3L3-OS-16a-6

新聞記事の時系列テキスト分析による株式市場の動向予測

Stock Market Prediction by Time-series Text Analysis for Newspaper

松井 藤五郎[*1*2] 和泉 潔[*3]
Tohgoroh Matsui Kiyoshi Izumi

[*1]中部大学 生命健康科学部 臨床工学科
Department of Clinical Engineering, College of Life and Health Sciences, Chubu University

[*2]中部大学 工学部 情報工学科
Department of Computer Science, College of Engineering, Chubu University

[*3]東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

This paper describes a new method to predict market by analyzing text data. This method analyses economic newspaper as time-series text data. It focuses on the difference between the newspaper published in the day that we want to predict market and that published in the previous day. We show the experimental results for predicting whether TOPIX (Tokyo Stock Price Index) goes up or down. The accuracy of our proposed method was 71.4% and the average annual return was 149% from 2008 to 2013.

1. はじめに

金融市場の動きを予測する研究は、古くから行われている。これらの研究では、過去の時系列データから将来の値を予測するテクニカル分析と、財務諸表や経済指標などから値を予測するファンダメンタル分析が主流であったが、これらの定量的な分析だけでは市場の動きを予測できない。たとえば、ある企業が不正会計していたことが発覚するとその企業の株価は急落するが、このような株価の動きはテクニカル分析やファンダメンタル分析では予測することができない。そこで、インターネットや新聞などの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 [和泉 12, 和泉 11]。

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目的変数が $y_1, y_2, \ldots$ 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 $\mathbf{x}_1, \mathbf{x}_2, \ldots$ 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、予測対象の動きを予測する。本研究では、これを時系列テキスト分析と呼ぶ。時系列テキスト分析のアイディアは、筆者らが [松井 11] で提案したものであるが、本論文の内容は特徴語の選択方法や予測する対象が [松井 11] とは異なっている。

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・データとその直前のテキスト・データを比較し、新たに出現した語、続けて出現している語、消滅した語を特徴として抽出して特徴ベクトルを作成し、SVM [Vapnik 95] 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) の日中の騰落を予測した結果を示す。

連絡先: 松井藤五郎, TohgorohMatsui@tohgoroh.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、$\mathbf{x}_t$ を時刻 t におけるテキストを表す特徴語ベクトル、$y_{t+\Delta}$ を時刻 $t+\Delta$ における市場の値（株価、金利、出来高、トレンドなど予測したい値）として、

$$y_{t+\Delta} = f(\mathbf{x}_t)$$

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

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

$$y_{t+\Delta} = f(\mathbf{x}_{t-m+1}, \ldots, \mathbf{x}_t)$$

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 前処理

提案手法について説明する前に、特徴語を抽出するための前処理について説明する。ここで説明する前処理は、結果に大きな影響を与えるものであるが、提案手法は前処理に依存するものではない。

本論文では、日本経済新聞を分析対象とし、記事の見出しとリード（第 1 段落）を抽出する。予測対象日の前営業日の夕刊から予測対象日（当日）の朝刊までを結合し、予測対象日の値動きに影響を与えるテキストであると仮定している。ただし、このようにすると、休み明けの営業日（例えば、月曜日）のテキストが、連続した営業日（例えば、前日が休日でない火曜日）のテキストよりも長くなるが、その影響についてはまだ検討できていない。

このようにして、記事ごとにテキスト・ファイルを作成し、全文検索システムの Hyper Estraier を用いて全文検索用のインデックスを作成する。予測対象日と記事 ID を組み合わせた

ファイル名をつけることによって、全文検索によってヒットした語がどの日に出現したかを調べることができる。

また、抽出したテキストに対し、MeCab [Kudo 04] を用いて形態素解析を行い、その結果を TermExtract [中川 03] に入力して専門用語を抽出し、特徴語とする。TermExtract は、「日本経済新聞」のように MeCab による形態素解析によって「日本」「経済」「新聞」のように分割された名詞を結合して出現頻度を評価し、専門用語として抽出するものである。

このようにして抽出された特徴語に対し、全文検索を行い、訓練データの期間内に出現した回数 (document frequency) を調べ、k 回以上出現したものだけを取り出す。また、訓練データのの期間内に予測対象が上昇した回数を調べ、k 回以上出現したものの中から上昇した割合が θ 以上のものと $1-\theta$ 以下のものを取り出す。

2.3 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。

2.3.1 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。

パターン p が前営業日のテキスト $\mathbf{x}_{t-1}$ に出現しておらず、かつ、当日のテキスト $\mathbf{x}_t$ に出現しているとき、これを新出と呼ぶ。p が $\mathbf{x}_{t-1}$ に出現しており、かつ、$\mathbf{x}_t$ にも出現しているとき、これを続出と呼ぶ。p が $\mathbf{x}_{t-1}$ に出現しており、かつ、$\mathbf{x}_t$ 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 p の出現パターンと呼び、p が新出するパターンを (a)、続出するパターンを (c)、消滅するパターンを (d) と表す。

2.3.2 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 [Vapnik 95] 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 l 個の特徴語出現パターン $p_1, \ldots, p_l$ に対して、特徴語出現パターン 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、l 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。f の出力は、予測対象日の日中（寄りから引けにかけての）利益率が正またゼロとき 1、そうでないときは -1 とする。

3. 実験結果

提案手法の有効性を確認するため、評価実験を行った。日本経済新聞を対象として、予測対象日の前営業日の夕刊から予測対象日の朝刊までを一つのテキストとし、その見出しのみを用いた。予測対象は 2008 年からの 2013 年までの東証株価指数 (TOPIX) 連動型上場投資信託 (ETF)[*1] とし、予測対象日の寄りから引けにかけて TOPIX ETF の取引価格が上昇するか下落するか（終値が始値よりも高いか低いか）を予測した[*2]。訓練データの期間は、予測対象日の直近の過去 5 年間とした。

*1 証券コード 1306

*2 分析する対象が新聞であるため、提案手法による予測が可能となるのは朝刊が発行された後であり、この予測に基づいた取引ができるのが最も早くても前場の寄りである。また、TOPIX そのものは売買できないが、TOPIX 連動型の ETF（上場投資信託）は市場で売買できる。

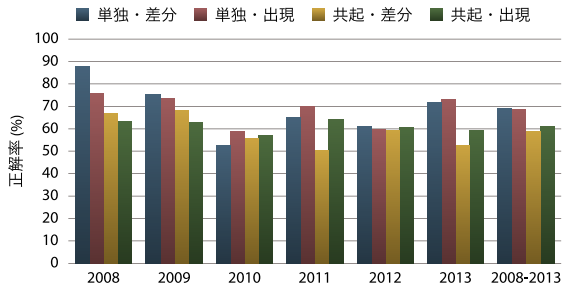

図 1: 手法による予測精度の違い

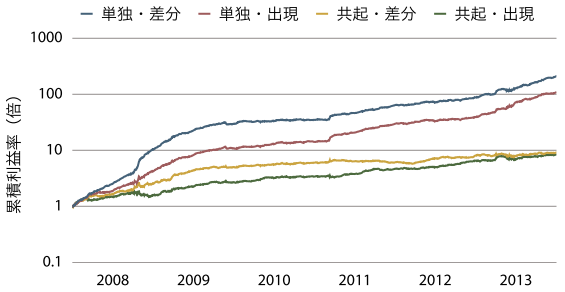

図 2: 手法による累積利益率の違い

出現パターンごとに出現数をカウントし、50 回以上出現したパターンのみを対象とし、パターンが出現した日に上昇した割合が 0.55 以上または 0.45 以下のパターンを抽出して特徴として用いた。SVM ツールには LIBSVM [Chang 11] を使用し、デフォルトのパラメーターを使った RBF カーネルを用いた学習を行った。

比較として、共起する語の出現パターンを用いた場合と、提案手法である時系列テキスト分析を行わずに語が出現したかどうかを特徴として用いた場合を用意し、これらを組み合わせた。

予測の正解率を図 1 に示す。「単独」は共起する語を用いずに語を単独で用いた場合、「共起」は共起する語を用いた場合を表す。「差分」は提案手法である時系列テキスト分析を行って新聞記事の差分を特徴として用いた場合、「出現」は提案手法である時系列テキスト分析を行わずに出現したかどうかを特徴として用いた場合を表す。グラフは 2008 年から 2013 年の年ごとと全体の正解率を表している。

2008 年から 2013 年までの全体では、時系列テキスト分析において語を単独で用いた場合の正解率が 68.9%で最も高かった。時系列テキスト分析で語の共起を用いた場合の正解率は 58.7%であった。時系列テキスト分析を使わなかったときは、語を単独で用いた場合の正解率が 68.3%、語の共起を用いた場合の正解率が 61.0%であった。

また、予測に基づいて取引を行った場合に獲得できる利益率による累積利益率を図 2 に示す[*3]。累積利益率は、違いがよくわかるように縦軸を対数軸にしている。提案手法で共起を用いなかった場合の累積利益率は 208 倍で最も高かった。これを

*3 手数料はなしとしている。

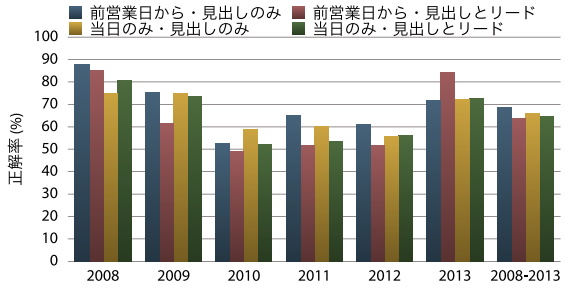

図 3: データによる予測精度の違い

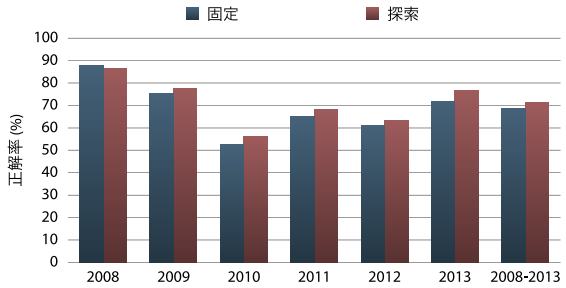

図 4: パラメーター探索による予測精度の違い

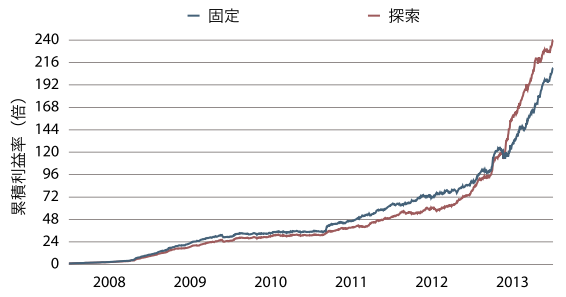

図 5: パラメーター探索による累積利益率の違い

平均年間利益率に換算すると 2.44 倍となる。次いで累積利益率が高かったのは時系列テキストを使わずに語を単独で用いた場合の 109 倍で、年間利益率に換算すると 2.18 倍 (118%) であった。

続いて、時系列テキスト分析において語を単独で用いた場合について、前営業日の夕刊から当日の朝刊までを分析した場合と当日の朝刊だけを分析した場合、見出しのみを分析した場合と見出しとリード（第一段落）を分析した場合を組み合わせて比較した。その結果を図 3 に示す。前営業日の夕刊から当日の朝刊までの見出しのみを分析したときの正解率が最も高かった。

最後に、これまでの実験において最も全体の予測精度が高かった、前営業日の夕刊から当日の朝刊までの見出しのみに対する語を単独で用いた時系列テキスト分析において、LIBSVM の機能を用いて SVM のパラメーター C と RBF カーネルのパラメーター γ の探索を行った。この機能を用いると、訓練データに対してクロス・バリデーションを行って最適なパラメーターを求め、それをテスト・データに適用する。その予測精度を図 4 に、累積利益率を図 5 に示す。累積利益率のグラフは、違いがよくわかるように、図 2 と異なり縦軸を対数軸にしていない。パラメーター探索を行った場合、全体の正解率は 71.4%、累積利益率は 238 倍、平均年間利益率は 2.49 倍 (149%) で、これまでの中で最も良い結果だった。

表 1: 上昇する確率が高い出現パターン

語	パターン	上昇確率	上昇/出現
常識	○新出 (a)	0.769	40/52
リビア	×消滅 (d)	0.695	41/59
牛肉	×消滅 (d)	0.684	39/57
Data	○新出 (a)	0.683	41/60
主要経済指標	○新出 (a)	0.680	34/50
日系企業	×消滅 (d)	0.677	42/62
電子版アンケート	×消滅 (d)	0.661	37/56
日経平均	×消滅 (d)	0.656	42/64
産業界	×消滅 (d)	0.655	38/58
天候不順	×消滅 (d)	0.654	34/52
LME 非鉄在庫	○新出 (a)	0.654	83/127
アジア株式市場	○新出 (a)	0.650	78/120
環境省	—続出 (c)	0.649	63/97
増収	×消滅 (d)	0.649	72/111
国際商品市況	○新出 (a)	0.644	76/118
規制委	×消滅 (d)	0.644	38/59
写真	—続出 (c)	0.643	45/70
県内企業	○新出 (a)	0.641	41/64
テクノロジー	○新出 (a)	0.640	57/89
定期預金	○新出 (a)	0.638	37/58

4. 考察

4.1 時系列テキスト分析で抽出された特徴語出現パターン

予測対象日が 2013 年 12 月 30 日のときに抽出された特徴語は、全体で 1,024 語あった。このうち、出現時に株価が上昇する確率が高いパターンを表 1 に、上昇する確率が低い（下落する確率が高い）パターンを表 2 に示す。

「常識」や「公開講座」など、一見すると株価の動きには影響がなさそうな語であっても、上昇確率に大きな偏りがある。日本経済新聞は、特定の曜日のみに掲載される記事もあるため、このような一見すると株価の動きには影響かなさそうな語であっても、曜日を推定するのに利用できるパターンとなる可能性がある。

また、「主要経済指標」が新たに見出しに載った日は上昇しやすく、「CME 日経平均先物」が新たに見出しに載った日は下落しやすいなど、経済の動きを伝える記事が掲載されると上昇確率に大きな偏りが生じている。

従来の出現パターン（出現しているか否か）に基づく手法では、新出 (a) と続出 (c) を同一視し、消滅 (d) を考慮できない。例えば、「LME 非鉄在庫」は、全体では 429 回出現し、そのうち上昇したのは 211 回で上昇確率は 0.492 であった。このように、従来手法では上昇確率にほとんど偏りがない（むしろ下落する確率の方が高い）という語でも、提案手法では上昇するときに出現する重要な語として抽出することができる。

また、「日経平均」が見出しに載っていない日（正確には、前日には見出しに載っていて、かつ、当日は見出しに載っていない日）は上昇しやすい、「FRB 議長」が見出しに載っていない日は下落しやすいというような出現パターン、つまり、消滅 (d) は従来手法では考慮できないが、提案手法では重要な語として抽出することができる。

4.2 予測精度と利益率

本論文の実験結果における予測精度は、最も高いもので 71.4% であった。スパム・メール・フィルターの分類精度などに比べると低い正解率だが、金融業界の実務家によると、市場予測に

表 2: 上昇する確率が低い（下落する確率が高い）出現パターン

語	パターン	上昇確率	上昇/出現
公開講座	×消滅 (d)	0.313	21/67
高松	一続出 (c)	0.327	18/55
怒り	○新出 (a)	0.333	17/51
こどもランキング	○新出 (a)	0.333	19/57
CME 日経平均先物	○新出 (a)	0.338	22/65
彩	一続出 (c)	0.345	19/55
中国銀	×消滅 (d)	0.347	26/75
ニュース	○新出 (a)	0.347	26/75
鋼材	○新出 (a)	0.350	21/60
FRB 議長	×消滅 (d)	0.351	27/77
技術開発	○新出 (a)	0.352	19/54
東京地裁	一続出 (c)	0.353	18/51
支払い	○新出 (a)	0.354	34/96
地球	○新出 (a)	0.356	26/73
暮らし	○新出 (a)	0.357	35/98
首脳会談	一続出 (c)	0.358	19/53
主要経済指標	×消滅 (d)	0.360	18/50
創論	○新出 (a)	0.360	18/50
新刊	○新出 (a)	0.360	18/50
躍起	×消滅 (d)	0.362	25/69

おいては、常に 55%以上を正解率を保つことができれば有用であると考えられる。提案手法は、年ごとにばらつきがあるものの、最も予測精度が悪い 2010 年でも 56.3%の正解率があり、この基準を満たしている。

また、予測精度に違いがなくても、利益率には大きな差が生じることがある。実際に、図 1 の比較において、単独の語を用いて時系列テキスト分析を行った場合（正解率 68.9%）と単独の語を用いて従来のテキスト分析を行った場合（正解率 68.3%）の予測精度の差は大きくないが、最終的な累積利益率はそれぞれ 208 倍と 109 倍であり、約 2 倍もの差がついた。これは、予測精度が同じでも、大きく変動している日に多く正解すると大きな利益率を得ることができるためである。今後は、上昇したか下落したかだけでなく、大きく上昇したか大きく下落したかを分類することなどが考えられる。

5. まとめ

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、テキストに連続性があることやテキストが時系列データであることを利用していなかった。そこで、本論文では、テキストが時系列データであることに着目し、テキストの差分から特徴的なパターンを抽出して量子化し、それに基づいてテキストを分類する手法を提案した。

提案手法を用いて日本経済新聞に対して時系列テキスト分析を行い、東証株価指数 (TOPIX) 連動型上場投資信託 (ETF) の日中の騰落を予測したところ、全体の正解率が 71.4%であり、最も予測精度が低い年でも正解率が 56.3%であった。また、この予測に基づく運用シミュレーションの平均年間利益率は 149%であった。金融業界の実務家によると正解率が常に 55%以上あれば有用であると考えられることから、提案手法は有用であるといえる。

本論文では金融市場の動きを予測することを目的として時系列テキスト分析を行っているが、時系列テキスト分析の概念は、国際情勢の変化や長期的な気候変動の検出などにも利用可能であり、今後の研究の発展が期待できる。

参考文献

[Chang 11] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, pp. 27:1–27 (2011)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying 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)

[Vapnik 95] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer (1995)

[和泉 12] 和泉 潔, 松井 藤五郎：金融テキストマイニングの紹介, 石田 基広, 金 明哲（編）, コーパスとテキストマイニング, 第 2 章, pp. 15–25, 共立出版 (2012)

[中川 03] 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)

[松井 11] 松井 藤五郎, 石田 智也, 中嶋 啓浩, 和泉 潔, 吉田 稔, 中川 裕志：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, 第 7 回人工知能学会ファイナンスにおける人工知能応用研究会 (SIG-FIN), pp. 44–47 (2011)

[和泉 11] 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, 第 70 号, pp. 15–19 (2011)