みずほFT__共同研究

2020年 10月 14日
中田研究室　笹尾知広

【本日アジェンダ】
・研究の問題設定/ 方向性について
・テンソル変換のモデル説明
・実験

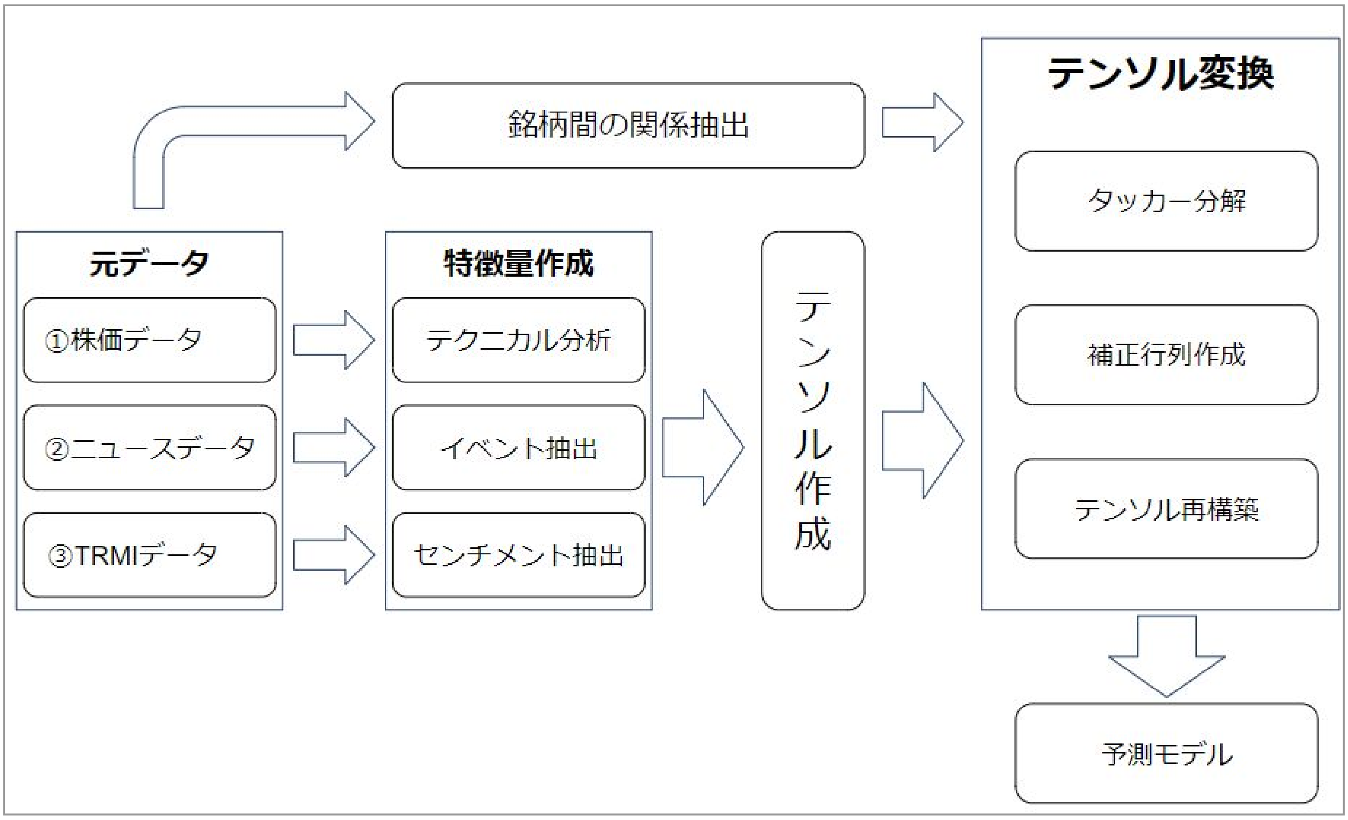
■研究の問題設定・方向性について

研究テーマ：テンソルを用いたマルチモーダル学習

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

全体像



■テンソル変換の説明

銘柄sの時刻tにおけるテンソル（株価データ，ニュースデータ，センチメントデータ）を$\mathcal{X}_{s,t}$とする.
$\mathcal{X}_{s,t}$の各要素を$x^{s,t}_{i,j,k}$とすると，（株価データ，ニュースデータ，センチメントデータ）は次のように入っている.

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

このテンソルは以下のようにタッカー分解することができる.

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

最終的に欲しいもの

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

$\mathcal{X}_{s,t} \in R^{I_1 \times I_2 \times I_3}$: 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

この V_k を得るための最適化式について

①銘柄内補完

$$\min_{V_{s,k}} J(V_{s,k}) = \frac{\mu}{sumW_s} \sum_{i=1}^{T} \sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - V_{s,k}^T U_k^{s,j}||^2 w_{s,i,j} + \frac{1}{sumW_s} \sum_{i=1}^{T} \sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - I_{s,k} U_k^{s,j}||^2 w_{s,i,j}$$

$$w_{s,i,j} = \begin{cases} 1 & (\text{時刻 } i,j \text{ がの結果が似ている場合}) \\ 0 & (\text{その他}) \end{cases} \quad (W_s \text{は対称行列})$$

$$sumW_s = \sum_{i=1}^{T} \sum_{j=1}^{T} w_{s,i,j}$$

$$V_{s,k} = sumW_s \left(\frac{2\mu + 1}{sumW_s} D_{U_k^s} - \frac{2\mu}{sumW_s} W_{U_k^s} \right)^{-1} D_{U_k^s}$$

$$d_{s,i} = \sum_{j=1}^{T} w_{s,i,j} = \sum_{m=1}^{T} w_{s,m,i}$$

$$D_{U_k^s} = \sum_{i=1}^{T} d_{s,i} U_k^{s,i} U_k^{s,i\,T}, \quad W_{U_k^s} = \sum_{i=1}^{T} \sum_{j=1}^{T} w_{i,j} U_k^{s,i} U_k^{s,j\,T}$$

②関連銘柄間の補完

$$\min_{V_{s,k}} J(V_{s,k}) = \frac{\gamma}{sumE_s} \sum_{t=1}^{T} \sum_{m=1}^{C} ||V_{s,k}^T U_k^{s,t} - V_{s,k}^T U_k^{m,t}||^2 z_{s,m} e_{s,t,m} \tag{68}$$

$$+ \frac{1}{sumW_s} \sum_{i=1}^{T} \sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - I_{s,k} U_k^{s,j}||^2 w_{s,i,j} \tag{69}$$

$$e_{s,t,m} = \begin{cases} 1 & (\text{時刻 } t \text{ の銘柄 } s, m \text{ のデータがある場合}) \\ 0 & (\text{その他}) \end{cases} \tag{70}$$

$$z_{s,m} \;:\; \text{銘柄 } s \text{ と } m \text{ の関連度 } (Z \text{ は対称行列}) \tag{71}$$

$$sumE_s = \sum_{i=1}^{T} \sum_{m=1}^{C} e_{s,t,m} \tag{72}$$

$$V_{s,k} = sumW_s(\frac{1}{sumW_s} D_{U_k^s} + \frac{\gamma}{sumE_s}(A_k^s + 2B_k^s + C_k^s))^{-1} D_{U_k^s} \tag{73}$$

$$A_k^s = \sum_{t=1}^{T} \sum_{m=1}^{C} U_k^{s,t} U_k^{s,t^T} e_{s,t,m} z_{s,m},\;\; B_k^s = \sum_{t=1}^{T} \sum_{m=1}^{C} U_k^{s,t} U_k^{m,t^T} e_{s,t,m} z_{s,m},\;\; C_k^s = \sum_{t=1}^{T} \sum_{m=1}^{C} U_k^{m,t} U_k^{m,t^T} e_{s,t,m} z_{s,m}$$

③銘柄内・関連銘柄間の補完

$$\begin{aligned} \min_{V_{s,k}} J(V_{s,k}) = & \frac{\mu}{sumW_s} \sum_{i=1}^{T} \sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - V_{s,k}^T U_k^{s,j}||^2 w_{s,i,j} \\ & + \frac{\gamma}{sumE_s} \sum_{t=1}^{T} \sum_{m=1}^{C} ||V_{s,k}^T U_k^{s,t} - V_{s,k}^T U_k^{m,t}||^2 z_{s,m} e_{s,t,m} \\ & + \frac{1}{sumW_s} \sum_{i=1}^{T} \sum_{j=1}^{T} ||V_{s,k}^T U_k^{s,i} - I_{s,k} U_k^{s,j}||^2 w_{s,i,j} \end{aligned}$$

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

□w_{s,i,j} の種類

・Up：時刻 i, j のResidual Returnが共に正の場合
・Down：時刻 i, j のResidual Returnが共に負の場合

□Z_{s,m}の種類

①終値の値動きの方向性の一致度
N_{s,m} : 銘柄s, mの終値が共に上昇・下落した日数
M : 全取引日数
Z_{s,m} = N_{s,m} / M

②ResidualReturnでの動きの方向性の一致度
③終値の変化率のピアソン相関係数
④Residual Returnのピアソン相関係数
（⑤記事の中の共起度合）

■実験

前回の総合ゼミまでの結果は、コードミスで正確な実験ができていませんでした。
今実験をやり直していますが、今回は実験の一部のみ報告します。

【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の終値のResidual Returnが上昇・下落するかを予測

・予測対象
NY証券取引所に上場されている**2業界13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、%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

・予測モデル
ランダムフォレスト

今までの実験では、３つのデータソースがそろっている日のみをpick upして実験を行っていた。
今回からは、株価データ+(ニュースデータ or TRMI)のデータがある日をpick upして実験を行う。
nullはトレーニングデータを使って、平均で埋める。
⇒ 実験のデータ数が増える & より実用的

■テンソル補完で前処理　VS　前処理なし
①銘柄内補完
実験未完了

②関連銘柄間の補完
実験未完了

③銘柄内・関連銘柄間の補完

300					
テンソル補完	**あり**				**なし**
μ, γ, W, Z	0.1, 1, Up, ②	0.1, 1, Up, ③	0.1, 1, Down, ②	0.1, 1, Down, ③	-
素材	0.5576	0.5556	0.5564	0.5554	0.5161
景気連動型消費財	0.5031	0.5009	0.4995	0.4991	0.4953
All	**0.5324**	**0.5303**	**0.5302**	**0.5294**	**0.5065**

20					
テンソル補完	**あり**				**なし**
μ, γ, W, Z	0.1, 1, Up, ②	0.1, 1, Up, ③	0.1, 1, Down, ②	0.1, 1, Down, ③	-
素材	0.5510	0.5520	0.5509	0.5507	0.5293
景気連動型消費財	0.5045	0.5037	0.5067	0.5034	0.5025
All	**0.5295**	**0.5297**	**0.5305**	**0.5289**	**0.5169**

前処理としてテンソル補完をしたら精度がよくなった。

■今後の予定

・上記の追加の実験の実施
・テンソル補完はデータが不足している日の予測に強い？
⇒テストデータをデータが不足している日にしてみての実験

・企業間の類似度を、記事内の共起度合にする

・テンソルベースの予測モデルでの実験