株価を用いたニュース記事評価と学習モデル間の比較

Estimating news articles' negative-pozitive from stock prices

五島圭一 [1*]　高橋大志 [2]　寺野隆雄 [1]
Keiichi Goshima [1]　Hiroshi Takahashi[2]　Takao Terano[1]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] 慶應義塾大学
[2] Keio University

Abstract: This study analyses the relationship between textual information and financial markets in Japan, focusing on headline News, a source of information that has immediate influence on the money market, and also which is regarded as an important source of information when making investment decisions. In particular we propose the objective way to estimate news articles' negative-positve by using mechine learning and statistics.

1　はじめに

投資家は，新聞やテレビ，各企業のプレスリリース，ソーシャルメディアなど，様々なメディアからニュースを入手し，投資先となる企業を選定する．ニュースには数値情報だけでなく，テキスト情報も含まれており，それらを活用することで数値情報だけでは説明することが難しい資産価格の変動やマーケットメカニズムなどの分析や予測ができる可能性がある．そのため，2000年代中頃から、資産価格の分野において，ニュースやソーシャルメディアといったテキストデータを，資産価格評価の分析に用いる試みが模索されている．例えば，Tetlock (2007) は Wall street Journal column から悲観度を抽出し，ダウ工業平均株価との関連性を見出している [6]．また，ソーシャルメディアと株価の関連性に言及している研究も存在する．Bollen et al. (2011) は，twitter の投稿内容を利用し，ダウ工業平均株価の変動を 87.6%の精度で予測できたとしている [5]．

このようにテキスト情報を用いることで，より正確な資産価格評価の試みがなされている．テキスト分析を行う際には，辞書の精度が重要となる [1]．Loughran and McDonald (2011) では，ファイナンスの文脈に沿ったテキスト評価の重要性を指摘しており，彼らは金融用の辞書を作成し，より精度の高い結果が得られたと報告している [7]．

しかしながら一方で，資産価格分析における文脈に沿ったテキスト内容の評価を行う際には，人の手によって，経験的に行われることになり，評価者の主観に強く依存してしまう可能性がある．それに対する解決策の一つとして，実際の資産価格からニュース記事を評価する方法があり，Healy and Lo (2011) では，外国為替を用いてニュース記事の評価を行い，リスク指標の作成を試みている [4]．

そこで本稿では，日本株式市場を対象に，ニュースと個別銘柄の株価情報を用いることで，より客観的かつ資産価格分析の文脈に即したニュース記事の評価方法を提示し，また，それらの分析方法を用いた株式投資戦略を構築し，本分析方法の有効性の検証を行った．特に、機械学習モデルと統計モデルによる評価方法をそれぞれ行うことによって、本分析方法の有効性の検証を試みる．次章は，データに触れ，3 章では分析方法，4 章では分析結果を記す．5 章は，まとめである．

2　データ

2.1　マーケットデータ

本稿では，個別銘柄の株価データについて，Thomson Reuter Datastrem から，トータルリターンの日次データを用いた．また，マーケットファクターのデータについては「日本上場株式　久保田・竹原 Fama-French 関連データ」からマーケットリターン (Rm)，リスクフリーレート (Rf)，バリューファクター (HML)，サイズファクター (SMB) の日次データを使用した．

*連絡先：
東京工業大学大学院総合理工学研究科知能システム科学専攻
〒 226-8502 神奈川県横浜市緑区長津田町 4259-J2-1705
E-mail: goshima.k.aa@m.titech.ac.jp

[1]本稿では，テキスト情報に極性（ポジネガ）を付与をするためのリストのことを辞書と呼んでいる．

2.2 ニュースデータ

ニュースデータについては，ロイターニュースを用いた．ロイターニュースは，トムソンロイター社の提供するニュースであり，本稿では，日本証券市場に関する日本語のニュース記事のみを分析対象とした．主に利用したタグ情報は，ニュースの発信日時・ニュースの見出し・各ニュースと関連する企業名（証券コード）を利用した．

本稿で用いるロイターニュースは，日本証券市場に参加している数多くの機関投資家がリアルタイムで閲覧するメディアであり，新聞やテレビニュースに比べ，イベントからニュース発信までのラグが小さく，ニュース発信時点において，資産価格に織り込まれていない情報を相対的に多く有すると考えられる．分析対象期間は 2009 年から 2010 年とし，分析対象企業は東証 1 部上場企業のみを分析対象とした．

3 分析方法

3.1 分析手順について

ここでは，本稿の分析手順の概略を記す．図 1 は，分析の流れを図にしたものである．

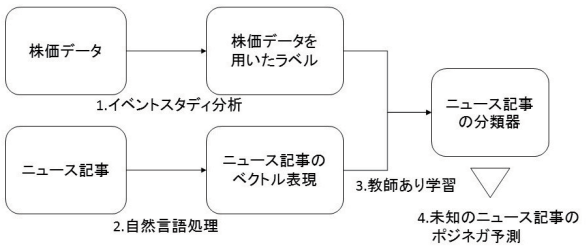

図 1: 本分析における手順の概略図

(1) はじめに，株価データを基に，ニュース記事にラベル（ポジティブ-ネガティブ）の付与を行った．株価データを基にした評価を行うことにより，客観的な記事評価を行うことが可能となる．本稿では，日本証券市場を対象としてイベントスタディ分析によって株価を教師情報としたニュース記事のラベルの生成を試みた．(2) 次いで，各ニュース記事を，bag-of-words に基づき，記事内容のベクトル表現を行った．(3) 更に，株価データからラベルを付与したニュース記事を訓練データとし，機械学習によってニュース記事へのポジネガ付与を行う分類器を作成し，(4) テストデータとなるニュース記事へのラベル付与を行った．2009 年のニュース記事を訓練データとし，2010 年のニュース記事をテストデータとした．以上の手順によって，本分析を進めた．次節以降において，それぞれの分析方法について詳細を記述する．

3.2 株価データからのラベル付与について

本稿ではイベントスタディ分析 [2] によって，株価データからニュース記事へのラベルの付与を試みた．

正常リターンを算出するためのモデルについては，Fama-French の 3 ファクターモデル [3] によって行った．また，モデルのパラメータを推定する際の推定期間に関しては，イベント日から 125 日前から 6 日前の 120 日間において推定を行った．イベントウィンドウに関してはニュース発信日の当日から 1 日後までの間とした．これは，ニュース記事が包含する情報を要因とした株価変動のみを抽出するためである．本稿で使用したニュースデータであるロイターニュースは報じられた日時が明確でイベント日を特定しやすいため，可能となると考えた．15 時以降に発信されたニュース記事については次の市場営業日に編入し，日付が市場休業日のニュースに関しても同様に，次の市場営業日に編入し，分析を進めた．

ここで，標準化を行い，ニュース発信日当日から 1 日後までの標準化された累積異常リターン $SCAR_i(0,1)$ を，当該ニュース記事が包含する情報を要因とした株価変動とし，ニュース記事の教師ラベルとした．

3.3 ニュース記事のベクトル表現について

テキスト分析をする際には，文書をベクトル表現することが求められる．本稿では、bag-of-words で表現を行うため，形態素解析，tf-idf 法，正規化を行った．そして，本稿においては名詞，動詞，形容詞の 3 つの品詞に注目し，抽出した．また，数値情報に関する名詞は除去をし，テキスト情報のみをベクトルの素性としている．

3.4 機械学習モデルと統計モデル

未知ニュースへのラベル付与については，機械学習モデルのひとつであるサポートベクトル回帰と統計モデルであるリッジ回帰と Lasso 回帰によって試みた．サポートベクトル回帰の学習器のパラメータチューニングについては，グリッドサーチによってハイパーパラメータの最適化を行っている．

4 分析結果

図 2 は、サポートベクター回帰によるテストデータのニュース記事のポジネガ予測結果である。横軸はニュー

ス記事が持つと予測される個別銘柄の標準化された累積異常リターン，縦軸は実際に実現した標準化された累積異常リターンを表している．平均 2 乗誤差は 10.2 であった。

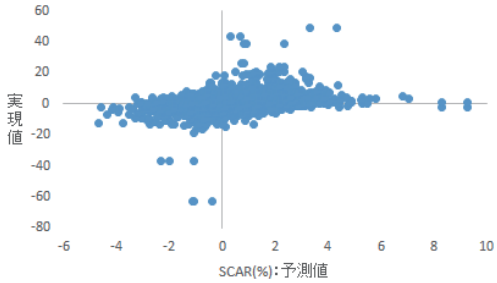

図 2: サポートベクター回帰による予測

図 3 と図 4 は、それぞれリッジ回帰と lasso 回帰によるテストデータのニュース記事のポジネガ予測結果である。平均 2 乗誤差はそれぞれ 10.3， 10.4 であった。

機械学習モデルと統計モデルどちらも、第二象限と第四象限にプロットされる事例が相対的に少ないのが見て取れる．特に，リターンを大きくプラスあるいはマイナスと推測したニュース記事については，正負を大きく間違える事例は少ないことが示唆される．また，0%付近の超過リターンの予測が困難であることが示唆される．分類手法の改善をはじめとしたより詳細な分析は，今後の課題である．

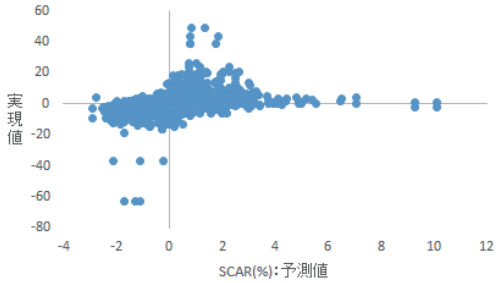

図 3: リッジ回帰による予測

5 まとめ

本稿では，ニュースと個別銘柄の株価情報を用いることで，より客観的かつ資産価格分析の文脈に即したニュース記事評価分析方法を提示した．特に、機械学習モデルと統計モデルによる評価方法をそれぞれ行うことによって、本分析方法の有効性の検証を試みた．分析の結果、機械学習モデルと統計モデルどちらもニュース記事の評価を通して、将来の株価予測ができる可能

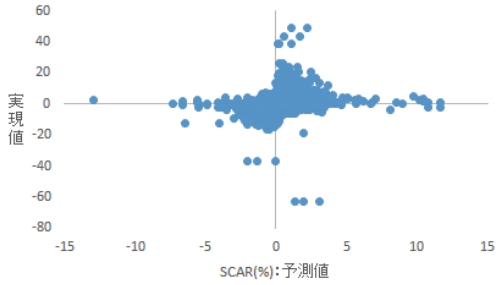

図 4: lasso 回帰による予測

性を見出した。今後の課題としては，バックテストによる有効性の検証や分析期間および分析対象資産の拡大などが挙げられる．

参考文献

[1] Bishop, Christopher M.: Pattern Recognition and Machine Learning, Springer (2006).

[2] Campbell, J. Y., A. W. Lo, and A. C. MacKinlay.: The Econometrics of Financial Markets, Princeton University Press (1997). 祝迫・大橋・中村・本多・和田訳: ファイナンスのための計量分析, 共立出版 (2003).

[3] Fama, E. F. and K. R. French.: Common risk factors in the returns on stock and bonds, *Journal of Financial Economics*, Vol. 33, pp. 3–56 (1993).

[4] Healy, Alexander and Andrew W. Lo,: Managing Real-Time Risks and Returns: The Thomson Reuters NewsScope Event Indices. In: Mitra, G. and Mitra L. (eds.), The Handbook of New Analytics in Finance, John Wiley & Sons, West Sussex, UK (2011).

[5] John Bollen, Hunia Mao and Xiaojun Zeng.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011).

[6] Paul C. Tetlock.: Giving Content to Investor Sentiment:The Role of Media in the Stock Market, *The Journal of Finance*, Vol. 62, No. 3, pp. 1139–1168 (2007).

[7] T. Loughran and B. McDonald.: When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65 (2011)