テンソルを用いたマルチモーダル学習による株式価格の変動予測

A tensor based multimodal learning for predicting stock price fluctuations

笹尾知広 [1*]　中田和秀 [1]
Tomohiro Sasao[1]　Kazuhide Nakata[1]

[1] 東京工業大学工学院経営工学系
[1] Graduate Major in Industrial Engineering and Economics, Tokyo Institute of Technology

Abstract: 近年では株価データと他のデータを組み合わせたマルチモーダル学習が頻繁に行われている．しかし，複数の情報源からそれぞれ特徴量ベクトルを作成した後，それらを単純に連結させたベクトルを用いて予測するのでは，情報源間の共起関係が十分に考慮されないという問題があった．このような問題に対して，高次元テンソルを用いて対処しようとする研究がされている．本研究はそれらの研究を発展させることで，複数の情報源を使ったときに，株式価格の変動をより精度よく予測するためのテンソルを用いたマルチモーダル学習フレームワークを提案する．そして，３種類の情報源としてトムソン・ロイター社が発信した記事，トムソン・ロイター市場心理指数（TRMI）のセンチメントデータ，ニューヨーク証券取引所の株価データを用いて評価実験を行った．その結果，一日後の株価の予測・残差リターンの正負を予測する実験において，提案手法により予測性能の有意な改善が見られた．

1　はじめに

今日ではビックデータを用いた研究が注目されており，様々な研究や開発に適用されている．株式市場予測の分野でもインターネット上で取得することができる SNS やニュース記事などのオルタナティブデータを用いる研究がされてきた．一般的に複数の情報源を用いた機械学習のことをマルチモーダル学習といい，近年では株価データと他のデータを組み合わせたマルチモーダル学習による株価予測（回帰問題）や株価の変動方向予測（二値分類問題）が頻繁に行われている．しかし従来の手法は複数の情報源から特徴量ベクトルを作成した後，それらを一つのベクトルとして連結させて処理することがほとんどであり，そのような連結させたベクトルを用いた予測では，情報源間の共起関係が十分に考慮されないという問題があった．このような問題に対して，特徴量の組み合わせに高次元テンソルを用いて対処しようとする研究がさまざまな分野でされており，株式市場予測の分野では Li ら [6] が「テンソルを用いたマルチモーダル学習フレームワーク」という手法を提案した（図 1）．

本研究では複数の情報源を使ったときに株式価格の変動をより精度よく予測するための,「改良したテンソルを用いたマルチモーダル学習フレームワーク」を提案する．提案手法は既存手法と比べて，補助行列の作成と予測手法に工夫を加えたフレームワークになっている．具体的に予測手法に関して，学習時間が短い予測・二値分類問題を考慮した予測を行うために，画像・動画分析の分野で提案された２種類のテンソル予測モデル（リッジ回帰ベースのテンソル回帰モデル [3]，ロジスティック回帰ベースのテンソル分類モデル [11]）を導入し，アルゴリズムの終了条件に改良を行った．フレームワーク中の処理に関して，補助行列を作成するために解く必要のある最小化問題において，新たに特殊な正則化項を加えた最小化問題を定式化した．

実験では株式データ，ニュースデータ，センチメントデータを用いた株価予測と株価の変動方向予測の実験を行った．その結果それぞれの情報源から作成した特徴量を１つのベクトルとして連結して予測を行うのではなく，提案したテンソルを用いたマルチモーダル学習のフレームワークを用いることによって，より精度良く予測できることを示した．

*連絡先： 東京工業大学工学院経営工学系
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail:sasao.t.aa@m.titech.ac.jp

2　関連研究

機械学習を用いた株式価格の変動予測において，さまざまなデータや機械学習手法が使われてきた．Liu ら [8] は株式データとニュースデータの２つを用いて，それぞれから特徴量を作成し LSTM モデルで，Apple 社の株価の変動方向予測を行った．Bollen ら [1] は株式データと Twitter の投稿で表現された国民感情を DJIA 指数の変動方向予測に利用した．Suwa ら [10] はヤフー株式掲示板のデータのトピック分析を行い，トピック分析から得られた特徴量と株式データを用いて，VI 指数の変動方向予測を行った．また Li ら [7] は株式データ，ニュースデータ，掲示板データの３つを組み合わせて個別銘柄の株価予測を行った．このようにさまざまなデータを用いた将来の株式価格やインデック指標，またその変動方向予測が行われている．

一般的に複数の情報源を用いた機械学習のことをマルチモーダル学習といい，近年では株式データと他のデータを組み合わせたマルチモーダル学習が頻繁に行われている．しかし従来の手法は複数の情報源から特徴量ベクトルを作成した後，それらを一つのベクトルとして連結させて処理することがほとんどであった．そのような連結させたベクトルを用いた予測では，情報源間の共起関係が十分に考慮されないという問題がある．

この複数の情報源間の組み合わせ方の問題に着目し，特徴量の組み合わせに高次元テンソルを用いた研究が様々な分野でされている．特に株式市場予測の分野では，Li ら [6] が株価データ，ニュースデータ，掲示板データから作成した特徴量を 3 次元テンソルを用いて組み合わせ予測を行う「テンソルを用いたマルチモーダル学習フレームワーク」を提案し，個別銘柄の株価予測で高次元テンソルによる組み合わせの有

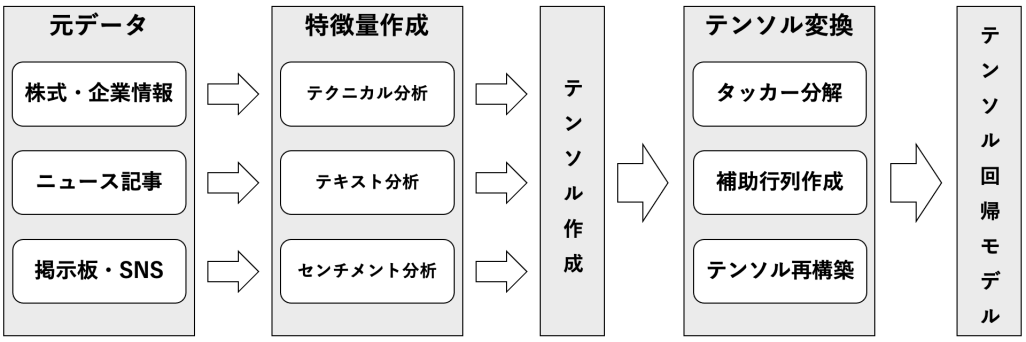

図 1: 「テンソルを用いたマルチモーダル学習フレームワーク」の概要図

効性を示した．さらに Huang ら [4] はこのモデルをさらに改良し，他の企業の情報も加味した手法を提案した．

しかし Li ら [6] と Huang ら [4] のモデルには，大きく分けて２つの問題点がある．１つ目は予測手法に関して，従来のモデルではサポートベクトル回帰をベースにしたモデルを用いており学習時間が遅いこと，また回帰問題しか想定しておらず二値分類問題が考えられていないことである．２つ目はフレームワークを実行する中で解く必要のある最小化問題において，求める最適解が一意に定まらず不安定なモデルであること，また制約条件が強すぎるためうまく学習できない場合があることである．

3 改良したテンソルを用いたマルチモーダル学習フレームワーク

本章では提案する「改良したテンソルを用いたマルチモーダル学習フレームワーク」について説明する（図 2）．この手法は Li ら [6] の手法を発展させたフレームワークであり，大きく分けて「テンソル作成」,「テンソル変換」,「テンソル予測」の３つの部分から構成されている．

3.1 テンソル作成

テンソルを用いたマルチモーダル学習フレームワークでは，企業 s，日付 t 毎に情報源からテンソル $\mathcal{X}_{s,t}$ を作成する．テンソルの階数は扱う情報源の数に対応している．本研究においては，企業データ，ニュースデータ，センチメントデータの３つの情報源を用いるため３階テンソルを用いており，以降３階のテンソルとして考える．テンソルは以下のような手順で作成する．

まず初めに企業データ，ニュースデータ，センチメントデータから企業 s，日付 t における特徴量ベクトル $\boldsymbol{f}_1^{s,t} \in R^{I_1}, \boldsymbol{f}_2^{s,t} \in R^{I_2}, \boldsymbol{f}_3^{s,t} \in R^{I_3}$ を作成する．テンソル $\mathcal{X}_{s,t} \in R^{I_1 \times I_2 \times I_3}$ の各要素を $x_{i,j,k}^{s,t}$ とすると，各情報源の特徴量を次のようにテンソルに格納する．

- $x_{i,1,1}^{s,t}$ ，$1 \leq i \leq I_1$ ： 株価データの特徴量 $\boldsymbol{f}_1^{s,t}$
- $x_{2,j,2}^{s,t}$ ，$1 \leq j \leq I_2$ ： ニュースデータの特徴量 $\boldsymbol{f}_2^{s,t}$
- $x_{3,3,k}^{s,t}$ ，$1 \leq k \leq I_3$ ： センチメントデータの特徴量 $\boldsymbol{f}_3^{s,t}$
- それ以外は 0

このようなテンソル $\mathcal{X}_{s,t}$ を，各企業・日付毎に作成する．

3.2 テンソル変換

作成した企業 s，日付 t のテンソル $\mathcal{X}_{s,t}$ から①タッカー分解，②補助行列作成，③テンソル再構築の３つの手順を介して，情報源間の共起関係の増強・ノイズが除去された新しいテンソル $\tilde{\mathcal{X}}_{s,t} \in R^{I_1 \times I_2 \times I_3}$ を作成する．具体的には，類似した結果になる２つのテンソルについて，それらのテンソルを分解して得られる因子行列同士が類似するような処理を行う．

まずはじめに，タッカー分解 [5] をもちいてテンソル $\mathcal{X}_{s,t}$ を次のように分解する．

$$\mathcal{X}_{s,t} = \mathcal{C}_{s,t} \times_1 U_1^{s,t} \times_2 U_2^{s,t} \times_3 U_3^{s,t} \tag{1}$$

ここで $U_k^{s,t} \in R^{I_k \times D_k}$ $(D_k \leq I_k,\ k = 1,2,3)$ は企業データ，ニュースデータ，センチメントデータそれぞれを表現する因子行列になっており，$\mathcal{C}_{s,t} \in R^{D_1 \times D_2 \times D_3}$ はそれぞれの情報源間の関係の強さを表すコアテンソルである．

続いて情報源間の共起関係の増強・ノイズを減少させるような，各因子に対応する補助行列 $V_{s,k} \in R^{I_k \times I_k}$ $(k = 1,2,3)$ を導入する．Li ら [6] はこの補助行列 $V_{s,k}$ を求めるために次のような最小問題を定式化した．

$$\min_{V_{s,k}} \ J(V_{s,k}) = \sum_{i=1}^{N} \sum_{j=i}^{N} ||V_{s,k}^T U_k^{s,i} - V_{s,k}^T U_k^{s,j}||^2 w_{s,i,j} \tag{2}$$

$$s.t. \ \sum_{i=1}^{N} \sum_{j=i}^{N} ||V_{s,k}^T U_k^{s,i}||^2 w_{s,i,j} = 1 \tag{3}$$

$$w_{s,i,j} = \begin{cases} 1 & \text{if } i \leq j \text{ and } |y_{s,i} - y_{s,j}|/y_{s,j} \leq 5\% \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

ここでの目的は，モード k の因子行列 $U_k^{s,i}|_{i=1}^N$ の本来の性質を保持しつつ，終値が類似している日付 i, j 同士の２つの行列 $V_{s,k}^T U_k^{s,i}$，$V_{s,k}^T U_k^{s,j}$ の値の差を小さくすることである．式 (2) で時刻間の類似性を考慮しており，解の個数を制限させ $V_{s,k}$ のオーバーフィッティングを避けるために式 (3) のような制約式を設けている．

本研究ではこの最小化問題に２つの問題があることを数式的に示すことができた．詳しい証明は割愛するが具体的に１つ目は，最小化問題の解が集合の形で与えられ，一意の解に定まらず不安定なもでるであること．２つ目は，制約式 (3) がデータ数 N や $w_{s,i,j}$ の密度によって変化しないため，データ数 N が多い場合や $w_{s,i,j}$ が密な行列の場合，$V_{s,k}$ が非常に小さい数値になることである．

そこで本研究ではこの補助行列 $V_{s,k}$ を求めるために，次のような制約なしの凸関数最小化問題を新たに定式化した．

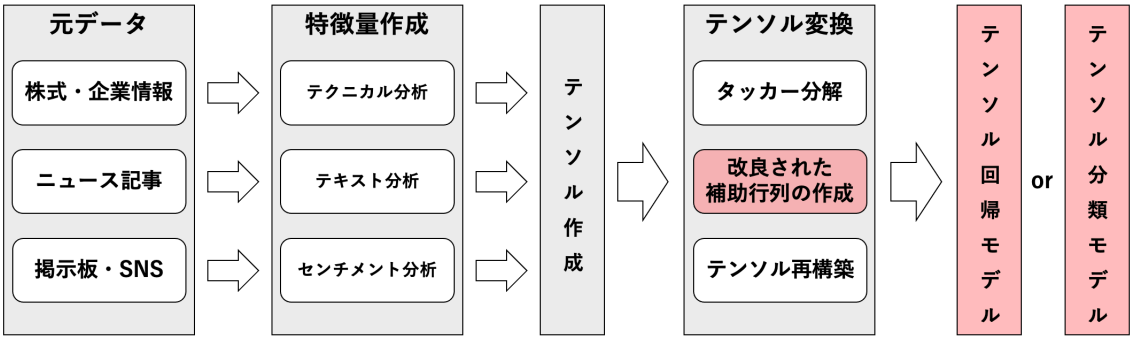

図 2: 「改良したテンソルを用いたマルチモーダル学習フレームワーク」の概要図

$$\min_{V_{s,k}} \ J(V_{s,k}) = \frac{\mu}{\mathrm{sum}W_s}\sum_{i=1}^{N}\sum_{j=1}^{N}||V_{s,k}^T U_k^{s,i} - V_{s,k}^T U_k^{s,j}||^2 w_{s,i,j} + \frac{\gamma}{\mathrm{sum}E_s}\sum_{t=1}^{N}\sum_{c=1}^{C}||V_{s,k}^T U_k^{s,t} - V_{s,k}^T U_k^{c,t}||^2 z_{s,c} e_{t,s,c} + \frac{1}{\mathrm{sum}W_s}\sum_{i=1}^{N}\sum_{j=1}^{N}||V_{s,k}^T U_k^{s,i} - I_{s,k} U_k^{s,i}||^2 w_{s,i,j} \quad (5)$$

$$w_{s,i,j} = \begin{cases} 1 & \text{if } i \le j \text{ and } |y_{s,i} - y_{s,j}|/y_{s,j} \le 5\% \\ 1 & \text{if } i > j \text{ and } |y_{s,i} - y_{s,j}|/y_{s,i} \le 5\% \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

または二値分類問題の場合，予測する二値数 $y^*_{s,i} = \{+1, -1\}$ に対して

$$w_{s,i,j} = \begin{cases} 1 & \text{if } y^*_{s,i} = +1 \text{ and } y^*_{s,j} = +1 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$z_{s,c} : \text{企業 } s, c \text{ の類似度} \quad (8)$$

$$e_{s,t,c} = \begin{cases} 1 & \text{時刻 } t \text{ の銘柄 } s, c \text{ のデータが存在} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

式 (5) の第 1,2 項目で時刻間・企業間の類似性を考慮しており，第 3 項目で $V_{s,k}$ が単位行列に近くなるという正則化項を加えている．さらに式 (5) は制約なしの凸関数最小化問題になっているため，$(D_{s,k} - G_{s,k})$，$G_{s,k}$，$Q_{s,k}$ のいずれかが正定値行列という仮定をおけば，以下のように解析的に一意の解を求めることができる．

$$V_{s,k} = \frac{1}{\mathrm{sum}W_s}\left\{\frac{2\mu}{\mathrm{sum}W_s}(D_{s,k} - G_{s,k}) + \frac{1}{\mathrm{sum}W_s}D_{s,k} + \frac{\gamma}{\mathrm{sum}E_s}Q_{s,k}\right\}^{-1} D_{s,k} \quad (10)$$

$$D_{s,k} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{s,i,j} U_k^{s,i} U_k^{s,i^T}$$

$$G_{s,k} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{s,i,j} U_k^{s,i} U_k^{s,j^T}$$

$$Q_{s,k} = \sum_{t=1}^{N}\sum_{c=1}^{C}\left(U_k^{s,t} - U_k^{c,t}\right)\left(U_k^{s,t} - U_k^{c,t}\right)^T e_{t,s,c} z_{s,c}$$

最後に，タッカー分解で得られたコアテンソルと各情報源の因子行列，補助行列を用いて，次のようにテンソルを再構築する．

$$\tilde{\mathcal{X}}_{s,t} = \mathcal{C}_{s,t} \times_1 (V_{s,1}^T U_1^{s,t}) \times_2 (V_{s,2}^T U_2^{s,t}) \times_3 (V_{s,3}^T U_3^{s,t}) \quad (11)$$

このように各企業 s，各日付 t について新しいテンソル $\tilde{\mathcal{X}}_{s,t}$ を作成する．新しく作られたテンソル $\tilde{\mathcal{X}}_{s,t} \in R^{I_1 \times I_2 \times I_3}$ は，補助行列 $V_{s,k}$ によって，元のテンソル $\mathcal{X}_{s,t}$ よりも情報源間の共起関係の増強・ノイズ除去されたテンソルとなる．

3.3 テンソル予測

提案する「改良したテンソルを用いたマルチモーダル学習フレームワーク」ではテンソルを共変量とする回帰問題と二値分類問題を扱う．各企業 s，各観測値 $t = 1, \ldots, N$ ごとに終値である実数 $y_{s,t}$ と共変量 $\tilde{\mathcal{X}}_{s,t}$ が与えられているとし，次のような回帰問題を考える．

$$\hat{y}_{s,t} = \langle \tilde{\mathcal{X}}_{s,t}, \mathcal{W}_s \rangle + b_s \quad (12)$$

$$= \tilde{\mathcal{X}}_{s,t} \times_1 \boldsymbol{w}_s^{(1)} \times_2 \boldsymbol{w}_s^{(2)} \times_3 \boldsymbol{w}_s^{(3)} + b_s \quad (13)$$

また二値分類問題であれば，日次収益率の正負である二値数 $y^*_{s,t} \in \{+1, -1\}$ と共変量 $\tilde{\mathcal{X}}_{s,t}$ が与えられているとし，次のような二値数の条件つき確率を考える．

$$\Pr(y^*_{s,t} = +1|\tilde{\mathcal{X}}_{s,t}) = \frac{1}{1 + \exp\left(-\langle \mathcal{W}_s, \tilde{\mathcal{X}}_{s,t}\rangle - b_s\right)} \quad (14)$$

$$= \frac{1}{1 + \exp\left(-\tilde{\mathcal{X}}_{s,t} \times_1 \boldsymbol{w}_s^{(1)} \times_2 \boldsymbol{w}_s^{(2)} \times_3 \boldsymbol{w}_s^{(3)} - b_s\right)} \quad (15)$$

ここでの目的は，回帰係数テンソル $\mathcal{W}_s \in R^{I_1 \times I_2 \times I_3}$，切片 $b_s \in R$ を推定し，新しい入力 $\tilde{\mathcal{X}}$ に対して y，$\Pr(y^*_{s,t} = +1|\tilde{\mathcal{X}}_{s,t})$ を予測することである．

従来の「テンソルを用いたマルチモーダル学習フレームワーク」ではこのサポートベクトル回帰をベースとした予測モデルしか考えられていなかった．本研究では学習時間が短いリッジ回帰をベースにしたテンソル回帰モデル [3] と，二値分類問題を考慮するためのロジスティック回帰をベースにしたテンソル分類モデル [11] をこのフレームワークに導入する．これらのテンソル予測モデルは，画像・動画分析の分野で提案された手法である．回帰式は式 (12)，(14) であり，リッジ回帰ベースの手法では損失関数式 (16)，ロジスティック回帰ベースの手法では損失関数式 (17) を用いて，最適な $\mathcal{W}_s, b_s$ を求める．

$$L = \frac{1}{2}\sum_{t=1}^{N}\left(y_{s,t} - \tilde{\mathcal{X}}_{s,t} \times_1 \boldsymbol{w}_s^{(1)} \times_2 \boldsymbol{w}_s^{(2)} \times_3 \boldsymbol{w}_s^{(3)} - b_s\right)^2$$

$$+\frac{\lambda}{2}\left(||\boldsymbol{w}_s^{(1)}||^2 + ||\boldsymbol{w}_s^{(2)}||^2 + ||\boldsymbol{w}_s^{(3)}||^2 + ||b_s||^2\right) \tag{16}$$

$$L = \sum_{t=1}^{N}\log\left[1 + \exp\left\{-y_{s,t}^{*}\left(\tilde{\mathcal{X}}_{s,t} \times_1 \boldsymbol{w}_s^{(1)} \times_2 \boldsymbol{w}_s^{(2)} \times_3 \boldsymbol{w}_s^{(3)} + b_s\right)\right\}\right]$$

$$+\frac{\lambda}{2}\left(||\boldsymbol{w}_s^{(1)}||^2 + ||\boldsymbol{w}_s^{(2)}||^2 + ||\boldsymbol{w}_s^{(3)}||^2 + ||b_s||^2\right) \tag{17}$$

これらの損失関数は変数が多いため，変数 $\boldsymbol{w}_s^{(1)}, \boldsymbol{w}_s^{(2)}, \boldsymbol{w}_s^{(3)}$ のうち２つを固定し通常のリッジ回帰・ロジスティック回帰の問題に帰着させ，交互に最適な $\boldsymbol{w}_s, b_s$ を求めていく．また今回２つの予測モデルをフレームワークに導入するにあたって，終了条件の改良を行った．具体的には従来の終了条件は「$\mathcal{W}_s$ の収束または，一定回数以上の反復」であったが,「損失関数の値の収束または，一定回数以上の反復」という終了条件にすることで，過学習が起きづらい学習モデルにすることができた．

4 数値実験

4.1 使用データ

今回使用するデータは企業データ，ニュースデータ，センチメントデータの３種類である．

「企業データ」：米国のニューヨーク証券取引所に上場している 10 業界 82 社のデータ．取引日・企業ごとに「終値」,「出来高」が記録されており，この二つとテクニカル指標指標の「ストキャスティクス%K」，”RSI”の４種類の特徴量を用いる．またインデックス指標として S&P500 の「終値」を用いる．

「ニュースデータ」：トムソン・ロイター社が提供しているニュースデータを用いる.「記事 ID」,「タイトル」,「記事本文」,「公開日時」,「更新日時」,「企業 ID」などが１セットとなったデータである．これらのデータから Zhang ら [12] の手法を参考に，タイトルから動詞を抽出し，それを学習済みの word2vec[9] と主成分分析を用いて 20 次元の分散表現を得る．

「センチメントデータ」：トムソン・ロイターマーケットサイク指数 (TRMI) を用いる．トムソン・ロイター社が世界中のニュースや公開情報，ソーシャルメディアの掲示板，ツイッターから情報集め，各企業や各国通貨，資源などの資産クラスごとに様々な種類の TRMI 指数を算出している．今回はニュースデータと区別するために，TRMI の中でもニュース以外のインターネット掲示板やツイッターをソースとして計算された TRMI 指数を用いる．１日に一度 (ニューヨーク時間の 15:30)，各企業の過去 24 時間に対してのセンチメントの値が取得される．今回は”buzz”と TRMI 指数である ”sentiment”, ”emotionVsFact”の３つの指標を用いる．”buzz”はその銘柄のが「どれくらい話題になっているかの情報量を図る」指標であり [2]，”sentiment”は「ポジティブセンチメントとネガティブセンチメントの差」，”emotionVsFact”は「すべての感情的なセンチメントとすべての事実や時事問題の差」を図る指標である [13].

4.2 実験概要

本研究はニューヨーク証券取引所に上場している 10 業界 82 社を対象に実験を行う．ニューヨーク証券取引所は現地時間の 9:30-16:00 が取引時間であり，また実験に用いる TRMI データが NY 時間の 15:30 に過去 24 時間のセンチメントの値を算出しているため,「各企業ごとに前日の 15:30 から当日の 15:30 までのデータを用いて，当日の終値または残差リターンの正負を予測する」というタスクを用いて手法の比較を行う．二値分類の予測対象に残差リターンの正負を用いる理由は，各銘柄は市場全体の流れで価格が変動することがあり，市場全体の影響を取り除いた残差リターンを考えることで，今回実験に用いる３つのデータからその銘柄の予測しやすくなると考えたからである．

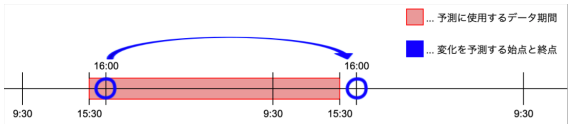

図 3: 予測に用いるデータ期間と対象

4.3 結果

表 1: サポートベクトル回帰ベースの予測

決定係数(R^2)				
μ	γ	類似度		
		(1)	(2)	(3)
1	1	-0.48123	0.92536	0.91818
1	0.1	0.89571	0.92155	0.92084
1	0.01	0.91906	0.92127	0.92116
1	0	0.92121		
0.1	1	**0.97357**	0.97344	0.97344
0.1	0.1	0.97347	0.97342	0.97342
0.1	0.01	0.97342	0.97342	0.97342
0.1	0	0.97342		
0.01	1	**0.97357**	0.97336	0.97336
0.01	0.1	0.97340	0.97336	0.97336
0.01	0.01	0.97337	0.97336	0.97337
0.01	0	0.97336		
0	1	**0.97357**	0.97337	0.97336
0	0.1	0.97339	0.97335	0.97335
0	0.01	0.97336	0.97335	0.97334
テンソル変換なし		0.97336		
(ベクトルで予測)		0.97323		

表 1，表 2，表 3 において，テンソル変換を行って予測精度が向上したものを赤く塗っている．それぞれの表の μ, γ,(1),(2),(3) はテンソル変換のハイパーパラメータを表しており，特に (1)(2)(3) は類似度 $z_{s,c}$ の種類を表している．(1) は銘柄同士の残差リターン正負の一致度，(2) は残差リターンの相関係

表 2: リッジ回帰ベースの予測

決定係数(R^2)				
μ	γ	類似度		
		(1)	(2)	(3)
1	1	0.97241	0.97248	0.97247
1	0.1	0.97242	0.97228	0.97237
1	0.01	0.97243	0.97235	0.97246
1	0	0.97252		
0.1	1	0.97240	0.97239	0.97224
0.1	0.1	0.97238	0.97233	0.97248
0.1	0.01	0.97241	0.97243	0.97230
0.1	0	0.97232		
0.01	1	0.97246	0.97237	0.97239
0.01	0.1	0.97245	0.97228	0.97242
0.01	0.01	0.97244	0.97235	0.97235
0.01	0	0.97241		
0	1	**0.97251**	0.97248	0.97237
0	0.1	0.97233	0.97232	0.97230
0	0.01	0.97238	0.97238	0.97228
テンソル変換なし		0.97244		

(ベクトルで予測)	0.97186

表 3: ロジスティック回帰ベースの予測

正解率(AC)				
μ	γ	類似度		
		(1)	(2)	(3)
1	1	0.51293	0.52166	0.52189
1	0.1	0.52165	0.52191	0.52202
1	0.01	0.52185	0.52201	0.52202
1	0	0.52202		
0.1	1	0.52511	0.52303	0.52294
0.1	0.1	0.52376	0.52263	0.52280
0.1	0.01	0.52269	0.52279	0.52280
0.1	0	0.52278		
0.01	1	**0.52531**	0.52326	0.52353
0.01	0.1	0.52264	0.52420	0.52415
0.01	0.01	0.52396	0.52413	0.52409
0.01	0	0.52409		
0	1	0.52530	0.52345	0.52374
0	0.1	0.52225	0.52418	0.52424
0	0.01	0.52426	0.52415	0.52412
テンソル変換なし		0.52409		

(ベクトルで予測)	0.52210

表 4: 既存のサポートベクトル回帰ベースのテンソル回帰モデルと提案したリッジ回帰ベースのテンソル回帰モデルの比較

トレーニングデータ数	企業数	決定係数(R^2)		学習時間[s]	
		SVRベース	リッジ回帰ベース	SVRベース	リッジ回帰ベース
300~400	5	0.97724	0.97555	2.366	0.164
400~500	11	0.97329	0.97213	3.657	0.237
500~600	15	0.97285	0.96963	5.023	0.281
600~700	11	0.97305	0.97272	6.032	0.307
700~800	8	0.98187	0.98196	7.751	0.305
800~900	10	0.97881	0.97858	9.395	0.307
900~1000	7	0.97232	0.97191	11.812	0.346
1000~1100	2	0.98415	0.98371	16.653	0.447
1100~1200	9	0.95960	0.96003	19.666	0.371
1200~1300	4	0.96821	0.96805	17.905	0.334
平均	82	**0.97336**	0.97244	8.711	**0.299**

数，(3) は銘柄同士が同じ記事に共に出た割合を銘柄間の類似度としてる．

これらの結果からまず，ただ単に特徴量を１つのベクトルに連結することや３次テンソルに格納したデータで予測を行うのではなく，テンソル変換を用いてテンソル再構築したデータで予測を行った時の方が精度よく学習できていることがわかる．これは提案した補助行列が時刻間・企業間の類似性を考慮したテンソル変換を行い，情報源間の共起関係の増強・ノイズ除去された新しいテンソルを作れたからだと考えられる．さらにハイパーパラメータごとの精度の違いを見ると，μ が小さく，γ が大きいほど精度よく予測できる傾向があることがわかる．このことから同一銘柄の時刻間の類似性を考慮するよりも，同一時刻の他企業の類似性を考慮したほうが精度が良いテンソル変換が行えるということがわかる．企業間の類似度については，終値予測タスク・残差リターンの正負予測タスクの両方において (1) の「残差リターンの正負の一致度」を用いたものが全体的に精度が良いことがわかる．

表 4 は 82 企業をトレーニングデータ数で分けて結果をまとめたものである．この結果から全体の株価予測精度は，既存手法のサポートベクトル回帰のテンソル回帰モデルの方が優れていることがわかる．しかし学習時間を見ると，トレーニングデータ数が増えていくごとにサポートベクトル回帰ベースの予測モデルの学習時間が増えていることがわかる．これはサポートベクトル回帰の学習の遅さによるものであり，実際にこれらのモデルを用いる際は予測精度と学習時間のトレードオフでどちらのモデルを用いるかを検討すべきである．

5 おわりに

本研究の目的は，株式価格の変動予測のためのマルチモーダル学習において，複数の情報源の情報を十分に考慮しより精度の高い予測を行うということであった．そこで本研究では「改良したテンソルを用いたマルチモーダル学習フレームワーク」を提案し，株価データ，ニュースデータ，センチメントを用いた株価予測 (回帰問題) と残差リターンの正負予測

(二値分類問題) の実験を行った．新たに導入したリッジ回帰ベースのテンソル回帰モデルは，既存のサポートベクトル回帰ベースのテンソル回帰モデルよりも予測精度は劣るものの学習時間を大幅に削減した予測モデルとなった．また株価予測・残差リターンの正負予測の両方において，それぞれの情報源から作成した特徴量を１つのベクトルとして連結して予測を行うのではなく，提案したテンソルを用いたマルチモーダル学習のフレームワークを用いることによって，より精度良く予測できることが示された．

今後の課題として次の２つが考えられる．１つ目の課題はさらなる予測精度の向上である．今回の手法・実験ではわずかな精度の向上しか確認できなかった．これは従来より株式市場の予測は難しいタスクであるとともに，今回の問題設定が１日間隔の予測を行っていたからだと考える．一日の間に複数の出来事があり，相場のモメンタムが激しく変わる場合などは予測が難しくなる．そこで実験の時間間隔をより短くすることで精度が向上の可能性があると考える．２つ目の課題はテンソルの各軸の次数を削減できるようなテンソル変換手法の考案である．従来の手法・今回提案した手法では高次元テンソルを扱うため，学習・予測の際に利用するリソースが必要になる．従来の「テンソルを用いたマルチモーダル学習フレームワーク」ではテンソル変換の際に各軸の次元削減が可能であったが，今回の提案手法では $V_{s,k}$ が正方行列という条件を追加したため，次元削減の要素を失ってしまった．よって次元削減と予測精度向上の両立できるテンソル変換モデルを考案することが望ましい．

謝辞

本研究はみずほ第一フィナンシャルテクノロジー株式会社にデータや実験環境の提供，研究の相談など多大なるご協力をいただきました．この場を借りて厚く御礼申し上げます．

参考文献

[1] Johan Bollen, Huina Mao and Xiaojun Zeng, “Twitter Mood Predicts the Stock Market,” *Journal of Computational Science*, 2(1), pp.1-8, 2011.

[2] Baoqing Gan, Vitali Alexeev, Ron Bird and Danny Yeung, “Sensitivity to Sentiment: News vs Social Media,” *International Review of Financial Analysis*, 67(101390), 2020.

[3] Weiwei Guo, Irene Kotsia and Ioannis Patras, “Tensor Learning for Regression,” *IEEE Transactions on Image Processing*, 21(2), pp.816-827, 2012.

[4] Jieyun Huang, Yunjia Zhang, Jialai Zhang and Xi Zhang, “A Tensor-Based Sub-Mode Coordinate Algorithm for Stock Prediction,” *Proceedings of 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pp.716-721, 2018.

[5] Tamara G. Kolda and Brett W. Bader, “Tensor Decompositions and Applications,” *SIAM Review*, 51(3), pp.455-500, 2009.

[6] Qing Li, Yuanzhu Chen, Li L.Jiang, Ping Li and Hsinchun Chen, “A Tensor-Based Information Framework for Predicting the Stock Market,” *ACM Transactions on Information Systems*, 34(2), 2016.

[7] Qing Li, Tiejun Wang, Ping Li, Ling Liu, Qixu Gong and Yuanzhu Chen, “The Effect of News and Public Mood on Stock Movements,” *Information Sciences*, 278, pp.826-840, 2014.

[8] Yang Liu, Qingguo Zeng, Huanrui Yang and Adrian Carrio, “Stock Price Movement Prediction from Financial News with Deep Learning and Knowledge Graph Embedding,” *Proceedings of Knowledge Management and Acquisition for Intelligent Systems*, pp.102-113, 2018.

[9] Radim Řehůřek and Petr Sojka, “Software Framework for Topic Modelling with Large Corpora,” *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp.45-50, 2010.

[10] Hiroki Suwa, Yuki Ogawa, Eiichi Umehara, Kento Kakigi, Keiichi Yasumoto, Tatsuto Yamashita and Kota Tsubouchi, “Develop Method to Predict the Increase in the Nikkei VI Index,” *Proceedings of 2017 IEEE International Conference on Big Data (Big Data)*, pp.3133-3138, 2017.

[11] Xu Tan, Yin Zhang, Siliang Tang, Jian Shao, Fei Wu and Yueting Zhuang, “Logistic Tensor Regression for Classification,” *Proceedings of the Third Sino-Foreign-Interchange Conference on Intelligent Science and Intelligent Data Engineering*, pp.573-581, 2012.

[12] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang and Philip Yu, “Improving Stock Market Prediction via Heterogeneous Information Fusion,” *Knowledge-Based Systems*, pp.236–247, 2017.

[13] リチャード・L・ピーターソン，[監修] 長尾慎太郎，[訳] 井田京子,『市場心理とトレード -ビッグデータによるセンチメント分析』，パンローリング株式会社，東京都，2017.