経済に関するオルタナティブ・データを考慮した金融マーケット予測手法の開発

富山県立大学情報システム工学専攻　　指導教員：奥原浩之
1855001　麻生　到

1　はじめに

現在, ツイッターなどのマイクロブログには, 様々なニュースやそれに対する人々の反応が書かれており, その情報量は膨大かつ, 増加し続けている. この膨大な情報を実世界の動きを観測するためのソーシャルセンサーとして利用する研究の数は増加しており, 観測する対象を予め設定し, それについて詳細な分析を行ったものが多く見られる. 特に, 経済動向を分析対象としたものとして, ツイッターからキーワードを用いて株式に関する情報を収集し, 株価動向との関連の分析に取り組んだ事例があるなど, ツイッター情報は経済動向の分析に大いに用いられている.

また, 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となっている. その分析は金融経済現象へ応用されている. 現在では, 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.

既存研究としては, Bollen らが, ツイートを対象に OpinionFinder(OF) と Google-Profile of Mood States を用いて,「calmess」などの 6 つの心的状態を表す指数を抽出し, ダウ平均株価の予測を行った. しかし, 分析対象となるツイートは”I feel”, “I’ m”といった心的状態を明言したものに限られていることに加えて, ツイート情報はダウ平均株価の過去の数値データによる予測を補うものとして用いられている.

また，一般的には為替の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.

しかし，為替の価格のみであったりテクニカル指標を加えた予測では過去の情報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という問題点がある.

投資判断の分析手法として，ファンダメンタル分析とテクニカル分析がある．ファンダメンタル分析は，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である．一方，テクニカル分析では，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である.

従来のテクニカル分析を用いた予測手法では，ファンダメンタル分析のような市場外的要因が影響した市場の動きを把握できていないため市場外的要因も考慮した分析手法が必要であると考えられる.

そこで，本研究では, 従来の予測手法に加えて 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　分析手法

市場外的要因を考慮した予測手法の提案の流れを以下に示す.

[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

2.1　ツイートによる為替の影響

まず，ツイートが為替に影響することがあるのかということを検証する必要がある．本研究では k-Shape によるクラスタリングを行うことによって為替の影響を検証する.

2.1.1　ツイートの取得

本研究では，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

図 1: 1. Twitter API により取得したツイートの例

2.1.2　k-Shape によるクラスタリング

k-Shape は，時系列の形状に着目した時系列クラスタリング手法である．このクラスタリング手法では，距離尺度として規格化した相互相関を用いている.

時系列データをクラスタリングする際には，何に重きを置いてクラスタリングするかが重要となる．この手法ではスケールと時間軸に重きを置いてクラスタリングを行う．つまり，データ同士がスケールした際に性質が近いかどうかと時間軸をずらした時に似ているかどうかを判断指標としている.

また，この手法では距離尺度として Shape-based distance (SBD) を用いている．一般的に距離尺度として用いられる Euclidean Distance(ED) と Dynamic Time Warping(DTW) と異なり SBD は相互相関を用いている.

2 つの時系列データ $\boldsymbol{x}$ と $\boldsymbol{y}$ における SBD は以下のように求めれられる

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

ただし，

$$CC_w = \mathcal{F}^{-1}\{\mathcal{F}(\boldsymbol{x}) * \mathcal{F}(\boldsymbol{y})\}$$

$$R_k(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} \sum_{l=1}^{m-k} x_{l+k}\cdot & (k \geq 0) \\ R_{-k}(\boldsymbol{y},\boldsymbol{x}) & (k \leq 0) \end{cases}$$

そして，SBD を用いて時系列データ $\boldsymbol{x}$ と各クラスタの重心ベクトル $\boldsymbol{\mu}_k$ との距離 $\boldsymbol{\mu}_k^*$ を求める.

$$\boldsymbol{\mu}_k^* = \arg\max_{\boldsymbol{\mu}_k} \sum_{\boldsymbol{x}_i \in P_k}\left(\frac{\max_w CC_w(\boldsymbol{x}_i,\boldsymbol{\mu}_k)}{\sqrt{R_0(\boldsymbol{x}_i,\boldsymbol{x}_i)\cdot R_0(\boldsymbol{\mu}_k,\boldsymbol{\mu}_k)}}\right) \tag{2}$$

クラスタリングまでの流れと以上のような数式を用いて，以下のように k-mean 法とアプローチ方法は同じである.

[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

2.2 為替予測の分析手法

本研究では，為替予測の分析手法として Long short-term memory(LSTM) を用いる．LSTM は名前の通り，短期記憶を長期に渡り適応できることから連続データの分析に用いられることが多い．そのため，為替のデータも連続データでありこの分析手法が適していると考え，本研究でも分析に用いる．

2.2.1 LSTM

従来の Recurrent Neural Network(RNN) では， 長期の系列における勾配消失・発散が問題となっていた．そこで，LSTM では入力と隠れ層，出力の重みを調節することにより長期の記憶を可能としている．

RNN では活性化関数に以下のような単純な tanh 関数を用いていたことが勾配の問題の原因であった．

$$h_t = tanh(Wx_t + Uh_{t-1} + b) \tag{3}$$

ここで，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．

$$i_t = \sigma(w_i * [x_t, h_{t-1}] + b_i) \tag{4}$$
$$f_t = \sigma(w_f * [x_t, h_{t-1}] + b_f) \tag{5}$$
$$o_t = \sigma(w_o * [x_t, h_{t-1}] + b_o) \tag{6}$$
$$u_t = tanh(w_u * [x_t, h_{t-1}] + b_u) \tag{7}$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot u_t \tag{8}$$
$$h_t = o_t * tanh(c_t) \tag{9}$$

2.3 テクニカル指標

本研究では，為替の予測に用いる特徴量として終値だけではなくテクニカル指標も用いる．そこで，予測の入力に用いるテクニカル指標を以下に示す．

- EMA
- ボリジャーバンド
- ストキャスティクス
- MACD
- シグナル
- William's % R
- アルティメット・オシレータ
- ADX
- 加重移動平均 (WMA: Weighted Moving Average)
- ２重指数移動平均 (DEMA: Double Exponential Moving Average)
- ３重指数移動平均 (TEMA: Triple Exponential Moving Average)
- 三角移動平均 (TMA: Triangular Moving Average)
- Kaufman の適応型移動平均 (KAMA: Kaufman Adaptive Moving Average)
- MESA の適応型移動平均 (MAMA: MESA Adaptive Moving Average)
- トレンドライン (Hilbert Transform - Instantaneous Trendline)
- MidPoint over period
- 変化率 (ROC: Rate of change Percentage)
- モメンタム (Momentum)
- RSI: Relative Strength Index
- APO: Absolute Price Oscillator
- PPO: Percentage Price Oscillator
- CMO: Chande Momentum Oscillator
- ヒルベルト変換

2.4 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．

2.4.1 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．

$$y_t^i = b_0 + b_1 x_1^i + b_1 x_2^i + \cdots b_n x_n^i \tag{10}$$

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．

2.5 センチメント分析による感情スコアの算出

本研究では，Twitter などの SNS の情報が為替に影響することを考慮して分析を行う．そこで，センチメント分析で感情スコアを算出し，その値を為替を予測する際の入力として扱う．そのため，本研究の特徴量はテクニカル指標だけではなく Twitter などの SNS のテキストをセンチメント分析を行い感情スコアを算出した．

2.5.1 センチメント分析

Web 上のユーザー生成コンテンツの量は，主に SNS やブログ，マイクロブログサイト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他のプラットフォームの出現したことで，ますます急速に増加している．また，それらのユーザー生成コンテンツは意見や感情が豊富である．そのユーザーの意見は感情は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．しかし，オンライン上のそれらの情報は Web 上に無数にあり手作業で分析することは不可能である．そのため，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．そこで，センチメント分析により膨大な人々の感情を把握できることが期待されている．

従来のセンチメント分析は，感性辞書を用いたポジネガ分析など感情の軸が 1 つしかなく表現力に乏しいといった問題があった．そこで，心理学者プルチックなどが提案している感情の環に基づいた分析を行うことで，より人々の感情を表すことができる分析が必要であると考えられる．

本研究では，センチメント分析により SNS 上の情報を為替予測に適用することで市場の動きを把握することにより，従来の為替予測では捉えきれなかった市場の状況を考慮し精度の向上を図る．本研究の手法では，Niko Colneric らが提案している Twitter のデータを元に作成したセンチメント分析を参考にシステムを構築した．Niko Colneric らは，Twitter の情報を元に深層学習を用いてセンチメント分析を行なっている．

2.6 主成分回帰分析を用いる GMDH

本研究では，為替予測手法として主成分回帰分析を用いた GMDH を適用する．この手法は，評価基準を用いて変数の逐次選択を行い最適な部分表現式を自己選択する GMDH である．

システムの完全表現式として，Kolmogorov-Gavor の多項式

$$\Phi = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \cdots \tag{11}$$

を想定する．そして，システムの部分表現式は 2 変数の 2 次多項式を以下に示す．

$$y_k = b_{0k} + b_{1k} x_i + b_{2k} x_j + b_{3k} x_i x_j + b_{4k} x_i^2 + b_{5k} x_j^2 \tag{12}$$

この部分表現式に対して変数選択の評価規準を用いて変数選択を行い，多重共線性を起こさない最適な部分表現式を自己選択する．本研究では評価基準として情報量規準 AIC を用いる．AIC は以下のように表される．

$$AIC = n \ln S_m^2 + 2(m+1) + C \tag{13}$$

ここで，

$$S_m^2 = \frac{1}{n \sum_{\alpha=1}^{n} (\Phi_\alpha - y_{ka})^2} \tag{14}$$

また，n はデータ数，m は変数の個数，C は m に無関係な定数，Φ_α は α 番目のデータの出力値，y_{ka} は中間変数 y_k の α 番目のデータに対応する値である．

システムの完全表現式は，最適な部分表現式を積み重ねて構成し，その次数は各選択層において変数を組み合わせることにより増加する．層を通過して計算を続けていくと多重共線性を起こす変数の組み合わせが増加して最適な部分表現式に含まれる変数の数が減少する．最終層では，すべての 2 変数の組み合わせについて最適な部分表現式の形が，

$$y_k = b_{0k} + b_{1k} x_i \tag{15}$$

となり，前層と同じ中間変数を発生し，予測精度が改善されなくなるため部分表現式の積み重ねを停止する．

このように，評価規準 AIC を用いて変数の逐次選択を行い最適な部分表現式を自己選択する GMDH を構成できる．

2.6.1 主成分回帰分析による最適部分表現式の作成

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．

$$\boldsymbol{z} = C \cdot \boldsymbol{x}_{ij} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{15} \\ c_{21} & c_{22} & \cdots & c_{25} \\ \vdots & \vdots & \ddots & \vdots \\ c_{51} & c_{52} & \cdots & c_{55} \end{bmatrix} \begin{bmatrix} x_i \\ x_j \\ x_i \cdot x_j \\ x_i^2 \\ x_j^2 \end{bmatrix} \tag{16}$$

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．

$$R \cdot C = C \cdot \Lambda \tag{17}$$

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．そして，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．

$$y_k = \boldsymbol{z}^{\mathrm{T}} \cdot \boldsymbol{d}_k = \begin{bmatrix} z_1 z_2 \cdots z_5 \end{bmatrix} \begin{bmatrix} d_{k1} \\ d_{k2} \\ \vdots \\ d_{k5} \end{bmatrix} \tag{18}$$

$\boldsymbol{d}_k$ 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．

$$Z^{\mathrm{T}} \cdot \boldsymbol{y}_k = (Z^{\mathrm{T}} \cdot Z) \cdot \boldsymbol{d}_k \tag{19}$$

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

$$d_{ki} = \frac{(Z^{\mathrm{T}} \cdot \boldsymbol{y}_k)_i}{r_i} \tag{20}$$

また，r_i $(i = 1, 2, \cdots, 5)$ は対角行列 $(Z^{\mathrm{T}} \cdot Z)$ の第 (i, i) 成分である．

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．変数 $\boldsymbol{z}$ が直交化されて無相関であるために変数の逐次選択方法としては，どの方法を採用しても作成されるモデルは同じになる．

AIC の計算に用いる残差の自乗平均は，

$$S_m^2 = S_{m-1}^2 - \frac{\frac{(Z^{\mathrm{T}} \cdot \boldsymbol{y}_k)_i}{d_{ki}}}{n} \tag{21}$$

$$S_0 = \frac{\boldsymbol{y}_k^{\mathrm{T}} \cdot \boldsymbol{y}_k}{n} \tag{22}$$

となり，第 2 項は変数 z_i を部分表現式の中に取り込んだ場合の自乗平均の減少量を示す．部分表現式の中に取り込んだ場合の残差自乗平均和を計算して，残差自乗和を最も小さくする変数から部分表現式の中に順次取り込む．そして，AIC の値が増加するときに変数の取り込みを停止する．このようにして，有用な変数のみを用いて最適な部分表現式を作成する．

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．

2.6.2 多層構造の計算停止方法

GMDH では，層を通過して計算を続けていくと多重共線性を起こす変数の組み合わせが数多く発生するため，主成分分析を行うときに変数の次元が縮小する．変数の次元は 5 次から 2 次い縮小して，最後の固有値の絶対値の大きいものが 2 つ残り，他の 3 つの固有値はすべて 0 に近く．

そこで，多重共線性を起こす中間変数同士を組み合わせても予測精度の改善が行われていないため，多重共線性の発生により変数の次元が縮小する層で多層構造の層の積み重ねを打ち切り，計算を停止する．

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，

$$\frac{\lambda_{k,max1} + \lambda_{k,max2}}{\sum_{i=1}^{5} \lambda_{k,i}} > E \tag{23}$$

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．

3 数値実験

本研究の数値実験は以下の流れで行う．

[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)

3.1 ツイートによる為替の影響

クラスタリング手法である k-Shape を用いてツイートが為替に影響しているかを検証する．

実験に用いるデータは，トランプ大統領の 2018.9.1 から 2018.11.1 の 2ヶ月間のツイートを対象として分析を行った．また，クラスタ数はエルボー法で決定した．エルボー法での結果を以下の図 2 に示す．

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．

クラスタ数を決定した後，実際にトランプ大統領がツイートした直後 20 分間の為替の価格に対してクラスタリングを行った．分析結果を以下の図 3 に示す．

以上の図 3 より，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ことがわかる．また，クラスタリングの結果から 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．よって，トランプ大統領がツイートした直後から 7.5 分後あたりから為替が変動することが多いと考えられる．

また，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図 4 と図 5 にそれぞれ示す．

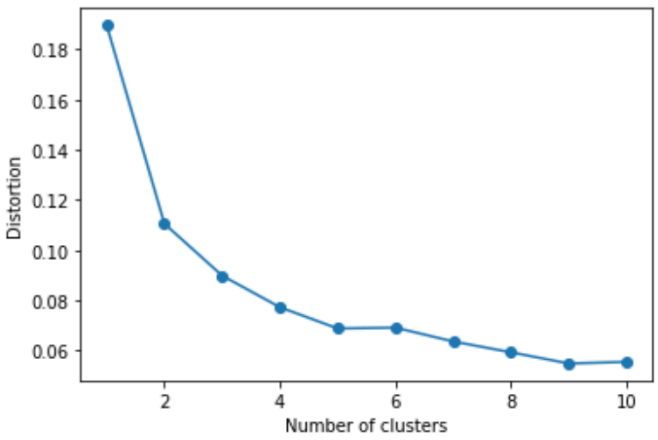

図 2: 2. エルボー法の分析結果

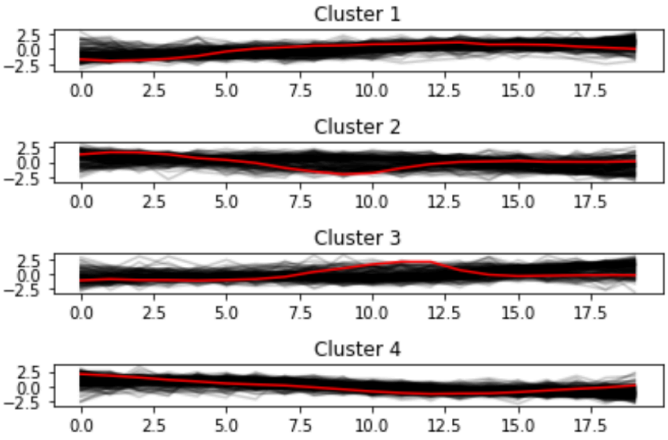

図 3: 3. k-Shape による分析結果

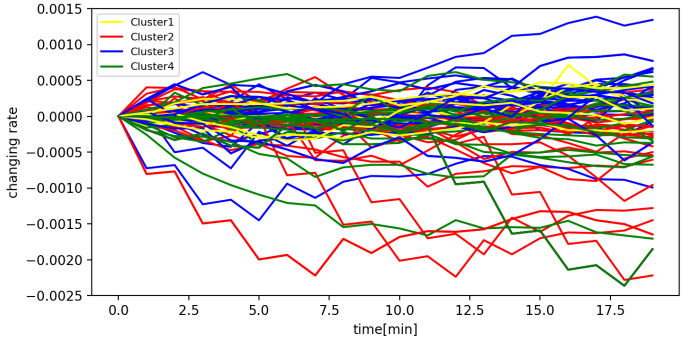

図 4: 4. トランプ大統領がツイートした直後の為替の変動

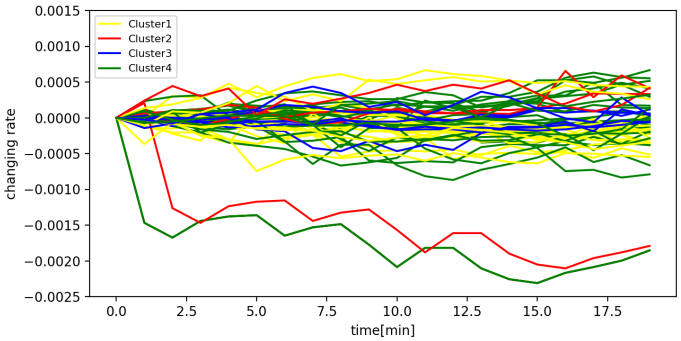

図 5: 5. ランダムな日時の為替の変動

3.2 為替予測の検証 1(特徴量:為替の終値のみ)

次に，実際に為替の価格を終値のみで予測を行う．実験には，学習データとして 2018.9.1 から 2018.9.31 を用いて，テストデータとしては 2018.10.10 から 2018.10.17 の期間のデータを用いた．

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．

- 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)

以上のネットワークを用いて予測を行い，その結果を図 6 に示す．

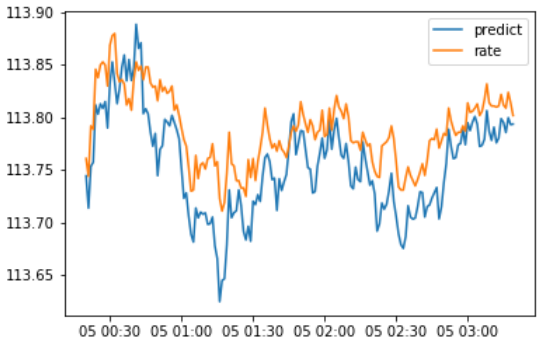

図 6: 6. 検証 1 の予測結果

3.3 為替予測の検証 2(特徴量:為替の終値とテクニカル指標)

まず，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．

以上の表の有意となる特徴量を検証 1 の特徴量に追加して検証を行う．

3.4 為替予測の検証 3(特徴量:為替の終値とテクニカル指標，感情スコア)

特徴量となる感情スコアをツイートからセンチメント分析により求める．例として，「A years ago, A star was Born, and here we are 6 times pink platium」といったツイートに対して本研究で用いるセンチメント分析によって感情スコアを求め，その結果を図 8 に示す．

図 8 のように，センチメント分析によって 6 つの感情軸に対してスコアを導出して，1 ツイートあたりに感情スコアを求めていく．

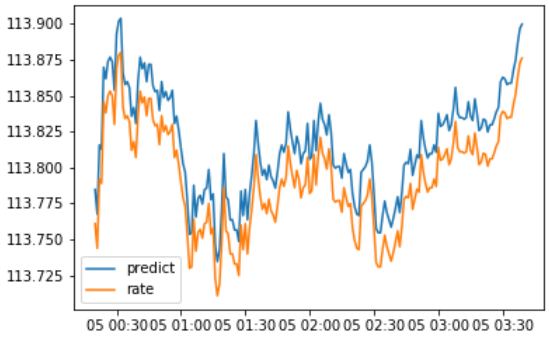

図 7: 7. 検証 2 の予測結果

表 1: 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

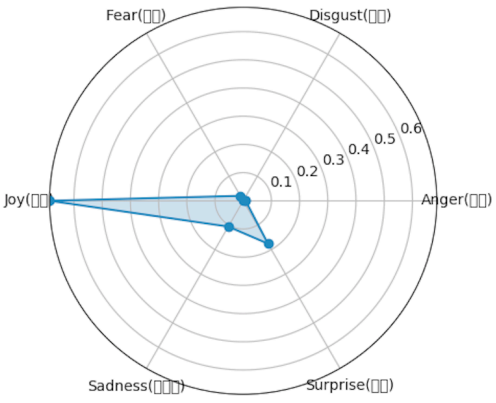

図 8: 8. 感情スコアの例