

# 経済テキスト情報を用いた 長期的な市場動向推定

島崎 圭介

富山県立大学 情報システム工学科

2023 年 10 月 27 日

## はじめに

近年、機械学習を用いたテキストマイニング手法によって、テキスト情報と市場変動の関係性を発見し市場分析に応用する研究が増えてきた。

しかし、既存の研究には二つの欠点が大きくあげられる。

**第一に**数分から 24 時間以内の短期的な市場の反応を分析対象としていることである。

**第二に**特定のキーワードがもたらす市場への単発的なインパクトのみを扱っていることが多いことである。

## 本研究の目的

本研究では、金融実務家から要望が高い、数週間以上の長期的でさらに個別銘柄より広範な市場分析に CPR 法と呼ばれるテキストマイニング手法を加えて分析を行い既存のテキストマイニング法よりも優位性を検証する。

## 背景知識

今回先行研究に基づき、比較的少ない次元数の特徴量を抽出するため共起解析と主成分分析、回帰分析のステップからなる CPR 法を使用する。

## 共起関係に基づく主要単語の抽出

$$co(w_i, w_j) = \frac{D(t) \text{ のうち } w_i, w_j \text{ が共に出現する段落数}}{w_i, w_j \text{ の少なくとも一方が出現する段落数}} \quad (1)$$

上記の式によって  $w$  が用いられる条件式確率  $key(w)$  を計算する。

$$key(w) = [1 - \Pi(1 - \frac{\text{語 } w \text{ と土台 } g \text{ 中の語の共起度}}{\text{土台 } g \text{ 中の語の出現頻度}})] \quad (2)$$

これらの式によって上位  $N$  個をキーワードとして主要単語に加える。

## 背景知識

### 主成分分析による単語のグループ化

$$P(y_{ui} = k \| p_{ui}) = \begin{cases} 1, w_i \in \{HighFreq(t), HighKey(t)\} \\ 0, \text{それ以外} \end{cases} \quad (3)$$

上記の式により N 個の合成変数にまとめる. このデータを時間列順に並べることによって N 次元の時系列データが作成される.

### 回帰分析による市場データの動向分析

最後に、各主成分スコアから月次での市場科価格の動きを分析する.

## 全体の流れ

今回 PCR を用いて実験を行うが運用テストでは売買を月次として、毎月の金融経済月報が発表された時点で取引ルールにしたがって売買することとし平均損益を PCR 法以外での手法を用いて実験を行う。

今回実験で用いる PCR 法以外は、TF-SVR と Co-SVR、BOJ、EXT と また、今回 X やブログなどの多様な内容のテキスト情報など色々なものがあるが、長期的な市場分析に有効なテキストデータとして、日本銀行の金融経済月報を選んだ。この月報は、日本の銀行の中心の日本銀行の経済状況に対する態度が含まれているとされており、実際に多くの金融市場のトレーダの多くが着目している重要なテキスト情報であるので今回はこのテキスト情報のみとした。

## 運用テスト

実験モデルの運用モデルの実行期間を 2008 年 1 月から 2010 年 5 月までとして、各月の取引を決定するために 1998 年 1 月から前月までのテキスト情報・価格・数値指標を訓練データとして用いた。

## 評価

$$\begin{cases} 1 \text{ 単位の資本を買う, } \tilde{\Delta} > \Delta'(t) \text{ の場合,} \\ 1 \text{ 単位の資本を売る, } \tilde{\Delta} < \Delta'(t) \text{ の場合} \end{cases} \quad (4)$$

## 実験結果

表 1: 高変動期間での価格変動の正答率 (%)

|           | CPR          | TF-SVR | Co-SVR       | BOJ   | EXT   |
|-----------|--------------|--------|--------------|-------|-------|
| 日本国債 2 年  | <b>85.71</b> | 71.43  | 71.43        | 71.43 | 57.14 |
| 日本国債 5 年  | <b>85.71</b> | 57.14  | 42.86        | 28.57 | 42.86 |
| 日本国債 10 年 | 71.43        | 71.43  | <b>85.71</b> | 42.86 | 57.14 |

太字は各市場での最高精度。

CPR 法は日本国債 2,5 年での正答率は最高精度を出すことができた。

また、CPR 法は、日本国債 2 年物と 5 年物、Co-SVR 法では日本国債 10 年物で有意水準 5% で正答率 50 % の有意差を見られた。

## 改善点

今回の実験では日本国債について重点を当てて行ったが、同じ手法を用いて日本の株式市場や外国為替市場に適用させた結果、株式市場では有効であったが為替市場ではあまり有効ではなかった。また、今回は修飾語による表現の調整をあまり行えていなかったの  
で、他分野のテキスト情報にも応用できるようにしなければならない。

## まとめ

- CPR 法という新しい手法を提案し、他の手法よりも有意性があるということが示せた。
- CPR 法は日本銀行の態度変化に関する兆候や意図的なメッセージをうまく抽出できた。