Vol.2015-MPS-102 No.4
2015/3/3

ニュース記事の時間的特性を考慮した株価動向予測

吉原 輝[1]　関 和広[2]　上原 邦昭[1]

概要：投資家が投資を行う際，株価等の数値情報の他に，新聞記事等の言語情報を基に株の売買を判断する．この判断を支援するため，これまで様々な研究が行われており，数値情報を対象にした研究では，株価の時系列データの特性が多く利用されている．これに対し，言語情報を対象にした研究では，その特性がほとんど利用されていない．これは，言語情報が株価に与える影響の時間的な変化を人手でルール化することが困難だからである．一方で，画像認識や音声認識などの分野において近年注目を集めている深層学習（Deep Learning）は，大規模なデータから有益な特徴の抽出が可能である．そこで本研究では，深層学習のアプローチを応用し，時間的な変化を考慮した再帰的なネットワークを構築することで株価動向の推定を行う手法を提案する．入力に新聞記事のデータを用いることで，言語情報が与える影響の時間的な変化を捉えることができる．実際の新聞記事と株価のデータを用いて 10 銘柄の株価動向推定を行い，本手法の有効性を示す．

キーワード：再帰型ニューラルネットワーク，深層学習，自然言語処理，金融テキストマイニング

Yoshihara Akira[1]　Seki Kazuhiro[2]　Uehara Kuniaki[1]

1. はじめに

近年，機械学習の手法を応用し，株価を含む膨大な金融情報を分析することで，投資家の判断を支援する技術が注目されている．投資を行う際，投資家は新聞記事やマイクロブログなどの言語情報から表出する有益な情報を基に，市場の動きを分析・予測する．しかし，発信される全ての言語情報を投資家自身が全て分析することは困難である．そこで，Lavrenko ら [12], [13] や Schumaker ら [15] の研究など，これらの言語情報を自動的に解析することで株価の動きを予測する試みが盛んに行われてきた．

一方，深層学習を用いた手法が，自然言語処理や画像認識などの分野において高い精度を上げており，近年注目を浴びている [3], [10]．例えば，Socher ら [18] は，深層学習の手法の一つであるオートエンコーダを拡張したモデルを用いることで，文章の感情予測での精度の高さを示している．

このことから，株価動向推定に関しても，深層学習を言語情報の解析に利用することにより，精度が向上する可能性があると考えられる．しかし，深層学習を用いた手法のうち，言語情報に対する手法の大半が時間的な変化を考慮していない．株価は，様々な事象によって時々刻々と変動する．また，ある株価に影響を与える事象が生じ，それが言語情報として現れたとき，その情報は，長期に渡って株価に影響を与え得る．例えば，2008 年 9 月 15 日にリーマン・ショックが起きたとき，多くの株価が 10 月下旬まで下落している．同日，日本経済新聞朝刊にリーマン・ショックに関する記事が複数あり，これらの記事を考慮すれば，長期に渡る株価の下落を推定できる可能性がある．従って，時間的な変化を含む情報に対しては，それらを考慮したモデルを構築することが必要であると考えられる．

本研究では，Recurrent Neural Networks-Restricted Boltzmann Machine (RNN-RBM) [2] を用いて，言語情報が株価に与える影響の時間的な変化を捉え，深層学習の枠組みで株価動向推定を行う手法を提案する．

本論文の構成は以下の通りである．まず，2 章で深層学習の一般的なモデルと時系列データを扱うモデルの構造を説明する．3 章では，RNN-RBM の概要と本手法の構造を述べ，4 章では，日本経済新聞朝刊を用いて複数銘柄の株価動向推定についての評価実験を行い，その結果を考察する．最後に，5 章で本論文のまとめと今後の課題を述べる．

[1] 神戸大学大学院 システム情報学研究科
Graduate School of System Informatics, Kobe University
[2] 甲南大学 知能情報学部
Faculty of Intelligence and Informatics, Konan University

2. 関連研究

本章では，まず株価動向推定に関連する研究を述べ，次に一般的な深層学習のモデルについて説明する．その後，時系列データを扱う深層学習のモデルの構造について述べる．

2.1 言語情報を用いた株価動向推定

これまで，数値情報を用いた株価動向推定が数多く行われており，その研究には Support Vector Machine (SVM) 等の機械学習の手法がよく用いられている [7], [9], [16], [21]. Tay ら [21] は，5 日前との相対的な株価の変動率を SVM の入力変数として用い，評価実験でニューラルネットワークを上回る結果を示した．しかし，金融危機等の急激な株価の下落を伴うイベントが生じたとき，このようなモデルは株価の動きを正しく予測することができない．このようなイベントは，過去の数値情報に表出しないためである．

一方，言語情報を用いた手法であれば，報道されるイベントを考慮することができる [5], [8], [12], [13], [14], [15]. 例えば金融危機が生じたとき，言語情報には「金融危機」など，株価の下落を示唆する言葉が表出する．つまり，このような言葉を考慮して予測することにより，より良い予測が得られると考えられる．Lavrenko ら [13] は，金融のニュース記事と株価のトレンドを結びつけ，記事の bag-of-words からベイズの定理を用いて近い将来に発生するトレンドを予測することで，株価の変動を予測した．実験では予測したトレンドを基に実際に各銘柄の売買を行うシミュレーションを行い，利益を得られることを示した．しかし，この研究では，言語情報から有益な情報を抽出することが課題となっている．この問題を解決するように，Ding らは [4] は，Open IE（Information Extraction）手法を用いて，大規模なニュース記事から記事本文の主述関係を保持した状態でイベントを抽出し，それらを特徴量として，S&P 500 の株価を予測した．また，Ding らは予測器に深層学習のモデルを用いており，評価実験では SVM を上回る精度を示した．しかし，彼らを含む言語情報を用いた多くの研究では、株価が時系列データであるにも関わらず，そのような特性を考慮していない．

そこで，本研究では再帰的なネットワークを持つ深層学習のモデルを利用する．深層学習は自動的に有益な情報を抽出する事が可能であり，また，再帰的なネットワークを利用することで，時系列データが持つ特性を考慮することが可能になる．

2.2 一般的な深層学習モデル

2.2.1 Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBM) [17] は，確率的深層学習モデルのひとつである．RBM では，可視層 $\vec{v}$ と隠れ層 $\vec{h}$ の結合分布をエネルギー関数 E を用いて定義する．

$$P(\vec{v},\vec{h}) = e^{-E(\vec{v},\vec{h})}/Z \quad (1)$$

$$E(\vec{v},\vec{h}) = -\vec{b_v}^T\vec{v} - \vec{b_h}^T\vec{h} - \vec{h}^T W \vec{v} \quad (2)$$

ここで，$\vec{b_v}$，$\vec{b_h}$，W はモデルのパラメータであり，それぞれ，可視層のバイアス項，隠れ層のバイアス項，重み行列を表す．Z は正規化項である．また，$\vec{v}$ が与えられたときの隠れ層のノードの状態 $h_i \in \{0,1\}$，$\vec{h}$ が与えられたときの可視層のノードの状態 $v_j \in \{0,1\}$ はそれぞれ以下のように計算される．

$$P(h_i = 1|\vec{v}) = \sigma(b_{h_i} + \vec{W}_i\vec{v}) \quad (3)$$

$$P(v_j = 1|\vec{h}) = \sigma(b_{v_j} + \vec{W}_j^T\vec{h}) \quad (4)$$

ただし，$\sigma(\cdot)$ はシグモイド関数を表す．$\vec{v}$ の周辺確率 $P(\vec{v})$ は，free-energy $F(\vec{v})$ を用いて，

$$P(\vec{v}) = e^{-F(\vec{v})}/Z \quad (5)$$

$$F(\vec{v}) = -\vec{b_v}^T\vec{v} - \sum_i \log(1 + e^{(b_{h_i} + \vec{W}_i v)}) \quad (6)$$

と表される．この周辺確率 $P(\vec{v})$ の対数尤度を最大化することにより，パラメータの学習を行う．

2.2.2 Deep Belief Networks

Deep Belief Networks (DBN) [6] は，RBM を階層的に積み上げて構成される深層学習モデルの一種である．DBN における可視層と隠れ層の結合分布は，

$$P(\vec{v},\vec{h}^1,\ldots,\vec{h}^l) = \left(\prod_{k=0}^{l-2} P(\vec{h}^k|\vec{h}^{k+1})\right) P(\vec{h}^{l-1},\vec{h}^l) \quad (7)$$

で表される．ここで，$\vec{v} = \vec{h}^0$ とし，$P(\vec{h}^{k-1}|\vec{h}^k)$ は，第 k 層目の RBM の条件付き確率を表す．RBM と違い，DBN は隠れ層同士で結合しているため，異なる表現能力を持つことが期待される．DBN のパラメータの学習は，RBM を利用した事前学習（pre-training）とファインチューニングに分けられる [1].

2.3 時系列データを扱う深層学習モデル

時系列データを扱う場合，そのデータが持つ時間的特性を考慮できるようなモデルを利用することがある．Temporal Restricted Boltzmann Machine (TRBM) [19] や Recurrent Temporal Restricted Boltzmann Machine (RTRBM) [20] などがそのようなモデルとして挙げられる．RTRBM では，モデルを再帰的に構築し，時刻が 1 つ前の隠れ層との結合を考えることで，ある時刻 t の隠れ層 $\vec{h}_t$ は，可視層 $\vec{v}_t$ の隠れ表現の他に，t 以前の隠れ層の時間的な変化の表現も可能となる．時刻 t において，$\vec{h}_{t-1}$ が与えられたときの $\vec{v}_t$ と $\vec{h}_t$ の条件付き結合分布は，次のように表される．

$$P(\vec{v}_t,\vec{h}_t|\vec{h}_{t-1}) = \frac{\exp\left(\vec{v}_t^T\vec{b_v} + \vec{h}_t^T W \vec{v} + \vec{h}_t^T(\vec{b_h} + W'\vec{h}_{t-1})\right)}{Z(\vec{h}_{t-1})} \quad (8)$$

ここで，$\vec{b_v}$，$\vec{b_h}$，W は式 (2) と同様であり，$\vec{W}'$ は，$\vec{h}_{t-1}$ から $\vec{h}_t$ への重み行列である．この式を用いて，RTRBM における結合分布 $P(v_1^T, h_1^T)$ は，

$$P(v_1^T, h_1^T) = \prod_{t=1}^{T} \sum_{h'_t} P(\vec{v}_t, \vec{h}'_t | \vec{h}_{t-1}) P(\vec{h}_t | \vec{v}_t, \vec{h}_{t-1})$$
$$= \prod_{t=1}^{T} P(\vec{v}_t | \vec{h}_{t-1}) P(\vec{h}_t | \vec{v}_t, \vec{h}_{t-1}) \quad (9)$$

である．

3. 言語情報の時間的特性を考慮した株価動向予測

本研究では，新聞等の言語情報から時間的に変動する株価の上昇・下落を予測することを目的とする．言語情報が株価に与える影響の時間的な変化を捉えるために，ここでは Recurrent Neural Networks-Restricted Boltzmann Machine (RNN-RBM) を利用する．RNN-RBM は，時系列情報を考慮した深層学習のモデルであり，2.3 節で述べた RTRBM を拡張したモデルである．本章では，3.1 節で，RTRBM の問題点と RNN-RBM の概要について述べ，その後，本手法の概要及び本手法で用いる素性について述べる．

3.1 RNN-RBM の概要

RTRBM では，ある時刻の隠れ層が必ず次の時刻の隠れ層に影響を与えるという制約が生じている．時系列データにおいて，全てのデータが必ずしも未来に影響を与えるとは限らず，時刻によって，その影響は変動すると考えられる．実際に新聞記事においても，株価に影響を与える記事が数多く存在する．例えば，リーマンショックに関する記事が株価に与える影響は長期的である．一方で，株式市場に関する記事が株価に与える影響は短期的と言われており，これらの記事は未来の株価には影響しないと考えられる．もし，RTRBM を株価動向推定に用いると，ある記事が必ず未来の株価に影響するという制約が生まれ，結果として予測を誤る可能性がある．RNN-RBM は，そのような制約を解消するため，RTRBM を更に拡張した手法である．

RNN-RBM の概要を図 1 に示す．RNN-RBM は，パラメータ W，$\vec{b}_h^{(t)}$，$\vec{b}_v^{(t)}$ を持つ RBM と，W'，W''，W_2，W_3，$\vec{\hat{h}}^{(t)}$，$\vec{b}_{\hat{h}}$ を持つ RNN で構成されている．RBM の隠れ層とは異なる隠れ層を加えることで，時間的な変化を表す隠れ表現 $\vec{\hat{h}}$ と観測データの隠れ表現 $\vec{h}$ を区別し，先に述べた制約を解決している．

単純化のために，1 層のみの RNN-RBM を考えると，隠れ表現 $\vec{\hat{h}}^{(t)}$ は次のように表される．

$$\vec{\hat{h}}^{(t)} = \sigma(W_2 \vec{v^{(t)}} + W_3 \vec{\hat{h}}^{(t-1)} + \vec{b}_{\hat{h}}) \quad (10)$$

また，各バイアス項 $\vec{b}_h^{(t)}, \vec{b}_v^{(t)}$ は，

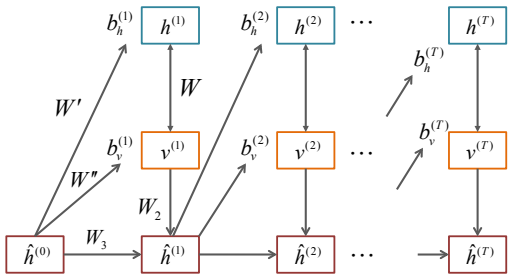

図 1 RNN-RBM のグラフィカルモデル

$$\vec{b}_h^{(t)} = \vec{b}_h + W' \vec{\hat{h}}^{(t-1)} \quad (11)$$
$$\vec{b}_v^{(t)} = \vec{b}_v + W'' \vec{\hat{h}}^{(t-1)} \quad (12)$$

で求められる．これらのパラメータを用いて，このモデルの学習は行われる．学習は 2 つのステップに分かれており，まず CD 法により RBM のパラメータにおける導関数を求め，その後，Backpropagation Through Time (BPTT) アルゴリズム [22] を適用することにより RNN のパラメータにおける導関数を求める．

3.2 DBN への導入

提案手法の概要を図 2 に示す．本手法では，DBN を拡張したモデルを利用する．拡張するモデルに DBN を選択した理由は 2 点あり，まず 1 点目は，DBN が多層構造であることによる．深層学習では，層の数を増やすことによって，表現力が増すと考えられている．2 点目は，RNN-RBM が RBM を拡張したモデルであるからである．DBN が RBM を階層的に積み上げて構成されていることから，その構成要素である RBM を RNN-RBM に拡張することは容易である．このように拡張することで，DBN は多層構造であり時系列情報を考慮したモデルとなる．

本手法は，DBN の層のうち，入力層 $\vec{v}$ と隠れ層 $\vec{h}^1$ からなる RBM の代わりに，RNN-RBM を用いる．このことにより，入力データの時間的な変化を考慮した隠れ表現が得られると考えられる．また本研究は，株価の上昇あるいは下落を推定する 2 値分類問題を扱うため，出力層の次元は 1 となる．

本手法の学習は，DBN と同様，事前学習とファインチューニングに分けて行う．事前学習では，先に述べた RBM と RNN-RBM の学習法により，それぞれのパラメータを求める．ファインチューニングでは，誤差逆伝播法によりそれらのパラメータを更新する．

3.3 本手法で用いる素性

本研究は言語情報が株価に与える影響の時間的な変化を捉えることが重要となるため，入力データは言語情報であり，モデルが扱えるようにこれらをベクトルで表現する必要がある．本手法では，日単位で記事をグループ化し，そ

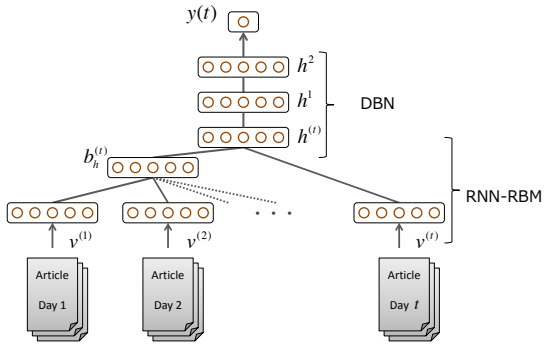

図 2 提案するモデルのグラフィカルモデル

れぞれの記事群を bag-of-words を用いてベクトルで表現する．このベクトルは特徴語を要素とし，それぞれの特徴語が記事群に出現していれば 1，していなければ 0 で表す．

次に，出力の扱いについて述べる．本手法では，日単位での株価の上下を予測する 2 値分類問題を扱う．出力を y としたとき，以下の式に従って予測を行う．

$$y = f\left(\sigma(W^o\vec{h}^2 + \vec{b^o})\right) \tag{14}$$

$$f(x) = \begin{cases} 1 & (x > 0.5) \\ 0 & (x \leq 0.5) \end{cases} \tag{15}$$

ここで，$\vec{h}^2$ は，DBN の最終層での出力を表し，W^o, b^o は，出力層のパラメータである．また，1 は株価の上昇，0 は株価の下落を表す．

4. 評価実験

本論文では，評価実験として 2 通りの実験を行った．最初の実験では，株価の上昇・下落の二値分類を行い，テストデータにおけるエラー率を報告する．

4.1 実験設定

本実験で扱う言語情報は，日本経済新聞の本紙朝刊であり，1999 年から 2008 年までの全 1,033,277 記事を用いた．このうち，1999 年から 2006 年までの 8 年間の 834,882 記事を訓練データ，2007 年の 98,667 記事を検証データ，2008 年の 99,728 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 10 銘柄と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 Moving Average Convergence Divergence [*2] (MACD) 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，直近の株価に重みを付けた EMA・指数平滑移動平均を使用している．平滑化が行われることにより，株価の微小な変化を無視することができる．

それぞれの銘柄の訓練データ，検証データ，テストデータにおける株価の上昇と下落の割合を表 1 に示す．また，各データのインスタンス数は，1,894 件，236 件，236 件である．提案手法や比較手法の学習や評価時に，これらのデータを利用する．また，本実験での提案手法の隠れ層のユニット $\vec{h}^1$, $\vec{h}^2$,$\vec{h}$ の数は，検証データを用いた予備実験により，3750，2500，200 とした．

表 1 データセットの上昇 / 下落の割合

brand	train(1/0)	valid(1/0)	test(1/0)
日経平均	0.51/0.49	0.49/0.51	0.5/0.5
日立製作所	0.39/0.61	0.37/0.63	0.37/0.63
東芝	0.37/0.63	0.42/0.58	0.4/0.6
富士通	0.41/0.59	0.42/0.58	0.42/0.58
シャープ	0.44/0.56	0.45/0.55	0.48/0.52
ソニー	0.5/0.5	0.47/0.53	0.49/0.51
日産自動車	0.4/0.6	0.46/0.54	0.45/0.55
トヨタ自動車	0.48/0.52	0.45/0.55	0.48/0.52
キャノン	0.48/0.52	0.42/0.58	0.47/0.53
三井物産	0.4/0.6	0.48/0.52	0.48/0.52
三菱商事	0.43/0.57	0.44/0.56	0.49/0.51

表 2 カイ二乗検定クロス表

	Uptrend over 1%	Downtrend over 1%	Neutral	Sum
Appear	U_{w+}	D_{w+}	N_{w+}	A_{w+}
Not Appear	U_{w-}	D_{w-}	N_{w-}	A_{w-}
Sum	U	D	N	A

4.2 入力に用いる語彙

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する．形態素解析器として Mecab [11] を用い，Wikipedia の見出し語，および日経新聞キーワードを形態素辞書に追加して用いた．なお，形態素解析で抽出された単語の内，記号や助詞は株価に影響を与えないものとし，無視している．

また，計算時間の関係上，入力として用いる特徴語の数を 5,000 語とした．これらの語は，日経平均に採用されている 225 銘柄の各銘柄に関して各語と株価動向についてのカイ二乗統計量を算出した結果，スコアの高かった上位 5,000 語である．カイ二乗統計量はクロス表（表 2）を用いて算出した．なお，U_{w+}・U_{w-} は上昇トレンド時に，各単語が一日の新聞記事中に存在した・存在しなかった日数，N_{w+}・N_{w-} は下落トレンド時に，各単語が一日の新聞記事中に存在した・存在しなかった日数，A_{w+}・A_{w-} は全てのトレンドに関して各単語が一日の新聞記事中に存在した・存在しなかった日数を示している．

[*2] 移動平均収束拡散手法とも呼ばれ，株価の将来の値動きを予想するテクニカル分析の指標の一つである．

上昇トレンドに対するスコア $\chi^2_{uptrend}$，下落トレンドに対するスコア $\chi^2_{downtrend}$ は次のように算出する．

$$\chi^2_{uptrend} = \frac{(|\Theta_{uptrend}| - \frac{A}{2})^2 \cdot A}{A_{w+} \cdot A_{w-} \cdot U \cdot D} \quad (16)$$

$$\chi^2_{downtrend} = \frac{(|\Theta_{downtrend}| - \frac{A}{2})^2 \cdot A}{A_{w+} \cdot A_{w-} \cdot U \cdot D} \quad (17)$$

$$\Theta_{uptrend}=U_{w+}\cdot(D_{w-}+N_{w-})-U_{w-}\cdot(D_{w+}+N_{w+}) \quad (18)$$

$$\Theta_{downtrend}=D_{w+}\cdot(U_{w-}+N_{w-})-D_{w-}\cdot(U_{w+}+N_{w+}) \quad (19)$$

上位 5000 語を選択する際は，上昇トレンドと下落トレンドに対するスコアの区別はしていない．

4.3 実験結果

本手法により株価動向を推定し，MACD により評価を行った結果を表 3 に示す．表 3 が示すそれぞれの数値は，テストデータにおけるエラー率（%）である．本実験では，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をベースラインとする．比較手法には，ベースラインの他，SVM，時間的な変化を考慮しない DBN を用いた．SVM は線形カーネルを用いており，検証データにおける予備実験で得られた最適なパラメータ値（$C = 0.0001$）を用いた．また，DBN の隠れ層のユニット $\vec{h}^1, \vec{h}^2$ も検証データを用いた予備実験により，2500,1250 とした．ベースライン，SVM との比較で，深層学習を導入することの有効性を検証し，DBN との比較で，時系列情報を考慮する必要性を検証する．なお，ベースラインに示す値は表 1 に示す値と少々異なっている．これは，本手法の学習にミニバッチ学習を採用しており，ミニバッチに分割する際，余ったデータは入力として扱っていないことによる．

ベースラインおよび SVM と比較すると，全ての銘柄において提案手法のエラー率が下回っている．平均値においてもそれぞれ約 7.5 ポイント，約 3.5 ポイントのエラー率減少を実現した．また，DBN と比較すると，11 銘柄中 6 銘柄において提案手法のエラー率が下回り，3 銘柄においては同等のエラー率，平均値においては約 1 ポイントのエラー率減少を達成した．性能向上の有意性については，次節で考察する．

4.4 考察

表 3 に示す精度が妥当であるかどうか評価するため，t 検定を行った．その結果，SVM とベースラインに対しては，有意水準 1 %において提案手法の有効性が確認できた．一方，DBN に対しては p 値が 0.076 となり，有意差が見られなかった．この原因として，一年間を通して，長期的に株価へ影響を与える事象が非常に少ないということが考えられる．実際に，テストデータである 2008 年の新聞記事において，1 ヶ月近くの株価の下落に影響を与えたと思われる事象は，リーマンショックのみであった．従って，その事象を考慮していたとしても，一年間全体の精度では差が現れにくいのではないかと推測される．

表 3 テストデータにおける株価動向推定のエラー率

銘柄	baseline	SVM	DBN	RNN-RBM +DBN
日経平均	49.57	48.73	45.50	**43.62**
日立製作所	35.71	37.29	**32.00**	**32.00**
東芝	39.52	41.95	**38.50**	**38.50**
富士通	40.00	40.25	**32.00**	34.00
シャープ	42.00	47.88	**40.00**	**40.00**
ソニー	43.00	47.46	41.43	**40.95**
日産自動車	40.00	45.34	39.50	**37.00**
トヨタ自動車	44.29	53.39	43.81	**42.38**
キャノン	43.81	53.39	43.00	**39.11**
三井物産	46.96	47.88	**41.43**	**41.43**
三菱商事	43.81	49.15	43.33	**40.43**
平均	42.61	46.61	40.05	**39.04**

そこで，リーマンショックが長期的に株価へ与えた影響を考慮できているか検証するため，一年間全体ではなく，特定の期間において再度実験を行った．ここで，特定の期間とは，リーマンショックが生じた 2008 年 9 月 15 日から 10 月 28 日までの約 1 ヶ月間である．その結果を表 4 に示す．

表 4 リーマンショックが生じてから 1 ヶ月間でのエラー率の比較

銘柄	DBN	RNN-RBM+DBN
日経平均	51.61	**38.70**
日立製作所	61.29	**32.25**
東芝	54.83	**38.70**
富士通	45.16	**32.25**
シャープ	58.06	**45.16**
ソニー	41.93	41.93
日産自動車	**29.03**	35.48
トヨタ自動車	48.38	**45.16**
キャノン	54.83	54.83
三井物産	41.93	**38.70**
三菱商事	29.03	**25.80**
平均	46.92	**39.00**

表 4 から，特定の期間に注目した場合，ほとんどの銘柄において提案手法のエラー率が DBN を下回っており，t 検定でも有意水準 5%で有意差が確認できた（$p = 0.025$）．このことから，RNN-RBM を用いることで，リーマンショックという長期的に株価に影響を与える事象を捉えることができ，且つ，その事象を考慮することが株価動向の推定に有効であったものと判断できる．

5. まとめ

本論文では，株価という時系列情報の特性に着目し，深層学習によって新聞記事が株価に与える影響の時間的な変化を捉え，株価動向推定を行う手法を提案した．本手法では，時系列情報を考慮した再帰的なモデルである RNN-RBM と RBM を階層的に積み上げて構成されたモデルである DBN を組み合わせたモデルを利用し，素性として，日単位で分けられた記事群を bag-of-words で表現したベクトルを用いた．

評価実験では，このモデルを用いて 2008 年の株価動向の推定を 10 銘柄に対して行い，従来研究の多くで用いられている SVM と比較したところ，平均して 8.26%の精度向上を実現した．各銘柄の株価の上昇・下落のうち，多い方を常に選択する手法と比較しても，平均して 3.56%の精度向上を達成し，どちらの手法に対しても，t 検定により有意差が確認され，株価動向推定に深層学習を導入することの有効性を示した．また，リーマンショックが生じてからの約 1 ヶ月間に期間を限定した実験では，DBN との平均エラー率の差が 7.92%となり，1 年間の実験と比べて，大きく差が開いた．この結果から，本手法は長期的に株価に影響を与える事象を捉えることができ，且つ，その事象を考慮することが株価動向の推定に有効であったと考えられる．

今後の課題としては，bag-of-words の他の文書の表現方法として，言語情報から必要な情報を抽出・統合するような文書の圧縮表現方法を検討していく．

参考文献

[1] Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H.: Greedy layer-wise training of deep networks, *Proceedings of the twenty-first international conference on Neural Information Processing Systems*, pp. 153–160 (2007)

[2] Boulanger-Lewandowski, N., Bengio, Y. and Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription, *Proceedings of the twenty-ninth International Conference on Machine Learning*, pp. 1159–1166 (2012)

[3] Dahl, G. E., Yu, D., Deng, L. and Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 30–42 (2012)

[4] Ding, X., Zhang, Y., Liu, T. and Duan, J.: Using Structured Events to Predict Stock Price Movement: An Empirical Investigation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 1415–1425 (2014)

[5] Gidofalvi, G. and Elkan, C.: Using news articles to predict stock price movements, *Department of Computer Science and Engineering, University of California, San Diego* (2001)

[6] Hinton, G. E., Osindero, S. and Teh, Y.-W.: A fast learning algorithm for deep belief nets, *Neural computation*, Vol. 18, No. 7, pp. 1527–1554 (2006)

[7] Huang, W., Nakamori, Y. and Wang, S.-Y.: Forecasting stock market movement direction with support vector machine, *Computers & Operations Research*, Vol. 32, No. 10, pp. 2513–2522 (2005)

[8] Izumi, K., Goto, T. and Matsui, T.: Trading Tests of Long-Term Market Forecast by Text Mining, *Proceedings of the tenth IEEE International Conference on Data Mining Workshops*, pp. 935–942 (2010)

[9] Kim, K.-j.: Financial time series forecasting using support vector machines, *Neurocomputing*, Vol. 55, No. 1, pp. 307–319 (2003)

[10] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks., *Proceedings of the twenty-fifth international conference on Neural Information Processing Systems*, pp. 1106–1114 (2012)

[11] Kudo, T.: MeCab: Yet another part-of-speech and morphological analyzer, *http://mecab. sourceforge. net/* (2005)

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

[13] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J.: Mining of concurrent text and time series, *Proceedings of the KDD-2000 Workshop on Text Mining*, pp. 37–44 (2000)

[14] Mittermayer, M.-A.: Forecasting intraday stock price trends with text mining techniques, *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, IEEE, pp. 10–pp (2004)

[15] 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 (TOIS)*, Vol. 27, No. 2, pp. 12:1–12:19 (2009)

[16] Shin, K.-S., Lee, T. S. and Kim, H.-j.: An application of support vector machines in bankruptcy prediction model, *Expert Systems with Applications*, Vol. 28, No. 1, pp. 127–135 (2005)

[17] Smolensky, P.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281 MIT Press (1986)

[18] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y. and Manning, C. D.: Semi-supervised recursive autoencoders for predicting sentiment distributions, *Proceedings of the sixteenth Conference on Empirical Methods in Natural Language Processing*, pp. 151–161 (2011)

[19] Sutskever, I. and Hinton, G. E.: Learning multilevel distributed representations for high-dimensional sequences, *Proceedings of the eleventh International Conference on Artificial Intelligence and Statistics*, pp. 548–555 (2007)

[20] Sutskever, I., Hinton, G. E. and Taylor, G. W.: The recurrent temporal restricted boltzmann machine, *Proceedings of the twenty-second international conference on Neural Information Processing Systems*, pp. 1601–1608 (2008)

[21] Tay, F. E. and Cao, L.: Application of support vector machines in financial time series forecasting, *Omega*, Vol. 29, No. 4, pp. 309–317 (2001)

[22] Werbos, P. J.: Backpropagation through time: what it does and how to do it, *Proceedings of the IEEE*, Vol. 78, No. 10, pp. 1550–1560 (1990)