研究論文

尤度比を用いたモデル数削減と予測精度の維持

Method of keeping model accuracy while reducing number of models by likelihood ratio test

津　田　高　治*　　今　泉　　忠**
Takaharu TSUDA　　Tadashi IMAIZUMI

Abstract : There are cases where news and its responses lead to debate and its shift explains socioeconomic changes.
Tsuda（2015）introduced a method to:
（1）numerically represent tendency of comments about chosen subject, and
（2）extract differential features of news and its quotation in social media.
In this paper, we propose an approach to:
（1）predict stock price using the differential features,
（2）keep predictive accuracy with reduced number of predictive models, and
（3）adopt as simple method as possible for predictive models.
So far many relevant studies have been carried out to predict stock price by statistical model utilizing publicly available text data such as those on internet, but while they have achieved predictive accuracy, they are not ready to use for real world purpose due primarily to lack of real-world applicability. In this paper we suggest an approach to apply likelihood ratio test to 2 periods of explanatory variables' differential features to identify significant change between the 2 periods so as to limit frequency of rebuilding of predictive model, and the approach is applied to publicly available text data of news and its responses regarding Honda motors as well as its stock price data. The result turns out that while limiting number of models, predictive accuracy keeps as good as those reported in past relevant studies. It is also numerically represented that the more number of models the more accurate prediction becomes.

Keywords：Canonical regression analysis, SVD, Text Mining, stock price prediction

* 多摩大学大学院経営情報学研究科　Graduate School of Business, Tama University
** 多摩大学経営情報学部　School of Management and Information Sciences, Tama University

1. はじめに

経済社会の多くの事象がその事象に関係する評価によって影響を受ける場合がある。私企業を例に取れば当該企業のニュースという評価情報、またニュースに対する投資家や消費者の評価が株価動向や製品売上に影響する場合があり、それがインターネットという誰もが自身の考えを表明する場が提供されたことでより容易に起こり得ている。

経済社会の事象として代表的なものに株価などの金融商品があり、その値動きをインターネット上の主にテキスト形式の評価情報で予測する試みが数多くなされ、先行研究として成果が挙がっている。和泉・後藤・松井（2011）は金融機関の発行した経済リポートを対象に共起解析、主成分分析、回帰分析のステップからなる金融テキスト・マイニング手法を活用することで、日本国債の変動の方向性に関する高い予測精度を実現している。また、Tetlock（2007）は Wall Street Journal column のテキストデータで Dow Jones Industrial Average などの指標の予測を実施している。しかし、これらの研究の焦点は予測の精度に当てられていて内容に関する記述、具体的にはテキストデータを複雑な金融商品の値動きを予測するに足るだけの特徴を備える情報に変換する過程や、またその情報が値動きにどのように関わっているのかなどの記述に関しては部分的であることが多い。本論文ではこれら内容に関する記述のことを「説明性」と呼ぶが、通常、市場全体や個別企業の数多くの情報に接して投資判断をしている投資家にとっては、値動き予測がいかに精確であっても説明性が十分でなければ投資判断の根拠とはなりえない。

テキスト形式の評価情報で株価など金融商品の値動きを予測するモデルのこれまでの取り組みでは「テキストデータによる金融商品の値動き予測で十分な精度が得られるか」が最大の関心事であったため先行研究の結果の報告も予測精度に重きを置いたものである場合が多かったが（和泉他、2011）（岡田 , 2014）（松井・石田・中嶋・和泉・吉田・中川、2013）、本論文ではモデルの「説明性」を高く保つことを第一に注意を払った上で実用可能なレベルに精度を高める方針を採った。また金融商品としては個別銘柄の株価を想定する。「説明性」を考慮し本論文の予測モデルでは以下の重回帰式（1）を採用する。

$$Y = X\beta + \varepsilon \tag{1}$$

式（1）の X は複数の説明変数、β は式（1）の説明変数に係る係数と切片、Y は株価を指す。重回帰式を採用したのはテキストから生成した説明変数（X）が株価（Y）にどのように関わっているのか説明することが比較的容易であるためである。また説明変数（X）をテキストから生成するにあたっては津田（2015）が提案した方法に則り「評価行列」を作成しその値を活用する方式を採る。「評価行列」とは財やサービスの供給側と需要側が存在する場合に適用可能な、供給側を評価するニュースやそのニュースに対する主に需要側のブログ・掲示版に於ける評価を一般的な適用性を考慮した固定的なカテゴリーに分類し行列とするものである。この固定的なカテゴリーは「評価行列」が主題とする財やサービスに係る 5W2H を表したもので解釈が容易なものになっている（津田、2015）。

株価（Y）と「評価行列」からなる説明変数（X）の関係は環境変化に応じて時間と共に変化し続ける。環境変化の例として、ある企業の業績が好調で企業として高評価を維持し株価も上昇傾向を維持していたにも関わらず、製品の品質問題やそれに対する消費者からの不満から評価も下がり、株価も下降してしまう事象が過去発生した（財経新聞、2011）。この際、株価

に関わっていた「評価」は上昇傾向の時点では企業の評判、下降期の時点では製品品質に関係するもので時点間の観点が異なっていることが考えられ、同一の株価と「評判」との関係がこの2時点間で有効であるとは考えにくい。また例えば同じ上昇傾向の時点間でも株価に関わる評価の観点・主観的評価（津田、2015）が時間の経過と共に変化することも考えられる。このことを考慮すると、株価（Y）と説明変数（X）の関係は例えば以下の式（2）のように考えることができる。

$$Y_{t+1} = X_t\beta_t + \varepsilon \tag{2}$$

予測モデルの「説明性」に関連してモデルの数の少なさに関しても考慮する必要がある。式（2）は毎時点モデルを作り変えることで変化への対応は出来るが、時点数と同じ数のモデルが作成されることになってしまう。モデルの数が多くなりすぎると投資家が全モデルの内容を読んで理解することが困難になってしまい1つ1つのモデルの「説明性」が高い場合でも投資判断を困難にする可能性がある。

モデルの数を少なく抑えることを考慮しつつ予測精度を維持するには各時点で新たにモデルを作る必要があるか否かを判断することが必要になる。その判断の基準に関しては次章以降で詳述するが、モデルの数を少なくすることを考慮した式は以下の（3-1）、（3-2）になる。

$$t^* = F(t) \tag{3-1}$$

F は表1で例示するステップ関数である。

$$Y_{t+1} = X_t\beta_{t*} + \varepsilon \tag{3-2}$$

式（3-1）の時点（t）と（t^*）の関係を模式的に例示すると以下の表1のようになる。

表1．時点（t）と（t^*）の関係

時点（t）	時点（t^*）
1	1
2	1
3	1
4	1
5	1
6	6
7	6
8	6
9	9
10	9

表1の例示に於いては、時点 t=1 でモデルが最初に作られ、時点 t^*=1 の係数として $\beta_{t*=1}$ が求められ、それが時点 t=5 まで有効なモデル係数として使われる。時点 t=6 に於いて $\beta_{t*=1}$ はもはや有効でないと判断され t=6 の時点で得られるデータを用いてモデルが作成されることで係数として $\beta_{t*=6}$ が求められ、それはその後有効でないと判断されるまで使われ続ける。表1の例示では毎時点でモデルを再作成する場合（モデル数は（t）の時点数と同数の10）に比較すると、モデル数は3で済み7モデル分内容を理解する負担を減らすことができる。表1の時点 t=6、時点 t=9 のことを本論文では「変化点」と呼ぶ。

本論文では「説明性」の高さを第一にしたモデル構築方針を採り、モデルの数を可能な限り少なくしながら予測精度を活用可能なレベルに維持することをテーマとして取り組む。その中で「変化点」の判断のために「評価行列」（津田 , 2015）の特徴を活かした新たな方法論を提案し、

その具体的な適用例と結果を提示する。

2.「評価行列」の生成と株価予測モデル式の構築・再構築

2.1 「評価行列」の生成

「評価行列」は、供給側の製品・サービスに主題を置き、関連ニュース記事やそれに対するインターネット上の SNS（Social Networking Services）に於ける評価記事を情報ソースやニュース記事の引用の有無、観点や主観的判断、時点別に件数を取りそれに基づき生成したものである。生成過程に関しては津田（2015）で提示された方法に則る。「評価行列」は記事を最小単位として、個々の記事の特徴、また観点・主観的評価・データソースなどでグループにした複数記事の特徴などを捉えることを第一義的な目的としている。

2.2 「評価行列」の 2 次元行列配置と説明変数（*X*）としての活用

「評価行列」は 7 データソース（情報ソースとニュース記事の引用の有無の組合せ）、6 観点、3 主観的評価、時点の 4 つの次元の配列であり、その個々の要素には減衰効果も加えた記事の件数が入っている（津田、2015）。この「評価行列」を式（3-2）の説明変数として活用するために以下のように配置の仕方を変える。「評価行列」は 4 つの次元の配列だが、これを各時点の特徴量として表現するために、時点を行として縦方向、それ以外の配列項目（7 データソース、6 観点、3 主観的評価）の組合せ 126 項目を列変数として横方向に展開する 2 次元行列とする（津田 , 2015）。

2.3 「評価行列」を基にした株価予測モデルの構築・再構築に係る仮説

前章で述べた「変化点」を捉えるために、「評価行列」の顕著な変化を捉えることを方法論として提案する。前章の例で示したように株価（*Y*）の上昇・下降の変化はその企業の評価のされ方の変化に関わっている場合があり（前章の例では上昇期では企業の評判、下降期では製品品質）、上昇から下降に転ずる際には「評価行列」の観点で示せば「企業の評判」の観点で「肯定」的な主観的評価から「製品品質」の観点で「否定」的な主観的評価へと「評価行列」の値の比重が移っていることが考えられる。また、その際には何が株価（*Y*）に相関関係の深い「評価行列」を基にした説明変数（*X*）であるかも、前述の値の比重と同様に変化している可能性が考えられる。この場合、後述のように説明変数（*X*）で株価（*Y*）を予測する予測モデルを再構築すべき「変化点」であると考える。このように説明変数（*X*）で株価（*Y*）の密接な関係を想定するためには「評価行列」が株価と相関が大きいことが前提となる。

「評価行列」に顕著な変化が見られる際に、株価（*Y*）と「評価行列」を基にした説明変数（*X*）との相関関係（株価（*Y*）を説明変数（*X*）で予測する予測モデル）には以下のことが可能性として考えられる。

(1)（*Y*）と（*X*）の相関関係には変化はない。説明変数（*X*）に顕著な値の変化があっても、その値を既存の予測モデルに与えることで正しく株価（*Y*）を予測できる

(2)（*Y*）と（*X*）の相関関係には変化がある。「評価行列」を基にした説明変数（*X*）の顕著な変化は株価（*Y*）と説明変数（*X*）の相関関係の変化を伴っており、既存の予測モデルでは正しく株価（*Y*）を予測できない

上記の（1）の場合は新たな予測モデルを再作成する意味はなく、（2）の場合に新たな予測モデルを再作成する意味が有ることになる。「評価行列」の変化を見ているだけでは上記（1）、（2）のどちらが当てはまるのかは分からないが、本論文では「評価行列」に顕著な変化が見られる際は（2）の可能性を含むため常に新たな予測モデルの再作成を実施することを提案する。上記（1）の場合では予測モデルの再作成は既存の予測モデルと大きく異ならないモデルを別途作ることになる。モデルの数の少なさはここで1モデル分損なわれ、予測の精度では得るところがない。（2）の場合はモデルの数の少なさは1モデル分損なわれるが、予測の精度の悪化を防ぐことができる。本論文では「評価行列」に顕著な変化が見られる際に常に新たな予測モデルの再作成を実施することでどれだけモデルの数の少なさを維持しつつ予測精度を保つことができるのか、次章以降で適用例と結果を示す。

次節のプリ・アナリシスで、株価と「評価行列」の間に相関があると認められた場合、以下を仮説として設定する。

- 「評価行列」を基にした説明変数（X）に顕著な変化が認められた場合に常にモデル再作成をすることでモデルの数を少なくしながら株価（Y）の予測精度を維持できる

説明変数側のみに着目してモデル再構築の判断をする点が本論文の独自性である。分析者の手許にある説明変数の直近のデータの変化を見ることでモデル再構築の判断をすることは、目的変数側の実測値と予測値から得られる予測精度を検証してからモデル再構築をするよりも精度が低い状態で予測をし続ける時点数を小さくできる可能性がある。

また、「説明変数（X）の顕著な変化」を把握する上で説明変数を「評価行列」から作成することは有利に作用する。「評価行列」のカテゴリーは観点の数を6、主観的評価の数を3と数も内容も固定化されテキスト・マイニングの分類処理がなされているため説明変数は複数時点をまたがって等質的であり、時点間の比較をより容易にしている。また「評価行列」の各カテゴリーは供給者と需要者が存在する主題であれば適用可能であるように設計され一般的な適用性を念頭に定義付けられている（津田 , 2015）ためカテゴリーが時間と共に陳腐化する可能性が低く、そのカテゴリーが基となって作られる説明変数が時点間を跨って等質の内容を保持することが可能で、その点でも時点間の比較をより容易にしている。

2.4 プリ・アナリシスとしての正準相関分析

前節で立てた仮説は説明変数（X）と株価（Y）との密接な関係を前提としており、その前提の検証のために、「評価行列」が株価と相関が十分に大きいかを検証する。

一般に「株価」の定義の中には最高値・最安値・始値・終値等々1つの変数だけで表現し切れない内容が含まれており、「評価行列」に由来する前述の126項目を126変数として捉えたデータセットと、「株価」を構成するデータセット、計2データセットの間の関係を見るために正準相関関係を実施する。この手順で2データセット間に相関があると判断された場合は次節のモデル構築以降の処理に入り、ないと判断された場合は以降の処理を中止するという判断をする。

2.5 モデル作成・再作成・スコアリング・尤度比のステップ

以下の図1にモデル作成･再作成･スコアリング･尤度比のステップを示している。[a] ステップとして予測モデル作成･再作成･スコアリング、[b] ステップとしてスコアリング、[x] ステップとして次節で詳述する尤度比を用意し、全体を構成している。

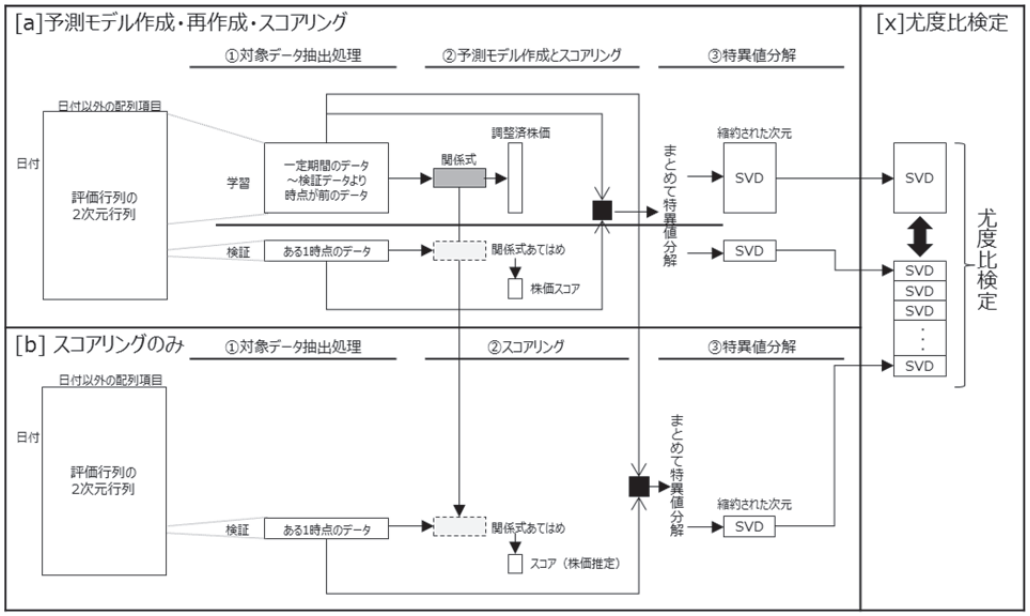

図 1. モデル作成・再作成とスコアリング、尤度比のステップ

[a] ステップに於いては株価（Y）を前述の「評価行列」に由来する 126 説明変数（X）で説明する線形回帰モデルである式（3）を構築する。[a] ステップで実行される時点を（$t=a$）とすると、それより前の一定期間分の株価（Y）と説明変数（X）からなるデータを学習データとし式（3）の作成に使う。式（3）が求まったら [a] ステップ時点の説明変数のデータを $X_{t=a}$ としてモデルに適用して $(X_{t=a}\beta_{t*=a})$ スコアを生成する。また、[b] ステップではその時点のデータ $X_{t=b}$ に [a] ステップで構築済みのモデルを適用する $(X_{t=a}\beta_{t*=a})$ ことでスコアを生成する。[x] ステップの尤度比については次節で詳述するが、その目的は 2.3 節で述べた「説明変数（X）の顕著な変化」を把握することである。ここでは 2.2 節で紹介した 126 個の行列項目全てを尤度比に用いるのではなく、特異値分解でそれを次元縮約した特異ベクトルから主要なものを用いる。特異値分解を実施する際の考慮点は、特異値分解の等質性を確保することである。具体的には [a] ステップ時点の「評価行列」を 2 次元行列にしたものを特異値分解したものと、[b] ステップ時点の同様の特異値分解が同じ基準で為されていること、また複数の [b] ステップ時点間での等質性である。図 1 に示す通り、[a] ステップに於いて、モデル構築に使う（$t=a$）時点より前の一定期間分のモデル作成用の学習用データとその時点（$t=a$）1 時点分のデータとを 1 つのデータセットにした上でまとめて特異値分解をし、[b] ステップではその時点（$t=b$）の 1 時点分のデータと [a] ステップで使った一定期間分のモデル作成用のデータとを 1 つのデータセットにした上でまとめて特異値分解を実施する。このことで [a] に於ける特異値分解と [b] での特異値分解とで等質的な基準で特異ベクトルを作ることができる。このことは [b] ステップを何度か繰り返す時にも同様である。

各ステップの実行順序であるが、[a][b] の各ステップの完了後に [x] ステップを実施して「説明変数（X）の顕著な変化」が認められれば次ステップは [a]、そうでなければ次ステップは [b] を選択する。尤度比では前述の学習データの説明変数（X）を特異値分解して得られた特異ベクトルと、各ステップでスコア生成の為に使った説明変数（上述の $X_{t=a}$ や $X_{t=b}$ が例）の特異ベクトルを比較することになる。これは 2.3 節で述べた、モデル式再作成を必要とする基準を「評

価行列」を基にした説明変数（X）に顕著な変化が認められるか否かに求めることができるという仮説に従っている。[x] ステップ完了後､時点（t）を 1 進めて [a] か [b] の選択されたステップに進む。

2.6 尤度比によるモデル再構築の基準設定

「評価行列」の顕著な変化を、尤度比によって検知する。一般的に尤度比とは 2 群のデータが与えられたときにモデルに関して帰無仮説と対立仮説を

$$H_0: \mu_0 = \mu_1 \quad (= \mu \text{ と置く})$$
$$H_1: \mu_0 \neq \mu_1$$

とし、各仮説のもとでのσ^2の推定をそれぞれ$\widehat{\sigma_0^2}, \widehat{\sigma_1^2}$と置いたときの尤度関数の比である$\frac{L^0\left(\hat{\mu},\widehat{\sigma_0^2}|x\right)}{L^1\left(\widehat{\mu_1},\widehat{\mu_2},\widehat{\sigma_1^2}|x\right)}$を以下の形にした際に漸近的に$\chi^2$分布に従う。自由度を m、信頼区間を α とした際のχ^2分布を用いて$T > \chi^2(m, 1-\alpha)$のときに平均値が異なることを受容する。

$$T = -2\log\left(\frac{L^0\left(\hat{\mu},\widehat{\sigma_0^2}|x\right)}{L^1\left(\widehat{\mu_1},\widehat{\mu_2},\widehat{\sigma_1^2}|x\right)}\right) \tag{4}$$

本論文に於いては上記の μ_0 は [a] ステップの一定期間分の学習データの説明変数（X）を特異値分解して得られた特異ベクトルの主要なものの中から選ばれたある 1 つの特異ベクトルの値の学習データの一定期間分の時間を跨った平均、μ_1 は各 [a][b] ステップでスコアリングに使った説明変数（上述の $X_{t=a}$ や $X_{t=b}$ が例）の特異ベクトルの中で μ_0 と対応する同じ特異ベクトルの時間を跨った平均である。また、$\widehat{\sigma_0^2}$は学習データ由来の上記の特異ベクトルと各 [a][b] ステップでスコアリングに使った説明変数（上述の $X_{t=a}$ や $X_{t=b}$ が例）に対応する特異ベクトルとの間に差異はないという仮説のもとで計算された σ^2 の推定値で、$\widehat{\sigma_1^2}$ はそれに差異があるという仮説のもとで計算されたσ^2の推定値である。

尤度比を今回の目的､「評価行列」から作った学習データの説明変数と最新の説明変数の間に顕著な差異が認められるか見ること､に使うことに関しての考慮点を述べる。（t）時点の「評価行列」の配列要素は（*t-1*）時点から定義された時点数 d まで過去時点（*t-d*）の同一配列要素の件数を減衰効果の考慮をしながら合計した値を算出し、その上で時点毎に各データソースの合計値で割った比率の値が入っていて、それがそのまま説明変数の値になっている。尤度比の μ_0 の基となる期間とμ_1の基となる期間の間隔が d 以下である場合は、その 2 期間のデータは独立ではなく、尤度比の独立性の前提を満たさない。また、上記が d より大きい場合であっても「評価行列」はインターネットなど一般的に利用可能なテキストデータによる評価に基づくもので過去の「評価」が d 時点数を超えてのちの「評価」に影響を与えることも考えられる。従って、2 期間のデータの間隔 d の定義に関わらず独立性の前提が担保されるかは不確定である。このことにより尤度比を統計的な差異の検定のための統計量としては使うことが出来ないが、本論文では後述のようにそれをデータの変化度合いを指し示す指標として取り扱う。

尤度比の対象として「評価行列」の 126 項目をそのまま使うのでなく､「評価行列」を特異値分解した特異ベクトルの中で主要なものを用いるのは、主要な特異ベクトルは「評価行列」全体の主要な特徴を表しており、126 全体を対象にするのに比較してより簡潔に尤度比による比較を可能にするからである。また、主要な特異ベクトルを 2 期間のデータから取り出し尤度

比を求め、T が大きいかどうか検証する。T が大きな特異ベクトルの個数が対象とした主要な特異ベクトル総数に比較し一定以上の割合である際に「評価行列」に顕著な変化があるとする。

2.7　尤度比の 2 期間のデータの選び方

尤度比計算をする目的は直近の「評価行列」から作った説明変数がモデル式を構築する際に使ったものと比較してどの程度変化しているか見ることである。従って、1 群は「直近」の何時点かの [a][b] ステップでスコアリングに使った説明変数（上述の $X_{t=a}$ や $X_{t=b}$ が例）を特異値分解したもの、もう 1 群は [a] ステップでの学習データとして使った時点の説明変数を特異値分解したものである。

「直近」とは将来の株価予測に直接関係すると思われる、[x] ステップを実施する時点から何時点か遡った時点までの一定期間を指す。遡る時点数をどうすれば「直近」というものにふさわしいかは通常のビジネス上の慣行から求める方法、ないしは後述する本論文の適用例で示すデータから求める方法が考えられる。

2.8　評価

本論文ではモデルの数を可能な限り少なくしながら精度を一定レベルに維持することを提案している。したがって評価として、モデルの数の少なさの評価を精度評価と共に実施する。モデルの数の少なさは、[a] ステップ（予測モデル作成･再作成･スコアリング）と [b] ステップ（スコアリング）の回数の合計に比較した [b] ステップの比率、つまり全体の時点数に対するモデルを作成しなかった時点数を比率として取ることで指標とする。精度評価に関しては、次時点の騰落予想が実際の騰落と一致した比率を取る。モデルの数の少なさと精度は、モデルの数の少なさが高まれば精度が低まるという背反の関係になることが多いと予想される。

3.　適用例

3.1 「評価行列」の基となるテキストデータの獲得

「評価行列」の基となるテキストデータは本田技研工業に対して記述されたニュースやブログサイト、掲示板サイトの記事である。データの獲得元は津田(2015)で論じたものと同一だが、再掲する。当企業はハイブリッド・カーを世界で始めて市場に供給した先進企業として、先んじた技術開発や品質問題、需要者側のメリットやコスト、トヨタ自動車と比較した販売状況や企業活動など多岐に渡り評価をされており、適用例として最適である。

表 2. データの獲得源の代表例

サイトの種類	サイト名称
ニュース	朝日新聞、毎日新聞、日本経済新聞、MSN 産経ニュース、YOMIURI ONLINE 等
ブログ	アメブロ、ヤプログ、ライブドアブログ、はてなブログ、ヤフーブログ、ツイッター、みんなのカーライフ ブログ等
掲示板	2 チャンネル

データを獲得する際に用いたツールは Effyis Inc. の Boardreader.com で、使ったキーワードは「ホンダ」、「本田技研」、「本田」、「Honda」であった。結果、2010 年から 2012 年 6 月までのデータが 87,870 件得られた。

3.2 「評価行列」の作成

「評価行列」の作成方法の詳細に関しては津田（2015）で説明した方法と同一なので再掲しないが、その過程で各記事の観点や主観的評価が以下のような分布で分類された。月あたりの記事件数が安定して多い 2011 年 7 月 1 日から 2012 年 6 月 30 日までに時期を絞った 51,088 件のうちの記事件数となる。

表 3. 観点の分類

観点の分類	記事件数
企業の評判（供給主体）	10,334
車の品質問題（供給物）	16,760
販売活動（供給主体の需要主体への働きかけ）	4,473
車の価格や燃費（需要主体の担うコスト）	12,429
車の使い勝手（需要主体のベネフィット）	5,806
その他	1,286

記事とは 1 単位のテキストのことだが、具体的にはひとつのニュース記事、ひとつのブログエントリー、掲示板では 1 スレッドの中の複数の書き込みの中の 1 つを 1 単位とし、それぞれを記事と呼ぶ。

表 4. 主観的評価の分類

観点の分類	記事件数
肯定	31,101
中立	14,928
否定	5,059

これらの分類の精度測定の目的で、2012 年 1 月 1 日のデータ 99 件の Naïve-Bayes ロジックによる分類済みのデータを人間の目で誤判別率チェックを実施した結果、観点は 21%（主題と無関係な 23 記事を取り除いた 76 記事中に誤判別された記事が 16 件）、主観的評価は 18.2%（99 記事全てを対象として 18 件の誤判別）であった。

またデータソースは（津田、2015）と全く同様にデータソースはニュース、ニュースを引用するブログ（車関係)、ニュースを引用するブログ（車以外)、ニュース引用のないブログ、ニュースを引用する掲示板（車関係)、ニュースを引用する掲示板（車以外)、ニュース引用のない掲示板とし、また時点は日単位とした。

3.3 目的変数の元データの獲得

目的変数の元データとして時系列の株式データを獲得することとし、具体的には本田技研工業㈱の個別銘柄の日次データを Yahoo! Japan ファイナンスのサイトからデータ獲得をした。本田技研工業に関する「評価行列」によって説明する対象は、同じ企業の個別銘柄とするのが自然である。株式の時系列データとしては日次で始値・高値・安値・終値・出来高・調整後終値が上述のサイトによって提供されている。

3.4 正準相関分析による「評価行列」と株価時系列データとの相関分析

正準相関分析で相関分析をするのは前節の株式時系列データと「評価行列」を基とした説明変数である。株式時系列データには始値・高値・安値・終値・出来高・調整後終値の 6 項目があるが、終値と調整後終値は同一の値が入ることが多く、調整後終値の方が株式分割の際にも使えるなど用途が幅広いため、終値は除き 5 項目とした。また「評価行列」を基とした説明変数に関しては 2.2 節で記述したように 2 次元行列配置をした 126 項目を指すが、その全てを対象とするのではなく、7 つのデータソース別にそれぞれ 18 項目（6 観点× 3 主観的評価）を取得して株式時系列データとの正準相関分析を実施、これを 7 つのデータソースそれぞれに実施した。例としてニュースのデータソースの「評価行列」から得た 18 変数と株式時系列データの 5 変数の 2 データセットの間で正準相関分析を実施した結果を示す。表 5 の 1 行目の「Pr>F」の値が 0.05 未満であることから 5 つの正準相関全てが 0 であるという帰無仮説が棄却され、2 行目の「Pr>F」の値が 0.05 未満であることから 2 番目以降の 5 番目までの正準相関全てが 0 であるという帰無仮説が棄却されている。ここから上位 2 つの正準相関が 0 という帰無仮説は棄却される。また表 6 に於いては株式データの 5 項目とニュースをデータソースにした「評価行列」18 項目の間の正準相関がすべて 0 であるという帰無仮説を 4 つの統計量すべてに於いて P 値が小さいことから棄却されることを示している。

表 5. 正準相関分析

	正準相関	調整正準相関	近似標準誤差	平方正準相関	固有値とその比率				H0: その行以降における正準相関係数がすべて 0				
					固有値	差	比率	累積	尤度比	近似 F 値	分子の自由度	分母の自由度	Pr>F
1	0.868	0.847	0.022	0.754	3.062	2.152	0.697	0.697	0.088	4.43	75	502.36	<.0001
2	0.690	0.641	0.047	0.476	0.910	0.657	0.207	0.905	0.357	2.22	56	410.6	<.0001
3	0.450	0.342	0.072	0.202	0.253	0.145	0.058	0.963	0.681	1.12	39	314.64	0.3001
4	0.313	0.169	0.081	0.098	0.108	0.052	0.025	0.987	0.854	0.73	24	214	0.8156
5	0.231	0.114	0.085	0.053	0.057		0.013	1.000	0.947	0.55	11	108	0.8613

表 6. 正準相関分析の多変量統計量

統計量	値	F 値	分子の自由度	分母の自由度	Pr>F
Wilks のラムダ	0.088	4.43	75	502.36	<0.001
Pillai のトレース	1.584	3.34	75	540	<0.001
Hotteling-Lawley のトレース	4.390	6.00	75	361.41	<0.001
Roy の最大根	3.062	22.04	75	108	<0.001

当適用例のデータに於いて 7 つすべてのデータソースに対して同様に分析をした結果、表 6 と同様に 4 つの統計量の P 値が小さいことが 7 つすべてに当てはまり、2 データセットはデータソース別に分析した場合相関がある。またこのことは「評価行列」を基とした説明変数 126 項目全てと株式データの 5 変数の 2 データセットの間でも同様である。

プリ・アナリシスの結果 2 データセットは相関があるという条件が満たされたので 2.3 節で述べた仮説の検証として次節以降に進む。

3.5 予測モデルの構築と再構築の処理

式（3）のモデルを構築するにあたって目的変数（Y）は 3.4 節の調整後終値とし、説明変数は「評価行列」に由来する 126 変数とした。変数選択は 95% 信頼区間のステップワイズによる自動選択とし、それに基づいてモデル式を構築し、そのモデル式を [a] ステップで用意したスコアリングデータに適用 $(X_{t=a}\beta_{t*=a})$ してスコアを生成、[b] ステップではその時点で用意したスコアリングデータに [a] ステップで構築済みのモデルを適用 $(X_{t=b}\beta_{t*=a})$ することでスコアを生成した。また学習データは [a] ステップの時点（$t=a$）以前の 6 ヶ月とした。

また [x] ステップの尤度比に於いて、主要な特異ベクトルの選び方としては特異値の大きいものに対応した特異ベクトルから順に一定数の特異ベクトルを選ぶが、今回の適用例に於いては上位 20 の特異ベクトルを選んだ。上位 20 の特異ベクトルに対応する特異値合計はすべての [x] ステップに共通して全体の特異値合計の 80% 以上であることから、全体の傾向を捉えるには十分であると判断した。

全体のステップで用意したデータは 2011 年 7 月から 2012 年 6 月まで、予測期間として用意したデータは 2012 年 1 月から 6 月迄で、6 月末までの予測が完了した時点で処理を終了した。

3.6 尤度比の対象である「直近」のデータの捉え方

ここでは前節で記述した [a] ステップ（予測モデル作成･再作成･スコアリング）と [b] ステップ（スコアリング）を選択する基準としての [x] ステップ（尤度比）の詳細について記述する。

2.6 節でその中の 1 群は [a] ステップでの学習データの説明変数であるが、もう 1 群はいわゆる「直近」のデータの説明変数であると述べた。ここでの問題は「直近」とは [x] ステップの時点（$t=x$）から何時点遡った時点までをそう呼ぶのかである。具体的には [x] ステップの時点（$t=x$）から遡る日数を n^* として、n^* をどのように求めるかであるが、この適用例ではデータからそれを求める方法を採用しその内容と結果を紹介する。

この設問に対するアプローチとして以下のステップを取った。目的変数の時点を（t）時点とするとそれに対しそれぞれ 1 時点前の（$t\text{-}1$）時点の説明変数で予測モデルを作り予測株価と実際の株価の比較をする。これを毎時点繰り返して全時点の予測株価と実際の株価から予測

精度を数値化する。加えて同様に（t）時点の目的変数に対して（t-2）時点の説明変数を使った予測モデルを作り全時点繰り返すことで予測精度を取得する。以降、（t-3）、…、（t-n）時点の説明変数で（t）時点の目的変数を予測するモデルで同様に精度を取得する。精度の指標としては翌日の株価の騰落予想が実際の騰落と一致する件数を全体件数で割った騰落予想一致率を取ることにした。n の値に対して予測精度がどのように変化するかグラフを描き、n が 1 増える際に顕著に予測精度が落ちる時点を捉え、顕著に精度が落ちる 1 つ前の n を n^* とする。（t-1）から（t-n^*）までの情報が（t）時点を説明するのに有効であることから、現在と関わりの深い「直近」とは n^* 日前までの情報と考える。

なお、以上の n^* を求める処理には 2011 年データのみを使った。その結果が図 2 である。（t-1）時点の説明変数では精度が低いが、概ね（t-5）までは精度が一定レベルを維持できており、（t-5）から（t-6）で 9% も低まりそれ以降は概ね精度が低まる傾向にある。ここからこのデータに於ける n^* は 5 であると判断した。5 日前までを直近とすること、またそれに基づいた尤度比を今回の目的に活用することは通常のビジネスの感覚からも違和感なく受容可能であると判断した。

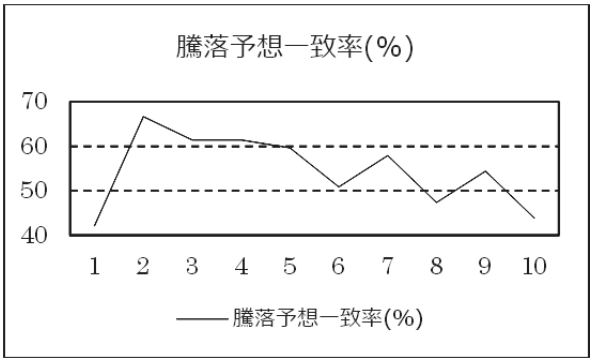

図 2.（t-n）時点の説明変数を使った株価予測に於ける n と騰落一致率の関係

3.7 尤度比による判定

尤度比の対象の変数は特異ベクトルの上位 20、つまり特異値の大きさの上位 20 に対応する特異ベクトルを選択した。個々の特異ベクトルで比較する 2 群間で違いがないとする帰無仮説が棄却されない場合は 1、棄却される場合は 0 としてそれが全体の 20 の特異ベクトルに対する割合（p）を算出する。上位 20 の特異ベクトルの中の 1 つを（i）とすると、（p）は以下の（5）（6）式で算出される。

$$v_i = \begin{cases} 0, & (H_1: \mu_0 \neq \mu_1 \text{のとき}) \\ 1, & (H_0: \mu_0 = \mu_1 \text{のとき}) \end{cases} \tag{5}$$

$$p = \frac{1}{20}\sum_{i=1}^{20} v_i \tag{6}$$

また、しきい値の p^* を設定し、p の値の p^* との比較で以下のように、[x] ステップの尤度比の次に選択されるステップを選択することと設定した。

（[a] ステップ、($p < p^*$のとき)

（[b] ステップ、($p \geq p^*$のとき)

この適用例を実施するに当たって、p^* の値を高く設定した場合は [a] ステップを実施する頻

度が高まり p^*=1 とした場合は [a] ステップが毎期の 1 期先予測に近づき、予測精度は高まるがモデルの数の少なさは損なわれると予想される。また、p^* を低く設定した場合は [a] ステップを実施する頻度が低まり p^*=0 にすると [a] ステップは最初の 1 回のみの実施となり予測精度が低まると予想されるがモデルの数の少なさを維持できる。次章では、p^* を変化させることで予測の精度やモデルの数の少なさにどのような変化が見られるか、仮説が今回の適用例で支持されるかなどを具体的に考察していく。

4. 分析結果

4.1 「評価行列」閾値と関係式構築の回数

前節の「閾値」p^* の値が大きくなるに従って [a] ステップのモデル再作成を実施する回数は多く [b] ステップのスコアリングの回数は少なくなることが予想されるが、時点ごとに観察される変化にモデルを適合させる機会が多くなるため予測精度は全体として高まると予想できる。実際のデータ分析の結果もその通りであった。以下、p^*=0.6 から p^*=0.9 まで 0.1 刻みで変化させた上記 2 指標の変化を図 3 に示す。横軸で p^*=0.6 から p^*=0.9 まで 0.1 刻みで変化させ、左の縦軸で騰落一致率、右の縦軸で [b] ステップによるスコアリング頻度の率（モデルの数の少なさ）を示す。p^* が増えるほど実線で示される精度の曲線は上昇し、点線で示される [b] ステップによるスコアリング頻度の率は低まることが見て取れる。

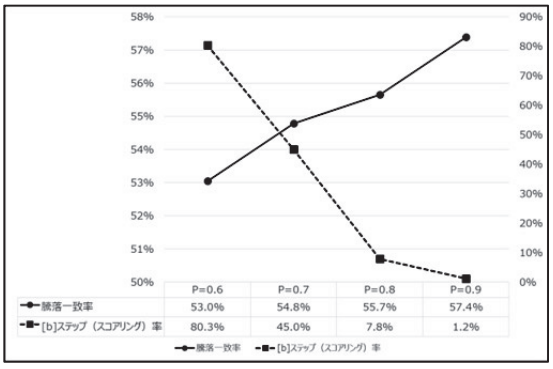

図 3.「評価行列」閾値と騰落一致率・[b] ステップによるスコアリング頻度の率

本論文の適用例としては株価時系列データの中の調整後終値を目的変数として予測モデルを構築しており、相関係数や平均絶対誤差率など一般的に使われる指標ではなく騰落一致率を精度の指標として採用することにした。モデルによる次時点（*t+1*）の株価予想値を現時点（*t*）実測値と比較して大小を算出、それを次時点（*t+1*）と現時点（*t*）の実測値から算出した大小と比較し、一致している場合は 1、そうでない場合は 0 として合計しそれを全時点数で割ることで算出する。また、今日まで出てきたモデルの多くが一致率 57-58% であること（和泉他、2013）（松井、2013）から、本論文で採用している「評価行列」を基にした予測モデルの精度も p^*=0.9 の場合の騰落一致率 57.4% はそれと同等であると言える。

モデル再作成の頻度を抑えてモデルの数の少なさを維持しつつ予測精度を一定に保つこと、再作成の契機を説明変数の直近の値の学習データからの差異に求めることが本論文のテーマであるが、p^*=0.6 に於ける結果を見ると 2012 年 1 月から 6 月までの全 167 日の取引市場の営業

日のうち 134 日は [b] ステップによるスコアリングのみで、スコアリングする日数が 2 日でしかない p^* =0.9 での場合に比較すれば、モデルの数の少なさは高い。また、p^* =0.6 に於ける騰落一致率は53%という数字であり、52.3%の騰落一致率で投資利益が得られるとした事例（岡田, 2014）に比較しても、モデルの精度としては充分である。ここから、p^* =0.6 の設定で尤度比を実施する場合に、「評価行列」を基にした説明変数（X）に顕著な変化が認められた場合に常にモデル再作成をすることでモデルの数を少なくしながら株価（Y）の予測精度を維持できるという仮説が支持されたと評価する。

また、参考までに再作成の契機を目的変数の予測と実測の値の乖離に求めた結果を示す。上述の「乖離」は平均絶対パーセント誤差を指標として採用し、その許容値を設けて過去 5 日分の値が許容値と同じかより低ければ既存の予測モデルによるスコアリング、許容値より大きければモデル再作成とし、許容値の変動が予測精度とモデルの数の少なさにどう影響するか評価した。過去 5 日間としたのは 3.6 節で採用したのと同様に直近の値を 5 日間と設定して可能な限り同じ条件で比較をする意図からである。精度は前述の騰落一致率、モデルの数の少なさは構築済みのモデルを使ったスコアリングにより予測値を出した日数の全体日数に対する比率で表現した。

表 7. 平均絶対パーセント誤差の許容値と騰落一致率、モデルの数の少なさの関係

	再モデル構築の判断となる平均絶対パーセント誤差の許容値上限				
	9%	7%	5%	3%	1%
騰落（一致率）	51.0%	54.9%	47.1%	54.9%	54.9%
スコアリング（構築済モデル使用）率	94.2%	69.2%	63.5%	36.5%	5.8%

許容値の上限を小さくするほど、モデル再作成の頻度が高まりモデルの数の少なさは損なわれるが、騰落一致率で表される精度は図 3 ほど明確に向上していない。目的変数側の乖離を捉えてからモデル再構築をすることは、説明変数の直近のデータの変化を見ることに比較して精度が低い状態で予測をする時点数がより多くなりうることが本適用例に於いて再作成の頻度を高めても図 3 ほどには精度があがらなかった原因の一つと考えられる。

4.2 株価予測線形回帰モデルの例

以下の表 8 に 2012 年 1 月 16 時点のモデル式を例示する。2012 年 1 月 16 日は株価が上り調子になる好調期である。説明変数の係数は小数点以下を切り捨てて表示している。「評価行列」が基となる各変数には各データソース別の合計値で割り込んだ率を値として持っており、率の値が小さい変数である場合はその係数は率が小さい分大きくなる傾向がある。したがって係数の大きさは変数の重要性を直ちに意味しない。

表 8. 2012 年 1 月 16 時点で構築された株価予測線形回帰モデル

変数番号	説明変数	係数
1	切片	2187
2	ニュース　企業の評判　肯定	1247
3	ニュース　車の品質問題　肯定	855
4	ニュース　販売活動　否定	2446
5	ブログ　ニュース引用（車関係）　車の使い勝手　肯定	302
6	ブログ　ニュース引用（車関係）　車の価格や燃費　否定	286
7	ブログ　ニュース引用（車関係）　車の使い勝手　否定	210
8	ブログ　ニュース引用（車以外）　車の品質問題　肯定	292
9	ブログ　ニュース引用（車以外）　車の使い勝手　肯定	-786
10	ブログ　ニュース引用（車以外）　販売活動　中立	-738
11	ブログ　ニュース引用（車以外）　企業の評判　否定	204
12	ブログ　引用なし　企業の評判　肯定	-1709
13	ブログ　引用なし　車の品質問題　中立	3133
14	ブログ　引用なし　販売活動　中立	8350
15	ブログ　引用なし　車の価格や燃費　否定	-10942
16	ブログ　引用なし　その他　否定	228315
17	掲示版　ニュース引用（車関係）　企業の評判　中立	1786
18	掲示版　ニュース引用（車関係）　企業の評判　否定	-314
19	掲示版　ニュース引用（車関係）　車の価格や燃費　否定	1095
20	掲示版　ニュース引用（車関係）　販売活動　否定	-1162
21	掲示版　引用なし　企業の評判　肯定	-2371
22	掲示版　引用なし　車の価格や燃費　中立	-3627

変数番号 2 と 3 に該当するニュース記事を確認したところ、コンパクトカーの市場予測としてハイブリッド車が主流になるとの経営者としての見方や、またフリードなどのハイブリッドモデルの投入やその技術などが好感を持って受け止められている様子が観察できた。また変数番号 4 に関連して、当社が販売面で苦戦している否定的なニュースがモデル式において株価上昇の方向に関係している点に関してだが、これら記事群に於いてはホンダ車の苦戦はありつつも記事全体としてはハイブリッド・コンパクトカーという当社が得意とする市場の盛り上がりを伝える内容となっており、その点がプラスに作用したと考えることができる。

説明変数とモデル係数を見ると、主観的評価が肯定的だが株価モデルの係数がマイナスであるなど主観的評価（好評・不評）とその係数の符号（プラス・マイナス）が整合的でない場合が見られる。2.1 節で述べたように「評価行列」は記事を最小単位として主観的評価の分析をしており、例えば記事の題材が主題と直接関係ないものを主題と平行して扱っている場合、それらに対する主観的評価をも総合して記事全体の主観的評価を決定する。この場合、記事全体の主観的評価が主題（本論文の適用例では本田技研工業）に対する主観的評価と異なる可能性があり、この点が主観的評価とその係数の符号との不整合となって表れる背景を為すと考えられる。

また、p^* =0.6 に於いて再構築すべきと判断されたときにモデルにどのような変化が見られたかに言及する。再構築されたモデルにはその直前のモデルと比較して新しく追加された変数、

削除された変数、継続で使い続けられた変数の 3 種類があるが、全期間を通じてのそれぞれの種類の個数の平均は追加の変数が 6.7 個、削除が 6.7 個、継続が 10.8 個であった。追加と削除の和の全体に対する比率は 55.5% で、この値は p^* =0.9 に於ける 51.8% と比較して全体的なモデルの数を少なくした分 1 回当たりの内容の変化が大きい様子を示している。

モデル式全体を通して、説明変数の観点として最も数多く選ばれているのが「企業の評判」であり、該当する記事を確認したところ、当時の円高等を受けてカナダ向けのフィットの生産を中国工場にシフトしたこと、タイ工場の浸水で大量の完成品を廃棄したこと、N BOX などの新車種市場投入、ライバル車アクアの登場などがニュース報道としてなされブログや掲示板がそれに反応していることを確認できた。

5. まとめ

本論文に於いて、本田技研工業に関する「評価行列」を説明変数として、当該企業の調整後終値を予測する試みを通じて、「評価行列」から作られた説明変数の値変化の顕著さを尤度比で捉えモデル再作成の契機とすることでモデルの数の少なさを維持しつつ一定の実用可能な予測精度を達成することができた。「評価行列」は一般的な適用性を考慮して作られた固定的な行列項目であることから時点間を跨って同一の内容を保持しやすく陳腐化しにくいという特徴を持ち、この特徴が説明変数側の複数の時点間の尤度比による判定をより容易にした。また内容的な概要を類推しやすい「評価行列」の行列項目を元とする説明変数からなるモデル式の内容は理解可能なもので、本論文で提案する式（3）のモデルは実用可能な予測精度を維持しつつ「説明性」を高く保つことができたと考える。

参考文献

和泉潔・後藤卓・松井藤五郎（2011）。経済テキスト情報を用いた長期的な市場動向推定。情報処理学会論文誌、52, 3309-3315

松井藤五郎・石田智也・中嶋啓浩・和泉潔・吉田稔・中川裕志（2013）。新聞記事を対象とした時系列テキスト分析による市場予測。人工知能学会研究会資料、SIG-FIN-007-08、44-47

岡田克彦（2014）。ビッグ・データで株価を読む。中央経済社。

Tetlock, P. C., (2007). Giving content to investor sentiment: The role of media in the stock market. *The journal of finance, 62*, 1139-1168

津田高治（2015）。情報ソースの種別を考慮した記事の特徴分析。データ分析の理論と応用、5, 49-66

財経新聞（2011）。花王が急落、増額なかったことで「売方」が一気に攻勢、一段安も。