdots, x_n$ に対し, BIC 値は以下のように定義される [6].

$$BIC = -2\log L(\hat{\theta_p}) + q\log n$$

ここで, $\hat{\theta_p} = [\hat{\mu_p}, \hat{V_p}]$ は, p 変量正規分布の最尤推定値とする. q はパラメータ θ の次元数で L は尤度関数で $L(\theta) = \prod f(x_i;\theta_p)$ である.
2 分割したモデルにおける BIC' は以下のようになる.

$$BIC' = -2\log L(\hat{\theta'_p}) + q'\log n$$

ここで, $\hat{\theta_p}' = [\hat{\theta_p}^1, \hat{\theta_p}^2]$ は, 2 分割されたそれぞれの p 変量正規分布の最尤推定値とする. q' はパラメータ θ' の次元数である.

3.4 判別アルゴリズム

判別アルゴリズムには株価から抽出したデータを入力し, 株価が次時点で上がる場合には"1"を, 次時点で下がる場合には"0"を出力する.

3.4.1 決定木

決定木とは, 目的変数と説明変数のデータから木構造の分類器を生成して入力データを分類していく手法である [10]. 分岐処理をツリー状に形成し、トップから再帰的に対象データを分岐させて最終的なクラスを決定する.

3.4.2 ランダムフォレスト

ランダムフォレストとは, 複数の決定木を組み合わせてより強力なモデルを構築するアンサンブル手法である [7]. 個々の決定木がお互いに異なるように, 決定木の構築過程で乱数を導入し, 少しずつ異なる決定木をたくさん集める. 個々の決定木は比較的うまく予測できているが, 一部のデータに対して過剰適合してしまっていると言う考えに基づいている. それぞれ異なった方向に過剰適合した決定木をたくさん作れば, その結果の平均を取ることで過剰適合の度合いを減らすことができる.

3.4.3 勾配ブースティング

勾配ブースティングとは, 複数の決定木を組み合わせてより強力なモデルを構築するアンサンブル手法である [7]. 勾配ブースティングは, 1 つ前の決定木の誤りを次の決定木が修正するようにして, 決定木を順番に作っていく. 勾配ブースティングのポイントは, 浅い決定木のような, 簡単なモデル (弱学習機) を多数組み合わせることにある. それぞれの決定木はデータの一部にしてしか良い予測を行えないので, 決定木を繰り返し追加していくことで, 性能を向上させる.

3.4.4 サポートベクターマシン

SVM とは, Support Vector Machine の略称で, パターン識別学習アルゴリズムである [8]. SVM では, 学習データと識別境界の距離 (マージン) が最大化されるように識別境界を求める. クラス分離が線形で行われる場合は, 線形分離と同様だが, 線形分離ができない場合, カーネルトリックを用いる. カーネルトリックとは, 入力 x を高次元の特徴空間 $\psi(x)$ へ写像し, 線形分離を行う方法である. 2 つの入力点 $x1$, $x2$ の非線形写像, $\psi(x_1)$, $\psi(x_2)$ の内積が

$$K(x_1, x_2) = \psi(x_1) \cdot \psi(x_2)$$

のようにカーネル関数 K を用いて計算できると, 非線形変換を行うことなく内積を計算することができる. このカーネル関数を使い, 線形分類を行い予測する.

第4章　提案手法

本研究では, 機械学習を用いて対象銘柄と類似度が高いデータを特徴量として予測モデルに組み込むことで, 対象銘柄の株価が次時点で上がるか下がるかを予測する精度を上げることができるのではないかという仮説を立てた. なぜなら, 株価というものは自身の業績などだけで変化するものではなく他の企業や為替, 政治などの外部からの影響を受けて変化する時系列データであるからだ. そこで, 対象の株価を推定できるような説明変数を見つけ出し, 予測モデルを作成すれば, 予測精度は向上すると考える. その説明変数として, 本研究では他の企業の株価から抽出した情報を用いる. 目的変数には, 株価が次時点で上がる場合には”1”を, 次時点で下がる場合には”0”を用いる. 影響を与えあっている企業を探索する方法としてクラスタリングを援用する. クラスタリングとは, あるデータをデータ間の類似度に従って, いくつかのグループに分けるデータ分析の手法であり, 対象と関係のありそうなデータを選択するために用いる. 本研究では, 最初に株価時系列データの形状の類似性を考慮できる DDTW を用いて各企業間の株価時系列データの類似度を測った. 次に, 代表的なクラスタリング手法である k 平均法に比べ, 最適なクラスタ数を推測することができる x-means 法でクラスタリングを行った. また, 株価予測に関係のありそうなデータとして, クラスタ平均を用いる理由は, 平均値は代表的な統計量であり, そのデータの特徴をつかむ数値としてよく用いらているからである. 提案手法の流れは以下である.

1. 以下で株価時系列データを変換したのち, DTW 距離を用いて各企業間の類似度を計算する.
$$x'_i = x_i - x_{i-1}$$
2. 各企業間の類似度行列を作成し, x-means 法でクラスタリングを行う.
3. 株価を予測したい企業の株価データと, 対象の企業が属するクラスタのクラスタ平均などの株価予測に関係のありそうな特徴量を用いて予測モデルを構築する.

3 で用いる特徴量とは, データにどのような特徴があるかを数値化したものである. 機械学習を行う際には, 入力された学習データから特徴量と呼ばれる数値ベクトルを抽出する必要があり, 抽出した特徴量を元にパターン・経験則をモデルを使って学習する. 機械学習モデルの予測パフォーマンスを良くするには元データの特徴量を増やしたり, 変更を加えたり, 組み合わせることで学習に適した特徴量を生成することが重要である. 本研究では機械学習モデルの予測精度の改善を目的として, 訓練データの中からターゲットの予測により強い関連がある特徴量として, 始値, 終値, 出来高, 株価の上下などあらゆる特徴量を検討する.

第5章　実験と評価

本章では, 行った実験とその結果について述べる. 構成は次のようになっている.

- 5.2 : 各クラスタ平均の次時点での上下を各クラスタに属する銘柄の次時点での予測値として実験を行った. 異なる学習期間のモデルを組み合わせたアンサンブル予測も行った.
- 5.3 : 各クラスタ平均の次時点での上下を各クラスタに属する銘柄の次時点での予測値として実験を行った. 本実験では, ほんの少しの上昇下落のようなノイズを取り除いて学習させることで予測精度の向上を目指した.
- 5.4 : 個別銘柄の株価の次時点の上下を予測した. 予測対象の銘柄の株価予測をする際に, 銘柄の株価時系列データに予測対象の銘柄の属するクラスタのデータを加えることで予測精度の向上を目指した.

5.1　実験に使用したデータ

日経平均株価に採択されている 225 銘柄の 2018/4/1 から 2018/10/1 までの 125 日分の始値, 終値, 出来高を用いた. また, 予測対象のデータの前半 7 割を学習用トレーニングデータに, 後半 3 割を評価用テストデータにし, 結果を評価した.

5.2　クラスタ平均値予測

各クラスタ平均の次時点での上下を各クラスタに属する銘柄の次時点での予測値として実験を行なった.

5.2.1　予測モデル

225 銘柄の終値を用いて DDTW 距離行列を作成し, x-means 法で銘柄をクラスタリングし, 以下を組み合わせて予測モデルを作成した.

- 学習日数: 1 日, 5 日, 25 日
- 株価データ: 始値のクラスタ平均, 終値のクラスタ平均, 出来高のクラスタ平均
- 学習アルゴリズム: 決定木, ランダムフォレスト, サポートベクターマシン, 勾配ブースティング

また, 1 日, 5 日, 25 日の学習日数の予測を組み合わせたアンサンブル予測も加えた.

5.2.2 評価方法

予測精度には以下を用いた.

$$\text{期待値} = \frac{\max\{\text{株価上昇回数}, \text{株価下降回数}\}}{\text{株価上昇回数} + \text{株価下降回数}}$$

$$\text{予測精度} = \frac{\text{期待値を上回ったクラスタ数の合計}}{\text{総クラスタ数}}$$

5.2.3 予測結果

終値を用いた予測結果を表 5.1 に示す. 出来高, 始値を用いた予測結果も同様の結果となった. 予測精度のばらつきが大きく上手く予測できていないと考えられる. 学習日数を長くすることによって予測精度を向上させることはできないと考えられる.

表 5.1: 終値のクラスタ平均を用いた予測精度

	学習日数			
学習アルゴリズム	1 日	5 日	25 日	1 日 5 日 25 日
決定木	36%	31%	26%	26%
ランダムフォレスト	26%	26%	31%	26%
サポートベクターマシン	31%	26%	26%	31%
勾配ブースティング	31%	15%	31%	15%

5.3 クラスタ平均値 2 段階予測

各クラスタ平均の次時点での上下を各クラスタに属する銘柄の次時点での予測値として実験を行なった. 本実験では, ほんの少しの上昇下落のようなノイズを取り除いて学習させることで予測精度を向上することができると考え, 翌日の株価の変動が大きい (1%以上の上昇または-1%以上の下落) か小さい (1%未満の上昇または-1%未満の下落) かというデータを予測モデルに組み込んだ. 225 銘柄の終値を用いて DDTW 距離行列を作成し, x-means 法で銘柄をクラスタリングし, 以下を組み合わせて予測モデルを作成した.

- 学習日数: 1 日,$\cdots$,30 日
- 株価データ: 終値のクラスタ平均
- 学習アルゴリズム: ランダムフォレスト

予測手法の流れは以下である.

1. 翌日の株価の増減率によって上昇 (1%以上の上昇), 中間 (1%未満の上昇または-1%未満の下落), 下落 (-1%以上の下落) の 3 つに分け, 上昇下落の大きく変動する場合は”1”, 中間の小さく変動する場合は”0”とするデータを作成する.

2. 翌日の株価が大きく変動 (上昇, 下落) するのか小さく変動 (中間) するのかを学習させた予測器 $g(x)$ を作成する.

3. 2 で学習した予測器 $g(x)$ によって翌日の株価が大きく変動 (上昇, 下落) すると判断された株価を対象として, 次の時点での上下を学習させ, 予測器 $f_1(x)$ を作成する. また, 予測器 $g(x)$ によって翌日の株価が小さく変動 (中間) すると判断された株価を対象として, 次の時点での上下を学習させ, 予測器 $f_2(x)$ を作成する.

4. $g(x)$, $f_1(x)$, $f_2(x)$ の 3 つの予測器を評価用テストデータに適用する. 手順としてはまず, x 日分 $(x = 1, \cdots, 30)$ の終値のクラスタ平均のデータに対し $g(x)$ を適用する. 次にその $g(x)$ の判断によって $f_1(x)$ または $f_2(x)$ を翌日の株価の上下のデータに適用し, 結果を評価する.

5.3.1 評価方法

予測精度には以下を用いた.

$$\text{期待値} = \frac{\max\{\text{株価上昇回数}, \text{株価下降回数}\}}{\text{株価上昇回数} + \text{株価下降回数}}$$

$$\text{予測精度} = \frac{\text{期待値を上回ったクラスタ数の合計}}{\text{総クラスタ数}}$$

5.3.2 予測結果

2 段階ではない場合の予測と 2 段階の場合の予測の結果を図 5.1 に示す. 縦軸が予測精度, 横軸が学習日数, 赤が 2 段階ではない場合の結果とその回帰直線 , 青が 2 段階の場合の予測の結果とその回帰直線である. 結果としては, 2 段階で予測することで予測精度が上がる傾向は見られなかった.

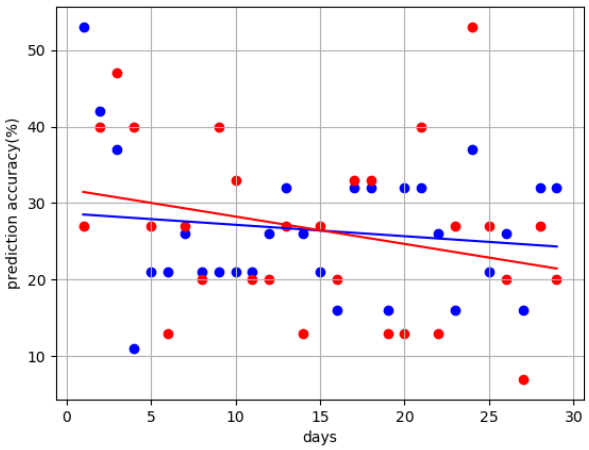

図 5.1: 学習日数と予測精度

5.4 個別銘柄予測

予測対象の銘柄の株価予測をする際に, 銘柄の株価時系列データだけを用いたモデルの予測精度よりも銘柄の株価時系列データに予測対象の銘柄の属するクラスタのデータを加えたモデルの方が予測精度が上がるのではないかという仮説を検証した.

5.4.1 予測モデル

225 銘柄の終値を用いて DDTW 距離行列を作成し, x-means 法で銘柄をクラスタリングし, 以下を組み合わせて予測モデルを作成した. また, 銘柄の連動類似度を用いた予測結果を評価するため 1 銘柄のみ属しているクラスタは除いた.

- 学習日数: 1 日, $\cdots$, 30 日
- 株価データ: 個別銘柄の終値, 個別銘柄の属するクラスタの終値の平均
- 学習アルゴリズム: ランダムフォレスト

5.4.2 評価方法

クラスタ平均の効果を確かめるために, 予測精度を以下の量で評価した.

$$\text{評価量 } 1 = y_t - \hat{y}_t$$

$$\text{評価量 } 2 = y_t - \tilde{y}_t$$

ここで, y_t は対象銘柄の株価と属するクラスタの平均との差分を用いた予測結果, $\hat{y}_t$ は対象銘柄の株価のみを用いた予測結果, $\tilde{y}_t$ は対象銘柄の株価と属するクラスタの平均との差分を用いた際, 訓練データにおける株価の上昇回数と下落回数で多い方を予測としてテスト期間を予測した時, つまりチャンスレベルでの予測結果である. 評価量 1, 評価量 2 は正の値を取るとき, クラスタ平均の効果がある.

5.4.3 予測結果

各クラスタの評価量 1 と学習日数の箱ひげ図を図 5.2, 図 5.4, 図 5.6, 図 5.8, 図 5.10, 図 5.12, 図 5.14, 図 5.16, 図 5.18, 図 5.20, 図 5.22, 図 5.24, 図 5.26, 図 5.28, 図 5.30 に示す. 各クラスタの評価量 2 と学習日数の箱ひげ図を図 5.3, 図 5.5, 図 5.7, 図 5.9, 図 5.11, 図 5.13, 図 5.15, 図 5.17, 図 5.19, 図 5.21, 図 5.23, 図 5.25, 図 5.27, 図 5.29, 図 5.31 に示す.1, 4, 5, 8, 16 の 5 つのクラスタは 1 銘柄のみ属しているため除いた. クラスタの平均との差分のデータを加えるとクラスタや学習日数によって予測精度にばらつきがあり, 効果があるとは言えない.

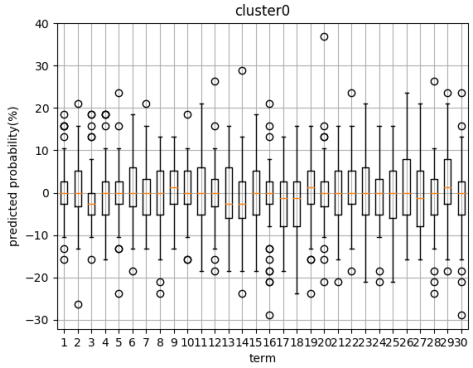

図 5.2: クラスタ 0 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

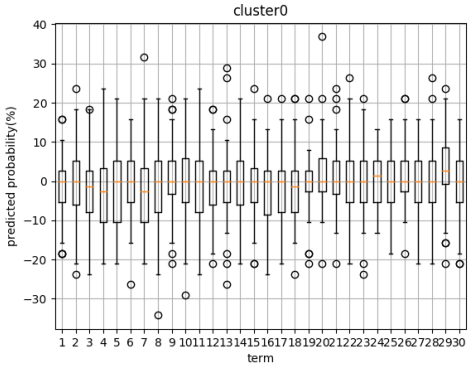

図 5.3: クラスタ 0 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

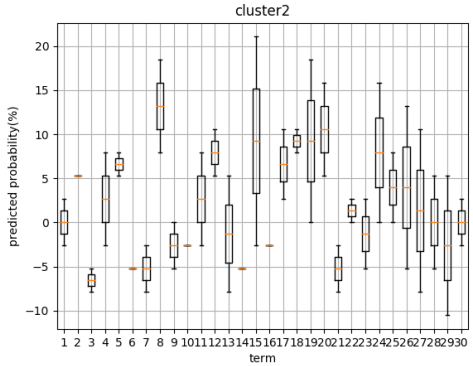

図 5.4: クラスタ 2 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

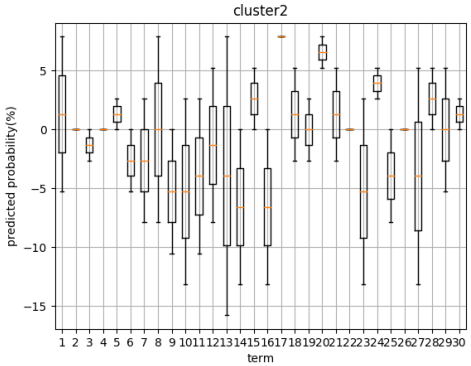

図 5.5: クラスタ 2 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

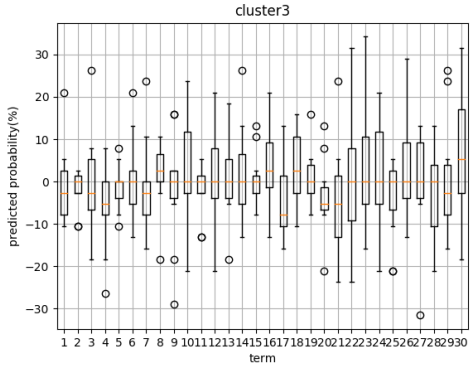

図 5.6: クラスタ 3 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

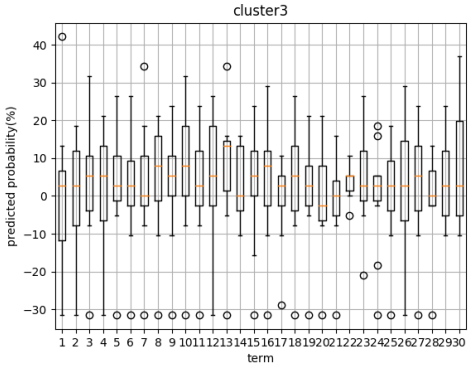

図 5.7: クラスタ 3 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

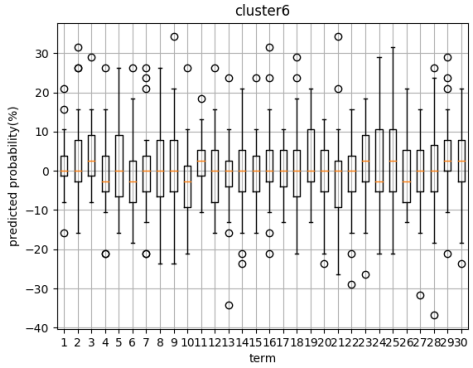

図 5.8: クラスタ 6 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

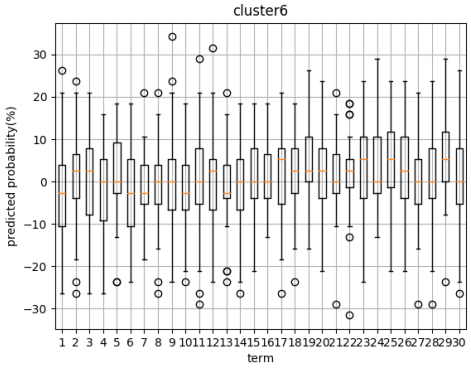

図 5.9: クラスタ 6 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

図 5.10: クラスタ 7 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

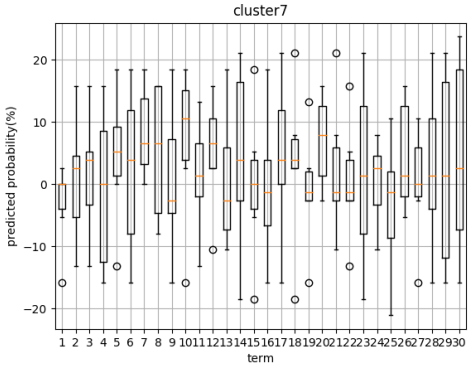

図 5.11: クラスタ 7 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

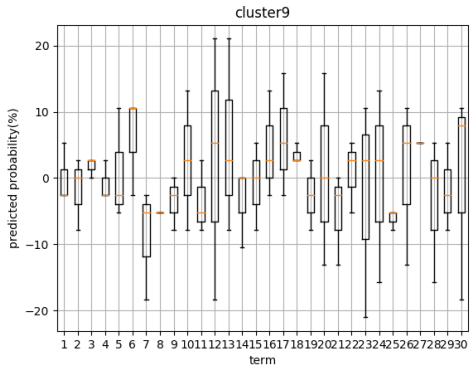

図 5.12: クラスタ 9 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

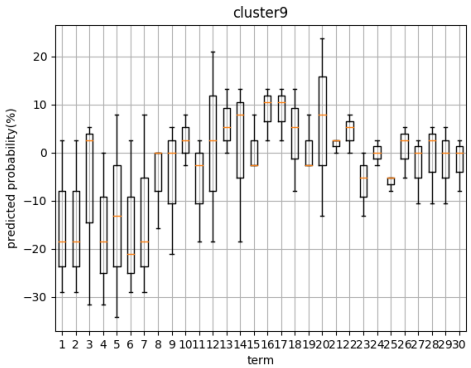

図 5.13: クラスタ 9 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

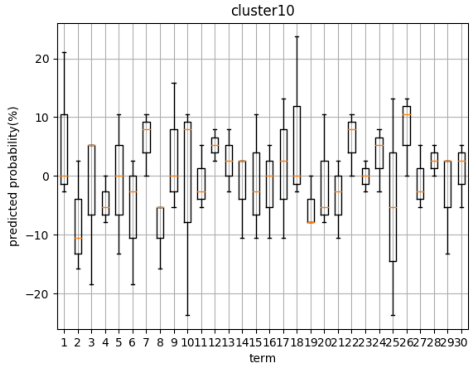

図 5.14: クラスタ 10 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

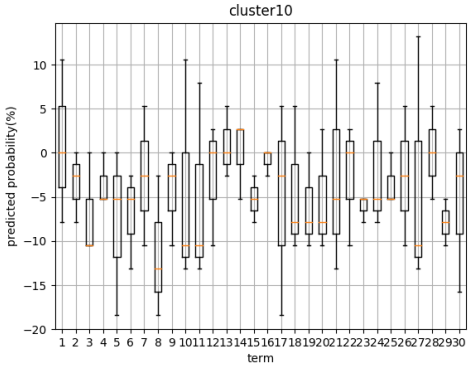

図 5.15: クラスタ 10 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

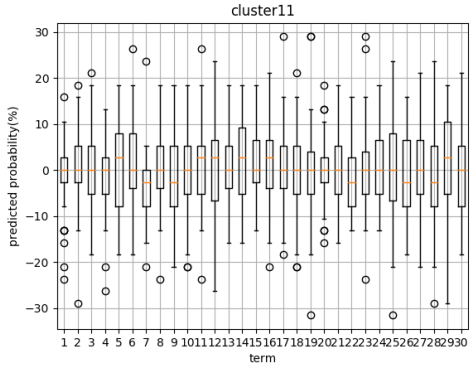

図 5.16: クラスタ 11 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

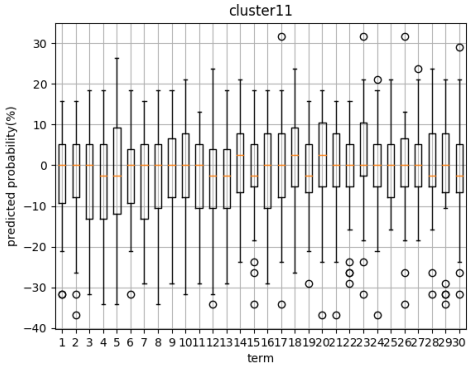

図 5.17: クラスタ 11 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

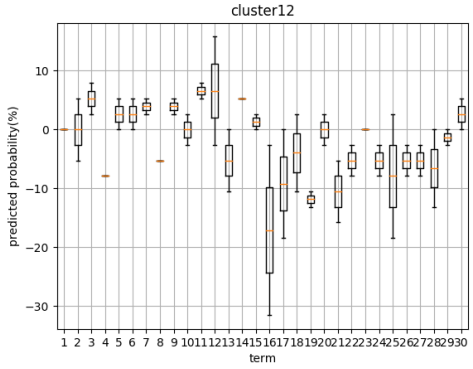

図 5.18: クラスタ 12 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

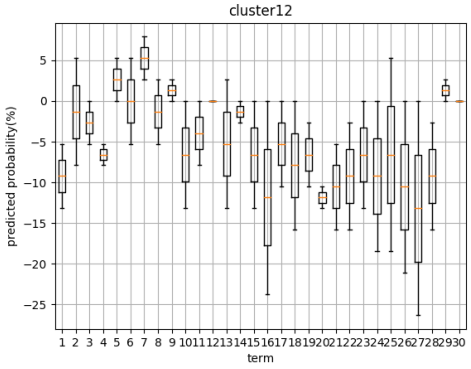

図 5.19: クラスタ 12 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

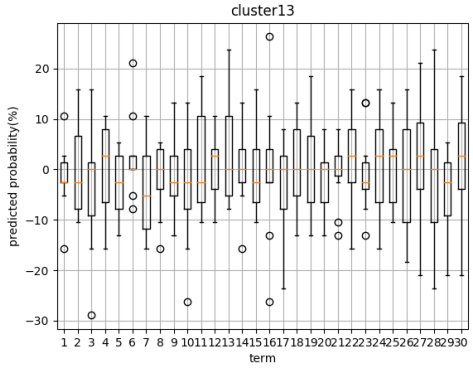

図 5.20: クラスタ 13 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

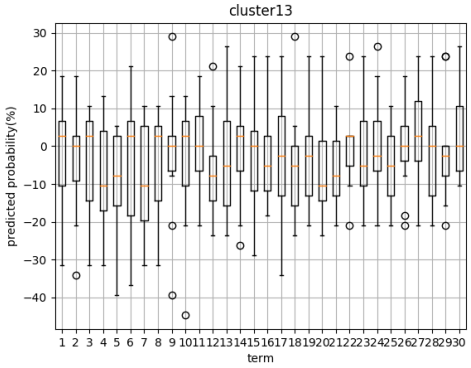

図 5.21: クラスタ 13 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

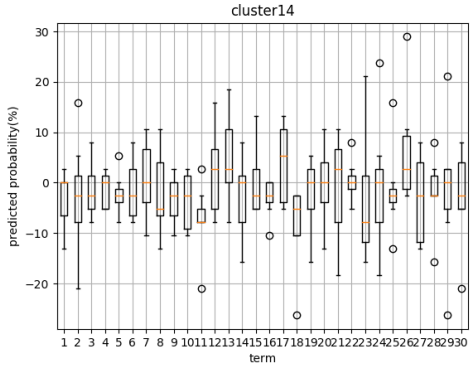

図 5.22: クラスタ 14 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

図 5.23: クラスタ 14 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

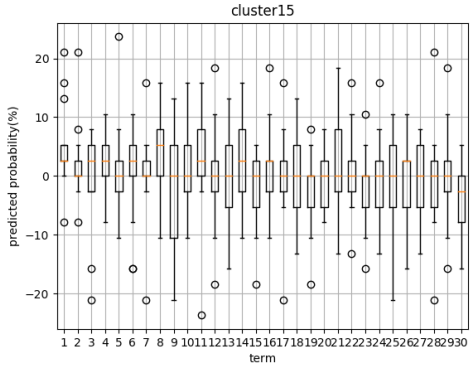

図 5.24: クラスタ 15 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

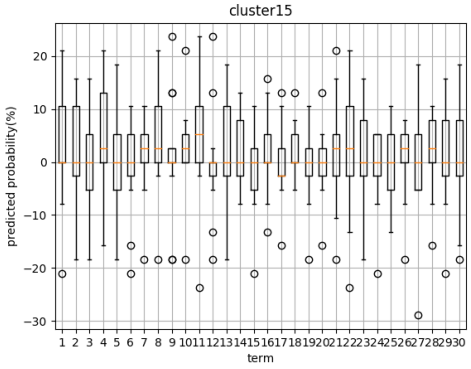

図 5.25: クラスタ 15 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

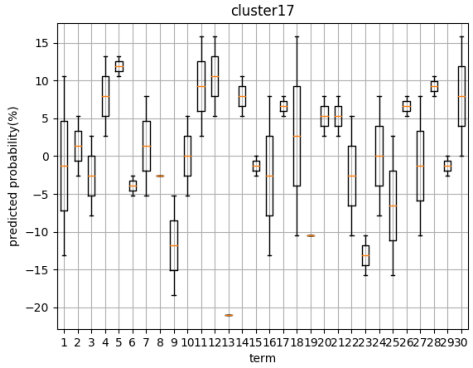

図 5.26: クラスタ 17 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

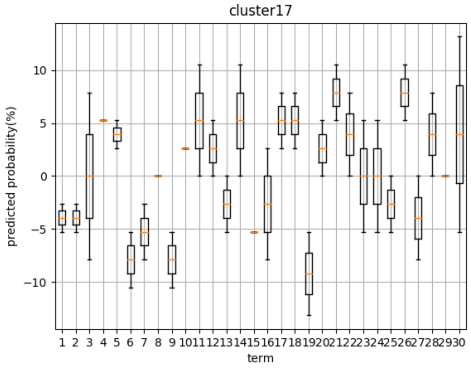

図 5.27: クラスタ 17 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

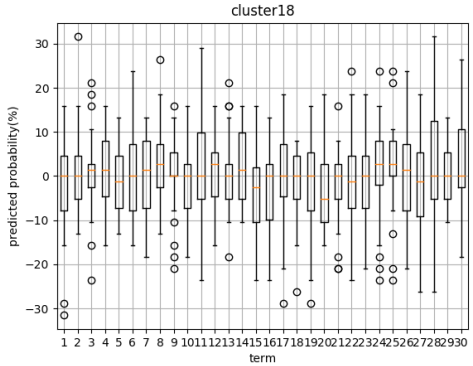

図 5.28: クラスタ 18 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

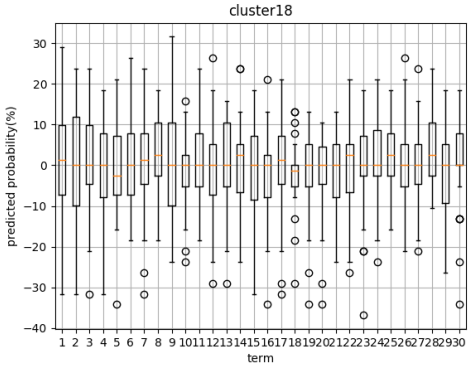

図 5.29: クラスタ 18 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

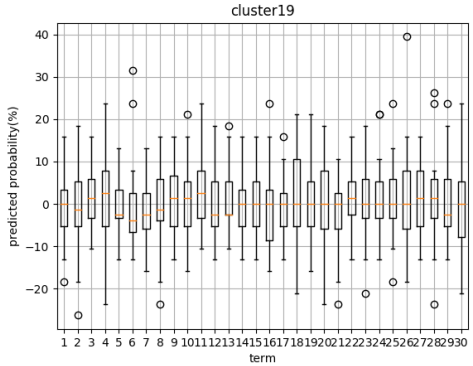

図 5.30: クラスタ 19 の評価量 1 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 1.

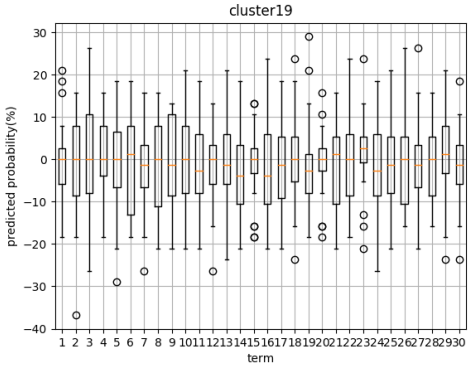

図 5.31: クラスタ 19 の評価量 2 の箱ひげ図. 横軸は学習日数, 縦軸は評価量 2.

5.5 考察

クラスタ平均を用いることで, 対象銘柄の株価のみで予測した時よりも精度が上がることがあったが, 一方で精度があまり変わらなかったり下がることもあった. つまり, クラスタ平均を用いただけでは株価の動向をすべて説明することができない. 原因として考えられるのは, 株価以外の影響や過去のデータにない株価の急激な変動などが考えられる. また, 株価の類似度を DTW 距離によって決定していたが, 他の類似度を用いることでクラスタリング結果, 予測精度に影響が出ると思われる. 加えて, クラスタのデータとして平均値を用いたが, 他の統計量を用いることで予測精度に影響が出ると思われる.

第6章　結論

本研究では, 時系列解析とクラスタリングを組み合わせることで株価の予測精度が向上するという仮説の下で実験を行った. 結果として, クラスタ平均を用いることで予測精度が上がる傾向は見られなかった. これは, 株価が自身の業績だけではなく様々な外部要因の影響を受けて形成されていることや過去の情報だけでは説明できない動きをするためだと考える. 今までの実験の考察からクラスタ平均のデータを組み合わせただけでは予測精度が上がる傾向は見られないことがわかったので他の統計量を用いるなどデータの抽出方法を変えたり, 為替や経済指標などの株価以外のデータを用いることが今後の方針である.

謝辞

本研究において多大なるご指導をしてくださった竹之内先生, 共に頑張ってきた同研究室の波多さん, 1 年間ありがとうございました.

参考文献

[1] 平林明憲, 伊庭斉志：遺伝的アルゴリズムによる外国為替取引手法の最適化, 人工知能学会全国大会論文集, 3H1-02, 2008.

[2] 宮崎邦洋, 松尾豊： 深層学習を用いた株価予測の分析, 人工知能学会全国大会論文集, vol.31, 2017.

[3] 中川慧, 今村光良, 吉田健一：株価変動パターンの類似性を用いた株価予測, 第 31 回人工知能学会全国大会, 2D1-1, 2017.

[4] 小沢育実, 関和広：関連銘柄同定のための時系列データ類似度尺度の提案, 第 78 回情報処理学会全国大会, 6H-08, 2016.

[5] 志津綾香, 松田眞一：クラスタ分析におけるクラスタ数自動決定法の比較, アカデミア情報理工学編 11, 17-34, 2011.

[6] 赤池弘次：AIC と MDL と BIC, オペレーションズリサーチ, 1996 年 7 月号, 375-378, 1996.

[7] Andreas C. Muller, Sarah Guido：Python ではじめる機械学習 -scikit-learn で学ぶ特徴量エンジニアリングと機械学習の基礎, オーム社, 2017.

[8] 津田博史, 嶋田 康史： ディープラーニング入門 -Python ではじめる金融データ解析, 朝倉書店, 2018.

[9] 佐藤賀一：テクニカル分析に基づくペアトレードの有効性と日本の株式市場の効率性, 行動経済学第 10 巻, 22-49, 2017.

[10] 荒木大, 小島昌一： 数値データによる決定木の帰納学習, 人工知能学会誌, 7 巻 6 号, 992-1000, 1992.

[11] 海野一則, 山田隆志, 寺野隆雄：機械学習を用いたポートフォリオの最適化, 第 27 回人工知能学会全国大会, 1D5-6, 2013.