ニュース記事で報道される社会的イベントを考慮した株価動向予測の補正

Adjusting Stock Price Prediction Considering Web News Articles

中菅 章浩[1*]　関 和広[1]
上原 邦昭[1]
Akihiro Nakasuga[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: One is numerical information, including past stock prices, currency exchange rates, and interest rates. The other is textual information, mainly news stories covering statements of government dignitaries, consumer trends, and miscellaneous events. Although numerical information has been proven useful for predicting stock prices, its predictive power is limited in a sense that much information, such as the statements of government dignitaries mentioned above, resides only in textual information. Given a stock for which one would like to predict the future price, textual data provides different—but much wider coverage—of information from numerical information, which may be beneficial in prediction. This study exploits public Web news articles and attempts to estimate the residue that cannot be explained only by numerical information through a simple additive regression model. In addition, to distinguish between different types of news articles, such as those specific to particular companies or types of industry and more general top news, the framework of multiple kernel learning is adopted. The validity and effectiveness of the proposed approach is evaluated on the real-world data consisting of share prices of Nikkei 220 companies and 47 thousand Web news articles.

1　はじめに

今日，投資家が投資を行う際には，投資商品の価格や政策金利などの経済指標，さらには，日本銀行や金融機関の発表や，為替，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選択し，市場の動きを分析・予測している．市場の分析に用いられるこのような情報は，主に 2 種類に大別することができる．一つは，企業の株価や物価指数，政策金利といった数値情報，もう一つは，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・事故などの社会的イベントを伝えるニュース記事に代表される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事実上不可能である．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，情報処理技術を市場分析に適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挙げてきたものの，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．

また，数値情報は指標化された情報であり，それ以上の情報を含まない．将来の株価は，過去の株価の傾向やテクニカル指標等の数値情報のみに影響されるのではなく，その企業の行う発表や関わった事件，さらにトップニュースとなるような社会的イベントにも影響されて変化する．例えば，ある電機メーカーが大きく期待される新製品を発表すると，その電機メーカーの株価は上昇する．また，食品メーカーが製品に異物を混入させてしまったと報道されれば，その食品メーカーの株価は下落する．さらに，直接企業に関係のない事件として，大震災の話題が出れば，その被災地域にある拠点や工場が被害を受け，株価が下落する企業もある．このような企業の動向や社会的イベントの情報は，企業の発表するレポートや一般のニュースとして発信されるため，数値情報には含まれず，テキスト情報の中に存在する．

*連絡先：(神戸大学大学院システム情報学研究科)
(兵庫県神戸市六甲台町 1–1)
E-mail: nakasuga@ai.cs.kobe-u.ac.jp

本論文では，株価などの数値情報だけでなく，一般に配信されているニュース記事を利用した株価動向推定を試みる．前述の例のように，企業の株価は，ニュース記事として表出する現実世界の事象の影響によって変動する場合がある．そこで，数値情報のみから学習した回帰関数から得た推定株価と実際の株価との誤差を，ニュース記事の情報を用いて補正することで，数値情報のみからは推定できないニュース記事の影響を考慮した株価の変動を推定する．本研究ではテキスト情報として，Web 上で一般に配信されているニュース記事を使用する．

2 関連研究

これまで様々な経済指標を用いて株価動向推定などの市場分析を行う研究が数多く行われてきた．ここでは，関連研究として，テキスト情報を用いて国債価格の予測を行った研究，数値情報とテキスト情報を用いて株価予測を行った研究の例をあげ，本研究の位置付けを示す．

Izumi et al. [Izumi 10] は，日銀月報を分析して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．モデルを式 (1) に示す．

$$\hat{p}_t = a_0 + \sum_{i=1}^{n} a_i x_{i,t} \tag{1}$$

$\hat{p}_t$ は，t 月の予測価格であり，a_i は回帰パラメータ，$x_{i,t}$ が t 月における日銀月報を分析し抽出したテキストの特徴量，n が特徴量の数（次元）である．

この研究で利用されている特徴量の抽出手法は，KeyGraph [Ohsawa 98] 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た．次に，抽出した単語に主成分分析を適用し，累積寄与率 0.6 を基準に 30 の主成分が特徴として採用された．

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回帰や，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もテキスト情報が有効である事が示された．

Tang et al. [Tang 09] は，数値情報とテキスト情報の両方を用いて，株価の予測を試みた．具体的には，数値情報である株価データを入力とした MA（Moving Average）モデルとテキスト情報であるニュース記事から抽出した素性ベクトルを入力とした MA モデルの線型和によって，株価回帰モデルを構築した．定義式を以下に示す．

$$\hat{y}_t = \sum_{k=t-T}^{t-1} \beta_k \cdot y_k / T \tag{2}$$

$$\hat{x}_t = \sum_{k=t-T}^{t-1} \theta_k \cdot x_k / T \tag{3}$$

$$\bar{y}_t = \hat{y}_t + \alpha \cdot \hat{x}_t \tag{4}$$

式 (2) が過去の株価を入力とする MA モデル，式 (3) が過去のニュース記事を入力とする MA モデル，そして式 (4) が最終的な予測モデルである．ここで，t は特定の時間（日），y_k は時間 k における実株価，x_k は時間 k に利用可能なテキスト情報から得た特徴量，T は考慮する時間窓の大きさ（日数）を表す．各パラメタ α，β，θ は，訓練データから最小二乗法によって決定する．実験に用いたデータは上海 A 株の日足データであり，エネルギー業界，情報通信業界，不動産業界の指数を予測した．三つの業界について実験を行ったのは，関連するニュース記事がほとんどないという状況を避けるためである．実験では，数値情報のみから学習した MA モデルと提案モデルを絶対予測誤差について比較し，全ての業界において提案手法の予測誤差が下回っていた．つまり，ニュース記事の情報を利用することによって，株価予測精度を向上させることができた．しかしこの研究では，テキストから特徴として抽出する単語（ニュースの中でよく使用されている単語，あるいは推定対象銘柄に関連している単語）の候補を手作業で決定している．このため，推定対象銘柄が変わった場合には，改めて単語の候補を作り直す必要があり，コストが高い．

3 社会的イベントを考慮した株価予測補正

3.1 概要

将来の株価を推定する場合，過去の株価やテクニカル指標といった数値情報のみからでは予測できない要因が存在する．例えば，2011 年 10 月 14 日にオリンパスのマイケル・ウッドフォード元社長が解任されたことからオリンパスの過去の粉飾決算が発覚し，株価は 2011 年 10 月 14 日から 24 日の間に，2045 円から 1099 円まで下落した（図 1）．このようなニュースの影響を受けた場合，過去の株価の動きなどの数値情報のみから将来の株価を予想することは難しい．

本研究では，上記のような数値情報のみからは説明できない変動をニュースの影響と仮定し，ニュース記事から抽出したデータを用いて，株価時系列の推定誤差を回帰する．図 2 に，提案モデルの概念図を示す．左の

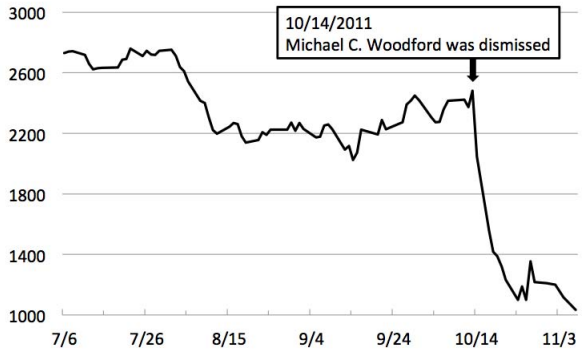

図 1: マイケル・ウッドフォード元社長解任前後のオリンパスの株価の推移.

グラフは実際の株価時系列を表し，中央のグラフは数値情報のみを利用して推定した株価時系列を表す．右のグラフは数値情報のみを利用して推定した株価と実際の株価との誤差の時系列を表す．中央と右のグラフの和が左のグラフに等しく，右のグラフで表した推定誤差がニュースの影響で生じたものと仮定する．すなわち，数値情報のみで推定した株価にニュース記事の情報から推定した値を加えることで，株価動向の予測を補正する．

3.2 株価時系列回帰モデル

本研究では，株価時系列回帰にサポートベクトル回帰（以下 SVR）[Drucker 96] をベースラインとして利用する．SVR は汎化能力に優れており，経済学分野を含め，これまで多くの分野で利用されている [Chang 10, Long 11, Van Gestel 01, Schumaker 09] ．提案モデルでは加法回帰モデル [Witten 11] の考えを採用し，SVR の回帰式にテキスト情報による補正項を加える．つまり，テキスト情報から予測される項を加えることで，SVR の推定誤差を補正する．実株価（推定日の終値）を z としたとき，次式に予測株価 $\hat{z}$ の推定式を示す．

$$\hat{z} = f(\mathbf{x}, \mathbf{y}) \tag{5}$$

$$= \sum_{i=1}^{n} \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x}) + \sum_{j=1}^{m} w_j \mathrm{y}_j z\,(\mathbf{x}, \mathbf{y}) \tag{6}$$

$$= \sum_{i=1}^{n_1} \alpha_i k\left(\mathbf{x}^{(i)}, \mathbf{x}\right) + \sum_{j=1}^{K} \sum_{l=1}^{n_2} \beta_j k'_j \left(\mathbf{y}^{(l)}, \mathbf{y}\right) \tag{7}$$

式 (7) の第一項は，SVR の定義式である．$\alpha_i\ (i = 1, 2, ..., n)$ は SVR の回帰パラメタであり，$k(\mathbf{x}^{(i)}, \mathbf{x})$ はカーネル関数である．$\mathbf{x}$ は説明変数であり，推定日の前 d 営業日の終値を用いる．なお，通貨レートや NY ダウなどの数値情報も株価予測に有効であると考えられる．しかし，本研究の主目的は，株価予測におけるニュース記事の有効性を提案モデルにおいて実証することであるため，簡単化のため本研究では用いない．式 (7) の第二項は MKL（Multiple Kernel Learning）[Gönen 11] による回帰式である．$\mathbf{y} = (y_1, y_2, \ldots, y_m)$ は，株価動向を推定する日の前日の Web ニュース記事から生成した特徴ベクトルである．

続いて，モデルの学習と評価において，株価データとニュースデータをどのように利用するかを架空の株価時系列を使って図 3 に示す．まず，図の左側の期間 t_0 における株価データを訓練データとして，回帰モデル f_{num} を学習する．続いて，学習されたモデル f_{num} を利用して，中央の期間 t_1 の株価を予測する．この予測結果と t_1 における実際の株価との残差（residue）を数値情報外の社会的イベントによる影響と捉え，残差を非説明変数，この期間の Web ニュース記事から得られた特徴量を説明変数として，回帰モデル f_{news} を学習する．最後に，このように学習された二つのモデル f_{num}, f_{news} の和によって，期間 t_1 の株価を予測する．期間 t_0 については株価データだけ，期間 t_1（と期間 t_2）については，株価データとニュースデータの両方が必要になる．

3.3 株価データからの回帰

期間 t_0 の株価データを用いて，式 (7) の第一項のモデル f_{num} を学習する．これは，式 (8) における通常の SVR の回帰パラメータ α_i を学習することに相当する．なお，この時点ではテキスト情報は一切用いない．

$$f_{num}(\mathbf{x}_{t_0}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_{t_0}^{(i)}, \mathbf{x}_{t_0}) \tag{8}$$

ここで，$\mathbf{x}_{t_0}$ は説明変数であり，期間 t_0 内の d 日分の株価終値の系列（ベクトル）である．式 (8) のパラメタ α_i は，式 (9) で表す ϵ 不感応損失関数から導かれる経験的リスクを最小化することで求めることができる [Drucker 96].

$$L_\epsilon(z, f_{num}(\mathbf{x}_{t_o})) = \begin{cases} 0 & \text{if } |z - f_{num}(\mathbf{x}_{t_o})| \le \epsilon \\ |z - f_{num}(\mathbf{x}_{t_o})| - \epsilon & \text{otherwise} \end{cases} \tag{9}$$

3.4 Web ニュース記事による推定誤差回帰

次に，推定した SVR の推定誤差をテキスト情報を用いて MKL [Gönen 11] で回帰する．MKL は，サポートベクトル回帰などのカーネルを用いた機械学習手法

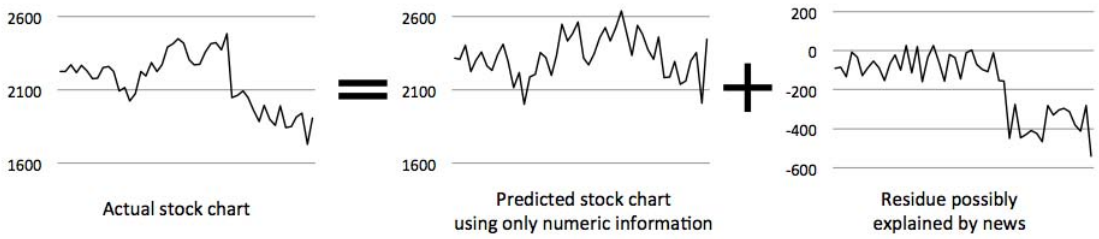

図 2: 提案モデルの概念図

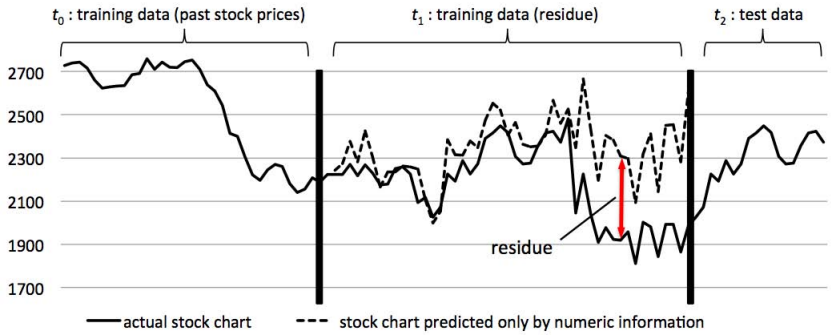

図 3: 二種類の訓練データとテストデータに分割された株価時系列データ

において，複数のカーネルの線形結合によって得られるカーネル関数を用いた手法であり，それぞれのカーネルの最適な重みを学習する．結合カーネルは以下の式で表される．

$$\mathrm{K}(x, x') = \sum_{j=1}^{K} \sum_{l=1}^{n} \beta_j k_j \left(x^{(l)}, x' \right) \qquad (10)$$

K 個のカーネルのうち，j 番目のカーネル k_j の重みが $\beta_j (\geq 0)$ であり，その総和 $\sum_{j=1}^{K} \beta_j$ は 1 である．本研究では，銘柄に関するニュース，銘柄と同じ業種に関するニュース，トップニュースの 3 つの種類のニュースを組み合わせてニュースによる誤差回帰を行う．したがって，それぞれの種類のニュースに対してカーネルの最適な重みを学習する．

4 Web ニュース記事の取得とデータ表現

4.1 Web ニュース記事

本研究で用いたテキスト情報は，日経平均に採用されている銘柄に関する Web ニュース記事である．これらの Web ニュース記事は，Google アラート[1]に銘柄の正式名称を設定し，この名称を内容に含むニュースを RSS フィードで受信することで取得した．また，政治や国債情勢等，企業とは直接関係ない世間一般の情報も株価に影響を与えると考えられるため，毎日のトップニュースも取得した．トップニュースは，Google ニュース[2]の RSS フィードから取得した．取得した Web ニュース記事の内容は，RSS フィードに含まれる各ニュースの見出しと要約のみである．図 4 に，パナソニックに関して取得した Web ニュース記事の例を示す．

```
<news id="1">
        <url>http://news.kakaku.com/prdnews/cd=kaden/ctcd=2120/id=20010/</url>
        <title>パナソニック、幅60cm未満のスリムタイプ冷蔵庫</title>
        <date>20120126</date>
        <body>パナソニックは、冷蔵庫の新モデルとして、321Lモデル「NR-C32AM」「NR-C32AM-CK」、365Lモデル「NR-C37AM」の3機種を発表。2月下旬より発売する。いずれも、3ドアで幅60cm未満のスリムタイプを実現したモデル。クリーンな冷気で庫内を冷やすAg抗菌脱臭や、外して丸洗い
        </body>
</news>
```

図 4: パナソニックに関して取得したニュース記事の例.

4.2 データ表現

取得した Web ニュース記事から特徴を抽出するため，形態素解析器 Sen によって単語の同定，基本形への変換を行った．そして，文章の意味をよく表す内容語として，名詞・動詞・形容詞を抽出した．

[1] http://www.google.co.jp/alerts

[2] http://news.google.co.jp/

次に，抽出した単語集合に TFIDF [Sparck Jones 72] を適用し，各ニュース記事を表現する単語ベクトルを生成した．TFIDF は，情報検索や文書分類で一般的に利用される単語の重み付け法であり，局所的に頻出する語により大きな重みを与える．本研究では，実験に用いるニュース記事の集合として以下の 3 カテゴリを利用し，それぞれについて TFIDF を計算した．

- 推定対象の銘柄に関するニュース（以下，銘柄ニュース）
- 各業界に関するニュース（以下，業界ニュース）
- トップニュース

ここでいう業界とは，日経平均 225 において分類された業種のことであり，全 35 業種ある[3]．上述のそれぞれのカテゴリで TFIDF を計算することによって，各カテゴリに適した語の重みを算出することができる．なお，テキストデータからの特徴量の抽出（生成）には，主成分分析や Latent Dirichlet Allocation [Blei 03] などの方法も考えられる．しかしながら，予備実験においてこれらの手法は予測性能向上につながらなかったため，本研究では利用しない．

4.3 欠損データ

ある特定の銘柄（たとえばパナソニック）のニュースは，毎日報道されるわけではない．もし，推定対象銘柄の Web ニュース記事が存在しない場合，特徴量 $\mathbf{y}$ が取得出来ず，提案モデルによって株価の推定を行うことはできない．このような欠損データを補間する方法として，すべての特徴量をゼロとしたデータを用いる方法，推定対象日の前日の Web ニュース記事から抽出した特徴量を流用する方法（前日もニュースが存在しなければ，ニュース記事が存在する日までさかのぼる.），過去のニュースデータから抽出された特徴量の平均値を利用する方法などが考えられる．予備実験では，推定対象日の前日の Web ニュース記事から抽出した特徴量を流用する方法が最も優れていたため，本研究では，この方法を利用する．

なお，期間 t_1（3 参照）の学習時には，存在するニュース記事の影響を適切に学習するために，ニュースデータが存在している場合だけを学習に用いる．テスト時には，全ての日付について推定を行う必要があるので，それぞれの方法で補間したニュースデータを用いて株価動向推定を行う．

[3] http://www.nikkei.co.jp/nkave/index.html

5 評価実験

5.1 実験設定

推定対象の銘柄は，日経平均に採用されている 225 銘柄のうち，後述する方法で選択した 3 銘柄である．これら 3 銘柄の過去の株価は，インターネット上で公開されている各銘柄の日足データを利用した[4,5] ．Web ニュース記事は，Google アラートにおいて，各銘柄の正式名称で登録して RSS にて配信されたニュースと，Google ニュースの RSS にて配信されたトップニュースを利用した．取得した記事数は，銘柄に関連するニュース記事が 40,859 件，トップニュースが 6,346 件で，合計 47,205 件である．

また，実験に使用するデータは，株価時系列回帰の訓練のためのデータ，Web ニュース記事による推定誤差回帰の訓練のためのデータ，評価テストのためのデータの 3 つに分割した．株価時系列回帰の訓練に使用する株価データは，2009 年 8 月 10 日から 2011 年 4 月 26 日までのデータであり，Web ニュース記事による推定誤差回帰に利用する株価データとニュースデータは，2011 年 4 月 27 日から 2011 年 8 月 24 日までのデータである．評価テストにて利用する株価データとニュースデータは，2011 年 8 月 25 日から 2011 年 10 月 27 日のデータである．また，テストデータにおけるニュースデータは欠損データを 4.3 節で述べた方法により補完して用いた．

株価時系列回帰の訓練時には，SVR のパラメータは，10 分割交差検定により決定し，Web ニュース記事による推定誤差回帰時に用いるパラメータは実験的に $C = 1.0$ とした．

実験結果の評価には，二乗平均平方根誤差比率（Root Mean Square Error Rate; RMSER）と株価動向適合率（Up Down Correct Rate; UDCR）を用いた．RMSER は，各銘柄における各日の推定誤差を二乗してから平均し，平方根をとった値が，その銘柄の平均株価の何%になるかを示す値である．また，実験結果の表に示す RMSER の値は，銘柄で求めた RMSER の平均値である．実験評価に RMSER を利用する理由は，株価の桁が銘柄によって大きく異なるためである．例えば，2012 年 1 月 23 日において，ソニーの終値は 1,422 円であったが，KDDI の終値は 482,500 円であった．このように株価の桁が違うほど異なると，各銘柄に対する推定誤差が比較できない上，平均誤差を算出することができない．そこで，平均株価に対する誤差比率を見ることで，各銘柄における推定誤差の比較や手法の有効性の検証を行う．UDCR は，株価推定日の前日から株価推定日当日の実際の株価の上がり下がりと，提案手法

[4] http://www.mujinzou.jp/
[5] http://homepage1.nifty.com/hdatelier/data.htm

表 1: 比較実験の結果.

手法	RMSER	UDCR
SVR	2.148	0.786
単一カーネル	3.245	0.794
MKL（B+C+T）	2.190	0.794
MKL（B+C）	2.176	0.794
MKL（B+T）	2.178	0.786
MKL（C+T）	2.210	0.786

での上がり下がりの一致率を表すため，前日の株価から上昇するか下落するかの動向を推定する精度を測ることができる．

6 評価実験

実験は，ニュースの種類別にまとめたニュースデータを用いて，以下に述べるそれぞれの方法での株価動向推定への有効性を検証した．本研究は，ニュース記事を用いて推定誤差を補正することを目的としているため，221 銘柄の中でも特にコンスタントにニュース記事が発信されている銘柄に絞って行った．1 日あたりの平均ニュース数が 10 記事を超える上位 3 銘柄に関して評価を行う．まず，TF-IDF をニュース記事の特徴としたデータを用いて MKL の枠組みで実験を行った．次に，ニュース情報を用いた誤差補正の有効性を確認するため，SVR により数値情報のみから株価の動向推定を行った．最後に，MKL の有効性を確認するため，銘柄ニュース，業界ニュース，トップニュースのそれぞれから抽出した単語を別の単語として区別し，一つのニュースデータとして扱うことで，一般的な SVR を用いて複数のニュースを組み合わせる手法で株価の動向推定実験を行った．

各実験の結果を RMSER と UDCR を用いて比較し，Web ニュース記事による推定誤差回帰によって，SVR 単独で回帰を行うよりも的確な株価動向推定が可能であるかを検証した．また，MKL を用いることで，異なる別のニュースソースからなるカーネルの重みを自動的に最適化する手法の有効性を確認した．

6.1 TF-IDF を用いた提案手法の推定誤差比率比較

実験で用いるニュースデータは，ニュース記事中に出現する単語の TF-IDF をニュース記事の特徴としたデータである．表 1 に，実験結果を示す．

表 1 中の Method 列の「SVR」は数値情報のみから株価を予測する手法,「単一カーネル」は銘柄ニュース，業界ニュース，トップニュースのそれぞれから抽出した単語を別の単語として区別し，一つのニュースデータとして扱うことで，単一の SVR を用いて複数のニュースを組み合わせる方法,「MKL」は MKL を用いて複数のニュースを組み合わせる方法である．ここで，B は銘柄ニュース，C は業界ニュース，T はトップニュースを表す．

表 1 から，RMSER の指標で SVR と他の手法を比較すると，SVR のエラー率が低く，テキスト情報の利用が必ずしも有効に働いていないことが分かる．一方，UDCR で比較すると，単一カーネルと MKL（B+C+T），MKL（B+C）の精度の方が SVR よりもわずかながら高い．MKL（B+T），MKL（C+T）では性能向上は見られなかったことから，銘柄ニュースとカテゴリニュースの組み合わせが特に重要であることが示唆される．

6.2 事例分析

ここでは，SVR による回帰の誤差をニュースデータを用いて補正に成功した事例を紹介する．図 5 に KDDI 株式会社の 2011 年 9 月 8 日から 2011 年 10 月 4 日までの株価と，SVR による予測株価，MKL（B+C+T）による予測株価の予測補正結果を示す．2011 年 9 月 21 日は，台風 9 号の影響を受け，KDDI の au 携帯電話のサービスに障害が発生したことがニュースにて報じられた．この結果，実際の KDDI 株価は 2011 年 9 月 21 日を境に急落していることが分かる．この例からも分かるように，投資家はニュースの情報も投資動向の判断材料にしていることが分かる．ここで注目したいのが，MKL（B+C+T）による株価の推定誤差補正の結果である．2011 年 9 月 22 日には実際の株価が下がっているにも関わらず，数値情報のみから回帰を行なっている SVR では，株価急落が予測できていない．しかし，MKL（B+C+T）に注目すると，回帰結果をニュース情報を元にして SVR での回帰結果を下方に補正した結果，KDDI 株が急落することを予測できている．このことより，ニュース情報を元にした社会的イベントによる株価補正がうまく機能していることが分かる．

さらに，図 6 にソフトバンク株式会社の 2011 年 9 月 8 日から 2011 年 10 月 4 日までの株価と，SVR による予測株価，MKL（B+C+T）による予測株価の予測補正結果を示す．2011 年 9 月 16 日付近を境に株価が急落していることが分かる．2011 年 9 月 16 日に日本経済新聞が，KDDI 株式会社が米 Apple 社のスマートフォンである iPhone を扱うという一報を報じた．それまで iPhone はソフトバンク株式会社が独占して販売していた製品であるため，この報道はソフトバンク株

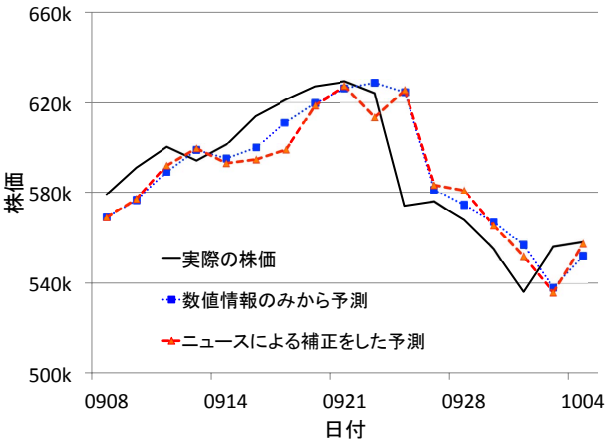

図 5: KDDI の株価予測.

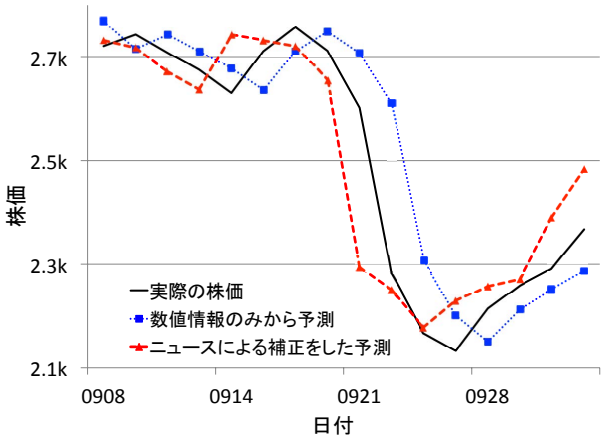

図 6: ソフトバンクの株価予測.

式会社にとってマイナスの要因と受け取られたものと考えられる．ここで SVR の動きを見てみると，実際の株価が下落しているにも関わらず，SVR は株価の上昇を予測している．しかし，MKL（B+C+T）による補正により，最終的な予測は株価が下落すると正しく推定ができた．ここで特に重要なことは，KDDI 株式会社が iPhone を扱うという報道はソフトバンク株式会社に関する報道ではない．しかし，KDDI 株式会社とソフトバンク株式会社は同じ業界に属しており，MKL（B+C+T）の業界ニュース，トップニュースの情報より株価の動向を推定することができた．

7　まとめ

本論文では，Web 上のニュース記事を利用し社会的イベントを考慮した株価動向の推定誤差の補正手法を提案した．本手法では，過去の株価を学習した回帰式から得た推定株価の誤差を社会的イベントによるものであると仮定し，Web ニュース記事から得られる特徴量を用いてその誤差を回帰することで，より正確な株価動向の推定を試みた．

評価実験では，本論文で提案したニュース記事に表出する社会的イベントを考慮した株価予測補正の手法を用いて，2011 年 8 月 25 日から 2011 年 10 月 27 日の株価推定を行った．その結果，株価動向適合率にわずかながら改善が見られた．また，銘柄に関するニュース，業種に関するニュース，トップニュースの 3 種類の異種ニュースを組み合わせる手法として MKL を導入し，単純な SVR を用いて異種ニュースから学習をする場合と比較し，推定誤差比率が改善することを確認した．

一方，現在用いているモデルでは，株価情報から学習した SVR モデルが古いまま使用されるため，さらに未来の予測を行う場合，SVR のモデルが必ずしもその時点で適切なモデルであるとは限らないという問題がある．従って，常に最新の株価データ，ニュースデータからモデルを学習できるようにするため，SVR の学習期間，またニュースデータの学習期間を調整できるように枠組を見直す必要がある．

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Chang 10] Chang, C.-Y.: Application of support vector regression for phyciological emotion recognition, in *Proceedings of the 2010 International Computer Symposium*, pp. 12–17 (2010)

[Drucker 96] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V.: Support Vector Regression Machines, in *Proceedings of the 10th annual conference on neural information processing systems*, pp. 155–161 (1996)

[Gönen 11] Gönen, M. and Alpaydin, E.: Multiple Kernel Learning Algorithms, *Journal of Machine Learning Research*, Vol. 12, pp. 2211–2268 (2011)

[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*, pp. 935–942 (2010)

[Long 11] Long, N., Gianola, D., Rosa, G. J. M., and Weigel, K. A.: Application of support vector regression to genome-assisted prediction of quantitative traits, *Theoretical and Applied Genetics*, Vol. 123, No. 7, pp. 1065–1074 (2011)

[Ohsawa 98] Ohsawa, Y., Benson, N. E., and Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, in *Proceedings of the Advances in Digital Libraries Conference (ADL 98)*, pp. 12–18 (1998)

[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, pp. 12:1–12:19 (2009)

[Sparck Jones 72] Sparck Jones, K.: Statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol. 28, No. 1, pp. 11–20 (1972)

[Tang 09] Tang, X., Yang, C., and Zhou, J.: Stock Price Forecastiong by Combining News Mining and Times Series Analysis, in *Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pp. 279–282 (2009)

[Van Gestel 01] Van Gestel, T., Suykens, J., Baestaens, D.-E., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., and Vandewalle, J.: Financial time series prediction using least squares support vector machines within the evidence framework, *IEEE Transactions on Neural Networks*, Vol. 12, No. 4, pp. 809–821 (2001)

[Witten 11] Witten, I. H., Frank, E., and Hall, M. A.: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Massachusetts (2011)