深層学習による複数文書の圧縮表現の獲得と株価動向推定への応用

Generating Uni ed Representation for Multi-Documents by Deep Neural Network and its Application to Stock Price Prediction

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

[1] 神戸大学大学院システム情報学研究科
[1] Graduate School of System Informatics, Kobe University

Abstract: 本論文では，新聞記事で報道される情報をもとに，深層学習による株価動向推定の手法を提案する．一日の新聞記事中には，様々な銘柄に直接的・間接的に関係するニュース記事が混在しており，それらに出現する語彙を全て同等に扱うのは適切でない．そこで，記事ごとに作成した特徴ベクトルを深層学習によって統合することで，一日の出来事に関する圧縮表現を獲得する手法を提案する．さらに，このように獲得した特徴ベクトルを用いて，数種の銘柄の株価動向予測を行った結果について報告する．

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の他に，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，株式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，主に二種類に分類することができる。一方は企業の株価や物価指数，政策金利といった数値情報，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．

これら数多くの情報は日々発信されているものの，投資家たちがこの膨大な数の情報全てに目を通し，市場分析に利用することは容易ではない．そこで，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．例えば，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である．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全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を独立に扱い，bag-of-words モデルで表現することが多い．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 日の日本経済新聞朝刊では,「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る．

このような問題を回避する方法として，扱うニュース記事を銘柄名や銘柄の属するカテゴリー内の銘柄名のキーワードによってフィルタリングすることで，無関係の記事を除去することが考えられる．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，あるニュース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，新聞記事をテキスト情報として利用しつつ，記事ごとに特徴抽出を行って株価動向推定を行う手法を提案する．特徴抽出には教師なし学習である Restricted Boltzmann Machine（RBM）を利用し，得られた隠れ層の出力が新聞記事の圧縮表現であると考える．この方法で，同日に発表された各記事ごとに圧縮表現を獲得し，それらの max-pooling を行った結果を一日の出来事の圧縮表現として考える．これを

連絡先：神戸大学システム情報学研究科
〒 657-8501 神戸市灘区六甲台町 1-1
E-mail: fujikawa@ai.cs.kobe-u.ac.jp

以降の教師あり学習の入力として利用することで，株価動向の推定を行う．

2 関連研究

本章では，本研究に関連する従来研究について説明する．まず深層学習と自然言語処理ドメインへの応用に関する研究の例を挙げる．次にテキスト情報を用いて株価動向の予測を行った研究の例を挙げ，問題点をまとめる．

2.1 深層学習と自然言語処理への応用

近年，事前学習でデータから抽象度の高い内部表現を獲得させて，多層ニューラルネットワークの精度を向上させる深層学習を用いた手法が，特に音声認識や画像認識などの分野において高い成果を上げたことで注目を集めている [1, 2, 3]．

自然言語処理ドメインにおいては Collobert ら [4] が深層学習による統一的なフレームワークを用いて，自然言語処理の一般的なタスクである品詞のタグ付け，チャンキング，固有表現抽出，意味役割付与などを行った．より具体的には，まず単語から先頭が大文字かどうか等の情報で構成された素性を抽出し，畳み込みによって周辺文脈を考慮したベクトル表現を獲得する．これによって得られたベクトルに関して文単位で max-pooling を行うことにより，必要な情報を選択して定次元ベクトルで表現することができる．そして，得られたベクトルを用いた多層ニューラルネットワークを複数のタスクに適用した．実験では，複数タスクを同時に学習させることにより一般的な性能を向上させることを示した．

2.2 テキスト情報を用いた株価動向予測

Schumaker ら [5] 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 bag-of-words 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 [6] 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 2gram や 同一ウィンドウ内での 2 単語の組み合わせを表す 2-word combination が用いられ，各銘柄の株価に関して以下の式で定義されたカイ二乗検定を行ったスコアによって素性選択を行った．

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i \quad E_i)^2}{E_i} \tag{1}$$

ここで，O は観測値，E は期待値を表す．実験では，カイ二乗検定による素性選択のアプローチが精度向上に貢献し，過学習も緩和させられることを示した．しかし，この研究では銘柄ごとに素性選択を行うため，推定対象銘柄が変わった場合には，改めてカイ二乗検定により素性選択をし直す必要がある．また，この研究で用いられているデータは対象銘柄に関する情報であることが保証されているため，記事の内容が対象銘柄に関する内容かどうかを判別することを必要としないものの，一般的な新聞記事を対象とする場合には，記事の内容が対象銘柄に関する内容かどうかを判別する必要がある．

3 提案手法

本章では，新聞記事を用いた株価動向推定の手法について述べる．まず，複数（一日分）の新聞記事に対する圧縮表現を獲得するモデルについて概要を説明する．次に，それらを用いて株価動向推定を行う学習の手法について説明する．

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何らかの方法でベクトル表現することが一般的である．Schumaker ら [5] 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ものの，日次の株価動向を推定する際には，一定期間の（例えば推定の前日に報道された）複数記事から固定長のベクトルを生成する必要がある．その際に，いずれかの記事に特定の語彙が含まれるかどうかだけでベクトル表現すると，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，不適切な表現が得られてしまう場合がある．

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，各記事に関して個別にその圧縮表現を獲得し，そして個別の圧縮表現の max-pooling によって統合する．これにより，図に示すように「トヨタ」の「中間連結決算」は「過去最高」であって,「住友不動産」に関してはその事実がないことを表現することが期待できる．次に，このようにして得られた一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 stacked denoising auto-encoders を用いることで，株価変動の動向を推定する．

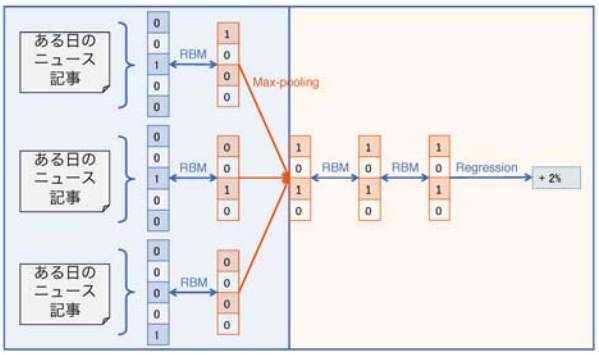

図 1: システム概要.

3.2 複数記事に対する圧縮表現の獲得

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の獲得を行う．まず，個々の新聞記事に関し，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，教師なし学習のアルゴリズムである Restricted Boltzmann Machine (RBM) を用いて特徴抽出を行う．RBM では以下の尤度を最小化するようにパラメータを学習する．

$$p(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z}\exp(\quad E(\boldsymbol{v},\boldsymbol{h})) \tag{2}$$

ただし，Z は正則化項，E はエネルギー関数であり，以下で定義される．

$$Z = \sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}}\exp(\quad E(\boldsymbol{v},\boldsymbol{h})) \tag{3}$$

$$E(\boldsymbol{v},\boldsymbol{h}) = \quad \sum_i b_i v_i \quad \sum_j c_j h_j \quad \sum_i\sum_j v_i W_{ij} h_j \tag{4}$$

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に関する圧縮表現と見なすことができる．図 2, 3 は，新聞記事中の語彙を日ごと・記事ごとに集計したベクトルを入力として学習した場合の隠れ層の出力の一例を示している．破線の重みを 0, 実線の重みを正の数値とすると，前述の例に関するユニット 1 の出力は「トヨタ ⋀ 中間連結決算 ⋀ 過去最高」，ユニット 2 の出力は「トヨタ ⋀ 中間連結決算 ⋀ 下方修正」を表現していると考えられる．記事の内容を考慮すると，ユニット 1 の出力は 1，ユニット 2 の出力は 0 に近い値が望ましい．

この時，図 2 のような単一の単語ベクトルを用いると，ユニット 1・ユニット 2 のいずれの値も 1 となってしまうため，隠れ層の表現が不適切となる．一方で，図 3 のように単語ベクトルを記事ごとに生成すると，ユニット 1・ユニット 2 の値が理想的な出力となる．これらの max-pooling を求めることで，二つの新聞記事中に「トヨタ ⋀ 中間連結決算 ⋀ 過去最高」に関する記述は存在し，「トヨタ ⋀ 中間連結決算 ⋀ 下方修正」に関する記述は存在していないことが表現できる．

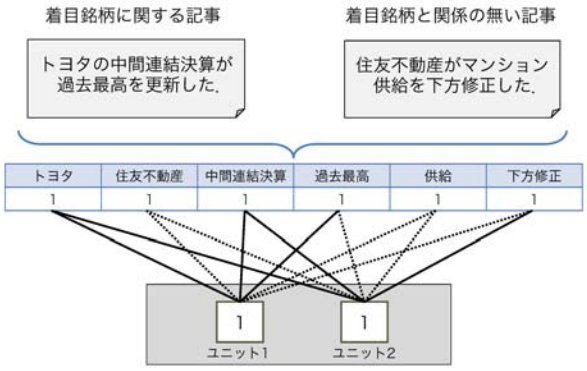

図 2: 複数記事から単一の単語ベクトルを生成した場合の RBM の隠れ層の出力例.

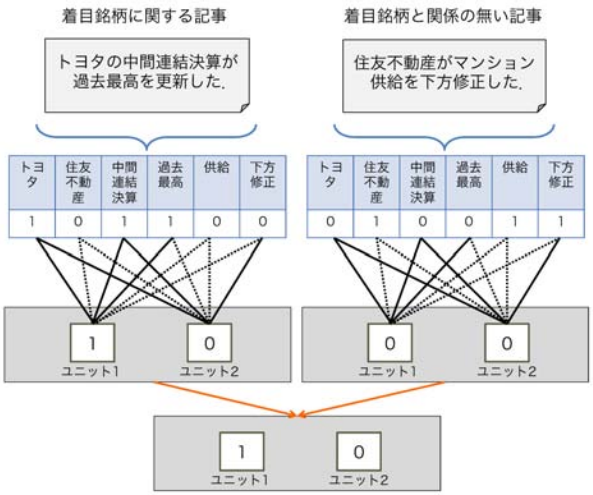

図 3: 記事ごとに単語ベクトルを生成した場合の RBM の隠れ層の出力例.

3.3 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 stacked denoising auto-encoders (SdA) を利用し，素性には最初のステップで得られたベクトルを用いる．

SdA は多層ニューラルネットワークの事前学習に denoising auto-encoder を用いたもので，層の数を増加さ

せた際にパラメータが拡散する課題を解決している [7]. 事前学習では入力データとして $\boldsymbol{x}$ の一部を損傷させたデータ $\tilde{\boldsymbol{x}}$ を用い，次式で定義される符号化・復号化を通して得られた $\boldsymbol{z}$ が元の入力データ $\boldsymbol{x}$ に近い値を出力するように訓練を行う.

$$\boldsymbol{y} = s(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}) \tag{5}$$

$$\boldsymbol{z} = s(\boldsymbol{W'}\boldsymbol{y} + \boldsymbol{b'}) \tag{6}$$

ただし，$s(\cdot)$ はシグモイド関数であり，$\boldsymbol{W'} = \boldsymbol{W^T}$ とする．学習では，以下のクロスエントロピー誤差関数を用い，確率的勾配降下法によって誤差が最小となるようにパラメータの更新を行う.

$$L(\boldsymbol{x}, \boldsymbol{z}) = \quad \boldsymbol{x} \log \boldsymbol{z} \quad (1 \quad \boldsymbol{x})(1 \quad \log \boldsymbol{z}) \tag{7}$$

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 2004 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，銘柄名を含む記事が存在する日数が最も多い 5 銘柄とした.

実験結果の評価には，新聞記事の発行された日の始値と終値の差に関する株価動向適合率（Up Down Correct Rate; UDCR）を用いた．これにより，報道された新聞記事の内容を受けて株価がどのように動いたのかを評価することができる.

また，計算時間の都合上，入力として用いる語彙数（辞書サイズ）を 5000 語とした．これらの語は，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検定を行い，スコアの高かった上位 5000 語である.

4.2 評価実験

本手法により株価動向を推定し，UDCR により評価を行った結果を表 1 に示す．表 1 中の Baseline は，テストデータ中の株価上昇・下落を集計し，多い方を常に選択した方法である.

表 1 で示す通り，提案手法が全ての銘柄において Baseline を上回ることができ，平均値において 3.9 ポイントの精度向上を実現した．一方で，全体を通して大きな精度改善は見られなかったため，以下で原因の考察を行った.

原因として考えられることは,「 ne tuning が上手く作用していない」ということである．表 1 で示した結果の多くは epoch の値が 3 より小さい時点で達成しており， ne tuning の学習が進むにつれてエラー率が増加するような状態になっていた．この問題に対する解決策としては，ハイパーパラメータのグリッドサーチによる最適化や，最上層の回帰モデルに時系列データの予測に有効とされる CRF や Recurrent Neural Network を適用させることが考えられる.

表 1: 評価実験の結果.

	提案手法	Baseline
トヨタ自動車	0.427	0.494
ソニー	0.427	0.475
東芝	0.433	0.453
日産自動車	0.435	0.470
日立製作所	0.453	0.477
平均値	0.435	0.474

5 結論

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場合にかかる素性エンジニアリングや学習のコストを軽減した.

評価実験では，同じモデルを用いて素性を抽出し，異なる銘柄の予測に対応できることを示した．一方で， ne tuning による精度改善があまり見られず，予測部分に課題を残した．株価という時系列データの特性を活かし，今後は時間的な文脈情報を考慮したモデルの構築を検討していく.

参考文献

[1] Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H.: Greedy layer-wise training of deep networks, *Advances in Neural Information Processing Systems (NIPS)*, Vol. 19, p. 153 (2007).

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.: Imagenet classi cation with deep convolutional neural networks, *Advances in Neural Information Processing Systems (NIPS) 25*, pp. 1106–1114 (2012).

[3] Dahl, G. E., Yu, D., Deng, L. and Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,

Audio, Speech, and Language Processing, IEEE Transactions on, Vol. 20, No. 1, pp. 30–42 (2012).

[4] Collobert, R. and Weston, J.: A uni ed architecture for natural language processing: Deep neural networks with multitask learning, *Proceedings of the 25th international conference on Machine learning (ICML)*, ACM, pp. 160–167 (2008).

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

[6] Hagenau, M., Liebmann, M., Hedwig, M. and Neumann, D.: Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Speci c Features, *System Science (HICSS), 2012 45th Hawaii International Conference on*, IEEE, pp. 1040–1049 (2012).

[7] Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th international conference on Machine learning (ICML)*, ACM, pp. 1096–1103 (2008).