新聞記事を対象とした時系列テキスト分析による市場予測

松井藤五郎 [1*] 石田智也 [2] 中嶋啓浩 [2] 和泉潔 [3,4] 吉田稔 [3] 中川裕志 [3]
Tohgoroh Matsui[1] Tomonari Ishida[2] Akihiro Nakashima[2] Kiyoshi Izumi[3,4] Minoru Yoshida[3] Hiroshi Nakagawa[3]

[1] 中部大学 [1] Chubu University
[2] 野村證券株式会社 [2] Nomura Securities Co.,Ltd.
[3] 東京大学 [3] The University of Tokyo
[4] JST さきがけ [4] PRESTO, JST

Abstract: 本論文では，新聞記事を対象とした時系列テキスト分析の手法を提案する．本手法では，分析する時点のテキストとその直前のテキストを比較し，新たに出現した語，続けて出現している語，消滅した語を抽出して特徴ベクトルを作成し，SVM を用いてテキストの変化と市場の変化の関係を学習する．また，本手法を日本経済新聞の記事に適用し，東証株価指数（TOPIX）の騰落を予測した結果を報告する．

1 はじめに

テキスト・マイニング技術の進歩に伴い，インターネットや新聞などの豊富なテキストを分析することによって市場を予測する研究が盛んに行われるようになった．これらの金融テキスト・マイニングの研究については，筆者らの文献 [5] にまとめられている．

新聞や経済月報など，定期的に発行されるテキスト・データは，一種の時系列データである．すなわち，株価や金利など市場データである被説明変数が $y_1, y_2, \ldots$ と時系列を成すのと同様に，説明に用いられるテキストも $x_1, x_2, \ldots$ と時系列を成している．過去のデータから将来のデータを予測する時系列分析の分野では，直前のデータとの差分に着目することはよく行われている．

そこで，本研究では，定期的に発行されるテキスト・データを時系列データと捉えることによって，テキストの差分に着目した分析を行い，非説明変数の動きを予測する．本研究では，これを時系列テキスト分析と呼ぶ．

本論文では，新聞記事を対象とした時系列テキスト分析の手法を提案する．本手法は，分析する時点のテキストとその直前のテキストを比較し，新たに出現した語，続けて出現している語，消滅した語を抽出して特徴ベクトルを作成し，SVM [3] を用いてテキストの変化と市場の変化の関係を学習する．また，本手法を日本経済新聞の記事に適用し，東証株価指数（TOPIX）の騰落を予測した結果を報告する．

* 連絡先: TohgorohMatsui@tohgoroh.jp, http://とうごろう.jp

2 新聞記事を対象とした時系列テキスト分析による市場予測

2.1 時系列テキスト分析

従来のテキスト分類を用いた市場予測では，x_t を時刻 t におけるテキスト，$y_{t+\Delta}$ を時刻 $t+\Delta$ における市場の値（株価，金利，出来高，トレンドなど）として，

$$y_{t+\Delta} = f(x_t)$$

となるような関数 $f : X \to Y$ を，$x_t \in X$ と $y_{t+\Delta} \in Y$ の組 (x_t, y_t) の集合から学習する．ここで，X はテキストの集合，Y は市場の値の集合を表す．

本研究では，テキストの時系列性に着目し，直近 m 個のテキスト $x_{t-m+1}, \ldots, x_t$ から $y_{t+\Delta}$ を出力する関数 $f : X^m \to Y$，すなわち

$$y_{t+\Delta} = f(x_{t-m+1}, \ldots, x_t)$$

となるような関数を学習する．本論文では，これを時系列テキスト分析と呼ぶ．

2.2 提案手法

本論文では，新聞記事を対象として，時系列テキスト分析によって市場の動きを予測する手法を提案する．

2.2.1 テキストの差分に基づく出現パターン

本論文では，テキストが時系列データであることを利用して，テキストの差分に基づいた特徴語選択と量子化

を行う．

まずはじめに，営業日ごとに新聞記事に対して MeCab [2] を用いて形態素解析を行い，TermExtract [4] を用いて専門用語を抽出する．

語 T が前営業日のテキスト x_{t-1} に出現しておらず，かつ，当営業日のテキスト x_t に出現しているとき，これを新出と呼ぶ．T が x_{t-1} に出現しており，かつ，x_t にも出現しているとき，これを続出と呼ぶ．T が x_{t-1} に出現しており，かつ，x_t には出現していないとき，これを消滅と呼ぶ．本論文では，これらを T の出現パターンと呼び，T が新出するパターンを T^{a}，続出するパターンを T^{c}，消滅するパターンを T^{d} と表す．

予測対象日とその前営業日のテキストから抽出された専門用語の新出，続出，消滅について，それぞれ次節で説明する重み付きロジット差を計算し，重み付きロジット差が大きい語と小さい語を選択して特徴語とする．

例えば，「金融危機」という語が抽出されているとき，過去の新聞記事において「金融危機」という語が新出した場合 (a)，続出していた場合 (c)，消滅した場合 (d) のそれぞれについて，上昇日に生じる確率と下落日に生じる確率に基づく重み付きロジット差を計算し，特徴語を選択する．

2.2.2 重み付きロジット差

二つの事象の起こりやすさの違いを表す尺度としてロジット差がある．本研究では，リターンの絶対値が大きいときに起こりやすい事象を重視するため，リターンの絶対値を用いて重み付けしたロジット差を用いる．

本論文では，重み付きロジット差を次のように定義する．

$$WLD(T^p) = \log \frac{\Pr(T^p|+)}{1-\Pr(T^p|+)} - \log \frac{\Pr(T^p|-)}{1-\Pr(T^p|-)}$$

ここで，$\Pr(T^p|+)$ は上昇日における出現パターン T^p の重み付き出現確率，$\Pr(T^p|-)$ は下落日における T^p の重み付き出現確率であり，次のように求められる．

$$\Pr(T^p|+) = \frac{\sum_{+\wedge T^p} |R_i|}{\sum_{+} |R_i|}$$
$$\Pr(T^p|-) = \frac{\sum_{-\wedge T^p} |R_i|}{\sum_{-} |R_i|}$$

ここで，$\sum_{+} |R_i|$ は上昇日のリターンの絶対値の合計，$\sum_{+\wedge T^p} |R_i|$ は特徴語出現パターン T^p が起こった上昇日のリターンの絶対値の合計，$\sum_{-} |R_i|$ は下落日のリターンの絶対値の合計，$\sum_{-\wedge T^p} |R_i|$ は T^p が起こった下落日のリターンの絶対値の合計を表す．

2.2.3 訓練データの絞り込み

過去データのリターンには絶対値が大きいものも小さいものも含まれている．このうち，リターンの絶対値が非常に小さいものは何らかのノイズによって符号が逆転している可能性がある．例えば，リターンが 0.001% だったとき，同じ状況でリターンが（0 に近い）負になることは十分に考えられる．本手法はリターンの正負を分類するため，このようなノイズはラベル誤りとなって学習に大きな影響を及ぼす．そこで，過去データからリターンの標準偏差 σ を求め，リターンの絶対値が $k\sigma$ 未満のものを訓練データから取り除く．ここで，k は閾値を表すパラメーターである．

2.2.4 出現パターンに基づくテキスト分類

本手法では，テキスト分類には SVM [3] を用いる．SVM はマージン最大化に基づく機械学習の手法であり，高次元のデータに対しても利用可能なこととカーネル・トリックを用いることによって非線形分離問題を扱えることから，多くの研究で分類器として用いられている．

選択された m 個の特徴語出現パターン $T^p_1, \ldots, T^p_m$ に対して，特徴語出現パターン T^p_i が生じているときに第 i 次元の特徴量を 1，そうでないときは 0 とする．このようにして，m 次元の特徴量ベクトルを作成し，これを SVM が学習する関数 f への入力とする．SVM が学習する関数 f の出力は，リターンが正のとき 1，そうでないときは -1 とする．

3 実験結果

提案手法を用いて，評価実験を行った．新聞記事は日本経済新聞の朝刊および夕刊とし，その見出しだけをテキストとして使用した．予測対象は東証株価指数 (TOPIX) の日次終値とした．

2006 年から 2008 年までの市場を予測対象として，予測対象日の直近の過去 5 年間の新聞記事と市場データを用いて提案手法を適用して学習を行い，学習した予測モデルを用いて予測対象日の終値が前営業日の終値に対して上昇するか下落するかを予測した．

SVM のツールは SVM$^{\mathrm{light}}$ [1] を用い，カーネルは線形カーネルとした．特徴ベクトルの次元数は 5,000 とし，重み付きロジット差の大きい出現パターンと小さい出現パターンを 2,500 個ずつ選択した．

結果を図 1 に示す．2006 年から 2008 年までの 3 年間について，年ごとの平均予測精度と，3 年間の平均予測精度を表している．リターンの絶対値が小さい訓練デー

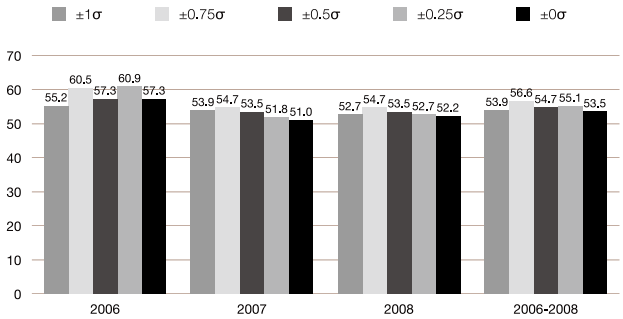

図 1　予測結果．縦軸は予測精度（%）を表す．

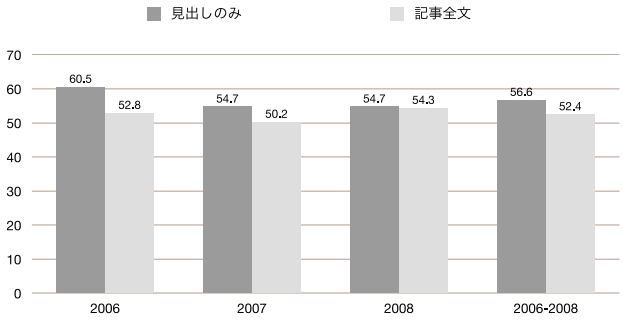

図 2　見出しだけを用いた場合と記事全文を用いた場合の比較．縦軸は予測精度（%）を表す．

タを除外する閾値を $k = 0, 0.25, 0.5, 0.75, 1$ として，閾値による予測精度の違いを調べた．$k = 0$ は訓練データを除外しないで全データを使用することを意味する．

予測精度は $k = 0.75$ のときが最も良かった．この結果から，リターンの絶対値が小さい訓練データを除外することが有効であることがわかる．そこで，これ以後の実験では全て $k = 0.75$ とした．

次に，見出しだけでなく記事全文をテキストとして用いた場合との比較を行った．この結果を図 2 に示す．記事全文を用いた場合の結果は，見出しだけを用いた場合よりも悪くなった．この結果から，見出しだけを用いることが有効であることがわかる．

続いて，特徴語選択の対象となる専門用語を予測対象日とその前営業日のテキストだけとした場合と，予測対象日を含む過去 5 年間全てのテキストとした場合の比較を行った．この結果を図 3 に示す．特徴語選択の対象を予測対象日を含む過去 5 年間全てのテキストとした場合の結果は，予測対象日とその前営業日のテキストだけを対象とした場合の結果よりも悪くなった．この結果から，特徴語選択の対象を予測対象日とその前営業日のテキストだけに限定することが有効であることがわかる．

最後に，予測対象日のテキストと前営業日のテキストの差分を用いた場合と従来どおり予測対象日のテキストに出現するか否かを用いた場合の比較を行った．この結

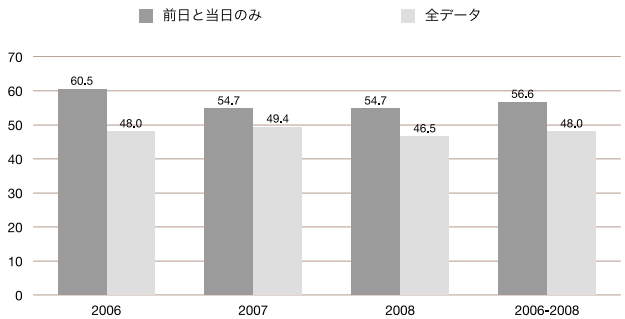

図 3　特徴語選択の対象となる専門用語を予測対象日とその前営業日のテキストだけとした場合と予測対象日を含む過去 5 年間全てのテキストとした場合の比較．縦軸は予測精度（%）を表す．

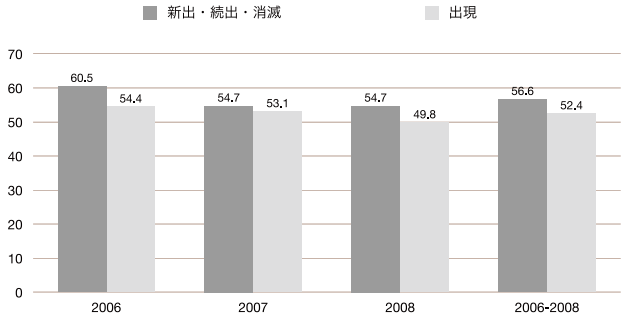

図 4　予測対象日のテキストと前営業日のテキストの差分を用いた場合と従来どおり予測対象日のテキストに出現するか否かを用いた場合の比較．縦軸は予測精度（%）を表す．

果を図 4 に示す．従来どおり予測対象日のテキストに出現するか否かを用いた場合の結果は，予測対象日のテキストとその前日のテキストの差分を用いた場合の結果よりも悪くなった．この結果から，テキストの差分に基づく分析が有効であることがわかる．

表 1 と表 2 に，2008 年 12 月 30 日を予測対象日としたときに選択された特徴語の出現パターンの上位 20 語と下位 20 語を示す．上位の出現パターンは上昇日に起こりやすく，下位の出現パターンは下落日に起こりやすいものである．表より，予測対象日のテキストには出現しないため従来手法で使われることがなかった「消滅した語 (d)」も多く使われていることがわかる．

4　考察とまとめ

見出しだけを用いずに記事全文を用いた場合に精度が悪くなるのは，記事全文を用いると含まれる語が多すぎて出現する確率が高くなるためだと考えられる．たとえば，「円高」という語が見出しに出現する確率と記事全文に出現する確率を比較すると，記事全文には見出しも

表 1　2008 年 12 月 30 日を予測対象日としたときに選択された特徴語の出現パータン上位 20 語．出現パターンの a は新出，c は続出，d は消滅を表す．

順位	語/出現パターン	$WSD(T)$
1	コロンブス/d	7.4
2	老女/a	7.1
3	金融危機下/d	5.7
4	静岡銀行/c	4.8
5	防災格付け融資/a	4.6
6	財源移譲/d	4.4
7	ＷＢＡフライ級/d	4.4
8	北島康介選手/d	4.4
9	粉ミルク汚染/d	4.3
10	ＧＳユアサ株/d	4.2
11	大坪/d	4.1
12	交通局/a	4.1
13	注入/c	4.1
14	協定廃止/d	4.1
15	オーシャンズ/d	4.0
16	ナイ/a	3.9
17	ＡＩＧエジソン/c	3.9
18	谷社長/a	3.9
19	谷村新司/a	3.8
20	函館タクシー/a	3.8

表 2　2008 年 12 月 30 日を予測対象日としたときに選択された特徴語の出現パータン下位 20 語．出現パターンの a は新出，c は続出，d は消滅を表す．

順位	語/出現パターン	$WSD(T)$
1	畜産価格/c	−7.8
2	蜃気楼/d	−6.0
3	うた/c	−5.3
4	関西経済同友会/c	−5.3
5	美肌師佐伯チズ/a	−4.8
6	撃退/c	−4.8
7	３つ/c	−4.6
8	キユーピー/c	−4.5
9	１０代/c	−4.3
10	国補助金/d	−4.1
11	オバマ次期米大統領/c	−4.1
12	遺棄容疑/c	−4.1
13	出稼ぎ農民/a	−3.9
14	再生医療/c	−3.9
15	一般事務/a	−3.8
16	業種別日経平均/d	−3.7
17	あいりん地区/d	−3.7
18	紀陽銀/c	−3.7
19	献身/d	−3.6
20	おせち/c	−3.5

含まれるため後者の方が常に大きい．重要な語ほど何らかの形で記事全文に含まれる可能性が高く，出現パターンは続出 (c) になりやすい．これが原因で，記事全文を用いたときに精度が低くなってしまうと考えられる．

特徴語を過去データの全てのテキストを対象として選択した場合，予測対象日にはあまり関係のない語が特徴語として選択されてしまう．たとえば，「解散総選挙」という語が過去 5 年間のある期間で非常に重要だったとしても，予測対象日が解散総選挙に関係ない場合はこれを特徴語としても意味がない．このような理由から，特徴語の選択対象を予測対象日とその前営業日のテキストに限定することが有効に働くと考えられる．

これまでのテキスト・マイニングを用いた市場予測の研究ではテキストそのものを量子化して分類や回帰を行っていたが，本論文では，テキストの差分を量子化し，それに基づいて分類を行う手法を提案した．テキストの差分を用いることによって，時系列テキストの変化に基づく分析が可能となった．本論文ではテキストの差分だけに基づいて分類を行ったが，今後は従来手法や被説明変数の過去の値を組み合わせてより高い予測精度が得られる手法の開発に取り組みたい．

参考文献

[1] Thorsten Joachims. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

[2] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Appliying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 230–237, 2004.

[3] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[4] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[5] 和泉潔, 松井藤五郎. Web 上のテキストから金融市場が予測できるか―金融テキスト・マイニング研究の紹介―. 信学技報, 第 111 巻, pp. 15–19, 2011.