数法則発見による オルタナティブデータを用いた 金融マーケット予測手法の開発

Development of Financial Market Forecasting Method Considering Alternative Data on Economy

Itaru Aso

Graduate School of Information Engineering, Toyama Prefectural University
t855001@st.pu-toyama.ac.jp

L205, AM 9:00-9:25, Friday, December 8, 2018,
Toyama Prefectural Univ.

1.1 背景

本研究の背景

1. 計算機科学の発展により，ビッグデータの蓄積や蓄積したデータを機械学習を用いて分析することが可能.
2. ビッグデータを用いたデータ分析は，金融経済現象にも応用.
3. オルタナティブ・データを活用することで新たな金融工学の地平が切り開かれている.

既存研究

1. OpinionFinder(OF) と Google-Profile of Mood States を元にツイートから心的状態を表す指数からダウ平均株価の予測
2. 金融テキストマイニングを用いた市場動向推定

1.2 本研究の概要

投資判断の分析手法

1. ファンダメンタル分析
   → 国際的な経済の動きや個別の企業の財務情報など市場外的要因を考慮する手法
2. テクニカル分析
   → 現在の市場のトレンドを把握する方法であり，テクニカル指標を用いて市場内的要因を考慮する手法

従来の為替予測の問題点

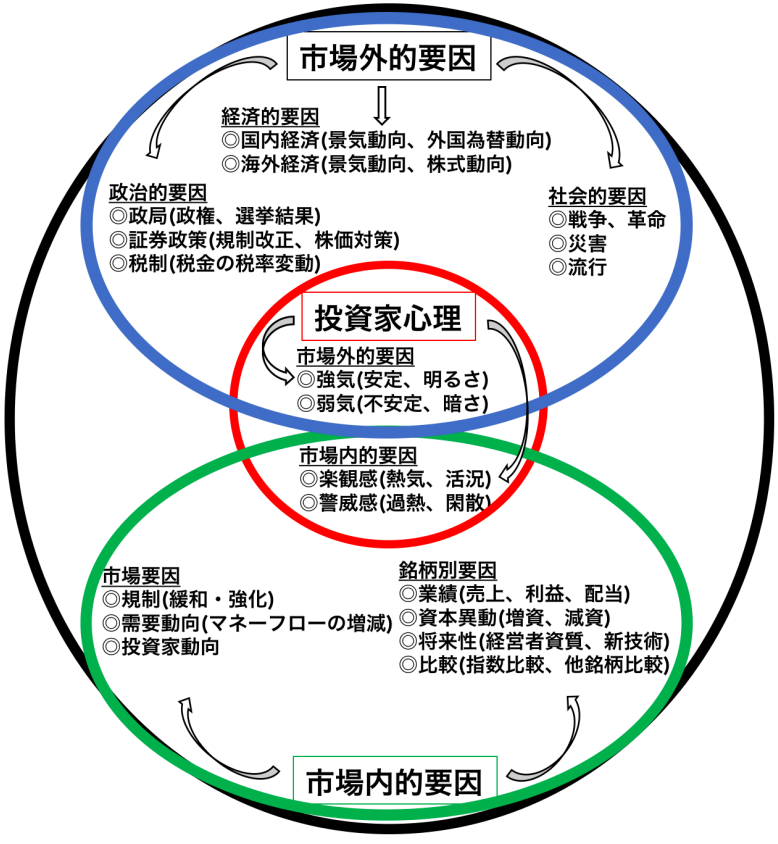
- 従来の為替予測では，予測を行う際には為替の価格のみだったり，テクニカル指標を用いた予測手法が行われている．

  → 市場内的要因しか考慮していないためマーケットの挙動の変化を予測して対応できない

1.3 本研究の目的

本研究の目的

従来のテクニカル分析を用いた予測手法に加えて Web 上の景気に関する情報や SNS からの情報のテキストからセンチメント分析を行い, 金融マーケットの状況の把握も可能な予測手法の提案を行う.



2.1 本研究の流れ

市場外的要因と市場内的要因を考慮した予測手法の提案の流れを以下に示す.

本研究の目的

1. ツイートが為替に影響しているのかどうかの検証.
2. 予測に用いるテクニカル指標の重回帰分析を用いた選択
3. Twitter のデータをセンチメント分析による感情スコアの抽出
4. 2, 3 で求めたテクニカル指標と感情スコアを用いた為替予測

2.2 ツイートの為替への影響

まず，実際にツイートが為替に影響しているかどうかを検証する必要がある

特定のアカウント

本研究では，k-Shape によるクラスタリングを行うことでツイートの為替への影響を検証する．

Twitter API を用いてツイートの取得を行った．Twitter API からはタイムスタンプやツイート，リツイート数，いいねの数など様々な情報を取得することが可能である．本研究ではタイムスタンプとツイートのみ扱うことにする．

text	created_at	retweet_count	favorite_count
“If the Fed backs off and starts talking a little more Dovish I think we’re going to be right back to our 2800 to 2900 target range that we’ve had for the S&P 500.” Scott Wren Wells Fargo.	10-30-2018 12:53:03	14962	61498
The Stock Market is up massively since the Election but is now taking a little pause - people want to see what happens with the Midterms. If you want your Stocks to go down I strongly suggest voting Democrat. They like the Venezuela financial model High Taxes & Open Borders!	10-30-2018 12:33:39	30334	112637
Congressman Kevin Brady of Texas is so popular in his District and far beyond that he doesn’t need any help - but I am giving it to him anyway. He is a great guy and the absolute “King” of Cutting Taxes. Highly respected by all he loves his State & Country. Strong Endorsement!	10-30-2018 12:25:07	14233	57144

Figure: 1. Twitter API により取得したツイートの例

2.3 k-Shape

k-Shape

時系列の形状に着目した時系列クラスタリング手法

クラスタリングまでの流れ

1. クラスタの核となる k 個の重心ベクトルを決める.
2. 各時系列データと各 k 個のクラスタの重心ベクトルと比較して，重心の距離が最も近いクラスタに割り当てる.
3. 各クラスタの重心ベクトルを更新する.
4. 重心ベクトルの値が変化しなければ終了
5. 重心ベクトルの値が変化したならば，[1] に戻る

2 つの時系列データ x と y 距離尺度 Shape-based distance (SBD)

$$SBD(\boldsymbol{x}, \boldsymbol{y}) = 1 - \max_{w} \left(\frac{CC_w(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{R_0(\boldsymbol{x}, \boldsymbol{x}) \cdot R_0(\boldsymbol{y}, \boldsymbol{y})}} \right) \tag{1}$$

2.4 為替予測

本研究の為替予測手法

為替予測の分析手法として Long short-term memory(LSTM) を用いる

LSTM セル

時刻 t の入力 x の重みを W, 時刻 $t-1$ の隠れ層 h_{t-1} の重みを U, 切片を b とした.
LSTM では，入力ゲート i_t, 忘却ゲート f_t, 出力ゲート o_t を用いることにより，入力と隠れ層，出力の重みを調節して長期的な記憶を実現している.

$$i_t = \sigma(w_i * [x_t, h_{t-1}] + b_i) \tag{2}$$

$$f_t = \sigma(w_f * [x_t, h_{t-1}] + b_f) \tag{3}$$

$$o_t = \sigma(w_o * [x_t, h_{t-1}] + b_o) \tag{4}$$

$$u_t = tanh(w_u * [x_t, h_{t-1}] + b_u) \tag{5}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot u_t \tag{6}$$

$$h_t = o_t * tanh(c_t) \tag{7}$$

2.5 重回帰分析を用いた特徴量の選定

本研究で用いる特徴量に対して，実際の未来の為替への影響が大きい特徴量を選ぶ．そして，為替の予測に選定された特徴量を入力として分析を行う．
入力にする特徴量，重回帰分析の p 値が有意かどうかを基準にして選ぶ．

重回帰分析

現在の時刻 t から i 分後の為替の終値目的変数 y_t^i, n 次元のテクニカル指標を目的変数 $\boldsymbol{x}$ とすると，以下の式に表すことができる．

$$y_t^i = b_0 + b_1 x_1^i + b_1 x_2^i + \cdots b_n x_n^i \tag{8}$$

ここで，b_0 は定数，$b_1 \cdots b_n$ は偏回帰係数である．
そして，有意確率 p 値が 5 %で有意であった特徴量を為替予測の入力として用いる．

2.6 センチメント分析による感情スコアの算出

感情スコア

センチメント分析によって，Pultick の感情の環に基づいてスコアを算出する

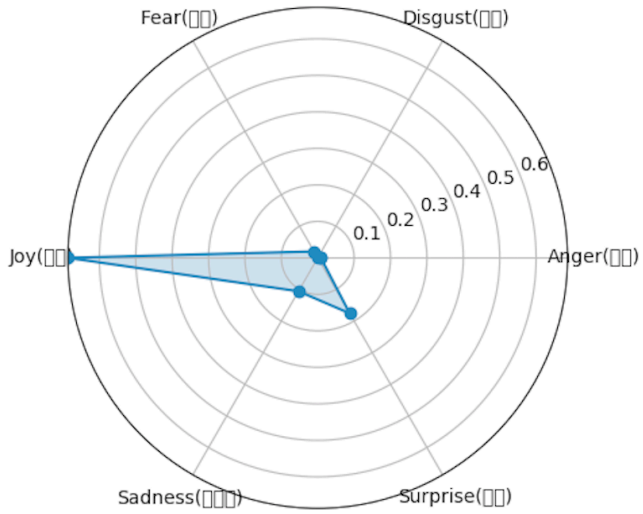


Figure: 2. Pultick の感情の環に基づいた感情スコア

「A years ago, A star was Born, and here we are 6 times pink platium」

2.7 センチメント分析

センチメント分析

1. Web 上のユーザー生成コンテンツの量は，主に SNS やブログなど自分の個人コンテンツを共有することを可能にする無数のプラットフォームの出現により，ますます急速に増加している．
2. ユーザー生成コンテンツは意見や感情が豊富であり，株式市場の変動の予測であったり，マーケティングの戦略の参考にされることが多い．
3. オンライン上の人々の情報を自動的に分析する手法が重要となってきている．

2.7 感情スコアの導出

センチメント分析

感情スコアを Niko Colneri らの Twitter での感情認識を参考にシステムを作成した.

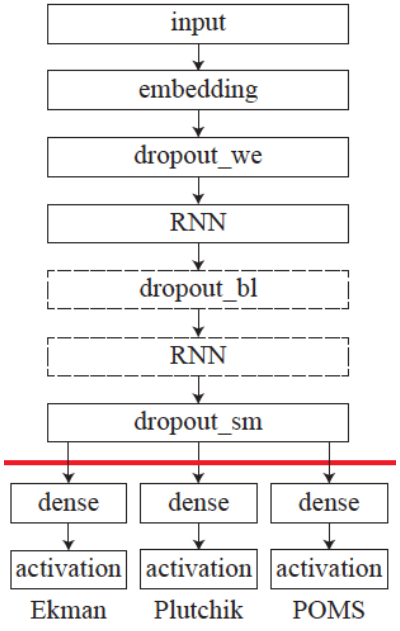


Figure: 2. 参考にしたネットワーク構成

2.8 主成分回帰分析を用いる GMDH

GMDH

1. 少ない入出力データからシステムに関する先験的な知識を必要とすることなくモデリングできる
2. 階層型ニューラルネットワークと同じく，多層構造と並列処理を特徴としたアルゴリズム

主成分回帰分析を用いる GMDH のメリット

1. 最適な変数の組み合わせを自己選択する.
2. 層の積み重ねを自動停止する.

2.9 定式化

GMDH

システムの完全表現式として，Kolmogorov-Gavor の多項式を想定する．

$$\Phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \cdots \tag{9}$$

システムの部分表現式は 2 変数の 2 次多項式

$$y_k = b_{0k} + b_{1k} x_i + b_{2k} x_j + b_{3k} x_i x_j + b_{4k} x_i^2 + b_{5k} x_j^2 \tag{10}$$

以上の部分表現式における変数に対して，評価規準 AIC を用いて変数の逐次選択を行う

2.9 最適部分表現式の作成

最適部分表現式

はじめに，最適部分表現式を作成する前に入力データを平均 0，分散 1 に規準化し，出力データを平均 0 に規準化する．
規準化した変数 $\boldsymbol{x}_{ij}$ を次のように直行変換し，新しい変数 $\boldsymbol{z}^{\mathrm{T}} = (z_1, z_2, \cdots, z_5)$ を次のように求める．

$$\boldsymbol{z} = C \cdot \boldsymbol{x}_{ij} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{15} \\ c_{21} & c_{22} & \cdots & c_{25} \\ \vdots & \vdots & \ddots & \vdots \\ c_{51} & c_{52} & \cdots & c_{55} \end{bmatrix} \begin{bmatrix} x_i \\ x_j \\ x_i \cdot x_j \\ x_i^2 \\ x_j^2 \end{bmatrix} \tag{11}$$

ここで，C は 5×5 の正規直交行列を示す．C は，$\boldsymbol{x}_{ij}$ から構成する相関行列 R の固有値問題を解くことにより求められる．

$$R \cdot C = C \cdot \Lambda \tag{12}$$

また，R は 5×5 の相関行列，Λ は固有値 $\lambda_1, \lambda_2, \cdots, \lambda_5$ を対角要素にもつ対角行列である．

2.10 最適部分表現式の作成

最適部分表現式

作成した新しい変数 $\boldsymbol{z}$ を入力変数として直行回帰分析を行い多重共線性を起こさない最適部分行列を以下のように求める.

$$y_k = \boldsymbol{z}^{\mathrm{T}} \cdot \boldsymbol{d}_k = \begin{bmatrix} z_1 z_2 \cdots z_5 \end{bmatrix} \begin{bmatrix} d_{k1} \\ d_{k2} \\ \vdots \\ d_{k5} \end{bmatrix} \tag{13}$$

$\boldsymbol{d}_k$ は y_k に対応する係数ベクトルを示す.
係数ベクトル $\boldsymbol{d}_k$ は，入力変数 $\boldsymbol{z}$ が直交化されているため以下のように求めることができる.

$$Z^{\mathrm{T}} \cdot \boldsymbol{y}_k = (Z^{\mathrm{T}} \cdot Z) \cdot \boldsymbol{d}_k \tag{14}$$

データ数 n 個とすると，$\boldsymbol{y}_k^{\mathrm{T}} = (y_{k1}, y_{k2}, \cdots, y_{kn})$, Z は n 個の $\boldsymbol{z}^{\mathrm{T}}$ からなりデータ行列である.

2.11 最適部分表現式の作成

最適部分表現式

最適部分表現式を作成するときに，有用な変数のみを自己選択するために情報規準量 AIC を用いて変数の逐次選択を行う．
AIC の計算に用いる残差の自乗平均は，

$$S_m^2 = S_{m-1}^2 - \frac{\frac{(Z^{\mathrm{T}} \cdot \boldsymbol{y}_k)_i}{d_{ki}}}{n} \tag{15}$$

$$S_0 = \frac{\boldsymbol{y}_k^{\mathrm{T}} \cdot \boldsymbol{y}_k}{n} \tag{16}$$

中間変数の自己選択は，最適な部分表現式によって発生される中間変数に対して AIC の値の小さいものを自己選択する．

2.12 多層構造の計算停止方法

多層構造の計算停止方法

変数の次元が縮小する層は固有値を用いて以下のように判定する．各選択層のすべての中間変数い対して，

$$\frac{\lambda_{k,max1} + \lambda_{k,max2}}{\sum_{i=1}^{5} \lambda_{k,i}} > E \tag{17}$$

が満たされたときに変数の次元が縮小したとみなし多層構造の積み重ねを打ち切る．$\lambda_{k,max1}, \lambda_{k,max2}$ は第 k 番目の中間変数に対する固有ベクトル $\boldsymbol{\lambda}_k = [\lambda_{k,1}, \lambda_{k,2}, \lambda_{k,3}, \lambda_{k,4}, \lambda_{k,5}]^{\mathrm{T}}, (k = 1 \sim L)$ の 5 個の要素の中で 1 番目と 2 番目に大きな値をとる固有値を示す．L は中間変数の選択個数を示す．E は多層構造の打ち切り判定規準値を示す．

2.13 簡単な同定問題への適用

適用する式

$$\Phi = (1.0 + 1.1x_1 + 1.2x_2 + 1.3x_3)^4 + \varepsilon \tag{18}$$

ここで，ε は $N(0, 25^2)$ の正規性白色雑音を示す.

入力変数：$x_1 \sim x_4$(4 変数)

$\rightarrow$ 適用する式において x_4 は含まれていないが，入力に不必要な変数が含まれている場合においてもシステムの同定が可能であることを示すために x_4 を入力に含めている.

同定に用いる入出力データ：20 個

モデルの評価に用いる入出力データ：20 個

同定されたモデル

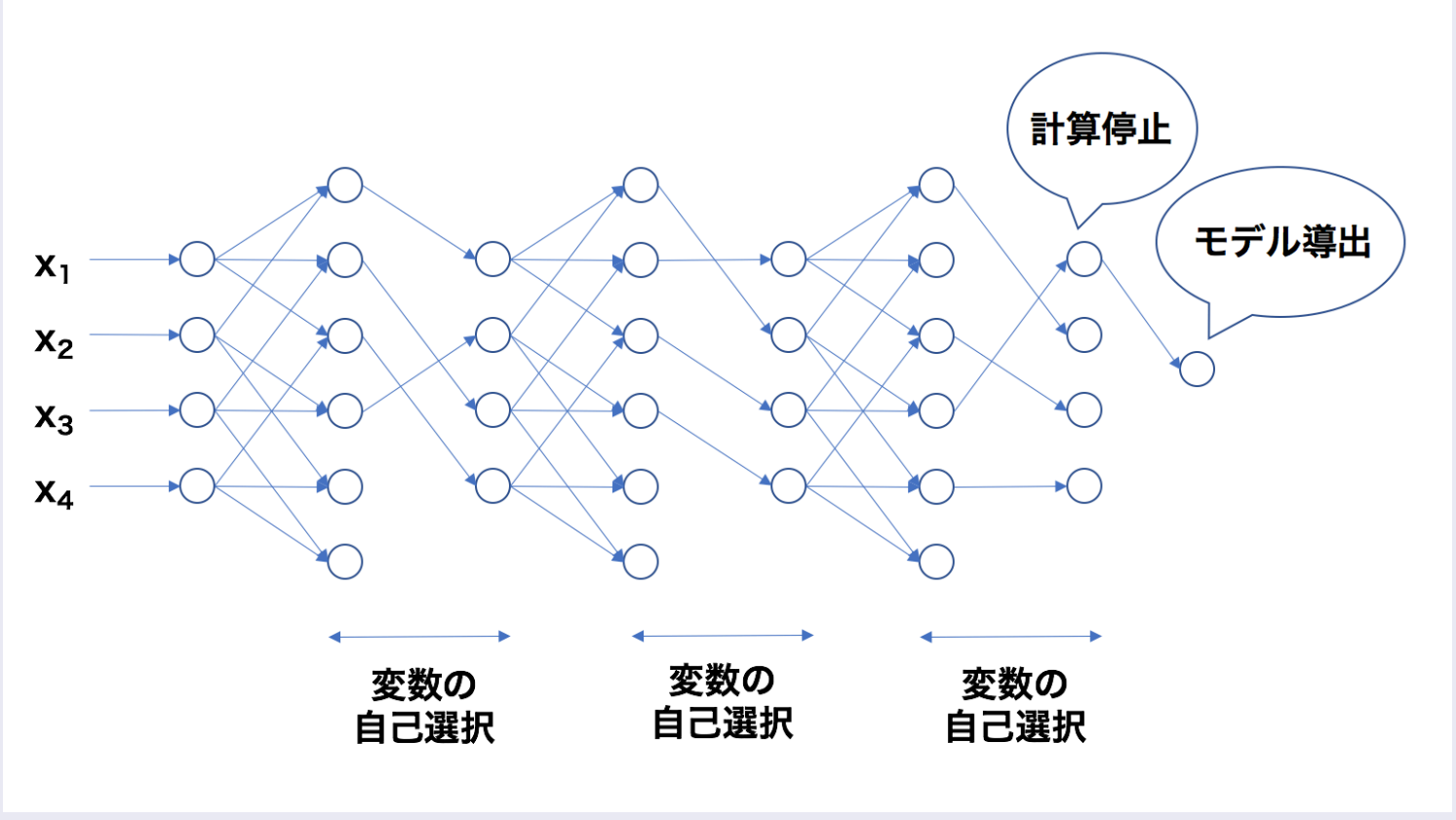


Figure: 同定されたモデル

2.14 同定結果

予測精度

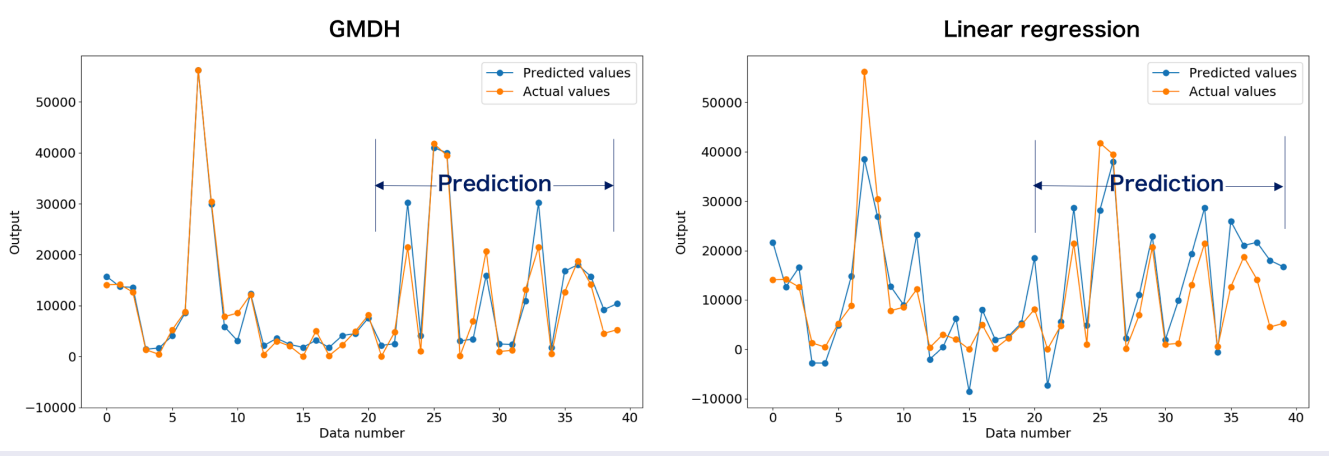


Figure: GMDH と線形回帰の予測精度

3.1 実験の概要

本研究の数値実験は以下の流れで行う.

実験の流れ

1. ツイートによる為替の影響の検証
2. 為替予測の検証 1(特徴量:為替の終値のみ)
3. 為替予測の検証 2(特徴量:為替の終値，テクニカル指標)
4. 為替予測の検証 3(特徴量:為替の終値，テクニカル指標，感情スコア)

3.2 ツイートによる為替の影響の検証

実験対象のデータ

トランプ大統領の 2018.9.1 から 2018.11.1 の 2ヶ月間のツイート

クラスタリングによる検証を行う前に，クラスタ数をエルボー法により決定する

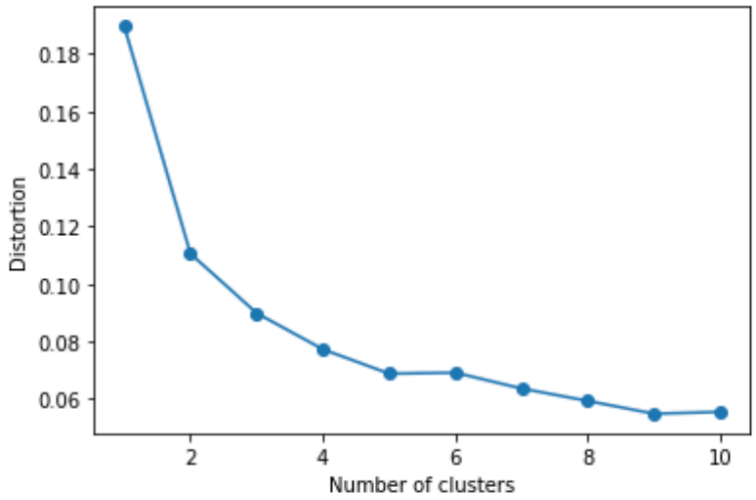


Figure: 2. エルボー法の分析結果

エルボー法は肘のように SSE 値が曲がった点が適しているクラスタ数である．以上の図より，クラスタ数を 4 に決定した．

3.3 k-Shapeによるクラスタリング

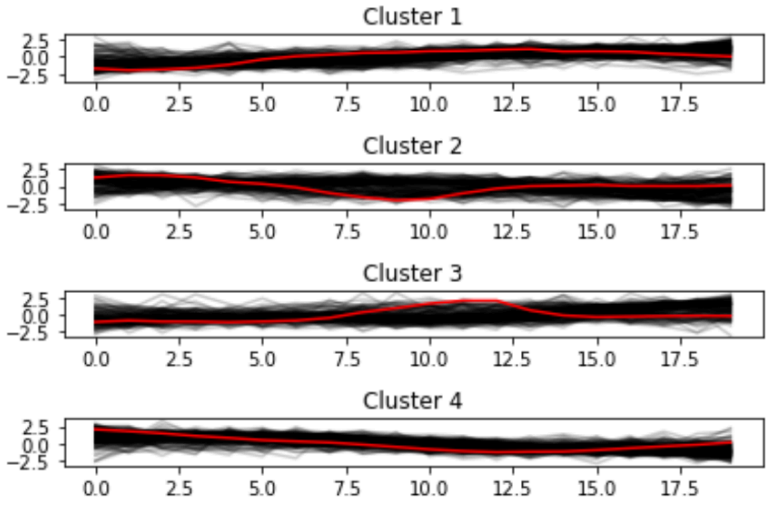


Figure: 3. k-Shapeによる分析結果

考察

1. Cluster2とCluster3は7.5分を過ぎたあたりからそれぞれ上下に大きく変動している
2. Cluster2とCluster3にはトランプ大統領のツイート直後の為替の価格の変動が多くみられた

3.3 ツイートの為替の影響

トランプ大統領がツイートした直後の為替の価格とランダムな日時の為替の価格の変動をクラスタごとに色分けしたグラフを以下に示す

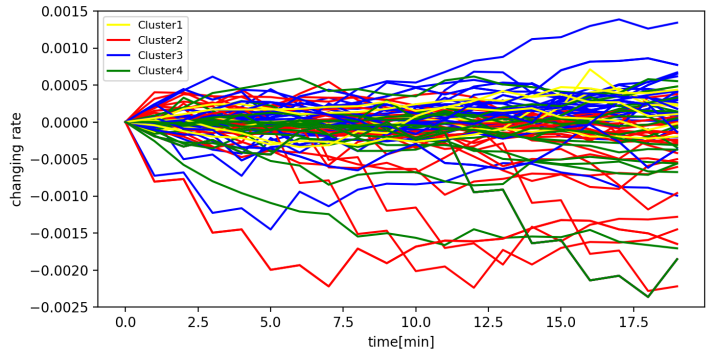


Figure: ツイート後のレート

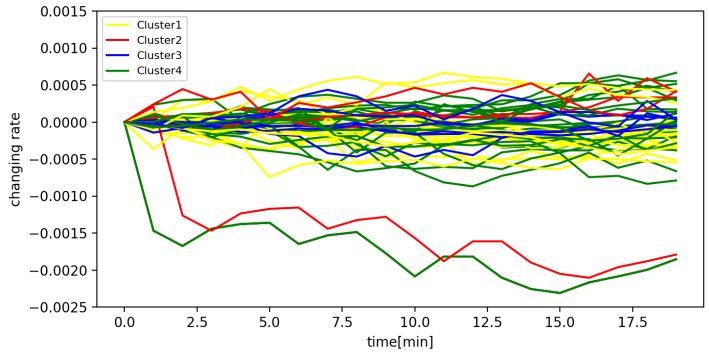


Figure: ランダムのレート

3.4 為替予測の検証 1(特徴量:為替の終値のみ)

実験対象のデータ

1. 学習データ：2018.9.1 ～ 2018.9.31
2. テストデータ：2018.10.10 ～ 2018.10.17

分析手法としては，LSTM を用いて以下にネットワークの仕様を示す.

実験対象のデータ

- 隠れ層:5
- 活性化関数:線形
- 最適化手法:Adam
- 誤差関数:平均絶対誤差 (MAE)

3.5 為替予測の検証 1(特徴量:為替の終値のみ) の結果

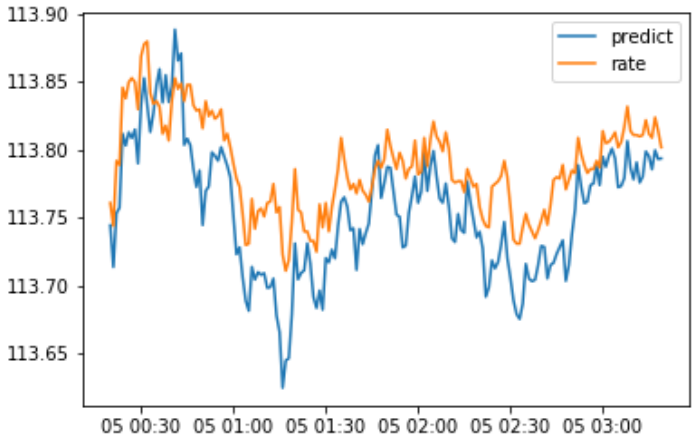


Figure: 6. 検証 1 の予測結果

考察

1. 予測と実測値の動きが全く逆の場合が多々見られた
2. 10 分後の予測は終値だけでは難しい可能性が大きい
   → 1 分後の値の予測は，決定係数が 0.85 で精度が高めであった

3.6 為替予測の検証 2(特徴量:為替の終値とテクニカル指標)

予測の際に有意だと思われる特徴量を重回帰分析によって抽出する.

抽出した特徴量

Table: p 値

特徴量	p 値
Close	0.041
perd(ストキャスティクス)	0.011
ADX(トレンド)	0.016
fama(MESA の適応型移動平均)	0.011
midpoint	0.010
htdcperiod(ヒルベルト変換 - Dominant Cycle Period)	0.02
signal(MACD)	0.013

3.7 為替予測の検証 2(特徴量:為替の終値とテクニカル指標)

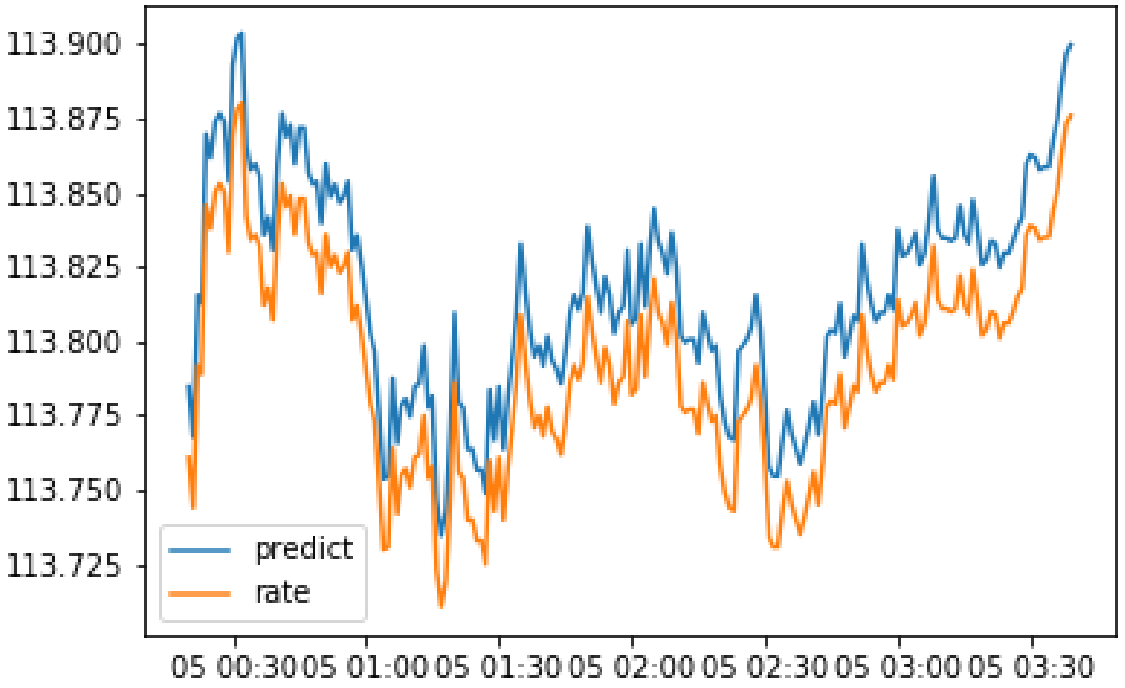


Figure: 7. 検証 2 の予測結果