機械学習と言語処理による株価予測と知識獲得

村田 真樹[†]　中原 裕人[†]　馬 青[‡]
[†]鳥取大学 工学部 電気情報系学科
[‡]龍谷大学 理工学部 数理情報学科
[†]murata@tottori-u.ac.jp,b16t2078y@edu.tottori-u.ac.jp
[‡]qma@math.ryukoku.ac.jp

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．本稿では，まず，教師有り機械学習と自然言語処理に基づいて行った株価予測の試みについて述べる．しかし，株価予測は簡単なものでない．そこで，株価予測に役立つような経済に関わる知識の獲得の研究も行っており，これについても本稿で述べる．

まず，2 節と 3 節で株価予測の研究について述べる．2 節では，新聞において行った株価予測の研究について述べる．株価予測を主とした研究であるが，経済に関わる知識獲得も少し行っておりそれについてもふれる．3 節では，ツィッターにおいて行った株価予測の研究について述べる．

次に，4 節と 5 節で経済に関わる知識獲得の研究について述べる．4 節では，機械学習を利用してウェブと新聞を用いて行った経済に関わる知識獲得の研究 [3] について述べる．2 節でも，機械学習による経済に関わる知識獲得を行っているが，それをより丁寧に行った研究である．また，新聞における知識獲得のみならず，ウェブからも知識獲得をする．5 節では，新聞を用いて構築した単語ネットワークを利用した知識獲得の研究 [4] について述べる．

2　新聞データを利用した株価予測と知識獲得

2.1　手法

実験データに 2007 年から 2018 年の毎日新聞のデータを用いる．株価の予測では当日の朝刊から当日の始値と終値の差の予測を行う．

教師あり機械学習を利用して，どのような語が記事中にあれば株価が上昇下降するのかを学習し，その学習結果を利用して株価の騰落を推定する．実験データには，2007 年から 2018 年の毎日新聞のすべての記事のタイトル，本文中の特定の単語 (前日終値比，前日比など) を含む段落を用いる．入力には，当日の上記データを利用する．テストデータに 2010 年から 2018 年を利用し，学習データはそれぞれの年をテストデータとする場合その直前 3 年のデータを学習データとする．株価の変化が上昇，変化なし，下降のいずれであったかを，入力に対する分類先として機械学習を行う．また，0.5 %以上の変動で上昇下降と定義する．それ以下の変動の場合は変化なしとする．教師あり機械学習には最大エントロピー法を利用する．機械学習の素性には TermExtract[5] により抽出した専門用語[*1]を用いる．

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，機械学習の素性を分析することで株式相場や経済に関わる知見を獲得する．

表 1: 新聞データでの実験結果

手法	正解率	総数	1 年間の平均利益 (円)	1 年間の損失最大値 (円)	t 検定 p 値
手法 1	0.451	2207	661	-1375	0.046
手法 2	0.438	819	279	-1775	0.207
手法 3	0.453	2207	778	-1716	0.057
ベースライン			1006	-3866	0.143

2.2　株価の予測実験

素性は専門用語とし，入力は以下の手法 1 から手法 3 のものを用いて実験を行った．テストデータは，手法 1 と手法 3 のデータ数は 2,207 件，手法 2 のデータ数は 819 件である．

- 手法 1：当日の朝刊のすべての記事のタイトル
- 手法 2：当日の朝刊の「前日比」「前日終値比」を含む段落
- 手法 3：当日の朝刊のすべての記事のタイトルと「前日比」「前日終値比」を含む段落

また，ベースラインは Buy&Hold(年始に買って年末まで保持し続ける方法) とする．それぞれの手法での正解率，1 年間の平均利益，1 年間の損失の最大値 (1 年毎で見た時，最も大きくなった場合の累積損失額) を表 1 に示す．ここでの利益および損失は日経平均 1 株あたりの値である．提案手法での株の売買方法は，上昇 (または下降) と判断した場合当日の始値で日経平均 1 株を買って（また売って）それを当日の終値で売る（または買い戻す）ものとし，変化なしと判断した場合は取引をしないものとする．利益の結果をもとに有意差を検定した．それぞれの手法について，各年の利益と取引をしなかった場合 (損益 0) のデータを利用して，両側検定の t 検定を行った．その際の p 値を表 1 に示す．

[*1] TermExtract は名詞の頻度や連接頻度を用いて専門用語を取り出すツールである．本研究の実験で取り出された専門用語には例えば「国連総長」「野党」「陸上」「信頼回復」「議員宿舎」などがあった．表 3 と表 4 にある表現も TermExtract で取り出された専門用語の例となる．

表 2: 新聞での手法 1 の実験結果の詳細

年	正解率	総数 (U,D,N)	1 年間の平均利益 (円)	1 年間の損失最大値 (円)	買い	売り	ベースラインの 1 年間の平均利益 (円)	ベースラインの 1 年間の損失最大値 (円)
10	0.346	243(57,63,123)	728	-409	14	96	-381	-1776
11	0.453	245(47,60,138)	597	-136	25	77	-1897	-2214
12	0.524	248(53,52,143)	39	-95	9	34	1835	-271
13	0.351	245(88,69, 88)	-464	-464	7	2	5687	-119
14	0.414	244(51,60,133)	1722	-21	92	15	1303	-2178
15	0.443	244(63,52,128)	2825	-442	35	58	1708	-517
16	0.392	245(69,65,111)	-178	-1376	23	33	296	-3866
17	0.657	248(38,33,177)	15	-614	18	10	3466	-1059
18	0.477	245(55,54,136)	672	-702	34	18	-2959	-3711
平均	0.451		661				1006	

手法 1 では p 値が 0.046 であり，有意水準 0.05 で有意差がみられた．

実験結果より，新聞のタイトル情報だけでも予測ができることがわかった．手法 2 では他の手法と比較して利益が少なくなっているが，これはデータ数が少なかったことが原因であると考えられる．データ総数が手法 1 と手法 3 では 2,207 であるのに対し，手法 2 では 819 であるため，これらの数を揃えた場合手法 2 の利益が手法 1 より多くなることも考えられる．

表 1 より，提案手法はベースラインと比較して利益は劣っているが 1 年間の損失の最大値は小さくなっている．また，手法 1 では有意差ありとなっているので，提案手法でもある程度の予測はできているといえる．

比較的性能の高かった手法 1 について，より詳細な実験結果を表 2 に示す．表の各列は左から順に, 予測を行った年, 各年の正解率, データ総数[*2]，利益, 累積の損益が最低となったときの値, 買い注文の回数, 売り注文の回数, ベースライン手法の利益, ベースライン手法の累積の損益が最低となったときの値である．表を見ると，利益が特に多い年があるなど偏りがあることがわかる．

松井らの研究 [2] では，日本経済新聞を利用して，7 割の正解率，年平均の投資額の増額比率 2.5 倍の結果を得ていた．それに比べるとかなり性能は劣るが，毎日新聞を利用した実験でもある程度予測ができる場合があることがわかった．

2.3 素性分析による知識獲得実験

株式相場を予測する上で，どのような素性が役に立つのかを明らかにするために，素性の分析を行う．機械学習の分類性能が高い素性は特に役立つ．そこで，機械学習の正解率を上げるために本節では 2 日前の終値と前日の終値の差を予測する．これは，株価の本当の意味での予測でない．株価の騰落を記載する際には概ね前日のことを新聞は書く．その新聞が書くであろう頃の株価を予測することで，株価の予測の正解率があがりやすくなり，知識獲得に役立つ．

[*2] U,D,N は，上昇、下降、変化なしが正解の場合のデータ数である．

表 3: 新聞データでの素性分析

素性	正規化 α 値	素性	正規化 α 値
主力	0.82	…	…
東京株	0.79	円高傾向	0.21
高値	0.78	円買い	0.21
雇用情勢	0.78	会合	0.20
続伸	0.78	反落	0.11
…	…	リスク	0.11

表 4: 2 節の実験での新聞データで得られた有用な素性

株価上昇	株価下降
回復，需要，中央銀行，景気，原油先物，大幅安，成長期待，経営難，円安傾向，懸念材料	不透明感，日銀，動揺，交錯，急騰，減速懸念，米経済，債権価格，深刻化

10 分割クロスバリデーションでもとめた正解率は 0.711 であった．

最大エントロピー法で求まる α 値を全分類先での合計が 1 となるように正規化した値をここでは正規化 α 値と呼ぶ．この値が高いほど，その分類先であることを推定するのに重要な素性であることを示す．

表 3 に新聞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．正規化 α 値が上位のものは株価上昇に役立つことが，下位のものは株価下降に役立つことがわかる．正規化 α 値の上位 100 個と下位 100 個を人手で考察することにより，得られた有用な素性を表 4 に示す．

3 ツィッターデータを利用した株価予測

実験データに 2018 年 11 月 5 日から 2019 年 10 月 21 日のツィッターデータを用いる．ツィッターデータは，日経という表現を含むものを集めた．学習データには，2018 年 11 月 5 日から 2019 年 4 月 30 日のツィッターデータを用いる．テストデータには，2019 年 5 月 1 日から 2019 年 10 月 21 日のツィッターデータを用いる．各日では 1,000 行のデータのみを利用する．学習データは 119 個、テストデータは 115 個である．

株価の予測では前日夜までのツィッターのデータを入力として，当日の始値と終値の差の予測を行う．株価の

表 5: ツィッターデータでの実験結果

手法	正解率	利益 (円)	損失最大値 (円)
提案手法	0.783	0	0
ベースライン		365	-1923

表 6: 知見獲得での機械学習の実験結果

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

変化が上昇，変化なし，下降のいずれであったかを，その入力データに対する分類先として機械学習を行う．また，0.5 %以上の変動で上昇下降と定義する．教師あり機械学習には最大エントロピー法を利用する．機械学習の素性には TermExtract[5] により抽出した専門用語を用いる．

実験結果を表 5 に示す．表のベースラインは 2 節と同じく Buy&Hold(初日の始値で買って最後まで保持し続ける方法) である．提案手法では機械学習の推定先はすべて変化なしであった．データ中では上昇，下降に比べて変化なしの頻度が多い．また実験に用いたデータが 1 年であり少ない．これらがすべて変化なしと推定した原因と思われる．今後データを増やして実験を行いたい．

4 ウェブと新聞を利用した機械学習による知見獲得

本節では，機械学習を利用してウェブと新聞を用いて行った経済に関わる知識獲得の研究 [3] について述べる (詳細は文献 [3] を参照)．2 節でも，機械学習による経済に関わる知識獲得を行っているが，それをより丁寧に行った研究である．また，新聞における知識獲得のみならず，ウェブからも知識獲得する．

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 に相当する部分をデータとする．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 に相当する部分をデータとする．このデータから，株価の騰落を教師有り機械学習で学習する．機械学習には最大エントロピー法を用いる．素性はデータ中の 1 から 3 個までの単語連続を利用する．

クロスバリデーションの性能を表 6 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．これを行うことで新たな有用な知見を獲得できる．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．

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 7 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 8 に示す．

表 7: ウェブデータで得られた有用な素性

株価上昇	株価下降
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 8: 新聞データで得られた有用な素性 (1 回目)

株価上昇	株価下降
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 9: 新聞データで得られた有用な素性 (2 回目)

株価上昇	株価下降
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

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，1 回目の実験で有用とされた 200 件の素性を省いて実験をする新聞データ (2 回目) の実験を行った．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ものである．

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 9 に示す．1 回目の新聞データの実験の表 8 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．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 単語ネットワークを利用した知見獲得

Bollen らは，ツィッターでの感情分析を利用して，株価の予測がある程度可能であることを示した [1]．

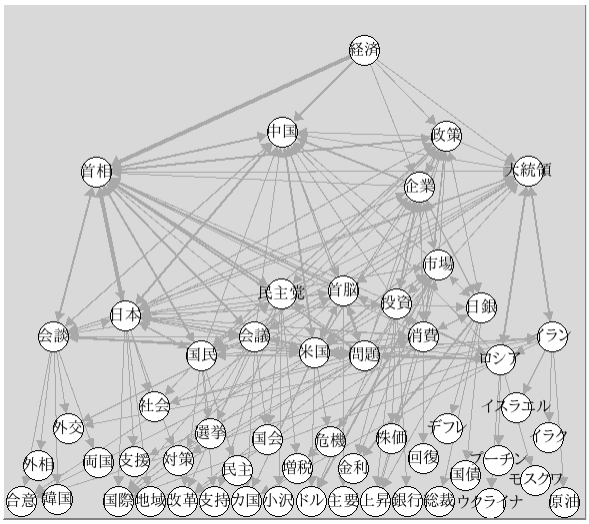

図 1: 単語ネットワークを利用した「経済」に関わる知見獲得

筆者らは，文献 [4] で，経済と感情の概念が特に重要と考えて，経済と感情に関わる概念について分析した (詳細は文献 [4] を参照)．具体的には，経済と感情に関わる表現を，単語ネットワーク [6] を用いて分析した．実際にその分析を，「経済」「景気」「財政」「感情」に関わる単語を対象に行った．

「経済」に関わる分析では，図 1 の単語ネットワークが得られた．「経済」の単語に対して比較的早く「中国」がつながっている．中国の影響力がわかる．また，「首相」や「大統領」や「政策」も経済と関係があることがわかる．すなわち，「経済」に関わる分析では，「中国」「首相」「大統領」が経済と関係が深く見えることがわかった．

同様な分析を，「景気」「財政」「感情」に関わる単語を対象に行った．「景気」に関わる分析では，景気がよいと消費が増える，金利が上昇すると株価は下落すること，さらに国債に影響することを示唆する結果が得られた．「市場」「企業」「日銀」が景気と関係が深いこともわかった．「財政」に関わる分析では，消費を増やすために財政を利用していること，財政のために増税が必要ということ，財政の観点でギリシャ支援が必要であることを示唆する結果が得られた．「感情」に関わる分析では，「事件」「映画」「ドラマ」「戦争」「テロ」「判決」「受賞」「恋愛」など，感情と関係がある事柄を把握する際に役立つ単語群が得られた．

2.3 節と 4 節で行った株価の騰落を元にした機械学習による知識獲得では，株価の騰落に関わる知識しか得られなかったが，単語ネットワークでは株価の騰落の知識が得られるのみならず株価の騰落以外の広範な知識が獲得できることがわかる．

6 おわりに

本稿では，機械学習と言語処理を用いて行った株価予測と経済に関わる知識獲得の研究について述べた．

株価予測の研究では，新聞を用いた場合の予測とツィッターを用いた場合の予測を行った．

新聞を用いた場合の予測では，タイトルを用いるだけでもある程度の予測ができることがわかった．提案手法は，利益はベースラインに劣ったが，最大累積損失は，ベースラインよりも少なかった．使い方によっては提案手法も役立つ場合があると思われる．

ツィッターを用いた場合の予測では，すべて変化なしと予測してしまい、売り買いをしないものとなった．これはまだデータが少ないためと思われる．今後はデータを増やして実験していきたい．

経済に関わる知識獲得の研究では，機械学習の素性分析による知識獲得と，単語ネットワークを用いた知識獲得を行った．

機械学習の素性分析による知識獲得では，種々の興味深い知見を獲得できた．特にウェブデータでの機械学習では，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．

単語ネットワークを用いた分析では，「中国」「首相」「大統領」が経済と関係が深いこと，景気がよいと消費が増える，金利が上昇すると株価は下落すること，さらに国債に影響すること，「市場」「企業」「日銀」が景気と関係が深いこと，消費を増やすために財政を利用していること，財政のために増税が必要ということ，財政の観点でギリシャ支援が必要であることを示唆する結果が得られた．また，「事件」「映画」「ドラマ」「戦争」「テロ」「判決」「受賞」「恋愛」など，感情と関係がある事柄を把握する際に役立つ単語群が得られた．

株価の騰落を元にした機械学習による知識獲得では，株価の騰落に関わる知識しか得られなかったが，単語ネットワークでは株価の騰落の知識が得られるのみならず株価の騰落以外の広範な知識が獲得できることがわかった．

謝辞

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．

参考文献

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8, 2011.
[2] 松井藤五郎, 和泉潔. 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, 中原裕人, 馬青. 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. 情報処理学会第 82 回全国大会, 2020.
[4] 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.
[5] 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.
[6] Yuta Doen, Masaki Murata, Ryuta Otake, Masato Tokuhisa, and Qing Ma. Construction of concept network from large numbers of texts for information examination using tf-idf and deletion of unrelated words. In *Proceedings of SCIS-ISIS 2014*, pp. 1108–1113, 2014.