研究内容と論文紹介

2020年 10月 8日（木）

中田研総合ゼミ　笹尾知広

目次

1. 論文紹介

2. 共同研究の内容について
   a. 研究概要
   b. 提案手法
   c. 実験

論文情報（メイン）

- タイトル
  A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News

- 雑誌
  IEEE Transactions on Knowledge and Data Engineering
  Jan 2020

- 著者
  Qing Li, Jinghua Tan, Jun Wang, HsinChun Chen

論文情報(サブ)

- タイトル
  A Tensor-based eLSTM Model to Predict Stock Price Using Financial News

- 学会
  Hawaii International Conference on System Sciences 2019

- 著者
  Jinghua Tan, Jun Wang, Denisa Rinprasertmeechai , Rong Xing, Qing Li

論文概要

- **市場データ、ニュース**の2つを組み合わせて、**2階テンソル**で株価予測を行う
  ※ 前回の発表では3階テンソル
- 各銘柄の予測に**相関性が高い銘柄**の共通性を利用
- **不規則なニュースの情報**を処理できる、テンソルベースのLSTMモデル、**Multimodal Event-driven LSTM**を提案
- 実験において、提案手法がほかの手法よりも優れた予測ができていたことを確認

論文背景

- 根底にある考え：**情報が株式の動きを形成する**

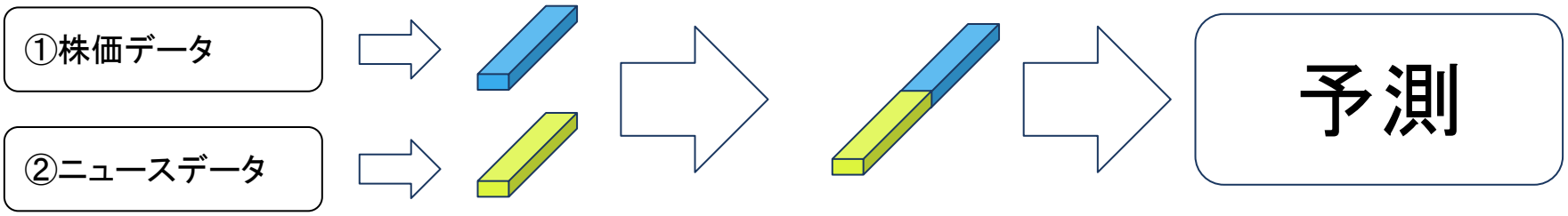
- 金融情報：**定量的データ**（企業規模、キャッシュフローなど）
  定性的データ（ニュース、センチメントなど）

- **異なる情報源**が株価に影響を与えている
  （[1]Francis, Dou-glas Hanna, and Philbrick 1997）

[1]https://ideas.repec.org/a/eee/jaecon/v24y1997i3p363-394.html

論文背景

- 従来のよくある手法：
  複数の情報源から特徴量を作成し、一つの連結したベクトルを作成する



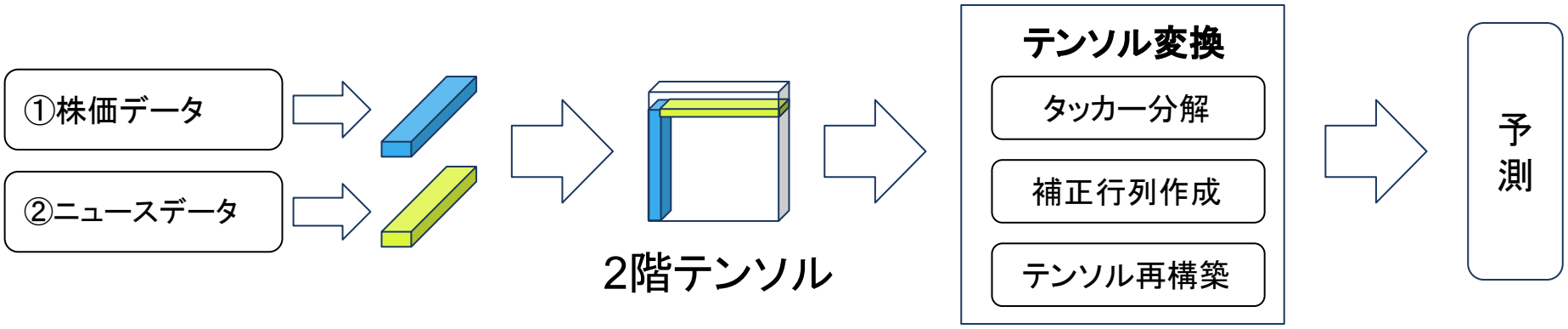
- 連結ベクトルの問題点：
  ①次元の呪いを引き起こす（[2]Bellman and Dreyfus 1962）
  ②異なる情報源間の共起関係は弱められる

- 複数の情報源を扱うときの問題：
  情報間でデータ取得の頻度が異なる

[2]https://www.rand.org/content/dam/rand/pubs/reports/2006/R352.pdf

論文背景

- テンソルベースの予測モデル(※前回発表 **[3]Q.Li et al., 2014,** [4]J.Huang et al., 2018)



- 課題:
  - ニュースない日のデータを無視している
  - TeSIAという(SVRを拡張した)予測モデル
- 提案:
  - **類似銘柄を用いた情報の補完**
  - **テンソルベースのLSTMを用いて予測**

[3]https://www.semanticscholar.org/paper/Tensor-Based-Learning-for-Predicting-Stock-Li-Jiang/dd0801bc0ea9294b3ec7bdca9b59a10541a15f21
[4]https://arxiv.org/abs/1805.07979

手法概要

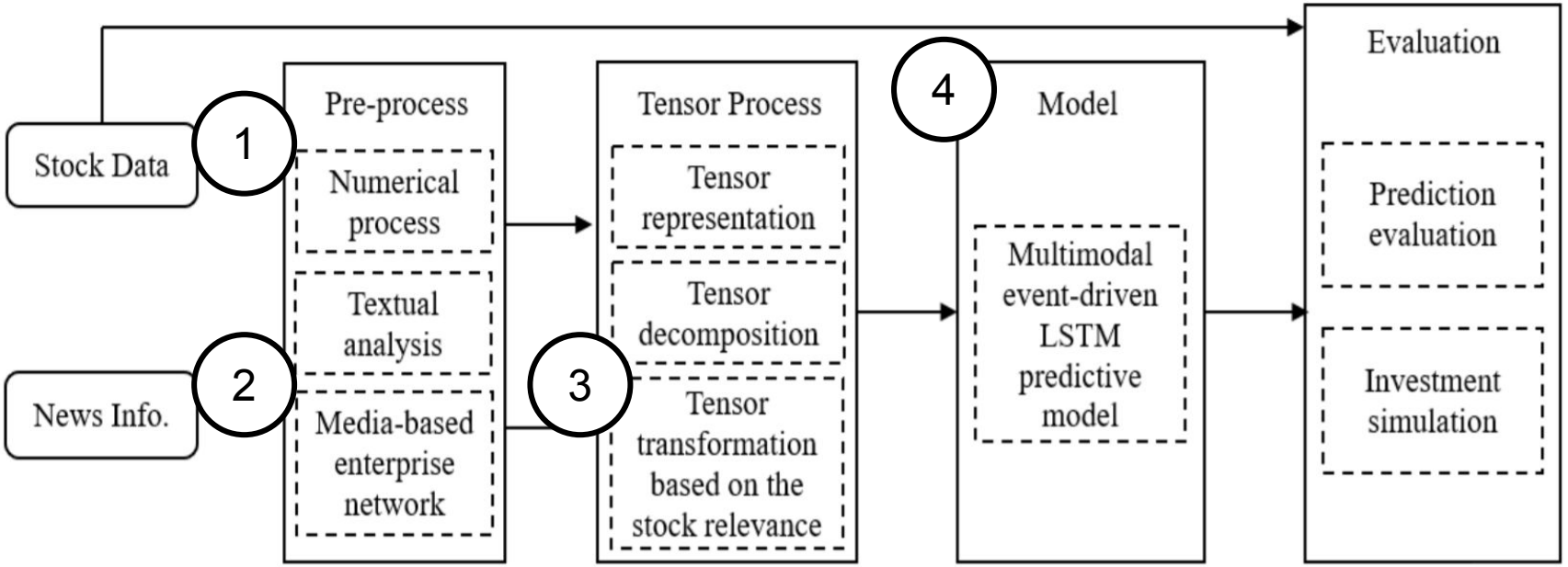
① 2種の情報源から特徴量を作成しテンソルとして表現

② 銘柄間の関係行列を取得

③ 関係銘柄を用いてテンソル変換を行う

④ 変換されたテンソルを用いて、

Multimodal Event-drivenLSTMで予測



提案手法 - ①2種の情報源をテンソルとして表現

- 2種の情報源（2015/01/01 - 2015/12/31）
  - **企業情報**
    中国の証券市場(CSI)の91銘柄の株価、売上高、PER、PBR、その他指標
  - **メディア情報（ニュース）**
    91銘柄に関連付けられているニュース45,021件
    ポジ、ネガ、その差

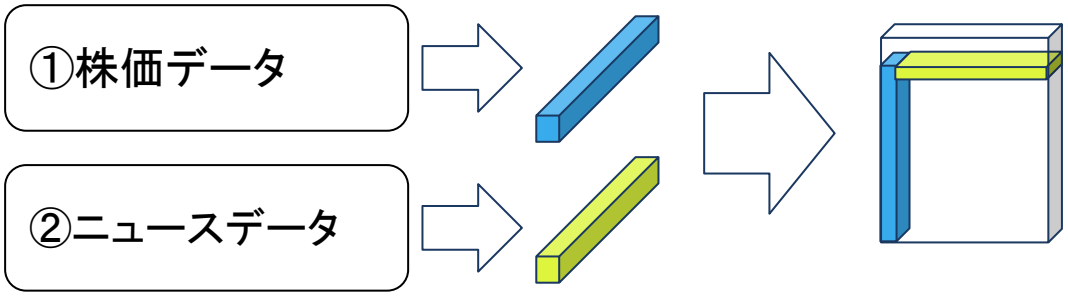
$$P_t^+ = \frac{N_t^+}{N_t^+ + N_t^-}, P_t^- = \frac{N_t^-}{N_t^+ + N_t^-}, D_t = \frac{N_t^+ - N_t^-}{N_t^+ + N_t^-}$$

Nt+(Nt-) : t日のニュースに含まれるポジティブ(ネガティブ)ワード数

提案手法 - ①2種の情報源をテンソルとして表現

- 三階テンソルを $X_t \in R^{I_1 \times I_2 \times 2}$ で表す。 I_1, I_2 はそれぞれ企業、メディアの次元数

- テンソルの要素 a_{i_1,i_2} は次のように構成される
  - $a_{i_1,1},\ 1 < i_1 \leq I_1$ ： 企業特徴量
  - $a_{2,i_2},\ 1 < i_2 \leq I_2$ ： メディア特徴量
  - その他の成分は0



提案手法 - ② 銘柄間の関係行列を取得

背景：**企業間の取引や業界の関係で、他社に関するニュースによって株価が変動する**

ゴール：**自社に関係する企業を見つける**

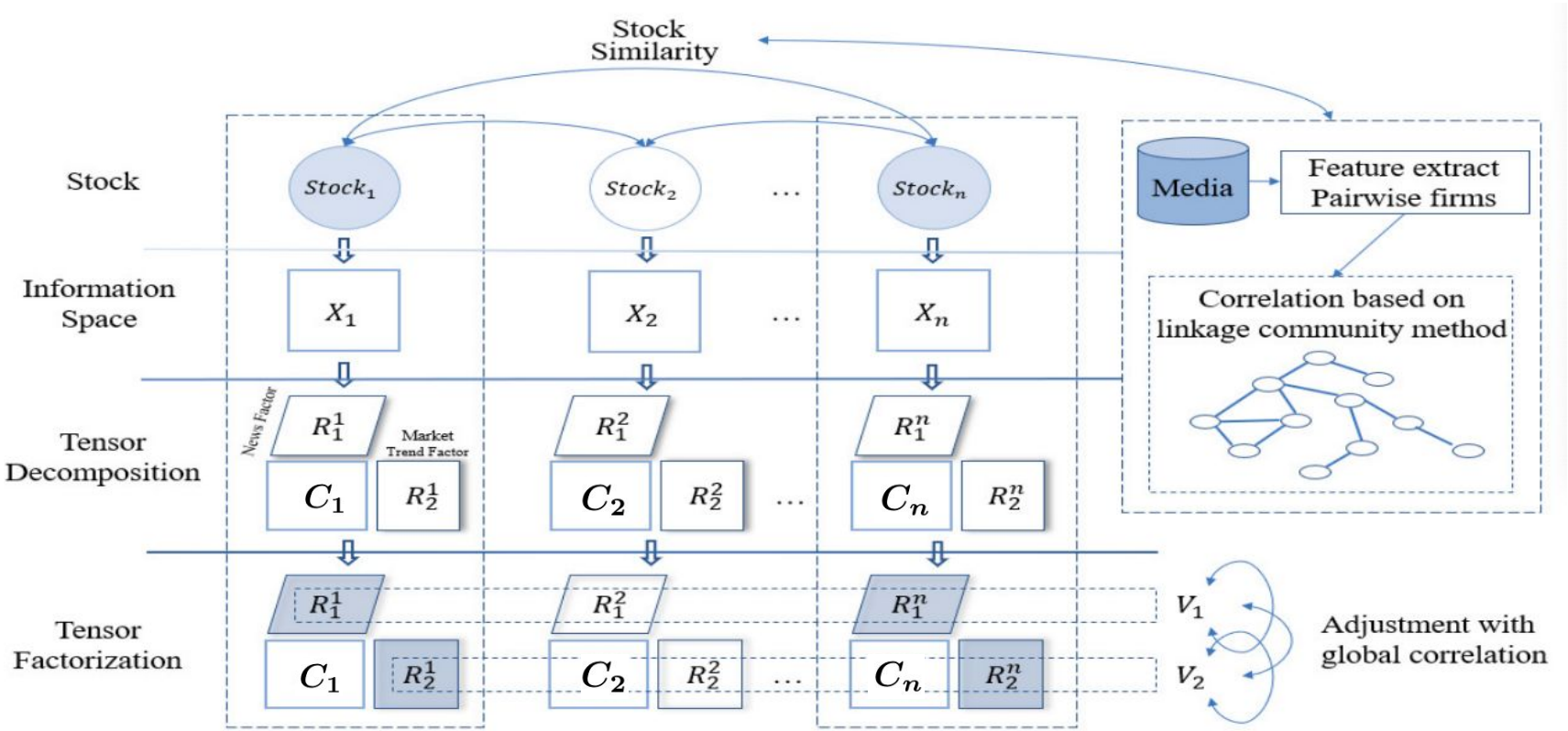
方法：**ニュースに同時に記載されている銘柄をカウント**

$$\bar{s}_{i,j} = \frac{\text{銘柄}i,j\text{が同時に載ったニュース数}}{\text{銘柄}i,j\text{のニュースの合計}}$$

$$s_{i,j} = \begin{cases} 1 & \text{if } i \leq j \text{ and } \bar{s}_{i,j} \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

提案手法 - ③ 関係銘柄を用いてテンソル変換を行う

- 各銘柄 i において、t 日のテンソル v_i^t を作成
- 関連する銘柄の情報を補完するため、テンソル変換を行う



提案手法 - ③ 関係銘柄を用いてテンソル変換を行う

テンソルの変換（次元削減）の流れ

①テンソルをタッカー分解

$$X_i = C_i \times_1 R_1^i \times_2 R_2^i$$

②銘柄の関係性が高いテンソル同士が同じようなテンソルになるように、補助行列 V_1, V_2 を導入

③テンソルを再構築

$$\tilde{X}_i = C_i \times_1 (V_1^T R_1^i) \times_2 (V_2^T R_2^i)$$

提案手法 - ③ 関係銘柄を用いてテンソル変換を行う

②銘柄の関係性が高いテンソル同士が同じようなテンソルになるように、補助行列 V_1, V_2 を導入

$$\min_{V_{k,k=1,2}} L(V_k) = \frac{\lambda}{2}\sum_{i=1}^{N} ||X_i - C_i \times_1 (V_1^T R_1^i) \times_2 (V_2^T R_2^i)||^2$$
$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} ||V_1^T R_1^i - V_1^T R_1^j||^2 s_{i,j}$$
$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=i}^{N} ||V_2^T R_2^i - V_2^T R_2^j||^2 s_{i,j}$$

提案手法 - ③ 関係銘柄を用いてテンソル変換を行う

②銘柄の関係性が高いテンソル同士が同じようなテンソルになるように、補助行列 V_1, V_2 を導入

$$\nabla_{v_1} L = \lambda \sum_{i=1}^{N} (C \times_1 (V_1^T R_1^i) \times_2 (V_2^T R_2^i) R_1^i) + (D_{R_1} - S_{R_1}) V_1$$

$$\nabla_{v_2} L = \lambda \sum_{i=1}^{N} (C \times_1 (V_1^T R_1^i) \times_2 (V_2^T R_2^i) R_2^i) + (D_{R_2} - S_{R_2}) V_2.$$

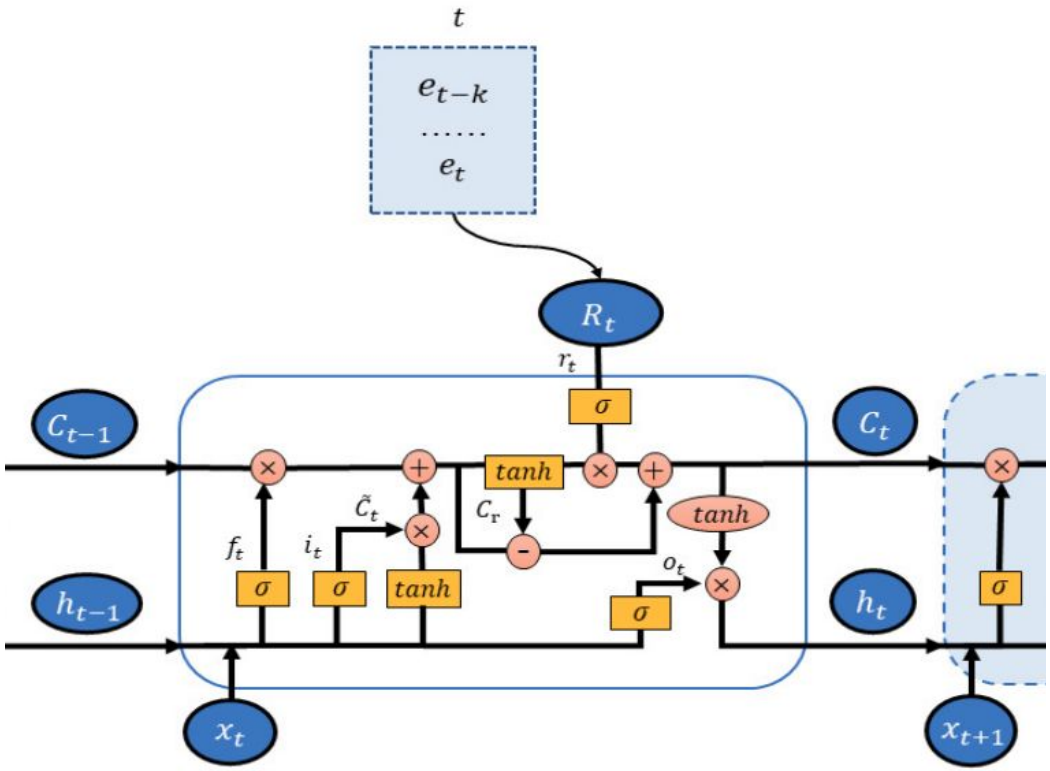
$D_{R_k} = \sum_{i=1}^{N} (R_k^i R_k^{iT}) d_{i,j}$, $S_{R_k} = \sum_{i=1}^{N} \sum_{j=i}^{N} (R_k^i R_k^{iT}) s_{i,j}$

$d_{i,i} = \sum_{m=1}^{N} s_{m,i}$

提案手法 - ④ Multimodal Event-drivenLSTMで予測

Multimodal Event-drivenLSTM

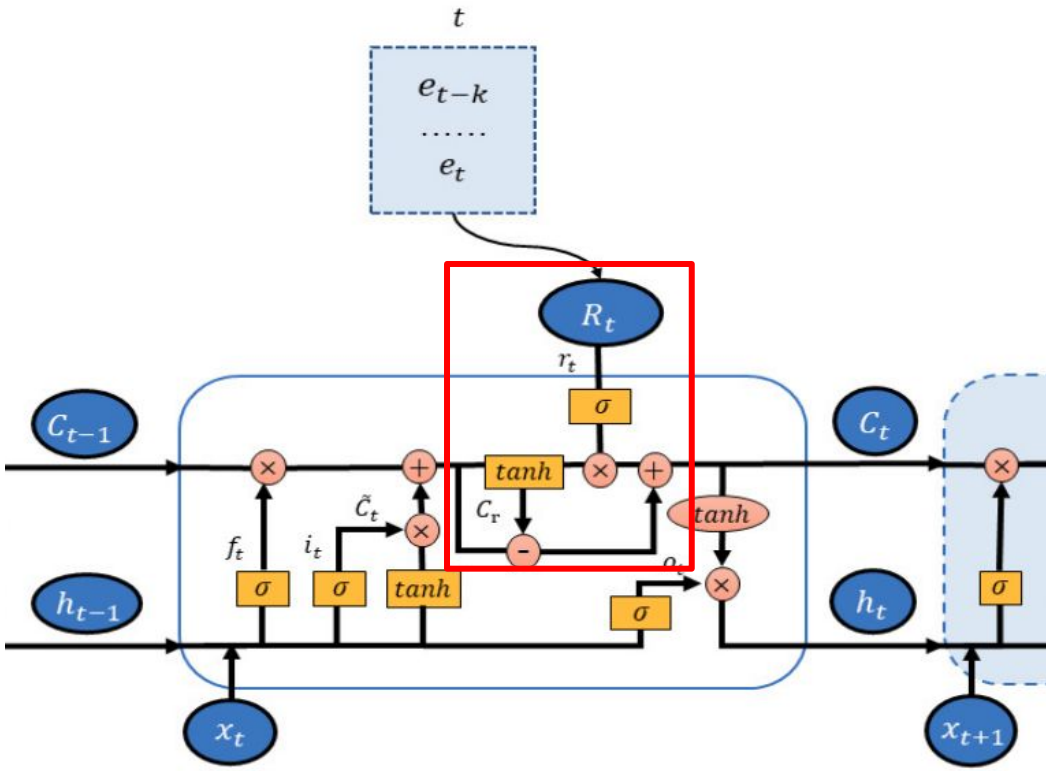
- 畳み込みを含んだLSTMモデル
- 不規則なニュースの情報に対処できる



提案手法 - ④ Multimodal Event-drivenLSTMで予測

Multimodal Event-drivenLSTM

- 畳み込みを含んだLSTMモデル
- 不規則なニュースの情報に対処できる



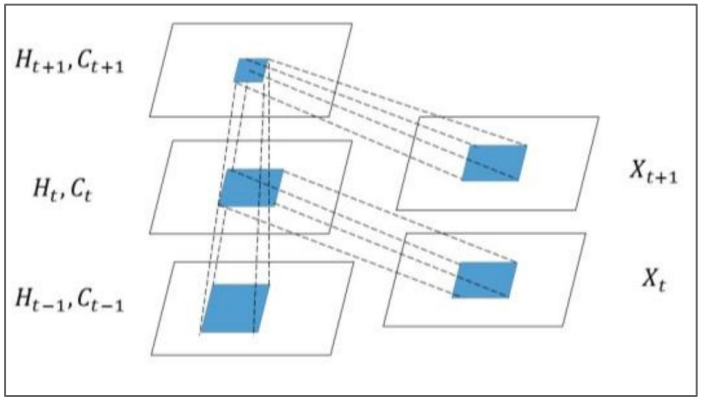
提案手法 - ④ Multimodal Event-drivenLSTMで予測

畳み込みを含んだLSTMモデル

Convolutional LSTM ([5] X. Shi, et al., 2015)

入力：行列 (n×m) 、出力：行列 (n×m)

CNNと同じように畳み込み計算を行う



[5] https://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf

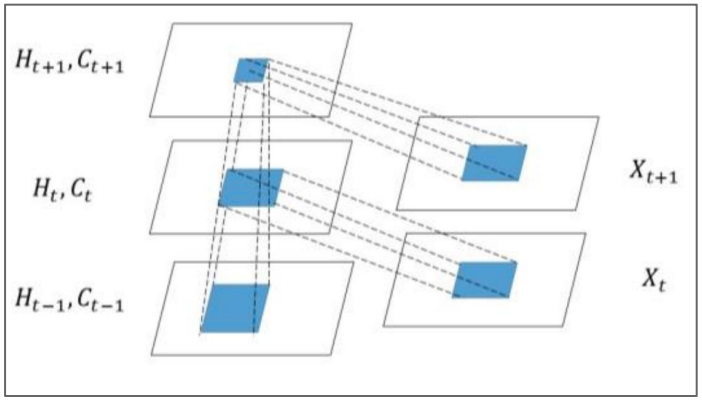
提案手法 - ④ Multimodal Event-drivenLSTMで予測

畳み込みを含んだLSTMモデル

Convolutional LSTM ([5] X. Shi, et al., 2015)

入力：行列 (n×m) 、出力：行列 (n×m)

CNNと同じように畳み込み計算を行う

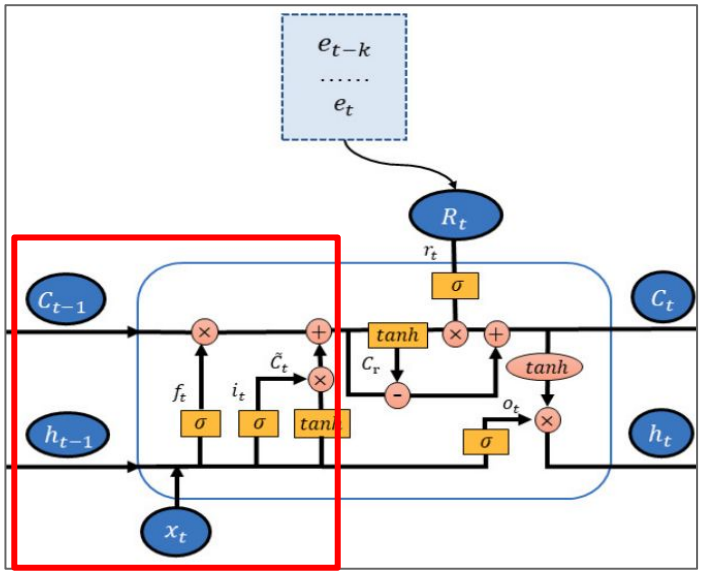


$$\tilde{C}_t = \tanh(W_c * X_t + U_c * H_{t-1} + V_c * R_t + B_c)$$
$$f_t = \sigma(W_f * X_t + U_f * H_{t-1} + V_f * R_t + B_f)$$
$$I_t = \sigma(W_i * X_t + U_i * H_{t-1} + V_i * R_t + B_i)$$
$$\hat{C}_t = f_t \circ C_{t-1} + I_t \circ \tilde{C}_t$$

$$R_t = (e_{t-k}, \ldots, e_t)^T$$

e_t ： t日でのニュース記事の総数

＊は畳み込み計算、○はアマダール積計算



[5] https://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf

提案手法 - ④ Multimodal Event-drivenLSTMで予測

ニュースの量で、メモリセルの扱いを変化させる

$$r_t = \sigma(V_r * E_t + B_r)$$

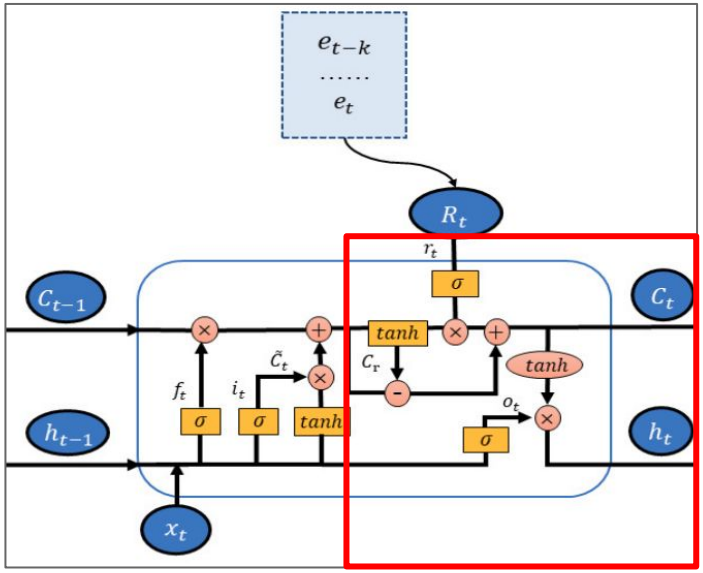
$$C_r = tanh(\hat{C}_t)$$

$$C_t = \hat{C}_t + (C_r \circ r_t - C_r)$$

Event driven

$$O_t = \sigma(W_o * X_t + U_o * H_{t-1} + V_o * E_t + B_o)$$

$$H_t = O_t \circ tanh(C_t)$$



$R_t = (e_{t-k}, \ldots, e_t)^T$

e_t ： t日でのニュース記事の総数

＊は畳み込み計算、○はアマダール積計算

[5] https://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf

実験

- 期間
  トレーニング：2015年1-9月
  テスト：2015年10-12月

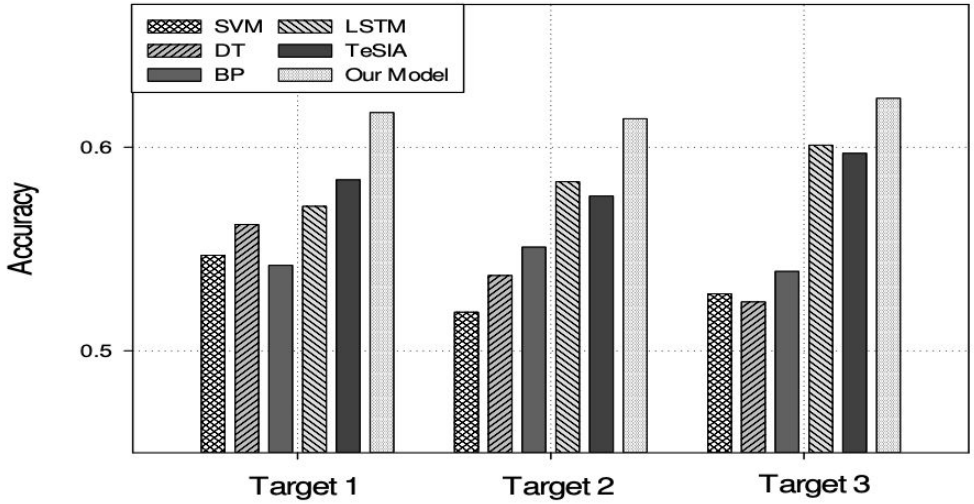
- データ
  - 中国の証券取引所での91銘柄データ
  - 91銘柄に関する45,021の記事

- ターゲット

Target 1	Open (t+6) - Open (t)
Target 2	Close (t+6) - Close (t)
Target 3	Close (t+6) - Open (t)

実験

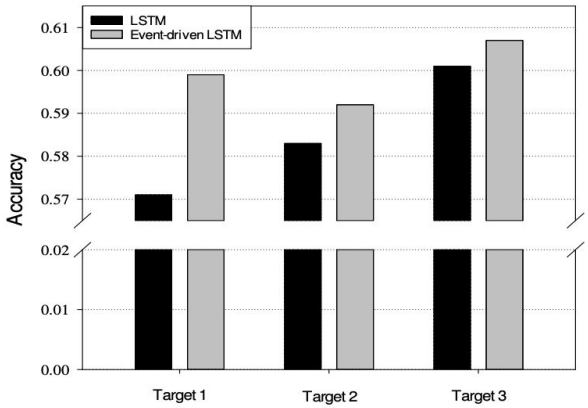
- 比較手法
  - サポートベクタルマシン (SVM)
  - 決定木(DT)
  - ニューラルネット(BT)
  - LSTM
  - テンソルベース回帰 (TeSIA)
  - 提案手法
- 結果



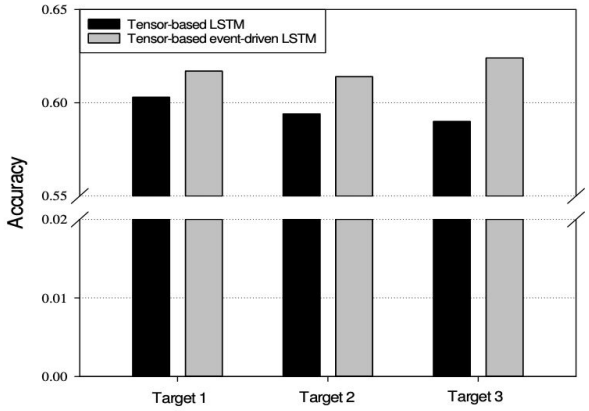
(a) Directional accuracy

実験

- **ベクトル** VS **テンソル**

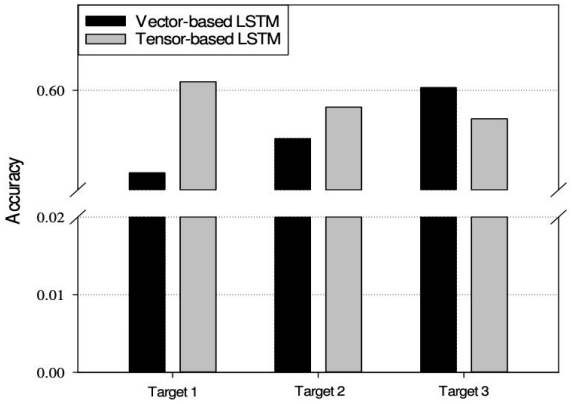


(a) Vector-based methods

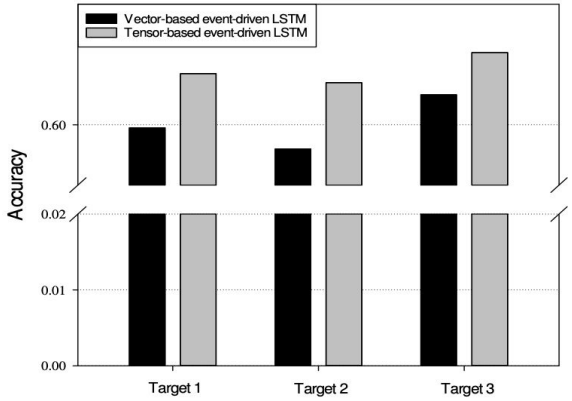


(b) Tensor-based methods

- **通常のLSTM** VS **Event-driven LSTM**



(a) LSTM models



(b) Event-driven models

共同研究の概要

共同研究先：

みずほフィナンシャル・テクノロジー（みずほFT）

概要：

①株価データ
②ニュースデータ
③市場のセンチメントデータ

を組み合わせた株価予測

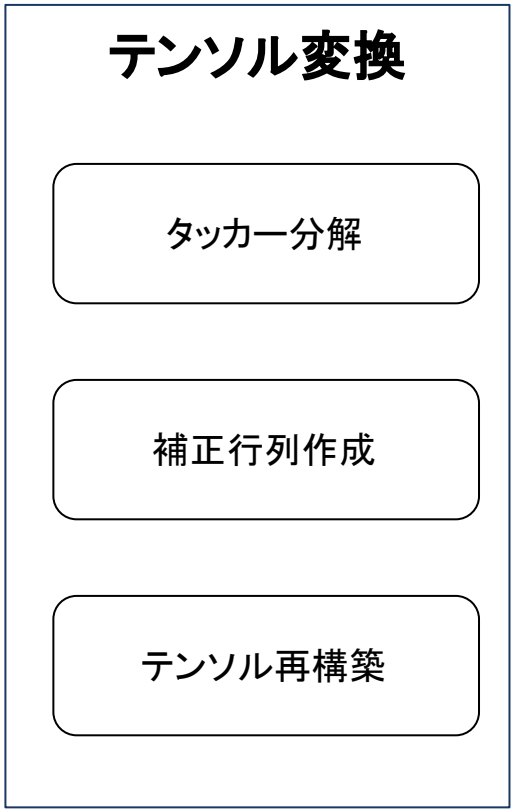
②：ロイター社が提供する英語のニュース記事
③：ロイター社が提供するTRMI(トムソンロイターマーケットサイク)というセンチメントデータ

共同研究について

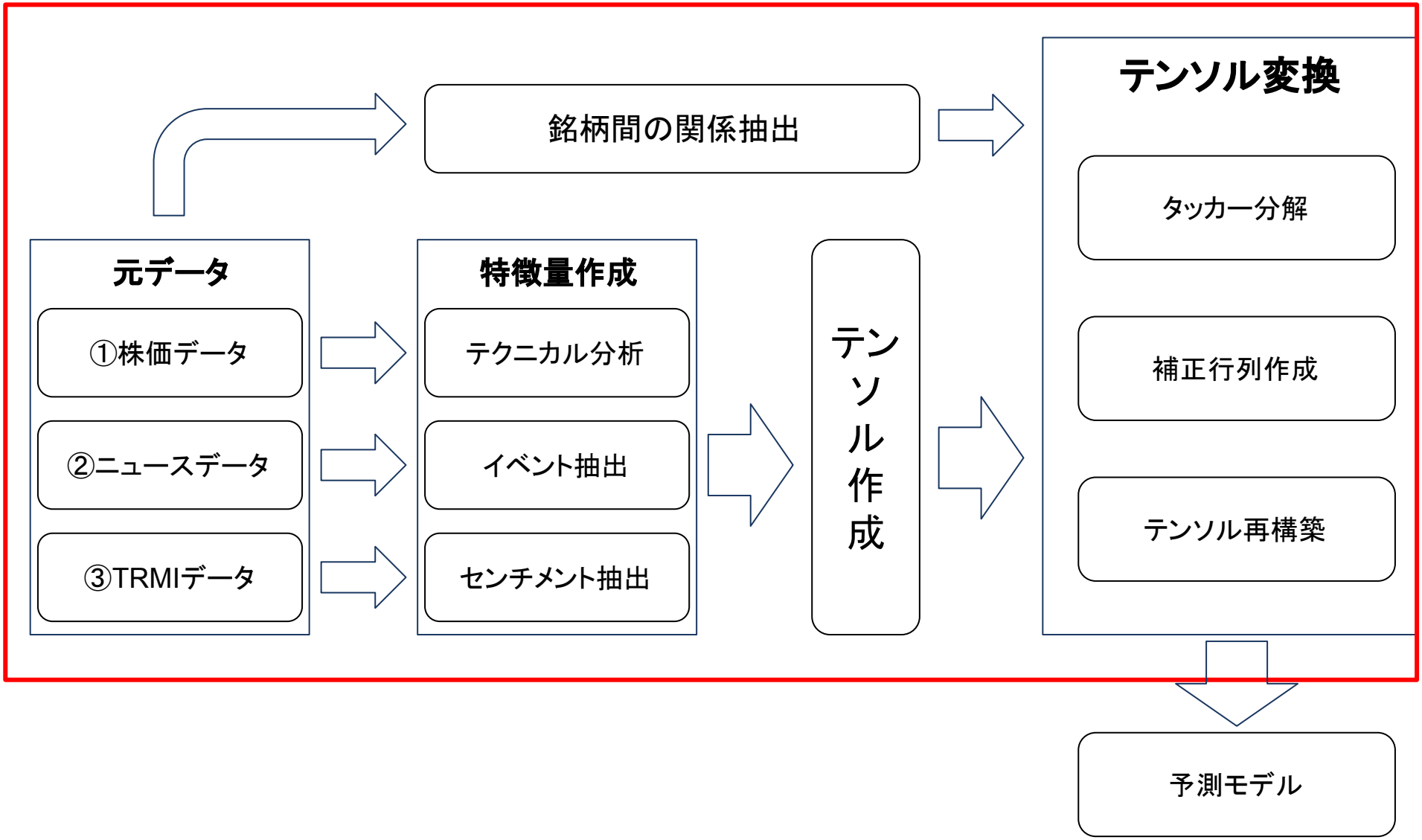
- 取り組んでいるテーマ
  テンソルを用いたマルチモーダル学習

- 具体的な研究内容
  - **テンソル変換を用いた前学習**
  - **テンソルベースの予測**

- モチベーション
  - **モーダル間の情報を捉える**
  - **モーダルごとに異なるサンプル頻度にうまく対処できる**



研究の概要（全体像）


銘柄間の関係抽出
テンソル変換
タッカー分解
元データ
①株価データ
②ニュースデータ
③TRMIデータ
特徴量作成
テクニカル分析
イベント抽出
センチメント抽出
テンソル作成
補正行列作成
テンソル再構築
予測モデル

先行研究

A. Q.Li, et al., “Tensor-Based Learning for Predicting Stock Movements”, 2014.
- テンソルベースのマルチモーダル学習の**最初の論文**
- **同一銘柄内で情報を補完**するようなテンソル変換
- テンソルを入力とする予測モデル（SVRベース）

B. J.Huang, et al., “A Tensor-Based Sub-Mode Coordinate Algorithm for Stock Prediction”, 2018
- **類似している銘柄間で情報を補完**するようなテンソル変換
- テンソルを入力とするLSTM

先行研究の問題点（テンソル変換の部分）

A, Bの両論文においても、テンソル変換の補助行列　　作成の部分が不十分

- Aの論文のV_kを求める最適化問題

$$\min_{V_k} J(V_k) = \sum_{i=1}^{T}\sum_{j=i}^{T} ||V_k^T U_k^i - V_k^T U_k^j||^2 w_{i,j}$$

$$s.t. \sum_{i=1}^{T} ||V_k^T U_k^i||^2 d_{i,i} = 1$$

$$w_{i,j} = \begin{cases} 1 & (i \leq j \text{ かつ } |y_i - y_j|/y_j \leq E_1) \\ 0 & (\text{その他}) \end{cases}$$

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

$\mathcal{X}_{s,t} \in R^{I_1 \times I_2 \times I_3}$: 銘柄s，時刻tのテンソル

$\mathcal{C}_{s,t} \in R^{D_1 \times D_2 \times D_3}$: 銘柄s，時刻tのテンソルをタッカー分解したときのコアテンソル

$U_k^{s,t} \in R^{I_k \times D_k}$: 銘柄s，時刻tのテンソルをタッカー分解したときのファクターk

$V_{s,k} \in R^{I_k \times J_k}$ $(J_k \leq I_k)$: 銘柄sのファクターkの補助行列

- Bの論文のV_kを求める最適化問題

$$\min_{V_k} J(V_k) = \sum_{s=1}^{S}\sum_{i=1}^{T}\sum_{j=i}^{T} ||V_k^T U_k^{s,i} - V_k^T U_k^{s,j}||^2 w_{s,i,j} + \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{m=s}^{S} ||V_k^T U_k^{m,t} - V_k^T U_k^{s,t}||^2 z_{t,s,m}$$

$$w_{s,i,j} = \begin{cases} 1 & (i \leq j \text{ かつ } |y_{s,i} - y_{s,j}|/y_{s,j} \leq E_1) \\ 0 & (\text{その他}) \end{cases}$$

$z_{t,s,m}$ ： 時刻 t における銘柄 s と m の相関

提案した手法

$$\min_{V_{s,k}} J(V_{s,k}) = \mu \sum_{i=1}^{T}\sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - V_{s,k}^T U_k^{s,j}||^2 w_{s,i,j} + \sum_{i=1}^{T}\sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - I_{s,k} U_k^{s,j}||^2 w_{s,i,j} + \gamma \sum_{t=1}^{T}\sum_{m=1}^{S} ||V_{s,k}^T U_k^{s,t} - V_{s,k}^T U_k^{m,t}||^2 z_{s,m}$$

$$w_{s,i,j} = \begin{cases} 1 & (|y_{s,i} - y_{s,j}|/y_{s,j} \leq E_1) \\ 0 & (\text{その他}) \end{cases} \quad (W_s \text{は対称行列})$$

$z_{s,m}$ ： 銘柄 s と m の相関 (Z は対称行列)

解析的に求まる

$$V_{s,k} = ((2\mu + 1)D_{U_k} - 2\mu W_{U_k} + \gamma(A_k + 2B_k + C_k))^{-1} D_{U_k}$$

実験

テンソル補完の前処理　VS 前処理なし

【問題設定】
(前日の15:30 , 今日の15:30] のデータを利用して、今日の16:00の終値を予測

・予測対象
NY証券取引所に上場されている3業界17銘柄(景気連動型消費財6社、生活必需品4社、素材7社)
※ニュースの数が多い企業を選択

・使用データと取得のタイミング

	項目	データ期間	データ取得のタイミング
株価データ	終値、出来高、%R、%K、RSI	2013 - 2019年	NY時間 16:00
ニュースデータ	ニュースタイトルの単語をword2vecで分散表現(300次元)。PCAで20次元に減らす	2006 - 2019年	グリニッジ標準時間 ニュースが出たタイミング
TRMI	sentiment、buzz、emotionVsFact	1998 - 2019年	NY時間 15:30

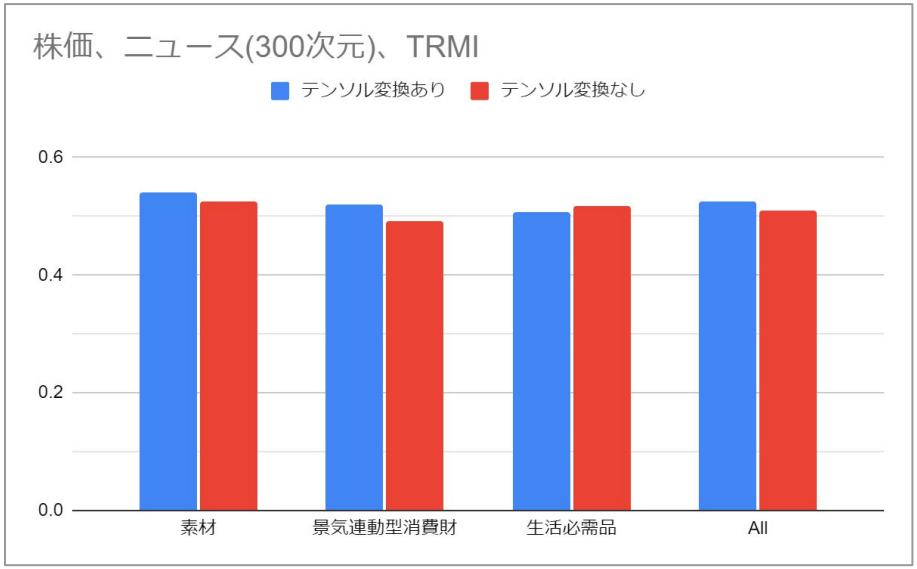
※NY証券取引所の取引時間はNY時間 9:30～16:00　サマータイム等を考慮して、すべてのデータをNY時間に合わせる

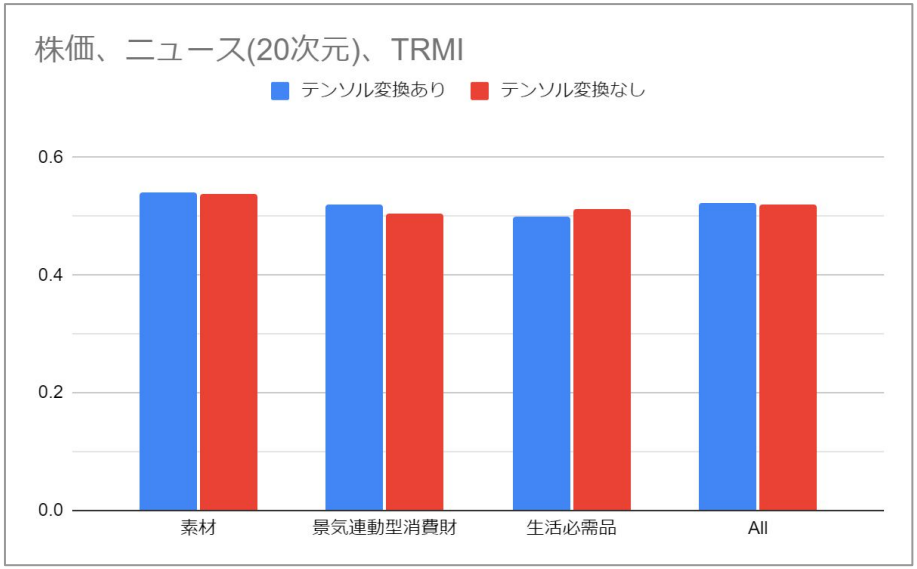
・予測期間
トレーニング：2013 - 2017年
テスト：2018 - 2019年

実験

- 予測手法
  ランダムフォレスト

- 結果





	テンソル変換あり	テンソル変換なし
素材	0.5408	0.5239
景気連動型消費財	0.5190	0.4908
生活必需品	0.5065	0.5174
All	**0.5251**	**0.5107**

	テンソル変換あり	テンソル変換なし
素材	0.5400	0.5387
景気連動型消費財	0.5190	0.5042
生活必需品	0.4996	0.5121
All	**0.5231**	**0.5203**

今後の予定

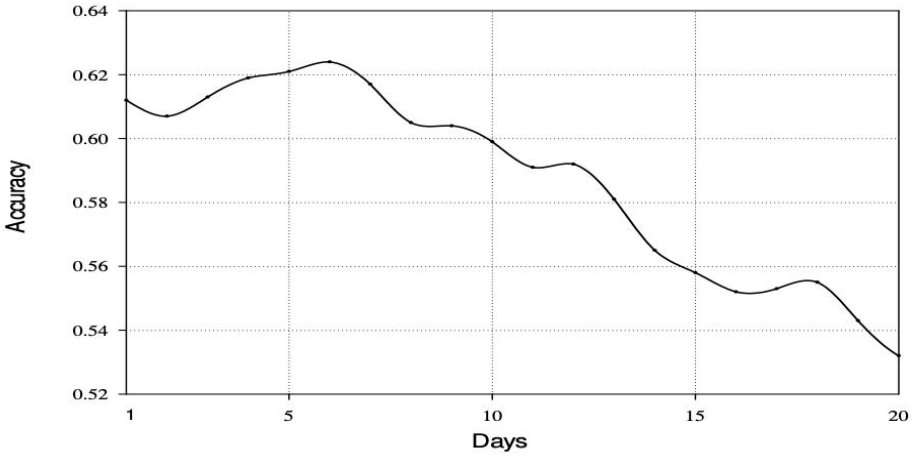
- テンソル変換の方法をさらに改良していく
  - 企業間の類似度行列
  - 最適化式の変更
  - パラメタ調整

- 予測方法の模索
  - LSTMベースで実験する...?

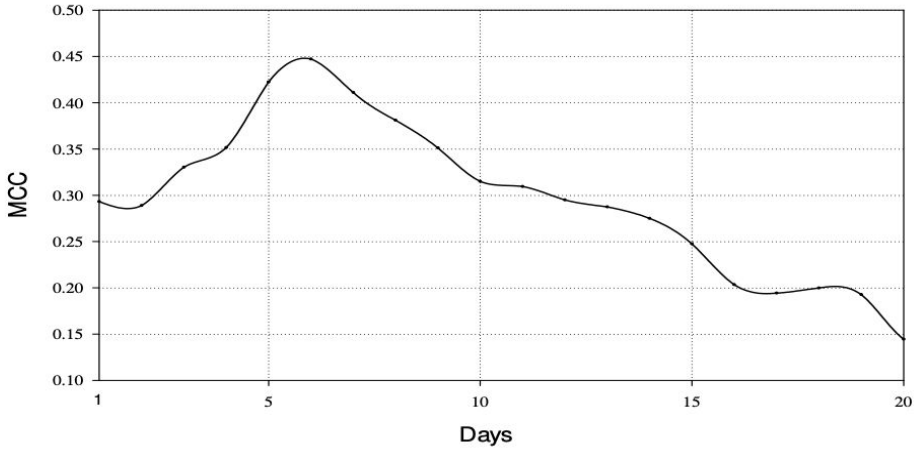
Appendix

予測ターゲット

６日目の予測精度が一番高い



(a) Directional accuracy



(b) Matthews correlation coefficient