3L4-OS-26b-3

言語情報を用いた経済指標の予測と分析

Financial Trend Prediction and Analysis through Textual Data

藤川 和樹 *1　関 和広 *1　上原 邦昭 *1
Kazuki Fujikawa　Kazuhiro Seki　Kuniaki Uehara

*1神戸大学大学院システム情報学研究科
Graduate School of System Informatics

This study proposes an approach to financial trend prediction, specifically a stock market, based on integrated multi-document representation by deep neural networks. The representation allows us to convert multiple texts reporting independent events during a day into a single vector of values, yet distinguishes the concepts of the original events. The validity and effectiveness of the proposed approach for market trend prediction are evaluated on real-world data for Nikkei 165 companies.

1. はじめに

投資家は，投資商品の価格や政策金利などの数値情報，日本銀行や金融機関の発表，為替や企業に関するニュースなど，数多くの情報の中から有用な情報を発見し，投資の意思決定に役立てている．市場分析に用いられるこれらの情報は，主に 2 種類に分類することができる．1 つは企業の株価や物価指数などの数値情報，他方は市場に対して影響力を持つ人物の発言や企業の動向，事件・事故を知らせるニュース記事といったテキスト情報である．しかしながら，量的，時間的な制約から，投資家たちがこの膨大な数の情報すべてに目を通し，市場分析に利用することは難しい．そこで，市場情報を推論するエキスパートシステムの構築や，ニューラルネットワークや遺伝的アルゴリズムを用いた市場分析など，人工知能分野の手法や技術を金融市場予測へ応用する研究が行われてきた．これらの研究は一定の成果をあげてきたものの，多くの研究は数値情報のみを利用しており，市場分析に有効な情報をすべて活用しているとは言えない．

一方，新聞記事等のテキスト情報を市場分析に用いる従来研究では，テキストに現れる単語の独立性を仮定した Bag-of-Words（BoW）モデルを採用していることが多い．しかしながら，新聞には様々な銘柄に関係する記事が混在しているため，別々の記事に出現した単語をすべて独立に扱うのは適切ではない．例えば，2007 年 8 月 6 日の日本経済新聞には,「トヨタの中間連結決算が過去最高を更新した」という内容の記事と,「住友不動産がマンション供給を下方修正した」といった内容の記事が含まれており，これらの記事から抽出される「トヨタ」「中間連結決算」「過去最高」「住友不動産」「下方修正」などのキーワードを文脈を無視して扱ってしまうと，どの銘柄が中間連結決算で過去最高だったのかという情報が失われてしまう．

このような問題を回避する簡単な方法は，関心のある銘柄名やその銘柄に関係するキーワードによって記事をフィルタリングし，無関係と思われる記事を除去することである．しかしながら，銘柄ごとにフィルタリングのルールを作成したり，特徴量抽出を行ったりすることは手間がかかる．また，ある記事がどの銘柄の株価に影響を与えるかは必ずしも明らかではない．

そこで本論文では，複数記事の情報を銘柄に依存しない内部表現に圧縮し，これを株価動向推定に用いる．より具体的には，教師なし学習の 1 つである Restricted Boltzmann Machine（RBM）を用いて個々の記事から特徴量を抽出し，それらを複数記事で統合する．そして，統合された圧縮表現を入力として Deep Belief Network（DBN）を用いることで株価動向の予測を行う．

連絡先: 関和広，神戸大学，神戸市灘区六甲台町 1-1，078-803-6480，seki@cs.kobe-u.ac.jp

2. 関連研究

Lavrenko ら [Lavrenko 00] は，金融ニュース記事を用いて株価の変動の予測を行った．まず株価を区分的線形回帰によって平滑化し，平滑化後の各区間をトレンドとして定義する．次に，区間の長さや傾きなどの素性を基にトレンドをクラスタリングする．同時に，各トレンドの発生する 5 時間前までのニュース記事がトレンドに繋がるニュース記事であると定義し， Yahoo Finance から取得した 127 銘柄のタグの付いたニュース記事と結びつける．テスト期間に新たなニュース記事が出現すると，記事の BoW からベイズの定理を用いて近い将来に各トレンドが発生する確率を求め，予測を行う．予測したトレンドを基に各銘柄の売買を行うシミュレーションにより，投資利益を得ることができることを示した．

Schumaker ら [Schumaker 09] は，ニュース記事に対してあらかじめ準備しておいた語彙が含まれるかどうかを記事ごとに集計した BoW を素性とし，各記事に関して SVR による記事発行 20 分後の株価動向の推定を行った．また，Hagenau ら [Hagenau 12] は，DGAP，EuroAdhoc と呼ばれるドイツとイギリスの企業報告書データを用いて，当日の株価の始値と終値の差分の正負の予測を行った．素性には連続する 2 単語を表すバイグラムや同一ウィンドウ内での 2 単語の組み合わせをが用いられ，各銘柄の株価に関してカイ 2 乗統計量によって素性選択を行った．

日本語のテキスト情報を用いた経済指標の予測に関する研究も進んでおり，和泉ら [Izumi 10] は，日銀月報を分析し抽出した特徴量を説明変数，国債価格を被説明変数として回帰モデルを構築し，長期国債の価格の予測を試みた．なお，この研究で利用されているテキストマイニング手法は，KeyGraph と主成分分析である．KeyGraph は共起に基づいて重要な単語を抽出する手法であり，日銀月報から重要単語を抽出するために利用されている．次に，抽出した単語を元にして主成分分析を行い，この主成分を日銀月報の特徴量としている．利用さ

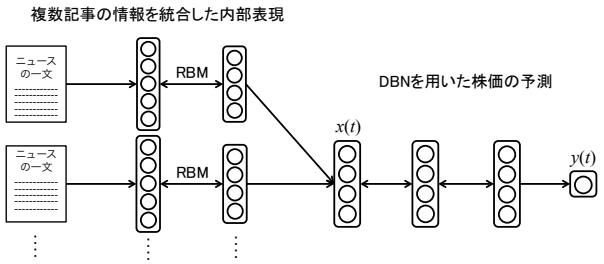

図 1: 提案手法の概念図.

れた主成分の数は累積寄与率 0.6 を基準に選択され，結果として 30 次元の特徴ベクトルとなっている．実験では，日銀月報内の名詞，動詞，形容詞の出現数を説明変数とした SVR や従来から国債価格予測に利用されている経済モデルとの比較の結果，提案手法の誤差が最も低くなった．このことから，提案手法の有効性や，国債価格のような市場情報の予測に対してテキスト情報が有効である事を示している．

辻ら [辻 13] は，日本の全国紙の新聞記事を基に，SVM と SVR を用いて株価を予測する研究を行った．新聞記事から対象企業名と日経シソーラス用語の両方が出現する記事を抽出し，それらの一致語の有無，構文解析によって得られた企業名の係り受け語等を素性とした．実験では，株価のトレンドと変動率を予測する実験を別々に行い，トレンド予測に関しては企業名と単語の生起情報の組み合わせの時に最大の 58.89%となり，ベースラインと比べて正答率が 5 ポイント程度上昇した．

3. 提案手法

3.1 概要

テキスト情報を用いて株価動向を推定する際には，テキスト中に出現する語彙を BoW モデル等でベクトル表現することが一般的である．BoW モデルでは，1 つの文書 d は形態素解析によって切り出された単語によって，単語ベクトル $\{w_1, w_2, \cdots, w_M\}$ ($w \in \mathbb{R}$) で表現される．

Schumaker ら [Schumaker 09] のように，単一の記事を基に株価動向の推定を行う場合には，記事ごとに素性を作成することが可能である．しかし，日次の株価動向を推定する際には，一定期間の（例えば推定の前日の）複数記事から 1 つの単語ベクトルを生成する．しかし，それら複数記事には様々な銘柄に関する情報が混在しているため，得られる単語ベクトルも様々な語が混在した不適切な表現になってしまう．

この問題を解決するため，図 1 に示す手法を提案する．まず，各記事中の各文に関して個別にその圧縮表現を獲得し，得られた複数の圧縮表現を 平均プーリング によって統合する．これにより,「トヨタの中間連結決算が過去最高」と「住友不動産がマンション供給を下方修正」のような異なるイベントを個別に表現可能となることが期待できる．次に，このようにして得られた複数新聞記事の圧縮表現を素性とし，深層学習のアルゴリズムの一つである Deep Belief Network（DBN）を用いることで，株価変動の動向を推定する．以下の節で，提案手法の詳細を述べる．

3.2 複数記事に対する文単位での圧縮表現の獲得

最初のステップでは，複数の新聞記事中で報道されている個々のイベントを統合して表現するベクトル（図 1 の $x(t)$）の獲得を行う．まず，個々の新聞記事の文に関し，あらかじめ準備した辞書に含まれる単語だけを抽出し，単語ベクトルを生成する.（辞書の作成については以降で述べる.）そして，得られたベクトルを入力として，教師なし学習アルゴリズムである Restricted Boltzmann Machine（RBM）を用いて特徴抽出を行う．この学習を行うと，入力データの共通因子を捉えるようにパラメータが更新されるため，隠れ層 $\boldsymbol{h}$ の出力は，入力された新聞記事（文）に関する圧縮表現と見なすことができる．

このように単語ベクトルを文ごとに生成し，これらを平均プーリングすることで，前出のトヨタと住友不動産のような 2 つの記事があった場合も,「トヨタ ⋀ 中間連結決算 ⋀ 過去最高」というイベントが存在し,「トヨタ ⋀ 中間連結決算 ⋀ 下方修正」というイベントは存在しないということが表現できる．なお，ベクトルの統合に最大プーリングを利用した場合，一日に発行された記事の中にどの程度「トヨタ ⋀ 中間連結決算 ⋀ 過去最高」のような記述があったのかを考慮することができないため，本研究では平均プーリングを採用する．

なお，予備実験において隠れ層の各ユニットの出力値にはばらつきがあった（例えば「ある，ない，する，こと，できる，ため」のような高頻度語に対して重みの大きなユニットが存在していたが，このユニットは他のユニットと比較して大きな値を出力しやすい.）このようなユニットの値は重要ではないため，各ユニットの出力値を訓練期間における各々の平均値によって正規化することで重みを調整する．

3.3 Deep Belief Networks を用いた株価動向推定

次のステップでは，実際に新聞記事の内容を受けて株価がどのように変動するのかを推定する．学習のアルゴリズムには Deep Belief Network（DBN）を利用し，素性には最初のステップで得られたベクトル $x(t)$ を用いる．このベクトルは連続値で表現されているため，通常の RBM の更新式を用いると不具合が生じる．そのため，可視層が連続値をとっても問題のないように修正した Gaussian Bernoulli RBM [Hinton 06] を用いて事前学習を行う．式 (1) に Gaussian Bernoulli RBM におけるエネルギー関数 $E(\boldsymbol{v}, \boldsymbol{h}; \theta)$ を示す．

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = \frac{1}{2}\sum_i v_i^2 - \sum_i b_i v_i - \sum_j c_j h_j - \sum_i \sum_j v_i W_{ij} h_j \quad (1)$$

ここで，$\boldsymbol{v}$ は可視層から隠れ層への入力，$\boldsymbol{h}$ は隠れ層から可視層への入力，$\boldsymbol{b}, \boldsymbol{c}$ はバイアス項を表す．この場合，条件付き確率は式 (2)，式 (3) で与えられる．ここで，$N(x; \mu, \sigma^2)$ はガウス関数である．

$$p(v_i|\boldsymbol{h}) = N(v_i; b_i + \sum_j W_{ij} h_j, 1) \quad (2)$$

$$p(h_j = 1|\boldsymbol{v}) = \sigma(c_j + \sum_i v_i W_{ij}) \quad (3)$$

また，性能と速度の改善のため，通常の DBN に以下の手法を用いた．

- Dropout による過学習の緩和 [Hinton 12]
- Momentum を考慮した勾配法の収束速度の向上 [Polyak 64]
- 活性化関数 Rectified Linear Units の利用による計算速度の改善 [Nair 10]

4. 実験

4.1 実験設定

テキスト情報としては日本経済新聞の本紙朝刊を用い，1999 年から 2006 年までの 6 年間の 635,886 記事を訓練データ，2005 年から 2006 年までの 2 年間の 198,996 記事を検証データ，2007 年から 2008 年までの 2 年間の 198,395 記事をテストデータとした．株価動向推定の対象とした銘柄は，日経平均に採用されている 225 銘柄のうち，1999 年から 2008 年までの 10 年間に欠損データの無かった 164 銘柄，および日経平均株価を用いた．

実験結果の評価には，新聞記事の発行された日の MACD（Moving Average Convergence Diverence）と翌日の MACD に関する株価動向適合率（Up Down Correct Rate; UDCR）を用いた．MACD は，EMA・指数平滑移動平均を使用した株取引のテクニカル指標である．ここで，時系列の平滑化を行ったのは株価の微小な変化を無視するためであり，Lavrenko ら [Lavrenko 00] の用いた区分的線形怪奇による平滑化よりも一般的な手法として MACD を選んだ．

入力として用いる語彙は，新聞記事に形態素解析を行うことで獲得した．ここで形態素とは，文章や言葉を，最小単位の語の連なりに分割したときに得られる 各々の要素のことである．形態素解析器には MeCab [工藤 04] を用い，形態素辞書には Wikipedia の見出し語・日経新聞キーワードを追加した．また，計算時間の都合上，モデルへの入力として用いる語彙数（辞書サイズ）を 10,000 語に制限した．これらの語は，ストップワードを除いた後，日経平均に採用されている 225 銘柄のそれぞれに関して各語と株価動向についてカイ二乗統計量を算出した結果，値の大きかった上位の語である．

4.2 圧縮表現の獲得に関する評価

まず，複数の新聞記事中で報道されている個々のイベントを統合して表現するベクトルの獲得に関しての評価を行った．ニューラルネットワークにおける隠れ層の出力値は，入力のベクトルの重み付き和であるため，各ユニットの出力は，重みが大きくなっている入力の単語が（多く）出現しているかどうかを表現していると解釈できる．そのため，各ユニットに対して重みの大きい語を調べることで，圧縮アルゴリズム（RBM）が意図したように機能しているかどうかを評価できる．

表 1 は，隠れ層のサイズを 1,000 とした際に獲得できたコンセプトの一例を示している．右側の列はそのユニットにおいて重みが大きかった語を，左側の列は，それらの語から著者らが解釈したコンセプトの意味を表すラベルである．獲得できたコンセプトの中には，各銘柄の株価動向に対して影響を与えるであろう各銘柄の業績に関するコンセプトが発見された．具体的には,「伊藤忠商事・最終損益・赤字」「NTT・経常損益・訂正」「キリン・地震・特別損失」などである．このようなコンセプトを表現しているユニットが 1 に近い値を出力する時，各銘柄の株価には負の影響を与えることが予想される．

また,「日産自動車・ヤマハ発動機」「日本たばこ産業・JT・NTT」などを表現するユニットは，各銘柄と業種の近い銘柄であるものと推測される．日産自動車，ヤマハ発動機は売上構成の海外比率の高い輸出関連銘柄として知られ，円安時に業績が上がって株価も上昇する傾向がある．JT と NTT は業界は異なるものの，共に不景気時や投資家が投資対象を選べない状況にある際に安定を求めて買われやすいディフェンシブ銘柄として知られている．そのため，一方の業績に関する記事は他方の業績予想に対して有益となることが考えられる．

さらに,「ドル高・円安」「地震・被害」などを表現するユニットは，日本経済全体に影響を与えるコンセプトを捉えているものと考えられる．輸出関連銘柄は円安時に業績が上がり，株価も上昇傾向になり，地震などの災害時には日本全体の株価が下落傾向になることが多い．そのため，このような日本経済全体に影響を与える記事の有無は重要な情報であると考えられる．

表 1: RBM で獲得されたコンセプトの例.

意味（解釈）	語
業績	インター，最終損益，赤字，伊藤忠商事，参照，マイカル，訂正，経常損益，民事再生法，小学校，NTT，地震，キリン，ニューフェース，特別損失，戸建て，ビール
関連銘柄	止め，日産自動車，日産，住所，指揮，歯，ヤマハ発動機．日本たばこ産業，JT，止め，NTT，たばこ
日本経済	情報，センター，ドル高，気象庁，経常益，抄，円安，地震，沖，被害，JT，ケーブル
高頻度語	ない，する，こと，ある，できる，月，する，ある，日本，東京，ため

表 2: SVM と DBM による株価トレンドの予測結果.

モデル	UDCR (%)	向上数	低下数
SVM	58.13	–	–
DBN	59.96 (+1.83)	157	8

4.3 株価動向推定に関する評価

本節では，深層学習を導入することの有効性，および文ごとに素性を作成して統合することの有効性の 2 点を検証した．前者については，分類問題でよく用いられる SVM（support vector machine）を比較手法とした．入力としては,（SVM と DBN の両方で）日ごとの記事をすべて BoW モデルでまとめた単語ベクトルを利用した．なお，SVM と DBN のパラメータはグリッドサーチにより銘柄ごとに有効なパラメータを決定した．SVM では，カーネルの選択（線形あるいは RBF），γ $(1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4})$，$C$ $(1, 10, 10^2, 10^3, 10^4)$ に関して，DBN ではファインチューニング時の学習率 (1, 0.1, 0.01)，バッチサイズ (30, 50, 100)，エポック数 (100, 200, 300, 400) に関して最適な設定を探索した．

4.3.1 SVM と DBN の比較

SVM と DBN で同一の素性（一日分の全記事をまとめた BoW）を用いて株価動向の予測を行った．165 銘柄の予測結果の平均値を表 2 に示す．DBN を用いて各銘柄の株価動向予測を行った場合，SVM と比較して UDCR で平均 1.83 ポイント精度を向上させることができ，t 検定で有意な差が見られた（$p < 0.01$）．また，各銘柄の予測結果の精度の上位と下位 5 銘柄について表 5 に示す．有意水準 0.01 において有意差が確認できたものに上添字 ▲，有意水準 0.05 において有意差が確認できたものに上添字 △ で示す．アドバンテストにおいて 9.80 ポイントの予測精度向上を実現し，14 銘柄に対して $p = 0.05$ での有意差を確認することができた．また，有意に精度が低下した銘柄は存在せず，165 銘柄中 8 銘柄において精度が低下したものの，157 銘柄において精度向上を実現した．

4.3.2 文ごとに獲得した圧縮表現を用いた株価動向推定

文ごとに圧縮表現を獲得・統合する提案手法の有効性を評価するため，一日分の全記事を BoW でまとめる方法（表 2 の DBN）と提案手法を比較した．以降では，前者を「DBN（日ごと）」，後者を「DBN（記事ごと）」と呼ぶ．両手法で株価動向推定を行った結果を表 4 に示す．提案手法の DBN（記事ご

表 3: SVM と DBN の比較において予測精度（UDCR）の向上率と下落率が高い銘柄.

銘柄	変化率	銘柄	変化率
アドバンテスト	+9.80▲	関西電力	−0.48
古河電気工業	+7.59▲	旭化成	−0.13
デンソー	+7.31▲	川崎重工業	−0.13
JT	+6.43△	中部電力	−0.11
武田薬品工業	+5.53△	東京電力	−0.09

表 4: 複数記事の表現の違いによる株価トレンドの予測結果.

モデル	UDCR (%)	向上数	低下数
DBN（日ごと）	59.96	–	–
DBN（記事ごと）	60.17 (+0.21)	58	42

と）は DBN（日ごと）と比較して，UDCR で平均 0.21 ポイント精度が向上し，t 検定で有意な差が確認できた（$p < 0.01$）．ここで，前節と同様に各銘柄の予測結果の精度の上位と下位 5 銘柄の詳細を表 5 に示す．ここで，$p = 0.05$ で有意差が確認できたものに上添字 △ を記載する．表 5 に示すように，ホンダに関して 5.00 ポイント予測精度が向上し，計 2 銘柄に関して有意水準 0.05 で有意な差を確認することができた．また，有意に精度が低下した銘柄は存在せず，165 銘柄中 42 銘柄において精度が低下したものの，58 銘柄において精度向上を実現した．

5. まとめ

本論文では，新聞記事から深層学習によって必要な情報を抽出・統合し，株価動向推定を行う手法を提案した．本手法では，複数の新聞記事中で報道されている個々のイベントを自動的に抽出・統合することにより，推定対象銘柄が変わった際の素性エンジニアリングや学習のコスト軽減した．評価実験では，深層学習を用いて 2007 年〜2008 年の株価動向を 165 銘柄に対して予測し，従来研究の多くで用いられている SVM との比較で，平均して 1.83 ポイントの精度向上を実現した．また，RBM を用いて文ごとに圧縮し，平均プーリングを用いて日ごとに統合する枠組を提案した．これにより，さらに平均 0.21 ポイントの精度向上が確認できた．

今後は，株価という時系列データの特性を活かし，時間的な特性を考慮したモデルの構築を検討していく．

参考文献

[Hagenau 12] Hagenau, M., Liebmann, M., Hedwig, M., and Neumann, D.: Automated news reading: stock price prediction based on financial news using context-specific features, in *Proceedings of the 45th Hawaii International Conference on System Science (HICSS)*, pp. 1040–1049 (2012)

[Hinton 06] Hinton, G. E. and Salakhutdinov, R. R.: Reducing the dimensionality of data with neural networks, *Science*, Vol. 313, No. 5786, pp. 504–507 (2006)

[Hinton 12] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R.: Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint arXiv:1207.0580* (2012)

[Izumi 10] Izumi, K., Goto, T., and Matsui, T.: Trading tests of long-term market forecast by text mining, in *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 935–942 (2010)

[Lavrenko 00] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., and Allan, J.: Language models for financial news recommendation, in *Proceedings of the 9th international conference on Information and knowledge management*, pp. 389–396 (2000)

[Nair 10] Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, in *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 807–814 (2010)

[Polyak 64] Polyak, B. T.: Some methods of speeding up the convergence of iteration methods, *USSR Computational Mathematics and Mathematical Physics*, Vol. 4, No. 5, pp. 1–17 (1964)

[Schumaker 09] Schumaker, R. P. and Chen, H.: Textual analysis of stock market prediction using breaking financial news: the AZFin text system, *ACM Transactions on Information Systems*, Vol. 27, No. 2, p. Article No. 12 (2009)

[工藤 04] 工藤 拓, 山本 薫, 松本 裕治：Conditional random fields を用いた日本語形態素解析, 情報処理学会自然言語処理研究会 SIGNL-161, pp. 89–96 (2004)

[辻 13] 辻 洋平, 古宮 嘉那子, 小谷 善行：Web ニュース中の複数企業に対応した株価予測, 電子情報通信学会技術研究報告 IBISML, pp. 109–113 (2013)

表 5: 日ごと・記事ごとの DBN の比較において予測精度（UDCR）の向上率と下落率が高い銘柄.

銘柄	変化率	銘柄	変化率
ホンダ	+5.00△	コナミ	−2.14
NTT ドコモ	+4.22△	デンソー	−1.75
NTT データ	+3.11	信越化学工業	−1.56
日経平均株価	+2.89	協和発酵キリン	−1.56
イオン	+2.67	中外製薬	−1.56