

ニュース記事解析とその投資支援への応用

坂見 耕輔[†] 馬 強[†]

[†] 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

E-mail: [†]sakami@db.soc.i.kyoto-u.ac.jp, ^{††}qiang@i.kyoto-u.ac.jp

あらまし 現在日本政府は「貯蓄から投資へ」というスローガンを掲げ、投資活動を推進している。投資活動を活発化させるために、本研究では、株価の予測という手段を用いて投資家の支援を行う。株価を予測することは、ビジネスにおける意思決定や、個人の投資先の選定などに際して重要な役割を果たす。本研究では、隠れマルコフモデル、LDA を拡張し、複数銘柄の相互影響及びニュース記事による影響を組み込んで予測を行うモデルを提案する。提案モデルを用いることで、関連企業やニュースによる影響を考慮した企業の株価予測が可能になる。

キーワード グラフィカルモデル、トピックモデル、時系列データ、投資情報分析、金融工学

1 はじめに

現在日本政府は「貯蓄から投資へ」というスローガンを掲げ、投資活動を推進している。投資活動を活発化させるために、本研究では、株価の予測という手段を用いて投資家の支援を行う。株価の予測は、ビジネスにおける意思決定や、個人の投資先の選定などに際して重要な役割を果たす。特に個人投資家に関しては、株価予測に用いることのできる情報源は限られており、株価等の数値データやニュース記事等のテキストデータは非常に重要なものとなっている。

株価は企業価値の市場での評価を反映した指標であり、一方、ニュース記事は、投資の判断材料となるような企業動向などの情報を示していることも多い。ニュース記事が株価に影響を与えている可能性は高い。

利用されるデータのタイプから、株価予測の研究は、ニュース記事などのテキストを対象としたテキストマイニング系 [1] [2] と株価などの数値データを用いた時系列データ解析系 [3] に分けられる。

本研究では、企業の状態はニュース記事のトピックと株価の両方に影響を与える可能性が高いと仮定し、各時点、各企業の背後に状態を置いて、時間発展していく状態からニュースのトピックと株価が生成されるモデルを考える。この仮定を表現するために、テキストデータをトピッククラスタに分解して解析することのできる LDA (Latent Dirichlet Allocation) [4]、及び、マルコフ性を満たして時間発展していく潜在状態から観測値が得られるとする隠れマルコフモデルを用いて、テキストデータと数値データを統合的に扱う。

ここでの仮定をまとめると、

- 各状態に対応した分布から株価が生成される。
 - 各状態に対応した分布からニュースのトピック分布が生成される。
 - 各状態は 1 時点前の状態に依存し、2 時点以上前の状態とは条件付き独立である。
- の 3 点をおいている。

先行研究 [5] では、トピックモデルを用いて、各トピックが株価に与える影響をモデリングし、株価予測に用いている。ニュースが公開された時刻からちょうど 1 日後までの株価の変化率を、ニュース記事による影響であるとし、ニュースの各トピックを説明変数とした線形回帰問題を解くことによって予測値を算出している。

Park ら [6] らは、各時点で 1 つの観測値と 1 つのテキストデータが存在する時系列データに対して、トピック分布のハイパーパラメータと目的変数との間に線形な関係性を仮定したうえで、ハイパーパラメータの時間発展をマルコフ性により表現している。

Zhang ら [3] は、企業間での相互作用を、隠れマルコフモデル (Hidden Markov Model) を拡張した Extended Coupled Hidden Markov Model を用いて記述している。各状態には rise, fall の 2 種類を仮定し、観測値として、300 クラスのいずれかに分類されたニュース記事、及び 1 日の始値を用いて、EM アルゴリズムによって状態の尤度を推定している。

本研究では、株価の背後に存在する状態が時間発展していく、というモデルを考えることによって、ニュース記事の影響を時系列的に取り扱うことができ、同時に、これらの潜在状態が交互に影響し合うことで、関連企業からの影響の波及を考慮に入れることができる。また、本研究では、トピックモデルによるクラスタを用いてニュース記事を表すことにより、あるニュースに関してポジティブな意味合いとネガティブな意味合いの混合比率などの情報を取り入れつつ、分析を行うことができる。

本論文の構成は次の通りである。2 節ではニュースが株価に与える影響や株価予測に関する関連研究を示し、3 節では本研究で提案するモデルについて記す。4 節では実験の概要について説明する。そして、5 節は本研究のまとめである。

2 関連研究

ニュースやソーシャルネットワークなどのテキスト情報が株価に影響を与える、とする先行研究は数多く存在する。Li ら [7] はニュース記事が肯定的か否定的か、などの感情分析を行い、

その結果を用いて、株価予測を行っている。Lee ら [1] はニュース記事内の出現単語に、その記事による株価変動分を加算していく方法で、企業の株価の上昇、下降を予測している。単語と株価の変動の対応関係を分析して、記事内の単語に対応する株価変動を集計して、記事の報道対象である企業の株価の上昇・下落を予測する手法を提案している。

Li ら [2] は株価・ニュース・感情の3種類のデータを各軸の要素に持つテンソルを分解し、再構築することで、それぞれの交互作用を算出し、株価予測に用いている。感情の抽出は、ソーシャルメディアに含まれる感情語を用いて行っている。馬場ら [5] はトピックモデルを用いて、各トピックが株価に与える影響をモデリングし、株価予測に用いている。縮小市場を用いてそれぞれでパラメータを推定し、最後にコサイン類似度を用いて対応するトピックごとにパラメータ統合することで、計算負荷を分割している。このように、テキスト情報は株価の変動を予測するのに重要な指標のひとつになっている。

本研究では、トピックモデルとして LDA (Latent Dirichlet Allocation) [4] を用いて、ニュース記事に含まれる潜在トピックの分布を推定する。トピックモデルに関する研究は盛んに行われている。Blei ら [8] は文書に付属する評価値をトピックの混合割合から予測する Supervised topic model を提案している。これは、文書のトピックの分布と目的変数との間に線形な関係性を仮定し、回帰により係数を求める手法である。同じく Blei ら [9] は、一連の文書のトピックが時間発展していくようなモデルを提案している。トピック分布、単語分布のハイパーパラメータにマルコフ性を仮定し、分布の時系列を表現している。

3 提案モデル

本研究では、企業の状態はニュース記事のトピックと株価の両方に影響を与える可能性が高いと仮定し、各時点、各企業の背後に状態を置いて、時間発展していく状態からニュースのトピックと株価が生成されるモデルを考える。この仮定を表現するために、テキストデータをトピッククラスタに分解して解析することのできる LDA (Latent Dirichlet Allocation) [4]、及び、マルコフ性を満たして時間発展していく潜在状態から観測値が得られるとする隠れマルコフモデルを用いて、テキストデータと数値データを統合的に扱う。

本研究で提案するモデルでは、株価が以下のように、状態の各次元の重み付き線形和に従って生成されると仮定する。

$$S_t^i = \mathbf{b}^{iT} \boldsymbol{\alpha}_t^i + \epsilon_t^{iS} \quad (1)$$

$$\epsilon_t^{iS} \sim N(0, \sigma^{iS}) \quad (2)$$

ここで、 S_t^i は、時刻 t 時点の企業 i の株価、 $\mathbf{b}^i$ は $\boldsymbol{\alpha}$ の各次元にかかる重み、 $\boldsymbol{\alpha}_t^i$ は、時刻 t 時点で企業 i に言及するニュース記事のトピック分布のハイパーパラメータを表す。

トピック分布が確率分布の要件を満たすために、状態のスケールを変換し、相対的な割合でトピック分布を表現する必要がある。そこで、先行研究 [6] に習って、 $\boldsymbol{\alpha}$ から生成した正規

分布に softmax 関数をかけてトピック分布として扱う。

$$P(\theta_t^i | \boldsymbol{\alpha}_t^i) \sim N(\boldsymbol{\alpha}_t^i | \boldsymbol{\alpha}^{i2} I_k) \quad (3)$$

$$\pi_k(\theta) = \frac{\exp(\theta_k)}{\sum_{i=1}^K \exp(\theta_i)} \quad (4)$$

ここで、 $\pi_k(\theta)$ がトピック分布に相当する。 $\boldsymbol{\alpha}$ の時間発展は、以下の式で表す。

$$\boldsymbol{\alpha}_t^i = \boldsymbol{\alpha}_{t-1}^i + \sum_{j \in C_i} w_j (\boldsymbol{\alpha}_{t-1}^j - \boldsymbol{\alpha}_{t-2}^j) + \epsilon_t^{i\alpha} \quad (5)$$

$$\epsilon_t^{i\alpha} \sim N(0, \sigma^2 I) \quad (6)$$

ここで、 C_i は企業 i と関連する企業の集合を表す。第二項で、注目する企業と関連企業の状態の相対座標が、重み付きで足し合わされることによって関連する企業の状態が波及してくる様子を表す。関連する企業の抽出は、学習データの期間を通じて、注目している企業の株価との相関係数が閾値を超える企業群を関連企業として扱う。他系列からの影響の重み w_j としては、この相関係数をそのまま使うことを考えている。

提案モデルの生成過程は以下ようになる。

Algorithm 1 提案モデルの生成過程

```

1: for all  $t$  do
2:    $\boldsymbol{\alpha}_t \sim N(\boldsymbol{\alpha}_{t-1} + \sum_{j \in C_i} w_j (\boldsymbol{\alpha}_{t-1}^j - \boldsymbol{\alpha}_{t-2}^j) | \boldsymbol{\alpha}_{t-1}, \sigma^2 I_k)$ 
3:    $S_t^i \sim N(\mathbf{b}^{iT} \boldsymbol{\alpha}_t^i, \sigma^{iS})$ 
4:   for all  $d$  do
5:      $\theta_{t,d} \sim N(\boldsymbol{\alpha}_t, a^2 I)$ 
6:     for all  $n$  do
7:        $z_{t,d,n} \sim \text{Multi}(\pi(\theta_{t,d}))$ 
8:        $w_{t,d,n} \sim \text{Multi}(\beta z_{t,d,n})$ 
9:     end for
10:  end for
11: end for

```

各時刻について、時間発展により状態が変化した後、状態から株価が生成される。その後、各ドキュメントについて状態からトピック分布を生成し、各単語に相当するトピックの抽出及び単語の抽出が行われる。

提案モデルのグラフィカルモデルを以下に示す。

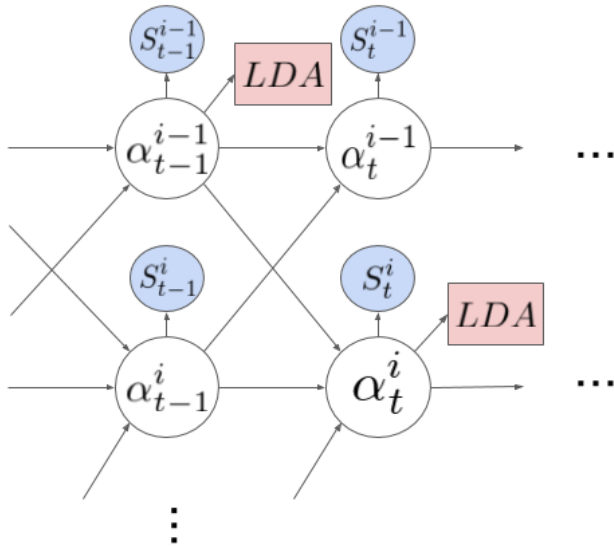


図1 グラフィカルモデル

表1 図1の変数

変数	意味
α_t^i	企業 i の時刻 t における潜在状態
S_t^i	企業 i の時刻 t における株価

ここでは、右方向に時間軸を置いて、時刻 t における企業 i の状態 α_t^i が、関連企業の影響を受けながら時間発展している様子を表している。また、各状態から株価が生成され、イベントに応じてニュース記事の生成も起こる。

図内の LDA 部分は下図のように表される。

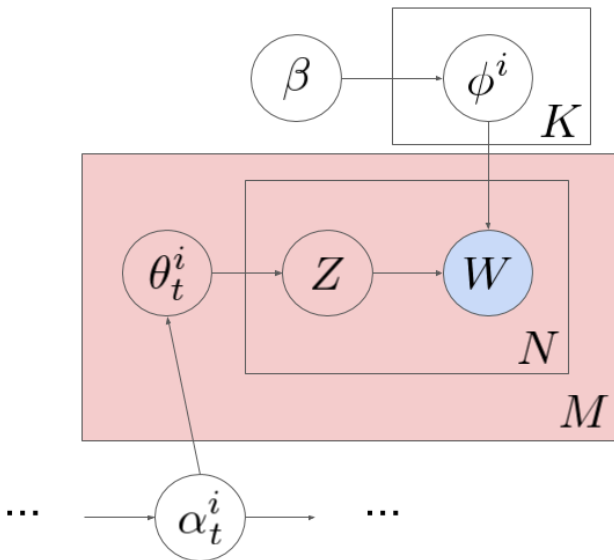


図2 LDA

表2 図2の変数

変数	意味
α_t^i	企業 i の時刻 t における潜在状態
θ_t^i	時刻 t に企業 i に言及したニュース記事のトピック分布
Z	各単語に対応するトピック
W	単語
ϕ^i	企業 i に対応するトピッククラスターの単語分布
β	単語分布のハイパーパラメータ

提案モデルの懸念点として、計算量が大きくなってしまいうことが考えられるが、ニュースを言及企業だけに絞っているため、実際に LDA を使って計算を行う箇所は限られていることや、企業数を、例えば、情報・IT 業界の大手 5 社などに絞って考えることで、計算負荷を軽減することができる。

4 実験

本節では、実際の株価データとニュース記事データを用いてパラメータを学習させた結果と考察を示す。また、表3に用いたハイパーパラメータを示す。

4.1 実験方法

実験では、1ヶ月分の株価データ及びニュース記事データを用いてパラメータを学習させた後に、翌月のデータに対して株価の予測を行う。また、単純に提案モデルをデータに適用したところ、パラメータが収束しなかったため、今回の実験では LDA を事前学習させて、計算量を削減した。事前学習では、2015 年 1 月から 2016 年 12 月までの日経 QUICK ニュースの全コーパスを使用し、各ニュースのトピッククラスター及びその単語分布を推定した。

4.2 推定

推定には MCMC サンプリングの手法の一つであるハミルトニアンモンテカルロを用いる。3 チェイン、5000 イテレーションでサンプリングを行い、最初の 500 イテレーションをバーンインとして棄却した。

4.3 評価

実験の評価は、テスト期間の株価データに対して、株価の値そのものの予測値ではなく、前日の株価に比較して上昇するか、下降するかの 2 値分類で行う。これは、多くの先行研究の評価が 2 値分類で行っているため、評価尺度を統一する目的である。

4.4 データセット

株価データは東京証券取引所¹の日時四本値を使用する。日

表3 ハイパーパラメータ

変数	値
β	各次元 0.5 を持つ、10 次元ベクトル
K (トピック数)	10

1: <http://db-ec.jp.co.jp/>

表 4 抽出した企業の相関係数

	コムシス	ライオン	サンケン	日立造船	島津製作所
コムシス	1.	0.95	0.91	0.94	-0.91
ライオン	0.95	1.	0.92	0.88	-0.84
サンケン	0.91	0.92	1.	0.81	-0.77
日立造船	0.94	0.88	0.81	1.	-0.91
島津製作所	-0.91	-0.84	-0.77	-0.91	1.

時四本値は、始値、高値、安値、終値、売買高、売買代金等を記録しており、今回の実験ではこの内の始値を用いる。具体的には、2016 年 5 月の 1ヶ月間のデータを用いた。

ニュース記事のデータは、日経 QUICK²のニュース記事を使用する。日経 QUICK は株式投資金融情報サイトであり、市場の動向を即座に反映している。

4.5 データの前処理

● 株価データ

本研究で推定する株価は、企業間でスケールを揃えるために、前日の株価に対する変動率を使用する。

$$S_t = \frac{S_t - S_{t-1}}{S_{t-1}} \quad (7)$$

今回の実験では、ある程度取引量が多く、互いに相関もしくは逆相関する企業群の中からランダムに、コムシスホールディングス株式会社、ライオン株式会社、サンケン電気株式会社、日立造船株式会社、株式会社島津製作所、の 5 社を選択した。それぞれの相関係数を表 4 に示す。

今回はこの中から、相関係数の絶対値が 0.9 以上のものだけを選んで関連企業とした。

● ニュース記事データ

ニュース記事をモデルの入力として用いるために、文章を単語に分割する必要がある。そのために、日本語形態素解析のソフトである MeCab [10] を利用する。使用する辞書としては、比較的最新の固有名詞等も登録されている、mecab-ipadic-neologd³を用いる。また、企業に影響のない単語は省くために、使用する品詞を名詞・動詞・形容詞に限定し、URL や日付、記号・数字などの情報を削除して、モデルの入力に用いる。データとしては本文のみを使用し、ニュースタイトルは使用しない。

4.6 考 察

この条件下で実験を行ったが、各状態のパラメータは収束せず、予測まで行うことは難しい、という結果となった。ただし、株価を求める際の、状態の各次元に対する線形回帰の係数に関しては収束が見られたため、その結果を示す。図 3 は、一部の係数に関して推定した分布を示したものである。 $b[i, j]$ は、 i 番目の系列（企業）、トピック j に対応する係数を表す。また、色によってチェーンごとの推定結果を示している。

結果を見ると、図 3 の一番上のプロットのように、0 から少し外れた値を持ちやすい分布になっているものもあるが、9 割

近いパラメータは下の 3 つのプロットのように、0 を中心とする正規分布に似た確率分布を持っていた。

原因としては、以下の点が考えられる。

- LDA を事前学習させたため、株価変動に影響を与えるトピッククラスタをうまく抽出することができなかった

教師あり LDA [8] のように学習させることによって、観測値の回帰に適したトピッククラスタを作る効果が期待できる。今回の実験では LDA を事前学習させ、分布の独立性を仮定して計算量を削減したことにより、株価変動に影響を与えるトピッククラスタをうまく抽出することができなかったと考えられる。

- トピック数として、今回は 10 を選択したが、株価を表現する状態として 10 次元は冗長すぎた

2016 年 5 月の 1 月の時点数は土日・祝日を除くと 17 しかなく、データ点の数に対してモデルの表現力が高すぎたことが問題として大きかったと考えられる。

- 日経 QUICK ニュースは株価予測に適したコーパスではなかった

今回の実験ではコーパスの準備ができなかったため、日経 QUICK ニュースを用いて実験を行ったが、日経 QUICK ニュースはあるイベントによって株価が上昇・下降した、という「変化後」の記事が多くを占めるため、株価予測に適したコーパスではなかった。

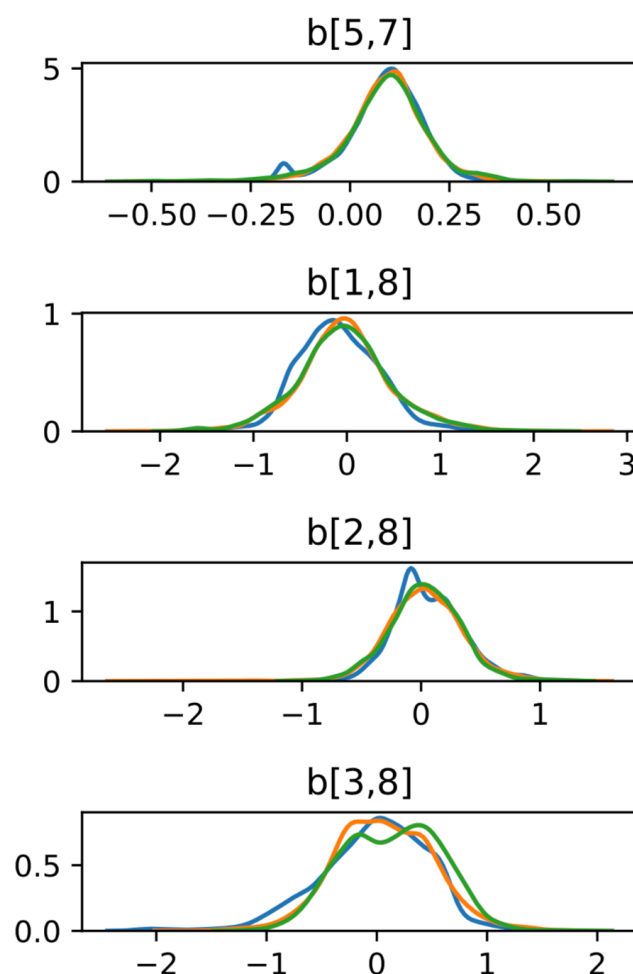


図 3 実験結果

² : <https://www.quick.co.jp/>

³ : <https://github.com/neologd/mecab-ipadic-neologd/>

5 おわりに

本研究では、複数の企業の株価データとニュースを、隠れマルコフモデル及びトピックモデルを組み合わせて統一的に分析できるモデルを提案した。提案モデルを用いれば、株価を推定し、ビジネスの意思決定支援や、個人投資家の銘柄選定などに役立たせる効果が期待できる。

今後の予定を以下に示す。

- トピック数に対して適切なデータ数を用意し、実験を行う。
- 予測に対して適切なコーパスを用意し、実験を行う。

考察で示した問題点を解消したのちに、扱える企業数を増やしたり、逐次的にパラメータを更新していくことができるように実装、モデルの見直しを行っていく。また、関連企業の抽出においても、Granger 因果推定などの方法を用いてより正確に行っていく。

謝 辞

本研究の一部は科研費（16K12532）と 京都大学教育研究振興財団の助成による。

文 献

- [1] Kari Lee and Ryan Timmons. Predicting the stock market with news articles.
- [2] Qing Li, LiLing Jiang, Ping Li, and Hsinchun Chen. Tensor-based learning for predicting stock movements. In *AAAI*, pp. 1784–1790, 2015.
- [3] Xi Zhang, Yixuan Li, Senzhang Wang, Binxing Fang, and S Yu Philip. Enhancing stock market prediction with extended coupled hidden markov model over multi-sourced data. *Knowledge and Information Systems*, pp. 1–20, 2018.
- [4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.
- [5] Satoshi Baba and Qiang Ma. A hetero-topic-model and its application to investment support.
- [6] Sungrae Park, Wonsung Lee, and Il-Chul Moon. Associative topic models with numerical time series. *Information Processing & Management*, Vol. 51, No. 5, pp. 737–755, 2015.
- [7] Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. The effect of news and public mood on stock movements. *Information Sciences*, Vol. 278, pp. 826–840, 2014.
- [8] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pp. 121–128, 2008.
- [9] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006.
- [10] 工藤拓, 山本薫, 松本裕治ほか. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理 (NL), Vol. 2004, No. 47 (2004-NL-161), pp. 89–96, 2004.