D-022

テキストマイニングを用いた金融時系列変化点の要因分析
Attribution analysis of change points on stock price series using text mining

東明翔 † 山崎高弘 ‡ 大野麻子 ‡ 常盤欣一朗 ‡
Akito Azuma Takahiro Yamasaki Asako Ohno Kin-ichiroh Tokiwa

1. はじめに

株式市場の動向に影響を与える投資家は，投資対象の価格や指数などの数値情報，あるいは企業からの広報情報や金融機関からの発表などのテキスト情報を判断しながら投資活動を行っている．投資家心理を表す指標として，ニュース記事などのテキスト情報からムードやセンチメントを獲得する試みがなされている．この手法は，投資家がニュース記事の中にある言語表現に対して反応し，投資活動を行っているという仮定のもとに成り立っている．Schumaker ら [1] は，ニュース記事中で 3 回以上存在する固有名詞とあらかじめ用意していた語彙をもとに，配信された記事のセンチメントから 20 分後の株価予測に影響を与えるかの推定を行っている．

一方で，時系列データに対して解析を行うことで，有益な情報を発見する手法も株式市場の分析に適用できると考えられている．ネットワークセキュリティ分野では，アクセスログの時系列データに対して異常検知の手法が用いられている．この手法では，通常と異なる挙動を示す箇所を発見することができ，株価時系列における暴騰・暴落のような突発的に起こる事態を予測することが可能となる．本研究では，異常検知手法の中で，時系列上の急激な変化を発見する変化点検出の手法に着目し，その手法を株価データに用いたときの有効性について検討する．また，テキスト情報である新聞記事から市場のセンチメント分析を行い，時系列変化点と楽観的な記事の割合の関連性を調べる．

2. 時系列分析とセンチメント分析

本研究の目的は，時系列上で何らかのイベントが発生したとき，変化点前後の日付の記事を評価することで，市場センチメントから時系列変化点の要因を分析することである．まず，時系列上の急激な変化の時点を発見するために Takeuchi ら [2] が提案している変化点検出手法の一つである Change Finder を用いる．この手法は，時系列データに対してリアルタイムに変化点の度合いを計算するため，自己回帰モデルを使用したオンライン 2 段階学習に基づくアルゴリズムによって，時系列データの各時点における変化点スコアを求める．第 1 段階では時系列中の外れ値しか検出できないため，外れ値スコアの平均を計算し，ノイズに反応した外れ値を除去している．そして，第 2 段階で本質的な変動のみの検出を行っている．

次に，新聞記事の評価を行うためにサポートベクターマシン (Support Vector Machine : SVM) を用いる．SVM は自然言語処理において良く用いられる線形二値分類器である．記事はあらかじめ定めた特徴ごとにベクトル化したものを扱う．まず，人手で 1 年間の株式市場に関する記事が楽観的な記事であるのか，慎重な記事であるのか，あるいは中立的な記事であるのかを分類し，SVM に入力する教師データとして作成した．そして，教師データとして用意した慎重な記事，楽観的な記事にそれぞれ評価値「1」と「2」を与えて SVM による機械学習を実行し，分類モデルを作成する．これにより楽観的な記事クラスと慎重な記事クラスを分離する評価関数を決定する．この評価関数を使って回帰計算を実行し，分類対象となっている記事の評価値が「1」に近ければその記事は慎重と判断し，「2」に近ければその記事は楽観的と判断する．

3. 実験

3.1. 対象データ

本研究では，2005 年 1 月 4 日から 2010 年 12 月 30 日までの株価時系列データから変化点検出を行う．検出のための学習期間を 2ヶ月間とする．一方，新聞記事 2005 年 1 月 1 日から 2010 年 12 月 31 日までの日本経済新聞を対象とする．株式市場とは無関係の記事を除くため，全記事から銘柄名をキーワードとして記事を識別した．本稿では，東証一部に上場かつ新聞記事数が多い上位 5 位の銘柄を対象とした．各銘柄ごとの 1 年間の新聞記事数は 500 件から 1,800 件である．

3.2. 変化点検出

図 1 は，上位 1 位の銘柄の株価時系列推移と変化点検出によって算出された変化点スコアを示したものである．図から変化点スコアが最も高く示されている日付の前後の時系列を見ると，株価が 7,000 円台から 5,000 円台まで変動していることがわかる．

しかし，2005 年から 2007 年までの期間で見ると，株価は上昇していると判断することもできる．ゆえに，数値データのみのでは，その変化点が下降傾向に転じて

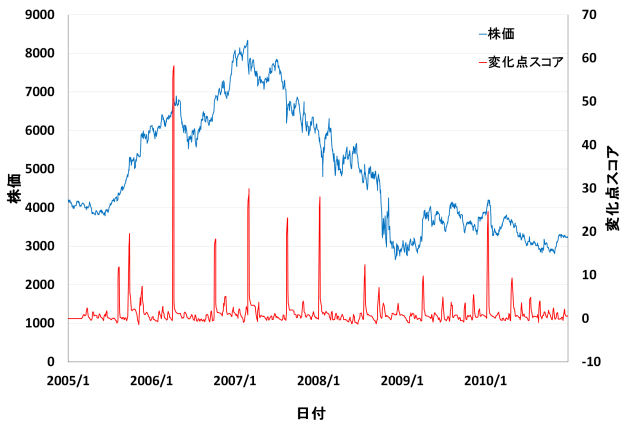

図 1: 株価時系列推移と変化点スコア

†大阪産業大学大学院工学研究科
‡大阪産業大学工学部

いるのか，上昇傾向を指し示しているのかを判断することは困難である．そこで，変化点スコアが高い値で示されている 2006 年の 1 年間に絞り，変化点前後の市場のセンチメントに変化が起きているのかどうかを検証する．

3.3. 市場センチメント

記事のセンチメントを分類し，市場のセンチメントを判定する手順を以下に示す．

1. SVM による回帰計算により，新聞記事の評価値を算出し，1 年間の記事の評価値の平均と標準偏差 (σ) を求める．
2. 平均を基準値とし，個々の記事ごとに評価値が基準値から σ 以上大きい記事を楽観的な記事，σ 以上小さい記事を慎重な記事として分類する．
3. 市場のセンチメントを判定するために，楽観記事率を直近 7 日間の楽観記事と慎重記事の数を集計して計算する．

$$\text{楽観記事率} = \frac{\text{楽観記事数}}{\text{楽観記事数} + \text{慎重記事数}} \quad (1)$$

4. 日付ごとに集計期間を変更しながら，楽観記事率を求める．

式 (1) から楽観記事率が 0.5 を超えれば，市場のセンチメントは楽観的とみなし，0.5 未満であれば慎重と判断できる．図 2 は，上述の手順によって求めた楽観記事率と変化点スコアである．本研究では変化点が検知されており，かつ楽観記事率が 0.75 より大きい日付について検証を行った．その条件をもとに取り出した値を表 1 に示す．変化点が検出された 4 月 6 日から 11 日までの変化点スコアは 55.2 から 58.3 と高くなっている．また，変化点スコアが最も高い日付の楽観記事率は 0.92 となっており，市場のセンチメントは，かなり楽観的であると確認できた．しかし，8 月 17 日から 22 日までの市場のセンチメントは楽観記事率が高いにもかかわらず，変化点が検出されていなかった．そこで，変化点が検出されている日付とされていない日付の記事数を調べた結果，8 月 17 日から 22 日の楽観記事数は

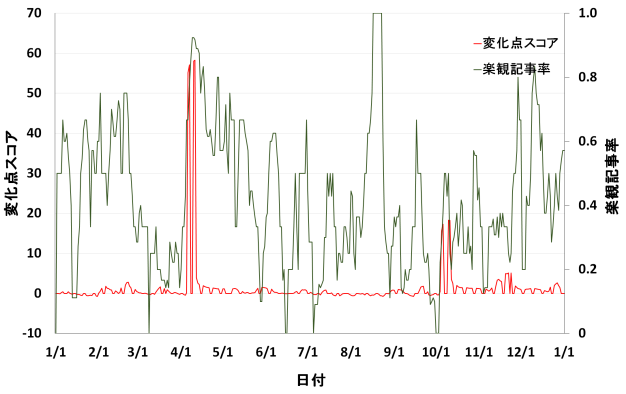

図 2: 2006 年の楽観記事率と変化点スコア

表 1: 2006 年の楽観記事率と変化点スコアの検証

日付	変化点スコア	時系列	楽観記事率	楽観記事数	慎重記事数
4/6	55.2	上昇	0.67	6	3
4/7	57.1		0.78	7	2
4/10	58.1		0.92	12	1
4/11	58.3		0.92	11	1
8/17	-0.41	上昇	1.00	3	0
8/18	-0.28		1.00	4	0
8/21	-0.24		1.00	2	0
8/22	-0.60		1.00	3	0

2 件から 4 件と少なく，慎重記事数は 0 件とまったく検出されていなかった．このことから楽観記事率が大きく，楽観記事数も多い条件のときの変化点スコアは比較的高く検出され，時系列も上昇傾向となっていた．

同様に，記事数が多い他の上位 4 銘柄についても同じ条件で検証を行った．変化点スコアが最も高い日付の楽観的記事率は上位銘柄順に 0.78，0.50，0.13，0.67 であった．楽観記事率 0.78 の日付の株価時系列は下降から上昇の変化点と一致していた．また，楽観記事率 0.13 の日付の株価時系列では上昇から下降の変化点と一致していた．この二つの銘柄で時系列変化点との関係性が発見できた．残りの楽観記事率 0.50 と 0.67 では，十分な有意差は見られなかった．

4. まとめ

本研究では，まず，非定常データである株価の時系列推移に変化点検出手法を用いて，株価の動向が変化する変化点を検出した．次に，個別銘柄ごとの新聞記事の評価値から，市場のセンチメントを求めた．算出された変化点と市場のセンチメントが，どのように影響を与えているかの検証を行った．結果として，時系列推移が上昇/下降，かつ変化点スコアが最も高く検出された日付前の記事は，楽観的/慎重な記事が多いことが分かった．また，高く割り出された楽観記事率の日付の変化点スコアは比較的高いことが分かった．記事数が多い 5 銘柄のうち 3 銘柄は比較的同じ結果になった．今回は楽観記事率の期間を 7 日間として検証したが，今後は期間を変更し，変化点から何日前後の新聞記事が，株式市場に影響を与えるかを分析する．

参考文献

[1] R. P. Schumaker, Y. Zhang, C. N. Huang, and H. Chen,“Evaluating Sentiment in Financial News Articles,” Decision Support Systems, vol.53, issue 3, pp.458–464, June 2012.

[2] J. Takeuchi and K. Yamanishi, “A Unifying framework for detecting outliers and vhange points from time series,” IEEE Transaction on Knowledge and Data Engineering,vol.18, issue 4, pp.482–492, 2006.