経済テキスト情報を用いた金融市場動向の分析に関する研究

1X12C002-1　芦澤章太
指導教員　後藤正幸

1　研究背景・目的

近年，金融市場動向に影響を及ぼす多種多様の情報を用いて市場動向の分析・予測をすることが重要な課題となっている．市場動向と密接な関係をもつ情報の中でも，テキスト情報は非常に多く提供されているが，これらの情報は経済指標とは異なり数値化されていない．通常，これらのテキスト情報は大量に存在するため，人手による分析は困難であり，テキスト情報を半自動的に数量化して市場動向の予測に用いる分析モデルが必要となる．このような背景のもと，テキストマイニング技術と市場動向を分析するための統計的手法を組み合わせた方法が多く提案されている．

本研究では，それらの手法の一つである CPR 法 (Co-occurrence, Principal component analysis, Regression analysis) [1] に着目し，その改良を試みる．CPR 法は，各月の中旬に日本銀行によって発行される金融経済月報のテキストデータから特徴語を抽出し，その出現の有無によって数量化されたデータを主成分分析により圧縮したうえで，回帰モデルに当てはめ，得られる各月末の日本国債市場価格の予測値をもとに，日本国債市場動向を予測する方法である．

しかしながら，CPR 法には大きな 2 つの問題点が存在する．1 つ目は，特徴語抽出では単語の出現順序を考慮しないため，修飾語による表現の差異まで考慮できないという点である．2 つ目の問題は，CPR 法では目的変数である日本国債市場価格とは独立に主成分分析により次元が圧縮されるうえ，回帰モデルに適用する際に変数選択が行われるため，有用な情報を失う可能性がある点である．

本研究では，これらの問題を解決するために係り受け解析を用いた特徴語の組み合わせの抽出とリッジ回帰分析 [2] を組み合わせた新たな日本国債市場動向の分析手法を提案する．この手法により，先行研究における上記の 2 つの問題点が解決され，予測精度の向上が期待される．提案手法の有効性を示すために，実際の日本国債市場動向の予測シミュレーションを行う．

2　従来手法

CPR 法は，金融経済月報を用いてテキスト情報を数値化し，日本国債市場を予測するための手法である．この手法は，特徴語抽出と主成分分析，回帰分析の 3 つのプロセスから構成される．

2.1　特徴語の抽出 (C)

まず，金融経済月報を特徴語出現ベクトルで表現するために，各 t 月 $(1 \leq t \leq T)$ のテキストデータ $D(t)$ の特徴を表す単語として 2 種類の単語 $HighFreq(t)$, $HighKey(t)$ を抽出する．$HighFreq(t)$ とは，t 月の文書内において多く出現する単語であり，$HighKey(t)$ とは $Keygraph\ Algorithm$ [3] に基づいて抽出される，t 月において高い頻度で出現しないが，意味的に重要な単語である．CPR 法では，全テキストデータで抽出された全ての t 月における $HighFreq(t)$ と $HighKey(t)$ を要素とした d 個の特徴語の集合 $\mathcal{A}$ $(|\mathcal{A}| = d)$ を定義している．

2.2　主成分分析による特徴語出現行列の圧縮 (P)

まず，各月のテキストデータ $D(t)$ において各特徴語が出現すれば 1，出現しなければ 0 を取る要素からなる $T \times d$ の特徴語出現行列を定義する．ここで，抽出される特徴語数は文書データの数に対して多くなる．そのため，金融市場動向を予測するために特徴語を説明変数として重回帰分析を行う際，偏回帰係数を求めることができない．そこで，CPR 法ではこの特徴語出現行列に対し主成分分析を適用し，累積寄与率が 60% を超えるような最小の N_{pc} $(N_{pc} \leq d)$ に次元圧縮する．これにより，d 個の変数からなる特徴語出現行列を，N_{pc} 個の主成分スコアからなる行列へと圧縮する．

2.3　回帰分析による市場データの動向分析 (R)

テキストデータから抽出された特徴語と日本国債市場動向の関係を明らかにするために，主成分分析によって圧縮された特徴語出現行列の主成分スコアを説明変数，月次の市場価格データを目的変数とした重回帰モデルを推定する．また，t 月中旬に発行される金融経済月報から得られた特徴語出現行列の主成分スコアを説明変数として，得られた回帰式に当てはめ，t 月末における日本国債市場価格の予測値を推定する．ここで，$\tilde{y}_t$ を t 月末での日本国債市場価格の予測値とすると，日本国債市場動向は $\tilde{y}_t - \tilde{y}_{t-1}$ の値が正であれば上昇，負であれば下落すると推定することができる．

3　提案手法

3.1　概要

従来手法の特徴語の抽出方法では，テキスト情報に対して単語ごとに数量化するため，形容詞とそれに修飾される名詞の組み合わせによって意味を持つ単語の組み合わせが，単語ごとに別々に抽出されるという問題が存在する．また，CPR 法における主成分分析では，目的変数とは独立に説明変数の次元が圧縮される．そのため，目的変数に対して強い影響を持つ情報を損失する可能性がある．そこで本研究では，複数の単語の組み合わせである係り受けデータを従来の特徴語に加える．さらに，説明変数の数に対してデータの数が少ない場合においても偏回帰係数を推定することができるリッジ回帰分析を用いた日本国債市場動向の分析モデルを提案する．これにより，従来手法の問題点が解決され，日本国債市場動向についての予測精度の向上が期待される．

3.2　係り受け解析を用いた特徴語の抽出

本研究では，隣接する全ての形容詞と名詞の組み合わせを係り受け解析により抽出し，従来手法における特徴語，および抽出された形容詞と名詞の組み合わせを新たに「係り受け特徴語」として定義する．さらに，係り受け特徴語の数量化では，それぞれの係り受け特徴語の出現頻度を要素とする係り受け特徴語頻度行列を定義する．係り受け特徴語の出現頻度を説明変数，日本国債市場価格を目的変数とした回帰モデルを構築することにより，形容詞と名詞の組み合わせの違いによる国債市場価格への影響を捉えることができる．例えば,「強いドル」と「弱いドル」では「ドル」という名詞に対して正反対の意味の形容詞に修飾されることで,「ドル」が与える国債市場動向への影響は大きく異なるものと考えられる．提案手法を適用することで，形容詞と名詞の組み合わせの違いによる国債市場動向への影響を定量的に捉えることが期待される．

3.3　リッジ回帰分析による市場データの動向分析

係り受け特徴語それぞれの出現頻度を説明変数，日本国債市場価格を目的変数とした回帰モデルを推定する場合，デー

タの数に対して説明変数の数が多いため，偏回帰係数を求めることができない．これに対して，通常の回帰モデルに正則化項を加えることにより，変数選択を行わずに偏回帰係数を求めることが可能なリッジ回帰モデルが多くの事例に適用されている [2]．

いま，抽出された係り受け特徴語の数を d^* とすると，t 月における切片推定のための 1 と d^* 個の係り受け特徴語の頻度を表す項からなる d^*+1 次元のベクトル $\boldsymbol{x}_t = (1, x_{t1}, x_{t2}, \ldots, x_{td^*})^{\mathrm{T}}$ を説明変数ベクトルとする．このとき，Tヶ月間における係り受け特徴語の出現頻度行列を $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]^{\mathrm{T}}$, 偏回帰係数のベクトルを $\boldsymbol{w} = (w_0, w_1, \ldots, w_{d^*})^{\mathrm{T}}$，日本国債 10 年物市場価格を $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)^{\mathrm{T}}$ とすると，回帰モデルは以下の式 (1) で表される．

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon} \tag{1}$$

ただし $\boldsymbol{\varepsilon}$ は T 次元の残差ベクトルである．ここで，リッジ回帰分析に基づき，式 (2) を最小化するような偏回帰係数ベクトル $\boldsymbol{w}$ を求める．

$$J_\lambda = \frac{1}{2}||\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}||^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2 \tag{2}$$

式 (2) の第 1 項は通常の重回帰分析における二乗誤差項であり，第 2 項は正則化項である．この正則化項を加えることにより，データの数より説明変数の数が多い場合でも偏回帰係数を推定することが可能となる．よって，CPR 法に対して，主成分分析を行わずに係り受け特徴語頻度を説明変数，月次の国債市場価格データを目的変数とした回帰分析を行い，日本国債市場価格の予測を行うことができる．市場動向の予測方法は従来手法 [1] に従い，$\tilde{y}_t - \tilde{y}_{t-1}$ の値が正であれば上昇，負であれば下落するとして，日本国債市場動向を予測する．

4 提案手法を用いた市場動向の予測実験

提案手法の有効性を示すため，実際の日本国債市場価格データを用いて市場動向を解析した．

4.1 解析データおよび運用ルールと評価手法

1998 年 1 月から 2007 年 12 月までの 120ヶ月間の金融経済月報のテキストデータと，日本国債 10 年物の市場価格データ（月末終値）を学習データとして，CPR 法および提案手法によって，市場価格の予測モデルを構築した．得られたそれぞれの予測モデルに対して，2008 年 1 月から 2014 年 12 月における 72ヶ月間の同テキストデータをテストデータとし，予測モデルに当てはめて市場動向を分析した．さらに，以下のような運用ルールを適用し，累積運用損益率 U と変動動向の予測精度 V を評価指標とした．

運用ルールの売買は月次とし，毎月の金融経済月報が発表された時点で式 (3) に従って買いまたは売りのポジションを持つ取引と，月末にポジションを解消して損益を確定する取引を行う．取引量は毎月決まった資本量に固定し，売買量の調整は行わない．また，取引手数料は考慮しない．

$\tilde{y}_t$ を予測された t 月での月末価格，$\acute{y}_t$ を金融経済月報が公開された時点の価格とし，y_t を実際の月末の価格とする．つぎに，前月からの予測価格の変動幅を $\tilde{\Delta}_t = \tilde{y}_t - \tilde{y}_{t-1}$，月報発表時に実現している変動幅を $\acute{\Delta}_t = \acute{y}_t - y_{t-1}$ として，これらの値を比較し取引を決定する．

$$\begin{cases} 1\text{ 単位の資本を買う}, & \tilde{\Delta}_t > \acute{\Delta}_t \text{の場合}, \\ 1\text{ 単位の資本を売る}, & \tilde{\Delta}_t < \acute{\Delta}_t \text{の場合} \end{cases} \tag{3}$$

t 月末に月報発表時の取引と反対の売買を行い，損益 PL_t を決定する．t 月での損益 PL_t は，月報発表後の変動幅を $\Delta_t = y_t - \acute{y}_t$ としたとき，$\tilde{\Delta}_t - \acute{\Delta}_t$ との符号を比較し，以下の式 (4) のように表すことにする．また，累積運用損益 U は式 (5) のように表すことにする．

$$PL_t = \begin{cases} |\Delta_t|, & \Delta_t(\tilde{\Delta}_t - \acute{\Delta}_t) > 0 \text{ の場合}, \\ -|\Delta_t|, & \Delta_t(\tilde{\Delta}_t - \acute{\Delta}_t) < 0 \text{ の場合} \end{cases} \tag{4}$$

$$U = \frac{1}{72}\sum_{t=1}^{72}\frac{PL_t}{\acute{y}_t} \tag{5}$$

さらに，$\tilde{\Delta}_t$ の正負が $y_t - y_{t-1}$ の正負と一致する確率を変動動向の予測精度 V とした．

4.2 結果と考察

表 1，表 2 に，テストデータにおける各手法の累積運用損益 U と変動方向の予測精度 V の解析結果を示す．

なお，従来手法での特徴語抽出の結果，271 語が抽出された．提案手法での係り受け解析では，学習データでのすべての隣接する形容詞と名詞の組み合わせを抽出した結果，317 の組み合わせが抽出された．このうち，5%以上の学習データに出現する 77 の組み合わせを特徴語に加えて，それらの集合を係り受け特徴語とした．

表 1．各手法での累積運用損益率 (%)

	主成分＋重回帰分析	リッジ回帰分析
特徴語（従来）	35.60	37.95
係り受け特徴語（提案）	45.57	107.41

表 2．各手法での変動動向の予測精度 (%)

	主成分＋重回帰分析	リッジ回帰分析
特徴語（従来）	50.60	54.21
係り受け特徴語（提案）	56.62	60.24

表 1，表 2 より，係り受け解析とリッジ回帰分析のうち一方を適用した場合でも，改善が認められた．

さらに，2 つの手法を組み合わせた提案手法では，更なる改善を確認することができた．以上より，提案手法の有効性を示すことができた．

5 まとめと今後の課題

本研究では，金融経済月報を用いて，日本国債 10 年物の市場変動の方向性を予測する手法として，修飾語を考慮した係り受け解析とリッジ回帰分析を適用した新たな手法を提案した．また，実データでの実験により，提案手法の有効性を検証した．

今後の課題として，株価や為替などの他の予測対象への拡張や，価格変動の方向性だけでなく，価格水準に対する予測精度の向上などが挙げられる．

参考文献

[1] 和泉潔, 松井藤五郎, “経済テキスト情報を用いた長期的な市場動向推定”, 情報処理学会論文誌 52.12, pp.3309-3315, 2011.

[2] Hoerl Arthur E, Robert W Kennard, “Ridge regression: Biased estimation for nonorthogonal problems”,*Technometrics 12*, pp.55-67, 1970.

[3] Yukio Ohsawa, Nels E. Benson, Masahiko Yochida, “Keygraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor”,*Proc. Advanced Digital Library Confenrence(IEEE ADL’98)*, pp.12-18, 1998.