4D-04

パターンと教師あり機械学習と素性分析を利用したウェブと新聞からの株式相場に関わる知見獲得

村田 真樹† 中原 裕人† 馬 青‡

鳥取大学工学部† 龍谷大学理工学部‡

１．はじめに

株式相場の予測に関わる研究がなされている[1,2]。筆者らも株式相場の予測を行いたいと考えている。その手始めに株式相場に役立つ知見の収集を行っている。論文[3]では、新聞データを単語ネットワークを利用して分析することで、株式相場や経済に関わる様々な知見を取得した。本稿では、パターンと教師あり機械学習と素性分析を利用して、ウェブと新聞から株式相場における知見の収集を行う。

２．パターンに基づくデータ作成方法と機械学習の問題設定

データには、ウェブと新聞を利用する。株価の騰落に関わる文章をパターンで収集する。

ウェブでは、「X による株価上昇」「X による株価下落」のパターンでウェブから文章を収集し、X に相当する部分を入力テキストデータとする。例えば、「原油価格の混乱による株価下落」の表現から入力テキストとして「原油価格の混乱」を得る。

新聞では、「東京株式市場 X 日経平均株価…前日終値比」というパターンで抜き出し、X に相当する部分を入力テキストデータとする。ただし、X の個所が 5 文字以上あるもののみを使う。また、前日終値比の後の「…銭高」「…銭安」の個所を利用して、株価が上昇したデータかいなかを把握する。例えば、「１０日の東京株式市場は、前日の株価上昇に対し利益確定の売りが広がり日経平均株価は反落、一時、前日終値比３０６円９２銭安の１万６９３０円８５銭まで値を下げた。」という表現から、入力テキストとして、「は、前日の株価上昇に対し利益確定の売りが広がり」を得る。

入力テキストデータを入力として、株価上昇であるかいなかを出力として、教師あり機械学習を行う。

表１ 機械学習の正解率

データ	正解率
ウェブデータ	0.77
新聞データ	0.86
新聞データ（２回目）	0.64

３．機械学習による実験

機械学習には最大エントロピー法[4]を利用する。最大エントロピー法では、学習に役立つ素性の重みが得られるため、知見獲得に役立つ。

２節の方法でウェブデータを収集した。株価上昇と株価下落のデータの頻度を同じように調整したところ、株価上昇と株価下落のデータはそれぞれ 219 件得られた。

２節の方法で新聞のデータを収集した。毎日新聞(2007～2018 年)を用いたところ、株価上昇と株価下落のデータは 360 件と 349 件得られた。

最大エントロピー法で、株価上昇かいなかを推定する実験を行った。素性には、入力テキスト中のすべての 1 から 3 個の単語連続を用いた。新聞データでは、X に株価の変動理由以外に株価の変動そのものを示す表現が入りうる。変動理由のみを取り出したいため、新聞データでの実験では、株価の変動そのものに相当しやすい表現である、「好感」「嫌気」「全面安」「全面高」「買い」「売り」「買われ」「売られ」の素性を削除して実験を行った。推定の正解率を、10 分割クロスバリデーションでもとめた。

推定結果を表１に示す。表１の「新聞データ（２回目）」は、新聞データの実験において、有用とされた 200 個の素性の単語を含む素性を省いて実験を行ったものである（株価上昇の分類先の正規化α値（４節参照）の上位 100 個と下位 100 個のもので削除）。素性を省かない実験では、ウェブデータと新聞データの両方で 7,8 割という高い性能で、株価上昇かいなかを推定できている。

４．素性分析による知見獲得

全データを学習データとした場合の最大エントロピー法でもとまるα値を正規化した値(ここでは正規化α値と呼ぶ。株価上昇と株価下降の 2 個の分類先のうちどちらの方が重要かを示す二分類においてこのα値の和が 1 になるように正規化している。)をもとめた。この値が大きいほ

Acquisition of stock market knowledge from the web and newspapers using patterns, supervised machine learning, and feature analysis

†Masaki Murata, Yuto Nakahara, Faculty of Engineering, Tottori University

‡Qing Ma, Faculty of Science and Technology, Ryukoku University

本研究は、公益財団法人石井記念証券研究振興財団の助成金を受けて実施された。

表２　ウェブデータでの素性分析

素性	正規化α値	素性	正規化α値
アベノミクス	0.88	…	…
拡大	0.78	材料	0.28
政権	0.73	悪化	0.27
期待	0.71	不正	0.27
成長	0.69	円高	0.24
…	…	増資	0.16

表３　ウェブデータで得られた有用な素性

株価上昇	株価下落
アベノミクス、拡大、政権、期待、成長、買い、利益、緩和、円安、原油高、買収、自社株買い、トランプラリー、バブル、株式分割、仕手筋、好業績、業績拡大、黒田バズーカ	増資、円高、不正、悪化、懸念、リスク、金利、問題、影響、業績悪化、危機、下落、ショック、空売り、ＥＵ離脱、MSCB、利上げ、トランプリスク、崩壊、投げ売り、金利上昇、虚偽、粉飾、希薄、権利落ち、株式発行、安部退陣、TOB

表４　新聞データで得られた有用な素性

株価上昇	株価下落
円安、米国株高、上昇、一服、期待、反発、急伸、上昇、戻し、懸念が和らぎ、利下げ、ドル高、続伸、原油先物相場、大幅、改善	株安、円高、下落、利益確定、急落、懸念、反落、続落、先行き、問題、欧州、米国株安、大幅下落、円相場、大幅安

表５　新聞データ（２回目）で得られた有用な素性

株価上昇	株価下落
金利、追加、サブプライムローン、和らいだ、債務危機、堅調、急速な、利上げ	ドル安、金融不安、アジア株、進行、不透明、債務、急進、トランプ、膨らみ、中国の

どその素性が重要であり、小さいほど重要でないことを示す。

表２にウェブデータで株価上昇の正規化α値の上位 5 個と下位 5 個を示す。「アベノミクス」「期待」「成長」が株価上昇に役立つことがわかる。「増資」「円高」「不正」「悪化」が株価下落につながることがわかる。

ウェブデータで株価上昇の正規化α値の上位 100 個と下位 100 個を人手で考察して有用と思われた素性を表３に示す。

新聞データで株価上昇の正規化α値の上位 100 個と下位 100 個を人手で考察して有用と思われた素性を表４に示す。

ウェブデータでは、株価上昇で「株式分割」があり、株価下落で「増資、空売り、金利上昇、株式発行」などがあり、株式相場の初心者には、勉強になる知見が多い。新聞データよりも、ウェブデータの方が、多様な知見が獲得できていることがわかる。ただ、ウェブデータでの知見は、ウェブを記述している人の考えで書かれたものであり、実際にそれらの知見が正しいか、また役立つかはわからない。それに比べて、新聞データは、実際の株価のデータを扱っているため、実際の状況を知るのに役立つ知見となる。

新聞のデータでより多くの知見を獲得するために、３節で述べた新聞データ（２回目）の実験を行った。有用とされた 200 件の素性を省いて実験をすることで、有用とされた 200 件とは異なる素性により、新たな知見獲得を目指す。

新聞データ（２回目）の実験で株価上昇の正規化α値の上位 100 個と下位 100 個を考察して得られた有用な素性を表５に示す。１回目の新聞データの実験の表４で得られなかった新しい知見が得られている。「サブプライムローン」「債務危機」は本来は株価下落に関わる知見だが、「和らいだ」のような表現とともに出ることで株価上昇の知見となっている。「利上げ」は、本来は金利が上がることで株を買わず預金するという方向に向かい株価下落につながる事柄であるが、データを確認すると、日銀でのわずかな「利上げ」であり、景気が良好であることを日銀が認めたことによる利上げとして好感されて株価上昇につながったものであった。

新聞データ（２回目）の実験により、素性を省いて繰り返し機械学習を行うことで、新しい知見の獲得ができることがわかった。

５．おわりに

本稿では、パターンと教師あり機械学習と素性分析を利用して、ウェブと新聞から株式相場における知見の収集を行った。新聞データよりも、ウェブデータの方が、多様な知見が獲得できることがわかった。新聞データ（２回目）の実験により、素性を省いて繰り返し機械学習を行うことで、新しい知見の獲得ができることがわかった。ウェブデータと新聞から株式相場に関わる多くの知見が得られた。得られた知見を利用して、株式相場の予測を行う予定である。

参考文献

[1] Johan Bollen et al., Twitter Mood Predicts the Stock Market, Journal of Computational Science, Volume 2, Issue 1, pp. 1-8, 2011.
[2] 松井藤五郎ら，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，人工知能学会全国大会，2016.
[3] 村田真樹，金子徹，上東嵩，馬青，単語ネットワークを用いた経済と感情に関わる表現の分析，行動経済学会第12回大会，pp.1-6，2018.
[4] Masaki Murata, Kiyotaka Uchimoto, Masao Utiyama, Qing Ma, Ryo Nishimura, Yasuhiko Watanabe, Kouichi Doi and Kentaro Torisawa, Using the Maximum Entropy Method for Natural Language Processing: Category Estimation, Feature Extraction, and Error Correction, Cognitive Computation, Volume 2, Issue 4, pp.272-279, 2010.