金融情報学研究会（第20回）

日時　2018 年 3 月 20 日（火）

会場　東京証券取引所 2F 東証ホール

SIG-FiN
JSAI Special Interest Group on Financial Informatics

人工知能学会
金融情報学研究会

第20回 人工知能学会 金融情報学研究会(SIG-FIN)

2018 年 3 月 20 日(火) 東京証券取引所 2F 東証ホール

ダークネット観測情報を用いた 仮想通貨市場におけるリスクの考察 ～仮想通貨市場におけるオルタナティブ・データの活用～

Evaluation of Risk in Crypto Currency With Darknet Observation Information

中川　慧 [1,2*]　今村　光良 [1,3]　面 和成 [4]
Kei Nakagawa[1,2] Mitusyoshi Imamura[1,3] Kazumasa Omote[4]

[1] 野村アセットマネジメント株式会社
[1] Nomura Asset Management Ltd.
[2] 筑波大学 大学院 ビジネス科学研究科
[2] University of Tsukuba Graduate School of Business Sciences
[3] 筑波大学 大学院 システム情報工学研究科
[3] University of Tsukuba Graduate School of Systems and Information Engineering
[4] 筑波大学 システム情報系
[4] University of Tsukuba Faculty of Engineering, Information and Systems

Abstract: Bitcoin is a crypto currency that is a peer-to-peer(P2P) network systems based on distributed ledger technology and is being used as an alternative payment system. Reliability and safety are very important aspects of payment system. However, in recent years, with an increase in value, crypto currency becomes a target of a malicious users and the attacks that strike the vulnerability of the system are regarded as a problem. Such a problem significantly reduces reliability and safety as a payment system for crypto currency. Therefore, it is necessary to pay attention to cyber security risks inherent in the system, as well as price fluctuations usually focused on financial asset prices. In this research, we propose to use information observed in darknet as alternative data. It is useful in evaluating the risk in the crypto currency market. The darknet is a name of an IP address space unallocated by terminals or the like among spaces that can be assigned IP addresses. The darknet is mainly used for observing signs of security incidents. This is useful for investors to grasp potential risks of crypto currency markets and is important for service providers to explain security risks and measures. In addition, the darknet observation information has a prospect of utilizing not only the crypto currency but also the monitoring of the security risk of companys.

1　はじめに

近年、世界規模で急速に利用者数が拡大し、注目を集めているのが Bitcoin を代表とする仮想通貨である。仮想通貨とは、ウェブ上に投稿された論文 [20] を基に、有志により開発がすすめられている P2P 型の分散台帳 (ブロックチェーン) を用いたシステムを指す。台帳に記録される数値が金融資産として取引され、取引手数料が安価であり、決済代替手段としての活用がすすめられたことから「通貨」と呼ばれている。

仮想通貨はその市場規模が大きくなるにつれて、投資対象として投資家からの関心が高まり、学術方面では、金融資産としての分析が活発に行われるようになった。例えば、伝統的な資産である株や債券などに対して用いられる、時系列解析に基づく分析がある。時系列解析のモデルには、条件付き平均モデル、条件付き分散モデルがある。条件付き平均モデルの代表例としては、AR モデル、MA モデル、ARMA モデルがあり、

*連絡先： 野村アセットマネジメント株式会社
〒 103-0027 東京都中央区日本橋 1 丁目 12-1
E-mail: kei.nak.0315@gmail.com

これらは株価の水準あるいは収益率のモデリングおよび予測に用いられる。AR モデルは過去の株価の線形結合で将来の株価を予測するモデルであり、MA モデルは過去の株価の攪乱項の線形結合で将来の株価を予測するモデルである。ARMA モデルは AR と MA の両者を組み合わせたモデルである。一方で、条件付き分散モデルには、ARCH モデルや、ARCH をさらに一般化した GARCH モデルが提案されている [12]。

Bitcoin については、条件付き分散モデルを用いて、ボラティリティの分析や、ヘッジ手段としての有効性など分析した先行研究がある [7, 8, 16]。

時系列解析以外には、Bitcoin に関連するニュースの発信および拡散に用いられるソーシャルメディアを代表とした web 上から得られるデータをクロスセクショナルな特徴量として、Bitcoin 価格との関係性について調査した先行研究 [11, 19] もある。

その他、Bitcoin を伝統的資産と組み合わせた場合における分散効果について調査した先行研究 [4] などもあり、その関心の高さが伺える。

一方で、Bitcoin に対する関心は、投資対象である金融資産としてだけはなく、基盤技術であるブロックチェーンを用いた「システム」としての側面に注目が集まっている。そのため、Bitcoin のシステムとしての研究開発調査が拡大しており、学術方面においては、システムにおける動作プロトコルのメカニズムに焦点が当てられ、体系的な研究報告 [3] がある。近年は、その資産価格の上昇から、悪意あるユーザーの標的となり、攻撃の懸念や、マネーロンダリングの手段とされるなどの問題が指摘されている。そのため、特定ユーザーを識別する研究 [17] など、セキュリティ方面での学術的な取り組みが活発である。最近の観点としては、特に、先行研究 [13] で報告されている通り、ブロックチェーン上に記録されている情報を分析する場合に得られる情報は限定的であり、P2P ネットワークの通信情報を分析することの重要性が認識されている。

そこで本研究では、セキュリティ方面で、活用が検討されているネットワークの通信パケットを、仮想通貨市場におけるリスクを評価する上で、有効と考えられるオルタナティブ・データ[1]として、活用することを提案する。すなわち仮想通貨を金融資産ではなく、システムとして捉え、セキュリティの観点から有効な情報を用いて価格分析を行う。

以降、第 2 章では、関連研究として、本研究で利用するダークネット観測情報および分析手法である GARCHSK について紹介する。第 3 章では、ダークネット到達パケットを用いた Bitcoin 価格の実証分析を示し、第 4 章で結論を述べる。

[1]オルタナティブ・データとは価格や出来高などの従来金融資産の分析に用いていた公開情報以外のデータ群をいう。

2 関連研究

仮想通貨におけるオルタナティブ・データの活用としては、先行研究 [1] の、その日に確認できるユニークなアドレス数を用いた研究がある。当該研究は、ネットワークの価値がネットワークの規模より推定される利用者のネットワーク効果に着目している。一方で、ビットコインのネットワーク分析に用いるデータとして、ダークネットを用いることを提案した研究がある [24]。この研究では、従来のビットコインのネットワーク分析で活用される正常なネットワークにて観測される情報ではなく、イレギュラーなネットワークにて観測される情報を用いて、ビットコイン・ネットワークの分析を試みた研究である。そこで本研究では、このダークネット観測情報と、ビットコインの価格やリターンのモーメントとの関係を分析する。以下に、本研究で用いるダークネットを用いた先行研究および、分析手法である GARCHSK モデルについて紹介する。

2.1 ダークネットを用いた先行研究

一般的にダークネットという単語には、下記 2 つの意味合いで用いられることがある。

- Tor[6] などの匿名通信プロトコルや BitTorrent[5] や Napster[1]といった、違法行為に関与した技術やサービスを含む違法な薬物や個人情報などを取り引きするために用いるサーバーやプログラムによって形成されるネットワークの総称。
- IP アドレスの割り当てが可能な空間のうち、端末等が未割り当ての IP アドレス空間の呼称であり、スキャニング活動、分散型サービス拒否攻撃 (DDoS 攻撃)、マルウェア識別などのさまざまなサイバー脅威情報を生成するために活用されているものを指す [25]。

Bitcoin などの仮想通貨については、違法取引の決済手段として利用されることもあるため、主に前者の闇市場関連の意味として扱う先行研究 [10] 等もあるが、本研究で扱うのは後者である。

ダークネットに関する研究を調査した先行研究では、ダークネットの研究対象から、1. ダークネットの展開とセットアップ (展開)、2. 展開されたセンサーによるダークネットデータの測定と分析 (パケット分析)、3. パケットの可視化と表現のためのツールとテクニック (可視化) の 3 つのカテゴリに分類している。

ビットコイン・ネットワークとダークネットの関係について調査した先行研究 [24] は 3 つのカテゴリーのうち 2. のパケット分析に該当する。

[1]Peer-to-Peer によるファイル共有サービス

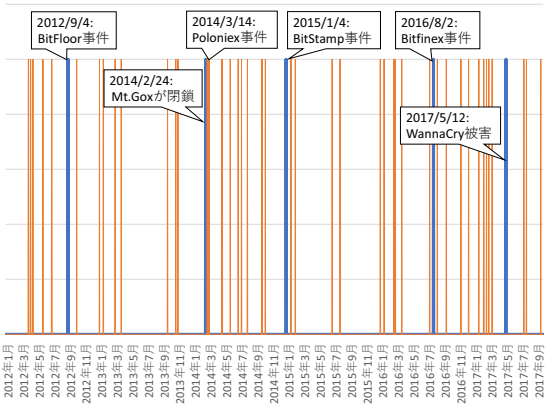


図 1: パケットの異常時とセキュリティ・インシデント

上述の研究では、まず、ダークネットを用いてビットコインネットワークを分析するにあたり、そもそもビットコイン・ネットワークで利用される通信がダークネットにて観測されるか確認している。本来ダークネットは端末等が未割り当ての IP アドレス空間であるため、通信が発生しないことが前提となる。しかしながら、通信が観測されないはずの空間に通信が観測されることを報告している。先行研究 [10] にて、世界中に観測点を持ち、大規模なプロジェクトに分類される国立研究開発法人情報通信研究機構 (NICT) の NICTER プロジェクト [15] にて観測しているダークネット上に観測される通信情報を用いて、2011 年 1 月 1 日から 2017 年 10 月 5 日の期間について分析した。具体的には、ビットコインなどの P2P 型のサービスがインターネットを介して端末間を接続する際に用いる、接続設定に着目した。そして、この接続設定に該当する通信に対する 1 日あたりの総パケット数の変化を確認している。また、図 1 の通り、観測値の異常値とセキュリティ・インシデントが関連する可能性について示唆した。

2.2 GARCHSK モデル

分散不均一性[2]を示す経済、金融時系列の条件付きボラティリティの時系列変化をモデリングするため、[9] は ARCH(AutoRegressive Conditional Heteroscedasticity) モデルを提案した。

[2] は ARCH モデルを一般化した GARCH(Generalized ARCH) モデルを提案した。(AR-)GARCH モデルは次式のように、資産リターン r_t の条件付き分散 h_t を過去の収益率のショック ε_{t-1} の 2 乗に、過去の分散 h_{t-1} の線形和で表現する。なお、定常性を満たすためには、$|\alpha| < 1, \beta_1 + \beta_2 < 1$ また分散の非負性から $0 < \beta_0, \beta_1, \beta_2$ の係数制約が必要である。

$$r_t = \alpha_0 + \alpha_1 r_t + \varepsilon_t \tag{1}$$
$$h_t = \beta_0 + \beta_1 \varepsilon_{t-1}^2 + \beta_2 h_{t-1} \tag{2}$$
$$\varepsilon_t | I_{t-1} \sim N(0, h_t) \tag{3}$$

[18] は条件付き分散をモデル化する GARCH モデルをさらに拡張して、条件付き歪度、条件付き尖度の変動も取り込んだ GARCHSK(GARCH Skewness-Kurtosis) モデルを提案した。また GARCHSK モデルによる実際の資産価格変動の実証分析を行い、株や為替のいくつかでは条件付き歪度、条件付き尖度の存在が確認された。このモデルの特徴は条件付き歪度、条件付き尖度の変動を GARCH モデルと同等のわかりやすい構造で明示的に捉えることができる。かつ、推定は容易である。GARCHSK モデルの具体的な定式化は次の通り。

$$r_t = \alpha_0 + \alpha_1 r_t + \varepsilon_t \tag{4}$$
$$h_t = \beta_0 + \beta_1 \varepsilon_{t-1}^2 + \beta_2 h_{t-1} \tag{5}$$
$$s_t = \gamma_0 + \gamma_1 \eta_{t-1}^3 + \gamma_2 s_{t-1} \tag{6}$$
$$k_t = \delta_0 + \delta_1 \eta_{t-1}^4 + \delta_2 k_{t-1} \tag{7}$$
$$\eta_t = h_t^{\frac{-1}{2}} \varepsilon_t \tag{8}$$
$$\eta_t | I_{t-1} \sim g(0, 1, s_t, k_t) \tag{9}$$

ここで、g は平均 0、分散 1、歪度 s_t、尖度 k_t を持つ確率密度関数である。なお、定常性を満たすためには、$|\alpha| < 1, \beta_1 + \beta_2 < 1, |\gamma_1 + \gamma_2| < 1, \delta_1 + \delta_2 < 1$ また分散と尖度の非負性から $0 < \beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2$ の係数制約が必要である。

彼らは、Chebyshev-Hermite 多項式を用いた Gram-Charlier 展開によって GARCHSK モデルの確率密度関数 $g(0, 1, s_t, k_t)$ が従う次のような分布を導出した。

$$g(\eta_t | I_{t-1}) = \frac{\phi(\eta_t)\psi^2(\eta_t)}{\Gamma_t} \tag{10}$$
$$\phi(\eta_t) = \frac{1}{\sqrt{2\pi h_t}} exp(\eta_t^2 - h_t) \tag{11}$$
$$\psi(\eta_t) = 1 + \frac{s_t}{3!}(\eta_t^3 - 3\eta_t) + \frac{k_t - 3}{4!}(\eta_t^4 - 6\eta_t^2 + 3) \tag{12}$$
$$\Gamma_t = 1 + \frac{s_t^2}{3!} + \frac{(k_t - 3)^2}{4!} \tag{13}$$

$\eta_t = h_t^{\frac{-1}{2}} \varepsilon_t$ より ε_t の確率密度関数は $f(\eta_t | I_{t-1}) = h_t^{\frac{1}{2}} g(\eta_t | I_{t-1})$ となる。したがって、定数項を除いた対数尤度関数 l_t は次のようにかける。

$$l_t = -\frac{1}{2} ln h_t - \frac{1}{2}\eta_t^2 + ln(\psi^2(\eta_t)) - \Gamma_t \tag{14}$$

よって、GARCHSK モデルの各パラメータは最尤法により l_t を最大化することで求めることができる。

[2] ある時期にはボラティリティが平均して小さく、別の時期にはボラティリティが平均して大きくなる傾向が観察される。このようなボラティリティが時期によって異なった水準を示すことを分散不均一性またはボラティリティ・クラスタリングと呼ぶ。分散不均一性は経済・金融時系列データに幅広く見られる現象である。

表 1: データ期間におけるビットコインの日次リターンの統計量

平均	標準偏差	歪度	尖度
0.43	4.58	0.43	15.65
サンプル数	最大値	最小値	Jarque-Bera (p-値)
2,102	41.59	-31.09	14,107 (0.000)

3 実証分析

前述の通り、本来ダークネットは端末等が未割り当ての IP アドレス空間であるため、通信が発生しないことが前提となる。しかしながら、通信が観測されないはずの空間に特に異常な量の通信が観測された場合、その前後で価格やモーメントにどのような影響を及ぼしているのかを分析する。各次数のモーメントの変動をとらえるために GARCHSK モデルを使用する。分散、歪度、尖度など通常の各次数のモーメントは、ある期間において一定の値をとるが、GARCHSK モデルを用いることで、各時点ごとの分散、歪度、尖度の時系列変化を捉えることができる。

3.1 データ

本分析に用いるデータとしては、先行研究 [24] と同様に、国立研究開発法人情報通信研究機構 (NICT) の NICTER プロジェクト [15] にて観測しているダークネット上に観測される通信情報を用いて、2011 年 1 月 1 日から 2017 年 10 月 5 日の期間について分析した。Bitcoin の価格については、coindesk[3]にて公開されている Bitcoin Index の値を用いた。

はじめに分析対象のデータ期間全期間における統計量を確認する。表 1 は日次リターンの統計量を整理した表である。日次の標準偏差は 4.5% と非常にボラタイルで、正の歪度を持ち、尖度も 15 と非常に大きい。当然ながら Jarque-Bera 検定による正規性はなく、通常の金融資産と同じくファットテールな分布を持つ。

次に、表 1 に示す通り日次リターンに対して AR(1) モデルを当てはめ、その残差についての自己相関を確認した。Ljung-Box 検定の結果、4 次までのすべての残差について自己相関があることがわかる。そのため、4 次の条件付きモーメントの自己相関をモデル化する GARCHSK モデルを当てはめる余地がある。具体的に

[3]https://www.coindesk.com/

表 2: AR(1) モデルの残差の Ljung-Box 統計量

LB(20) ε_t (p-値)	LB(20) ε_t^2 (p-値)	LB(20) ε_t^3 (p-値)	LB(20) ε_t^4 (p-値)
58.94 (0.0000)	881.43 (0.0000)	81.32 (0.0000)	77.47 (0.0000)

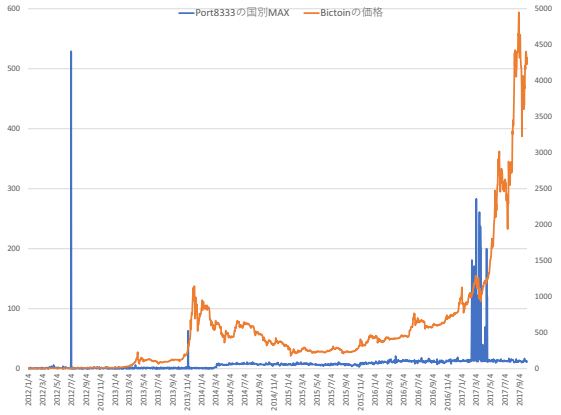


図 2: Price

日次リターンから推定した GARCHSK モデルのパラメータは表 3 の通りである。

次に、ダークネット観測情報から得られたパケットの「異常値」を定義する。図 2 はパケット観測値と Bitcoin の価格を合わせて表示した図である。明確にパケットの異常位置はわかるが、今回は 30 日間の標準化したパケット量が $+3\sigma$ を超えたら異常とした。また異常とされた日は 51 日存在した。

3.2 分析結果

以上の推定した条件付きモーメントおよび異常と定義したパケットを用いてそれらの関係を見ていく。図 3 から図 6 は、異常なパケットが発生した日とリターン、GARCHSK モデルで推定した条件付き分散、歪度、尖度を重ねて表示したグラフである。それぞれ分散は高いほうがリスクが大きいことを表し、歪度は低いほうが、尖度は高いほうが、極端な値をとるリスクが高いことを表している。グラフから条件付き分散、歪度、尖度がジャンプする前あるいは同時にパケットの異常値が検出されていることが確認できる。

さらに詳しくパケットの異常値が観測された前後の価格変動を包括的に確認する。表 4 は異常なパケットが観測される前 10 日の日次リターンのサマリーを示し、表 5 は観測後 10 日間示している。まず、パケットの観

表 3: GARCHSK モデルのパラメータ推定結果

	α_1	β_0	β_1	β_2	γ_2	γ_1	γ_2	δ_0	δ_1	δ_2
係数	0.0988	0.0008	0.0783	0.4509	-0.0900	0.0145	-0.1700	1.2787	-0.0000	0.6887
標準誤差	0.0142	0.0201	0.0159	0.0159	0.0159	0.0159	0.0142	0.0170	0.0160	0.0142
t 値	6.9323	0.0377	4.9192	28.3306	-5.6544	0.9082	-11.9379	75.3242	-0.0000	48.3634

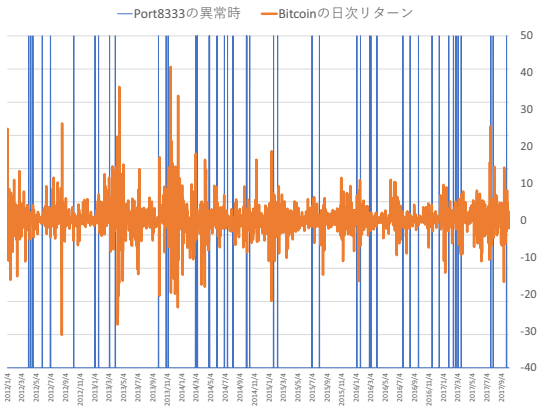


図 3: Return

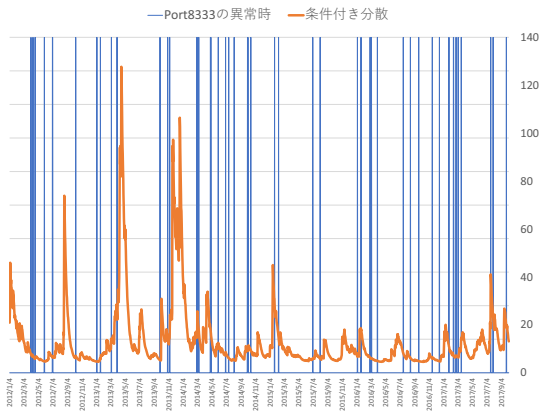


図 4: Vol

表 4: 過去 $+3\sigma$ を超える異常のパケットが観測された日までの前 10 日間のモーメント

	日次リターン	条件付き分散	条件付き歪度	条件付き尖度
平均	0.92	10.80	-0.06	4.07
標準偏差	5.22	5.05	0.28	0.47
歪度	0.81	1.07	-0.68	5.40
尖度	8.03	3.80	24.27	37.77
MIN	-10.97	4.72	-2.49	3.89
MAX	24.78	28.44	1.77	8.17

表 5: 過去 $+3\sigma$ を超える異常のパケットが観測された日の後 10 日間のモーメント

	日次リターン	条件付き分散	条件付き歪度	条件付き尖度
平均	-0.34	21.10	0.13	4.44
標準偏差	3.74	25.53	0.97	2.56
歪度	0.79	2.01	6.44	8.71
尖度	11.27	6.10	60.99	90.01
MIN	-16.17	4.57	-4.00	3.89
MAX	7.43	127.73	10.04	33.57

測前後で日次リターンが反転してマイナスとなり、さらに、パケットが観測された後は条件付きボラティリティが倍程度大きくなっている。図 7 はパケットの異常値が観察された時点を 0 として、前後 20 日の累積リターンをプロットした図である。明らかに、異常値が観察された後では価格変動幅が大きくなっていることがわかる。

以上の分析から図 1 の通り、ハッキングをはじめとした何らかの Bitcoin のセキュリティ・インシデントが影響を与えている可能性を示唆する。

4 まとめ

本研究では、セキュリティ方面で、活用が検討されているネットワークの通信パケットを、仮想通貨市場におけるリスクを評価する上で、有効と考えられるオルタナティブ・データとして、活用した。仮想通貨を金融資産ではなく、システムとして捉え、セキュリティの観点から有効な情報を用いて価格分析を行った。

具体的には、ダークネット観測情報から得られたパケットデータの異常を検知することで、以下の事象を確認し、セキュリティ・インシデントに起因すると思われる価格変動リスクを回避できる可能性をしめした。

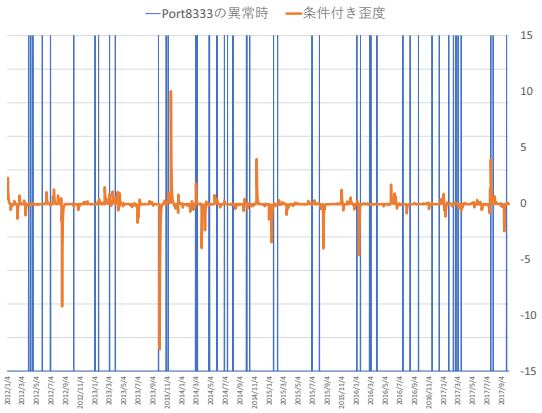


図 5: Skew

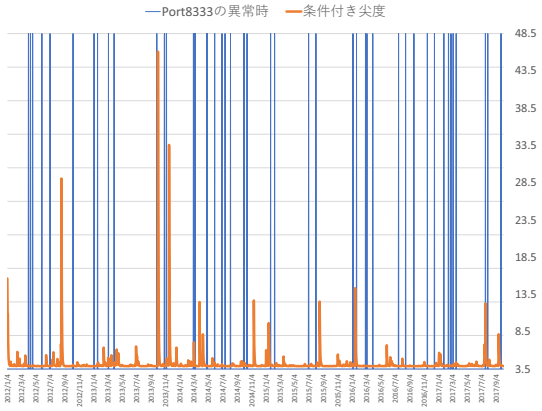


図 6: kurt

- パケットの異常観測の後、あるいは同時に条件付きモーメントがジャンプする。
- 異常観測後前後 10 日で日次リターンがマイナスになる。
- パケットが観測された後は条件付きボラティリティが倍程度大きくなる。

今後の展開としては、ダークネット観測情報を利用して、企業のセキュリティ・リスクもモニタリングすることが挙げられる。今回は Bitcoin を対象として分析のため、8333/tcp について確認したが、企業のセキュリティ・リスクとしは、well-known port と呼ばれる、一般的に特定サービス利用のためのポートを監視することで確認可能であると考える。セキュリティ・リスクが資産価格に与える影響については、主にイベント・スタディ法を用いた分析が主体であり、例えば、情報セキュリティ事故が企業価値 (株価) に与える負の影響をイベント・スタディの方法を用いて分析した研究 [26] や、脆弱性が発表されたときにソフトウェアベンダーの市場価値がどのように変化するかを検証した研究 [23] など

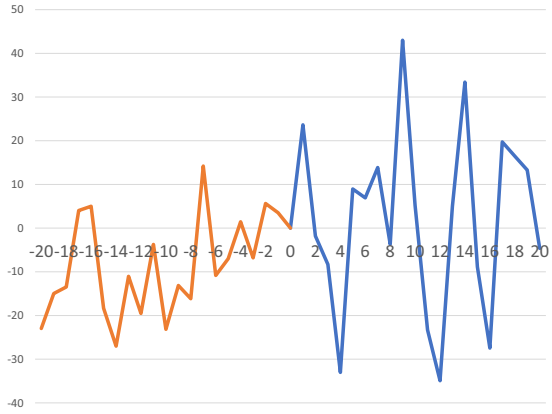


図 7: パケット異常値を基準とした累積リターン

がある。その他にも、脆弱性の発表が行われた際の株式市場の反応について纏めた研究 [22] や、ハッカーの攻撃を直接受けた企業および類似企業の株価がともに低下するといった研究報告 [14] もある。こうした情報セキュリティが株価に与える影響に関する体系的な調査をした研究 [21] では、企業の株価へのセキュリティ事象の影響の統計的有意性を報告している。

参考文献

[1] Ken Alabi. Digital blockchain networks appear to be following metcalfe's law. *Electronic Commerce Research and Applications*, Vol. 24, pp. 23–29, 2017.

[2] Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, Vol. 31, No. 3, pp. 307–327, 1986.

[3] Joseph Bonneau, Andrew Miller, Jeremy Clark, Arvind Narayanan, Joshua A Kroll, and Edward W Felten. Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In *Security and Privacy (SP), 2015 IEEE Symposium on*, pp. 104–121. IEEE, 2015.

[4] Marie Brière, Kim Oosterlinck, and Ariane Szafarz. Virtual currency, tangible return: Portfolio diversification with bitcoin. *Journal of Asset Management*, Vol. 16, No. 6, pp. 365–373, 2015.

[5] Bram Cohen. Incentives build robustness in bittorrent. In *Workshop on Economics of Peer-to-Peer systems*, Vol. 6, pp. 68–72, 2003.

[6] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor the second-generation onion

router. Technical report, Naval Research Lab Washington DC, 2004.

[7] Anne Haubo Dyhrberg. Bitcoin, gold and the dollar–a garch volatility analysis. *Finance Research Letters*, Vol. 16, pp. 85–92, 2016.

[8] Anne Haubo Dyhrberg. Hedging capabilities of bitcoin. is it the virtual gold? *Finance Research Letters*, Vol. 16, pp. 139–144, 2016.

[9] Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

[10] Claude Fachkha and Mourad Debbabi. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 2, pp. 1197–1227, 2016.

[11] David Garcia and Frank Schweitzer. Social signals and algorithmic trading of bitcoin. *Royal Society open science*, Vol. 2, No. 9, p. 150288, 2015.

[12] James Douglas Hamilton. *Time series analysis*, Vol. 2. Princeton university press Princeton, 1994.

[13] Jordi Herrera-Joancomartí. Research and challenges on bitcoin anonymity. In *Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance*, pp. 3–16. Springer, 2015.

[14] Oliver Hinz, Michael Nofer, Dirk Schiereck, and Julian Trillig. The influence of data theft on the share prices and systematic risk of consumer electronics companies. *Information & Management*, Vol. 52, No. 3, pp. 337–347, 2015.

[15] Daisuke Inoue, Masashi Eto, Katsunari Yoshioka, Shunsuke Baba, Kazuya Suzuki, Junji Nakazato, Kazuhiro Ohtaka, and Koji Nakao. nicter: An incident analysis system toward binding network monitoring with malware analysis. In *Information Security Threats Data Collection and Sharing, 2008. WISTDCS'08. WOMBAT Workshop on*, pp. 58–66. IEEE, 2008.

[16] Paraskevi Katsiampa. Volatility estimation for bitcoin: A comparison of garch models. *Economics Letters*, Vol. 158, pp. 3–6, 2017.

[17] Philip Koshy, Diana Koshy, and Patrick McDaniel. An analysis of anonymity in bitcoin using p2p network traffic. In *International Conference on Financial Cryptography and Data Security*, pp. 469–485. Springer, 2014.

[18] Ángel León, Gonzalo Rubio, and Gregorio Serna. Autoregresive conditional volatility, skewness and kurtosis. *The Quarterly Review of Economics and Finance*, Vol. 45, No. 4, pp. 599–618, 2005.

[19] Hsin-Ke Lu, Li-wei Yang, Peng-Chun Lin, Tzu-Han Yang, and Alexander N Chen. A study on adoption of bitcoin in taiwan: using big data analysis of social media. In *Proceedings of the 3rd International Conference on Communication and Information Processing*, pp. 32–38. ACM, 2017.

[20] Nakamoto Satoshi. Bitcoin:a peer-to-peer electronic cash system. *URL:http://www.bitcoin.org/bitcoin.pdf*, 2008.

[21] Georgios Spanos and Lefteris Angelis. The impact of information security events to the stock market: A systematic literature review. *Computers & Security*, Vol. 58, pp. 216–229, 2016.

[22] Georgios Spanos, Lefteris Angelis, and Kyriaki Kosmidou. Is the market value of software vendors affected by software vulnerability announcements? In *Strategic Innovative Marketing*, pp. 465–469. Springer, 2017.

[23] Rahul Telang and Sunil Wattal. An empirical analysis of the impact of software vulnerability announcements on firm stock price. *IEEE Transactions on Software Engineering*, Vol. 33, No. 8, pp. 544–557, 2007.

[24] 今村光良, 面和成. ダークネット観測情報を用いたビットコインネットワークの分析. SCIS2018:2018年暗号と情報セキュリティシンポジウム.

[25] 小出駿, 鈴木将吾, 牧田大佑, 村上洸介, 笠間貴弘, 島村隼平, 衛藤将史, 井上大介, 吉岡克成, 松本勉ほか. 通信プロトコルのヘッダの特徴に基づく不正通信の検知・分類手法. コンピュータセキュリティシンポジウム 2014 論文集, Vol. 2014, No. 2, pp. 48–55, 2014.

[26] 廣松毅. 情報セキュリティ事故が企業価値に与える影響の分析-イベント・スタディ分析を用いたリス

ク評価の試み. 情報セキュリティ総合科学, Vol. 3,
pp. 91–106, 2011.

ビットコイン市場におけるニュースの関係性における分析

Analysis on the relationship of news in bitcoin market

川田真也 [1,2,3]　尹熙元 [2,3]　藤原義久 [1]

Shinya Kawata[1,2,3], Hiwon Yoon[2,3], and Yoshi Fujiwara[1]

[1] 兵庫県立大学大学院　シミュレーション学研究科
[1]University of Hyogo graduate school of simulation study
[2] 株式会社シーエムディーラボ
[2] CMDlab Inc.
[3] 株式会社ビットアルゴ取引所東京
[3] bitARG Exchange Tokyo Inc.

Abstract: In this research, we use English-language news related to the Bitcoin posted on the Internet as a data source, and use LDA(Latent Dirichlet Allocation) which is one of probabilistic topic models, for each article. We judged what kind of topics (keyword group) the sentence is composed of, and quantitatively expressed the excitement of topics in the period using different topic distribution for each article obtained. Furthermore, by analyzing the relationship with the bitcoin's market price on the Internet, we try to evaluate the influence of the news. We show to the relation between the quantity concerning the bitcoin's price (BTC/JPY) in the target period and the excitement of the topic in the news article related to bitcoin.

1. はじめに

日本では 2017 年 4 月 1 日に改正資金決済法が施行され、世界で初めて交換業としての仮想通貨の取り扱いは、国の認可が必要となった。そのことによって、世界から大きな注目を集めている。また法整備等に関しては、金融商品取引法ではなく、資金決済法の中で扱われている。その一方で、現状の仮想通貨は、従来の金融商品のような性質を持ち、投資や投機目的で用いられている。そのため、従来の金融商品であれば、禁止されているような相場操縦行為や風説の流布といった事象に抵触するような情報等が流れていることも目にする。現在の仮想通貨市場において、ある一つの話題やニュースなどが引き金となり、価格変動を助長しているようにも見受けられる。また情報のリソースが SNS 等によるものであることを鑑みると、更新される新しい情報に対しての変化を捉え、健全な市場形成のために客観的な方法によって、提供される情報に関するガイドライン等の整備が必要であると考えられる。まず本研究では、仮想通貨に関する英字ニュースに対して、確率的トピックモデルの一つである LDA(Latent Dirichle Allocation)を用いて、記事データに話題の集積を数値的に評価し、価格データとの関係性について示す。

2. データについて

2.1. テキストデータ

今回は、インターネット上で公開されている英字の仮想通貨関連の記事を取得した。記事の取得期間は、2017 年 5 月 6 日から 2017 年 11 月 25 日までとなる。さらに期間ごとに連続する（本研究では、２連続または、3 連続）単語について、データセットごとに、TFIDF を用いて連続する単語の重要度を計算し、重要度の高い連続する単語を複合語として判断し処理を行う。

【複合語の例】

mining farms　=>　mining-farms
bitcoin cash　=>　bitcoin-cash
quantum computing　=>　quantum-computing
initial coin offerings　=> initial-coin-offerings

・データ取得先

https://blockchain.info/ja/charts/marketcap?timespan=all

総記事数：1349 記事

総単語数 : 850836 単語（ユニークな単語 : 60295 語）

2.2. ビットコイン[1]の価格データ

今回ビットコインの価格データに関しては、一日ごとの始値・終値・高値・安値を集計した国内の主要取引所の価格データを用いる。

3. 分析手法

3.1. 記事データにおける集積度

記事を用いた集積度（話題の集中度合）を定義する。単純な単語の出現数で判断することでは、同一のキーワードが出てきていないものであっても内容については、同義の内容を示しているものなどの判定が難しくなる。そのため、確率的トピックモデルの一つのである LDA[2,3]によって、データセットごと（今回は、過去 6 週間の記事データを 1 データセットとした）に解析を行った。LDA に用いたパラメータは、今回データセットごとでの記事数の違いから一律のパラメータ（表 1）を用いている。ただしパラメータの設定については、検討の余地を残している。 LDA によって得られた記事ごとの話題の分布 P(x)または Q(x)から分布の類似度計算を行い、記事の話題の分布が近しいものを特定することとした。記事間のつながり[4]の判定並びに類似度計算については、Jensen-shannon divergence を用いて計算を行った。以下に計算式(1)～(3)を表す。

＜表 1 ＞LDA に用いたパラメータ

α (記事毎の話題の分布を推定に用いる)	0.25
β (話題内の単語の分布の推定に用いる)	0.01

$$D_{KL} = \int_x P(x)(log\, P(x) - log\, Q(x))\, dx \quad (1)$$

$$M = \frac{P(X)+Q(X)}{2} \quad (2)$$

$$D_{js} = \frac{1}{2} D_{kl}(P||M) + \frac{1}{2} D_{kl}(Q||M) \quad (3)$$

得られた類似度計算から、データセット内でのすべての記事の組み合わせから類似度の高い組み合わせの上位 1%を抽出する。抽出した組み合わせを用いてデータセット間の記事のつながりから記事のつながりでコンポーネントを判定し、指標（表 2）を算出し、記事を用いた期間内での集積度の計算を行う。ただし、本研究においてコンポーネントは、グラフ構造内での分離されているネットワークを指すものとする。

＜表 2 ＞集積度計算のための指標

$n_{i,a}$	データセット i における類似度計算で上位 1.0%に含まれる記事の総数 (コンポーネントが 2 つの記事で構成されているものは除く)
$n_{i,b}$	データセット i における記事数
C_i	データセット i のコンポーネントの数
R_i	データセット i の話題の集積度

$$R_i = \frac{n_{i,b}}{n_{i,a}} \times \frac{1}{C_i} \quad (4)$$

データセット内の記事数が 30 あるとし、記事の組み合わせから類似度の高い組み合わせの上位 1%を抽出した組み合わせから作成した記事ネットワーク（図 1）について考える。
article00~article05,article07 は一つのコンポーネント、さらに article_11~article13 も一つのコンポーネントを形成している。定義より article09, article10（article14, article15）はコンポーネントであるが二つの記事のみでのコンポーネントであるために除外するとする。
この結果、各指標は表 3 のようになる。

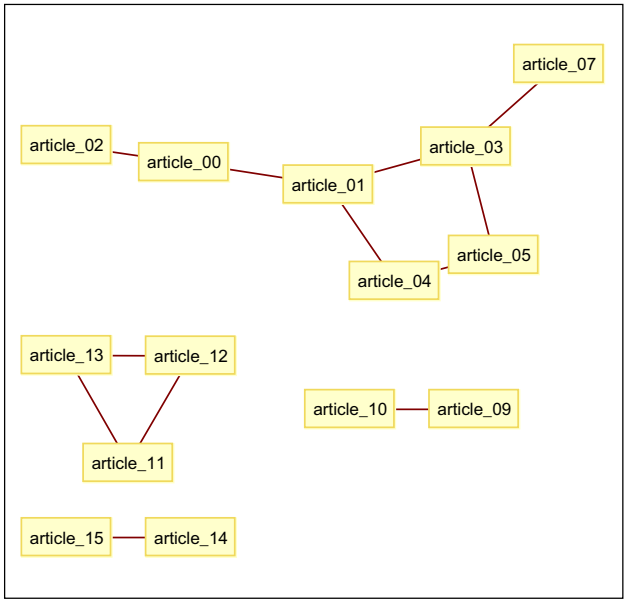


図 1 ：記事ネットワーク

＜表 2 ＞集積度計算のための指標

$n_{i,a}$	10
$n_{i,b}$	30
C_i	2
R_i	1.50

3.2. ビットコインの価格データの処理について

今回ビットコインの価格データに関しては、一日ごとの始値・終値・高値・安値を集計したデータから 1 週間ごとの始値・終値・高値・安値に再集計したデータを用いる。

4. 結果及び考察

ビットコインの価格（図２）が 2017 年 1 月から 2017 年 11 月にかけて 5 倍近く、その他の仮想通貨でも軒並み同様かそれ以上の価格の上昇が見られることを鑑みると価格の推移を比較するのではなく、1 週間ごとの高値・安値から 1 週間ごとの高値・安値を計算し、価格変動幅を求め停止した話題の集積度との比較を行った。

価格変動幅の変化率と話題の集積度の変化率との相関を取ると 0.49071 となり弱い相関を持つ。同様の解析を対象とするデータセットの期間を 1 週間から 2 ヶ月周期で行なった際には、相関が、0.813 と高い相関を持つことはわかっている。

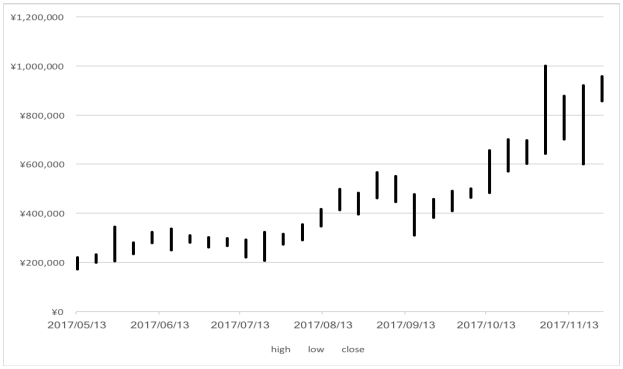


図 2：ビットコインの価格(BTC/JPY)の週足データ (2017/5/6-2017/11/24)

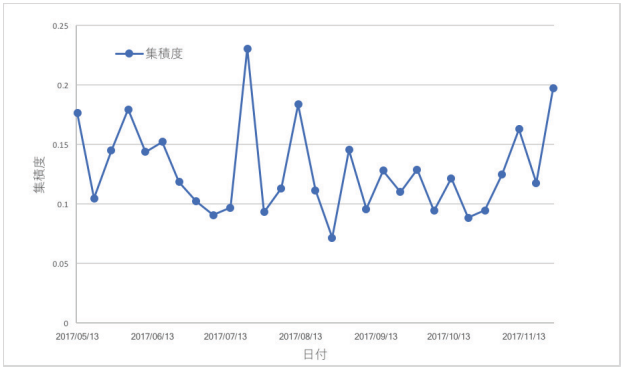


図 3：話題の集積度の週ごとの推移 (2017/5/6-2017/11/24)

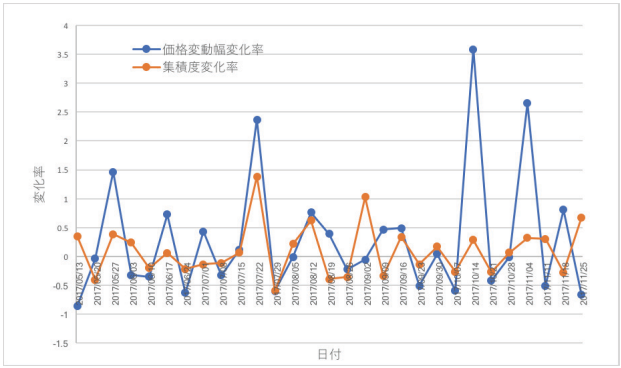


図 5：価格変動幅と集積度の変化率

単純な相関のみを評価したが、実際にテキスト情報と価格変動の因果関係については、さらなる検討が必要になる。因果関係を考えるためには、正確にどのタイミングで情報を受け取っているのかという点も考慮する必要がある。

また、ビットコインの価格のように変動が激しい時系列データに関しての処理等についても考慮する必要がある。

5. 今後の展望

本研究では、英字テキストとビットコイン価格（BTC/JPY）を用いての解析であったが、日本語によるデータソース及び国内の主要取引所等でのビットコイン価格（BTC/JPY）に関しての解析を今後行うこととしている。また単語数がある程度見込めるリソースでの解析であるが、実際の市場を鑑みると短文形式の SNS や仮想通貨に関するまとめサイトさらには、市場への注目度を鑑みると一部の専門的なメディアだけではなく、様々な媒体に対してのアプローチ方法について現在検討中である。

謝辞

本研究は、株式会社シーエムディーラボ並びに株式会社ビットアルゴ取引所東京(仮想通貨交換業　第 00011 号　関東財務局)によるデータの提供、協力によって行うことができました。ここに深く感謝いたします。

参考文献

[1] Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. (2008).

[2] Blei, D. M., and Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation, Journal of Machine Learning, 3, 993–1022 (2003)

[3] Phan, X.-H. and Nguyen, C. T., GibbsLDA++: A C/C++

Implementation of Latent Dirichlet Allocation(2008)

[4] Kawata, S., & Fujiwara, Y.. Constructing of network from topics and their temporal change in the Nikkei newspaper articles. Evolutionary and Institutional Economics Review, *13*(2), 423-436(2016).

感情によるマルチモーダル AI を利用した IPO 株価推定

Prediction of IPO stock prices by using multimodal emotion AI

河合 継 [1]，新田 翔 [1,2]，大川 尭郁 [1,3]，西山 昇 [4]

Kei Kawai[1], Sho Nitta[1,2], Takafumi Okawa[3], and Noboru Nishiyama[4]

[1] クリスタルメソッド株式会社

[1] Crystal method co.ltd

[2] 東京理科大学工学研究科経営工学専攻

[2] Department of Management Science, Graduate School of Engineering, Tokyo University of Science

[3] 東京大学工学系研究科物理工学専攻

[3] Department of Applied Physics, Graduate School of Engineering, The University of Tokyo

[4] 千葉商科大学会計大学院　客員教授

[4] MBA Program , Graduate School of Accounting & Finance, Chiba University of Commerce

Abstract: IPO 及び New Stage（市場変更）時に放映される StockVoice TV 出演社の社長や他出演者の表情・声の特徴量・テキストを基にした感情特徴量が株価に与える影響について、機械学習を通じた検証を行った。価格データについては、株式会社 Quick 様から提供して頂いたデータを利用し、検証対象として番組全体の感情特徴量と翌日の終値、及び番組中の一分ごとの感情特徴量と次点の一分足の終値の学習を行った。本研究では、目・耳・言葉の三系統のマルチモーダルな特徴抽出を行う事で予測の精度を上げることが出来るのではないか、という仮説のもと検証を行った。日足に関しては、SVM・ロジスティック回帰に特徴量を入力し二値検証を行い、１分足は５秒ごとの特徴量を入力とし、XGBoost 及び DNN を用いた３クラス分類による検証を行った。

1. はじめに

古典的な経済学において、人間の売買行動は需要と供給によって決定される価格に基づいて行われるものだと考えられている。しかしながら、近年の様々な研究によると、人間は感情によって売買行動を起こすという研究結果も報告されている。実社会においては、店員が顧客と会話を進めている際、「顧客が頻繁にうなずくような演出」をすることにより、購買意欲が上昇し、一方で首を振るような演出をすると購買意欲が低下するという報告がされている。これは、頷きによって人間は無意識に良い感情を抱くためであるとされている。同時に、プレゼンテーションの良し悪しも購買意欲や売上に大きな影響を及ぼすと考えられ、人々は何かものを買う際に論理的に買う理由を考える一方で、買うタイミングにおいては何かしらの感情の動きがあり、それに基づいて売買を行うことが多いとされている。

同様にして、IPO 及び市場変更の際に企業の社長や従業員が出演し会社の事業計画をスピーチする映像においても、人々が受ける印象や感情によって売買行動への影響があるのではないかと考え、検証を行った。映像による企業説明においても、一般的な購買行動と同様にセールスが店頭において行う営業活動と共通する部分があるのではないかと考えた。

このような IPO・市場変更に伴うプレゼンテーション映像は、株式会社 StockVoice 様によって上場・公開日当日の後場開始時に放映されたものを用い、番組では主に自社の事業計画について説明が行われる。（表 1 参照）

本映像は、キャスターと二人で話をするものや、プレゼンテーションの資料を基に事業内容の説明を行う 13 分程度のインターネット・CS 放映である。これらの映像に基づく感情の特徴量が株価の値動きにどのような影響を与えるのか、また影響があるのかということについて、機械学習を用い、検証を試みた。

2017年			コード	
	統合	AOI TYO Holdings(株)	3975	☆①
	統合	(株)FCホールディングス	6542	☆JQ
127	1030	(株)シャノン	3976	☆M
206	1430	(株)CDG※質問順番変更	2487	②→①
207	1330	森トラスト・ホテルリート投資法人:	3478	☆REIT
210	1030	(株)安江工務店(やすえこうむてん	1439	☆JQ
216	1430	(株)日宣(にっせん)	6543	☆JQ
217	1330	(株)TOKYO BASE	3415	M→①
223	1330	(株)レノバ	9519	☆M
223	1430	ユナイテッド&コレクティブ(株)	3557	☆M
227	:	(株)ハブ	3030	JQ→②
302	1330	イフジ産業(株)	2924	②→①
303	1030	(株)やまぜんホームズ	1440	☆PRO
306	:	(株)エイチワン	5989	②→①
307	1030	(株)ロコンド	3558	☆M

表 1. IPO・市場変更に伴う放映一覧の例

2. 前提・検証概要

今回の研究では、人間の感情特徴量をマルチモーダルに抽出し、機械学習を用いて行った。マルチモーダルに特徴量を抽出する理由としては、例えば、「さようなら」という単語の意味を扱う際、まず一次的には単語の意味と話し方、そして二次的には文脈によってその意味合いが大きく左右されると考えられ、どのような意味合いで「さようなら」と言っているのかをより精度良く判断したいと考えたからである。今回は、目・耳・言葉（テキスト）の三つのモーダルチャネルを用いることによって、感情のそれぞれにおける特徴量を抽出して学習を行った。

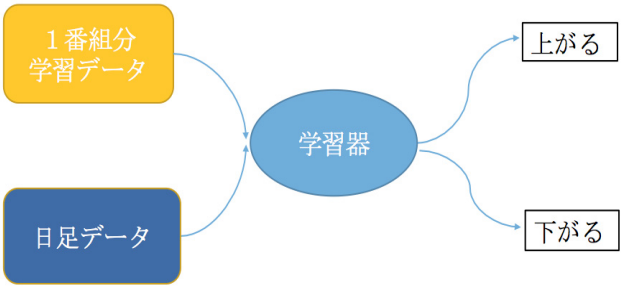
検証対象は、

・番組中全体の評価を受けて翌営業日の値動き（日足）

・番組中の値動き（一分足）

を用い、一分足は歩み値から生成を行った。

・日足



・一分足

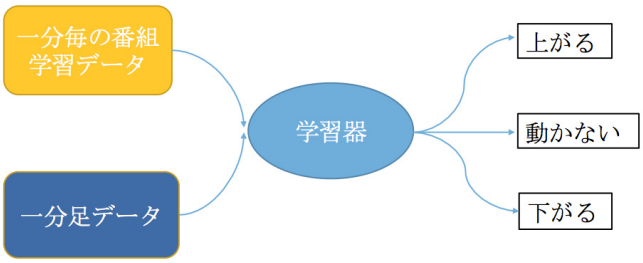


図 1. 日足と分足の二値・三値予測

2.1 目の特徴量（映像の顔表情）

視覚による感情特徴量抽出では、StockVoiceTV の 5 秒ごとの顔表情特徴量を取得した。

特徴量の抽出は、Microsoft の Emotion API を利用した。この際の特徴量のデータは、以下のようなベクトル量である。

emotion_indices=['disgust','happiness','surprise','anger','fear','neutral','sadness','contempt']

複数人が画面上に表れている場合は、全体の平均、プレゼンテーションの場合、画面上に顔が現れないため特徴量は neutral を表す要素のみが 1 となる One-Hot ベクトルとなる。日足、分足については同様のデータを利用した。

2.2 耳の特徴量（感情特徴量）

音声処理によって感情認識を行う場合、音声のパワー、周波数、MFCC が大きな影響を及ぼすとされる。これらの特徴は感情特徴量と呼ばれる。[1]

今回は、200 次元の MFCC とパワーの特徴量変数を学習に利用し、OpenSmile を用いて 5 秒ごとに情報抽出を行った。

2.3 言葉の特徴量（テキスト）

言葉による特徴量抽出では、StockVoiceTV を一分ごとに区切り、音声ファイルを作成した後、その音声ファイルを Google Speech API に送信し発話内容のテキストを受け取った。そして、Google Natural Language API によって感情極性を取得した。

また、日足検証用のテキストに関しては、Google

ストレージにマップした後、Google Speech API によって日本語化を行った。一分足のデータとの相違点としては、途中で文が切れること無く、文が区切られて表現されている。特徴量は、−1.0から+1.0の数値によって表現される。

また、東北大学で収録した感情音声コーパス（JTES）についても同様にテキスト起こしを行い、Google Natural Language API に渡した。この際の評価値は、以下の通りであり、感情について Google Natural Language API で判定ができると考えている。

・ang（怒り）：-0.312
・joy（喜び）：0.516
・sad（悲しみ）：−0.114
・neu（中立）：0.198

3. 入力データとモデル

3.1 価格

株式価格の入力データとしては、日足分 2015 年 1 月から 2017 年 12 月の分について株式会社 Quick 様より提供して頂いたものを用いた。一分足データについては、2017 年の歩み値を同様に提供していただき、2015、2016 年分については、JPX データクラウドのサービスからダウンロードしたものを PostgreSQL にロードし、今回検証分のみ抽出した。

またデータの欠損等により、最終的に分類器の作成に用いた企業数は 201 社となった。

3.2 分類器について

日足の分類器作成には、SVM とロジスティック回帰を用いた。これらは、多クラス分類で一般的に用いられているアルゴリズムである。

１分足の分類器作成には、XGBoost および DNN を使用した。XGBoost は高速かつ予測精度が高いアルゴリズムとして注目されており、機械学習のコンペティションでよく使用されている。DNN は昨今注目されている Deep Learning の手法の１つである。

4. 検証方法・結果

4.1 番組全体の値動きについて

日足データに関して、IPO（新規株式公開）のみの場合（91 社）と New Stage（上場変更）を含む場合（196 社）に分けて検証を行った。SVM、ロジスティック回帰それぞれにおいて分類器を作成した。

使用した特徴量は顔データ・音声・テキストの特徴量を会社ごとに一つのファイルにまとめ、当日の株価と翌営業日の株価の差がプラスであれば0、マイナス及び動きがない場合は1のクラスを設定して学習を行った。

7 : 3	LOGISTIC	SVM
1	0.6923	0.5128
2	0.4872	0.3333
3	0.4615	0.5128
4	0.5385	0.5128
5	0.4872	0.5385
6	0.4103	0.4359
7	0.4615	0.4872
8	0.5128	0.4872
9	0.5385	0.5641
10	0.5385	0.5128
Max	0.6923	0.5641
Min	0.4103	0.3333
Avg	0.5128	0.4897

表★．新規上場企業のみ（Train : Validation = 7 : 3）

8 : 2	LOGISTIC	SVM
1	0.7692	0.5385
2	0.5385	0.6923
3	0.6538	0.3846
4	0.5385	0.5385
5	0.5769	0.5000
6	0.4615	0.5000
7	0.3846	0.3846
8	0.3846	0.5385
9	0.4615	0.6154
10	0.5385	0.5000
Max	0.7692	0.6923
Min	0.3846	0.3846
Avg	0.5308	0.5192

表★．新規上場企業のみ（Train : Validation = 8 : 2）

7 : 3	LOGISTIC	SVM
1	0.4576	0.3390
2	0.4746	0.5593
3	0.4915	0.5254
4	0.5085	0.5424
5	0.4915	0.4407
6	0.4576	0.4407
7	0.4407	0.4746
8	0.5254	0.5085
9	0.4915	0.5085
10	0.5424	0.5424
Max	0.5424	0.5593
Min	0.4407	0.3390
Avg	0.4881	0.4881

表★．すべての企業（Train : Validation = 7 : 3）

8:2	LOGISTIC	SVM
1	0.5128	0.3590
2	0.4872	0.5385
3	0.5897	0.6154
4	0.4615	0.5641
5	0.4359	0.4359
6	0.3846	0.3590
7	0.4872	0.3846
8	0.5641	0.4872
9	0.4872	0.4872
10	0.4359	0.6667
Max	0.5897	0.6667
Min	0.3846	0.3590
Avg	0.4846	0.4897

表★．すべての企業（Train : Validation = 8 : 2）

4.2　リアルタイム中の値動きについて

IPO の開始時点で公募価格は決定しており、リアルタイムの値動きに変動があった場合にのみ配信される歩み値を基にデータを構築しているため、番組開始時点で価格配信がされていない場合もありえる。201 社中 91 社が番組開始時点で価格配信を受けていたため、まずその 91 社について検証を行った。（約定がない等、リアルタイムでのデータがとれないことが、公募価格を取り入れたことで、約定がない時間帯に、価格が上昇しているのか、下降しているのかがわかるようになった。）また、公募価格についても株式会社 Quick 様のご協力のもとデータを受け取り、価格配信がなされていない 40 社程度についても検証を行えるようにした。（合計 138 社）

会社ごとに５秒ごとの顔の表情・声・テキストの特徴量を入力としてデータを構築した。

その後 DNN を用い、分類器を作成した。ここで、３クラス分類のラベルは以下のように設定した。

0：0.01%以内の値動き or 約定なし
1：0.01%以上の上昇
2：0.01%以上の下落

検証の結果は以下に示す通りである。

4. 3　91 社のデータを用いた学習・分析

4.3.1　ランダムフォレスト

次に、学習が高速であることや説明変数が多数でも上手く働くという理由から、ランダムフォレストを用いて分類器を構築した。学習した分類器のテストデータでの正答率を以下に示す。

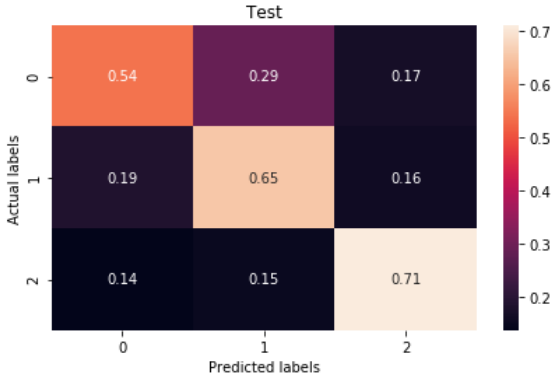


図 3．ランダムフォレストによる予測

4.3.2　XGBoost

さらに、予測精度の向上のため、XGBoost を用いて分類器の構築を行った。データはランダムフォレストと同じものを用いた。学習した分類器のテストデータでの正答率を以下に示す。結果は、ランダムフォレストと比較して予測精度が向上した。

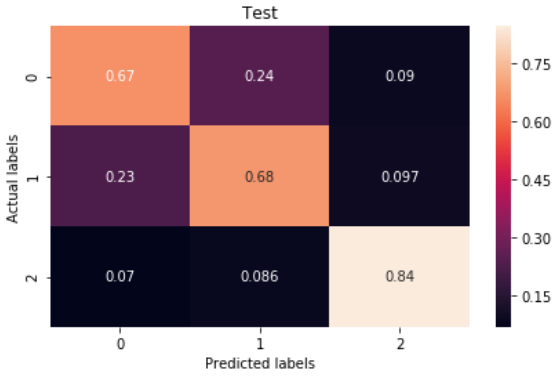


図 4．XGBoost による予測

また、XGBoost では、説明変数の重要度（寄与度）の算出を行った。結果を以下に示す。

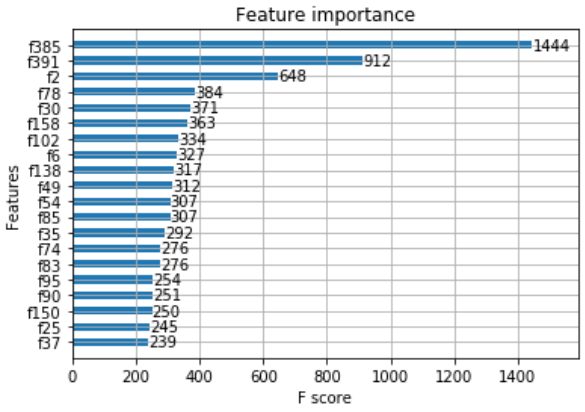
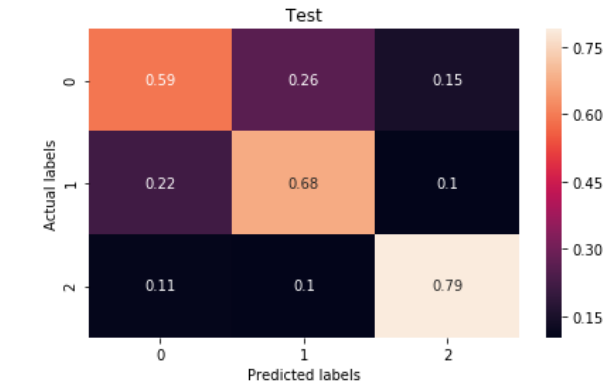


図 5. 説明変数の算出

この分析において、最も重要度が高い「f385」は、テキストの感情極性を示す。また、二番目の「f391」は、中間を表す顔表情感情特徴量である。一方で、テキストの感情極性は一分間同じ値を保持しているため、テキストの感情極性を除いて、再度勾配ブースティング決定木を用いた予測器の構築と説明変数の重要度（寄与度）の算出を行った。テストデータでの正答率と重要度算出の結果を以下に示す。



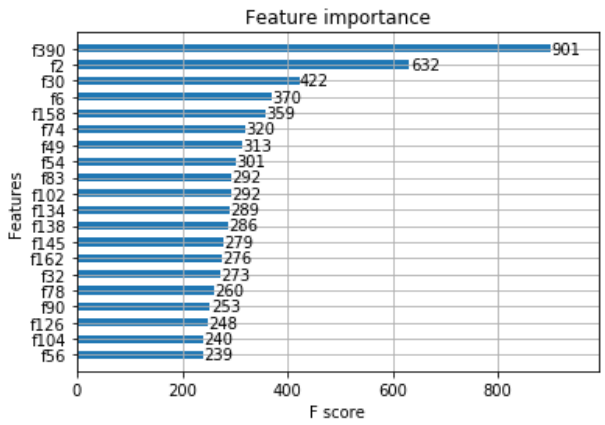


図 6. テストデータの正答率と説明変数の重要度

最も重要度の高い「f390」は中間を示す顔表情特徴量であり、二番目の「f2」は音量の二乗平均平方根値のデータ中の最小値である。三番目の「f30」は、二次メル周波数ケプストラム係数の算術平均を示す。「中間」以外の顔表情特徴量で高い重要度を示したものは無かった。

4.3.3 DNN

91 社の学習データ、および、138 社の学習データを用い、ニューラルネットワークワークの分類器構築を行った。今回の検証ではハイパーパラメータの設定が難しく、試行錯誤を行った。
無作為にデータ増幅した場合学習率は伸びたが、結果が伴っていない（オーバーフィッティング）。3 月 20 日の研究発表までに解決する見通し。

4.3.4 91 社のデータを用いた学習・分析

保有するデータのうち、Stock Voice TV の開始時刻から 10 分以内に約定を一度でもしている企業の数が 91 社であった。これらの企業のデータを用い、DNN による分類器を作成した。結果は以下に示す通りである。

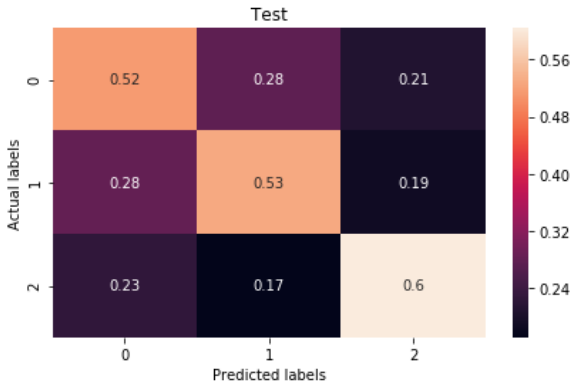
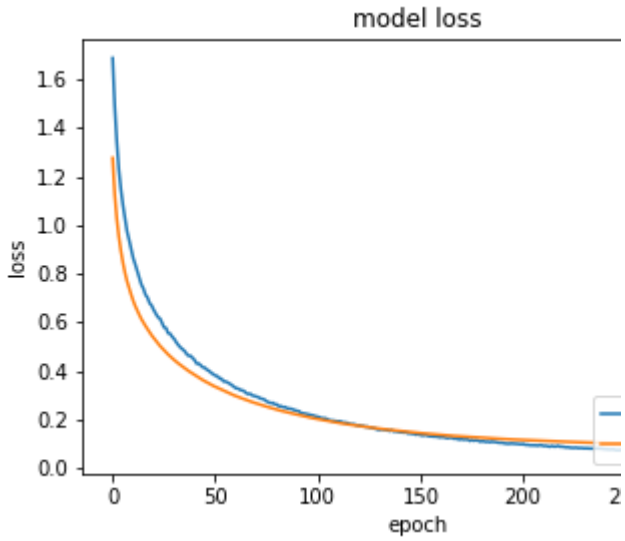
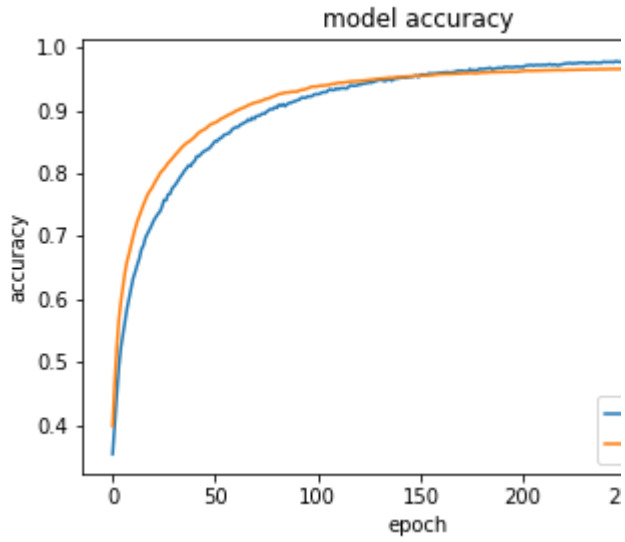


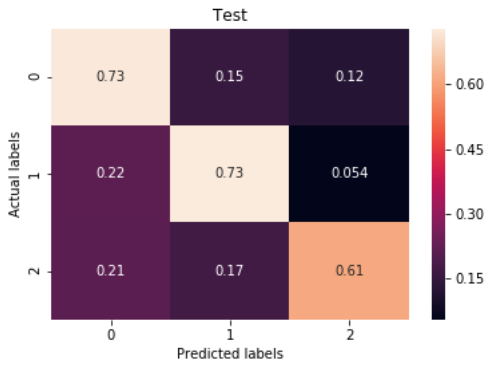
図 7. ニューラルネットワークを用いた正答率

4.3.5 138 社のデータを用いた学習・分析

138 社のデータを使用した場合、データ量が少ないため、学習に使用するデータでは、異なる時点の特徴量ベクトルを無作為に入れ替え、同じラベルをつけた。これらのデータにより学習した分類器を用い、無作為に入れ替えていないデータのクラス分類を行った。結果は以下に示す通りである。データ量をランダムに 100 倍増幅をした結果正答率は 67.7%となった。







4.3.6　LSTM での予測

・3 月 20 日の本会で発表いたします。

5. 課題と本技術の応用先について

5.1　発話区間・コンテキストについて

何らかの映像作品において、今回のように人が話をしているような場合、発話区間という概念があり、コンテキストとしては、発話区間内で整合性が取れているが、一分ごとという形を取ったことにより、このような発話区間の概念が失われている点に課題があると考えられる。
また、現在では、音声合成ソフトで利用する形で実装し、実用化を目指し開発を行っている。

5.2　音声の特徴量について

今回学習したスペクトログラムによる DNN の感情認識エンジンや、東北大学と共同で行っているエンジンも約 50%程度の精度となっており、今回は感情特徴量と呼ばれる値を取り出したが、今後はエンジンの精度をより上げる方向で研究を進めたいと考えている。

5.3　テキスト部分について

テキスト学習については、より精度を上げるため、決算短信や、有価証券報告書などを学習したものをエンジンとして価格推定を行うことを検討している。また、Google の自然言語処理 API は極性判定であるため、個別感情について分類できるものを目指している。

5.4　映像について

現在は表情推定を行っているが、色や OCR で読み取った文字列なども同時に学習評価が出来る仕組みの構築を目指している。

5.5　本技術の活かし方

本技術は、金融情報への応用だけでなく、例えばコミュニケーションロボットなどで、マルチモーダルな AI を用いることで、コンテキストを読み取り、今までより精度の高い認識を行うことが出来るロボットの開発に応用出来るのではないかと考えている。

現在開発を行っているものも、感情を認識し発話を行う段階には至っているが、コンテキストを読み取ることは難しいため、本技術を応用することで、より精度の高い製品を作ること目標としている。

6. 謝辞

本研究は、コンセプトをお話する中で、株式会社 Quick 様に株式会社 StockVoice 様をご紹介頂き、実現致しました。
また、株式会社 Quick 様より日足、歩み値などのレート情報と公募価格についてご提供を頂きました。
東北大学の伊藤・能勢研究室に JTES のデータ提供を受けております。共著に入っていない方で数多くの学生さん、先生の協力を得て実現しました。本当にありがとうございました。

参考文献

[1] 鈴木基之：音声に含まれる感情の認識，日本音響学会誌，Vol. 71，No. 9，pp.484-489，(2015)
[2] デイビッド・ルイス (著), 武田玲子 (翻訳)　(2013)『買いたがる脳』日本事業出版社
[3] 伊勢隆一郎　『人は感情で物を買う』(2015)　フォレスト出版
[4] 日経新聞ニュース https://www.nikkei.com/markets/kigyo/ipo/public-price/　（20180305）

欧州中央銀行総裁の表情解析から見る 量的金融緩和政策の縮小決定

ECB monetary policy analysis based on the facial expression analysis of the ECB presidents

水門善之[1]　勇大地[2]

Yoshiyuki Suimon[1], Daichi Isami[2]

[1]野村證券株式会社 金融経済研究所
[1] Nomura Financial and Economic Research Center
[2]マイクロソフト コーポレーション
[2] Microsoft Corporation

Abstract: The European Central Bank (ECB) decides the euro area's monetary policy at the monetary policy meeting of the Government Council. After the ECB's monetary policy meeting, the ECB president Mario Draghi and the vice-president Vítor Constâncio hold a press conference to explain the monetary policy management. In this research, using facial expression recognition algorithms based on deep learning, we analyzed the presidents' facial expression in the press conference and estimated the emotional indexes such as "Happiness", "Anger", "Sadness" and "Surprise". As a result, we found that the president Draghi's index of "Happiness" decreased and the index of "Sadness" increased just before making major policy changes in the phase of ECB policy normalization. In addition, the vice-president Vítor Constâncio's index of "Happiness" tends to change in the opposite direction to Draghi's index of "Happiness" regardless of the policy change. We believe that the inverse correlation may have the adjustment effect that the impression of the whole press conference will be neutral.

はじめに

欧州中央銀行（European Central Bank，以下 ECB）はユーロ圏の金融政策を，政策理事会で決定している．具体的には，6 週間ごとに開催される政策理事会の参加者（ECB の総裁，副総裁及び理事，ユーロ圏の各国中央銀行総裁）が，多数決で政策を決定する．政策理事会の議事録は，後日 ECB が公表している．

政策の透明性確保の観点から，ECB は政策理事会の後に，総裁及び副総裁の両名が出席する形で記者会見を行っている．日本銀行（以下，日銀）も金融政策決定会合の後に，記者会見を行っているが，日銀の場合は総裁一名によって会見が行われるという点では，ECB と異なる．

ECB の会見では，日銀同様に，金融政策運営に関する説明や質疑応答が行われる．そのため，ECB の景気認識や先行きの金融政策運営を読み解く上でも有用であることから，金融市場における関心は高い．

また ECB は，このような会見を通じて市場とのコミュニケーションを図ると同時に，記者会見の様子を記録した動画データの公表も行っている[1]．情報理論的な定義における情報量で見た場合，会見における発言の文章よりも動画・音声の方が，情報量は大きい．

これらを踏まえ，本研究では ECB の会見動画の解析を行い，文書データには含まれない情報の抽出を試みた．具体的には，記者会見における総裁及び副総裁の表情の変化を，深層学習（ディープラーニング）等の人工知能技術に基づく表情認識アルゴリズムを用いて「喜び」・「怒り」・「悲しみ」・「驚き」・「恐怖」等に分類することで，感情の起伏を指数化し，それらの変化を見ることで，金融政策運営との関係や，ECB のコミュニケーションスタイルの特性を解析した．また，本研究に先立って行った，日銀会見における総裁の表情分析[2]の結果との比較も行なった．

ECB の金融政策運営

本研究では，ECB の政策理事会後の記者会見動画

の分析を行ったが，分析内容の紹介を行う前に，最近の ECB の金融政策運営を振り返りたい. ユーロ圏インフレ率の低迷が続く中, ECB は 2015 年 1 月に，物価安定目標の達成に向けて，量的緩和政策の導入を決定した．具体的には，国債・政府機関債，カバードボンド，資産担保証券（ABS）等の資産の直接購入を通じて，金融政策の伝達メカニズムの活性化や，ユーロ圏における信用供与の円滑化を目指すものであり，非伝統的金融政策とも呼ばれる．このような金融政策の実施により，ユーロ圏の多くの国では内需の回復が促されたほか，世界的な景気回復の流れも相俟って, ユーロ圏の景気は 2016 年後半頃から堅調に拡大を続け，インフレ率にも上昇圧力がもたらされている．

更に，ECB は 2016 年 3 月に，資産購入額の追加的な増額を決める等，量的緩和効果の強化に努めてきた．しかしその後は，2016 年 12 月 8 日に資産購入額の減額を決定したほか，2017 年 10 月 26 日には資産購入額を更に半減させることを決定しており，現状，ECB は資産購入のペースを縮小させる方向に舵を切っている．

一般に，中央銀行が非伝統的金融政策において，資産購入額を減らすことはテーパリング（Tapering, 漸減）と呼ばれる．これは，金融市場の需給バランスを直接左右することから，市場に対する影響度合いが非常に大きい．そのため，量的緩和の縮小局面においては，中央銀行の市場とのコミュニケーションは特に重要となる．

なお，ECB が 2016 年 12 月 8 日に資産購入額の減額を決定した際，ドラギ総裁は記者会見で，資産購入額をゼロに向かわせることについては，政策理事会で協議されなかったと発言している[1]．そのため，2016 年 12 月 8 日の資産購入額の減額は，ECB の資産購入プログラム上保有できる国債の上限（各国の債務に対して ECB が保有できる額は 33%まで）を意識した技術的な調整に過ぎないという見方もできよう．ただし，これらの意思決定を経て，現在 ECB の国債購入のペースは着実に縮小方向に向かっている．

これらを踏まえ，本研究では，ECB が資産購入額の縮小を決定する半年前（2016 年 6 月）から直近までの期間を分析対象とすることで，量的緩和政策からの出口局面に向かう ECB のコミュニケーションスタイルの検証を行った．なお，本分析期間内の総裁はマリオ・ドラギ（Mario Draghi），副総裁はヴィトル・コンスタンシオ（Vítor Constâncio）である．

総裁会見動画の解析

次に，本研究で行った，ECB の政策理事会後の記者会見の動画データの解析内容を紹介したい．記者会見における総裁及び副総裁の表情の変化を解析するため，まず，ECB が公表している会見動画[1]を約 0.5 秒ごとにスクリーンショットを撮り, 解析の対象とする画像データを作成した．そして，作成した各画像データに対して，人工知能モデルを用いて表情の認識を行い，各画像について「喜び」,「怒り」,「悲しみ」,「驚き」,「恐怖」,「軽蔑」,「嫌悪感」,「中立」の各感情の度合いを指数化した．

昨今，Microsoft にて深層畳み込みニューラルネットワーク（DCNN）をベースとした表情認識アルゴリズムの研究が進められている点を踏まえ[3]，本研究では Microsoft の Cognitive Services における表情認識アルゴリズムを感情値の計測，及び登場人物の判別に用いた[4]．そのため，記者会見における表情データ自体を感情値の指数化アルゴリズム作成の為の学習に用いていない点には注意が必要である．

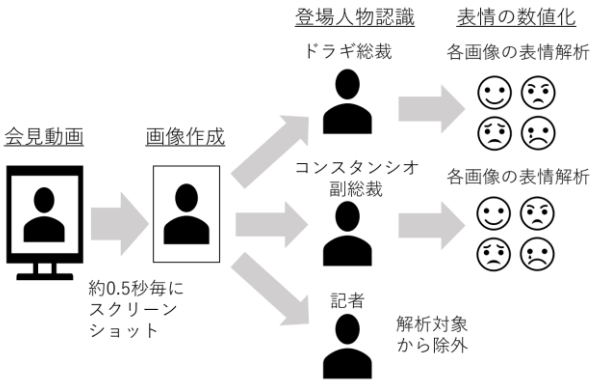


図 1: 会見動画の感情値の計測手順

今回分析対象とした ECB の会見動画が，日銀を対象とした分析[2]で用いた会見動画と異なる点は，映像に映し出されている人物が複数いることだ．具体的には，ECB の会見動画では，総裁，副総裁に加えて，質疑応答で質問を行う記者も動画内で大きく映し出される．日銀の会見動画の場合，表情が明確に映される人物は，基本的に総裁のみである．

複数の人物が映り込む動画を解析するにあたっては, 前述した Cognitive Services の Face API を用いて，

[1] 2016 年 12 月 8 日の会見で, ドラギ総裁は「Tapering has not been discussed today.」と発言. ただし, Tapering という単語の意味について問われた際に,「The word has several meanings depending on who is using it, but the natural way to look at a word like that is to have a policy whereby purchases would gradually go to zero. And that's not been discussed.」と回答しており，資産購入額をゼロに向かわせる政策という意味で，Tapering という言葉を用いているとしている．

総裁と副総裁の顔認識の為の学習を行い，モデル上，総裁・副総裁と認識された人物に対して感情値の計測を行った．なお，総裁・副総裁の両名が映像に映り込む場面では，両名に対して，それぞれ感情値を計測した．

このようにして，約 0.5 秒毎の画像に対して，「喜び」，「怒り」，「悲しみ」，「驚き」，「恐怖」，「軽蔑」，「嫌悪感」,「中立」の各感情値の割り振りを行った．そして，総裁及び副総裁の，各会見中の全感情値の総和に占める，各感情値の総和の割合を算出した．

$$\text{喜びの割合} = \frac{\sum_{t=Start}^{End}(\text{喜び}_t)}{\sum_{t=Start}^{End}(\text{喜び}_t + \text{悲しみ}_t + \text{中立}_t + \cdots)}$$

会見毎に算出した各感情値の割合を検証したところ，資産購入額の減額を発表する前回の記者会見では，総裁・副総裁の表情に関して特徴的な変化が確認された．

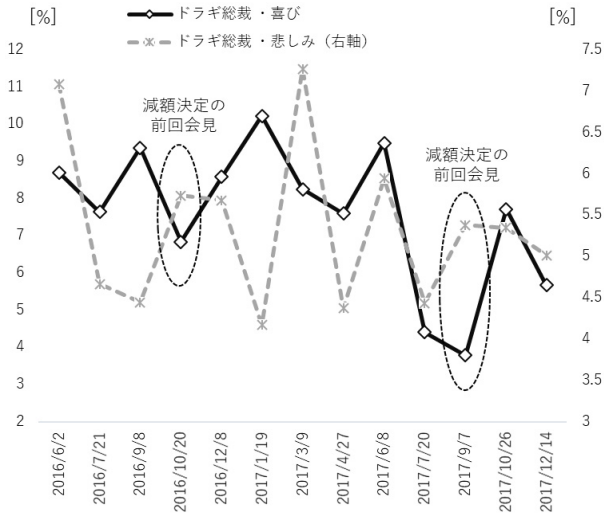


図 2: ドラギ総裁の感情値と政策変更タイミング

例えば，ドラギ総裁の場合，減額決定の前回の会見では,「喜び」の割合が低下すると同時に「悲しみ」の割合が上昇している．このような変化は，日銀を対象とした先行研究で見られたような[2]，政策変更の直前に，総裁のネガティブな表情の割合が高まるという特徴と共通している．ただし，日銀を対象とした先行研究で確認されたような，大きな感情の変化は見られず，ECB の場合，図 2 に示すように，ある程度のレンジの中での変化に留まっている．意識的か否かはさて置き，ドラギ総裁の表情の変化は，ある程度コントロールされているように見られる．

更に，前述の通り，ECB の記者会見が，日銀と大きく異なるのは，会見に総裁と副総裁の両名が出席している点だろう．そこで，今回計測した両名の表情値を比較すると，政策変更の有無に関わらず，ドラギ総裁の「喜び」の割合が低下する会見において，コンスタンシオ副総裁の「喜び」の割合が上昇する傾向が見られた（図 3）．

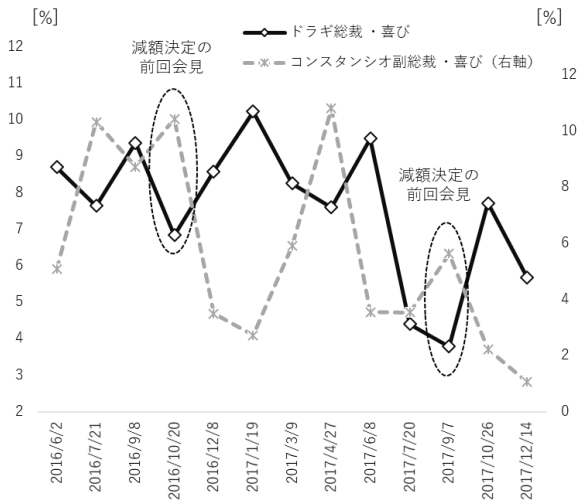


図 3: ドラギ総裁とコンスタンシオ副総裁の感情値

また，図 4 では，ドラギ総裁，コンスタンシオ副総裁の，前回会見からの「喜び」の割合の変化を比較した．これを見ても，ドラギ総裁の「喜び」の割合が低下している会見では，コンスタンシオ副総裁の「喜び」の割合が上昇する傾向が確認できよう．そもそも，コンスタンシオ副総裁は回によっては発言を行わないこともあるが，そのような場合においても，ドラギ総裁と併せて表情を分析することで，表情の変化に特徴が見出せる点は，興味深いと言えよう．

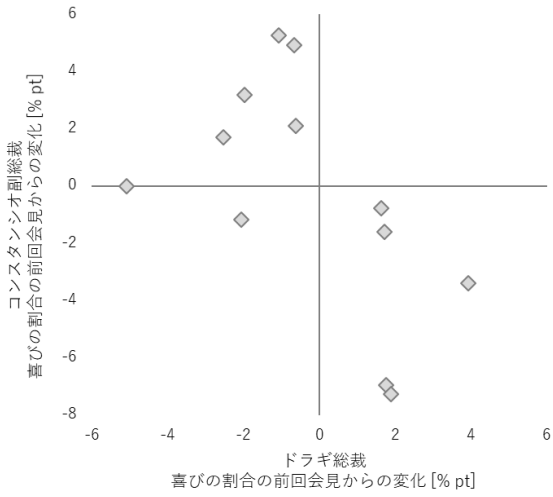


図 4: 感情値の前回会見から変化幅の比較

まとめと解釈

量的金融緩和政策からの出口（量的緩和の縮小局面）において，中央銀行と金融市場とのコミュニケーションは重要度合いを増す．

本研究では，現在，量的緩和の縮小に向けた金融政策運営を行っている ECB を対象に，市場との主要なコミュニケーションの場である政策理事会後の記者会見の動画分析を行った．その結果，ドラギ総裁の表情には，政策変更の直前には「喜び」の割合が低下し，「悲しみ」の割合が上昇するという変化が見られた．加えて，ドラギ総裁の隣に座っているコンスタンシオ副総裁は，政策変更の有無に関わらず，「喜び」の割合が，ドラギ総裁とは逆の方向に変化する傾向が確認された．両名の感情値の変化が逆相関関係にあるということは，ECB の記者会見において，全体としての印象をニュートラルに近づけるような調整効果がもたらされている可能性があると言えよう．更に，ECB のように，複数人によって記者会見が行われる方式では，会見での個人の情報発信スタイルの振れによって，市場に意図せざるメッセージが伝わるリスク等を軽減できる可能性もあることから，結果，中央銀行による金融市場へのメッセージ発信の安定性に資する可能性があると考えられよう．

留意事項

本稿は，著者の個人見解を表すものであり，野村證券株式会社および Microsoft Corporation の公式見解を表すものではありません．

参考文献

[1] European Central Bank, Press conferences https://www.ecb.europa.eu/press/pressconf

[2] 水門善之，勇大地：日銀総裁会見の表情解析に基づく感情値の計測と金融政策変更との関係，人工知能学会第 19 回金融情報学研究会，(2017)

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer and Zhengyou Zhang: Training deep networks for facial expression recognition with crowd-sourced label distribution, ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction, Pages 279-283

[4] Microsoft Cognitive Services Emotion API https://azure.microsoft.com/ja-jp/services/cognitive-services/emotion/

A possible approach to enhancing popular Japan equity market strategies with an emphasis on machine learning solutions

西山　昇

Noboru Nishiyama

Dragons' Desk Ltd. / 千葉商科大学（会計大学院）

Abstract: We analyze the impact of advanced machine learning methods on the performance and risk characteristics of popular Japan equity strategies over the last 10 years. We then propose a possible approach to enhancing each strategy through advanced risk control and we analyze the results.

Key words: Historical back-testing, Machine learning, EM algorithm, GARCH process, Optimization

1. はじめに

ここ数年､預金から投資への関心の高まりもあり、投資信託、ＥＴＦ市場が活況を呈している。その中でも機関投資家に注目されてきたファクター投資（スマートベータ）と呼ばれるポートフォリオ運用戦略がある。

近年 GPIF（年金積立金管理運用独立行政法人）に採用され、ファクター投資（スマートベータ）型の ETF（Exchange-Traded Fund:上場投資信託）が数多く開発されてきた。

ファクター投資（スマートベータ）とは、市場の変動を説明するシステマティックファクターにポートフォリオを連動させる運用手法である。

パッシブ型運用におけるベンチマークに多様性を持たせたところに特徴がある。さらに人間の意思決定に影響を受けないクォンツの最適化手法を活用することで運用手数料の低下にもつながっている。

本研究では、ファクター投資（スマートベータ）において最近ポピュラーとなっている最小分散戦略と複数の代表的な財務指標によるファンダメンタルファクターへの個別ティルト戦略のリスク・パフォーマンス特性をバックテストにより確認している。

今回のバックテストでは、リスクモデルとして GARCH プロセスを統合した EM アルゴリズムによる統計的マルチファクターモデルを使用する。

その特徴は次の 3 点である。バックテストのリスクモデルには特に（3）が関連する。[1]

(1)意思決定のための学習アルゴリズム
推定期間が経時的に進行するなかで一定の期間のウィンドウデータが更新される。EM algorithm は、そのたびに新しい統計的ファクターを自動的に探索する。

(2)人的なインプットなしのファクター選択
どのファクターになるかについて人的な関与がない。他の手法ではファンダメンタルのクロスセクションデータからファクター構造を人間が解釈する必要がある。EM アルゴリズムではファンダメンタルのクロスセクションデータを属性分析に使うものの、基本的には潜在ファクター構造を明らかにするが、ファクターを決定するのに人的なインプットは求められない。

(3)意思決定に有効な分散・共分散の予測
ファンドマネージャは近未来の分散・共分散行列の日次更新を受け取る。そしてリスクリターンプロファイルを改善する、あるいは求めるアロケーションをメンテナンスするために、どのようにリバランスするのかを意思決定する。

ファクター投資（スマートベータ）のバックテストの期間は 11 年 1 か月（2007 年 1 月 1 日～2018 年 2 月 1 日）、月次でのリバランスを実施する。

対象となる代表的な投資戦略は次のとおりである。

(1) 最小分散戦略(MinVar)
(2) 1 株あたりの純資産／調整済み株価（株価純資産倍率（ＰＢＲ）の逆数）(以下 B/P)
(3) 1 株あたりの配当金額（四半期分を合計）／調整済み株価（配当利回り）（以下 Div/P）
(4) 1 株あたりの売上高（四半期分を合計）／調整済み株価（株価売上高倍率の逆数）（以下 Sales/P）

最小分散戦略では、リバランスごとに分散共分散を最小化する最適化を行う。最小分散を基準にポートフォリオを月次で構築する。

分散には EM アルゴリズムを活用した統計的マルチファクターモデルのリスク予測値を採用する。リスク予測値は日次で更新される。

またファクターティルト戦略では、各ファクターを目的関数として最大化する一方、リスクモデルから算出する予測値のひとつであるベータ値を制約条件として使用する。

日本株ポートフォリオとして、ＴＯＰＩＸ採用銘柄をユニバースとしたポートフォリオ最適化を実行する。

2. 投資戦略とその組み合わせ

2.1 ファクター投資（スマートベータ）

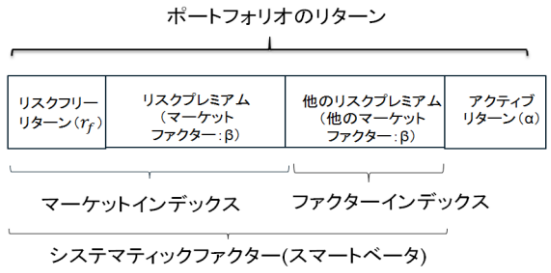
（図 2 - 1）では、ポートフォリオのリターン表現を分解したイメージを図で示してある。

最小分散の分散とは、ポートフォリオのリターン全体の分散共分散を指す。

またそれを最小にする考え方である。そのことにより、アクティブリターン(α)の最大化とトレードオフになっている。

ファンダメンタルデータを基礎とする、B/P、Div/P、Sales/P は、（図 3 - 1）のファクターインデックスにあたる部分であり、リターンの源泉として各ファクターを最大化する方向に調整している。

問題は、個別ファクターを最大化しているとしても、それが純粋に B/P ファクターにのみにティルトがかかっているのか、マーケットインデックスの影響を受けずに、純粋にアルファ効果を抽出しているのかの判断がむつかしい。そこにリスクモデルを考慮する効果が生じることになる。



（図 2 - 1）ファクター投資（スマートベータ）の概念図[2]

2.2 統計的マルチファクターモデル

リスク予測値（分散共分散、ベータ）を計算するモデルとして APT 型の統計的ファクターモデルを適用する。[3]

$$R: return, F: risk\ \ factors, \beta: sesitivity$$

$$R = \tilde{\beta}_1 F_1 + \tilde{\beta}_2 F_2 + \tilde{\beta}_3 F_3 + \cdots\cdots\cdots \alpha + \varepsilon$$

（式 2 - 1）統計的マルチファクターモデルのリターン表現

$$\Sigma: \text{var}\,iance - \text{cov}\,ariance(\text{分散・共分散})$$

$$\Sigma = \tilde{\beta}_1\,\tilde{\beta}_1'\sigma_1^2 + \tilde{\beta}_2\,\tilde{\beta}_2'\sigma_2^2 + \tilde{\beta}_3\,\tilde{\beta}_3'\sigma_3^2 + \cdots\cdots\cdots + D_\varepsilon$$

（式 2 - 2）統計的マルチファクターモデルのリスク表現

（式 2－2）において、Σはトータルリスクとしての分散共分散、右辺第二項 $D\varepsilon$ はアンシステマティック（非組織的）リスク、右辺の項全体から $D\varepsilon$ を除いた部分をシステマティック（組織的）リスクと呼ぶ。

よって（トータルリスク）＝（システマティックリスク）＋（アンシステマティックリスク）と読み替えることができる。

繰り返しになるが、最小分散とは分散共分散Σを最小化することで配分比率を決定している。また B/P、Div/P、Sales/P の各ファクターは（式 3－2）には潜在的にしか登場してこないが、統計的ファクターによりファンダメンタルファクターへのリスク配分を調整する。

2.3 最小分散戦略(MinVar)

バックテストの最適化条件は次のとおりである。

目的関数：各（月次）リバランス時に分散（予測値）最小化（月次リバランス）

バックテスト期間：2007 年 1 月―2018 年 2 月

ユニバース：TOPIX

期初ポートフォリオサイズ：10 億円

制約条件

・最大ターンオーバー：リバランス毎 10%（一方向）

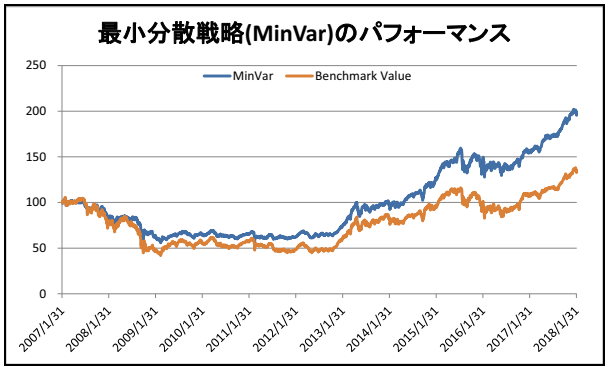
・個別銘柄の最大保有サイズ：最小（ポートフォリオの 5%、あるいは、ベンチマークウェイト×5）

・セクター制約(TOPIX 33 業種)

・フルインベストメント

リスクモデル：EM アプリケーションズ日本モデル

(図 2-2)は、最小分散戦略（MinVar）とユニバースである TOPIX をベンチマークとしたパフォーマンス推移のグラフである。全期間では、最小分散戦略がベンチマークをアウトパフォームしている。



（図 2-2）最小分散戦略(MinVar)のパフォーマンス

2.4 ファンダメンタル・ティルト戦略

ファンダメンタルファクターとして採用したのは次の 3 種類である。それぞれデータソース(S＆P Capital IQ)における定義(数式)を示す。

（1） B/P（株価純資産倍率（ＰＢＲ）の逆数）

Definition: Ratio of book value to market value of common equity

FORMULA

Data Source: Capital IQ PIT

$$BP_{i,t} = \frac{BVPS_{i,t}}{Close_{i,t}}$$

CloseM : Adjusted Closing Price (Adjusted Closing Price)
BVPS : Book Value Per Share (4020)

（2） Div/P（配当利回り）

Definition: The ratio of trailing four quarter dividends per share to current stock price.

FORMULA

Data Source: Capital IQ PIT

$$DivP_{i,t} = \frac{\sum_{j=0}^{3} DPS_{i,t-j}}{CloseM_{i,t}}$$

CloseM : Adjusted Closing Price (Adjusted Closing Price)
DPS : Dividends Per Share (3058)

（3） Sales/P(株価売上高倍率の逆数)

Definition: The ratio of trailing four quarter sales to average market value of common equity over the same period.

FORMULA

Data Source: Capital IQ PIT

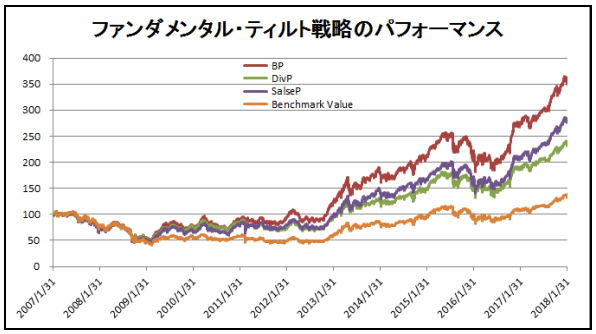
$$SP_{i,t} = \frac{\sum_{j=0}^{3} TRIN_{i,t-j}}{CloseM_{i,t} \times \frac{1}{4} \times \sum_{j=0}^{3} WASOF_{i,t-j}}$$

WASOF : ECS Common Shares Outstanding (24152)
TRIN : Sales/Turnover (Net) - Quarterly ([28])

目的関数（各ファンダメンタルファクターティルト）
B/P(株価純資産倍率の逆数)
Div/P（配当利回り）
Sales/P(株価売上高倍率の逆数)
最小分散戦略(MinVar)とは、目的関数と次の制約条件以外を同一とする。
各目的関数の最大化
ベータ＝1
ベータを算出するリスクモデル：EM アプリケーションズ日本モデル

(図 2-3)は、ファンダメンタルの B/P(株価純資産倍率の逆数)、Div/P（配当利回り）、Sales/P(株価売上高倍率の逆数)と、ユニバースである TOPIX をベンチマークとした場合のパフォーマンス推移である。

3 ファクターともバックテスト全期間(2007 年 1 月―2018 年 2 月)では、ベンチマークをアウトパフォームしている。同期間を通じてのパフォーマンスの順位は、B/P＞Sales/P＞Div/P＞TOPIX であった。



（図 2-3）ファンダメンタル・ティルト戦略のパフォーマンス

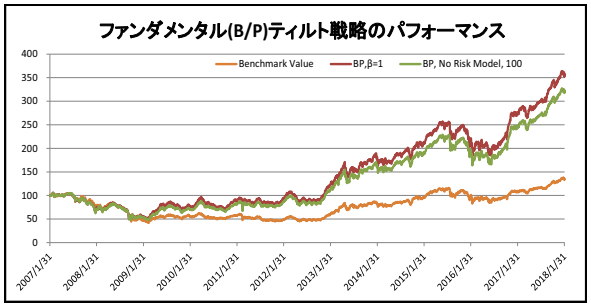
2.5 ティルト戦略とリスクモデル

ファンダメンタル・ティルト戦略のリターンの源泉を確認するのにベータ制約をつけた場合とつけなかった場合のバックテストの比較をおこなう。

ファンダメンタルデータがリターンの源泉であることは確実と考えられるものの、ファンド運用者にとって B/P はリスクファクターでもある。

そのためリスク低減させる機能を維持しつつアルファの源泉としても活用すべきファクターである。

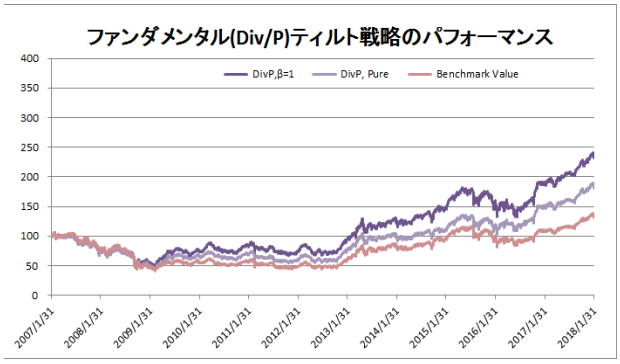
そこで最適化にリスクモデルを使用した場合（$\beta=1$）と使用しなかった場合のパフォーマンスの比較を全期間に対して実行する。



（図 2-4）B/P ティルト戦略（β 制約の有無）のパフォーマンス

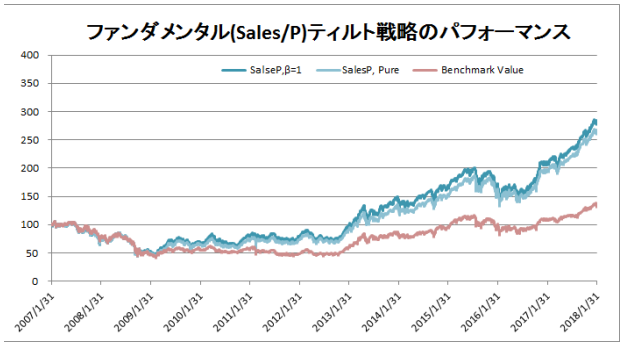
(図 2-4)では、B/P ティルト戦略のβ＝1 の制約有りと制約無し（リスクモデル不使用）のパフォーマンス推移をグラフ化している。

B/P ティルト戦略の制約あり（β＝1）は、制約なしのパフォーマンスに対してバックテスト全期間を通じて年率ベースで約 3%パフォーマンス優位である。



（図 2-5）Div/P ティルト戦略（β制約の有無）のパフォーマンス

同様に(図 2-5)の Div/P ティルト戦略の制約あり（β＝1）は、制約なしのパフォーマンスに対してバックテスト全期間を通じて年率ベースで約 2.4%パフォーマンス優位である。



(図 2-6) Sales/P ティルト戦略（β制約の有無）のパフォーマンス

同様に(図 2-6)の Sales/P ティルト戦略の制約あり（β＝1）は、制約なしのパフォーマンスに対してバックテスト全期間を通じて年率ベースで約 0.6%パフォーマンス優位である。

リスクモデルを使用してβ制約を設定した方が、数値の大小はあるもののβ制約を設定しなかった Pure（リスクモデル無し）の場合よりアウトパフォームする結果となった。

3. 各戦略のリスク特性

3.1 リスク・パフォーマンス指標

バックテストの結果をまとめると次のようになる。各戦略の中でリスク・パフォーマンス関連指標がもっともすぐれている数値がカラー（黄色）となっている。

たとえば、全期間を通じて Realized Standard Deviation（実現標準偏差）がもっとも低かったのは、18.1 の最小分散戦略（MinVar）であり、同戦略の最大ドローダウン幅が 46.6%と最小となっている。

一方 TOPIX は、最大ドローダウンが 60.2 とバックテストの中では最大の値となっている。

Realized Tracking Error がもっとも小さい値だったのは、Div/P の 5.9 だった。

またプラスに評価される項目がもっとも多かったのは、ファンダメンタル B/P（株価純資産倍率（ＰＢＲ）の逆数）ティルト戦略（β=1 制約）だった。

	Min Var	B/P, Pure	B/P, β=1	Div/P, Pure	Div/P, β=1	Sales/P, Pure	Sales/P, β=1	TOPIX
AR of Return	6.4	9.3	12.3	5.8	8.1	9.3	9.9	2.8
RStDeviation	18.1	22.4	22.2	21.0	21.7	22.2	22.5	23.4
Sharp Ratio	0.4	0.4	0.6	0.3	0.4	0.4	0.4	0.1
Excess Return	3.7	6.5	9.5	3.0	5.4	6.5	7.1	
RTrError	9.5	7.5	8.6	6.2	5.9	8.3	8.6	
Infor Ratio	0.4	0.9	1.1	0.5	0.9	0.8	0.8	
Total Return	98.9	166.6	258.0	85.5	136.7	165.1	181.8	35.3
Max Drawdown	46.6	55.9	54.0	58.4	54.3	57.9	56.2	60.2

AR of Return: Annual Rate of Return, RStDeviation: Realized Standard Deviation, RTrError: Realized Tracking Error, Infor Ratio: Information Ratio　2007 年 1 月 1 日～2018 年 2 月 1 日（11 年 1 か月）

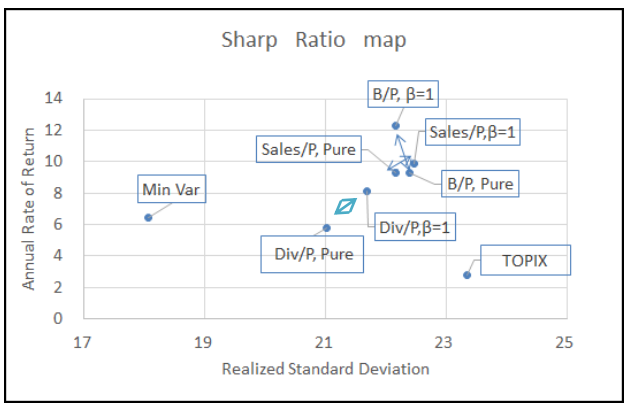
（表 3-1）各戦略のリスク・パフォーマンス指標の比較

3.2 リスク・リターン特性

リスク・リターン特性をみるのに全期間を通じた個別戦略のシャープレシオ（SR）とインフォメーションレシオ（IR）をグラフにプロットする。

SR＝Annual Rate of Return／Realized Standard Deviation

IR＝Excess Return／Realized Tracking Error

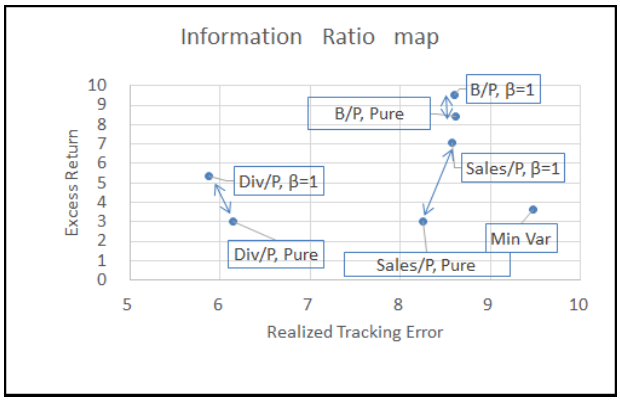


（図 3-1）各戦略のシャープレシオの分布

シャープレシオが、もっとも高かったのは、ファンダメンタル・ティルト戦略の中の Book To Price(株価純資産倍率の逆数)ティルト＆β=1 の 0.55 だった。次に大きい値は、Sales/P(売上高株価比率)ティルト＆β=1 の 0.44、それに続いて、Book To Price(株価純資産倍率の逆数)ティルトのリスク制約無し（B/P, Pure）が 0.42 となっている。

各ファンダメンタルファクターのリスク制約有りと無しに注目するといずれもリスク制約（β=1）有りのシャープレシオ（SR）の方が高い値となっている。

次にインフォメーション・レシオ（IR）の分布を確認する。



（図 3-2）各戦略のインフォメーションレシオの分布

インフォメーションレシオが、もっとも高かったのは、ファンダメンタル・ティルト戦略の中の Book To Price(株価純資産倍率の逆数)&(β=1 制約あり)の 1.10 だった。

それに続くのは、Book To Price(株価純資産倍率の逆数)&(β=1 制約有り)の 0.98 である。

Div/P（配当利回り）&(β=1 制約有り)、Sales/P(株価売上高倍率) &(β=1 制約有り)ファクターは、それぞれ 0.91、0.83 となっている。

こちらもリスク制約有りと無しに注目するといずれもリスク制約（β=1）有りのインフォメーションレシオ（IR）の方が高い値となっている。

4. 考察

今回は最適化の目的関数の違いとリスクモデルの有無によるリターンを比較するバックテストを行った。

最小分散戦略をベースとしてファンダメンタルファクターにティルトする代表的な戦略のパフォーマンスを確認している。

個別のファンダメンタルファクターの最適化ではβ＝1 の制約がある場合とない場合を比較している。

ここでのベータとは、対ベンチマークとの線形関係を前提とした感応度の値ではなく、リスクモデルから算出された各個別銘柄が持つリスク値である。

ファンダメンタルファクターのティルト戦略では、Pure として制約条件なしで設定した結果を提示した。β＝1 制約（リスクモデル）がある場合と無い場合とでパフォーマンスに差がついたのは、どこに要因があるのだろうか。

ベータを 1.0 に制約することは、BP ティルトをマーケットニュートラルに調整していると考えられる。

ファクターティルトは、マーケットに意図せざるベットをしており、リスクモデル無しではマーケットリスクとの相関を調整しきれていない。

現時点から過去 10 年を振り返れば、実現した相関を見ることができる、しかしこれをバックテストに利用することは現実的でない。それは、われわれはマーケット(TOPIX) とファンダメンタルファクターの将来の相関がどうなるのかを事前に知ることはないからである。

だからこそフォワードルッキングなファクターモデルを使うことが適切なソリューションになると考える。

5. おわりに

今回は長期にわたるバックテストを実施して各戦略の実現（シミュレーション）リスク・リターンを確認した。

大切なポイントは、ファンダメンタルファクターの中には、リスクファクターとアルファファクターの両面があることである。

通常のマーケットでは、運用者の視点からリスクファクターとして取り扱われる場合が多い。

そのためマーケットリスクをヘッジする適切な方法はポートフォリオ最適化とリスクモデルを適切なフレームワークで適用することにある。

今回のバックテストでは、長期にわたるシミュレーションであるために制約条件の設定に（解が収束しないケースがあり）工夫が必要だった。

またリスクモデルには、分散（共分散）を高い精度で予測すること、同時に新しい外部ショックが発生したときに、その波及効果をリスク値にすぐ反映することも求められる。

統計的マルチファクターモデルの枠組みにすることにより、リスクの日次エクスポージャをモニターすることができる。ただ今回のバックテストの設定では月次リバランスである。

リスクモデルによるベータ制約は、ファンダメン

タルファクターをマーケットベットから純粋化させる効果があることがパフォーマンスの数値により示された。

純粋化が実現するには、ダイナミックに実行されるフォワードルッキングな（短期）予測モデルを使う合理性があると考える。

純粋化についてポートフォリオ理論の側面から説明することができれば、意図せざるベットが発生していることに早く気付き、リスクエクスポージャを調整可能となる。

究極的には、アルファを常に最大化するようにファンダメンタルファクターを自動的に選択、最適化できる仕組みに向かうことになるだろう。

謝辞

本稿を作成するのにデータ処理面でのアシストをしてくれた David Andorosoni 氏に感謝したい。またバックテストにあたりスタンダード＆プアーズ（S＆P）Capital IQ（キャピタル IQ）様よりファンダメンタルファクターに使用した財務データと ClariFI（バックテストツール）の提供を受けたことに謝意を表する。

参考文献

[1] The impact of North Korea risk on the Japan equity market: what do AI based risk models tell us? Noboru Nishiyama 第 20 回 SIG-FIN 発表原稿, January, 2017. (http://sigfin.org/?plugin=attach&refer=019-06&openfile=SIG-FIN-019-06.pdf)

[2] 加藤康之, スマートベーター新時代の投資理論－, 応用経済時系列研究会第 23 回談話会資料, 2016 年 2 月 16 日

[3] EM Applications, Ltd.,

（https://emapplications.com/index.php?q=research/statistical-factor-model/stat-factor-model）

人工市場を用いた市場流動性に影響を与える要因の検出

Detection of factors influencing market liquidity using artificial market

益田 裕司[1*]　水田 孝信[2]　八木 勲[1]
Yuji Masuda[1]　Takanobu Mizuta[2]　Isao Yagi[1]

[1] 神奈川工科大学情報学部
[1] Faculty of Information Technology, Kanagawa Institute of Technology
[2] スパークス・アセット・マネジメント株式会社
[2] SPARX Asset Management Co. Ltd

Abstract: 市場の「流動性」に関心が高まっている．流動性は金融市場の盛況を表す目安とされ，「取引のしやすさ」ともいうことができる．実証研究では，それぞれの研究目的に沿うような流動性指標を用いて，その有用性を議論していた．しかし，それらの指標が市場内外のどの要因の影響を受けて変化するのかは明らかにされていない．そこで本研究では，市場内のどの要素が，流動性指標に影響を与えるのかを人工市場を用いて調査した．その結果，4 つの流動性指標（Volume，Tightness，Resiliency，Depth）は，人工市場のパラメータのうち，1）ティックサイズ，2）投資家の注文戦略を決める成分（ファンダメンタル成分，テクニカル成分，ノイズ成分）から影響を受ける可能性があることが分かった．

1　まえがき

金融市場の盛況を表す目安とされる「流動性」に関心が集まっている．一般に流動性が高い状況とは,「その時々で観察される『市場価格』に近い価格で，市場参加者が売りたい（あるいは買いたい）量を，速やかに売れる（あるいは買える）」状況が想定されることが多い [1].

流動性に関する研究は，特に実証研究の分野で多数行われ，さまざまな知見が得られている．しかしながら，何をもって流動性とみなすかは実証研究の調査目的ごとに異なることが多い．例えば，市場価格のボラティリティの大きさや，市場参加者の売買が市場価格に大きな影響を及ぼさないことを流動性と結びつけることも多く，調査目的によって流動性の定義は異なってくる．それに伴い，流動性を計測するために使用される指標も研究ごとに異なることが多く，ある研究で得られた流動性に関する知見が他の研究で得られた知見と整合が取れているのかどうか判断することは困難である．

このように実証研究では対処困難なものに対応する手法の 1 つに，人工市場を用いる手法がある．人工市場は，社会シミュレーションの 1 つであり，計算機上に仮想的に構築されたマルチエージェントシステムの金融市場のことを指す．人工市場におけるエージェントは仮想的な投資家であり，現実の投資家の特性がモデルとして組み込まれている．そして，エージェントらに金融資産の取引をさせることで市場がどのようなふるまいをするかを確認することができる．市場側にモデル化した規制や制約を組み込むことで，エージェントの振る舞いや市場にどのような影響が現れるかを検証することもできる．

これまでに人工市場を使用した研究においていくつかの有益な知見が得られているが [3][4]，人工市場シミュレーションを用いた研究では，流動性そのものに着目した研究は行われていない．

そこで本研究では，金融市場における実証研究で良く用いられている流動性指標の間にどのような関係があるかを人工市場を用いて調査した．具体的には，流動性の代表的な 4 つの評価軸（Volume，Tightness，Resiliency，Depth）を計測するための代表的な指標に注目して，それら指標の関係をティックサイズなどの人工市場内のパラメータを変化させながら調査する．

本論文の構成は以下のとおりである．まず 2 章において流動性について説明する．また流動性の実証研究についても説明する．3 章では本研究で用いた人工市場モデルについて説明する．4 章では本研究で行う実験の詳細や得られた結果について説明する．最後に 5

*連絡先：神奈川工科大学情報学部
神奈川県厚木市下荻野 1030
E-mail:s1421036@cco.kanagawa-it.ac.jp

章では，本研究のまとめと今後の課題について述べる．

2　流動性

2.1　流動性の定義

流動性には，確立された唯一の定義というものは存在していない．しかし「流動性の高い市場とは，大口の取引を小さな価格変動で速やかに執行できる市場である」[2] といった定義は良く引用されている．流動性を計測する実証研究では，この定義のもとで 4 つの評価軸（Volume，Tightness，Resiliency，Depth）が提示されることが多い．そして，1 つの評価軸にいくつかの種類の指標が用いられる．土川ら [5] は，4 つの評価軸を視覚的に整理し，図 1 のようにまとめた．また価格の騰落率（ボラティリティ）についても流動性を表す目安と使われることもあるが，以下 4 つの評価軸について説明する．

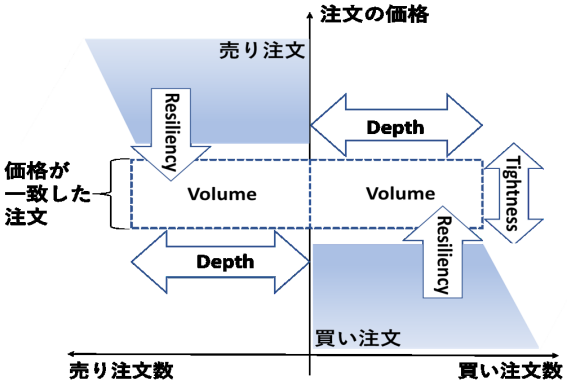


図 1: 流動性の概念整理 [5]

2.1.1　市場の取引量（Volume）

第 1 の評価軸として，市場の取引量（Volume）があげられる．市場の取引量が多ければ，頻繁な取引や短期間での大口の取引がより容易になると考えられる．そのため，Volume が大きければ流動性が高いといえる．

Volume の指標としては，出来高，売買回転率をみる方法と，取引がない時間間隔，ゼロリターン率をみる方法の 2 種類が存在する．

出来高は，市場の取引量を直接捉えることができ，データ取得や時系列比較も容易なため，最も頻繁に利用されている流動性指標である．しかし，出来高は観測期間中に偶然取引があったことを示しているにすぎない．そのため，出来高が高いため流動性があると考え注文を出したとしても，注文した時点では市場に残っている注文がなく，なかなか取引が成立することがないといったことが起こり得る．

2.1.2　買い手と売り手の提示価格の差（Tightness）

第 2 の評価軸は，買い手と売り手の提示価格の差（Tightness）があげられる．提示価格の差が狭ければ，市場参加者の意図する価格から離れず取引が行える．そのため Tightness が小さければ流動性が高いといえる．

Tightness の指標としては，買い手の 1 番高い注文の価格である最良買い気配（ベスト・ビッド）と売り手の 1 番安い注文である最良売り気配（ベスト・アスク）の乖離幅として定義されるビッド・アスク・スプレッドがよく用いられる．

ビッド・アスク・スプレッドも，出来高と同じくデータの取得が容易であり，出来高と同じようにより多くの市場参加者に利用されている流動性指標である．しかし，ビッド・アスク・スプレッドの算出に使うベスト・ビッド，ベスト・アスクは市場に残っている注文の中で，市場参加者に最も有利な価格のみを提示しており，その価格で取引できる量については提示していない．そのため，ビッド・アスク・スプレッドが小さいため流動性が高いと考え注文を出したとしても，注文量と比べ取引できる量が少なければ，市場参加者が望んだ量を取引することはできない．

2.1.3　市場の復元力（Resiliency）

第 3 の評価軸は，市場の復元力（Resiliency）があげられる．取引が成立し，上下した市場価格が元の価格に戻る早さのことを指す．価格に大きな騰落が起きた場合でも，迅速に騰落前の価格へ戻ると，取引を迅速かつ円滑に行える．そのため Resiliency は小さければ，流動性が高いといえる．Resiliency の指標としては，値幅･出来高比率，ベスト･ビッドの枚数回復速度，ILLIQ といったものが存在する．

値幅･出来高比率は，日中の市場価格の最高値と最小値の幅を 1 日の出来高で除した指標で，その日の取引での平均的な価格変化を示している．取引が成立することで板に残っている注文が少なくなり板が薄くなっても，市場に復元力があれば速やかに板は回復し，売買に伴う価格変化は小さくなる．そのため，1 つの取引の平均的な価格変化である値幅・出来高比率が小さければ，市場の流動性は高いといえる．しかしながら，値幅･出来高比率が低かったとしても，価格が最高値と最小値を行き来するような，日中の動きが激しく，市場参加がしにくい状況も存在する．

市場インパクト指標である ILLIQ は，非流動性指標ともいわれ，売買金額 1 単位あたりの価格変化の月中平

均値である [6]．ILLIQ の値が大きいと売買金額 1 単位で動くリターンが大きく，流動性が低いとされている．

2.1.4　市場の厚み（Depth）

第 4 の評価軸は，市場の厚み（Depth）があげられる．ベスト・ビッド，ベスト・アスクに近い価格で出されている注文量であり，現在の市場価格に影響を与えることなく取引ができる数量を示す．ベスト・ビッド，ベスト・アスクに近い価格での注文量が多い状況を市場が厚い状況と呼ぶ．市場に厚みがあれば，市場参加者が意図した価格と市場価格との差が小さくなる．また市場に残っている注文が多く，取引成立した際の市場価格が振れにくくなる．そのため，Depth が大きければ，流動性が高いといえる．

現実世界での Depth は，最良気配から 5，8，10 ティック離れた Depth 情報が公開され，流動性指標として利用されるようになった．土川ら [5] は，各営業日のビッド，アスクの価格ごとの注文量を示す注文板におけるベスト･アスク枚数の出現する頻度分布の中央値を Depth の指標として用いることを提案した．

2.2　実証研究

村永 [7] は，日本の株式市場を対象として流動性の動学的な側面を研究した．1995 年 10 月 2 日から 1996 年 9 月 30 日までの東京証券取引所の電気機器指数を構成している個別株式の市場データを用いてクロス・セクション分析を行い，Tightness，Depth，Resiliency に対応する代理指標について分析した．Tightness はビッド・アスク・スプレッド，Depth はマーケット・インパクト（取引執行によるビッド・アスク・スプレッドの変化率を出来高で割った値），Resiliency は市場弾力性（取引執行によるビッド・アスク・スプレッドの変化率を取引執行前の水準に戻るまでの時間割った値）で算出した．

分析の結果，ビッド・アスク・スプレッド，マーケット・インパクト，市場弾力性のいずれの流動性指標をみても，取引頻度と正の相関があることがわかった．

また 1998 年 4 月 13 日に東証がおこなったティックサイズ切り下げによる影響についても分析を行っている．結果，ティックサイズ切り下げはビッド・アスク・スプレッド，ボラティリティを小さくし，取引頻度を増加させたことがわかった．そのため，ティックサイズの変更は市場の流動性に影響を及ぼすことを指摘した．

Chung[8] は，呼値刻みが株価水準で変わるクアラルンプール証券取引所の時系列分析を行った．1996 年から 2001 年の上場銘柄の月次データを使い，Depth の対数を被説明変数，株価の逆数，出来高の対数，売買回転率，ボラティリティの 4 つを説明変数とし，クロス・セクション分析を行った．ここでの Depth は売買それぞれの和で千株単位の株数としている．

計測した結果，Depth は株価水準とボラティリティに対し負の相関を持ち，出来高や回転率には正の相関をもつことがわかった．

3　人工市場モデル

3.1　市場構成

本研究では，Mizuta ら [10] の人工市場モデルを基に，人工市場モデルの構築をおこなった．

本モデルは，1 つの資産のみを取引対象とする．エージェントは n 体おり，エージェント $j = 1$ から，$j = 2, 3, 4, \cdots$ と順番に注文を出す．最後のエージェント $j = n$ が注文を出すと，次の時刻にはまたはじめのエージェント $j = 1$ が注文を出していく．時刻 t はエージェント 1 体が注文を出すたびに，δt だけ増える．δt は注文が発生する時間間隔であり，注文は時間的にランダムに発生し，ポアソン分布に従うと仮定する．そのため，δt は平均 δo の指数乱数とする．注文数は常に 1 とした．各エージェントが持つキャッシュ量は無限とし，資産を何単位でも買うことができる．また空売りも可能とした．

3.2　価格決定メカニズム

価格決定メカニズムは買い手と売り手が価格を提示し，両者の提示価格が合致するとその価格で取引が成立する，連続ダブルオークション方式（ザラバ方式）とした．ティックサイズを ΔP とし，注文価格を求めるときに ΔP より小さい端数は買い注文の場合は切り捨て，売り注文の場合は切り上げる．買い注文価格より安い売り注文，または売り注文より高い買い注文が注文板に既に存在していれば，取引が即時成立する．取引が成立しなかった場合は注文を残す．本研究では，取引が即時成立する注文を成行注文，市場に残る注文を指値注文と呼ぶ．市場に残した指値注文がキャンセル期間 c だけ経過しても取引が成立せず残っていた場合，注文板から取り除く．

3.3　エージェントの注文プロセス

エージェントは以下の手順に従い，注文価格，買いと売りの判断を行う．エージェント j が時刻 t のとき

に予想する価格の変化率（予想リターン）$r^t_{e_j}$ は式（1）で求める．

$$r^t_{e_j} = \frac{1}{w^t_{1,j} + w^t_{2,j} + u^t_j}(w^t_{1,j} \log \frac{P_f}{P^{t-1}} + w^t_{2,j} r^t_{h_j} + u^t_j \epsilon^t_j) \quad (1)$$

ここで，$w^t_{1,j}$ は時刻 t，エージェント j の i 項目の重みであり，シミュレーション開始時に，それぞれ 0 から $w_{i,max}$ までの一様乱数で決める．この重みは，後で述べる学習プロセスにより変化する．u^t_j は時刻 t，エージェント j の 3 項目の重みであり，シミュレーション開始時に，それぞれ 0 から u_{max} までの一様乱数で決め，学習プロセスによって変化せず，その後も一定である．P_f は時間で変化しない一定のファンダメンタル価格，P^t は時刻 t での市場価格，ϵ^t_j は時刻 t, エージェント j の乱数項で，平均 0, 標準偏差 σ_e の正規分布乱数である．$r^t_{h_j}$ は時刻 t に，エージェント j が計測した過去リターンであり，$r^t_{h_j} = \log(P^{t-1}/P^{t-\tau_j})$ である．ここで τ_j は，シミュレーション開始時に 1 から τ_{max} までの一様乱数でエージェントごとに決める．

式 (1) の第 1 項目はファンダメンタル価値を参照し投資判断を行うファンダメンタル投資家の成分であり，ファンダメンタル価格と直前期の市場価格を比較し，市場価格が安ければプラス，高ければマイナスの予想リターンを表す．第 2 項目は過去の価格推移を参照し投資判断を行うテクニカル投資家の成分であり，過去のリターンがプラスならプラス，マイナスならマイナスの予想リターンを表す．第 3 項目はノイズの成分を表す．

予想リターン $r^t_{e_j}$ より予想価格 $P^t_{e_j}$ は式（2）で求める．

$$P^t_{e_j} = P^{t-1} \exp(r^t_{e_j}) \quad (2)$$

注文価格 $P^t_{o_j}$ は平均 $P^t_{e_j}$，標準偏差 P_σ の正規分布乱数で決める．ここで，P_σ は式（3）で求める．

$$P_\sigma = P^t_{e_j} \times \ Est \quad (3)$$

Est（$0 < Est \leq 1$）を便宜上,「ばらつき係数」と呼ぶ．買いと売りの判断は予想価格 $P^t_{e_j}$ と注文価格 $P^t_{o_j}$ の大小関係で決まる．

$$\begin{array}{l} P^t_{o_j} > P^t_{e_j} \text{なら 1 単位の買い} \\ P^t_{o_j} < P^t_{e_j} \text{なら 1 単位の売り} \end{array} \quad (4)$$

3.4 学習プロセス

状況に応じて戦略を切り替えるという学習プロセスを Yagi ら [4] のモデルを参考にモデル化した．学習はエージェントごとに注文の直前におこなわれ，ファンダメンタル投資家のみの場合の予想リターンを $r^t_{e_{1,j}} = \log(P_f/P^{t-1})$，テクニカル投資家のみの予想リターンを $r^t_{e_{2,j}} = r^t_{h_j}$ とする．これら $r^t_{e_{i,j}}$ を学習期間のリターン $r^t_l = \log(P^t/P^{t-t_l})$ と比較し，式（5）のように $w^t_{i,j}$ を書き換える．

$$\begin{array}{l} \text{同符号なら, } w^t_{i,j} \leftarrow w^t_{i,j} + k_l r^t_l q^t_j (w_{i,max} - w^t_{i,j}) \\ \text{異符号なら, } w^t_{i,j} \leftarrow w^t_{i,j} - k_l r^t_l q^t_j w^t_{i,j} \end{array} \quad (5)$$

ここで，k_l は定数，q^t_j は時刻 t，エージェント j に与えられる 0 から 1 までの一様乱数である．式（5）では，価格変化の方向の予測が現実と一致した戦略の重みを引き上げ，外れている戦略の重みを引き下げるようにしている．また式（5）の学習プロセスの他に，確率 m で $w^t_{i,j}$ を 0 から $w_{i,max}$ までの一様乱数にて再設定を行う．

4 シミュレーション結果

4.1 実験概要

3 章でモデル化した人工市場を用いて実験を行う．各パラメータ値を変更して流動性の評価軸（Volume, Tightness，Resiliency，Depth）の変動を検証する．

実験で用いるモデルの基準パラメータ値を表 1 に示す．このうち，変更するパラメータとその値を表 2 に示す．パラメータは 1 つ 1 つ変化させ，変更しないパラメータについては基準パラメータ値に固定している．各種パラメータ値でそれぞれ 5 試行行い，後述する Volume, tightness，Resiliency，depth を算出し，以後それらの平均値を求める．またシミュレーションは時刻 $t = t_{end} = 1,000,000$ までおこなった．

表 1: 基準となるパラメータ

パラメータ	値
n	1,000
$w_{1,max}$	1
$w_{2,max}$	10
u_{max}	1
τ_{max}	10,000
σ_e	0.06
Est	0.003
c	20,000
ΔP	0.1
P_f	10,000
m	0.01

表 2: 変更するパラメータと設定

パラメータ	値				
ΔP	0.0001	0.001	0.01	0.1	1.0
	3.0	5.0	10		
Est	0.003	0.005	0.01	0.02	0.03
$w_{1,max}$	1.0	3.0	5.0	8.0	
$w_{2,max}$	3.0	5.0	8.0	10	
σ_e	0.02	0.04	0.06	0.08	1.0
c	5,000	10,000	15,000	20,000	30,000

4.2 モデルの妥当性

実験に入る前に本人工市場モデルの妥当性を検証した．シミュレーションモデルは実証研究で得られている統計的性質（stylized fact）が満たされているかで判断される．今回は人工市場に用いられる代表的な stylized fact であるファット・テールとボラティリティ・クラスタリングを判断基準とした．

ファット・テールは，市場価格の騰落率の分布が正規分布ではなく，裾の厚い分布を取ることを指し，尖度が正のとき，ファット・テールが成立している．ボラティリティ・クラスタリングは，市場価格の騰落率の 2 乗の自己相関がラグがある場合でも正の相関を示すことを指す．表 1 の基準パラメータ値での尖度と騰落率の 2 乗の自己相関の統計値を表 3 に示す．この表からわかるように，尖度が正を示しているため，ファット・テールを満たしている．また騰落率の 2 乗の自己相関は，ラグがある場合でも正の相関を保っており，ボラティリティ・クラスタリングを満たしていることがわかる．以上より，本人工市場は妥当性があることが示された．

表 3: Stylized Fact

尖度		16.4796
価格騰落率の 2 乗の 自己相関	lag1	0.024897
	lag2	0.025015
	lag3	0.025879
	lag4	0.026114
	lag5	0.026321

4.3 本人工市場モデルの流動性指標

本節では，Volume，Tightness，Resiliency，Depth の 4 つの評価軸に対する流動性指標を示す．

4.3.1 Volume

Volume には出来高を用いる．実験開始から実験終了までの出来高を計測する．また後述の値幅・出来高比率の計算に 1 日の出来高の値を使用するため，1 日（20,000）ごとの出来高も計測する．概ね 20,000 期で実際の市場での 1 営業日の売買成立数に達するため，20000 期を 1 日とした．

4.3.2 Tightness

Tightness にはビッド・アスク・スプレッドを用いる．実験開始から実験終了までの間，1 期ごとのベスト・ビッド，ベスト・アスクを取得し，ビッド・アスク・スプレッドを求め，期間内のビッド・アスク・スプレッドの平均値を計測する．

4.3.3 Resiliency

実証研究においては，基準となる「元の価格」を定めることができないため，Resiliency を正確に計測することができない．そこで，実証研究と同様に便宜上 1 日の市場価格の高低差を出来高で除した「値幅・出来高比率」を用いている．

値幅・出来高比率を採用した理由は，次のとおりである．既述のように，取引が成立することで板に残っている注文が少なくなっても，市場に復元力があれば速やかに板は回復し，売買に伴う価格変化は小さくなる．そのため，1 つの取引の平均的な価格変化である値幅・出来高比率が小さければ市場の流動性は高いといえるためである．値幅・出来高比率は式（6）で求める．

$$\text{値幅・出来高比率} = \frac{\text{1 日の市場価格の最大値と最小値の差}}{\text{1 日の出来高}} \tag{6}$$

4.3.4 Depth

2.2 節で説明したように最良気配値の上下に最小ティックサイズ刻みで 5，8，10 ティックだけ離れた値が Depth 情報として公開されている．しかし，本研究では最小ティックサイズを変更して行う実験があることやティックサイズ刻みで Depth を取り出そうした場合，Depth の値が極端に小さく見えてしまう．そこで Depth は実験開始から実験終了までの間，1 期ごとの最良気配値から 50 離れた値までの注文枚数を求め，期間内の注文枚数の平均値を用いることとした．

4.4 実験結果

4.4.1 ティックサイズ ΔP 変更実験

ティックサイズ ΔP を，0.0001，0.001，0.01，0.1，1.0，3.0，5.0，10.0 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 4 に示す．ティックサイズが大きくなると，Volume，Tightness，Resiliency の値は大きくなり，Depth の値は小さくなる．

以下にこのようになった理由を述べる．

まず，Volume（出来高）が大きくなった理由は次のとおりである．ティックサイズが大きくなると，注文価格の刻みが荒くなり同じ価格の注文が複数残りやすくなる．最良気配値の注文も多くなるため，注文板から注文がなくなりにくく， 取引機会は減らず出来高は高くなるからである．

次に，Tightness（ビッド・アスク・スプレッド）が大きくなった理由は，ビッド・アスク・スプレッドの最小値は，0 の場合を除くとティックサイズの値と等しいため，ティックサイズが大きくなるとビッド・アスク・スプレッドも大きくなるからである．

さらに，Resiliency（値幅･出来高比率）が大きくなった理由は次のとおりである．値幅・出来高比率の計算式（式 (6) 参照）の分子は，1 日の市場価格の最大値と最小値の差であり，この値の変化の最小値はティックサイズのため，増加傾向になる．分母の出来高も増加傾向にあるが，ティックサイズの増加率に比べるとはるかに小さいため，ティックサイズが大きくなると値幅・出来高比率も大きくなる．

最後に，Depth が小さくなった理由だが，本研究の人工市場では注文数を常に 1 つとしているため，取引が成立し続ければ，注文板に残った指値注文は減り続ける．そのため，ティックサイズが大きくなるほど Depth は小さくなる．

表 4: ΔP 変更実験

ΔP	Volume	Tightness	Resiliency	Depth
0.0001	260,409	16.328	1.0169	2,185.13
0.001	263,467	16.841	1.0384	2,095.88
0.01	263,500	16.879	1.0360	2,094.33
0.1	265,140	17.122	1.0463	2,052.43
1.0	267,352	17.809	1.0596	2,037.84
3.0	272,635	19.409	1.0820	1,931.12
5.0	280,804	21.383	1.1370	1,906.47
10.0	289,561	24.849	1.1794	1,978.40

4.4.2 ノイズ成分の予想リターン計算 ϵ_j^t に用いる定数 σ_ϵ 変更実験

ノイズ成分の予想リターン計算 ϵ_j^t に用いる定数 σ_ϵ を 0.02，0.04，0.06，0.08，0.10 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 5 に示す．σ_ϵ が大きくなる（ノイズ成分の影響が強くなる）と，Volume，Tightness，Resiliency の値は大きくなり，Depth の値は小さくなる．

以下にこのようになった理由を述べる．

まず，Volume が大きくなった理由である．σ_ϵ の値が大きいほど，エージェントの注文価格予想が荒くなり，取引が成立する可能性が高くなるからである．

次に，Tightness が大きくなった理由であるが，エージェントの注文価格予想が荒くなり，取引が成立する可能性が高くなると，必然的にビッド・アスク・スプレッドは広がるからである．

さらに，Resiliency が大きくなった理由も同様で，エージェントの注文価格予想が荒くなることが原因で，値幅・出来高比率の分子が大きくなる．分母となる出来高も増加傾向にあるが，分子の増加割合の方が大きいと考えられる．

最後に Depth についてだが，Volume が大きくなる，すなわち，取引成立回数が多いということは，注文板上の指値注文は少なくなることを意味しているため，σ_ϵ の値が大きいほど Depth は小さくなる．

表 5: σ_ϵ 変更実験

σ_ϵ	Volume	Tightness	Resiliency	Depth
0.02	149,973	6.066	0.5940	5,474.57
0.04	225,272	11.825	0.8117	3,285.43
0.06	265,140	17.122	1.0463	2,052.43
0.08	291,861	22.997	1.3320	1,277.79
0.1	311,118	29.591	1.6646	811.41

4.4.3 ファンダメンタル成分の重みの最大値 $w_{1,max}$ 変更実験

ファンダメンタル成分の重みの最大値 $w_{1,max}$ を 1.0，3.0，5.0，8.0 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 6 に示す．本実験では 4.4.2 節の実験とは逆の結果が得られた．すなわち，$w_{1,max}$ が大きくなる（ファンダメンタル投資家が目立つようになる）と，Volume，Tightness，Resiliency の値は小さくなり，Depth の値は大きくなる．

このような傾向となった理由は，ファンダメンタル成分の重みを大きくさせるということは，テクニカル

成分やノイズ成分の重みを相対的に小さくしているからだと考えられる．その結果，ノイズ成分の影響を変化させた 4.4.2 節の実験と同様の相関が得られた．

表 6: $w_{1,max}$ 変更実験

$w_{1,max}$	Volume	Tightness	Resiliency	Depth
1.0	265,140	17.122	1.0463	2,052.43
3.0	244,241	14.906	0.9057	2,691.23
5.0	228,920	13.398	0.8222	3,158.24
8.0	208,670	11.589	0.7408	3,764.25

4.4.4 テクニカル成分の重みの最大値 $w_{2,max}$ 変更実験

テクニカル成分の重みの最大値 $w_{2,max}$ を 3.0，5.0，8.0，10.0 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 7 に示す．本実験では 4.4.2 節の実験とは反対の結果，かつ，4.4.3 節の実験とは同じような結果が得られた．すなわち，$w_{2,max}$ が大きくなる（テクニカル投資家が目立つようになると）と，Volume，Tightness，Resiliency の値は小さくなり，Depth の値は大きくなる．

このような傾向となった理由は，テクニカル成分の重みを大きくさせるということは，ファンダメンタル成分やノイズ成分の重みを相対的に小さくしているからだと考えられる．その結果，ノイズ成分の影響を変化させた 4.4.2 節の実験，ファンダメンタル成分の重みを変化させた 4.4.3 節の実験と同様の相関が得られた．

表 7: $w_{2,max}$ 変更実験

$w_{2,max}$	Volume	Tightness	Resiliency	Depth
3.0	332,824	45.201	2.3153	421.72
5.0	310,981	30.698	1.6621	815.62
8.0	282,196	20.825	1.2141	1,540.53
10.0	265,140	17.122	1.0463	2,052.43

4.4.5 ばらつき係数 Est 変更実験

エージェントごとの注文価格のばらつきを決めるばらつき係数 Est を 0.003，0.005，0.01，0.02，0.03 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 8 に示す．ばらつき係数が大きくなる（予想が荒くなる）と，Volume，Depth は小さくなり，Tightness，Resiliency は大きくなる．

以下にこのようになった理由を述べる．

まず，Volume が小さくなった理由だが，ばらつき係数が大きくなると注文価格のばらつきも広がるため，市場価格に対して安値の買い注文や高値の売り注文（いわゆる，指値注文）が発注されやすくなり，出来高が減少するからである[1]．

次に，Tightness が大きくなった理由については，ばらつき係数が大きくなると，注文価格のばらつきも広がり，その結果，ベスト・ビッドとベスト・アスクの価格の幅も広がるからである．

さらに，Resiliency が大きくなった理由だが，値幅・出来高比率の分母は出来高であるが，ばらつき係数が大きくなるにしたがい，出来高が急速に小さくなっていることがわかる．その結果，値幅・出来高比率が大きくなったものと考えられる．

最後に，Depth が小さくなった理由だが，ばらつき係数が大きくなると，注文価格のばらつきも広がるので，最良気配値周辺の指値注文がまばらになったものと考えられる．

表 8: Est 変更実験

Est	Volume	Tightness	Resiliency	Depth
0.003	265,140	17.122	1.0463	2,052.43
0.005	213,556	17.498	1.2660	2,107.08
0.01	143,536	18.934	1.9467	1,560.71
0.02	87,270	20.681	3.4044	924.41
0.03	63,228	21.826	4.9170	641.00

4.4.6 キャンセル期間 c 変更実験

キャンセル期間 c を 5,000，10,000，15,000，20,000，30,000 と変化させたときの Volume，Tightness，Resiliency，Depth の平均を表 9 に示す．

キャンセル期間 c が大きくなっても，Volume，Tightness，Resiliency には大きな変化は見られなかった．このような結果になった理由は，他の実験で変更したパラメータは，すべて注文価格に関連しているのに対し，キャンセル期間 c は価格に直接影響するパラメータではないため，キャンセル期間 c を変化させても注文価格に確かな影響が与えられなかったからだと思われる．

一方で，キャンセル期間が大きくなると Depth の値は大きくなる．キャンセル期間が大きくなると，注文

[1] ノイズ成分の予想リターン計算に用いる定数 σ_ϵ が大きくなる場合とよく似たメカニズムであるが，σ_ϵ が大きくなるときは，予想価格帯が広がるだけで注文価格のばらつきは一定なので，予想価格帯に注文が集中するため取引が多くなる（Volume が大きくなる）．しかし，ばらつき係数が大きくなるときは，予想価格は変わらず注文価格のばらつきが大きくなるため，本文中に記載したような取引が成立しないような注文が増えてしまう（Volume が小さくなる）．

板上に注文が残っている期間が長くなるため，Depth は大きくなっただけであり，流動性が変化したわけではない．そのため，キャンセル期間 c は流動性に影響を与えるパラメータではない．

表 9: c 変更実験

c	Volume	Tightness	Resiliency	Depth
5,000	261,767	16.858	0.9912	575.17
10,000	265,364	17.412	1.1365	1,036.47
15,000	263,216	16.920	1.0636	1,582.31
20,000	265,140	17.122	1.0463	2,052.43
30,000	266,204	17.297	1.0278	3,046.93

4.5 流動性指標の関係性

表 10 は各パラメータ値を変化させたときの流動性指標変化の結果の一覧である．各パラメータ値が大きくなった時のそれぞれの指標の変化を上段（増加：↑，減少：↓，変化なし：—），流動性の上下を下段（上昇：○，下降：×，変化なし：—）に記している．

表 10: 各パラメータ値を増加させたときの流動性の変化（上段：指標の増減，下段：流動性の上下）

パラメータ	Volume	Tightness	Resiliency	Depth
ΔP	↑	↑	↑	↓
	○	×	×	×
σ_ϵ	↑	↑	↑	↓
	○	×	×	×
$w_{1,max}$	↓	↓	↓	↑
	×	○	○	○
$w_{2,max}$	↓	↓	↓	↑
	×	○	○	○
Est	↓	↑	↑	↓
	×	×	×	×
c	—	—	—	↑
	—	—	—	○

表 10 からティックサイズ ΔP，ノイズ成分の予想リターン計算 ϵ_j^t に用いる定数 σ_ϵ，ファンダメンタル成分の重みの最大値 $w_{1,max}$，テクニカル成分の重みの最大値 $w_{2,max}$ を変化させたとき，Volume に対して，Tightness と Resiliency は正の相関をもち，Depth は負の相関をもつことが分かった．これは Volume を指標とするか，その他 3 つを指標とするかで流動性の方向性が正反対になることを示唆している．

この原因を注文の種類に着目して改めて考察する．まず，パラメータの値を変化させた際，成行注文のばらつきが大きくなるとする．すると，成行注文は注文板上の待機注文との取引がたくさん成立することになるため，Volume は上昇する．一方で，待機注文は減少するためその他の 3 指標は下降する．次に，成行注文のばらつきが小さくなるとする．すると，成行注文と注文板上の待機注文との取引が成立することに少なくなるため，Volume は下降する．しかし，成行注文は取引が成立せずに待機注文となるため，その他の 3 指標は上昇する．

以上より，原理的には Volume と Depth がともに上昇（もしくは下降）することは発生しにくいと思われる．しかし現実には，Volume が増えることにより Depth も上昇することが実証分析 [7][8] で分かっている．逆に言えば，その理由が本研究でモデル化していないメカニズムにある可能性が示唆された．本研究では Volume が高いことを理由に指値注文を増やすという行動をモデル化していないため，これが Volume と Depth が正の相関をもつメカニズムである可能性がある．これは今後の課題である．

ばらつき係数 Est を変化させたときは，Volume に対して，Tightness と Resiliency は負の相関，Depth は正の相関をもつことが分かった．このときは 4 指標全てにおいて流動性の方向性が一致していることが分かる．

ばらつき係数 Est の値を変更すると，取引が成立せずに注文板上に残った指値注文（待機注文）もばらついた状態で残ることになるので，他のパラメータ値を変化させたときとは異なる動きになると思われる．

最後にキャンセル期間 c を変化させたときは，Depth 以外の指標に明確な変化は得られなかった．このことから，キャンセル期間のような直接注文価格に影響を与えない要因は流動性にも明確な影響が出ないと思われる．

5 まとめと今後の課題

流動性の代表的な 4 つの評価軸（Volume，Tightness，Resiliency，Depth）を計測するための代表的な指標に注目して，それら指標の関係をティックサイズなどの人工市場内のパラメータを変化させることで調査をした．その結果，ティックサイズの大きさ，ノイズ成分の予想リターン計算に用いる定数，ファンダメンタル成分の重みの最大値，テクニカル成分の重みの最大値を変更した際，4 つの流動性指標の間に以下に示す傾向があることが確認された．Volume が増加すると，Tightness と Resiliency，Depth は悪化（流動性が減少）ことが分かった．一方で，ばらつき係数 Est では Volume が増加すると，Tightness と Resiliency，Depth は向上（流

動性が増加）することが分かった．しかし，キャンセル期間 c を変化させた場合には流動性指標の間に明確な傾向は見られなかった．以上より，キャンセル期間のような直接注文価格に影響を与えない要因は流動性にも明確な影響が出ないと思われる．

今後の課題は次の通りである．Volume と Depth がともに上昇することは原理的に発生しにくいと思われる．しかし現実には，Volume が増えることにより Depth も上昇することが分かっている．その理由が本研究でモデル化していないメカニズムにある可能性が示唆された．本研究では Volume が高いことを理由に指値注文を増やすという行動をモデル化していないため，これが Volume と Depth が正の相関をもつメカニズムである可能性がある．このメカニズムについて検証する必要がある．

留意事項

本論文はスパークス・アセット・マネジメント株式会社の公式見解を表すものではありません．すべては個人的見解です．

謝辞

本研究は SPS 科研費 15K01211 の助成を受けたものです．この場を借りてお礼申し上げます．

参考文献

[1] 黒崎 哲夫，熊野 雄介，岡部 恒多，長野 哲平：国債市場の流動性：取引データによる検証，日本銀行ワーキングペーパー，No.15-J-2，日本銀行，2015.

[2] 辰巳 憲一：市場の流動性と HFT ～約定時間を一指標として提案する～，学習院大学経済論集，第 53 巻，第 1 号，2016.

[3] Yamamoto, R. and Hirata, H.：Strategy switching in the Japanese stock market，Journal of Economic Dynamics and Control，Vol.37，No.10，pp.2010–2022，2013.

[4] Yagi, I.，Nozaki, A.，and Mizuta, T.：Investigation of the rule for investment diversification at the time of a market crash using an artificial market simulation，Evolutionary and Institutional Economics Review，Vol.14，No.2，pp451–465，2017.

[5] 土川 顕，西崎 健司，八木 智之：国債市場の流動性に関連する諸指標，日銀レビュー，2013-J-6，日本銀行金融市場局，2013.

[6] 海野利勝：JREIT の流動性リスクに関する研究，日本不動産金融工学学会，2009.

[7] 村永 淳：本邦株式市場の流動性に関する動学的考察 -東京証券取引所のティック・データ分析，日本銀行金融研究所ディスカッション・ペーパー，No.2000-J-18，日本銀行金融研究所，2000.

[8] Chung, K.H.：Liquidity and quote clustering in a market with multiple tick sizes，Journal of Financial Research，Vol.XXVIII，No.2，Summer，pp.177–195，2005.

[9] 辰巳 憲一：市場の厚みの分析 ～Depth の研究展望と HFT 解明に向けての考察～，学習院大学経済論集，第 52 巻，第 2 号，2015.

[10] Mizuta, T.，Noritake, Y.，Hayakawa, S.，and Izumi, K.：Impacts of Speedup of Market System on Price Formations using Artificial Market Simulations，JPX Working Paper，Vol.9，Japan Exchange Group，2015.

テキストマイニングによる
有価証券報告書からの因果関係文の抽出

Extraction of Causal Knowledge from Annual Securities Report by Text Mining

佐藤史仁 [1]　佐久間洋明 [1]　小寺俊哉 [1]　田中良典 [1]　坂地泰紀 [2]　和泉潔 [2]

Fumihito Sato[1], Hiroaki Sakuma[1], Shunya Kodera[1], Yoshinori Tanaka[1], Hiroki Sakaji[2], and Kiyoshi Izumi[2]

[1] 日興リサーチセンター株式会社　投資工学研究所

[1]Institute of Investment Technology, Nikko Research Center, Inc.

[2] 東京大学大学院工学系研究科

[2]Graduate School of Engineering, The University of Tokyo

Abstract: 有価証券報告書には，業績の他，リスク対策や企業の施策等，決算短信にはない情報の記載もある．また，先行研究では，多くの情報から重要な文を効率よく抽出する方法として，因果関係文を重要文とした手法が提案されている．しかし，抽出対象を決算短信等とした報告はあるが有価証券報告書とした報告はない．そこで本稿では，この手法を応用し，有価証券報告書専用の因果関係文を抽出する判別モデルを提案した．そして，判別モデルの評価等を行い，高い性能であることなどを示した．この判別モデルにより有価証券報告書独自の投資判断に有益な情報の効率的な抽出が期待できる．

1. はじめに

投資判断に利用できる情報として，財務データやマーケットデータなどの数値データの他，ニュースや新聞記事，決算短信，有価証券報告書などに含まれるテキストデータがある．これらテキストデータの特徴は，過去又は将来の業績に対する理由及び根拠や経営戦略に関する情報，進行中の施策，新商品発表情報，抱えているリスク，企業の不祥事に関する情報など，投資判断において数値データにはない重要な情報を含んでいる点が挙げられる．しかしながら，構造化することが難しく，そのほとんどは投資家が直接読まなければ投資判断に利用できない場合が多かった．近年，この問題に対し，テキストマイニングなどの人工知能分野の技術を，金融市場における分析に導入して解決を試みる研究が盛んに行われている．例えば，ある企業に関連するニュースがその株価にとってポジティブに働くかネガティブに働くか（極性）でニュースを定量化し，株式リターンとの関係を分析した研究[1][5]が挙げられる．これらの研究は，投資家が直接そのニュースを読まなくとも，ニュースに極性を付与し定量化することで，投資戦略に活用できることを示唆している．

テキストデータの定量化以外にも，テキストデータから投資判断等に関する重要文を抽出する手法の研究が行われている．例えば，坂地ら[8][9]は，文の表現と機械学習によって，過去の業績や製品の売れ行きなどに対する因果関係文を経済新聞の記事や決算短信から抽出する手法を提案している．因果関係文とは，出来事（結果）とその理由（原因）の組から構成される文と定義される．例えば，原因「猛暑」による結果「冷房需要の盛り上がり」等の因果関係を提示することで，「猛暑」の際には「冷房需要」が高まる可能性があるという情報を得られるとした．因果関係文の抽出以外にも，決算短信から業績の要因を含む文を抽出した研究[7]や，業績の予測を示す文を抽出した研究[2]がある．さらに，抽出された重要文に極性を付与する研究[6]も行われている．このように，テキストデータからの重要文の抽出手法は，投資家が投資判断に有益な情報を効率良く把握することを可能にし，また，ある特定のテーマを持った重要文の定量化データを既存の分析や投資戦略等へ導入することで，新しい投資戦略や手法の開発に役立つだろう．しかしながら，多様なテキストデータからある特定のテーマを持った重要文についての抽出手法やその定量化が多岐にわたり研究されているが，現在のところ，それらの統一的な抽出手法や数量化モデルは存在しない．言い換えれば，抽出対象

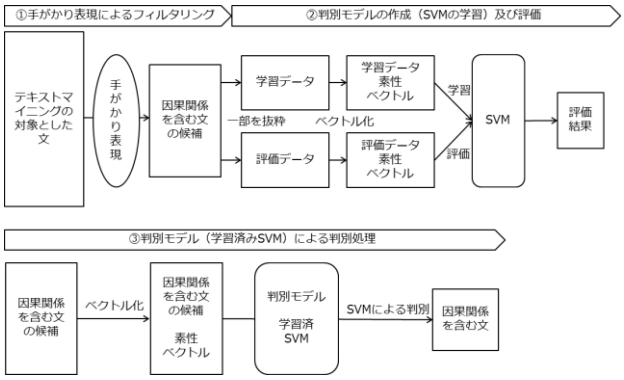
となるテキストデータや重要文によって，抽出手法や定量化手法に工夫が必要であると言える．

有価証券報告書は，上場会社が証券取引所から求められている適時開示資料である決算短信に対し，金融商品取引法により提出が定められている開示資料である．有価証券報告書は，決算短信と比較して速報性はないものの[1]，「業績等の概要」等の業績に関する情報だけでなく，「対処すべき課題」や「事業等のリスク」など投資判断に有益と考えられる情報をより多く含む．また，ニュースや新聞記事は限られた企業に関する情報が多いが，有価証券報告書は全ての上場企業から公表されているという利点もある[2]．投資判断に有益なテキスト情報としては，まずは業績に関する情報が考えられる．どんな事象が原因で，そのような業績結果となったのかを知るには，業績に関する項目の因果関係文を抽出すれば良い．また，何をリスク要因や経営課題として捉えて，それに対しどんな対策を講じているのかを知るには，リスクやその企業がもつ課題に関する項目の因果関係文を抽出することで把握できると考えられる．つまり，有価証券報告書の各項目から因果関係文を抽出することは，単純に業績に関する原因と結果だけでなく，その企業のリスク対策や目指している方向性などを把握するための有力情報を取得することになると言えるだろう．にもかかわらず，因果関係文の抽出に関する研究の中で，新聞記事や決算短信を対象とした研究はあるものの，有価証券報告書を対象にした報告はない．そこで，本稿では，有価証券報告書から「業績等の概要」,「対処すべき課題」,「事業等のリスク」を対象に因果関係文を抽出する判別モデルを提案する．具体的には，坂地ら[8]の手法を応用し，有価証券報告書専用の因果関係文を抽出する機械学習を用いた判別モデルを作成した．データは TOPIX1000 を構成する企業の 2008 年から 2016 年までの有価証券報告書（本決算）に含まれるテキストデータを用いた．

2. 因果関係文の判別モデル

本稿では，新聞記事から因果関係を含む文を抽出する研究[8]で用いられた手法を応用した．この手法は，より広範の因果関係文を対象とできることや高い抽出性能となったことが報告されている．本稿で行った因果関係を含む文の抽出の実際の処理として，まず，テキストマイニングの対象とした文から手がかり表現で因果関係を含む文の候補を抽出した．次に，この因果関係を含む文の候補の一部から学習データ及び評価データを作成した．そして，素性でベクトル化した学習データでサポートベクターマシン（以下，SVM）[3]を学習させ判別モデルを作成した．判別モデルの評価は，素性でベクトル化した評価データで行った．最後に，因果関係を含む文の候補を素性でベクトル化し，判別モデルで因果関係を含む文を抽出した．なお，SVM のカーネルは線形を用いた．手がかり表現と素性及び学習データと評価データについては後述する．図表 1 は因果関係文の抽出処理の概要を示す．また，抽出対象となる因果関係を含む文と抽出対象外である因果関係を含まない文の具体例を図表 2 に示す．ただし，太字は手がかり表現を示している．

図表 1　因果関係文の抽出処理の概要



2.1 手がかり表現によるフィルタリング

手がかり表現とは因果関係文を判定する上で重要な手がかりとなる表現を示す．例えば，「猛暑日が連続したため､飲料水の売上が伸びた｡」という文の「ため､」が手がかり表現となる．本稿では，決算短信と有価証券報告書は記載される文が類似していることから，決算短信の手がかり表現[9]を参考に 37 個の手がかり表現を選定した．これらの選定した手がかり表現を含む文が，因果関係を含む文の候補となる．ただし，2 文にまたがる因果関係や，手がかり表現が含まれていない文は対象外とした．選定した手がかり表現の例を図表 3 に示す．

[1]決算短信は，決算期末後 45 日以内の開示が適当とされ，30 日以内の開示がより望ましいとされている．一方，有価証券報告書は，やむを得ない場合を除いて事業年度末から 3 ヵ月以内の公表が求められている．

[2]有価証券報告書は，金融商品取引法第二十四条により提出が定められている．企業の HP の他，金融庁の EDINET などから入手可能．http://disclosure.edinet-fsa.go.jp/

[3]本稿では，python の機械学習のオープンソースライブラリである scikit-learn（http://scikit-learn.org/stable/index.html）を用いた．

図表 2　因果関係を含む文と含まない文の例

因果関係を含む文	シューズ部門では、ランニングブームの継続と、フィッティングの取組みを強化したこと**により**、ランニングシューズの販売が堅調に推移いたしました。
	また、新興国を中心とする旺盛な需要や新しいエネルギー資源の開発など**を背景に、**当社グループの事業環境は好転しております。
因果関係を含まない文	また、通商、独占禁止、特許、消費者、租税、為替管制、環境・リサイクル関連の法規制**を受けております。**
	平成 17 年 7 月 1 日**から、**製造たばこの販売に際しては、これらの規定に従っております。

図表 3　手がかり表現の例

から、	を背景に、	を受けております。
を反映し、	を反映して	に支えられて
によって	により	ためであります。
に伴う	に伴い、	を受け、

2.2 サポートベクターマシンによる因果関係文の抽出

（1）学習データ及び評価データの作成方法

SVM の学習データ及び評価データは因果関係を含む文の候補から下記の手順 1-1～手順 1-4 の手順で抽出した後，因果関係を含む場合に正例，含まない場合に負例とするラベル付与を人手で行った．文の抽出においては，精度の高い判別モデルを作成するため，学習データについてより広範の表現を抽出できるような工夫を行った．具体的には，有価証券報告書に記載される文が時期や業種，項目によって特徴が異なることが考えられるため，これらが均一に抽出されるようにした．

手順1-1： 分析対象の有価証券報告書を，有価証券報告書の発表日ベースで年ごとに振り分ける．

手順1-2： 手順 1-1 の各年の有価証券報告書から業種ごとにランダムに 3 社を抽出．計 459（3 社×17 業種×9 年）の有価証券報告書が抽出される．業種は東京証券取引所が定めた東証 17 業種分類を利用する．

手順1-3： 手順 1-2 で抽出した 459 の有価証券報告書から手がかり表現によるフィルタリングで因果関係を含む文の候補を抽出する．

手順1-4： 手順 1-3 で抽出した各有価証券報告書の文から，「業績等の概要」，「対処すべき課題」，「事業等のリスク」の各項目からランダムに 1 文ずつ抽出．合計 1377（459×3）文が抽出される．なお，1 文も抽出されなかった有価証券報告書があった場合は，既に抽出済みの有価証券報告書以外で，同年，同業種内でランダムに取得された有価証券報告書を使用して同じ項目の 1 文をランダムに抽出する．

学習データ及び評価データをそれぞれ 1377 文抽出した後，人手で正例と負例のラベルを付与した．正例と負例の誤判定の発生を抑えるために，少なくとも 3 人の判定が一致するような判定手順で作業を行った．判定員は金融業務に従事する実務者が担当した. 正例と負例のラベル付与の手順を手順 2-1～手順 2-4 に示した．学習データ及び評価データとも共通の手順となる．得られたラベル付きデータの内訳は，学習データが正例 782 文，負例 595 文，評価データが正例 733 文，負例 644 文となった．

手順2-1： 抽出した学習データ（または評価データ）1377 文を 3 グループ（各 459 文）に分割する．

手順2-2： 判定員計 9 人を各グループに 3 人ずつ割り当てる．

手順2-3： グループ内のそれぞれの文に対して，判定員 3 人が同じ判定だった場合はその判定を採用し正例または負例のラベルを付与する．

手順2-4： 手順 2-3 でラベルが付与されなかった文に対しては，別の判定員 2 人が判定を行う．5 つの判定のうち，同じ判定が 3 つ以上となった判定を採用しラベルを付与する．

（2）素性の作成

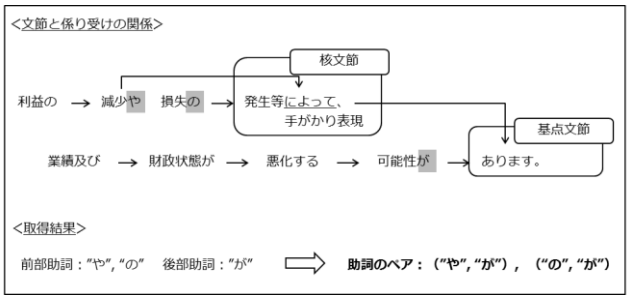
判別モデルとして SVM を用いる際，学習，評価及び学習後の判別処理において，文の特徴を表す素性が必要となる．本稿では坂地ら[8]を参考に 4 つの素性を利用した．各素性の概要を図表 4 に示す．ただし，図表 5＜文節と係り受けの関係＞は係り受け解析を行った文節と係り受けの関係を表しており，一塊の文字列が文節を，矢印が係り先を示している．因果関係を含む文の特徴を適切に捉えるために，文

図表 4　素性の概要

素性の名前	概要
助詞のペア	核文節に係る文節に含まれる助詞を前部助詞，基点文節に係る文節の助詞を後部助詞とし，前部助詞と後部助詞を合わせた全ての助詞のペア（重複を除く）． ただし，前部助詞が取得できない場合は，核文節より前の最も近い文節の助詞を前部助詞とする．存在しない場合は欠損値．また，後部助詞が取得できない場合は，核文節より後で基点文節に最も近い助詞を後部助詞とする．存在しない場合は欠損値． ※核文節は手がかり表現を含む文節．基点文節は核文節の係り先の文節． ※助詞のペアの取得イメージは図表 5 を参照．
文に含まれる手がかり表現	図表 3 を参照．
形態素ユニグラム	因果関係を含む文の候補を形態素解析器で分解した形態素のうち，頻度が 2 以上のものを抽出し重複を除いたもの．
形態素バイグラム	因果関係を含む文の候補を形態素解析器で分解し，隣り合った全ての形態素ペア（重複を除く）．

に含まれる手がかり表現と構文的な素性である助詞のペアが素性として含まれている．SVM に入力される最終的なデータは，取得した全ての素性を並べて，文に含まれている素性を 1，含まれていない素性を 0 としたベクトルとなる．素性の作成においては，形態素解析では形態素解析器 MeCab[4] [4]を，構文解析では係り受け解析器 CaboCha[5] [3]を用いた．

図表 5　助詞のペアの取得イメージ



3. 判別モデルの評価結果

SVM を学習させ作成した判別モデルの評価結果を図表 6 に示す．ただし，平均／合計の欄は，精度，再現率，F 値について，因果関係を含む文と含まない文の数で加重平均した値を，データ数については合計を示している．

因果関係を含む文の場合も，含まない文の場合も，精度，再現率，F 値 とも 0.8 を超える結果となった．また，図表 7 に抽出された因果関係を含む文の例を示す．

[4] http://taku910.github.io/mecab/
[5] https://taku910.github.io/cabocha/

図表 6　判別モデルの評価結果

	精度	再現率	F 値	データ数
因果関係を含む文	0.85	0.89	0.87	733
因果関係を含まない文	0.87	0.82	0.84	644
平均／合計	0.86	0.86	0.86	1377

4. 考察

判別モデルの評価は，因果関係を含む文の場合も，含まない文の場合も，精度，再現率，F 値 とも 0.8 を超え，良好な判別結果となったが，この要因としていくつかの理由が考えられる．1 つは，上手くモデルの学習が機能するような学習データが準備できたことが考えられる．実際に，年，業種，項目ごとに万遍なく文を取得していることや，手がかり表現でフィルタリングしたことで正例候補が絞られ，結果的に正例と負例の割合がおおよそ 6 対 4 となり，偏りなく広範の表現の学習データが準備できた．その他，本稿で対象とした有価証券報告書の因果関係を含む文の構文や単語は似たものが多かったという可能性や，本稿の素性で上手く捉えられるようなシンプルな特徴を持つ文が多かった可能性も考えられる．ここで，シンプルな特徴を持つ文は，因果関係を含むか否かの人による判別が容易であると考え，SVM の性能評価の追加確認を行った．評価データとしては，手順 2-3 において，最初の判定員 3 人の判定が一致した文を判別の容易な文としたデータ，反対に，3 人の判定が一致しなかった文を判別の難しい文としたデータのそれぞれを用いた．その結果を

図表 7　判別モデルの因果関係文抽出結果例

項目名	因果関係文として抽出された文
業績等の概要	投資活動によるキャッシュ・フローは、船舶の取得による支出などにより、当連結会計年度は 1,455 億 40 百万円のマイナスとなりました。
事業等のリスク	一方、9 月以降には取引先からの返品が発生するため、第 4 四半期の収益が低下いたします。
対処すべき課題	ロール事業では、事務機の構造変化や高耐久化の加速による補修品市場の縮小といった環境変化により販売の低迷が想定されます。

図表 8　判別モデルの追加の評価結果

判別難易	因果関係有無	精度	再現率	F 値	データ数
易	有	0.93	0.93	0.93	590
	無	0.91	0.91	0.91	454
難	有	0.58	0.74	0.65	143
	無	0.75	0.60	0.67	190

図表 8 に示す．判別の容易な文については，判別性能の評価が高く，一方で，判別の難しい文は判別性能の評価が低い結果となった．判別の難しい文は 5 人の判定員で判別されており，単純にルール化できるような因果関係を含むか否かの基準だけではなく，判定員それぞれの異なる経験と様々な観点から判別されている．他方，SVM では，限られた学習データと決められた素性で学習しているため，ルール化できるような基準は反映できても，個々人の経験や様々な観点までは反映しきれなかったのかもしれない．このようなニュアンスと呼べる要素を加味するには，まず，判別が難しい文の学習データを増やすことや，判別の難しい文に対する因果関係が含まれるか否かの判別基準となる特徴などを，素性として新たに加える等の工夫が必要になると考えられる．

5. まとめ

本稿では，有価証券報告書から因果関係を含む文を抽出する判別モデルを提案した．作成したモデルの評価では，因果関係を含む文の場合も，含まない文の場合も，精度，再現率，F 値 とも 0.8 を超え，良好な判別結果となった．この判別モデルにより有価証券報告書独自の投資判断に有益な情報の効率的な抽出が期待できるだろう．

また本稿における SVM による判別モデルについて，判別が難しい文の学習データを増やすことや，判別の難しい文の因果関係が含まれるか否かの判別基準となる特徴などを素性として新たに加える等の工夫をすることで，判別モデルのさらなる性能の向上が見込めるだろう．

参考文献

[1] 五島圭一，高橋大志，寺野隆雄: 「ニュースのテキスト情報から株価を予測する」，第 29 回人工知能学会全国大会 大会論文集, Vol.29, pp.1-3, (2015)

[2] 北森詩織，酒井浩之，坂地泰紀: 「決算短信 PDF からの業績予測文の抽出」，電子情報通信学会論文誌 (D), Vol.J100-D, No.2, pp150-161, (2017)

[3] 工藤拓，松本裕治: 「チャンキングの段階適用による日本語係り受け解析」，情報処理学会論文誌, Vol.43, No.6, pp1834-1842, (2002)

[4] 工藤拓，山本薫，松本裕治: 「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, pp89-96, (2004)

[5] 沖本竜義，平澤英司: 「ニュース指標による株式市場の予測可能性」， 証券アナリストジャーナル, Vol.52, No.4, pp.67-75, (2014)

[6] 酒井浩之，小林義和，坂地泰紀:「企業の決算短信ＰＤＦから抽出した業績要因への極性付与」，第 15 回金融情報学研究会, pp.7-12, (2015)

[7] 酒井浩之，西沢裕子，松並祥吾，坂地泰紀: 「企業の決算短信 PDF からの業績要因の抽出」，人工知能学会論文誌, Vol.30, No.1, pp.172-182, (2015)

[8] 坂地泰紀，増山繁: 「新聞記事からの因果関係を含む文の抽出手法」，電子情報通信学会論文誌 (D), Vol.J94-D, No.8, pp1496-1506, (2011)

[9] 坂地泰紀，酒井浩之，増山繁:「決算短信 PDF からの原因・結果表現の抽出」，電子情報通信学会論文誌 (D), Vol.J98-D, No.5, pp811-822, (2015)

経済テキストからの市況分析コメントの自動生成

Automatic generation of market analysis comments from financial articles

酒井 浩之 [1*]　坂地 泰紀 [2]　和泉 潔 [2]　松井 藤五郎 [3]　入江 圭太郎 [4†]
Hiroyuki Sakai[1]　Hiroki Sakaji[2]　Kiyoshi Izumi[2]　Tohgoroh Matsui[3]　Keitaro Irie[4]

[1] 成蹊大学 [1] Seikei University [2] 東京大学 [2] The University of Tokyo
[3] 中部大学 [3] Chubu University
[4] 三菱 UFJ 国際投信 [4] Mitsubishi UFJ Kokusai Asset Management

Abstract: 本研究では，経済新聞記事などの経済テキストから、日経平均株価などの市況について言及している文書のみを抽出し，それらの内容を自動的に要約することにより，ファンドの運用報告書における市況分析コメントを自動生成する手法の開発を行う．本手法では，まず経済新聞記事から深層学習により日経平均株価の市況について言及している記事を抽出する．次に抽出された記事の中から例えば「ギリシャへの金融支援協議が難航していることや、中国・上海株の値動きへの警戒感から、投資家のリスクオフの動きが強まった。」のような日経平均が大幅に変動した理由について言及している文を抽出する．そして，抽出された文を時系列順に並べることで市況分析コメントを自動生成する．

1　はじめに

近年，証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援をおこなう技術の必要性が高まっている．そのため，人工知能分野の手法や技術を，金融市場における様々な場面に応用することが期待されており，例えば，膨大な金融情報を分析して投資判断の支援を行う技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」や経済新聞記事をテキストマイニングの技術を用いて解析し，経済市場を分析する研究などが盛んに行われている [1][5].

本研究では，毎月のファンドの運用報告書に記載される市況分析コメントを自動生成する手法を提案する．市況分析コメントとは，例えば「8 月の国内株式市況は、中国の景気減速懸念が台頭したことなどを背景とした世界的な株安を受けて大きく下落しました。」のような，その月における株価が大きく変動したイベント（例えば,「人民元の基準値切り下げ」）について述べ，株価が変動した理由を分析した文書である．以下に，2015 年 8 月のファンド運用報告書に記載された市況分析コメントの一部を示す.

> 8 月の国内株式市況は、2015 年度第 1 四半期決算で好業績を発表した企業への期待などを背景に上昇して始まりました。しかしながら中国人民銀行が人民元の対米ドルでの基準値切り下げを実施すると中国経済への減速懸念が広がり、国内株式市況は下落しました。さらに、中国経済の減速が世界景気へ及ぼす影響などを警戒して投資家がリスク回避姿勢を強めると世界的に株式市況は急落しました。…

上記のような市況分析コメントを記述するために，ファンド運用の担当者は①日経平均株価が大きく動いた記事を調べ，②その前後にあったイベントを確認し，③その記事の中から株価が変動した理由について述べている文を選択し，④まとめる，という作業を毎月，行う必要がある．現在のところ，市況分析コメントの作成はファンドごとに運用担当者が行っており，ファンドの特色に従ってファンドごとに異なる．しかし，①～③については共通化しＡＩで自動化できれば，ファンドの運用担当者の負担を減らすことが可能である [1].

そこで，本研究では，上記の①日経平均株価が大きく動いた記事の判定，②その前後にあったイベントの確認，③その記事の中から株価が変動した理由について述べている文を抽出，といった処理を自動化し，③

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp
†本論文に示された所見は著者自らのものであり，所属する組織の公的な立場を代表するものではない.

[1]④についてはファンドの特色に合わせ，ファンド担当者がまとめてもよい.

の処理によって抽出された文を時系列順に並べることで市況分析コメントを自動生成することを目的とする．

2 関連研究

関連研究として，決算短信を対象として様々な情報を抽出する研究がある．酒井らは決算短信から業績要因を含む文（例えば「半導体製造装置の受注が好調でした。」）を抽出する手法を提案している [8][9]．坂地らは決算短信から原因・結果表現を抽出する手法を提案している [10]．北森らは決算短信から業績予測文 (今後の業績予測に関する情報が記述されている文) を抽出する手法を提案している [2][3]．いずれの研究も抽出対象である情報を抽出するために有効な手がかり表現に着目し，それらの表現をブートストラップ的に自動的に獲得，もしくは，人手にて用意している．また，深層学習を用いて情報を抽出している手法もあるが，手がかり表現を使用して抽出したデータを学習データとすることで，学習データの自動生成を試みている．本研究では，③その記事の中から株価が変動した理由について述べている文を抽出する処理のために手がかり表現を用いるが，その手がかり表現の獲得のために酒井らの手法 [8] を用いている．ただし，最初の入力する初期手がかり表現は，本研究における抽出対象にあわせ「で買い」「が買い」「で売り」「が買い」に変更している．

本研究は，複数の記事から 1 つの要約を生成する複数文書要約とみなすことができる．複数文書要約に関しては多くの研究があり，例えば酒井らは，ユーザが知りたい情報を「要約要求」と定義し，要約要求を反映した要約を生成するために，ユーザとのインタラクションを導入した複数文書要約システムを提案している [7]．複数文書要約では，入力として複数の記事が与えられ，それらの記事から重要な文を抽出し，重要な文同士の冗長性を排除してまとめるという処理を行い，入力された記事の内容をまんべんなく含むような要約を生成することが求められる．それに対して，本研究で与えられる記事には市況分析コメントを生成するためには不要な記事も含まれており（すなわち，日経平均株価の市況について述べてはいるが，大きく動いた日ではない記事），そのような記事を排除する必要がある点が異なる．さらに，市況分析コメントの重要文として株価が変動した理由について述べている文を抽出している点も異なる．

3 市況分析コメント自動生成手法

3.1 手法概要

本手法では，①日経平均株価が大きく動いた記事の判定，②その前後にあったイベントの確認，③その記事の中から株価が変動した理由について述べている文を抽出，の順で処理を行い，市況コメントを生成する．手法の概要を以下に示す [2]．

Step 1: ある期間の日経平均について言及している記事から，日経平均が大きく変動したことについて述べた記事（以降，分析記事と定義）を深層学習により抽出．

Step 2: ある期間の日経平均について言及している記事集合から，その期間における重要なキーワード（以降、重要キーワード）を抽出（例 : 人民元，中国人民銀行）．

Step 3: Step 1 で抽出した分析記事集合より，日経平均が大きく変動した要因について述べた文（以降，要因文と定義）を抽出．

Step 4: Step 3 で抽出した要因文集合と，Step 2 で抽出した重要キーワード集合を使用し，重要な文を判定

Step 5: Step 4 で判定された重要な文を時系列順にならべ，市況分析コメントとする

3.2 分析記事の抽出

市況分析コメントを自動生成するための情報源として日経平均株価について言及している記事を使用する．しかし，日経平均株価について言及している記事は，ほぼ毎日 1 つは存在するため，それらの記事全てを市況分析コメント自動生成のための情報源として使用すれば重要ではない情報も混ざる．日経平均株価の実データを使用して記事を選別する方法でもいいが，その場合は日経平均株価の変動とその日に対応する記事をセットで用意する必要があり，入力フォーマットが複雑になる．そこで，日経平均株価が大幅に変動したことに言及した記事（分析記事）を深層学習にて自動的に抽出する．以下に分析記事の一部を示す．

[2] Step 1 が①に，Spte 2 が②に，Step 3 が③に該当する．

１２日の東京株式市場で日経平均が大幅に続落し、下げ幅は一時３００円を上回った。中国の景気減速への警戒感が広がり、運用リスクを減らす動きが優勢になっている。人民元切り下げが発表になった午前１０時１５分すぎから先物に海外勢からとみられる大口の売りが断続的に出て、日経平均は下げを加速した。…

3.2.1 学習データの自動生成

深層学習のための学習データは，１４年分の日本経済新聞記事のタイトルに「日経平均」が含まれている記事から以下の手法で自動生成した（2772 記事を生成）．

正例：第 1 文に「日経平均株価は大幅」、「日経平均株価が大幅」が含まれている記事

負例：第 1 文に「日経平均」がない

上記の手法により，以下のような学習データが自動生成される．

十日の東京株式市場で日経平均株価が大幅続落し、終値で一万一〇〇円台に下落した。バブル経済崩壊後の安値を再び更新し、一九八四年八月以来の水準となった。主力のハイテク株や通信株が相場の下げを主導。銀行、鉄鋼、不動産など幅広い銘柄に売りが膨らんだ。…

負例としては，インタビュー記事などが抽出される．

株安が直ちに一九九八年のような金融システム不安を引き起こすことはない。大手銀行の自己資本比率は一〇％を超えており、株価下落による比率低下は限定的。 …

3.2.2 素性選択

自動生成された学習データから入力層の要素となる語（素性）を選択する．具体的には，自動生成された学習データにおいて正例の記事に含まれる内容語（名詞，動詞，形容詞）に対して，以下の式 1 にて重みを計算する．

$$W_p(t, S_p) = TF(t, S_p) \times H(t, S_p) \tag{1}$$

ただし，

S_p**:** 学習データにおいて正例に属する記事集合

$TF(t, S_p)$**:** 記事集合 S_p において，語 t が出現する頻度

$H(t, S_p)$**:** 記事集合 S_p における各記事に含まれる語 t の出現確率に基づくエントロピー

$H(t, S_p)$ が高い語ほど，正例の記事集合に均一に分布している語であることが分かる．$H(t, S_p)$ は次の式 2 で求める．

$$H(t, S_p) = -\sum_{s \in S_p} P(t, s) \log_2 P(t, s) \tag{2}$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \tag{3}$$

ここで，$P(t, s)$ は記事 s における語 t の出現確率を表し，$tf(t, s)$ は記事 s において語 t が出現する頻度を表す．

次に，負例の記事に含まれる内容語（名詞，動詞，形容詞）に対しても，同様に重みを計算する．

$$W_n(t, S_n) = TF(t, S_n) \times H(t, S_n) \tag{4}$$

ただし，S_n は学習データにおいて負例に属する記事の集合である．

ここで，ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ の 2 倍より大きければ，その語 t を素性として選択する．もしくは，語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の 2 倍より大きければ，その語 t を素性として選択する．上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，正例，負例，ともによく出現するような一般的な語を素性から除去する．上記の手法により，61,678 文の学習データから 4,549 語が素性として選択された．以下に選択された素性の一部を示す．

平均, 市場, 株価, 株式, 投資, 相場, 証券, 売り, 買い

3.2.3 深層学習による分析記事の抽出

深層学習のモデルについて以下に述べる．入力は，2772 記事の学習データから素性として抽出された 4,252 語を要素，語 t における $W_p(t, S_p)$，もしくは，$W_n(t, S_n)$ の大きいほうを要素値としたベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数と同じ 4,252 とし，隠れ層は，ノード数 1,000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素である．また，活性化関数として，ReLU を使用した．学習されたモデルにより，市況コメントを生成する期間の日経新聞記事から分析記事を抽出する．テストデータは，記事のタイトルに「日経平均」が含まれる記事である．

3.3 分析記事からの要因文の抽出

分析記事を並べただけでは市況コメントとして長すぎるため，分析記事から日経平均が大幅に変動した理由について言及している文（要因文）を抽出する．例えば，以下のような文を抽出する．

> ・ギリシャへの金融支援協議が難航していることや、中国・上海株の値動きへの警戒感から、投資家のリスクオフの動きが強まった。
> ・米消費者物価指数の上昇やイエレン米連邦準備理事会議長が年内の利上げに前向きな姿勢を示し、為替相場が円安に向かった。

要因文の抽出は,「強まった」のような手がかり表現や,「ギリシャ」「金融支援」「イエレン米連邦準備理事会議長」といった，その期間における重要なキーワードを使用して抽出する．しかし，期間ごとの重要キーワード，有効な手がかり表現は数多く，全て人手で用意することは困難である．そこで，手がかり表現と重要キーワードを自動獲得する．

関連研究でも述べたが，要因文の抽出は酒井らが決算短信から業績要因文を抽出するために開発した手法 [8] を使用する．手がかり表現は以下の手法で獲得される．

Step 1: 少数の手がかり表現（具体的には,「が買い」,「で売り」の 2 表現を用いる）を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現（「警戒感」,「国際優良株」など）を共通頻出表現と定義し，抽出する．

Step 3: 共通頻出表現が係る節を取得し，その中から新たな手がかり表現を抽出する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 1 を参照）．□

Step 2 において，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 5 で求め，その値が，ある閾値以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \qquad (5)$$

ここで，分析記事の集合において，

$S(e)$: 共通頻出表現 e が係る手がかり表現の集合．

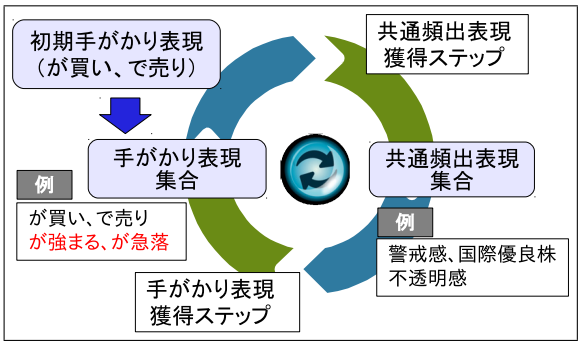


図 1: 共通頻出表現・手がかり表現自動獲得手法の概要

$P(e,s)$: 共通頻出表現 e が手がかり表現 s に係る確率．

同様に，Step 3 において，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを求め，その値が，ある閾値以上の手がかり表現を選別する．

以上の手がかり表現，共通頻出表現の選別処理を行うことで，例えば以下のような適切な手がかり表現を獲得する．

> が上昇，が強まる，が膨らんでいる，が急落，が堅調だった，が大幅下落した，で大幅続伸，が根強い，が急伸した，が後退，が加速した，が先行している

期間ごとの重要キーワードの抽出は，期間 t の分析記事における名詞 n に対して，以下の式 6 で重み $W(n, S(t))$ を計算することで行う．

$$W(n,S(t)) = (0.5 + 0.5 \times \frac{tf(n,S(t)}{\max tf(n,S(t)}) \times H(n,S(t)) \times \log_2 \frac{N}{df(n,N)} \qquad (6)$$

ここで，

$S(t)$: 期間 t の分析記事の集合．

$tf(n,S(t))$: $S(t)$ において，名詞 n が出現する頻度．

$H(n,S(t))$: $S(t)$ の各分析記事である d に名詞 n が出現する確率 $P(n,d)$ に基づくエントロピー．

$df(n,N)$: 記事集合 N で名詞 n を含む記事の数．

N: 日経平均について言及している記事の総数．

ここで，N の日経平均について言及している記事とは，2000 年から 2014 年までの日経新聞記事において，タイトルに「日経平均」を含む記事であり，その総数は 14,325 記事である．

表 1: 期間ごとの重要キーワードの例

期間	重要キーワード
1999 年 11 月（ＩＴバブル）	情報通信，情報通信関連株，年初来高値
2008 年 9 月（リーマンショック）	米政府，金融安定化，リーマン・ブラザーズ，総合金融安定化策，破綻
2015 年 7 月（ギリシャ問題）	ギリシャ，離脱，金融支援，中国株，国民投票，欧州連合

$W(n, S(t))$ は，情報検索で一般的な $tf \cdot idf$ 値を 1 つの期間の分析記事の集合を 1 つの文書とみなして求め，さらに，その期間の分析記事集合においてまんべんなく出現している場合に高い値をとる尺度を組み合わせたものである．表 1 に，上記の手法によって，期間ごとの分析記事から抽出された重要キーワードをいくつか示す．分析からの要因文の抽出は，分析記事から手がかり表現と重要キーワードがともに含まれている文を抽出することで行う．

3.4 市況分析コメントの自動生成

分析記事から抽出した要因文を時系列順にならべることで，市況分析コメントを生成する．ここで，分析記事から抽出した要因文を全て採用すると，市況分析コメントとして長すぎる場合がある．そのため，要因文にふくまれる重要キーワードのスコアの和を分析記事のスコアとし，スコアが上位の分析記事に含まれる要因文を時系列順に並べて，市況分析コメントとする．2015 年 8 月のタイトルに「日経平均」を含む 33 記事から本手法にて生成された市況分析コメントを以下に示す．

> 今月の国内株式市況は、◆１１日の日経平均株価は前日の米国株高を手掛かりに一時１３８円高を付けたが、人民元切り下げで一転して売りが優勢となり、２２６円安まで値下がりする場面があった。人民元安は米国の利上げ観測から軟調に推移していたアジア通貨にも波及。アジアの主要な株式市場でも売りが広がった。◆１１日の東京株式市場で日経平均株価は一時、年初来高値を上回る水準に上昇した。しかし、中国人民銀行による人民元の実質切り下げを機に下落に転じた。人民元の切り下げの影響は日本にとどまらず、欧米やアジアなどの株式市場にも及んだ。◆中国の景気減速から始まった世界市場の動揺がいったん収まり、株式や原油などリスク資産の買い戻しが活発になっている。２８日の日経平均株価は３日続伸し、直近安値の２５日に比べて７％上げた。一方、一部の新興国通貨への売り圧力はなお強く、混乱再燃の懸念はくすぶっている。

表 2: 評価結果

期間	本手法	Baseline	運用担当者
1 月	**0.2**	0.12	0.48
2 月	**0.1**	0.05	0.51
3 月	**0.2**	0.06	0.27
4 月	**0.13**	0.1	0.49
5 月	**0.27**	0.1	0.59
6 月	0.10	**0.13**	0.41
7 月	0.18	**0.2**	0.38
8 月	**0.35**	0.22	0.58
9 月	**0.14**	0.06	0.43
１０月	**0.17**	0.1	0.4
１１月	**0.25**	0.15	0.25
１２月	**0.21**	0.1	0.36

4 評価

本手法の評価を行うため，本手法を実装した．実装にあたり，形態素解析器として MeCab[3]，係り受け解析器として CaboCha[4] を使用した．評価方法は，実際に運用担当者が作成した市況分析コメント（2015 年 1 月～12 月）と，その期間の日経新聞記事を使用して本手法により自動生成された市況分析コメントとを比較して行った．具体的には，ある期間における運用担当者が作成した市況分析コメントと自動生成された市況分析コメントとの類似度を，その文書の名詞を要素，要素値として TF・IDF 値を使用したベクトル空間モデルで求める．そして，その期間における運用担当者が作成した市況分析コメントとの類似度の平均を評価値とした．ここで，比較手法として以下の手法と比較した．

Baseline: 入力された記事の第一文を連結して生成

運用担当者: 運用担当者が作成した市況分析コメントの 1 つを選択

運用担当者は理想的な結果に基づく評価値となる．結果を表 2 に示す．表の太字は，本手法と Baseline とを比較し，大きいほうを示す．

[3] http://taku910.github.io/mecab/

5 考察

本手法と Baseline とを比較すると，本手法のほうが概ね高い類似度を達成している．しかし，理想的な結果である運用担当者とは肉薄している月もあるが，まだ大きな差があることが分かる．5月，8月は重要な発言や株価に影響が大きなイベントがあったこともあり，高い類似度を達成しているが，6月は大きなイベントがなく（その月の日経平均の高値から安値を引いた変動幅が最も小さかった），そのような場合は類似度が低くなってしまう．

本研究における評価は，正解データである実際に運用担当者が作成した市況コメントとの類似度を測ることで行っている．しかし，本評価手法では，語の一致する割合が多いと類似度が高くなる傾向になるため，内容の冗長性や網羅性を評価できているわけではない．そのため，Baseline では冗長性が高い市況コメントが作成されているにもかかわらず，類似度が高くなることもある．テキスト自動要約では，ある文書から人間が作成した要約と自動生成された要約の間で一致する N グラムの割合で評価値を求める ROUGE[6] という評価手法が一般的によく用いられている．しかし，本タスクの場合，人間が作成した要約（運用担当者が作成した市況分析コメント）は，本研究でいうところの分析記事をもとに作成しているわけではないので，ROUGE で評価することが妥当ではなかった．理想的には，運用担当者が作成した市況分析コメントを評価者が読んで自動生成された市況分析コメントと比較し，どの程度，内容が一致しているか，冗長性が除かれているかを人手にて評価すべきであるが，それを行うにはある程度の専門知識が必要であることと再現性が困難であることから，今回は行えなかった．今後の課題として，本タスクにおける評価手法の確立が必要であると考える．

6 まとめ

本稿では，経済新聞記事などの経済テキストから、例えば日経平均株価などの市況について言及している文書のみを抽出し，それらの内容を自動的に要約することによりマーケットレポートにおける市況コメントを自動生成する手法について述べた．本手法では，まず経済新聞記事から深層学習により日経平均株価が大幅に変動したことについて言及している記事を抽出し，次に抽出された記事の中からその理由について言及している文を抽出した．そして，抽出された文を時系列順に並べることで市況コメントを自動生成した．評価の結果，入力された記事の第一文を連結して生成した文書より概ね高い類似度を達成した．

参考文献

[1] 和泉潔, 後藤卓, 松井藤五郎: 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, Vol. 52, No. 12, pp. 3309–3315 (2011).

[2] 北森詩織, 酒井浩之, 坂地泰紀: 決算短信 PDF からの業績予測文の抽出, 電子情報通信学会論文誌 D, Vol. J100-D, No. 2, pp. 150–161 (2017).

[3] Kitamori, S., Sakai, H. and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 67–73 (2017).

[4] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[5] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志: 新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296 (2013).

[6] Lin, C.-Y. and Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, *the 4th Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technology*, pp. 150–157 (2003).

[7] 酒井浩之, 増山繁: ユーザの要約要求を反映するためにユーザとのインタラクションを導入した複数文書要約システム, 日本知能情報ファジィ学会誌, Vol. 18, No. 2, pp. 265–279 (2006).

[8] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信 PDF からの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).

[9] 酒井浩之, 松下和暉: 決算短信からの業績要因文の抽出, 第 11 回テキストアナリティクス・シンポジウム, pp. 87–91 (2017).

[10] 坂地泰紀, 酒井浩之, 増山繁: 決算短信 PDF からの原因・結果表現の抽出, 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822 (2015).

ベクトル表現を用いた因果関係連鎖の抽出

Extraction of Causal Relation Chains using Vector Expressions

西村弘平 [1] *　坂地泰紀 [2] 和泉潔 [2]

[1] 東京大学工学部システム創成学科システムデザイン&マネジメントコース
[1] Department of Systems Innovation, Faculty of Engineering, The University of Tokyo
[2] 東京大学大学院工学系研究科
[2] Graduate School of Engineering, The University of Tokyo

Abstract: 複数のテキストデータから経済・金融事象を背景知識まで含めて可視化することは, 経済・金融事象の理解の助けになり有用である. しかしながら, 経済・金融事象の連鎖を手動で抽出することは非常に時間とコストがかかる. そこで, 本研究では経済・金融事象の連鎖を因果関係として扱い, 各事象を表したベクトル間の類似度を用いて因果関係の連鎖を構築する手法を提案する. また, 提案手法における問題点から今後の提案をまとめる.

1　はじめに

近年, 人工知能分野の手法や技術の金融市場の様々な場面への応用が期待されており, 膨大な金融情報を分析して投資・経営判断を支援する技術が注目されている. 特に, 投資家・経営者は経営・金融事象の情報を把握し, 各事象を正しく理解して投資や経営の判断をする必要があるため, 複数のテキストデータから抽出した因果関係を用いて因果ネットワークを構築し, 事象を背景知識とともに可視化する技術は有用である. 本論文では上記の背景を踏まえ, テキストデータから抽出した因果関係ノード間の原因表現と結果表現の表現類似度を測ることによってテキストデータから因果関係の連鎖を構築する手法を提案する.

2　先行研究

本章では, 提案手法に関連する因果関係抽出・因果関係連鎖構築の先行研究について述べる.

2.1　因果関係抽出

Khoo et al.[1], 乾ら [2] は接続関係や格フレームを用いて因果関係を抽出する手法を提案している. これらの手法は単文もしくは複文・重文からしか因果関係を抽出できないという問題点がある. 坂地ら [3] は文書中にある因果関係を, 手がかり表現を用いて抽出する手法を提案している. 坂地らの手法は因果関係が存在する構文パターンを列挙し, 手がかり表現を用いることによって単文, 複文・重文関係なく因果関係にある表現を抽出することができる. 本提案手法では坂地らの手法を利用して, 因果関係を抽出する. また, 本論文での因果関係の定義を坂地らと同様に, 「原因若しくは, 理由と結果を示し, 手がかり表現を伴って 1 文中, もしくは隣り合う 2 文中に表層的に出現するもの」とする.

2.2　因果関係連鎖構築

Ishii et al.[4] は SVO の構造と WordNet を用いてニュース記事から因果関係の連鎖を構築する手法を提案している. Ishii et al. は SVO 構造を利用することで注目する単語を決め, WordNet を用いた単語のマッチングによって概念的な単語の類似度を計算している. 津川ら [5], [6] は WebAPI を用いて要因結果検索を用いて, 事象間の共起度を測定し因果関係の連鎖を構築している. 津川らの WebAPI を用いる手法は検索エンジンのアルゴリズムによって影響を受けるため, 手法の再現性が低いと考えて, 比較手法には Ishii et al. の手法を用いた.

3　提案手法

本章では, テキストデータから因果関係の連鎖を構築する手法について述べる. 3.1 節から 3.4 節までで因果関係連鎖構築の概要を述べ, 3.5 節で因果関係連鎖構築に用いる因果ノード間類似度計算手法について述べる. 手法の概要は図 1 の通りである.

*東京大学工学部システム創成学科システムデザイン&マネジメントコース
〒 113 ー 8656 東京都文京区本郷 7 ー 3 ー 1 工学部 8 号館 530 室
E-mail: b2017knishimura@socsim.org

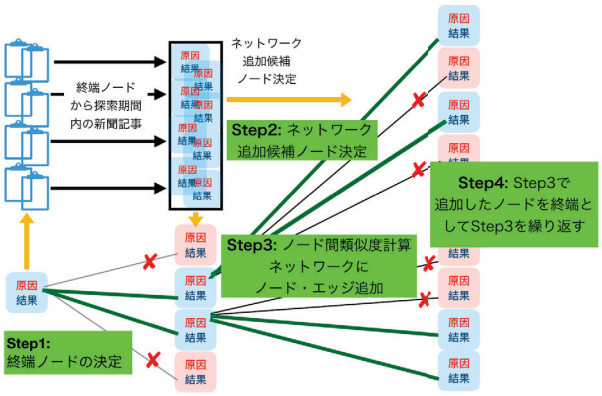


図 1: 因果関係連鎖構築手法の概略

3.1 因果関係連鎖の終端ノードの決定

決算短信のテキストから坂地ら [3] の手法を用いて因果関係表現を抽出する．抽出した因果関係ノードの中から市場の情勢や企業の業績を記した因果関係ノードを手動で選択し，因果関係連鎖の終端ノードとする．

3.2 探索対象の因果関係ノードの探索

3.1 節で定めた終端ノードを終端とする因果関係連鎖への追加候補の因果関係ノードを日本経済新聞のテキストから抽出する．

因果関係抽出は 3.1 節と同じく坂地ら [3] の手法を用いる．3.1 節で定めた終端ノードよりも過去のもので、終端ノードから探索期間 S 以内の因果関係を因果関係連鎖の追加候補とする．さらに坂地らの手法で抽出した因果関係ノードのうち，原因・結果表現の名詞・形容詞・動詞の単語数の合計がともに閾値 β 以上である因果関係ノードのみを因果関係連鎖の追加候補とした．

3.3 因果関係連鎖へのエッジ・ノードの追加

3.2 節で抽出した因果関係連鎖への追加候補の全ての因果関係ノードと終端ノードの組み合わせについて因果関係ノード間の類似度を計算する．因果関係ノード間の類似度が閾値 α 以上であるときにノードを追加して因果関係連鎖を拡張する．

3.4 因果関係連鎖の更新

3.2 節で抽出した因果関係ノードを因果関係連鎖への追加候補ノード，3.3 節で因果関係連鎖に追加したノードを終端ノードとして 3.2 節，3.3 節の処理を $n-1$ 回繰り返す．ここで，n はあらかじめ定めた因果関係連鎖の更新回数である．

3.5 因果関係ノード間の表現類似度計算手法

本節では，提案手法で用いている因果関係ノード間の計算方法について述べる．計算手法の概要は図 2 の通りである．因果関係ノード間の類似度は 2 種類の類似度を足し合わせることによって計算する．1 つは IDF 値を用いて作成したベクトル間のコサイン類似度で，もう 1 つは Bojanoski et al.[7] の FastText から求めた単語分散表現を用いた類似度である．IDF 値を用いたベクトル表現を用いることによって，因果関係が含まれている領域やドメインなどのトピック情報の類似度を計算し，FastText から得た分散表現からなるベクトル表現を用いることで比較する因果関係の原因・結果表現間の表層的な類似度を計算する．ベクトル間類似度はコサイン類似度を 0 から 1 の値取るように正規化したものとした．

$$cosine\_similarity = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

$$vector\_similarity = \frac{(cosine\_similarity + 1)}{2} \quad (2)$$

IDF 値を用いた類似度は，因果関係連鎖への追加対象のノードの結果表現と終端ノードの原因表現から IDF 値上位 3 つの単語を抽出し，IDF 値を用いてベクトルを作成しベクトル間の類似度を計算する．IDF 値は 1998 年から 2016 年までの全ての原因・結果表現から抽出した．FastText を用いた類似度は，因果関係連鎖への追加対象ノードの結果表現と終端ノードの原因表現内にある名詞・動詞・形容詞の分散表現を足し合わせてベクトルを作成し，ベクトル間類似度を計算する．FastText の類似度は cbow と skip-gram の 2 つのアルゴリズムに対して計算した．IDF, FastText の合計 3 つのベクトル類似度の平均を取ることによってノード間の類似度とした．

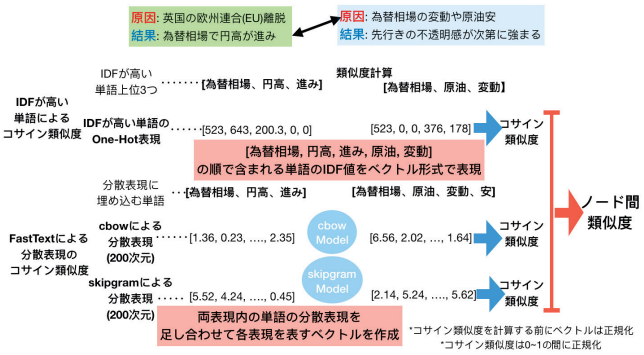


図 2: 因果関係ノード間類似度の計算方法

4 実験

実験には 1982 年から 2016 年までの 54206 件の決算短信テキストと 1998 年から 2016 年までの 6921 日分の記事テキストを用いた.

決算短信テキストから抽出した因果関係の中から 5 つの終端ノードを手動で選び, 日本経済新聞テキストから因果関係連鎖への追加候補ノードを抽出して因果関係の連鎖を構築する実験を行なった. 比較手法に Ishii et al.[4] の WordNet を用いたノード間類似度を計算する手法を用いた. 終端ノードは表 4 実験のパラメータは表 3 の通りである. 比較手法はノード間類似度の値が 0, 0.33, 0.66, 1 のいずれかなので 0.65 のみ実験を行なった.

S, 因果関係連鎖への追加候補ノード探索期間: 52 週間

n, 因果関係連鎖の更新回数: 4

α, 因果関係連鎖にマージするときのノード間類似度の閾値: 0.65, 0.70, 0.75

β, 因果関係連鎖への追加候補ノードの原因・結果表現の単語数についての閾値: 5

図 3: 因果関係連鎖構築に用いたパラメータ

ID1: [原因] **為替相場の変動や原油安** [結果] **先行きの不透明感が次第に強まる, 中央魚類株式会社, 2016-11-2**

ID2: [原因] **海外経済の減速懸念や個人消費の低迷, 資源価格安の長期化,** [結果] **景気の先行きに不透明感が出てまいりました, 矢作建築工業株式会社, 2016-5-9**

ID3: [原因]**3 中国や韓国を中心に全世界で鉄鋼生産能力増強が進行し, 過剰な生産設備による供給過剰問題が顕在化する,** [結果] **世界的な鉄鋼需給バランスが大きく崩れた, 合同製鐵株式会社, 2016-4-28**

ID4: [原因] **米国のゼロ金利政策解除による金融市場の変動, 中国経済の減速, 原油価格の下落などの影響,** [結果] **先行きが不透明な状況で推移しました, クエスト株式会社, 2016-4-1**

ID5: [原因]**EU 情勢不安や中国経済減速の影響,** [結果] **売り上げが減少しています, 石塚硝子株式会社, 2016-4-25**

図 4: 終端ノードに用いた因果関係ノード

5 実験結果と考察

手法の評価指標には Precision(精度) とノード数を用いた. 1 日の新聞テキストから約 300 件の因果関係ノードが抽出されるため、正解データを作成できず再現率の代わりにノード数を評価指標に加えた.

Precision の計算方法は, 構築した因果関係連鎖に対してランダムに 100 個のノードを抽出し, ノード間関係が「経済・金融事象を背景知識まで理解するために適切なつながりであれば正しい, そうでなければ誤り」という定義に従って, 因果関係の連鎖として適切か手動で判断し評価を行なった. 各ノード間関係の評価は著者 1 人のみが行なった.

Precision(精度), ノード数をそれぞれ表 1, 表 2 に記す. 手法名は比較手法, 提案手法をそれぞれ比較, 提案と記している.

精度, ノード数がともに高い手法がよりノード間の類似度をより正確に計算できていると言える. 提案手法では, 比較手法に比べて精度では全てのノードについて, ノード数にもおいても 3 つの終端ノードに対して比較手法よりも多くノード数を抽出できており, 提案手法が比較手法よりも精度高くノード間の類似度を計算できていると言える.

表 1: 各実験における精度の差

手法 (α)	ID1	ID2	ID3	ID4	ID5
比較 (0.65)	0.00	0.02	0.03	0.00	0.00
提案 (0.65)	**0.12**	**0.08**	**0.10**	**0.11**	**0.05**
提案 (0.70)	0.39	0.44	0.39	0.46	0.30
提案 (0.75)	1.00	0.48	0.83	0.56	0.60

表 2: 各実験におけるノード数の差

手法 (α)	ID1	ID2	ID3	ID4	ID5
比較 (0.65)	**82392**	24779	**1970587**	56027	11073
提案 (0.65)	80323	**97814**	90978	**100273**	**46429**
提案 (0.70)	44	424	41	288	115
提案 (0.75)	2	26	6	70	10

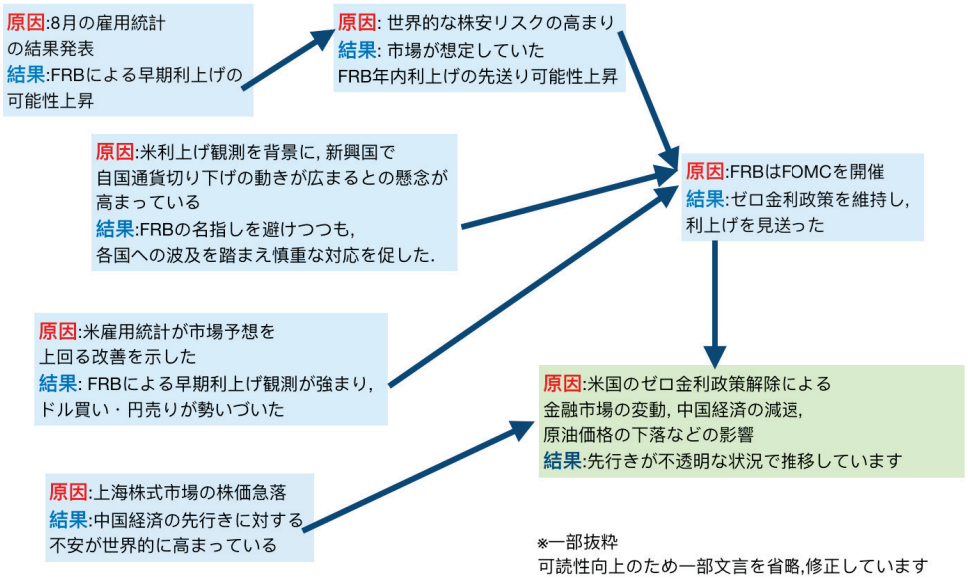


図 5: 構築した因果関係連鎖の例

終端ノードを ID5 としたときの具体的な因果関係連鎖の構築事例から抜粋したものを図 5 に示す.

構築した因果関係では, 「米国のゼロ金利政策解除による金融市場の変動, 中国経済の減速, 原油価格の下落」という経済事象の原因となったアメリカの金利政策の流れ, 上海市場の動向を因果関係の連鎖として取得できていることがわかる.

6 まとめ

本研究では, ベクトル表現の類似度を用いて決算短信と日本経済新聞のテキストから因果関係連鎖の構築手法を提案した. また, 提案手法が WordNet を用いた既存手法よりも精度高く因果関係間の類似度を計算できることを実験で確認した. 提案手法のベクトル表現は WordNet といった事前知識を必要としないため, 新たな単語を含むテキストからでも因果関係の連鎖を構築することができ有用である.

今後の課題としては類似だけでなく, 包含・例示・時系列推移といった類似以外の事象間関係を区別して因果関係の連鎖を構築すること, また, 因果表現のみだと経済・金融事象の情報量が少ないことがあるため, 周辺テキストから精度高く経済・金融事象の情報を抽出することの 2 点が挙げられる.

参考文献

[1] C. S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 336–343, Association for Computational Linguistics, 2000.

[2] 乾孝司, 乾健太郎, 松本裕治, *et al.*, "接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得," **情報処理学会論文誌**, vol. 45, no. 3, pp. 919–933, 2004.

[3] 坂地泰紀, 酒井浩之, and 増山繁, "決算短信 pdf からの原因・結果表現の抽出," **電子情報通信学会論文誌** *D*, vol. 98, no. 5, pp. 811–822, 2015.

[4] H. Ishii, Q. Ma, and M. Yoshikawa, "Incremental construction of causal network from news articles," *Journal of information processing*, vol. 20, no. 1, pp. 207–215, 2012.

[5] 津川敦朗, 新妻弘崇, and 太田学, "交絡事象の発見による因果関係ネットワークの改良." DEIM Forum, E2-3, 2015.

[6] 津川敦朗, 新妻弘崇, and 太田学, "共起関係に着目した因果関係ネットワークの拡張." DEIM Forum, F3-2, 2016.

[7] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

深層学習を用いて極性付与されたアナリストレポートと株式リターンとの関連性

平松賢士 [1]　酒井浩之 [2]　坂地泰紀 [3]

Kenji Hiramatsu [1], Hiroyuki Sakai[2], Hiroki Sakaji[3]

[1] 株式会社アイフィスジャパン，株式会社金融データソリューションズ
[1]IFIS JAPAN LTD.，Financial Data Solutions, Inc.
[2] 成蹊大学 理工学部 情報科学科
[2] Department of Computer and Information Science, Faculty of Science and Technology, Seikei University
[3] 東京大学
[3] The University of Tokyo

Abstract: アナリストレポートは，証券会社のアナリストが企業の経営状態や収益力などを調査してまとめたものである．以前の研究では，学習データを自動生成し，深層学習を用いてアナリストレポートに対して極性を付与する手法を提案した．そこで本研究では，深層学習を用いて極性付与されたアナリストレポートと株式リターンとの関連性について実証分析を行った．その結果，アナリストによる定量情報に対する予想変更が無いレポートについて，市場は該当レポートに付与された極性に沿って短期的に反応することや，レポートにて言及された銘柄は，極性の違いによってその後の長期にわたり株式リターンに違いが表れることが明らかになった．

１．はじめに

近年，投資家に対して投資判断の支援を行う技術の必要性が高まり，人工知能分野の手法や技術を金融市場における様々な場面に応用することが期待されている．例えば，決算短信から重要な情報を抽出して投資判断の支援を行うといった研究が行われている[3][4][5][6][7].

本研究において分析対象となるアナリストレポートは，証券会社のアナリストが企業の経営状態や収益力などを調査してまとめたものである．業績予測や事業の今後の展望などが記載されており，予想を元にレーティングが付与される．高度な専門知識をもつアナリストによるレポートは，投資判断のための重要な情報源のひとつであり，株価の変動要因にもなりうる．多い時には 1 日に 1200 本以上ものアナリストレポートが発表されることもあるため，全てのレポートに目を通し，内容を把握することは困難である．人工知能分野やテキストマイニングの手法を用いて投資判断を支援する技術が求められており，様々なアプローチで研究が行われている[1][2].

以前の研究では，アナリストレポートの重要度を判断し取捨選択するための支援技術として，学習データを自動生成し，深層学習を用いて極性（ポジティブ，ネガティブ）を付与する手法を提案した[1]．そこで本研究では，アナリストレポートに付与された極性と株式リターンとの関連性について実証分析を行う．

極性付与は，アナリストレポートに記載されたレーティングや目標株価，業績予想等の定量情報には寄らず，レポートのテキストデータのみを元に付与されるものである．レポートに付与された極性と株式リターンとの関連性を分析することは，アナリストレポート本文に記載された内容が市場にどのようなインパクトを与えているか理解することや，レポート本文にて記載のされ方が違う銘柄について株式リターンに違いがあるかといった事柄を理解することに繋がる．

アナリストレポート本文に対して実証分析を行った先行研究としては太田[8]があり，人手でヘッドラインを読み，調査する 14 項目毎に数値化していく方法をとっている．

２．アナリストレポートへの極性付与

本研究では，アナリストレポートへの極性付与については酒井らの手法[1]を使用する．以下，酒井らの手法[1]について簡単に述べる．分析対象とするレポートは，2012 年から 2016 年の間で主要証券会社 15 社から発行された銘柄レポート（個別に銘柄について言及しているレポート）166,223 本とする．またアナリストレポートからのテキストデータ抽出においては，アイフィスジャパン独自のクレンジング技術を使用して，本文以外のノイズとなるデータを可能な限り除去している．

2.1　学習データの自動生成

酒井らの手法[1]により学習データを生成する．アナリストレポートにはレーティングが付与されており，レーティングが上がっていれば，そのアナリストレポートはポジティブな内容が記述されていることが予想できる．同様にレーティングが下がっていればネガティブな内容と予想できる．そこで，2012 年から 2013 年の間に発行された，レーティングが上がったアナリストレポートを正例，レーティングが下がったアナリストレポートを負例とした学習データを生成する．この時，正例と負例の数を揃えるため，条件に合致するレポートのうち正例 1,630 本，負例 1,630 本の計 3,260 本を使用した．

2.2　素性選択

学習データから入力層の要素となる語（素性）を選択する．自動生成された学習データにおいて，正例に含まれる内容語（名詞，動詞，形容詞）に対して，式 1 で重みを計算する．

$$W_p(t,S_p) = TF(t,S_p)H(t,S_p) \qquad (1)$$

ただし，

S_p: 学習データにおける正例のアナリスト予想根拠文の集合．

$TF(t,S_p)$: 文集合S_pにおいて，語tが出現する頻度．

$H(t,S_p)$: 文集合S_pにおける各文に含まれる語tの出現確率に基づくエントロピー．

$H(t,S_p)$が高い語ほど，正例の文集合に均一に分布していることがわかる．$H(t,S_p)$は次の式 2 で求める．

$$H(t,S_p) = -\sum_{s\in S_p} P(t,s)log_2P(t,s) \qquad (2)$$

$$P(t,s) = \frac{tf(t,s)}{\sum_{s\in S_p} tf(t,s)}$$

ここで，$P(t,s)$は文sにおける語tの出現確率を表し，$tf(t,s)$は文sにおいて語tが出現する頻度を表す．次に，負例の文に含まれる内容語（名詞，動詞，形容詞）に対して，式 3 で重みを計算する．

$$W_n(t,S_n) = TF(t,S_n)H(t,S_n) \qquad (3)$$

ただし，S_nは学習データにおいて負例に属する文の集合である．

ある語tの正例における重み$W_p(t,S_p)$が負例における重み$W_n(t,S_n)$の 2 倍より大きければ，その語tを素性として選択する．もしくは，語tの負例における重み$W_n(t,S_n)$が正例における重み$W_p(t,S_p)$の 2 倍より大きければ，その語tを素性として選択する．すなわち，以下の式 4 の条件のどちらかを満たす語tを素性として選択する．

$$W_p(t,S_p) > 2W_n(t,S_n) \qquad (4\text{-}1)$$

$$W_n(t,S_n) > 2W_p(t,S_p) \qquad (4\text{-}2)$$

上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，正例，負例ともによく出現するような一般的な語を素性から除去する．選択された素性の一部を以下に例示する．

引き上げ，引き下げ，増額，減額，転換，ポジティブ，鈍化，原料，恩恵，苦戦，体質，ポテンシャル，着実，好転，リストラ

2.3　モデル

深層学習のモデルについて以下に述べる．入力は学習データから抽出された 2,474 語を要素，語tにおける$Wp(t,Sp)$，もしくは，$Wn(t,Sn)$の大きいほうを要素値としたベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数と同じとし，隠れ層は，ノード数 1,000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素である．また，エポック数は 50 回，活性化関数として，ReLU を使用した．上記のモデルを図 1 に示す．

酒井らの手法[1]に記載されている通り，レーティングが変動しなかったアナリストレポートに対して極性を付与した際の精度は 75.5%となっている．

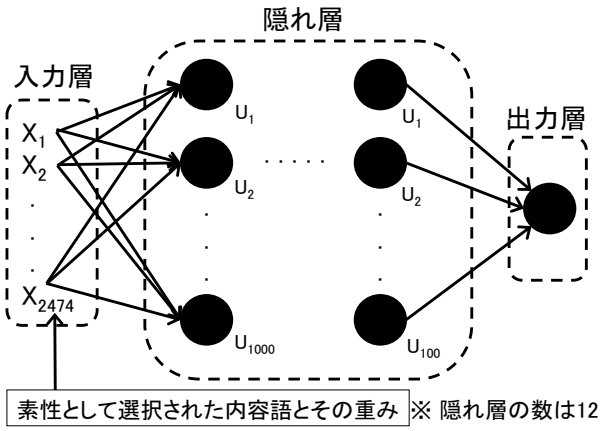


図 1: ニューラルネットワークのモデル

3. 極性と株式リターン

前述のモデルを使用し，2014 年から 2016 年の間に発行されたレポートに対し極性を付与し，株式リターンとの関連性を分析する.

3.1 イベントスタディ分析

3.1.1 レポート自体の市場へのインパクト

レポート発行前後 1 営業日間の累積超過リターン（$CAR_{-1,+1}$）を算出し，付与された極性の違いによってレポートに対する市場の反応に違いが表れるか検証する．異常リターンの算出には，金融データソリューションズが提供する「日本上場株式 Fama-French 関連データ」から取得した，Fama-French の 3 ファクターモデルを使用した．推定期間は，アナリストレポート発表日の 120 日前から 21 日前までの 100 日間とした．レポートの定量情報の変更があった場合は，変更による市場へのインパクトがノイズとなるため，定量情報の変更が無いレポートに絞り込む．また REIT、外国部上場銘柄に言及したレポートは対象外とした．また IPO 直後の銘柄に言及したレポートなど推定期間中のデータが確保できないものは除外した．最終的に$CAR_{-1,+1}$算出対象となったレポート本数は 45,756 本である．レポートに記載されたレーティングと付与された極性毎に$CAR_{-1,+1}$のサンプル数，平均値， t 値を表 1 に纏めた.

全対象レポートをポジティブが付与された群とネガティブが付与された群に分けた場合，平均値$\overline{CAR}_{-1,+1}$はそれぞれ，0.3%程度および-0.3%程度と対称の値が得られ，ポジティブな群とネガティブな群の$\overline{CAR}_{-1,+1}$の差分は 0.6%程度となった．またレポート内にて言及されているレーティングによって分類した場合も，ポジティブな群とネガティブな群の$\overline{CAR}_{-1,+1}$の水準は違えど，その差分は同程度となった．レーティングの取り扱いについては BUY，NEUTRAL，SELL 等の 3 段階評価で付与されることが多いが，一部証券会社では 5 段階評価で付与されている．今回は，5 段階評価で付与されたレポートは，5 段階評価の中間を NEUTRAL とし，それより高い評価のものは BUY ，低い評価のものは SELL として，3 段階評価に置き換えている．またここでは示さないが，レポート発行後 2 営業日以降の異常リターンはほぼ 0%の水準となっており，市場がレポートに対し短期間で反応していることがわかる．これらの結果は，先行研究と整合的な内容である[8][9].

表 1: 分類毎の$CAR_{-1,+1}$の記述統計量

レーティング	極性	サンプル数	平均値	t値
ALL	ALL	45756	0.176	7.228
	ポジティブ	34899	0.329	12.017
	ネガティブ	10857	-0.313	-5.913
	差分	45756	**0.642**	**10.765**
BUY	ALL	21313	0.293	8.521
	ポジティブ	17104	0.393	10.410
	ネガティブ	4209	-0.112	-1.361
	差分	21313	**0.505**	**5.574**
NEUTRAL	ALL	21023	0.092	2.542
	ポジティブ	15408	0.281	6.807
	ネガティブ	5615	-0.426	-5.763
	差分	21023	**0.707**	**8.349**
SELL	ALL	3420	-0.034	-0.324
	ポジティブ	2387	0.177	1.406
	ネガティブ	1033	-0.521	-2.747
	差分	3420	**0.697**	**3.067**

3.1.2 銘柄のパフォーマンス

次に，レポートにて言及された銘柄が，付与された極性の違いによって，レポート発行日以降について株式リターンに違いが表れるか検証する．検証方法は，ファクターによって得られるリターンを正常リターンと定義し，それら以外のリターンを異常リターンと定義して累積超過リターン（CAR）を算出する．先ほどの$CAR_{-1,+1}$と区別して添え字が無いものがこの定義で算出した累積超過リターンとする．3.1.1 での算出時と同様に，レポートの定量情報の変更があった場合は変更による市場へのインパクトがノイズとなるため，定量情報の変更が無いレポートに絞り込む．また REIT、外国部上場銘柄に言及し

たレポートは対象外とした．最終的に算出対象となったレポート本数は 45,889 本である．

CARの算出には，金融データソリューションズが提供する「日本株式資産運用業務支援サービス(NPMServices)」から取得した，ローゼンバーグ型マルチファクターモデル（リスクファクター12 個＋業種ファクター33 個）を使用した．リスクファクターの内訳は，規模，市場感応度，純資産／株価，利益／株価，財務健全性比率(一般)，財務健全性比率(金融)，米国株感応度，売買回転率，変動性，長期リターン，東証 1 部外フラグ，新興市場フラグの 12 個である．この時，異常リターン（AR）は式 5 で与えられ，累積することでCARが算出される．

$$AR_{i,t} = R_{i,t} - \beta_{i,k,t-1} R_{k,t} \qquad (5)$$

ここで，
$AR_{i,t}$: 銘柄iのt時点における異常リターン．
$R_{i,t}$: 銘柄iのt時点における超過リターン．
$R_{k,t}$: t時点におけるkファクターリターン．
$\beta_{i,k,t-1}$: 銘柄iの$t-1$時点におけるkファクターエクスポージャー．時点毎に推定している．
レポート発行日をイベント日$T=0$としその後 250 営業日までを検証対象期間とした．

対象レポート群のうち，極性がポジティブな群，極性がネガティブな群，それぞれの平均値$\overline{CAR}$ は下記の図 2 の通りとなった．

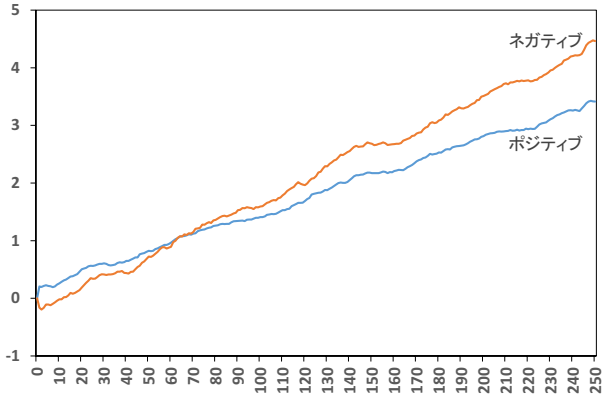


図 2: 累積超過リターンCARの平均値$\overline{CAR}$[%]

ポジティブな群とネガティブな群との間で，レポート発行直後に 0.4%程度の差が広がり，その後$\overline{CAR}$の差は縮まる．60 営業日後程度で差が無くなり，その後は差がマイナス方向に広がっていく．

次にレポートに記載されたレーティング情報によって分類した場合も，付与された極性によって結果に差が出るか確認する．下記の図 3 では，レーティング毎にポジティブな群とネガティブな群の$\overline{CAR}$の差をプロットした．レーティングは 3.1.1 での算出時と同様のルールで 3 段階評価に置き換えている．

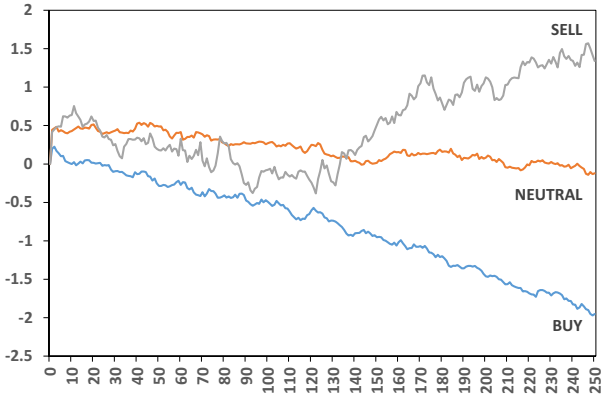


図 3: レーティング毎のポジティブ群の$\overline{CAR}$とネガティブ群の$\overline{CAR}$の差[%]

レーティングが BUY のレポート群において，ポジティブな群とネガティブな群の差は，レポート発行直後に 0.2%程度の差が広がり，その後$\overline{CAR}$の差は 10 営業日程度で無くなる．その後は差がマイナス方向に広がっていき，ネガティブな群の$\overline{CAR}$の方が高い値となる．レーティングが NEUTRAL のレポート群においては，ポジティブな群とネガティブな群の差は，レポート発行直後に 0.45%程度の差が広がり，その後は緩やかに差が縮まっていく．レーティングが SELL のレポート群においては，ポジティブな群とネガティブな群の差は，レポート発行直後に 0.4%程度の差が広がり，その後は一旦差が縮まるが 130 営業日経過後に急速に差が広がっていく．但し，SELL のレポート本数が少ないため統計的に有意な差とは言い切れない．分類毎の$\overline{CAR}$と t 値を表 2 に纏める．

表 2: 分類毎の$\overline{CAR}$と t 値

		CAR (T=60)	CAR (T=120)	CAR (T=180)	CAR (T=240)
BUY	ポジティブ	0.69 (7.57)	1.09 (8.42)	1.51 (9.48)	1.85 (10.07)
	ネガティブ	0.96 (4.89)	1.70 (6.10)	2.71 (7.97)	3.68 (9.33)
	差分	−0.27 (1.25)	−0.60 (1.97)	−1.20 (3.19)	−1.83 (4.21)
NEUTRAL	ポジティブ	1.07 (11.39)	1.99 (15.07)	3.12 (19.47)	4.06 (22.06)
	ネガティブ	0.65 (4.14)	1.74 (7.68)	2.96 (10.89)	4.09 (13.24)
	差分	0.42 (2.26)	0.25 (0.95)	0.16 (0.50)	−0.04 (0.11)
SELL	ポジティブ	2.26 (7.78)	4.05 (9.44)	6.08 (11.64)	8.32 (13.47)
	ネガティブ	1.93 (4.11)	4.29 (6.37)	5.29 (6.65)	6.98 (7.85)
	差分	0.33 (0.59)	−0.24 (0.30)	0.79 (0.83)	1.34 (1.24)

（図表注）各 T 時点における$\overline{CAR}$ [%]を表す．括弧内は t 値．

付与された極性によるリターンの違いについてより深く理解するために，レポート発行前後について250 営業日前から 500 営業日後までの長期のCARを算出する．前述の条件で絞り込んだ 2014 年から 2015 年までのレポート 30,128 本を対象とし，レポート発行 250 営業日前を算出基準日とする．下記の図 4 は，ポジティブな群とネガティブな群の長期の$\overline{CAR}$をプロットした．

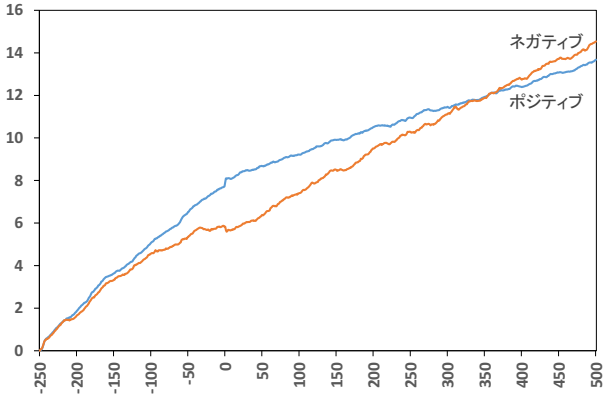


図 4：長期でのポジティブな群と
ネガティブな群の$\overline{CAR}$[%]

レポート発行前 100 営業日程度から 2 群間の$\overline{CAR}$の差が広がり，レポート発行直後頃に差が最大になる．その後は徐々に差が縮まり，レポート発行後 360 営業日程度で差が無くなる．その後はポジティブな群の$\overline{CAR}$とネガティブな群の$\overline{CAR}$が逆転し，ネガティブな群の$\overline{CAR}$の方が大きくなる．

同様に，レーティングと極性によって 6 分類し，長期の$\overline{CAR}$をプロットした図が下記の図 5 である．

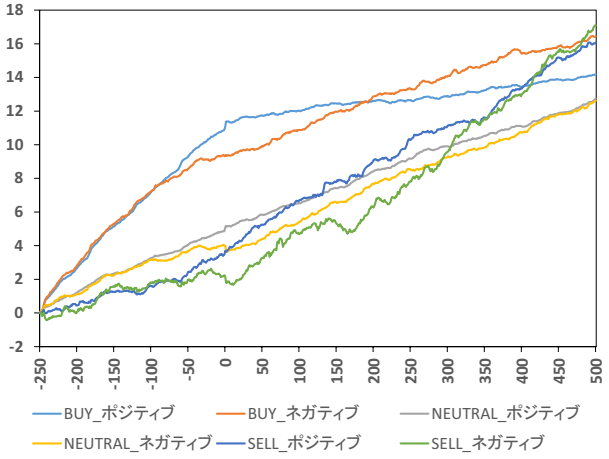


図 5：長期での分類毎の$\overline{CAR}$[%]

3 分類したどのレーティングの場合でも，レポート発行前にポジティブな群とネガティブな群の$\overline{CAR}$の差が広がり，レポート発行後は，徐々に差が縮まっている．またレポート発行前は，BUY のレポート群の$\overline{CAR}$が先行して大きな値をとるが，レポート発行後は他のレーティング群の$\overline{CAR}$の増加分の方が大きくなっている．

3.2　レポート発行後のレーティング変更率

アナリストレポートの極性によって，その後アナリストがレーティングを変更する確率に差が生じるか確認した．表 3 は，各分類において発行されたレポート数とその後 1 年 1 カ月後までに発行されたレポートにおいて，レーティング変更等のアクションの有無を示している．対象としたレポートは 3.1.2 と同様の条件を元に抽出した 45,889 本で，複数回変更があった際には，最初の変更内容を採用している．

表 3：分類毎レポート数とその後の変更状況

ポジティブ	カバー廃止	下方修正	変化無し	上方修正	総計
SELL	40 (1.7%)		1389 (58.2%)	959 (40.2%)	2388
NEUTRAL	303 (2.0%)	1038 (6.7%)	11093 (71.8%)	3009 (19.5%)	15443
BUY	264 (1.5%)	4496 (26.2%)	12418 (72.3%)		17178
総計	607 (1.7%)	5534 (15.8%)	24900 (71.1%)	3968 (11.3%)	35009

ネガティブ	カバー廃止	下方修正	変化無し	上方修正	総計
SELL	24 (2.3%)		615 (59.5%)	394 (38.1%)	1033
NEUTRAL	135 (2.4%)	475 (8.4%)	4014 (71.3%)	1004 (17.8%)	5628
BUY	64 (1.5%)	1171 (27.8%)	2984 (70.7%)		4219
総計	223 (2.0%)	1646 (15.1%)	7613 (70.0%)	1398 (12.8%)	10880

（図表注）レポート本数を表す．括弧内は分類内の比率[%]

若干の差ではあるが，極性がポジティブなレポートはレーティングが上方修正およびカバー継続されやすく、極性がネガティブなレポートはレーティングが下方修正およびカバー廃止されやすい傾向がみてとれる．

3.3　極性を元にしたポートフォリオ分析

前節までの分析は，全て個々のレポートについて分析を行った．ここでは，個々のレポートに付与された極性を元に，ある基準日時点において銘柄をポジティブ，ネガティブに分類し，分類された銘柄群による等金額ポートフォリオを作成して，その後のパフォーマンスを測定する．分類方法は下記の通り．

1. 銘柄毎に基準日の前営業日から 3 ヶ月前までのレポートを収集する．この時，同一証券会社から複数のレポートが発行されていた場合は，最新のレポートのみを対象とする．
2. レーティング変更による市場に対するインパクトを排除するため，レーティングが変更されたレポートを含む銘柄は排除する．
3. 極性がポジティブなレポート本数と極性がネガティブなレポート本数を集計し，ポジティブなレポート本数がネガティブなレポート本数より 2 本以上多いものをポジティブ銘柄，ネガティブなレポート本数がポジティブなレポート本数より 2 本以上多いものをネガティブ銘柄とする．

2014 年 6 月末を基準に各銘柄群の等金額ポートフォリオを作成し，その後は半年毎に銘柄を分類し直し，各ポートフォリオのリバランスを実施する．2014 年 6 月末から 2017 年 6 月末までの各ポートフォリオのパフォーマンスと同期間の TOPIX のパフォーマンスを図 6 で示す．2014 年 6 月末時点を 100 とした．

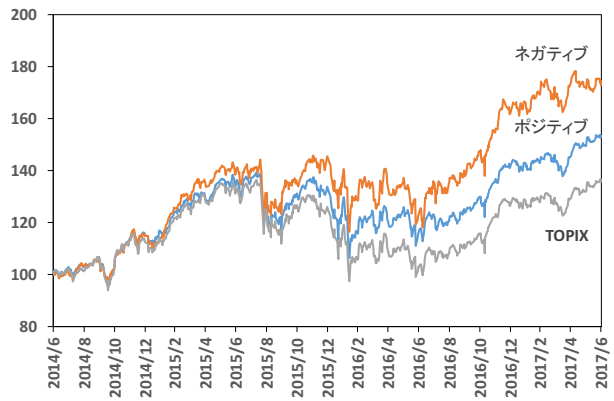


図 6：ポートフォリオ毎のパフォーマンス

期間中の TOPIX のリターンは 35%であったが，ポジティブ銘柄群のポートフォリオは 52%，ネガティブ銘柄群のポートフォリオは 72%となった．このときポジティブ銘柄群のポートフォリオの平均銘柄数は 250 銘柄程度で，ネガティブ銘柄群のポートフォリオの平均銘柄数は 23 銘柄程度であった．

4．考察

前章までの結果から，アナリストレポートに付与された極性の違いによって，言及された銘柄はその後の長期にわたり株式リターンの振る舞い方に違いが出ることが示された．図 4 および図 5 にて示したが，レポート発行前までは，ポジティブを付与された銘柄のパフォーマンスが良く，レポート発行後は，ネガティブを付与された銘柄のパフォーマンスが良くなる．これは，定性情報についてアノマリーが存在するかもしれないことを示していると考えられる．

また得られた極性情報を活用したポートフォリオを構築することで，TOPIX をアウトパフォーム出来る可能性が示唆された．ネガティブ銘柄群のポートフォリオのパフォーマンスが優れているのは，先ほど言及した定性情報についてのアノマリーが存在するならば，その寄与によると考えられる．

但し今回の分析は，2014 年から 2016 年までのレポートとその後の株式リターンを用いたものであり，分析対象とした期間に依存している可能性は否定できない．他の期間を対象に比較検証する必要がある．本文中では示さなかったが，今回分析対象とした期間を複数区切り，期間の違いで結果に差が生じるか確認したところ，統計的に有意な差は生じなかった．少なくとも今回分析対象とした期間中は，極性の違いによって株式リターンに差異が表れるという特徴は安定していると言えるだろう．また 3.1.2 では累積超過リターンによる結果を示したが，対 TOPIX 超過リターンの平均値や Fama-French の 3 ファクターモデルによる回帰分析等で検証しても同様の特徴が表れた．

5．まとめ

本研究では，アナリストレポートに付与した極性によって株式リターンに差が生じるか検証してきた．極性は，アナリストレポート内で記載されているレーティングや目標株価，業績予想などの定量情報にはよらず，本文のテキストデータのみによって算出されるもので，そのレポートの本文表現がポジティブなのかネガティブなのか分類しているとみなせる．極性と株式リターンの関連性は，アナリストレポート本文に記載された定性情報がどの様に市場に影響を与えているか，また定性情報自体に付加価値があるか理解することに繋がる．従来困難であった定性情報の計量化が，近年の自然言語処理技術や機械学習技術の発展により，大量のアナリストレポートに対しても簡便に計量化が出来る様になったことは大変興味深いことだろう．今回はポジティブ，ネガティブの 2 極性に分類したが，他の方法による計量化を実施することで，アナリストレポート本文が市場に与えている影響や，アナリストレポートの定性情報の持つ付加価値について，また別の特徴を捉えることも可能となるだろう．今後の研究課題としていきたい．

参考文献

[1] 小林和正，酒井浩之，坂地泰紀，平松賢士，“アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与”，第 19 回金融情報学研究会，pp.68-73, 2017.
[2] 工藤秀明，永島淳，宮崎義弘 “自然言語処理技術を用いたアナリストレポートの実証分析―センチメントの変化と株式市場反応について―” 証券アナリストジャーナル vol.55, no.9, pp.66-77, 2017.
[3] 酒井浩之，西沢裕子，松並祥吾，坂地泰紀，“企業の決算短信 PDF からの業績要因の抽出”，人工知能学会論文誌, vol.30, no.1, pp.172-182, 2015.
[4] 坂地泰紀，酒井浩之，増山繁，“決算短信 PDF からの原因・結果表現の抽出”，電子情報通信学会論文誌 D, vol.J98-D, no.5, pp.811-822, 2015.
[5] 北森詩織，酒井浩之，坂地泰紀，“決算短信 PDF からの業績予測文の抽出”，電子情報通信学会論文誌 D, vol.J100-D, no.2, pp.150‐161, 2017.
[6] 酒井浩之，松下和暉，“決算短信からの業績要因文の抽出”，第 11 回テキストアナリティクス・シンポジウム, pp.87-91, 2017.
[7] 室野莉沙，酒井浩之，坂地泰紀，ベネット ジェイスン，“決算短信から抽出した原因・結果表現の意外性の判定”，第 11 回テキストアナリティクス・シンポジウム, pp.93-98, 2017.
[8] 太田浩司 "アナリストレポートの実証分析 ―目標株価とレポート内容の分析を中心に―" 証券アナリストジャーナル vol.47, no.11, pp.48-62, 2009.
[9] 近藤江美，太田浩司 "アナリストによる株式推奨と利益予想の情報内容" 証券アナリストジャーナル vol.47, no.11, pp.110-122, 2009.

深層学習を用いた経済テキスト可視化の検証

伊藤友貴 [1*]　坂地泰紀 [1]　和泉潔 [1]
Tomoki Ito[1]　Hiroki Sakaji[1] Kiyoshi Izumi[1]

[1] 東京大学工学系研究科システム創成学専攻
[1] Graduate School of Engineering, The University of Tokyo

Abstract: 経済文書のような専門的な文書は非専門家にとって読みにくい場合が多い．そのため，非専門家を対象に経済文書上のセンチメントを単語単位で可視化するようなサポートシステムを構築することには一定の需要があると思われる．経済文書上のセンチメントを可視化する手段の一つとして近年提案された「ニューラルネットワークモデルの解釈」に関する手法，LRP (Layer-wise Representation Propagation) を用いるという手段がある．しかし現状 LRP が日本語の経済文書の可視化に有用かどうかは調査されておらず，その性質についての詳細な分析もあまりされていない．また，LRP の Attention RNN への適用方法は未だ提案されていない．本報告では LRP の Attention RNN への適用方法を提案し，また，LRP が日本語金融テキストの可視化に有用かどうかを検証する．さらに，実データを用いた検証の中で LRP を用いた日本語文書可視化の性質について分析する．

1　はじめに

経済文書のような専門的な文書は非専門家にとって読みにくい場合が多い．この原因の一つとして，経済文書のような専門的な文書を読み解くためにはドメイン特化の専門的な知識を要する必要がある場合が多々あることが挙げられる．例えば,「動きが悪い」という表現が景気動向に文書に出てきた場合，この表現の意味は「客足が悪い」,「消費が悪い」に近く，一般的な意味とは少し違う意味で使われる．このような問題を解決する策の一つとして，文書内の単語についてセンチメント値を与え，図 1 のように単語単位での可視化するという手がある．このような可視化によって非専門家であっても文書内のセンチメントを単語単位で簡単に把握することができる．これは非専門家が専門文書を読み解く上で大きな助けになると期待できる．本研究ではこのような経済文書の可視化を大規模なコメントとそのポジネガタグからなるデータセットを用いて行う手法について考える．このような問題設定の下で

競合 店 が 1 店舗 少ない なる 現 段階 で は 上向き だ ある

図 1: 非専門家のための経済文書可視化例: ネガテイブ単語を青色，ポジティブ単語を赤色に着色

文書の可視化を行う手法としてロジスティック回帰モデルなどのような線形回帰モデルの重みベクトルを利用する手法，[13]．Attention メカニズム [4, 11, 20] を利用する手法が考えられる．上記に加え，深層学習モデルの解釈を行う手法 [1, 2, 5, 6, 10, 14, 16, 17, 18] も有用であると考えられる．その中でも 特に Layer-wise Representation Propagation(LRP)[2] は LSTM をセルに持つ RNN モデルへも適用できる [1] state-of-the-art な深層学習モデルの解釈を行う手法の一つであり，文書の可視化にも有用だと考えられる．しかし，LRP には現状，以下のような課題がある．

- LRP が日本語の金融文書に適用できるかの検証がほとんどされていない．
- LRP を文書可視化に適用した場合の言語的な観点からの分析がほとんど行われていない．
- RNN Attention モデルは多くのタスクにおいて RNN よりも高い予測性能を出すことが知られているにも関わらず，LRP の RNN Attention モデルへの適用方法が提案されていない．

また，LRP に限定する話ではないが，今後深層学習を利用するケースが増えることが予想されることを踏まえると，深層学習モデルを解釈する手法について様々な観点から分析しておくこと，その汎用性を高めることには一定の研究意義があると考えられる．

そこで，本研究では以下について取り組む．

- LRP の RNN Attention モデルへの適用法を提案する．
- LRP が日本語金融文書の可視化に適用可能かどうかを検証する．
- LRP による文書可視化の性質を調査する．

*Email: m2015titoh@socsim.org

本研究の貢献は以下のようにまとめられる.

- LRP の RNN Attention モデルへの適用法を提案した (第 2 節).
- オリジナルセンチメントと文脈センチメントという文書の可視化を評価するための新しい指標を提案すると共にそれらを評価するためのデータセットを作成した (第 3 節).
- LRP を用いた文書可視化が日本語金融文書に適用可能か検証し，その性質を分析した (第 4 節).

2 LRP

LRP は提案されたニューラルネットワークモデルにおいて入力値が出力値に対して与える影響を計算する手法である [2] . 出力層から入力層にかけて Chain Rule に近い形で影響度を伝搬させることで求める手法である. LSTM[7], GRU[3] を含む RNN について計算するために linear connections と multiplicative connections という二種類の影響計算方法が提案されている [1].

2.1 Lnear connections

ノード z_j が $z_j = \sum_i z_i \cdot w_{ij} + b_j$ (z_i は z_j につながるノード) と 順伝搬時に計算できるとする. ここで, w_{ij} は重みベクトル, b_j はバイアスベクトルである. また, ノード z_j の 出力層に与える影響を R_j とする. このときノード z_i の影響度 R_i を以下のように計算する.

$$R_{i\leftarrow j} = \frac{z_j \cdot w_{ij} + \frac{\epsilon sign(z_j)) + \delta b_j}{N}}{z_j + \epsilon sign(z_j)}, R_i = \sum_j R_{i\leftarrow j}$$

N は z_j につながる下位層のノードの数, ϵ は十分に小さい値である. 先行研究 [1] と同様に $\epsilon = 0.001$, $sign(z_j) := (1_{z_j \geq 0} - 1_{z_j < 0})$, $\delta = 0$ とした.

2.2 Multiplicative connections

上位層のノード z_j が下位層のノード z_g, z_s によって $z_j = z_g \cdot z_s$ と計算される場合について考える. これは LSTM[7], GRU[3] などに見られる積の演算である. ここで, LSTM, GRU において sigmoid 関数によって $[0, 1]$ の値になる方を z_g, ならない方を z_s とする. このとき, $R_g = 0$, $R_s = R_j$ と計算する.

2.3 LRP for Attention RNN (提案手法)

LRP を LSTM cell を持つ Attention RNN に適用することを考える. 先行研究 [1] 通り, 線形結合については linear connections, LSTM における Gate 部分の結合については multiplicative connections を適用する.

さらに, Attention 部分についても multiplicative connections を適用する. この Attention 部分への 適用が本研究の提案部分となる.

3 実データによる可視化検証

本節では LRP が日本語経済文書の可視化に適用できるかどうかについて実データを用いて検証する. まず, LRP による日本語経済文書の可視化が妥当かどうかについて第 3.1 節にて定義するオリジナルセンチメント値, 文脈センチメント値を正しく割り当てられるかどうかという観点から検証する. その後, LRP による可視化の結果をいくつか紹介する.

3.1 単語センチメント

本研究では以下のようにオリジナルセンチメント値, 文脈センチメント値を定義する.

- オリジナルセンチメント値: 単語本来のセンチメント値.
- 文脈センチメント値: 文脈における反転情報考慮後のセンチメント値.

例として,「売上が伸びない」という文における「伸びる」という単語について考える. オリジナルセンチメント値をこの単語に与える場合はプラスの値がつくことが正しい. 一方, 文脈センチメント値をこの単語に与える場合には「伸びる」が「ない」によって反転を受けていることを考えてマイナスの値がつくことが正しい.

3.2 テキストコーパス

本実験においては内閣府から提供されている日経景気ウォッチャーデータ [1] を実テキストデータとして用いた. 期間としては 2002 年 1 月から 2017 年 4 月までのテキストデータを用いた. これらのデータはコメントとそのセンチメントタグからなるデータセットであり, センチメントタグの種類は {1 (悪い), 2 (やや悪い), 3 (変わらない), 4 (やや良い), 5 (良い) } の 5 つのタグがついている. 本研究では「悪い」,「やや悪い」

[1] http://www5.cao.go.jp/keizai3/watcherindex.html

のものをネガティブタグ,「良い」,「やや良い」のものをポジティグタグとして扱い，センチメント予測モデル構築に利用した．また，このコーパスの形態素解析には MeCab[9] を用いた．

3.3 モデル構築

各可視化手法の検証にあたって Logistic Regression model (LR), RNN model with LSTM cells (RNN)[7], Bidirectional RNN model with LSTM cells (BiRNN)[15], RNN Attention Model with LSTM cells (AttRNN)[11] の 4 つのセンチメントタグ予測モデルを構築した．このとき，3.2 節で紹介したデータのうちポジティブコメント及びネガティブコメント 20000 件ずつ (計 40000 件) を訓練データとして, ポジティブコメント及びネガティブコメント 2000 件ずつ (計 4000 件) をハイパーパラメータの探索及び学習の早期終了を行うための検証データとして用いた．

学習後にポジティブコメント及びネガティブコメント 4000 件ずつ (計 8000 件) からなるテストデータについてそのポジネガ予測力を検証した．LR, RNN, BIRNN, AttRNN それぞれの Macro F_1 score の結果はそれぞれ 0.878, 0.920, 0.921, 0.923 であった．

訓練データ，検証データ，テストデータの間に被りはない．その他の実験設定は以下の通りである．RNN, BiRNN, AttRNN における各隠れ層の次元数は 200 とし，埋め込みベクトルには 3.2 節で紹介したテキストコーパスをもとに skip-gram model (window size = 5, negative sampling を使用)[12] によって学習したものを使用した．また，RNN モデル，AttRNN モデルは共に埋め込み層 1 層，LSTM cell を含む 逆方向 RNN の層 1 層，線形結合層 1 層からなるモデルであり，BiRNN モデルは埋め込み層 1 層，LSTM cell を含む 両方向 RNN の層 1 層，線形結合層 1 層からなるモデルであった．AttRNN における Attention には dot 関数による global attention [11] を用いた．また，各 RNN モデルは Dropout 法 [19] (dropout rate = 0.5)，Adam Optimizer[8] を用い， 最大 epoch 数 = 50 という条件のもとでの学習を行った．

3.4 評価用データセット

可視化手法の評価のために評価用データセットを構築した．まず，テストデータからポジティブコメント 500 件，ネガティブコメント 500 件を抽出した．その後，人手で各コメント内の重要単語についてオリジナルセンチメント値が正か負かのタグ (オリジナルセンチメントタグ) と文脈センチメント値が正か負かどうかのタグ（文脈センチメントタグ）を付与した．オリジナルセンチメントタグについてはポジティブタグ 1,794 件，ネガティブタグ 1,062 件が付与され，文脈センチメントタグについてはポジティブタグ 1,340 件，ネガティブタグ 1516 件が付与された．

3.5 評価基準

次の 2 指標について macro F_1 値をもとに評価した．

オリジナルセンチメント: 各可視化手法によって各単語に付与するオリジナルセンチメント値の正負が人手でつけたオリジナルセンチメントタグの正負に一致する度合い．

文脈センチメント: 各可視化手法によって各単語に付与する文脈センチメント値の正負が人手でつけた文脈センチメントタグの正負に一致する度合い．

3.6 比較手法

オリジナルセンチメント，文脈センチメントの評価を以下の手法を用いて行い，結果を比較し，各手法の性質を調査した．

LR: LR Model の重みベクトルを用いて各単語へセンチメント値を付与する手法.

LRP with BiRNN: BiRNN に LRP を適用することで各単語へセンチメント値を付与する手法.

LRP with RNN: RNN に LRP を適用することで各単語へセンチメント値を付与する手法.

LRP with Attention RNN: AttRNN に LRP を適用することで各単語へセンチメント値を付与する手法．Attention 部分については 2.3 節で提案した方式に沿って LRP を適用する.

Gradient with RNN: RNN に Gradient 法 [5, 6] を適用することで各単語へセンチメント値を付与する手法.

Gradient with BiRNN: BiRNN に Gradient 法 を適用することで各単語へセンチメント値を付与する手法.

Gradient with Attention RNN: AttRNN に Gradient 法 を適用することで各単語へセンチメント値を付与する手法.

Attention RNN: Attention 層の計算に使われる dot 関数の演算結果をもとに各単語へセンチメント値を付与する手法.

4 検証結果

4.1 LRP の有用性

オリジナルセンチメント値，文脈センチメントスコアの結果は表 1 の通りである．これらの結果より RNN, BiRNN，AttRNN を用いた場合のどの場合においても LRP が Gradient 法 よりも正しくオリジナルセンチメントスコア，文脈センチメントスコア共に正しく付与できていることが確認でき，その有用性を確認できた．

4.2 Attention RNN への LRP 適用

今回提案した AttRNN への LRP への適用手法，LRP with Attention RNN が AttRNN を用いた他の可視化場合に比べ．オリジナルセンチメント，文脈センチメントの両面にて高性能で可視化でき，その妥当性を検証できた．また，本実験では LRP with Attention RNN はオリジナルセンチメントよりも文脈センチメントについてそのセンチメント値を正しく与えていた．

表 1: オリジナルセンチメント値，文脈センチメント値の付与結果 (macro F_1 スコア)

Method	オリジナル	文脈
LR	**0.910**	0.793
Grad with RNN	0.590	0.632
LRP with RNN	0.834	**0.816**
Grad with BiRNN	0.708	0.738
LRP with BiRNN	0.867	0.805
Attention RNN	0.633	0.713
Grad with Attention RNN	0.281	0.356
LRP with Attention RNN	0.680	**0.815**

4.3 各可視化手法の性質分析・考察

各可視化手法の性質を分析するためにセンチメントについて反転がある場合とない場合それぞれの場合における文脈センチメントの付与結果を見てみた（表 2）．本結果より，各手法を以下の四グループに大別できた．

- 反転がある場合にもできない場合にもそこそこ対応できるもの: LRP with RNN Attention.
- 反転がある場合は高性能はだが反転がない場合には低性能のもの: Grad with BiRNN, LRP with RNN Attention, Attention RNN.

表 2: 反転がある場合・ない場合における文脈センチメント付与性能検証結果 (macro F_1 スコア)

Method	反転あり	反転なし
LR	0.166	**0.938**
Grad with RNN	0.473	0.640
LRP with RNN	0.340	0.905
Grad with BiRNN	0.422	0.778
LRP with BiRNN	0.268	0.919
Attention RNN	0.537	0.715
Grad with Attention RNN	0.474	0.316
LRP with Attention RNN	**0.667**	0.809

- 反転がない場合は高性能はだが反転には対応ができないもの: LR, LRP with RNN, LRP with BiRNN.
- 反転があるなしに関わらず低性能のもの: Grad with RNN, Grad with BiRNN.

ただ，このグループ分けが他のデータセットでも同様かどうかは現段階では不明であり，これは検証すべき事項である．

各手法の特徴を可視化例から説明する．図 2 はセンチメント反転が起こっている場合の可視化例である．この例において「伸びる」は「ない」によってセンチメントが反転しているので「青く」色がつくのが正しい．この反転に LR, LRP with RNN は対応できていないが，Grad with RNN, Attention RNN, LRP with Attention RNN は対応できている．図 3 はセンチメント反転が起こっていない場合の可視化例である．この例において「少ない」に青く色がつく (ネガティブ) のが正しい．LR, LRP with RNN, LRP with Attention RNN は正しく色がついているが，Grad with RNN, Attention RNN は正しく色をつけることができていない．Attention RNN では全体的に赤く色がついており，コメント全体の極性がポジティブであることに引きづられて 失敗しているように見える．

5 結論

本研究では LRP が日本語金融文書の可視化に適用可能であることを実データを用いて検証し，LRP による文書可視化の性質を調査した．さらに深層学習の解釈の可視化手法 LRP の RNN Attention モデルへの適用法を提案した．実データを用いて今回提案した LRP の RNN Attention モデルへの適用法が他の RNN

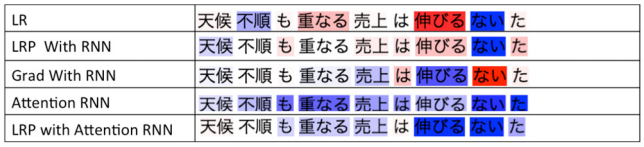


図 2: 反転が起こっている場合の可視化例

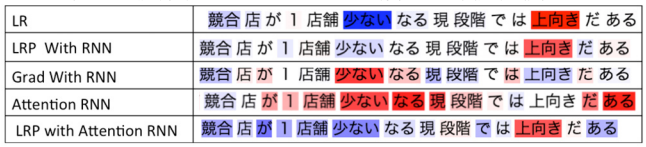


図 3: 反転が起こっていない場合の可視化例

Attention を解釈する手法に比べ，今回利用した日本語金融文書を可視化する上では有用であることを示した．

今後の課題として他のテキストデータによる解析も行い，LRP による可視化の性質について一般化すること，センチメントの反転があるなしに関わらず正しく文脈センチメントを割り振れるような可視化手法の構築，及びオリジナルセンチメント・文脈センチメントどちらにも柔軟に対応可能な可視化手法の構築が考えられる．

謝辞

本研究の一部は JSPS 科研費 JP17J04768 の助成を受けたものである．

参考文献

[1] L. Arras, G. Montavon, K. R. Muller, and W. Samek.: Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP Workshop* (2017)

[2] S. Bach, A. Binder, G. Montavon,F. Klauschen, K. R. Muller and W. Samek.: On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, Vol. 10, No. 7, 1–46 (2015)

[3] J. Chung, C. Gulcehre, K. Cho, Y. Bengio.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS Workshop* (2014)

[4] Y. Dong, H. Su, J. Zhu and B. Zhang.: Improving Interpretability of Deep Neural Networks with Semantic Information. In *CVPR* (2017)

[5] S. Karen, Ve. Andrea and A. Zisserman.: Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034* (2013)

[6] Y. Hechtlinger.: Interpretation of prediction models using the input gradient. *arXiv:1611.07634* (2016)

[7] S. Hochreiter and Jurgen Schmidhuber.: Long short-term memory. *Neural computation*, Vol. 9, No. 8 , 1735–1780 (1997)

[8] D. P. Kingma, J. L. Ba.: ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. *arXiv:1412.6980* (2014)

[9] T. Kudo, K. Yamamoto, Y. Matsumoto.: Applying Conditional Random Fields to Japanese Morphological Analysis. In *EMNLP*. 230–237.

[10] J. Li, D. W. Monroe, D. Jurafsky.: Understanding Neural Networks through Representation Erasure. *arXiv:1612.08220* (2016)

[11] M. Luong, H. Pham and C. D. Manning.: In *EMNLP*. Effective Approaches to Attention-based Neural Machine Translation 1412–1421 (2015)

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean.: Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 3111–3119 (2013)

[13] K. Ravi, V. Ravi.: 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*. **89**(C), 14–46 (2015)

[14] M. T. Ribeiro, S. Singh, C. Guestrin.: 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *KDD*

[15] M. Schuster and K. K. Paliwal.: Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, 2673–2681 (1997)

[16] A. Shrikumar, P. Greenside and A. Kundaje.: Learning Important Features Through Propagating Activation Differences. In *ICML* (2017)

[17] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller.: Striving for simplicity: The all convolutional net. In *ICLR Workshop* (2015)

[18] M. Sundararajan, A. Taly, Q. Yan.: Axiomatic Attribution for Deep Networks. In *ICML* (2017)

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, Vol. 15, No. 1, 1929–1958 (2014)

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio.: Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 77–81 (2015)

金融レポート、およびマクロ経済指数によるリアルタイム日銀センチメントの予測

Real time sentiment analysis of Bank of Japan using text of Financial report and macroeconomic index

余野　京登[1]　和泉　潔[1]　坂地　泰紀[1]

Kyoto Yono[1], Kiyoshi Izumi[1], Hiroki Sakaji[1]

[1]東京大学大学院

[1]Graduate School of Engineering,The University of Tokyo

１　はじめに

金融市場におけるテキストデータは、投資家にとって分析対象の一つであり、その重要性は日に日に増している。決算短信をはじめとする企業の業績について書かれたドキュメント、証券会社のアナリストが書いた個別企業についてのアナリストレポート、更にはツイッター等のSNSで発信している個人投資家のつぶやき等、多種多様なテキストデータが存在する。定性的な投資判断の材料になるとともに、これらのテキストを用いて、モデルを構築し、定量的な分析を行うこともある。

本研究では、中央銀行が発行する議事録やステートメントなどの公的な文章を対象に、定量的な数値化を目的としている。具体的には、日本銀行の発行する金融政策決定会合の議事要旨から物価、生産、雇用等の各トピックに対するセンチメント指数の構築を目指す。

1.1 センチメント付与に関する問題意識

公的な文章からセンチメントを導き出す研究は、各国の中央銀行を対象に行われてきた。山本らは、日本銀行のテキストを対象に、ニューラルネットワークを用いた深層学習でスコア付を行い、指数化を行った[1]。また、D. Wu らは、FOMCのテキストを対象に、LDAによりトピックわけを行い、各センテンスに関しては、極生語辞書を用いたスコア付を行った[2]。これらの研究においては、まず一文単位でスコアを付与した後、１ドキュメントに対しては、各文の平均スコアをその時点のセンチメントとしている。その際の課題点を３つ上げる。１点目は、あるドキュメントにおいて、文章数が少なく、かつ、一つの文のスコアが極値を取った場合、ドキュメント全体がその特定のスコアに大きく影響されてしまう。２点目は、ドキュメントには時系列性を有しているはずだが、それを特に考慮していない。３点目としては、モデルにより数値化しているが、それはドキュメント単体のみを評価しているに過ぎず、中央銀行の実際に行っている「マクロ経済指標等の各データを観察→各議員が議論→その結論として、ドキュメントが発行される」という一連の構造を捉えていない。

本研究では、これらの問題意識を踏まえ、マクロ経済指標も用いた生成モデルを用いて中央銀行のトピック別のセンチメント指数の構築を試みる。

２　モデル構築

本章では、実際の日本銀行における金融政策運営の概略について説明した後、その枠組ベースにした構築したモデルの概要について述べる。

2.1 日本銀行の金融政策運営

日本銀行の公式サイト[3]によると、金融政策運営は以下の通り、行われている。

金融政策運営の基本方針は、日本銀行政策委員会の「金融政策決定会合」とよばれる会合で決定します。会合では、金融経済情勢に関する検討を行うとともに、金融市場調節方針や当面の金融政策の運営方針を決定し、決定した内容は直ちに公表しています。（中略）金融政策決定会合では、年に8回、2日間かけて集中的に審議を行い、金融政策の方針を決定しています。議決は9名の政策委員（総裁、2名の副総裁、6名の審議委員）による多数決によって行います。

つまり、金融経済情勢に関して、各９名の審議委員が議論を行う。そして、その結果として議論の内容が議事要旨として公表される。

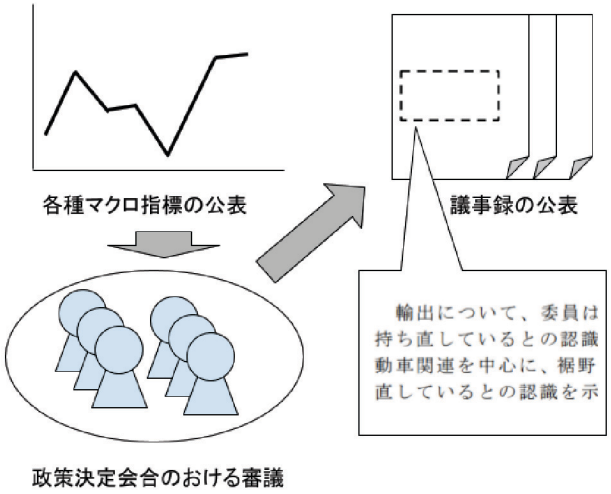


図１: 日本銀行における金融政策運営の概観

2.2 分析対象テキスト

日本銀行のトピック別のセンチメントを構築するために、対象とするテキストは金融政策決定会合の議事要旨とする。さらに、議事要旨内の「金融経済情勢に関する委員会の検討の概要」セクションのみを対象とする。対象のテキストの特徴としては、１点目としては、トピックについての記述が「～～について、」というように明示的に文頭に書かれる。２点目としては、「委員は、～～を共有した」や「一人の委員は、～～と述べた」というようにどの程度の人数の委員によって議論や意見がなされたかが文章内に必ず記述される形となっている。

わが国の景気について、委員は、所得から支出への前向きの循環メカニズムが働くもとで、緩やかに拡大しているとの見方で一致した。委員は、企業部門の動きについて、輸出は増加基調にあるほか、設備投資も、収益が過去最高水準を更新する中、緩やかな増加基調にあるとの認識を共有した。また、家計部門についても、委員は、個人消費

図２: 議事要旨の一例

2.3 スコア付

先行研究[1]で用いられた景気ウォッチャー調査を学習データに LSTMで構築したモデルでのセンチメントを各文に対して付与を行う。また、各文において、委員の意見の一致度により、そのスコアの重み付けを行う。

2.4 生成モデル

実際の日本銀行の金融政策運営を考慮し、以下のマルコフ性を考慮した生成モデルを構築する。

まず、各文のスコアSはその時点における日本銀行のセンチメントを示す潜在変数Zに基づき生成させる。その潜在変数Zはマクロ指数Mに依存しており、かつ１時点前の自身にも依存する時系列性を有している。

より詳細なモデルについては、研究会当日にて発表する予定である。

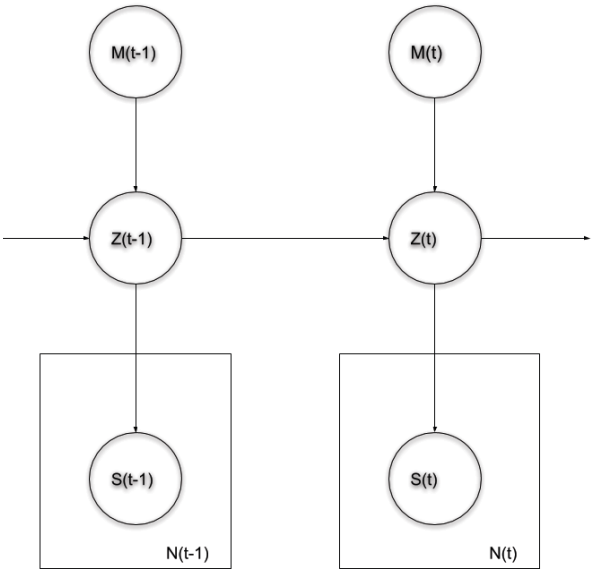


図３: 本モデルにおけるグラフィカルモデル

３　結果、およびまとめ

前章で述べたモデルを用いたシミュレーション結果、および考察を研究会当日にて発表する予定である。

参考文献

[1] 山本 裕樹,松尾 豊:景気ウォッチャー調査の深層学習を用いた金融レポートの指数化,第30回人工知能学会全国大会,(2016)

[2] Jegadeesh, N., and D. Wu: Deciphering Fedspeak:The Information Content of FOMC Meetings,AFA 2016 San Francisco Meeting Paper（2015）

[3] 金融政策決定会合の運営 ：日本銀行 Bank of Japan, www.boj.or.jp/mopo/mpmsche_minu

単語の類義性・対義性を考慮したドメイン特化極性辞書構築

Domain-specific dictionary construction method considering synonym and antonym

伊藤　諒[1*]　坂地　泰紀[1]　和泉　潔[1]　須田　真太郎[2†]
Ryo Ito[1]　Sakaji Hiroki[1]　Kiyoshi Izumi[1]　Shintaro Suda[2]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo
[2] 株式会社 三菱 UFJ トラスト投資工学研究所
[2] Mitsubishi UFJ Trust Investment Technology Institute Co.,Ltd.

Abstract: In recent years, textual information, which is unstructured data attracts attention as new analytical data in the financial and economic fields and it is expected to structure knowledge on this domain. One such knowledge is a sentiment polarity dictionary in which each word is representing positive or negative. In building the dictionary, it is costly to add the polarity value to a vast number of words manually. Therefore, in this research, we propose a the dictionary construction model especially considering the synonymity and symmetry of words. As a result of the experiment, the proposed method is a more accurate than the model of the previous research. In addition, we extended the conventional dictionary using the proposed method, and we showed that the extended dictionary has higher accuracy than the dictionary which is not extended.

1　はじめに

自然言語処理の根幹を支える資源として，語彙資源が存在し，これまでに語彙資源の構築に関する多くの研究がなされてきた [1]．語彙資源を構成するものとして，単語と極性が組みとなった極性辞書が存在し，このような極性辞書は，語彙ベースにおけるセンチメント分析を行う際に，不可欠なものである．

ここにおいて，極性辞書の構築を考えた際，膨大な数の単語に対して人手で極性値を付与していくことは，コストの観点から現実的ではない．また，単語の持つ極性はその単語が出現する背景・文脈によって異なり，解析対象となるテキストに適した極性辞書が必要である．

そこで本研究では，解析対象となるドメインに特化した，センチメント分析のための極性辞書を自動構築することを目的とし，とりわけ対象ドメインの知識，既存の知識ベースに含まれる単語の類義・対義性に関する知識を用いた，半教師あり学習による，ドメイン特化型極性辞書自動構築手法を提案する．また，提案手法のモデルを，既存の極性辞書に対する辞書構築精度の観点から評価を行う．さらに，提案手法を有用性を評価するために，対象ドメインを金融政策ドメインとし，本ドメインに対して人手で構築された辞書を拡張した場合に，センチメント分析の精度が向上するかという観点から評価を行う．

*連絡先：東京大学大学院工学系研究科システム創成学専攻和泉研究室，〒 113-8654 東京都文京区本郷 7-3-1，E-mail: m2016rito@socsim.org

†留意事項：本稿の内容は筆者が所属する組織を代表するものではなく, すべて個人的な見解である．また，当然のことながら，本稿における誤りは全て筆者の責に帰するものである．

2　関連研究

極性辞書構築手法に関する多くの研究がなされているが，コーパスベースのアプローチと，シソーラスベースのアプローチに大別される．コーパスベースのアプローチでは，単語の共起情報や文脈情報を用いて極性語を取得する方法が代表的である．シソーラスベースのアプローチとしては，シソーラスから語彙ネットワークを構築し，その語彙ネットワーク上に種表現を元にして極性を伝搬させる事で，全ての単語に対して，極性を付与する方法が代表的である．

さらに，コーパスベースのアプローチと，シソーラスベースのアプローチを統合した研究として，Alhothai and Hoey (2017) の研究がある [2]．Alhothai and Hoey

(2017) は，まず Skip-gram モデルまたは Glove モデルによって単語分散表現を学習する事で，単語分散表現から k-近傍グラフを構築し，次に得られた k-近傍グラフとシソーラスから構築された類義語ネットワークを合わせた上で，ネットワークに対してラベル拡散法を行う事で極性語を獲得する，SNWELP モデルを提案している．

しかしながら，Alhothai and Hoey (2017) の先行研究において，類義語に関する知識は用いられているが，対義語に関する知識は用いられていない．一般に，ある単語が極性語である際に，その単語に対する対義語は，元の単語とは反対の極性を有する場合が多いが，対義語に関する知識は，極性語獲得タスクにおいて重要な情報を含むため，単語間の類義性のみならず対義性も考慮する事で，より辞書構築精度が向上すると考える．

以上を踏まえ，対象コーパスに含まれるドメインの知識，既存の知識ベースに含まれる単語の類義・対義性に関する知識を用いた，半教師あり学習による，ドメイン特化型極性辞書自動構築手法を提案する．

3 ドメイン特化型極性辞書自動構築手法の提案

本章では，SMLS モデルと DLS モデルという，二つのドメイン特化型極性辞書自動構築手法の提案モデルについて述べる．

3.1 SMLS モデル

はじめに，SMLS モデルを提案する．SMLS モデルでは，はじめに対象コーパスのテキストに対して文分割をし，文分割されたセンテンスに対して形態素解析を行う．次に，単語分割された各センテンスを元に，Mikolov et al. (2013) による Skip-gram モデルを用いて単語分散表現を学習する [3]．そして，各単語に対して，類似度上位の単語 k 個をエッジで結んだ k-近傍グラフを構築する．ここで，単語 c と単語 d の分散表現をそれぞれ $\overrightarrow{u_c}$，$\overrightarrow{u_d}$ とした時，類似度をコサイン類似度とし，エッジの重みとしてコサイン類似度の値を付与する．

次に，得られた k-近傍グラフを元に，単語間がエッジで結ばれていれば，要素として，そのエッジ重みを，結ばれていなければ 0 を与えた，隣接行列 $\mathbf{M}$ を作成する．また，シソーラスから単語間の類義・対義関係を抽出し，単語間に類義関係もしくは対義関係が存在していればエッジで結んだ，類義語グラフ・対義語グラフをそれぞれ作成する．そして，得られた類義語グラフ・対義語グラフを元に，単語間がエッジで結ばれていれば 1 を，結ばれていなければ 0 を要素として格納した，隣接行列 $\mathbf{S}$，$\mathbf{A}$ をそれぞれ作成する．

ここで，分散表現から構築された隣接行列 $\mathbf{M}$ と，類義語・対義語グラフから構築された隣接行列 $\mathbf{S}$，$\mathbf{A}$ を結合した行列 $\mathbf{E}$ を作成する．なお，隣接行列 $\mathbf{E}$ の各要素 $E_{i,j}$ は，隣接行列 $\mathbf{M}$，$\mathbf{S}$，$\mathbf{A}$ の各要素を平均化した値とする．

$$E_{i,j} = \frac{M_{i,j} + S_{i,j} - A_{i,j}}{3}$$

次に，得られた行列 E に対して，シードを付与した上で，Zhou et al., (2004) によって提案されたラベル拡散法の手続きに基づき，ノードのラベル推定を行う [4]．

さて，V を行列 $\mathbf{E}$ の行数とした時，$\mathbf{p}$ を $\mathbf{p} \in \mathbb{R}^{|\mathbf{V}|}$ を満たす，単語の極性値ベクトルとする．ここで，極性値ベクトル $\mathbf{p}$ の要素は，$\frac{1}{|V|}$ で初期化されている．次に $\mathbf{D}$ を行列 $\mathbf{E}$ の次数行列とし，次数行列の各成分に絶対値をとった行列を $\mathbf{D}'$ とした際，以下の式に基づいて行列 $\mathbf{T}$ を計算する．

$$\mathbf{T} = \mathbf{D}'^{\frac{1}{2}}\mathbf{E}\mathbf{D}'^{\frac{1}{2}}$$

そして，得られた行列 $\mathbf{T}$ を用いて，以下の式を反復的に計算することで，ラベル拡散法を行う．

$$\mathbf{p}^{(t+1)} = \beta\mathbf{T}\mathbf{p}^{(t)} + (1-\beta)\mathbf{s}$$

ここで $\mathbf{s}$ は，$\mathbf{s} \in \mathbb{R}^{|\mathbf{V}|}$ を満たすベクトルであり，シードとして付与された単語に対応するベクトルの要素は $\frac{1}{|S|}$ を，シードとして付与されていない単語に対応するベクトルの要素は 0 を，要素として与えられたベクトルである．また，β は推定ラベルの，局所整合性・大域整合性を調整するパラメーターである．

そして，単語 w_i の推定極性値を得るために，ポジティブ単語．ネガティブ単語のシードセットを用いて，各々ラベル拡散法によって単語の極性値推定を行い，推定極性値 $\mathbf{P}^P(w_i)$ と $\mathbf{P}^N(w_i)$ を，それぞれ得る．さらに，得られた推定極性値を用い，以下の式によって単語 w_i の調整極性値 $\overline{\mathbf{P}}^P(w_i)$ を求める．

$$\overline{\mathbf{P}}^P(w_i) = \frac{\mathbf{P}^P(w_i)}{\mathbf{P}^P(w_i) + \mathbf{P}^N(w_i)}$$

最後に，得られた調整極性値 $\overline{\mathbf{P}}^P(w_i)$ を，各単語に対して，平均 0，分散 1 に標準化する．

3.2 DLS モデル

次に，DLS モデルを提案する．DLS モデルでは分散表現学習時に，コーパスに含まれるドメイン情報に加えて，類義・対義語関係の情報を分散表現として埋め込み，得られた分散表現を元に k-近傍グラフを作成し，ラベル拡散法によって単語の推定極性値を得る．

DLS モデルでは，はじめに K. A. Nguyen et al.(2016) によって提案された，d-LCE 法を用いて，コーパスにおける単語のドメイン情報と，単語の類義性・対義性に関する情報を埋め込んだ分散表現を学習する [5]．d-LCE 法は，Skip-gram モデルの目的関数に，単語の類義性・対義性に関する制約項を加えたモデルであり，d-LCE 法の目的関数は以下である．

$$\sum_{w\in V}\sum_{c\in V}\{(\#(w,c)\log\sigma(sim(w,c)) +k\#(w)P_0(c)\log\sigma(-sim(w,c))) +(\frac{1}{\#(w,u)}\sum_{u\in W(c)\cap S(w)} sim(w,u) -\frac{1}{\#(w,v)}\sum_{v\in W(c)\cap A(w)} sim(w,v))\}$$

ここで，V はコーパスに含まれる単語集合，$\#(w,c)$ は単語 w と単語 w に対するコンテキスト c との共起回数，k はネガティブサンプリングにおけるパラメーター値，P_0 はユニグラム分布，$sim(w_1,w_2)$ は単語 w_1 と w_2 のベクトル間のコサイン類似度，$W(c)$ はコンテキスト c に対する LMI 値が正の単語集合，$S(w)$・$A(w)$ は単語 w に対してシソーラスから抽出した類義語・対義語集合を表す．

次に d-LCE 法によって得られた分散表現を元に，SMLS モデルと同様に，各単語についてコサイン類似度上位 k 個の単語をエッジで結んだ k-近傍グラフを構築し，得られた k-近傍グラフを対象として，シードとして与えた単語の極性ラベルをラベル拡散法によって拡散する．

4 実験

本章では，提案手法の有効性を検証するための，各種実験設定と評価方法について述べる．実験は二通りの実験を行い，はじめに極性辞書構築実験を，次に極性辞書拡張実験を行った．

4.1 極性辞書構築実験

提案手法の優位性を検証するために，提案手法である SMLS モデル・DLS モデル，そして先行研究の手法である SNWELP モデルを用いて辞書構築を行い，辞書構築精度を比較した．ここでは，コーパスとして米国の株式公開企業において提出される Form 8-K を，シソーラスとして Wordnet を用い，これらを元に各種辞書構築手法を用いて辞書の自動構築を行い，ファイナンス分野のセンチメント分析において多く用いられている，Loughran and McDonald (2011) の辞書（以下 LM 辞書）中に含まれる各単語の極性の方向性を正解として，評価指標を AUC として評価した [6]．

各モデルにおいて，分散表現の次元数は 300，ネガティブサンプリングの負例数を 15，k-近傍グラフにおける k の値として 10 を用いた．また，コーパスとして用いる Form 8-K は，Lee et al. (2014) によって公開されているデータ[1]を用い，前処理の結果 20,198,170 センテンスを抽出し，各種モデルの入力データとした [7]．さらに単語の類義・対義関係を記述したシソーラスとして，Wordnet を用い，ラベル拡散法における反復処理の回数は 10,000，パラメーター β の値は 0.99 とした．また，ラベル拡散法におけるシード単語として，以下の表 1 に含まれるシード単語を用いた．

表 1: ラベル拡散法において用いたシード単語

Positive 単語	Negative 単語
successful, excellent, profit, beneficial, improving, improved, success, gains, positive	negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative

4.2 極性辞書拡張実験

提案手法の有用性を検証するために，金融政策ドメインを対象として人手で作成された辞書を対象に，提案手法である SMLS モデルを用いて辞書拡張を行い，辞書拡張の有無によって辞書ベースのセンチメント分析の精度が向上するかという観点から精度評価を行った．

コーパスとして，米国の金融政策を策定する委員会である Federal Open Market Committee（FOMC）[2]によって公開される議事録レビュー部分を用い，1993 年 1 月から 2016 年 12 月の間に公表された 192 件を対象とした．また辞書構築におけるシード単語として，人手によって作成された金融政策専門辞書を用いて，SMLS モデルによる辞書拡張を行った．この際，推定極性値が上位・下位 30 位以内に含まれ，かつ既存辞書に含まれず，名詞・動詞・形容詞の品詞に該当する単語を抽出し，既存辞書に追加をする事で辞書拡張を行った．また，SMLS モデルにおいて，分散表現の次元数は 300，ネガティブサンプリングの負例数を 15，k-近傍グラフにおける k の値として 10 を用いた．さらに，単語の類義・対義関係を記述したシソーラスとして，Wordnet を用い，ラベル拡散法における反復処理の回数は 10,000，パラメーター β の値は 0.99 とした．

次に，拡張された辞書を用いて伊藤ら (2017) の手法によって，経済成長・消費・生産・雇用・金融政策・金

[1] https://nlp.stanford.edu/pubs/stock-event.html
[2] https://www.federalreserve.gov

融市場・インフレ・貿易の 8 つのトピックに対する，トピック別センチメントの抽出を行った [8]．なお，伊藤ら (2017) のトピック別センチメントの抽出法は，トピック分類と，単語間の係り受け関係を考慮した辞書ベースのセンチメント分析を組み合わせた手法であり，文書を入力として，文書に含まれる各トピックに対するセンチメントスコアを算出する手法である．

ここで，FOMC 議事録におけるレビュー部分は，経済環境や金融市場の振り返りを行うセクションであるため，得られたトピック別センチメントが正確なものであれば，各トピックに対応するマクロ指標の実測値を，よく説明できるものと考えられる．そこで，辞書拡張の評価として，辞書拡張を行った場合と行わなかった場合とで，それぞれ算出されたトピック別センチメントの，各マクロ指標の実測値に対する説明力を以って評価を行った．説明力の算出においては，各トピックのセンチメントを説明変数，対応するマクロ指標を被説明変数として単回帰分析を行い，決定係数 R^2 による評価を行った．なお，レビューにおける各トピックのセンチメントを評価するマクロ指標は，以下の表 2 の対応となっている．

表 2: 各トピックに対応する検証用マクロ指標

トピック	マクロ変数
インフレ	インフレ率
雇用	非農業部門雇用者数
貿易	経常収支
消費	個人消費支出（PCE）
生産	鉱工業生産指数
経済成長	実質 GDP

5 結果と考察

本章では，各種実験の結果と考察について述べる．

5.1 極性辞書構築実験

表 3 は各モデルにおける，AUC の値を比較したものである．実験結果として，AUC の高い順に SMLS モデル，SNWELP モデル，DLS モデルとなり，提案手法である SMLS モデルが最も高い精度となった．

表 3: 各モデルにおける AUC の比較

SNWELP	DLS	SMLS
0.9096	0.8441	0.9190

これらの結果の理由として，まず SMLS モデルは SNWELP モデルと比較して高い AUC が得られているが，これは対義語の情報を考慮した上でラベル拡散を行なっているためだと考えられる．提案手法の章においても述べたように，ある単語が極性語である際に，その単語に対する対義語は， 元の単語とは反対の極性を有する場合が多いため，対義語の情報をモデルとして考慮することで，辞書構築の精度が向上したと考えられる．一方，DLS モデルでは他の 2 つのモデルと比較して，辞書構築の精度が低い結果となったが，これは分散表現の学習時において，単語の類義性・対義性に関する知識を十分に分散表現として埋め込む事が出来なかったためと考えられる．この点に関しては，d-LCE 法における，類義性・対義性に関する項の重要度を調整する事によって，より良い辞書構築精度を得る事ができると考える．

5.2 極性辞書拡張実験

SMLS モデルによる辞書拡張の結果，既存の辞書には含まれない 52 単語が抽出されたが，表 4 は，新たに辞書へ追加された単語の一部である．ポジティブ単語としては，gain・growth・lift などの，ポジティブな極性を有すると考えられる単語が追加され，一方ネガティブ単語としては，discourage・downgrade・depressing などの，ネガティブな極性を有すると考えられる単語が追加された．一方，substantial などのように，本来は極性を持たない単語も極性を持つ単語として，辞書に追加される結果となった．

表 4: SMLS によって，新たに辞書へ追加された単語の一部

Positive	Negative
gain, growth, foster, stability, accommodation, lift, substantial, strengthening	discourage, thin, cut, downgrade concerned, depressing, wan, lessening

図 1 は，辞書拡張の有無による，各マクロ変数の説明力の比較を示したものである．表から分かるように，辞書拡張の結果，鉱工業生産指数に対応する生産トピックを除いた全てのトピックにおいて，マクロ変数に対する説明力が向上するという結果が得られた．とりわけ，消費トピックに対応する PCE や，経済成長トピックに対応する GDP 成長率において，説明力を大きく向上させる事が出来ている．

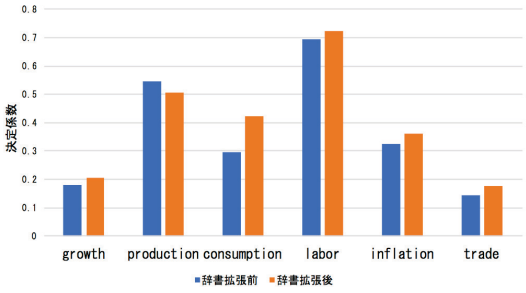


図 1: 辞書拡張の有無による，各マクロ変数に対する説明力の比較

辞書拡張によって精度が向上した理由として，既存の辞書には含まれていない極性語を獲得する事が出来，対象ドメインに対するより網羅性のある辞書を構築する事が出来たためと考えられる．一方で，本来極性語とは関係のない単語も極性語として追加されており，このような単語を機械的に排除する手法の検討や，さらなる辞書構築精度の向上は今後の課題である．

6 まとめ

本研究は，解析対象となるドメインに特化した，センチメント分析のための極性辞書を自動構築することを目的とする，ドメイン特化型極性辞書自動構築手法をした．提案手法である SMLS は，先行研究の SNWELP よりも辞書構築精度が上回る結果となり，提案手法の有効性が確認された．これは，ある単語が極性語である際に，その単語に対する対義語は， 元の単語とは反対の極性を有する場合が多いため，対義語の情報をモデルとして考慮することで，単語の極性値推定タスクにおいて，精度が向上したと考えられる．

また，SMLS を用いて既存の辞書を拡張した結果，拡張辞書を用いたトピック別センチメントのマクロ変数に対する説明力は一つのトピックを除き向上しており，提案手法のモデルの有効性が確認された．これは，辞書を拡張した結果，対象ドメインに対するより網羅性のある辞書を構築する事が出来たためと考えられる．

本研究の課題として，辞書構築精度向上のために，コパースの知識，単語の類義性・対義性に関する知識のみならず，センテンスの構文知識をモデルとして考慮する事が挙げられる．極性語どうしの文法的類似性から，単語間の係り受け関係には極性語としての特徴が含まれており，このような構文情報・コーパスのドメイン知識・単語の類義対義性を同時に分散表現として埋め込むことで，極性語獲得タスクにおいてよりよい分散表現になると期待される．

参考文献

[1] Ding, Y., and Foo, S. (2002). Ontology research and development, *Part 1-a review of ontology generation. Journal of information science* **28(2)**: pp.123-136.

[2] Alhothali, A., and Hoey, J. (2017). Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations, *arXiv preprint arXiv*:1703.09825.

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv preprint arXiv*: 1301.3781.

[4] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency, *In Advances in neural information processing systems*: pp.321-328.

[5] Nguyen, K. A., Walde, S. S. I., and Vu, N. T. (2016). Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction, *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*.

[6] Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance* **66(1)**: pp.35-65.

[7] Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction, *In LREC*: pp.1170-1175.

[8] 伊藤諒, 須田真太郎, 和泉潔 (2017). フォワードガイダンスの市場期待への影響分析 - テキストマイニング・アプローチ - , 第 *46* 回 *2016* 年度冬季 *JAFEE* 大会: pp.60-71.

テキストマイニングによる金融レポートの自動生成支援

Generation Support of Financial Reports by Textmining

丸澤 英将[1]　和泉 潔[1]　坂地 泰紀[1*]　田村 浩道[2]
本廣 守[2]
Hidemasa Maruzawa[1]　Kiyoshi Izumi[1]　Hiroki Sakaji[1]　Hiromichi Tamura[2]　Mamoru Motohiro[2]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo
[2] 野村證券株式会社
[2] Nomura Securities Co.,Ltd.

Abstract: Recently, with the increase of individual investors, the necessity of investment support technologies is increasing. Although analyst reports on which professional securities analysts forecast business performances or stock prices of companies are regarded as important investment decision materials, writing an analyst report is heavy burden. In this research, we summarize newspaper articles and support the generation of analyst reports by using knowledge of information features which are referred to as reasons for analysts' forecasts of business performances or stock prices in analyst reports.

1　はじめに

近年，我が国でも証券市場における個人投資家の比重が増大しており，個人投資家の投資判断を支援する技術の必要性が高まっている．個人投資家が重視する投資判断材料の一つに，証券会社が発行するアナリストレポートがある．アナリストレポートとは，証券市場調査・分析の専門家である証券アナリストが，企業の経営状態や収益力などを調査し，その結果をまとめたレポートのことである．アナリストレポートには，企業の業績や株価に対する証券アナリストの予想が示され，その根拠として，その企業の取り組む事業の近況・財務状況（企業のファンダメンタルズ）や事業に影響を与える経済・政治・社会状況（マクロ経済のファンダメンタルズ）などが言及される．

これら根拠として言及される情報は，証券アナリストの独自の調査によるものも含まれるが，規模の大きい企業のファンダメンタルズやマクロ経済のファンダメンタルズは，新聞などの媒体でも報じられるものである．ただし，媒体で報じられる様々な経済情報の中で，どの情報が企業の業績や株価に影響を与えるものであるかを見極めるには，証券アナリストの高度な専門知識を必要とする．アナリストレポートを参考にする個人投資家にとっては，企業の業績や株価に対する証券アナリストの予想だけでなく，証券アナリストがどのような情報を根拠として重視することで，その予想を導き出したのかという見極めの観点が重要である．例えば，酒井らは株式市場における次のような現象に注目している [1]．2012 年度上期のパナソニックの連結業績の発表では，前年同期比で売上高は減少したが，営業利益は増加したという内容であった．しかし，社長は「今回の大幅な業績の下ぶれの根本的な原因は，本業の不振にある」と語った．この発言が嫌気され，パナソニックの 2012 年 11 月 1 日の株価はストップ安となった．このように，株価を予想するためには，場合によっては決算発表中の営業利益の値ではなく，本業に関する社長発言を重視すべきという判断は，専門的な知見を必要とするものといえる．

そのため，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を捉え，新聞記事などの媒体から証券アナリストが注目するであろう情報に絞って自動的に要約する技術が重要である．この技術は，特に次のような点で有用である．アナリストレポートを作成してきた証券アナリストにとっては，レポート作成の支援に活用できる．執筆経験者によると，決算発表の時期には多くのレポート発行が集中し，膨大な経済情報の中から，企業の業績や株価の変動を引き起こす要因について人手で整理する作業は負担が大きいという．本技術により有用な経済情報を要約し，情報整理の作業時間が短縮できれば，より独自性の高い調査・分析に注力したり，執筆するレポートの数を増

*連絡先：東京大学大学院工学系研究科システム創成学専攻
和泉・坂地研究室
〒 113-8654 東京都文京区本郷 7-3-1
E-mail: staff@socsim.org

やしたりすることができる．アナリストレポートを参照してきた個人投資家にとっても，投資判断材料の充実に繋がる．アナリストレポートが発行される頻度は銘柄[1]によって大きく異なり，注目度の高い銘柄は四半期ごとの決算発表に合わせて度々発行されるが，発行されることが少ない銘柄もある．証券市場の上場企業数が東京証券取引所だけでも 3500 社近くに上る中で，特に個人投資家は保有する銘柄の選定理由が個々人で異なっており，各人が注目している銘柄のアナリストレポートが必ずしも頻繁に発行されているとは限らない．そこで，本技術により証券アナリストの執筆作業が効率化され，発行されるレポートの銘柄数・頻度が増えたり，個々のレポートの質がより高まったりすれば，個人投資家の投資判断材料を充実させることができる．

この目的に応用可能な技術として，近年，自然言語処理やテキストマイニング技術の進展により，テキストデータから自動的に重要な情報を抽出する技術が発達してきている [2, 3]．しかし，これらの要約技術は，そのままでは事象の背景にある因果関係を考慮できない．一方，文の因果関係の構造に注目し，原因表現を取り出す手法も提案され始めている [4]．丸澤らは，この文の原因表現を取り出す技術を応用し，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を学習した知識を，重要情報抽出技術に組み込むことで，業績変動要因文を新聞記事から抽出している [5]．ただし，抽出した文がアナリストレポートの執筆に有用であるかの実務家による評価は行われていない．本研究では，丸澤らの手法を用いて，新聞記事からアナリストレポート執筆に有用な要約文を自動生成するシステムを構築し，実務家の評価により実用性の検証を行う．

2　アナリストレポート執筆支援文自動生成の全体の流れ

アナリストレポート執筆支援文の自動生成に至るまでの流れを概説する．まず，アナリストレポートの文中で頻出する因果関係の構造を抽出する．次に，因果関係のうち原因表現を，証券アナリストによる企業業績・株価予想の根拠情報[2]として獲得する．その根拠情報と類似する内容を指す文を新聞記事中から探し出し，根拠情報となり得る企業業績要因を取得する．このようにして取得した企業業績要因をまとめ，アナリストレポート執筆支援文を自動生成する．全体の流れを図 1 に示す．

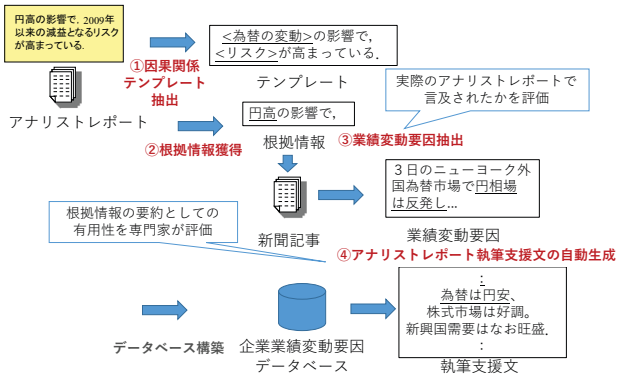


図 1: アナリストレポート執筆支援文自動生成の全体の流れ

本システムの実装例として，Web サーバー上のシステムとして実装したものの動作画面を図 2，3 に示す．

システムには，一定期間のアナリストレポートのテキストデータを与え，予め根拠情報の特徴を学習させておく．そして，データベース中に要約対象となる新聞記事のテキストデータを格納し，図 2 のように，要約対象として用いる新聞記事の期間，注目する銘柄の業種，自動生成文の並べ方を入力として指定し，Search ボタンを押す．システムは指定された期間（図の例では 2014 年 1 月〜3 月）の新聞記事のテキストデータを参照し，指定された業種についての学習した根拠情報と類似する内容を指す文を探し出す．それらの文のうち重要度の高いものを，図 3 のように，指定された並べ方によりいくつか並べて提示する．各文には，抽出元の記事の発行日と見出しを併記する．並べ方には，記事の発行日の時系列順，後述する業績関連度指数順のいずれかを用いる．提示された文を参考にして，証券アナリストがレポートを執筆する，という用途を想定する．

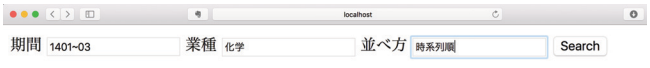


図 2: 提案手法によるシステムの入力画面

[1]証券会社に上場されている株式の企業名

[2]本稿では，酒井ら [6] に倣い，アナリストレポートの中で企業の業績や株価の変動を引き起こす要因として言及されている情報を，「（アナリスト予想の）根拠情報」と呼ぶ．また，一般に，企業の業績に影響を与える要因を「業績変動要因」と呼ぶ．

図 3: 提案手法によるシステムの出力画面

*原油安及び探鉱費の増加*__を主因に,__
（根拠部）　**（根拠部手がかり表現）**

*YY.M 期の純利益予想*__を下方修正した.__
（予想部）　**（予想部手がかり表現）**

図 4: アナリストレポート中の文の例

3　新聞記事からの根拠情報となり得る業績変動要因の取得

本システムに用いる，根拠情報となり得る業績変動要因の新聞記事からの取得には，丸澤らの手法 [5] を用いる．本節に，その手法を簡単に述べる．

3.1　アナリストレポートからの根拠部，予想部の抽出

アナリストレポート中の因果関係の抽出には，酒井らのブートストラップ法による手法 [4] を用いる．この手法では，アナリストの予想根拠文を特徴付ける手がかり表現と，手がかり表現に係る節の中で共通して頻繁に出現する共通頻出表現を定義する．最初に少数の手がかり表現と共通頻出表現を与えることで，互いに係り受け関係にある新たな共通頻出表現と手がかり表現が連鎖的に獲得される．

この手法を用いるに当たって，特にアナリストの予想を示す文の部分と，その予想の根拠を示す文の部分を分離して抽出する．前者を予想部，後者を根拠部と呼ぶ．アナリストレポート中の文の例を図 4 に示す．

この場合,「(を) 主因に,」を根拠部手がかり表現として，それに係る文の部分「原油安及び探鉱費の増加」を根拠部とする．一方,「(を) 下方修正した.」を予想部手がかり表現として，それに係る文の部分を予想部「YY.M 期の純利益予想」として，根拠部とは完全に分離して抽出する．なお，根拠情報は,「原油価格が下がった上に，探鉱にかかるコストが上がった」という経済イベントを指す．

3.2　根拠情報の業種別特徴の学習

次に，得られた根拠部手がかり表現から，根拠部がどのような業績変動要因を指し示しているかという意味的な特徴を学習する．まず，先に獲得した予想部の手がかり表現と係り受け関係にある文の部分を，根拠部として抽出する．この根拠部を形態素解析し，英単語を除く名詞に分類されるもののうち,「数，接尾，非自立」の下位分類を除いた形態素の組を取得する．この名詞の組を，全根拠部の名詞の組中での tf-idf 値を用いてベクトル化したものを，根拠部の特徴量とする．

すなわち，各組中の名詞について次の値を計算し，その組の特徴ベクトルの長さが 1 となるように正規化したものを特徴量とする．

$$v = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \left(\log \frac{|\mathrm{D}|}{|d : t_i \in d|} + 1 \right) \tag{1}$$

ここで，$n_{i,j}$ はアナリストレポート中の根拠部 d_j の名詞の組における名詞 t_i の個数，D はアナリストレポート中の根拠部の名詞の組全ての集合である．

ここで，根拠部とその根拠部を抽出したアナリストレポートが言及している銘柄が属する業種の関係に注目する．同じ業種に属する銘柄についての根拠部の集合には，似た根拠情報を指す集合が存在すると考えられる．逆に，似た根拠情報を指す根拠部の集合でも，特定の業種に偏って存在するものと，様々な業種に満遍なく存在するものがあると考えられる．

そこで，先に得た根拠部の特徴量を用いて根拠部を多クラス分類し，各クラスの根拠部がどの業種についての根拠部であるかの頻度分布を次の式のように計算する．

$$f_{n,m} = |v : v \in (\mathrm{C_n} \cap \mathrm{I_m})| \tag{2}$$

ここで，$f_{n,m}$ はクラス n の根拠部が業種 m についての根拠部である頻度，v は根拠部の特徴ベクトル，C は根拠部の特徴ベクトルを分類したクラスを表す集合 ($n = 1, 2, ...N_\mathrm{C}, N_\mathrm{C}$: クラスの総数)，I は根拠部の属する業種を表す集合 ($m = 1, 2, ...N_\mathrm{I}, N_\mathrm{I}$: 業種の総数) である．

さらに，この頻度分布のクラスごとの偏りを，次の式のように平均情報量 e を用いて定量化する．

$$e = -\sum_m f_{n,m} \log_2 f_{n,m} \tag{3}$$

この平均情報量が小さいほど，特定の業種に偏って存在する根拠部が属するクラスであり，平均情報量が

大きいほど，様々な業種に満遍なく存在する根拠部が属するクラスであると言える．

3.3 新聞記事からの業種別根拠情報の獲得

各クラスの代表点である重心ベクトルと頻度分布を用いて，新聞記事から新たな根拠情報を獲得する．

まず，新聞記事の文章から，アナリストレポートでの根拠部の特徴量を得るために使用した名詞を抽出する．ただし，単に特徴量に使用した名詞に一致する名詞のみを抽出した場合，抽出される根拠情報が限られてしまう．そこで，新聞記事の文章中の名詞を，構文上の出現位置の特徴を用いて分散表現を生成する word2vec 法 [7] を使用することで，文脈上の類似度の高い名詞まで抽出できるよう拡張する．

こうして抽出した新聞記事の文章中の名詞の組を，アナリストレポートでの根拠部の特徴量を得るために使用した tf-idf 値を用いてベクトル化することで，新聞記事の文章の特徴量とする．

すなわち，各組中の名詞について次の式の値を計算し，その組の特徴ベクトルの長さが 1 となるように正規化したものを特徴量とする．

$$v = \frac{n_{i,l}}{\sum_k n_{k,l}} \cdot \left(\log \frac{|\mathrm{D}|}{|d : t_i \in d|} + 1 \right) \tag{4}$$

ここで，$n_{i,l}$ は新聞記事の文章 a_l の名詞の組における名詞 t_i の個数，D はアナリストレポート中の名詞の組全ての集合である．

新聞記事のうち根拠情報として獲得するのにふさわしくない「観測記事」,「決算記事」を除いた新聞記事の文章の特徴ベクトルと，根拠情報を分類した各クラスの重心ベクトルとのコサイン類似度を次の式で求め，新聞記事の文章と各クラスとの類似度とする．

$$s_{l,n} = v_l \cdot g_n \tag{5}$$

ここで，$s_{l,n}$ は新聞記事の文章 a_l とクラス $\mathrm{C_n}$ との類似度，v_l は新聞記事の文章 a_l の名詞の組の正規化した特徴ベクトル，g_n はクラス $\mathrm{C_n}$ の重心ベクトルを長さ 1 に正規化したベクトルである．また，$\cdot$ は内積演算子である．

さらに，各クラスとの類似度と，そのクラスの根拠情報がどの業種の銘柄の業績予想の根拠となり得るかの頻度分布との加重平均を次の式のように計算することで，新聞記事の文章がどの業種に属する銘柄の業績予想の根拠となり得るかの指標とする．

$$c_{l,m} = \sum_n s_{l,n} \ f_{n,m} \tag{6}$$

ここで，$c_{l,m}$ は新聞記事の文章 a_l の業種 m への業績寄与度である．以下，この指標を新聞記事の文章の各業種への業績寄与度と呼ぶ．

様々な業種で満遍なく業績寄与度が高い根拠情報が混在してしまうことを防ぐため，各新聞記事の文章の全業績寄与度中，各業種への業績寄与度の値の偏差値を次の式で求める．

$$\begin{aligned} \mathrm{dev(c_{l,m})} = &\quad \frac{c_{l,m} - \mu_l}{\sigma_l} \cdot 10 + 50 \\ \mu_l = &\quad \tfrac{1}{N_\mathrm{I}} \textstyle\sum_m^{N_\mathrm{I}} c_{l,m} \\ \sigma_l = &\quad \tfrac{1}{N_\mathrm{I}} \textstyle\sum_m^{N_\mathrm{I}} (c_{l,m} - \mu_l)^2 \end{aligned} \tag{7}$$

ここで，$\mathrm{dev(c_{l,m})}$ は新聞記事の文章 a_l の全業績寄与度中，業種 m への業績寄与度の値の偏差値，N_I は業種の総数である．

最後に，いずれの業績寄与度もわずかしかないが，その業種への業績寄与度だけが少しだけ高いという新聞記事の文章が混在してしまうことを防ぐため，各業種への業績寄与度とその値の偏差値の調和平均を次の式のようにとったものを，業績関連度指数と定義する．

$$r_{l,m} = \frac{2 \ c'_{l,m} \ \mathrm{dev(c'_{l,m})}}{c'_{l,m} + \mathrm{dev(c'_{l,m})}} \tag{8}$$

ここで，$r_{l,m}$ は新聞記事の文章 a_l の業種 m への業績関連度指数，$c'_{l,m}$ は新聞記事の文章 a_l の業種 m への業績寄与度を平均 50，標準偏差 10 に正規化した値である．

4 アナリストレポート執筆支援文の自動生成

前節の手法により，各業種に属する銘柄の業績予想の根拠となり得る重要記事（業績関連度指数の高い記事）を取得できる．これらを，特定の時期・業種について抽出し，発行日・見出しを付して任意の数並べることで，アナリストレポートの執筆支援文として提示する．

手法全体の適用手順の具体例を示す．学習対象のアナリストレポートにおいて，化学業種に属する銘柄では,「原油安を主因に，純利益予想を下方修正した.」や「原油価格の持続的な下落により，純利益の減少が見込まれる.」のように，原油安が主な根拠情報として言及されることが特異的に多かったとする．この場合，まず 3.2 節の処理により,「原油安」や「原油価格の持続的な下落」というテキストが根拠部として抽出される．

そして，3.2 節の処理により，これらの根拠部を特徴ベクトルで表現して比較することで，これらが同じ原

油安について言及している類似するテキストであり，ある業種に偏って存在するテキスト群（クラス）であることが定量化される．

このように学習した情報を用いて，3.3 節の手法で新聞記事中の文を検索する．新聞記事の中に「原油価格続落」を報じる文があった場合，この文が「原油安」や「原油価格の持続的な下落」と類似したテキストであり，化学業種の根拠情報と関連が高い業績変動要因であることが計算される．この計算を特定期間の新聞記事の全文に渡って行うことで，その期間中の特定業種の根拠情報となり得る業績変動要因を含む記事を一覧できる．

さらに，本節のように，発行日・見出しを付して任意の数並べることで，化学業種について，表 1 のような文の列を提示することができる（表中の記事は架空のものである）．

表 1: アナリストレポート執筆支援文の自動生成例

文 (発行日 見出し)
会合後，原油先物は最安値を更新した．(2015/4/xx OPEC 会合，減産合意ならず)
国際原油相場の下落は xヶ月連続．(2015/4/xx 国際原油相場，xヶ月連続の下落)
原油価格の暴落が止まらない．(2015/5/xx 商品先物市場，総じて低調)

このような文の列を直近の新聞記事から自動生成することで，化学業種に属する銘柄についてのアナリストレポートを執筆する証券アナリストに対して，化学業種に属する銘柄の主要な業績変動要因である原油安についての要約文を提示し，執筆の支援を行うことができる．

5 評価

本手法で抽出した文の列について，表 2 の条件で，証券会社の実務家 3 名に評価を依頼した．

学習データには，野村證券株式会社の Global Markets Research レポート（2014 年下半期発行分，日本株 216 銘柄の表紙部分）を用いた．銘柄を分類する業種には，「野村 19 業種分類」（化学，鉄鋼・非鉄，機械，自動車，電機・精密，医薬・ヘルスケア，食品，家庭用品，商社，小売り，サービス，ソフトウェア，メディア，通信，建設，住宅・不動産，運輸，公益，金融）を用いた．新聞記事には，日本経済新聞本紙の朝・夕刊の地方面を除く 2015 年度の記事（スポーツ記事など，経済記事以外も含む）を用いた．

アナリストレポートからの根拠部の抽出では，初期の手がかり表現，共通頻出表現にそれぞれ「考慮し，反映し，評価し」,「増益，改善，成長」を用いた．予想部の抽出では，初期の手がかり表現，共通頻出表現にそれぞれ「継続する，予想する」,「利益，業績，売上」を用いた．形態素解析器に MeCab[3]を，係り受け解析器に CaboCha[8] を使用した．多クラス分類には，k-means 法を用い，k=100 とし，実装に Python ライブラリ scikit-learn 0.19.1[4]を利用した．word2vec 法のモデルには，ロイター社の 2003 年から 2013 年の経済記事の文章をコーパスとし，200 次元で分散表現を生成するよう学習したものを用いた．文脈上近い意味の名詞とみなす類似度の閾値には，0.7 を使用した．観測記事・決算記事を抽出する正規表現には，丸澤ら [5] と同じものを用いた．

比較対象には，因果関係の構造に注目して根拠部を分離することをせず，単にアナリストレポートの各文全体から名詞を抜き出してそれらの tf-idf 値を特徴量に用いた単純 bag-of-words（BOW）法を用いた．

評価基準は表 3 のように指定した（評価には専門的な判断が求められるため，負担を考慮し，全業種ではなく無作為に選んだ 5 業種のみの評価とした）．

表 2: 有用性評価実験の条件

期間	2015 年 4 月〜2016 年 3 月中の各四半期の計 4 期間
業種	無作為に選んだ 5 業種（化学，自動車，小売り，食品，住宅・不動産）
並べ方	時系列
抽出文の提示方法	各期間・業種の抽出文 5 つずつを，どちらが提案手法によるものかを伏せて比較手法によるものと並べて提示
参考情報の提示方法	当該期間・業種の実際のアナリストレポートの概要を事前に提示し，根拠情報の参考としてもらう

表 3: 有用性評価実験の評価基準

評価	当該期間における当該セクターの企業の業績に影響を与えそうな情報の抽出について，以下のいずれに当てはまるか
○	よく抽出できている．
△	一部含むが、要点にずれがある．
×	ほとんど含まない．

[3]http://taku910.github.io/mecab/
[4]http://scikit-learn.org/

6　結果

まず，3 名の実務家による評価の一致を測るため，Cohen の κ 値を 2 名の組み合わせごとに求めた．結果を表 4 に示す．

表 4: 有用性評価結果の Cohen の κ 値

評価者 A & 評価者 B	0.54
評価者 B & 評価者 C	0.72
評価者 C & 評価者 A	0.38
平均値	0.55

今回の評価結果は一致度が中程度と言える．評価がややばらついた原因は，何をアナリストレポートの根拠情報とするべきかは実務家でも判断が分かれることがある専門的な事項であり，一部の文で評価が 1 名は○，1 名は△，1 名は×と分かれたケースが見られたことが挙げられる（実際のアナリストレポートの執筆は，業種ごとに専門の担当者が行うが，今回は 3 名の評価者に 5 業種全てについてそれぞれ評価してもらった）．そのため，評価の多数決ではなく，評価○を 1 点，△を 0.5 点，×を 0 点と点数化して 3 名の評価の平均点を算出し，すべての文が○と評価された場合の点数を 100%として集計した．

業種ごとの集計結果を図 5 に示す．(図中の業種は，比較手法である単純 BOW 法での精度の降順に並べた．)

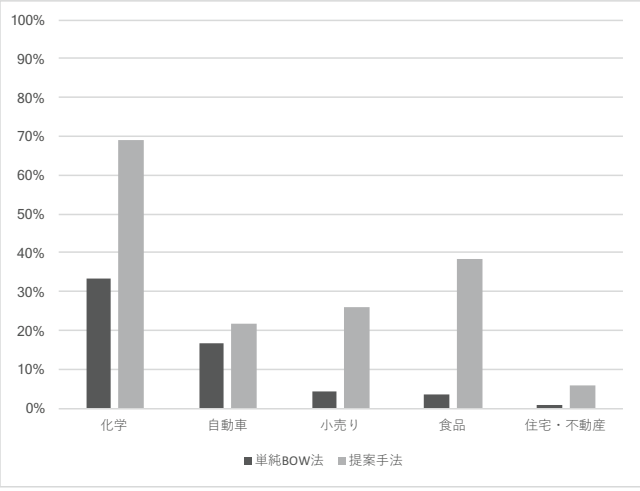


図 5: アナリストレポート執筆支援文の有用性評価による精度

比較手法では，化学業種での精度が 30%を超えたのみで，自動車業種では 20%，小売り，食品，住宅・不動産業種では 10%を下回る精度だった．一方，提案手法では，化学業種で 70%近い精度を達成したほか，食品業種で大幅に精度が改善した．自動車，小売り，住宅・不動産業種での精度も，いずれも比較手法のものを上回った．

比較手法での精度の低さは，アナリスト予想の根拠情報になり得る業績変動要因を新聞記事から抽出するという今回の問題設定に対しては，単なるキーワードマッチングでは，実務家による有用性評価に耐え得る精度に至らないことを示していると言える．

7　考察

提案手法によって良好な精度を達成した化学業種での抽出文例の比較を表 5 に示す．

表 5: 自動抽出文例（化学業種）

提案手法	石油製品の取引価格がアジアで軒並み下落している．
比較手法	各地の寺社などに油のような液体がまかれた事件は，警察庁によると 14 府県 43ヶ所（1 日現在）に被害が広がった．

いずれの文も「油」という，化学業種に属する銘柄のアナリストレポートの文中に出現するキーワードを含んでいる．原油・石油・灯油などの原料・加工品の需給動向は，化学業種に属する銘柄についての代表的な根拠情報である．しかし，比較手法で抽出された文中での油は，これらの需給動向とは関係が薄い．それに対し，提案手法では「取引価格」の「下落」という需給動向を表す名詞や,「アジア」という当該時期に主要な消費地としてアナリストレポートの文中で度々言及されていた名詞を含んだ文を抽出できている．これは，提案手法において，アナリストレポートの文中から原因表現の部分だけを取り出して特徴量を設計しているため，これらの根拠情報のキーワードを重視した新聞記事からの文抽出ができることの効果が現れているものと考えられる．

次に，提案手法によって大幅な精度の改善が見られた食品業種での抽出文例の比較を表 6 に示す．

表 6: 自動抽出文例（食品業種）

提案手法	伊藤忠商事はチョコレート原料の加工・販売事業に参画する．
比較手法	ショットピーニングは微小な金属粒を材料の表面にぶつけて耐久性を高める加工技術．

いずれの文も「原料・材料」や「加工」という，食品業種に属する銘柄のアナリストレポートの文中に出現するキーワードを含んでいる．食品の原材料の加工

事業や技術の動向は，食品業種に属する銘柄についての代表的な根拠情報である．しかし，比較手法で抽出された文中での材料加工は，食品ではなく金属についてである．それに対し，提案手法では食品の加工・販売事業についての文を抽出できている．これは，化学業種での考察と同様に,「販売」や「事業」への「参画」という根拠情報として同時に出現することが多いキーワードを重視できているためと考えられる．なお，食品業種では化学業種に比べて，提案手法・比較手法いずれの精度も低い．いずれも原料について言及されることが多い業種だが，化学業種では原油など限られた種類の原料が繰り返し言及されるのに対して，食品業種では多様な原料が言及され，キーワードとして学習することがより困難であることが一因と考えられる．

また，提案手法で抽出された文が，企業自体は食品業種に分類されていない「伊藤忠商事」の事業動向についてであることに注目されたい．新聞記事中に出現する企業名や企業との関連性が高いキーワードを基に，抽出文が関連する業種を特定することも可能だが，その場合，この文は「伊藤忠商事」という企業名を基に商社業種に分類されることになる．アナリストレポートの根拠情報として言及されるのは，当該業種の動向に限られず，原料の採掘・輸送・加工，製品の輸送・販売・宣伝，サービスの提供など一連のサプライチェーンで関わる業種全般の動向に及ぶ．この点を考慮した情報抽出が可能な点も，因果関係に注目する提案手法の特徴と言える．さらに，提案手法で算出する業績関連度指数は，業種ごとに関連度を連続値で評価するため，文が関連する先を単一の業種に限定せず，業績へ影響が及ぶ全業種について関連度を見ることが出来る．

最後に，提案手法・比較手法いずれも精度が低かった住宅・不動産，自動車業種について，抽出文例の比較をそれぞれ表 7，表 8 に示す．

表 7: 自動抽出文例（住宅・不動産業種）

提案手法	安倍晋三内閣は法人税や消費税の見直しを進めてきたが，所得税の抜本改革は放置してきた．
比較手法	財務省が２８日に発表した２０１４年～４月，１５年３月の税収実績は前年同期比１２・３％増と高い伸びになった．

住宅・不動産，自動車業種はそれぞれ税，外国為替の話題が多く抽出され，そのほとんどが実務家評価で根拠情報とは直接は関係ない文と判定された．表に掲載した提案手法による抽出文は，その中でも，実務家による評価が１名は○，１名は△，１名は×と分かれたもので，これらの業種の中では比較的関係ありと評価された文である．

表 8: 自動抽出文例（自動車業種）

提案手法	シティグループ証券は２０１７年の円ドル相場を年平均１ドル＝１２９円と予想していますが，２０年には１１６円まで円高になると予想しています．
比較手法	ドル／円１ドル＝１１２．１７～１１２．２０円（７０銭の円高）【中略（ユーロ／円，ユーロ／ドル相場の値)】（東京市場１２時時点）．

住宅・不動産業種では，この時期には消費税や相続税法改正の影響が根拠情報として盛んに議論されていた．不動産は高額な消費・相続の対象であるため，税の影響は大きい．そのため，税制・税収の話題が多く抽出されたと考えられる．政府の税収は関係が薄いが，税制改革の動向は根拠情報になり得るということである．

一方，自動車業種では，前章で述べたのと同様にこの時期も外国為替動向が根拠情報として大きく注目されていた．日々の具体的な相場の値を報じる文は関係が薄いが，長期の外国為替動向の予想は根拠情報になり得るということである．今回の提案手法では名詞のみを特徴量に用いているため，外国為替についての文を抽出することはできても，その中で，言及期間の長短や過去の事実と将来の予想の区別を付けることは困難である．このことから，新聞記事の文自体に，今回アナリストレポートの文に用いたような文構造解析を行えば，これらを区別し，より精度を向上できる可能性がある．

全般的に，小売り，食品，住宅・不動産のように消費動向が主な根拠情報となる業種では，キーワードに一般的な名詞が混在してしまうため，社会面の記事など業績変動要因の少ない文を誤って抽出してしまうことが多かった．この問題は，今回の提案手法では根拠情報を含む文から作った特徴量という正例のみを学習していることに一因があると考えられ，根拠情報を含まない文から作った特徴量を負例として与えることで改善する可能性がある．

8　まとめ

本研究では，証券アナリストの業務支援やそれによる個人投資家の投資判断材料の充実のため，アナリストレポートの中で企業の業績変動要因として言及されている根拠情報を学習し，根拠情報となり得る企業の業績変動要因を新聞記事から抽出することで，アナリストレポートの執筆を支援する要約文を自動生成する手法を提案した．アナリストレポートからの根拠情報

の学習では，文の係り受け解析を行い，背後にある因果関係に注目した処理を行った．また，業種別に根拠情報を学習・分類することで，業種ごとの企業業績への寄与度を定量化した．これにより，企業の業績変動要因を新聞記事から期間・業種別に抽出することが出来た．

本提案手法は，証券会社の実務家による評価で，実験を行った全業種について比較手法の単純 BOW 法より高い有用性を示した．これは，提案手法が文の因果関係の構造を反映した学習を行っており，因果関係の把握が重要な経済分野に適したテキストマイニング手法であるためと考えられる．一方，一部の業種での取得精度には課題が残り，文抽出の有用性においても改善の余地がある．根拠情報を含まない文を負例として与えて学習することなどで精度を改善し，新聞記事の即時的な取得や実際のアナリストレポートに近い形式での出力を行うことで，より実用的なアナリストレポート執筆支援文の自動生成システムを構築できることが期待される．

参考文献

[1] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀：企業の決算短信 pdf からの業績要因の抽出, 人工知能学会論文誌, Vol. 30, No. 1, pp. 172–182 (2015)

[2] Otterbacher, J., Erkan, G., and Radev, D. R.: Using random walks for question-focused sentence retrieval, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922 (2005)

[3] Filippova, K., Surdeanu, M., Ciaramita, M., and Zaragoza, H.: Company-oriented extractive summarization of financial news, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 246–254 (2009)

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE transactions on information and systems*, Vol. 91, No. 4, pp. 959–968 (2008)

[5] 丸澤英将, 和泉潔, 坂地泰紀, 田村浩道：業種別企業業績要因を含む新聞記事の抽出, 第 *19* 回金融情報学研究会, pp. 71–77 (2016)

[6] 酒井浩之, 柴田宏樹, 平松賢士, 坂地泰紀：アナリストレポートからのアナリスト予想根拠情報の抽出, 第 *17* 回金融情報学研究会, pp. 25–30 (2016)

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.:Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[8] 工藤拓, 松本裕治 :チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol 43, No. 6, pp. 1834–1842 (2002)

潜在トピック空間上でのマルチタスク学習による企業評価テキストデータを用いた財務指標予測

Predicting a financial index by articles: Multi-task learning on latent topic space

茂庭 綾香[1*] 中川 雄太[1] 江口 浩二[1†]
Ayaka Moniwa[1], Yuta Nakagawa[1], Koji Eguchi[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: This paper aims to predict a company's financial index by analyzing articles about the company. The authors propose MultiMedLDA, which is one of supervised topic models. MultiMedLDA assumes that each document has two types of labels, discrete value label and continuous one. It models relation between each document and these labels, and predicts an unknown label based on known labels and the documents. Making use of not only documents but also the known labels, it improves prediction accuracy. We evaluated our model with data from the "Japan Company Handbook". Using comments for each company as a document, the type of industry as a discrete value label and the company's ROE (Return On Equity) as a continuous value label, we predicted the ROE in the evaluation.

1 はじめに

企業の今後の業績を予想する際の手掛かりには，過去の業績に関する数値情報の他に，ニュース記事などの文書データが挙げられる．文書データには，事業展開や市場の動向など数値では表しきれない抽象的な情報が含まれている．本稿ではこのことに着目し，文書データをもとに企業の財務指標を予測する課題に取り組む．

文書データから数値を予測する既存のモデルには，教師ありトピックモデル (Supervised topic models)[1] や最大マージントピックモデル（Maximum Entropy Discrimination LDA : MedLDA）[2] が挙げられる．これらはトピックモデルの一種であり，文書を単語の多重集合（Bag-of-words : BoW）として捉える．Bag-of-words は文書中の各語彙の出現頻度にのみ着目した表現形式であり，語順は考慮しない．ここで挙げた既存モデルはいずれも，文書と予測対象数値ラベルの組を一対のデータとし，文書情報のみからラベルを予測している．しかし，企業の財務指標を予測する際には，文書情報以外にも業種や国籍といった付加情報の活用が期待できる．そこで本稿では，それらの付加情報もラベルとして文書に付与し，文書と複数のラベルが組になったデータを扱うマルチタスク最大マージントピックモデル（MultiMedLDA）を提案する．

*連絡先：神戸大学 大学院 システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: moniaya@cs25.scitec.kobe-u.ac.jp
†連絡先：神戸大学 大学院 システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: eguchi@port.kobe-u.ac.jp

2 関連研究

2.1 最大マージントピックモデル

最大マージントピックモデル（Maximum Entropy Discriminated LDA : MedLDA）[2] はトピックモデルの一種である．トピックモデルは文書データを表現するモデルの一種であり，文書一つをそこに含まれる単語の多重集合（Bag-of-Words : BoW）と捉える．それぞれの単語の背後には「トピック」と呼ばれる潜在変数を仮定し，これを推定することで文書の潜在特徴を表現する．すなわち，文書一つはそこに含まれる単語数と同じ個数のトピックの多重集合として表すことができる．

その上で MedLDA では，各文書に数値ラベルが一つずつ付与されていると想定する．MedLDA の生成過程を以下に示す．

1. 文書 $d(d \in 1, \ldots, D)$ に対して，$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択する．
2. 文書 d 内の N_d 個の単語 $w_{d,n}(n \in 1, \ldots, N_d)$ に対して，
   (a) トピック $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択する．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_{\mathrm{t}})$ $(t = z_{d,n})$ を選択する．
3. D 個の文書に対してラベル $y_d \sim F(\boldsymbol{\eta}, \boldsymbol{z}_d)$ を選択する．

MedLDA における変数同士の関係を表したグラフィカルモデルを図 1 に示す．図 1 中の y は各文書に付与された教師ラベルである．また，$\boldsymbol{\eta}$ はラベル評価時の各トピックに対する重み係数である．

なお，教師ラベル y が連続量であるときの MedLDA は回帰モデル，離散量であるときは分類モデルに位置づけられる．それぞれについて 2.1.1 節と 2.1.2 節にて後述する．

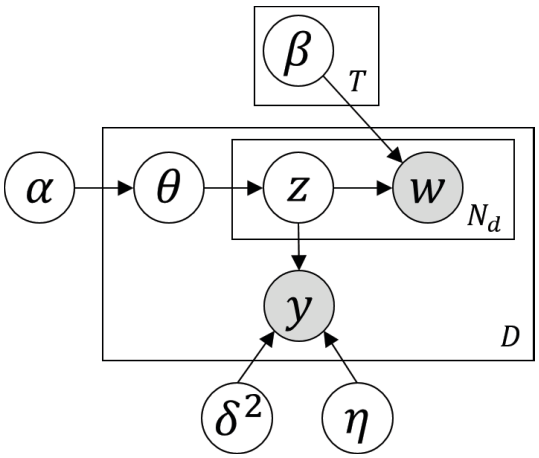


図 1: MedLDA のグラフィカルモデル

2.1.1 回帰問題を想定した最大マージントピックモデル

連続値ラベル $y \in \mathbb{R}$ を持つ文書データを扱う MedLDA Regression (MedLDA-Reg) について説明する．MedLDA-Reg では $y_d|\boldsymbol{z}_d, \boldsymbol{\eta}, \delta^2 \sim \mathcal{N}\left(\boldsymbol{\eta}^\top \bar{\boldsymbol{z}}_d, \delta^2\right)$ とし，マージン最大化 [3][4] を考慮することにより以下のような最適化問題が定義される．

P1(MedLDA-Reg) :

$$\min_{q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}),\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2,\boldsymbol{\xi},\boldsymbol{\xi}^\star} \mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})) + C\sum_{d=1}^{D}(\xi_d + \xi_d^\star)$$

$$\text{s.t. } \forall d : \begin{cases} y_d - \mathbb{E}\left[\boldsymbol{\eta}^\top \overline{Z}_d\right] \le \epsilon + \xi_d & [\mu_d] \\ -y_d + \mathbb{E}\left[\boldsymbol{\eta}^\top \overline{Z}_d\right] \le \epsilon + \xi_d^\star & [\mu_d^\star] \\ \xi_d \ge 0 & [v_d] \\ \xi_d^\star \ge 0 & [v_d^\star] \end{cases}$$

制約式中の $\xi_d, \xi_d^\star$ は訓練データの誤差を吸収する程度を示すスラック変数であり，ϵ は許容誤差である．定数 $C(> 0)$ は正則化パラメータ，$\mu_d, \mu_d^\star, v_d, v_d^\star$ はラグランジュ乗数である．上式右端の [ ] はラグランジュ関数を求める際の制約式とラグランジュ乗数の対応を表している．$\mathbb{E}[\cdot]$ は期待値を表す．また，各変数は $\boldsymbol{Z} := \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_D\}$,$\boldsymbol{z}_d := \{\boldsymbol{z}_{d,1}, \ldots, \boldsymbol{z}_{d,N_d}\}$, $\boldsymbol{\Theta} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$, $\boldsymbol{y} := \{y_1, \ldots, y_D\}$, $\boldsymbol{W} := \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D\}$, $\boldsymbol{w}_d := \{w_{d,1}, \ldots, w_{d,N_d}\}$, $\boldsymbol{\beta} := \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\}$, $\boldsymbol{\xi} := \{\xi_1, \ldots, \xi_D\}$, $\boldsymbol{\xi}^\star := \{\xi_1^\star, \ldots, \xi_D^\star\}$ を表す．$\boldsymbol{z}_{d,n}$ は $t = z_{d,n}$ 番目の要素のみ 1，それ以外の要素は 0 となる T 次元の指標ベクトルであり，確率変数 $Z_{d,n}$ のインスタンスである．$\overline{Z}_d := \frac{1}{N_d}\sum_{n=1}^{N_d} Z_{d,n}$，$\bar{\boldsymbol{z}}_d := (1/N_d)\sum_{n=1}^{N_d} \boldsymbol{z}_{d,n}$ である．

ここからは MedLDA-Reg の更新式の導出を行う．目的関数の $\mathcal{L}$ は

$$\mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})) = -\mathbb{E}_q\left[\log p(\boldsymbol{\Theta},\boldsymbol{Z},\ \boldsymbol{\eta},\boldsymbol{y},\boldsymbol{W}|\boldsymbol{\alpha},\boldsymbol{\beta},\delta^2)\right] - \mathcal{H}\left(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})\right) \quad (1)$$

である．$\mathcal{H}$ は事後分布 $q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta})$ のエントロピーであり，$\mathcal{H}(q) := -\sum q\,\log(q)$ である．最適化問題 P1 は一般的に解くことが困難であるため，変分近似を行いて q についての独立性を仮定する．

$$q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}) = q(\boldsymbol{\eta})\prod_{d=1}^{D} q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)\prod_{n=1}^{N_d} q(\boldsymbol{z}_{d,n}|\boldsymbol{\phi}_{d,n}) \quad (2)$$

ここで，$\boldsymbol{\gamma}_d$ および $\boldsymbol{\phi}_{d,n}$ は変分パラメータであり，$\boldsymbol{\gamma}_d$ はディリクレ分布パラメータの T 次元のベクトル，$\boldsymbol{\phi}_{d,n}$ は T トピックの多項分布パラメータである．$\mathbb{E}[Z_{d,n}] = \boldsymbol{\phi}_{d,n}$，$\mathbb{E}\left[\boldsymbol{\eta}^\top \overline{Z}_d\right] = \mathbb{E}[\boldsymbol{\eta}]^\top (1/N_d)\sum_{n=1}^{N_d}\boldsymbol{\phi}_{d,n}$ が成り立つ．

そして変分 EM アルゴリズムを行い，各パラメータを最適化する．変分 EM アルゴリズムでは次の 2 ステップを繰り返し行う．

1. **E-Step**：潜在変数の事後分布を推定
2. **M-Step**：未知変数を推定

更新式の導出では変分下限を最大化する各パラメータを求める．また，最適化問題 P1 の制約式を目的関数に

組み込み，ラグランジュ関数 L^r を定義する．

$$\begin{aligned}L^r=&\mathcal{L}(q)+C\sum_{d=1}^{D}(\xi_d+\xi_d^\star)-\sum_{d=1}^{D}\mu_d(\epsilon+\xi_d-y_d+\mathbb{E}\left[\boldsymbol{\eta}^\top\overline{Z}_d\right])\\&-\sum_{d=1}^{D}(\mu_d^\star(\epsilon+\xi_d^\star+y_d-\mathbb{E}\left[\eta^\top\overline{Z}_d\right])+v_d\xi_d+v_D^\star\xi_d^\star)\\&-\sum_{d=1}^{D}\sum_{n=1}^{N}c_{d,n}(\sum_{t=1}^{T}\phi_{d,n,t}-1)\end{aligned}\tag{3}$$

ここで，$c_{d,n}$ は制約 $\sum_{t=1}^{T}\phi_{d,n,t}=1$ に対するラグランジュ乗数である．この L^r を各パラメータに関して最適化することにより更新式を得る．

E-Step：

- γ に関して L^r を最適化：γ は α と ϕ から決定する．

$$\boldsymbol{\gamma}_d=\boldsymbol{\alpha}+\sum_{n=1}^{N_d}\boldsymbol{\phi}_{d,n}\tag{4}$$

- ϕ に関して L^r を最適化：$\partial L^r/\partial\boldsymbol{\phi}_{d,n}=0$ とし，次式が得られる．

$$\begin{aligned}\boldsymbol{\phi}_{d,n}\propto&\exp(\mathbb{E}\left[\log\boldsymbol{\theta}_{\mathrm{d}}\mid\boldsymbol{\gamma}_{\mathrm{d}}\right]+\log p(w_{d,n}\mid\boldsymbol{\beta})\\&+\frac{y_d}{N_d\delta^2}\mathbb{E}\left[\boldsymbol{\eta}\right]\\&-\frac{2\mathbb{E}\left[\boldsymbol{\eta}^\top\boldsymbol{\phi}_{d,-n}\boldsymbol{\eta}\right]+\mathbb{E}\left[\boldsymbol{\eta}\circ\boldsymbol{\eta}\right]}{2N_d^2\delta^2}\\&+\frac{\mathbb{E}\left[\boldsymbol{\eta}\right]}{N_d}(\mu_d-\mu_d^\star))\end{aligned}\tag{5}$$

なお，$\boldsymbol{\phi}_{d,-n}:=\sum_{i\neq n}\boldsymbol{\phi}_{d,i}$ であり，単語 $\boldsymbol{\phi}_{d,n}$ 以外の ϕ の総和を表す．$\boldsymbol{\eta}\circ\boldsymbol{\eta}$ はアダマール積であり，$\boldsymbol{\eta}$ 同士の各要素の積からなるベクトルである．

- $q(\boldsymbol{\eta})$ に関して L^r を最適化：A を，各行がベクトル $\overline{Z}_d^\top$ からなる $D\times T$ 行列と定義する．$\partial L^r/\partial q(\boldsymbol{\eta})=0$ として，次式を得る

$$\begin{aligned}q(\boldsymbol{\eta})=\frac{p_0(\boldsymbol{\eta})}{X}\exp\big(&\boldsymbol{\eta}^\top\sum_{\mathrm{d}=1}^{\mathrm{D}}(\mu_{\mathrm{d}}-\mu_d^\star+\frac{y_d}{\delta^2})\mathbb{E}\left[\overline{Z}_d\right]\\&-\boldsymbol{\eta}^\top\frac{\mathbb{E}\left[A^\top A\right]}{2\delta^2}\boldsymbol{\eta}\big)\end{aligned}\tag{6}$$

また，$\mathbb{E}\left[A^\top A\right]=\sum_{d=1}^{D}\mathbb{E}\left[\overline{Z}_d\overline{Z}_d^\top\right]$，$\mathbb{E}\left[\overline{Z}_d\overline{Z}_d^\top\right]=1/N_d^2(\sum_{n=1}^{N_d}\sum_{m\neq n}\boldsymbol{\phi}_{d,n}\boldsymbol{\phi}_{d,m}^\top+\sum_{n=1}^{N_d}\mathrm{diag}\{\boldsymbol{\phi}_{d,n}\})$，$X$ は定数である．得られた $q(\boldsymbol{\eta})$ を L^r に代入することによって，以下の双対問題が得られる．

$$\max_{\boldsymbol{\mu},\boldsymbol{\mu}^\star}-\frac{1}{2}\boldsymbol{a}^\top\Sigma\boldsymbol{a}-\epsilon\sum_{d=1}^{D}(\mu_d+\mu_d^\star)+\sum_{d=1}^{D}y_d(\mu_d-\mu_d^\star)\tag{7}$$

ここで，$\boldsymbol{\mu}:=\{\mu_1,\ldots,\mu_D\}$，$\boldsymbol{\mu}^\star:=\{\mu_1^\star,\ldots,\mu_D^\star\}$ である．E を単位行列とすると，$q(\boldsymbol{\eta})$ の $K\times K$ 共分散行列 $\Sigma=\left(E+1/\delta^2\mathbb{E}\left[A^\intercal A\right]\right)^{-1}$，$\boldsymbol{a}=\sum_{d=1}^{D}(\mu_d-\mu_d^\star+y_d/\delta^2)\mathbb{E}\left[\overline{Z}_d\right]$ であり，この双対問題を SVM-light[1] などのソルバーによって解くことで $q(\boldsymbol{\eta}),\mu,\mu^\star$ を得る．

M-Step：β と δ^2 の更新式は以下の通りである．

- β に関して L^r を最適化：

$$\beta_{t,w}\propto\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{I}(w_{d,n}=w)\phi_{d,n,t}\tag{8}$$

$\mathbb{I}(w_{d,n}=w)$ は，文書 d における単語 n の語彙が w である場合にのみ $\beta_{t,w}$ に加算することを意味する．

- δ^2 に関して L^r を最適化：

$$\delta^2=\frac{1}{D}(\boldsymbol{y}^\top\boldsymbol{y}-2\boldsymbol{y}^\top\mathbb{E}\left[A\right]\mathbb{E}\left[\boldsymbol{\eta}\right]+\mathbb{E}\left[\boldsymbol{\eta}^\top\mathbb{E}\left[A^\top A\right]\boldsymbol{\eta}\right])\tag{9}$$

なお，$\mathbb{E}\left[\boldsymbol{\eta}^\top\mathbb{E}\left[A^\top A\right]\boldsymbol{\eta}\right]=\mathrm{tr}(\mathbb{E}\left[\mathrm{A}^\top\mathrm{A}\right]\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^\top\right])$ であり，tr は行列の対角成分の和を表す．

2.1.2 分類問題を想定した最大マージントピックモデル

離散値ラベル $y\in\{1,\ldots,M\}$ を持つ文書データを扱う MedLDA Classification (MedLDA-Cla) について説明する．MedLDA-Cla では $y_d|\boldsymbol{z}_d,\boldsymbol{\eta}=\underset{y\in\{1,\ldots,M\}}{\arg\max}\,\mathbb{E}\left[\boldsymbol{\eta}^\top\mathbf{f}(y,\overline{Z}_d)|\boldsymbol{w}_d,\boldsymbol{\alpha},\boldsymbol{\beta}\right]$ とし，以下のような最適化問題が定義される．

P2(MedLDA-Cla):

$$\min_{q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}),\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}\mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}))+C\sum_{d=1}^{D}\xi_d$$

$$\text{s.t. }\forall d:\begin{cases}\widehat{y}\neq y_d\\\mathbb{E}\left[\boldsymbol{\eta}^\top\Delta\mathbf{f}_d(\widehat{y})\right]\geq 1-\xi_d\\\xi_d\geq 0\end{cases}$$

制約式中の $\widehat{y}$ はラベルの予測値，y_d は真値である．ξ は訓練データの誤差を吸収する程度を示すスラック変数であり，文書ごとに設定する．ここからは MedLDA-Cla の更新式の導出を行う．目的関数において，

$$\begin{aligned}\mathcal{L}(q(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\eta}))=&-\mathbb{E}_q\left[\log p(\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{W}\mid\boldsymbol{\alpha},\boldsymbol{\beta})\right]\\&-\mathcal{H}(q(\boldsymbol{Z},\boldsymbol{\Theta}))+KL(q(\boldsymbol{\eta})\parallel p_0(\boldsymbol{\eta}))\end{aligned}\tag{10}$$

[1] http://svmlight.joachims.org/

$$\Delta \mathbf{f}_d(\widehat{y}) := \mathbf{f}(y_d, \overline{Z}_d) - \mathbf{f}(\widehat{y}, \overline{Z}_d) \tag{11}$$

である．$\mathbf{f}(y, \overline{Z}_d)$ は，$(y-1)T+1$ から yT の要素がベクトル $\overline{Z}_d$ であり他の要素が 0 であるような特徴ベクトルである．

KL は 2 つの確率分布の差異を表すカルバック・ライブラー情報量であり，次式で表される．

$$KL(q(\boldsymbol{\eta}) \parallel p_0(\boldsymbol{\eta})) = \int q(\boldsymbol{\eta}) \log \frac{q(\boldsymbol{\eta})}{p_0(\boldsymbol{\eta})} \mathrm{d}\boldsymbol{\eta} \tag{12}$$

MedLDA-Reg と同様に最適化問題 P2 に関しても変分近似を行い，q についての条件付独立性を与える．

$$q(\boldsymbol{Z}, \boldsymbol{\Theta} \mid \boldsymbol{\gamma}_d, \boldsymbol{\phi}) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}) \prod_{n=1}^{N_d} q(z_{d,n} \mid \boldsymbol{\phi}_{d,n}) \tag{13}$$

また，$\mathbb{E}[Z_{d,n}] = \boldsymbol{\phi}_{d,n}$，$\mathbb{E}\left[\boldsymbol{\eta}^\top \mathbf{f}(y, \overline{Z}_d)\right] = \mathbb{E}[\boldsymbol{\eta}]^\top \mathbf{f}(y, 1/N_d \sum_{n=1}^{N_d} \boldsymbol{\phi}_{d,n})$ である．そして目的関数に制約式を含めたラグランジュ関数 L^c を次のように定義し，L^c を各パラメータに関して最適化することで更新式を得る．なお，γ, β に関しては MedLDA-Reg と更新式が同じであるため省略する．

$$\begin{aligned} L^c = {} & \mathcal{L}(q(\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\eta})) + C \sum_{d=1}^{D} \xi_d \\ & - \sum_{d=1}^{D} v_d \xi_d - \sum_{d=1}^{D} \sum_{\widehat{y} \neq y_d} \mu_d(\widehat{y}) (\mathbb{E}\left[\boldsymbol{\eta}^\top \Delta \mathbf{f}\, d(\widehat{y})\right] + \xi_d - 1) \\ & - \sum_{d=1}^{D} \sum_{n=1}^{N_d} c_{d,n} (\sum_{t=1}^{T} \phi_{d,n,t} - 1) \end{aligned} \tag{14}$$

E-Step：

- ϕ に関して L^c を最適化：$\partial L^c / \partial \boldsymbol{\phi}_{d,n}$ とし，次式が得られる．

$$\begin{aligned} \boldsymbol{\phi}_{d,n} \propto & \exp(\mathbb{E}[\log \boldsymbol{\theta}_{\mathrm{d}} \mid \boldsymbol{\gamma}_{\mathrm{d}}] + \log p(w_{d,n} \mid \boldsymbol{\beta}) \\ & + \frac{1}{N_d} \sum_{\widehat{y} \neq y_d} \mu_d(\widehat{y}) \mathbb{E}\left[\boldsymbol{\eta}_{y_d} - \boldsymbol{\eta}_{\widehat{y}}\right]) \end{aligned} \tag{15}$$

  最初の 2 項は MedLDA-Reg と同様である．

- $q(\boldsymbol{\eta})$ に関して L^c を最適化：

$$q(\boldsymbol{\eta}) = \frac{1}{X} p_0(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mu_\eta) \tag{16}$$

  ただし，$\mu_\eta = \sum_{d=1}^{D} \sum_{\widehat{y} \neq y_d} \mu_d(\widehat{y}) \mathbb{E}[\Delta \mathbf{f}_d(\widehat{y})]$.

2.2 双対分解

双対分解 (dual decomposition)[5] は，複雑な目的関数を効率的に求める手法である．直接的に求めることが困難な目的関数をいくつかの関数に分割でき，それぞれの関数の最適解が効率的に求まる場合に適用可能である．効率的に解くことができない次の関数を対象として，双対分解の例を示す．

$$\arg\max_{y} \; f(y) + h(y) \tag{17}$$

$\arg\max_y f(y)$，$\arg\max_y h(y)$ は効率的に求まると仮定する．このとき，次の問題は上の問題と同じ意味を持つ．

$$\arg\max_{y,z} \; f(z) + h(y) \tag{18}$$

$$\text{s.t. } y = z \tag{19}$$

この問題の解を $L^\star$ とする．そして，この問題に対してラグランジュ緩和を適用する．

$$L(u, y, z) = f(z) + h(y) + u(y - z) \tag{20}$$

u はラグランジュ乗数である．次に $L(u, y, z)$ に関して最大値をとるものを考える．

$$\begin{aligned} L(u) &= \max_{y,z} L(u, y, z) \\ &= \max_{z} (f(z) - uz) + \max_{y} (h(y) + uy) \end{aligned} \tag{21}$$

この関数は $y = z$ の制約を持たないため，最初の問題より広い解空間を持ち，$L^\star \leq L(u)$ が成り立つ．これにより，元の最適化問題の上限を与えている．よって，双対定理により以下が成り立つ．

$$L^\star = \min_{u} L(u) \tag{22}$$

$\min_u L(u)$ は凸関数であるので，u に関する勾配を求めることができれば，勾配降下法により最適化できる．よって，劣微分の 1 つである d_u は次のように求めることができる．

$$d_u = y^\star - z^\star \tag{23}$$

$$z^\star = \arg\max_{z} \; f(z) - uz \tag{24}$$

$$y^\star = \arg\max_{y} \; f(y) + uy \tag{25}$$

そして，勾配法に基づき以下のように u を更新する．

$$u \leftarrow u - \nu(y^\star - z^\star) \tag{26}$$

ν はステップ幅である．この更新を繰り返して $L(u)$ を小さくし，$y^\star = z^\star$ となる時が主問題と双対問題の値が一致したときなので，最適解を求めることができる．

3 双対分解を利用したマルチタスク最大マージントピックモデル

3.1 モデルの定義

2.1 節で述べたように，連続値または離散値の付加情報を持つ文書データの解析を行うためには MedLDA を利用すればよい．しかし，MedLDA では連続値と離散値の両方の付加情報を持つ文書データの解析を行うことができない．この問題を解決する為に，我々は双対分解を利用したマルチタスク最大マージントピックモデル (Multi-task MedLDA：MultiMedLDA) を提案する．MultiMedLDA は複数種類のラベルが付与された文書に対して適用可能なモデルであり，双対分解を利用して MedLDA を拡張している．以下に MultiMedLDA の生成過程を示す．

1. 文書 $d(d \in 1,\ldots,D)$ に対して，$\boldsymbol{\theta}_d \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ を選択．
2. 文書 d 内の N_d 個の単語 $w_{d,n}(n \in 1,\ldots,N_d)$ に対して，
   (a) トピック $z_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_d)$ を選択する．
   (b) 単語 $w_{d,n} \sim \mathrm{Multinomial}(\boldsymbol{\beta}_t)$ $(t = z_{d,n})$ を選択する．
3. D 個の文書に対して，連続値ラベル $y_d^r \sim F(\boldsymbol{\eta}^r, \boldsymbol{z}_d)$，離散値ラベル $y_d^c \sim F(\boldsymbol{\eta}^c, \boldsymbol{z}_d)$ を選択する．

なお，$\boldsymbol{\eta}^r, \boldsymbol{\eta}^c$ は重み係数である．
MultiMedLDA のグラフィカルモデルを図 2 に示す．

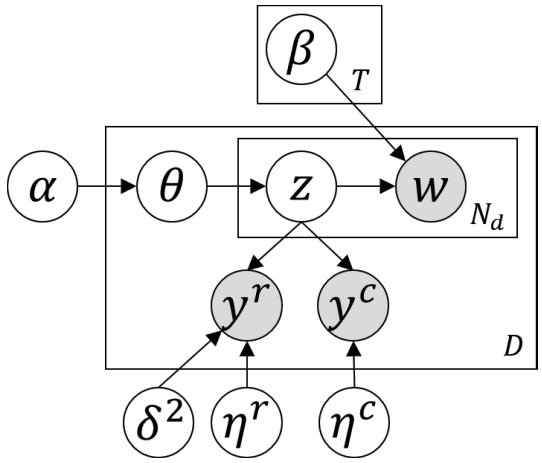


図 2: MultiMedLDA のグラフィカルモデル

3.2 モデルの推定

連続値ラベル $y^r \in \mathbb{R}$ と離散値ラベル $y^c \in \{1,\ldots,M\}$ が各文書に付加されている，データセットについて考える．このとき，MedLDA-Reg の最適化問題と MedLDA-Cla の最適化問題を統合することによって，以下のような最適化問題を定義することができる．なお，目的関数第 1，2 項が回帰タスクの目的関数，目的関数第 3，4 項が分類タスクの目的関数である．同様に制約式第 1，2，3 行が回帰タスクの制約式，制約式第 4，5 行が分類タスクの制約式である．第 6 行は双対分解のための制約である．以下，回帰タスクに関する変数は上付き文字の r，分類タスクに関する変数は上付き文字の c で表す．

$$\mathrm{P3(MultiMedLDA)}: \min_{q(\boldsymbol{Z}^r,\boldsymbol{\Theta}^r,\boldsymbol{\eta}^r),q(\boldsymbol{Z}^c,\boldsymbol{\Theta}^c,\boldsymbol{\eta}^c),\alpha,\beta,\delta^2,\boldsymbol{\xi}^r,\boldsymbol{\xi}^{r\star},\boldsymbol{\xi}^c}$$

$$\mathcal{L}^r(q(\boldsymbol{Z}^r,\boldsymbol{\Theta}^r,\boldsymbol{\eta}^r)) + C^r\sum_{d=1}^{D}(\xi_d^r + \xi_d^{r\star})$$

$$+ \mathcal{L}^c(q(\boldsymbol{Z}^c,\boldsymbol{\Theta}^c,\boldsymbol{\eta}^c)) + C^c\sum_{d=1}^{D}\xi_d^c$$

$$\text{subject to}\,\forall d \begin{cases} y_d^r - \mathbb{E}\left[{\boldsymbol{\eta}^r}^\top \overline{Z}_d\right] \le \epsilon + \xi_d^r \\ -y_d^r + \mathbb{E}\left[{\boldsymbol{\eta}^r}^\top \overline{Z}_d\right] \le \epsilon + \xi_d^{r\star} \\ \xi_d \ge 0, \xi_d^\star \ge 0 \\ \widehat{y}^c \ne y_d^c : \mathbb{E}\left[{\boldsymbol{\eta}^c}^\top \Delta\mathbf{f}_d(\widehat{y}^c)\right] \ge 1 - \xi_d^c \\ \xi_d^c \ge 0 \\ \boldsymbol{\phi}_d^r = \boldsymbol{\phi}_d^c \end{cases}$$

$\xi^r, \xi^{r\star}, \xi^c$ はそれぞれ訓練データの誤差を吸収する程度を示すスラック変数，ϵ は許容誤差である．$\boldsymbol{\phi}_d^r := \left\{\boldsymbol{\phi}_{d,1}^r,\ldots,\boldsymbol{\phi}_{d,N_d}^r\right\}$，$\boldsymbol{\Phi}^r := \{\boldsymbol{\phi}_1^r,\ldots,\boldsymbol{\phi}_D^r\}$，$\boldsymbol{\phi}_d^c := \left\{\boldsymbol{\phi}_{d,1}^c,\ldots,\boldsymbol{\phi}_{d,N_d}^c\right\}$，$\boldsymbol{\Phi}^c := \{\boldsymbol{\phi}_1^c,\ldots,\boldsymbol{\phi}_D^c\}$ である．

以下では回帰タスクに関する目的関数第 1，2 項を $\mathcal{L}(R)$，分類タスクに関する目的関数第 3，4 項を $\mathcal{L}(C)$ とする．この最適化問題に対してラグランジュ緩和を行い，次の最適化問題を得る．なお，簡単のため制約式は省略している．

$$L(\boldsymbol{U},\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c) = \mathcal{L}(R)\boldsymbol{I} + \mathcal{L}(C)\boldsymbol{I} + \boldsymbol{U}\circ(\boldsymbol{\Phi}^c - \boldsymbol{\Phi}^r)$$

$\boldsymbol{I}$ は全ての要素が 1 であるベクトルである．$\boldsymbol{U} := \{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_D\}$，$\boldsymbol{u}_d := \{\boldsymbol{u}_{d,1},\ldots,\boldsymbol{u}_{d,N_d}\}$ であり，$\boldsymbol{u}_{d,n}$ は $\boldsymbol{\phi}_{d,n}^r$，$\boldsymbol{\phi}_{d,n}^c$ に対応するラグランジュ乗数を表す．次に $L(\boldsymbol{U},\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c)$ を最小化する $\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c$ を考える．

$$\begin{aligned} L(\boldsymbol{U}) &= \min_{\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c} L(\boldsymbol{U},\boldsymbol{\Phi}^r,\boldsymbol{\Phi}^c) \\ &= \min_{\boldsymbol{\Phi}^r}(\mathcal{L}(R) - \boldsymbol{U}\circ\boldsymbol{\Phi}^r) + \min_{\boldsymbol{\Phi}^c}(\mathcal{L}(C) + \boldsymbol{U}\circ\boldsymbol{\Phi}^c) \end{aligned} \tag{27}$$

この関数には最適化問題 P3 の制約式第 6 行にある $\boldsymbol{\phi}_d^r = \boldsymbol{\phi}_d^c$ が考慮されていないので，より広い解空間を持つ．これにより，$L^\star \geq L(\boldsymbol{U})$ が成り立つので，最適化問題 P3 の下限を与えている．また，双対定理より $L^\star = \max_{\boldsymbol{U}} L(\boldsymbol{U})$ が成り立つ．よって，$L(\boldsymbol{U})$ の劣微分の 1 つである $\boldsymbol{d}_U$，および $\boldsymbol{\Phi}^{r\star}$，$\boldsymbol{\Phi}^{c\star}$，ラグランジュ乗数 $\boldsymbol{U}$ は以下のようになる．

$$\boldsymbol{d}_U = \boldsymbol{\Phi}^{c\star} - \boldsymbol{\Phi}^{r\star} \tag{28}$$

$$\boldsymbol{\Phi}^{r\star} = \arg\min_{\boldsymbol{\Phi}^r} \quad \mathcal{L}(R)\boldsymbol{I} - \boldsymbol{U} \circ \boldsymbol{\Phi}^r \tag{29}$$

$$\boldsymbol{\Phi}^{c\star} = \arg\min_{\boldsymbol{\Phi}^c} \quad \mathcal{L}(C)\boldsymbol{I} + \boldsymbol{U} \circ \boldsymbol{\Phi}^c \tag{30}$$

$$\boldsymbol{U} \leftarrow \boldsymbol{U} - \nu(\boldsymbol{\Phi}^{c\star} - \boldsymbol{\Phi}^{r\star}) \tag{31}$$

なお，ν はステップ幅であり，本研究では反復回 S の逆数を用いている．回帰タスクと分類タスクで潜在トピック $\boldsymbol{\Phi}^r$，$\boldsymbol{\Phi}^c$ の推定を行った後，この更新を繰り返すことで下限 $L(\boldsymbol{U})$ の最大化を行う．そして $\boldsymbol{\Phi}^{r\star} = \boldsymbol{\Phi}^{c\star}$ となった時が主問題と双対問題の値が一致したときなので最適解に到達したことが保証される．

これにより得られた $\boldsymbol{\Phi}^{r\star}$，$\boldsymbol{\Phi}^{c\star}$ をそれぞれの最適化問題に与えなおすことによって，もう一方の影響を考慮した潜在トピックの推定が可能となる．なお，MultiMedLDA の最適化問題は，MedLDA とは異なり制約式に $\boldsymbol{\phi}_d^r = \boldsymbol{\phi}_d^c$ が追加される．これにより，$-\boldsymbol{U} \circ \boldsymbol{\Phi}^r$，$\boldsymbol{U} \circ \boldsymbol{\Phi}^c$ の項が偏微分をした後にも残る．よって $\boldsymbol{\phi}$ の更新式は以下のようになる．

$$\begin{aligned}\boldsymbol{\phi}_{d,n}^r \propto \exp\Big(&\mathbb{E}\left[\log \boldsymbol{\theta}_{\mathrm{d}}^{\mathrm{r}} \mid \boldsymbol{\gamma}_{\mathrm{d}}^{\mathrm{r}}\right] + \log p(w_{d,n} \mid \boldsymbol{\beta}^r) \\ &- \frac{2\mathbb{E}\left[\boldsymbol{\eta}^{r\top} \boldsymbol{\phi}_{d,-n}^r \boldsymbol{\eta}^r\right] + \mathbb{E}\left[\boldsymbol{\eta}^r \circ \boldsymbol{\eta}^r\right]}{2N_d^2\delta^2} \\ &+ \frac{\mathbb{E}\left[\boldsymbol{\eta}^r\right]}{N_d}(\mu_d^r - \mu_d^{r\star})) + \frac{y_d}{N_d\delta^2}\mathbb{E}\left[\boldsymbol{\eta}^r\right] - \boldsymbol{u}_{d,n}\Big)\end{aligned} \tag{32}$$

$$\begin{aligned}\boldsymbol{\phi}_{d,n}^c \propto \exp\Big(&(\mathbb{E}\left[\log \boldsymbol{\theta}_{\mathrm{d}}^{\mathrm{c}} \mid \boldsymbol{\gamma}_{\mathrm{d}}^{\mathrm{c}}\right] + \log p(w_{d,n} \mid \boldsymbol{\beta}^c) \\ &+ \frac{1}{N_d} \sum_{\widehat{y} \neq y_d^c} \mu_d^c(\widehat{y}) \mathbb{E}\left[\boldsymbol{\eta}_{y_d^c}^c - \boldsymbol{\eta}_{\widehat{y}}^c\right]) + \boldsymbol{u}_{d,n}\Big)\end{aligned} \tag{33}$$

4 評価実験

4.1 データセット

本研究では，データセットとして東洋経済新報社が発行する会社四季報 [1] を使用した．これは四半期ごとに発表される経済記事であり，上場企業 3675 社 (2017 年度新春版) の，企業名をはじめとした上場コード，業種，営業利益，株価，短評などが載っている．2014 年度新春版から 2017 年度新春版までの 13 四半期分のデータを使用した．各企業には上場する際に登録された 32 種類の業種のうち 1 つが選ばれているが，その中で登録企業数の多い上位 10 種類 (サービス業，情報・通信業，小売業，卸売業，電気機器，機械，化学，建設業，食料品，輸送用機器) の業種の企業データを使用した．各業種の企業数には偏りが存在するため，1 業種につき 89 企業を無作為に選択している．

表 1: 四季報データセットの概要

年	2014	2015	2016
企業数	890	890	890
総単語数	164931	165272	164801
一企業あたりの単語数	185.3	185.7	185.2
語彙数	2862		
業種（離散値ラベル）の種類	10		

ある年の新春版から秋版までの短評を一つにまとめたものを文書データ，業種を離散値ラベル，文書データの翌年新春版の ROE（Return On Equity，自己資本利益率）を連続値ラベルとして使用した．また，2014 年と 2015 年のいずれかで 3 文書未満にしか出現しない低頻度語を除外している．なお，文書データは MeCab[2] を用いて形態素解析を行い，助詞や接続詞といった機能語を除外している．ROE の値は，年毎に平均が 0，分散が 1 になるよう正規化した．以上の処理を行ったデータセットの情報を表 1 に示す．

4.2 実験設定

連続値ラベルが未知であるという設定の下で，その値を予測する実験を行う．離散値ラベルを考慮しないモデルである MedLDA-Reg をベースラインとし，提案手法の MultiMedLDA と比較する．

学習用データでモデル構築を行い，モデルパラメータ β，η を得る．その後，学習で得られた β，η を用いてテストデータの潜在トピックを推定し連続値ラベルを予測する．テストデータの潜在トピック推定は連続値ラベルを隠した状態で行わなければならないため，MedLDA の実験では教師なしトピックモデルの一種である Latent Dirhlet Allocation（LDA）[6]，MultiMedLDA の実験では MedLDA-Cla を用いて行った．

2016 年のデータをテストデータとし，基本的に 2015 年のデータを学習用データとして用いる．モデル学習時の各パラメータの初期値は乱数で決定する．しかし，この初期値の与え方にも何らかの知見を活用したい．

[1] https://store.toyokeizai.net/cddvd/shikiho_cd/

[2] http://taku910.github.io/mecab/

そこで，2014 年のデータで学習して得られた β と η を，2015 年のデータでの学習時の初期値とする条件での実験も行った．この条件での MedLDA と MultiMedLDA をそれぞれ MedLDA-Seq，MultiMedLDA-Seq と呼び，初期値を乱数で決定する条件設定をそれぞれ MedLDA-Rand，MultiMedLDA-Rand と呼ぶことにする．MedLDA-Seq と MultiMedLDA-Seq において，2014 年のデータでの学習時の各パラメータの初期値は乱数で決定した．

ハイパーパラメータは $\alpha_t = 0.1 \quad \forall t$，損失パラメータは $l = 1$，正則化パラメータ $C^r = 1$，許容誤差 $\epsilon = 0.1$，ラグランジュ乗数 $\boldsymbol{U}$ を更新する際の反復回数 $S = 20$ に設定した．また，MultiMedLDA の学習において，MedLDA-Reg と MedLDA-Cla の特徴を活かすため，双方の計算結果を影響させない burn-in period を 5 回目の反復までに設定している．これにより，回帰タスクと分類タスクの特徴を活かした状態で潜在トピックの統合が図られている．学習の反復回数は 100 回とした．トピック数は $T \in \{20, 40, 60, 80, 100\}$ の 5 通り，MultiMedLDA における分類タスクの正則化パラメータは $C^c \in \{0.0625, 0.25, 1, 4, 16\}$ の 5 通りの条件で実験した．

4.3 評価尺度

4.3.1 Root Mean Squared Error : RMSE

Root Mean Squared Error(以下 RMSE) はモデルの予測能力を表す指標のひとつである．モデルの予測値と真値から算出される相対的な評価指標である．RMSE は以下の式で表される．

$$RMSE = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\widehat{y}_d - y_d)^2} \tag{34}$$

$\widehat{y}_d$ はモデルの予測値であり，y_d は真値である．予測値が真値から離れているほど大きい値をとるため，0 に近いほど優れている．

4.3.2 Mean Absolute Error : MAE

Mean Absolute Error（以下 MAE）も，モデルの予測値と真値から算出される相対的な評価指標である．MAE は以下の式で表される．

$$MAE = \frac{1}{D}\sum_{d=1}^{D}|\widehat{y}_d - y_d| \tag{35}$$

$\widehat{y}_d$ はモデルの予測値であり，y_d は真値である．予測値が真値から離れているほど大きい値をとるため，0 に近いほど優れている．

RMSE は MAE に比べて大きな誤差を重視する性質があり，MAE はその点で安定的な指標である．

4.4 実験結果

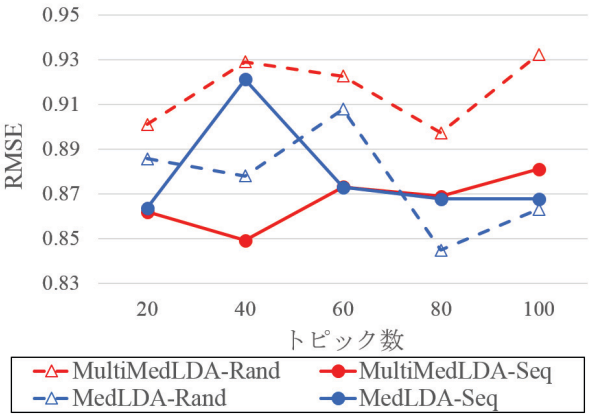


図 3: 各手法のトピックごとの RMSE

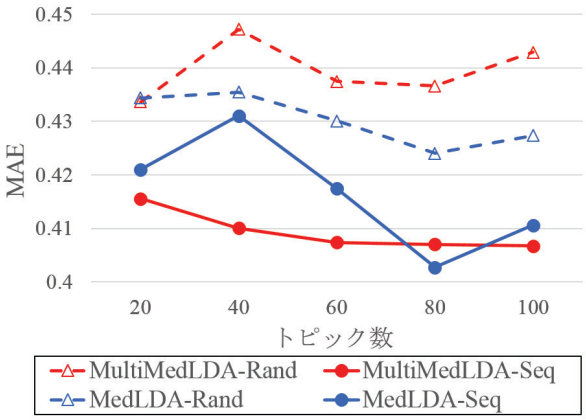


図 4: 各手法のトピックごとの MAE

結果のグラフを図 3，図 4 に示す．MultiMedLDA に関しては各トピック数において C^c の値が異なる 5 通りの実験を行ったが，その中で最も結果が良かった場合のものをプロットしている．この図から分かるように，MedLDA においても MultiMedLDA においても，初期値をランダムに与える-Rand 版よりも，過去の β，η で初期化する-Seq 版の方が高い精度を示した．各トピック数における最良の C^c を選択できれば，実験した 4 手法の内では提案手法の MultiMedLDA-Seq が安定的に最良の性能を示している．一方で MultiMedLDA-Rand は最も悪い結果となっており，既存手法の MedLDA に

比べて提案手法の MultiMedLDA は，初期値の与え方に性能が大きく左右されることが分かった．

5 むすび

本稿では，文書データをもとに企業の財務指標を予測する課題に取り組み，トピックモデルの一種である MultiMedLDA を提案した．このモデルは，双対分解を利用して既存手法の MedLDA を拡張したものである．MultiMedLDA の特徴は，離散値と連続値という 2 種類の数値ラベルを同時に持つ文書を扱うことができる点であり，文書に加えて既知ラベルの情報も考慮しつつ未知ラベルを推定することができる．『会社四季報』のデータを用いて評価実験を行った結果，既知ラベルを想定しない既存手法である MedLDA よりも良い性能を示した．また，特に提案手法に対しては，モデルパラメータを学習する際の初期値の与え方が精度向上に重要であることが分かった．

本稿ではいわゆる closed test を行ったが，今後は交差検証により最適な超パラメータを決定した上でテストを行い，より厳密にモデルの汎化性能を評価したい．また，現時点では一つの文書に離散値ラベルと連続値ラベルが一つずつ付随すると仮定しているが，これを複数個ずつに拡張することでさらなる性能の向上を図る予定である．

謝辞

本研究を行うにあたり有益な助言を頂いた神戸大学大学院経済学研究科の羽森茂之教授と金京拓司教授に感謝する．本研究の一部は科学研究費補助金基盤研究 (B)(15H02703) の援助による．

参考文献

[1] David M Blei and Jon D. McAuliffe.: Supervised topic models, *Advances in neural information processing systems*, pp. 121–128, (2008)

[2] Jun Zhu, Amr Ahmed, and Eric P Xing.: MedLDA: Maximum Margin Supervised Topic Models, *Journal of Machine Learning Reasearch*, Vol. 13, pp. 2237–2278, (2012)

[3] Edgar Osuna, Robert Freund, and Federico Girosi.: Support Vector Machines: Training and Applications, *Proceedings SVPR'97*, (1997)

[4] Alex J Smola and Bernhard Schölkopf.: A tutorial on support vector regression, *Statistics and Computing*, Vol. 14, pp. 199–222, (2004)

[5] S. Sra, S. Nowozin and S. Wright, “Optimization for Machine Learning”, Neural information processing series, MIT Press (2012).

[6] David M Blei, Andrew Y Ng, and Michael I Jordan.: Latent Dirichlet Allocation,*Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, (2003)

LSTMネットワークによる財務データの回帰分析

Regression Analysis of Corporate Financial Data using LSTM Networks

城内　光平[1*]　　江口　浩二[1]　　金京　拓司[2]　　羽森　茂之[2]

[1] 神戸大学大学院システム情報学研究科
[1] Graduate School of Systems Informatics, Kobe University
[2] 神戸大学大学院経済学研究科
[2] Graduate School of Economics, Kobe University

Abstract: In the economic and financial fields, there is a growing interest in obtaining new knowledge from large quantities of data, such as corporate financial data and international exchange transactions. On the other hand, as a technical trend of recent data analysis, deep learning-based models have been successfully applied to various data, such as images, text, and audio. Especially, Recurrent Neural Network (RNN) and its extension of Long Short-Term Memory Network (LSTM) have been developed as deep learning for sequential data or time series. However, regardless of its importance, LSTM has not applied to corporate financial time series, such as in the Financial Statements Statistics of Corporations, to the best of my knowledge. In this research, considering the above-mentioned trends, we conduct regression analysis using LSTM for corporate financial time series. For experiments, we obtain the capital investment rate and other financial indicators, such as the cash flow ratio, for each target company from the Financial Statements Statistics of Corporations, and then use them as the objective and explanatory variables, respectively. By changing the number and types of explanatory variables used in the experiments, we evaluate the contribution of each explanatory variable to regression power to the objective variable at several time steps ahead. Furthermore, as baseline methods for the regression tasks, we evaluate the regression power of classical methods: Autoregressive Integrated Moving Averaging (ARIMA), and discuss the comparative evaluation with the LSTM approach.

1　はじめに

近年，情報技術の発達に伴い，日々多くの情報が収集，蓄積されている．その分野は多岐にわたり，枚挙に暇がないが，その一つに，経済・金融分野における企業財務データや国際為替取引情報が挙げられる．それらの分野の専門家達の間でも，これらの大量に蓄積されつつあるデータを活かして新たな知見を得ることに関心が高まりつつある．とりわけ，法人企業統計調査[1]等の企業財務データは，企業の経営方針や業界の動向，さらには社会の景況感など，実世界の情報を多く含む複雑な時系列データとなっている．

最近のデータ分析の技術的な側面に目を向けると，画像やテキスト，音声など様々なデータに対して深層学習を用いたモデルが適用され，その有効性が示されている．その中でも，系列データを扱う深層学習のモデルとして，再帰型ニューラルネットワーク (Recurrent Neural Network：RNN)[1]，およびその拡張である長短期メモリネットワーク (Long Short-Term Memory Network：LSTM)[2] が挙げられる．RNN は長期間に渡る時間依存性に対しては勾配消失あるいは勾配爆発の問題のために，上手くこれらの関係性を捉えることができないことが少なくない．これらの欠点を改善し，長期間に渡る時間依存性を考慮することのできるモデルとしてLSTM が提案されている．また応用例として，Google 社の音響モデリングシステム [3] をはじめ，複雑なデータセットに対して複数層の LSTM からなるアーキテクチャを適用しこれまでのモデルを越える成果を残している．しかしながら，先に述べた企業財務に関する時系列データに対する LSTM の適用事例は，その重要性に関わらず筆者の知る限りない．

*連絡先：神戸大学大学院システム情報学研究科
〒 657-0029 1 兵庫県神戸市灘区六甲台町 1-1
E-mail: shirouchi@cs25.scitec.kobe-u.ac.jp
[1]http://www.mof.go.jp/pri/reference/ssc/index.htm

本論文においては，以上に述べた動向を踏まえた上で，法人企業統計調査における企業の財務報告に対して，LSTM を用いた回帰分析を行う．本実験においては，法人企業統計調査より，設備投資率と，キャッシュフロー比率を初めとした種々の財務指標を算出し，それぞれ被説明変数および説明変数として用いている．特に，設備投資率とキャッシュフロー比率との間の関係性に専門家の注目が集まっている [6]，[7]．また，実験に用いる説明変数の数や種類を変化させることで被説明変数に対する回帰精度への影響を評価し，回帰分析の対象とする財務指標と他の財務指標との依存性を推定する．

比較手法として，古典的な時系列解析手法である自己回帰和分移動平均 (Autoregressive Integrated Moving Average; ARIMA)[8] による分析を行い，比較評価について議論する．

2 関連研究

本研究においては，企業の活動の状態を表す財務指標に対して，深層学習の手法を用いて分析，および予測に関する検討を行う．そこで，本章では時系列データの解析によく用いられる深層学習の手法である，再帰型ニューラルネットワーク (Recurrent Neural Network: RNN)，およびその拡張である長短期記憶メモリネットワーク (Long Short-Term Memory network: LSTM) を紹介する．

2.1 再帰型ニューラルネットワーク

系列データは個々の要素が順序を持ち，その並びに意味が隠されているようなデータのことである．系列の並びが時刻の並びとなっている系列データを特に時系列データと呼ぶ．系列データの例としては音声の波形，動画，文章 (単語列) などがある．それにより，系列データの分類問題においてはある時刻，もしくは位置におけるデータの前後におけるデータとの関係性を上手く考慮することができれば，結果が向上することが知られている．そして順伝播型ニューラルネットワークを系列データを扱うために拡張したモデルとして再帰型ニューラルネットワーク (Recurrent Neural Network: RNN)[1] がある．RNN は入力の系列 $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots$ から正解の系列 $\mathbf{y}^1, \mathbf{y}^2, \mathbf{y}^3, \ldots$ を推定する問題として定式化することができる．このとき時刻 t における出力 $\mathbf{y}^t$ はそれ以前の入力 $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots, \mathbf{x}^t$ の影響を受けていると考えることができる．

RNN は内部に有向な閉路を持つニューラルネットの総称である．RNN はこの構造により情報を一時的に記憶し，また振る舞いを動的に変化させることができる．

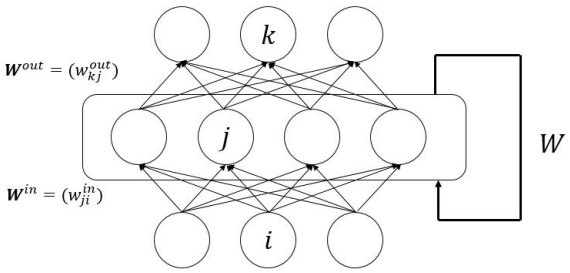


図 1: Graphical model of RNN.

これにより，系列データ中に存在する時間的な依存関係である「文脈」を捉えることができ，系列データに対して有効に処理することができる．RNN のグラフィカルモデルを図 1 に示す．

図 1 にあるように RNN は FFNN と同様の構造を持ち，ただし中間層のユニットの出力が自分自身に戻される「帰還路」を持っている．RNN の動作は各時刻 t につき 1 つの入力 $\mathbf{x}^t$ を受け取り，また同時に 1 つの出力 $\mathbf{y}^t$ を返すというものである．RNN は理論上，過去のすべての入力から 1 つの出力への写像を表現する．中間層に十分な数のユニットがあれば，任意の系列から系列への写像を，任意の精度で近似することができることが証明されている．

本研究では，出力層の活性化関数は恒等関数，誤差関数には二乗誤差関数を用いている．

2.1.1 順伝播計算

RNN における計算の過程を示す．ネットワークへの入力を $\mathbf{x}^t = (x_i^t)$，中間層への入出力をそれぞれ $\mathbf{u}^t = (u_j^t)$，$\mathbf{z}^t = (z_j^t)$，出力層ユニットへの入出力をそれぞれ $\mathbf{v}^t = (v_k^t)$，$\mathbf{y}^t = (y_k^t)$ とする．また目標出力を $\mathbf{d}^t = (d_k^t)$ とする．RNN の帰還路は，中間層の出力を自らの入力に戻し，この間の結合は全ユニット間で存在する．したがって，時刻 $t-1$ 中間層の任意のユニット j' から時刻 t 中間層の任意のユニット j へ重み $W_{jj'}$ の結合が存在する．よって，時刻 t における中間層の任意のユニット j への入力は以下のようになる．

$$u_j^t = \sum_i w_{ji}^{(in)} x_i^t + \sum_{j'} w_{jj'} z_{j'}^{t-1} \tag{1}$$

これを用いて，中間層の出力は活性化関数 f を用いて以下のように求めることができる．

$$z_j^t = f(u_j^t) \tag{2}$$

$$\mathbf{z}^t = \mathbf{f}(\mathbf{W}^{(in)}\mathbf{x}^t + \mathbf{W}\mathbf{z}^{t-1}) \tag{3}$$

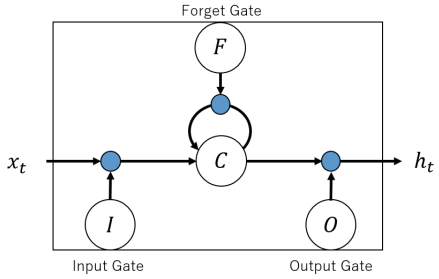


図 2: Memory unit of LSTM.

よって，出力層の入出力 $\mathbf{v}^t = (v_k^t)$，$\mathbf{y}^t = (y_k^t)$ は以下のように求められる．

$$v_k^t = \sum_j w_{kj}^{(out)} z_j^t \tag{4}$$

$$\mathbf{y} = \mathbf{f}^{(out)}(\mathbf{v}^t) = \mathbf{f}^{(out)}(\mathbf{W}^{(out)}\mathbf{z}^t) \tag{5}$$

2.2 長短期記憶メモリネットワーク

RNN は系列データの文脈を捉えて推定を行うことができる．このとき，捉えることのできる文脈の長さ，すなわち現時刻からどれだけ遠い過去の入力を出力に反映させることができるかは重要である．しかし，RNN で実際に出力に反映することができるのは高々過去の 10 時刻分程度であると言われている．この限界は勾配消失問題により生じている．層の数の多い深いネットワークにおいて，誤差逆伝播法によって勾配を計算するときに層をさかのぼるにつれて勾配の値が爆発的に大きくなるか，あるいは 0 に消滅してしまう．長期にわたる記憶を実現するために提案されたモデルが長短期記憶ネットワーク (LSTM) である．LSTM では上で述べたような RNN に対し，その中間層の各ユニットをメモリユニットと呼ぶ要素で置き換えた構造を持っている．その他の構造は通常の RNN と同様である．メモリユニットの構造を図 2 に示す．

中央にメモリセル (図中記号 C) があり，その周囲に入力ゲート，出力ゲート，忘却ゲートが配置されている．メモリセル C は状態 s_j^t を保持し，これを 1 時刻隔ててメモリセル自身に帰還させることで記憶を実現している．この帰還路には途中に忘却ゲートが挿入されており，ユニット F の出力がゲートの値 $g_j^{F,t} \in [0,1]$ となる．s_j^t に $g_j^{F,t}$ をかけたものが伝えられ，$g_j^{F,t}$ の値が 1 に近ければ現状態がそのまま記憶され，0 に近ければリセット (忘却) される．

メモリユニットへの外部からの入力は，外部からの入力と入力ゲートの値の積が入力される．ユニット I の出力がゲートの値 $g_j^{I,t} \in [0,1]$ となる．

メモリユニットからの外部への出力は，メモリセルと出力ゲートからの値の積が出力される．ユニット O の出力がゲートの値 $g_j^{O,t} \in [0,1]$ となる．ゲートの値が 1 に近ければメモリセルの出力は外部に伝達され，0 に近ければブロックされる．

これらの構造により短時間の記憶しか実現できないという RNN の限界を緩和することを目的としており，タイミングよくこれらのゲートの値を調整することにより，長い文脈を捉えたより高度な推定を実現する．

2.2.1 順伝播計算

時刻 t における入力ゲート，出力ゲート，忘却ゲートそれぞれの出力を $\mathbf{i}_t$，$\mathbf{o}_t$，$\mathbf{f}_t$，メモリセルの状態を $\mathbf{C}_t$，メモリユニットの出力を $\mathbf{h}_t$ とすると，それぞれ以下のように求めることができる．

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{b}_i) \\
\mathbf{C}_t &= \mathbf{i}_t \odot \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) + \mathbf{f}_t \odot \mathbf{C}_{t-1} \\
\mathbf{f}_t &= \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{b}_f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{V}_o\mathbf{C}_t + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{C}_t)
\end{aligned} \tag{6}$$

ここで，$\mathbf{W_i}, \mathbf{W_f}, \mathbf{W_c}, \mathbf{W_o}, \mathbf{U_i}, \mathbf{U_f}, \mathbf{U_c}, \mathbf{U_o}, \mathbf{V_o}$ はそれぞれ重み行列であり，$\mathbf{b_i}, \mathbf{b_f}, \mathbf{b_c}, \mathbf{b_o}$ はバイアスである．$\sigma(\cdot)$ はシグモイド関数，$\tanh(\cdot)$ はハイパボリックタンジェント関数である．

時刻 t における入力 $\mathbf{x}_t$ はメモリユニットに入力されると，入力ゲートの出力 $\mathbf{i}_t$ とのアダマール積がメモリセルに入力される．$\mathbf{i}_t$ は時刻 t における入力 $\mathbf{x}^t$，一時刻前の中間層の出力 h_{t-1}，バイアスによって求められる．また、時刻 t におけるメモリセルの状態 $\mathbf{C}_t$ は時刻 t における入力と時刻 $t-1$ のメモリセルの状態 $\mathbf{C}_{t-1}$ と忘却ゲートの出力 $\mathbf{f}_t$ の積の和となる．忘却ゲートの出力は $\mathbf{x}_t$，h_{t-1}，バイアスによって求められる．そして，メモリユニットの出力 $\mathbf{h}_t$ は $\tanh(\mathbf{C}_t)$ と出力ゲートの出力 $\mathbf{o}_t$ のアダマール積となる．出力ゲートの出力は $\mathbf{x}_t$，h_{t-1}，$\mathbf{C}^t$，バイアスによって求められる．

本論文においては，LSTM の重み更新式として RMSprop [4] を用いている．RMSprop は再帰型ニューラルネットワークにおいて定評のある重み更新アルゴリズムであり，以下の式で定義される．

$$\begin{aligned}
h_t &= \alpha h_{t-1} + (1-\alpha)\nabla E(\mathbf{w}^t)^2 \\
\eta_t &= \frac{\eta_0}{\sqrt{h_t} + \epsilon} \\
\mathbf{w}^{t+1} &= \mathbf{w} - \eta_t \nabla E(\mathbf{w}^t)
\end{aligned} \tag{7}$$

ここで，$\mathbf{w}$ は更新する重みであり，$E(\cdot)$ は誤差関数を表している．そして $\nabla E(\cdot)$ は誤差関数の勾配を表している．η が学習率である．本実験においては $\eta = 0.001$ に設定している．

3 実験

本節では，法人企業統計調査による企業の投資率および関連する指標の回帰分析に関する実験を行う．投資率と，その説明変数とする各財務比率との間の相関関係を調べるための実験設定として，時刻 $t+1$ における設備投資率を時刻 t における各財務比率を用いたニューラルネットワークによる回帰分析を行う．その際に，回帰に用いる説明変数の数および種類を変化させ，設備投資率の予測に対して有効な説明変数を推定する．最後に，各結果について評価する．本章において，説明変数の数を変化させる実験設定を特徴選択と呼ぶ．

3.1 データセット

本研究では，統計法に基づく基幹統計として「法人企業統計調査規則」(昭和 45 年大蔵省令 48 号) に基づいて，財務省により収集された法人企業統計調査を使用した．この調査は，営利法人等を調査対象とし，その中から無作為抽出により標本法人を選定している．総企業数は年次，および四半期別に異なるが 20000 社ほどとなっている．四半期別調査票データと年次別調査票データが存在し，本研究では一般業の四半期別調査票データの 2003 年 1 期から 2016 年 4 期までの 14 年間，56 期分のデータを使用している．まず 20000 社の内で対象の 14 年間の内に倒産またはその他の理由により調査対象から外れたものを除外した結果，対象の企業数は 2432 社となった．

法人企業統計には調査項目として，前期および当期における資本金や売上高，負債等を示した各企業の財務状況が報告されている．それらの企業のデータを用いて，種々の財務比率を算出し，説明変数として用いる．しかし，調査データのおいて多くの欠測値があり，算出することが困難な指標が多く存在した．したがって今回の実験においては，対象企業数全体の 50% 以上が欠測している財務比率については使用していない．過半数が欠測しているデータに関しては，その後補完処理を行ったとしても真の系列データとは大きく乖離していることが考えられるためである．財務比率を算出した際に 0 での除算などが発生した財務報告がなされている企業を除外した結果，最終的に実験に用いた企業数は 2330 社となった．最終的に実験に用いる欠測値を含む指標については，後述の補完処理を行い使用している．結果として，本研究において実際に使用している財務比率を示す．また以降はそれぞれの財務比率について以下に示すように $V1 - V12$ と呼ぶとする．

- V1: 総資本営業利益率
- V2: 総資本経常利益率
- V3: 売上高営業利益率
- V4: 売上高経常利益率
- V5: 総資本回転率
- V6: 有形固定資産回転率
- V7: 買掛金回転期間
- V8: 減価償却費
- V9: 労働装備率
- V10: キャッシュフロー率
- V11: 設備投資率
- V12: 総資本利益率 (ROA)

ここで，各財務比率は法人企業統計における財務営業比率の算式[2]を参考としている．また，キャッシュフロー率については以下の式で定義されている．

$$\text{キャッシュフロー率} = \frac{\text{当期純利益} + \text{減価償却費}}{\text{総資本}} \times 100 \tag{8}$$

3.1.1 データの補完処理

統計調査において，無回答や無記入により調査項目の一部の情報が得られない場合に欠測値が生じる．欠測を含むデータにおいて，観測された値のみを用いて推定を行う場合，欠測バイアスが生じることや，推定精度への悪影響が考えられる．また本研究で用いている，法人企業統計調査においても多くの欠測値が認められ，観測されているデータのみを用いて分析を行うことは困難であり補完処理を行うことが必要であった．また Table 1 に今回用いた企業データについて詳細を示す．補完処理については年次別データを用いて行った．四半期単位の財務指標の時系列 g_t において，四半期 j の値が欠測しているとする．その際には g_t と対応する，年次単位の財務指標の時系列 G_t を参照し，四半期 j が含まれる年次 J の値 G_J を用いる．

[2] http://www.mof.go.jp/pri/reference/ssc/results/calculation.htm

表 1: Overview of raw dataset.

	Missing rario[%]	AVE±STD
V1	0.03	0.010±0.029
V2	0.03	0.011±0.027
V3	0.04	0.031±0.654
V4	0.04	0.036±0.724
V5	0.04	0.275±0.242
V6	20.17	4.201±65.320
V7	6.41	6.828±39.400
V8	2.28	0.041±0.038
V9	20.25	57.499±498.640
V10	2.69	0.062±0.066
V11	7.04	0.008±0.028
V12	1.14	0.056±0.071

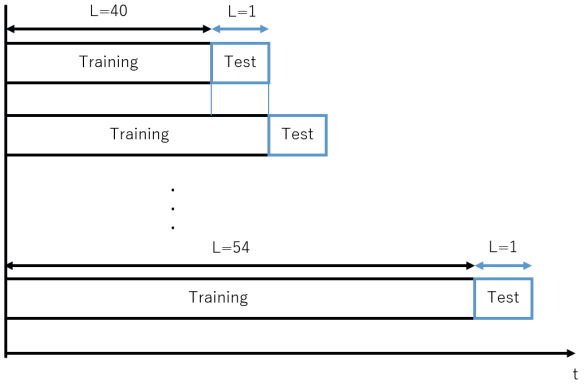


図 3: Overview of Experiment.

3.2 実験設定

財務指標間の関係性について，ある時刻 t の説明変数に当たる財務指標と時刻 $t+n$ における被説明変数に当たる財務指標への回帰タスクを行い結果を評価する．本論文では $n=1$ と設定し，1 時刻隔てた変数間の相関関係を考察する．比較実験として ARIMA モデルを用いた自己回帰を行い結果を比較する．学習期間として，40 期分のデータを与える実験から 54 期分のデータを入力として与え，その一つ先の期における被説明変数を予測する実験を行う．よって時間方向に 15 通りの実験を行い，予測対象の四半期の違いによる予測難易度の変化の比較実験もまた行った．概要を図 3 に示す．

予備実験は，本実験で時刻 t のデータから時刻 $t+1$ における被説明変数に対しての予測を行う実験設定においては時刻 $t-1$ のデータから時刻 t における被説明変数の予測を行う実験設定を予備実験とし，超パラメータの最適化を行った．探索範囲は中間層の数 L については $L \in \{1,2,3\}$，ユニット数 U については $U \in \{25,50,75,100,125\}$ である．なお，L については，全ての説明変数を用いた実験設定において $L \in \{1,2,3,4,5\}$ について実験を行ったが，探索した範囲において最適な値が $L=2$ となった．したがって，その他の説明変数の数を減じた実験設定においては $L \in \{1,2,3\}$ において行った．この実験設定においては，全財務比率を用いた実験およびキャッシュフロー比率のみを除外した実験についての予備実験を行い，そこで設定した超パラメータにより本実験を行った．その他の超パラメータについては，エポック数を 100，学習率を 0.001，バッチサイズを 10 に設定して実験を行った．

4 実験結果

結果を表 2, 図 4 に示す．それぞれの結果は学習期間が 40-54 の 15 種類の実験を行い，その MSE の平均値を示している．ARIMA モデルについては 40 期分のデータよりモデルを学習し，残りの四半期について予測を行った．本実験においてはキャッシュフロー比率を除外した実験において結果の悪化が確認された．来期の予測問題において，内部資金の利用可能性が設備投資の水準に作用しているという仮定に対して矛盾のない結果が観察された．

また，次に予測対象の四半期について注目した結果を表 3 に示す．これより，1-3 月の第 4 四半期の予測において誤差が大きい結果となった．これは，年度末における財務指標とその他の四半期における財務指標とで傾向が大きく異なっていることが考えられる．また財務分析において，どの期におけるデータかを示すダミー変数を追加することによって結果の向上を見込むことが考えられる．

5 おわりに

本論文では，企業の財務報告である法人企業統計に対して，設備投資率に対するその他の財務比率の相関を確認するために，LSTM を用いた回帰分析およびその比較手法として ARIMA モデルを用いて分析を行った．特定の財務比率を説明変数から除外することにより，相対的にその財務比率の影響力を評価することができると考えられる．特に設備投資率とキャッシュフロー比率との間の依存性に関心がある．実験では 1 時刻の先の被説明変数との分析を行い，結果として仮説と矛盾のない結果が示された．

今後の展望として，よりデータに則した処理を行うことが考えられる．例えば，説明変数に四半期の季節性を考慮したダミー変数を考慮することや，リーマン

表 2: Result of Experiment in the task of estimating the financial ratio of hidden companies in the one time step ahead.

Excluded financial ratio	MSE
V1	4.25E-04
V2	4.32E-04
V3	4.28E-04
V4	4.20E-04
V5	4.29E-04
V6	4.32E-04
V7	4.23E-04
V8	4.38E-04
V9	4.11E-04
V10	4.28E-04
Nothing excluded	4.20E-04
ARIMA	6.50E-04

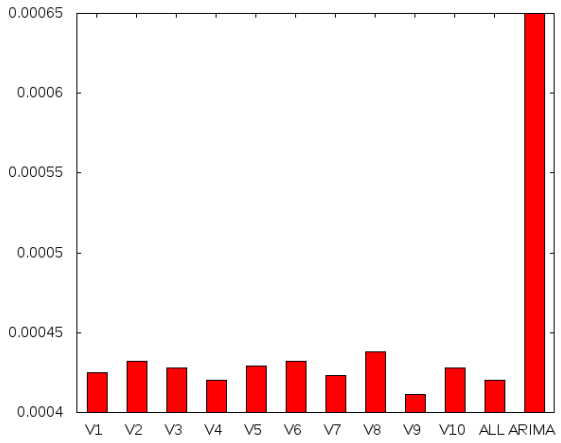


図 4: Result of Experiment in the task of estimating the financial ratio of hidden companies in the one time step ahead.

表 3: Quarterly Average Result of Experiment.

Target period	MSE
1(April-June)	4.02E-04
2(July-September)	3.15E-04
3(Octiber-December)	2.94E-04
4(January-March)	6.88E-04

ショックなど大きな事件が起こったことを表す変数を追加することが考えられる．経済データは社会動向により大きく変動することが考えられ，過去の系列情報だけでなく外因を考慮する必要があると考えられるからである．次に LSTM ネットワークについてもより最適化が可能であると考えられる．今回着目しなかった超パラメータについて最適化を行うことや，モデルの構造においてはシーケンス予測のための多くのアーキテクチャが提案されており今回のデータセットに対してより適切なアーキテクチャが存在している可能性がある．また深層構造を伴うネットワークを扱う際には，事前学習として Deep Belief Network を用いて事前学習を行うことなどにより性能の向上が期待できる．

謝辞

本研究の一部は科学研究費補助金基盤研究 (B)(15H02703) の援助による．

参考文献

[1] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur.: Recurrent neural network based language model, *Proceedings of the Annual Conference of International Speech Communication Association*, pp. 1045–1048, (2010)

[2] Sepp Hochreiter and Jurgen Schmidhuber.: Long short-term memory, *Neural computation*, pp. 1735–1780, (1997)

[3] Hasim Sak, Andrew Senior, and Franc，oise Beaufays.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling, *Proceedings of the Annual Conference of International Speech Communication Association*, (2014)

[4] Geoffrey Hinton.: Neural Networks for Machine Learning - Lecture 6a - Overview of mini-batch gradient descent.(2012)

[5] 桜井久勝，森脇敏雄．：財務比率による倒産リスク評価の有効性，*Discussion Paper Series*, (2016)

[6] 嶋恵一．：内部資金と投資 法人企業統計による企業規模別分析，財務省財務総合政策研究所「フィナンシャル・レビュー」平成 29 年第 2 (通巻第 130 号), (2017)

[7] 花崎正春，羽田徹也．：企業の投資行動の決定要因分析 投資の多様化の進展と内部資金の役割，財務省財務総合政策研究所「フィナンシャル・レビュー」平成 29 年第 4 号 (通巻第 132 号), (2017)

[8] 沖本竜義.: 経済・ファイナンスデータの計量時系列分析，朝倉書店，(2010)

高頻度注文情報の符号化と深層学習による短期株価予測

田代　大悟 [1*]　和泉　潔 [1]
Daigo Tashiro[1]　Kiyoshi Izumi[1]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo

Abstract: Predicting the price movements of stocks based on deep learning and high frequency data has been studied intensively in recent years. Especialy, limit order book which describes supply-demand balance of the market is used as feature of a neural network, however, these methods do not utilize the properties of market orders. On the other hand, order encoding method of our prior work can take advantage of these properties. In this paper, we apply some types of convolutional neural network(CNN) architectures to order-based features to predict the direction of mid-price movements. The results show that smoothing filters which we propose to employ over embedding features of orders improve accuracy. Furthermore, inspection of embedding layer and investment simulation are conducted to demonstrate the practicality and effectiveness of our model.

1　序論

めまぐるしく変動している金融商品の価格を予測することは可能であるのか．という問いに対して，実務家だけでなく，様々な分野の学者の間で多くの理論研究，実証研究が行われてきた．効率的市場仮説の提唱や実証研究によるそれへの反証などを経て，最近では情報工学の分野，特に機械学習，深層学習を用いた手法が増えつつある．これらの手法は，データの背後に潜む規則や知識を発見するパターン認識の能力を備えており，市場予測において一定の成果を上げている．

市場予測に関する研究が多く行われている中，市場の様相も大きく変化した．市場の電子化と高速化に伴い，アルゴリズム取引や高頻度取引 (High Frequency Trading: HFT) といった機械的な取引が台頭し，市場で観測される注文データ，取引データはサンプリング頻度が非常に高く，また膨大化している．これらは「高頻度データ」と呼ばれ，有効な利用が期待されている．

高頻度データに対しても，深層学習を用いた研究が行われている．特に，板を用いたものが多い [1][2]．板の注文の価格と量を売り (アスク)，買い (ビッド) それぞれ数本分を入力として，ニューラルネットワークによる仲値の動向予測を行い，既存の機械学習を上回る精度を上げている．しかしこれらの手法には，成行注文に関してより重大な課題が存在する．それは，板のベストアスクまたはベストビッドの数量が減少したとき，それが成行注文によるものかキャンセル注文によるものか区別がつかない，というものである．成行注文とは，即時約定かつコストを支払う注文でトレーダーの強い意思を表したものであり，マーケットへのインパクトも大きい．さらに，成行注文とリターンとの相関も強いため，キャンセル注文のもつ意味，情報とは異なる．これをニューラルネットワークに識別させるには，注文系列自身をモデルの入力とすれば良い．

一方，指値注文やキャンセル注文も価格への影響を持つと言われており，これを無視することはできない [3][4]．そこで本研究では，すべての注文タイプを含めた高頻度注文系列と，深層学習を用いた短期の価格動向予測を行う．まず注文の符号化手法について説明し，予測モデルとして価格予測に特徴的な注文を捉えるよう CNN を改良した A-CNN(Average Convolutional Neural Network) と，その課題を踏まえて拡張した A-CNN+を提案する．本研究の目的は，この手法によるアルゴリズム取引の支援である．本研究の達成により，その運用パフォーマンスを向上することができると考える．

2　注文の符号化手法

注文の特徴には，価格，数量，時刻は数値情報 (量的変数)，売り買いの別，注文タイプはカテゴリ情報 (質的変数) というように，質の異なる変数をもって表され

*連絡先：東京大学大学院工学系研究科システム創成学専攻和泉研究室，〒 113-8654 東京都文京区本郷 7-3-1，E-mail: m2016dtashiro@socsim.org

る．本節では，これらをニューラルネットワークへの入力とするために用いる手法，注文の符号化手法について説明する．注文の符号化は 2 つの段階に分けられる．まず，i) 注文に属する変数に対してすべて質的変数側に変換するカテゴリ化．ii) その後，すべての変数を一つの質的変数に変換する 2 つの処理である．まず，i) の処理を行う意図について説明する．

注文タイプ

売り買いの別，成行注文，指値注文，キャンセル注文の注文タイプに関しては，すでにカテゴリ化された質的変数であり，以降，本研究では一つにまとめて注文タイプとして扱うこととする．

価格

質的変数をニューラルネットワークの入力とする場合，正規化を行うことが一般的である．しかし，注文の価格の分布は非対称で多峰性を持っており，標準偏差を求めるのは難しい．そこで，絶対価格ではなく仲値からの距離という相対価格を用いるとする．この度数分布はべき性を持ち，平均と標準偏差に意味をなさない．同じく標準化が難しいことから，数値情報ではなくそれをカテゴリ化して用いる．また，相対価格だと，仲値から遠い指値注文は価格への影響が小さいものとして大きく一つとして扱うことが可能であるということから，カテゴリ化の方が正規化に優っていると考えている．

時刻差

高頻度データを分析するにあたり，どのような投資家がどのような注文を出したのかと識別することは難しい．しかし，注文市場参加者の思惑は様々であり，発注者の情報を注文に関連付けることにより市場予測の精度を高めることができると考えられる．そこで，その指標として一つ前の注文との時刻差を情報に加える．これは例えば，ミリ秒オーダー間隔での注文は機械的なトレーダーによる注文だと識別し，発注者がどのような取引を行っているのかというおおよその分類情報を加えることを意図している．

次に ii) について説明する．複数のカテゴリを統一し一つの符号として表す目的として，ニューラルネットワークへ入力する注文を均質にするということが挙げられる．また，同じ価格であっても，指値注文のそれとキャンセル注文のそれでは意味が異なる．さらに，ニューラルネットワークの第一層と注文の関係を一意にすることができ，後に行う埋め込み層の分析を容易にするという利点もある．

Figure 1 には，具体的な注文の符号化手法を挙げる．価格の影響の少ない仲値から遠い価格は粒度を粗くカテゴリ化する．時刻差カテゴリは，500 ms を境に機械的な注文か否か，さらに機械的な注文の中でも 20 ms を境に高速な注文か比較的低速な注文なのかを識別できることを期待して設定した．

$$
\begin{pmatrix} \mathrm{MarketOrder}^{\mathrm{ask}} \\ \mathrm{MarketOrder}^{\mathrm{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 20 \\ 20 \sim 500 \\ 500 \sim \end{pmatrix}
$$

$$
\begin{pmatrix} \mathrm{LimitOrder}^{\mathrm{ask}} \\ \mathrm{LimitOrder}^{\mathrm{bid}} \\ \mathrm{Cancel}^{\mathrm{ask}} \\ \mathrm{Cancel}^{\mathrm{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 1 \\ 1 \sim 2 \\ 2 \sim 3 \\ 3 \sim 5 \\ 5 \sim 7 \\ 7 \sim 10 \\ 10 \sim \end{pmatrix} \times \begin{pmatrix} \sim 20 \\ 20 \sim 500 \\ 500 \sim \end{pmatrix}
$$

注文タイプカテゴリ　　価格カテゴリ (円)　　時刻差カテゴリ (ms)

Figure 1: 本研究で用いた注文の符号化．

3　深層学習による価格予測

3.1　データセット

データは FLEX-FULL ヒストリカルデータを用いた．銘柄は全 20 銘柄を用いる．期間は各々2013 年 7 月 1 日から 2014 年 6 月 30 日までの 1 年間のうち，東京証券取引所の営業日 245 日分を使用した．それぞれの銘柄に対し，予測対象となる価格は約定価格と仲値，各予測対象に 2 クラス，3 クラスの分類問題を設定した．

本節では，アルゴリズム取引での運用を考慮して価格動向予測のデータセットの作成を行う．価格予測モデルの入力となる注文系列 (イベントドリブン) と出力である価格動向を決定するタイミングを，アルゴリズム取引と同じ時間スケール (タイムドリブン) に沿ったものにする．具体的には，符号化済みの一つの注文系列を 30 秒毎で間隔で区切りを設定し，複数の注文系列を獲得する (Figure 2)．このとき，分割された注文系列 $\mathcal{S}^j$ は可変長となるが，その取り扱いは予測モデル側で解消する．

$\mathcal{S}^j$ に対応する価格の動向クラス d^j は，間隔の終点から 30 秒後の価格を対象とする (Figure 2)．価格の上がり/下がりを予測する 2 クラス分類問題，上り/下り/変動しないを予測する 3 クラス分類問題を設定した．

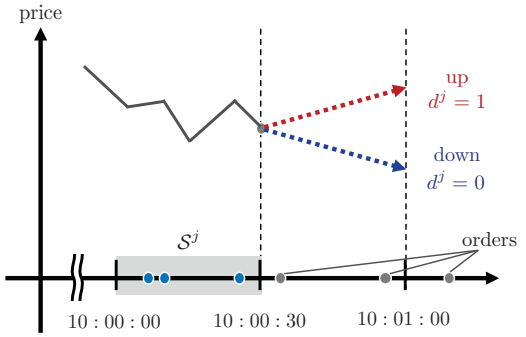


Figure 2: 系列 $\mathcal{S}^j$ のサンプリングと 2 クラス分類におけるラベリング．

3.2 CNN による予測モデル

CNN は画像認識の分野だけでなく，自然言語処理の分野においても，文書分類といったタスクで成功を収めている [5][6]．本研究で注文時系列に対して CNN を用いる理由として，CNN が位置に対して不変性を有している点が挙げられる．

CNN は畳み込みの後に系列方向に最大プーリングを行うため，指値注文やキャンセルが系列の後方に集中したとしても，成行注文を認識し，価格動向のシグナルとなる特徴やパターンを掴むのに有利だと考えられる．このような理由から，本研究では，局所的な注文系列を畳み込み，パターンを抽出する CNN を用いたモデルを用いる．

注文 x_t の埋め込みベクトル $\boldsymbol{x}_t$ を用いてパディングによって n で統一された系列 $\mathcal{S}$ は，$\boldsymbol{x}_{1:n} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ と表現する．これの局所的な行列 $\boldsymbol{x}_{i:i+h}$ を畳み込み，活性化関数を適用することによって新たな特徴 c_i を得る．ストライド幅 1 として，$(\boldsymbol{x}_{1:h}, \boldsymbol{x}_{2:h+1}, \ldots, \boldsymbol{x}_{n-h+1:n})$ に対してこの畳み込みを行うと，次のような新たな特徴ベクトル $\boldsymbol{c} = [c_1, c_2, \ldots, c_{n-h+1}]^T \in \mathbb{R}^{n-h+1}$ を得る．その $\boldsymbol{c}$ 最大プーリングを行うことによって，一つのフィルタから得られる一つの特徴量 $\hat{c} = \max(\boldsymbol{c})$ を得る．複数の畳み込みフィルタに対して，この処理を行う．k_{conv} 個のフィルタによる畳み込み演算と最大プーリングによって得られる特徴量 $\hat{\boldsymbol{c}} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_{k_{\text{conv}}}]^T$ を全結合層へ入力し，その出力をソフトマックス関数による変換後，出力クラスを得る．畳み込みフィルタのサイズを複数にするのは，その大きさによって，パターンを抽出するための系列の長さを変えるためである．つまり，大きいサイズのフィルタであれば，より多くの注文の関係性を捉え，反対に小さいフィルタであれば，局所的な関係性を捉えようとする．

3.3 注文の埋め込みの平均化を利用した予測モデル

本項では，注文の埋め込みの平均化を利用した予測モデル (Average Convolutional Neural Network: A-CNN) を提案する．金融市場での時系列データでは，一定の価格トレンドが観測されても，注文単位などのミクロ構造での明確なパターンは少ない．テキストデータとは異なり，注文の時系列では，畳み込みフィルタが捉える局所的な領域での，注文の相互作用が小さいと考えられる．そこで，一定数の注文の平均をとることで，これらの影響を小さくすることを考える．記号列に対して平均を取ることはできないので，埋め込み行列の局所的な範囲を対象とした平均化を行う．

埋め込み行列 $\boldsymbol{x}_{1:n}$ に対して平均プーリングを，窓幅 $1 \times l_{\text{pool}}$ で適用する．プーリング前の注文時系列方向上下それぞれのパディングサイズ l_{pad} によるパディングを行う．プーリング後の特徴行列は，床関数を $\lfloor x \rfloor$ とすると，$\boldsymbol{x}_{\text{pool}} \in \mathbb{R}^{e \times \lfloor \frac{l_{\text{pad}}+n}{l_{\text{pool}}} \rfloor}$ と表すことができる．その後は前項の CNN と同様に順伝搬および学習を行う．

また平均化による効果は，成行注文の相互作用を捉えるために一役買う．前項の CNN では，成行注文の間隔が，畳み込みフィルタのサイズより大きい場合に，成行注文の関係の特徴を抽出できない．しかし平均化を加えることで，注文系列に対するフィルタが捉える長さを $l_{\text{pool}} \times h$ に広げることができる．このようにすることで，出現頻度の低い成行注文間の相互作用を捉えることができる．しかし，成行注文がプーリングの窓 l_{pool} 内に複数存在する場合に，それらを平均化してしまうデメリットが生じる．

3.4 A-CNN の拡張モデル

本項では前項の A-CNN をさらに拡張したモデル (以下 A-CNN+) を提案する．Figure 3 にその図を示す．注文の埋め込み行列から，窓幅の異なる平均プーリングを行うことにより，複数の平均化行列を作る．そしてそれぞれの平均化行列に対して畳み込みフィルタを用意し，畳み込みと最大プーリングをとった後，ベクトルを連結し，上述の CNN と同様に全結合層へと入力する．

この操作の目的は，畳み込みフィルタのサイズを複数持たせるのと同様で，注文系列の文脈の大小に多様性を持たせるためである．つまり，平均プーリングのサイズが大きいほど，畳み込みフィルタはよりグローバルから特徴を抽出し，小さいほどローカルから特徴抽出を行う．ここでは，注文系列の中で価格変動に有用なパターンが異なる大きさで存在すると仮説を置いている．グローバルでは価格時系列でいうモメンタムのような大きなトレンドが，ローカルではよりミクロな注文の相互作用といったパターンがあると考えている．この複数のプーリングの窓幅と複数の畳み込みフィルタの窓幅を設定することで，注文間の相互関係の捉える幅を相乗的に変化させ，あらゆる長さのパターンに対応させる．また A-CNN での，プーリングの窓内に出現する複数の成行注文の平均化まで行ってしまう課題を解決する．

3.5 結果

各銘柄 1 年間のデータのうち，学習用データ，検証用データ，評価用データとして，古い順に 7 : 1.5 : 1.5 に分割した．20 銘柄の各実験，手法ごとに、学習用データを用いたモデルの学習とパラメータ探索を行い，検証用データ用で最も評価の高いモデルを選択する．評

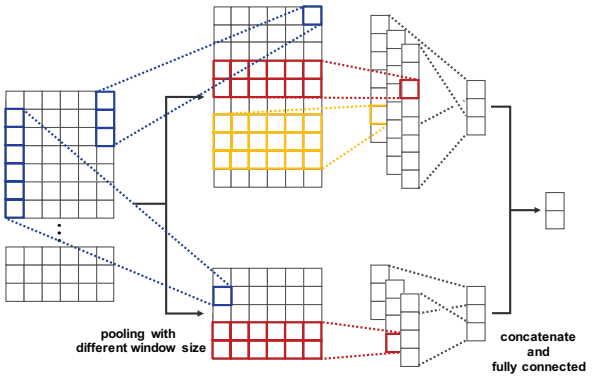


Figure 3: A-CNN+の構造.

価では，評価用データにおける各クラスに対する F1 値をそれぞれ求め，平均を取ったもので行う．

実験結果の一つを Figure 4 に示す．これは仲値の 2 クラス予測実験のモデルの評価用データでの F1 値である．すべての銘柄で提案手法がベースラインを上回る．A-CNN と A-CNN+では，実験や銘柄によって優劣が異なる．

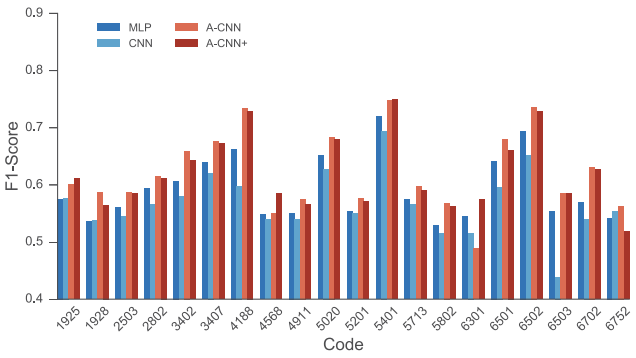


Figure 4: 仲値予測の 2 クラス分類問題における各手法の F1 値．ベースラインとしてロジスティック回帰，非線形 SVM(Support Vector Machine)，MLP(Multi Layer Perceptron)，3.2 節の CNN を用いた．ここでは，ベースラインの中で最もスコアの高かった MLP，CNN と，提案手法である A-CNN と A-CNN+を示す．

3.6 埋め込み層の分析

本節では，埋め込み層の分析を行う．各注文に対応したベクトルは各注文を表現したベクトルであり，そのノルムはニューラルネットの発火の強さであると考えることできる．Figure 5 は，2 クラス分類問題の，対象を約定価格とした実験と，仲値とした実験それぞれの，ノルムの平均を示している．約定価格予測では成行注文の重みが突出して高いのに対して，仲値予測では仲値に近いところに出された注文やキャンセルは成行注文と同等の重みを置かれている．

これは，仲値の予測が，約定価格の予測と比較して指値注文，キャンセル注文の影響を受けやすいためである．仲値は，ベストアスク，ベストビッド付近の指値注文とキャンセル注文が入ることにより変動するが，約定価格は成行注文が入ることでしか変動しない．仲値も成行注文によって変動するので，成行注文を含めすべての注文のタイプに反応するようなモデルになっていると考えられる．この分析結果は，約定価格と仲値の性質から見て納得のいくものとなっており，提案手法が意図の通りに動作していることを示している．

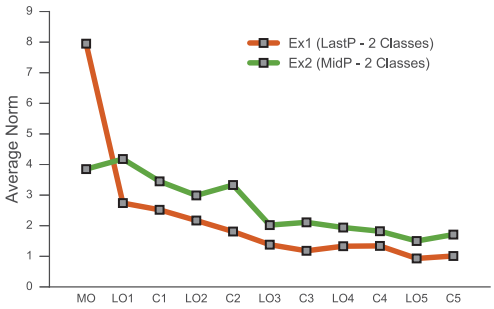


Figure 5: A-CNN モデルにおける埋め込みベクトルのノルム平均の詳細．成行注文と，各価格帯における指値注文，キャンセル注文の埋め込みベクトルのノルムを，すべての銘柄の対して平均を取る．MO は成行注文，LO は指値注文，C はキャンセル注文を指す．LO1 や C2 の数字は価格カテゴリを表す．1 は Figure 1 における価格カテゴリ ～ 1 を，2 はカテゴリ 1 ～ 2 を，3 はカテゴリ 2 ～ 3，4 はカテゴリ 3 ～ 5 を，5 はカテゴリ 5 ～ 7 である．LastP は約定価格を，MidP は仲値を指す．

4 投資シミュレーション

ニューラルネットワークの出力層にはソフトマックス関数を用いており，確率を出力する．モデルの出力するこの確率が高い場合のみ採用し投資行動をとることは，精度を高め，取引一回あたりのパフォーマンスを上げることができると考えられる．そこで，出力の確率の最大のものに閾値を設け，投資行動を決定することを考える．

この出力の確率が高いほど，うまく特徴を捉え信頼性の高い出力となることを裏付けるものととして，Figure 6 を示す．これらは，各評価用データにおいて，出力に閾値を設けた場合の Precision である．出力が各閾値を超えたサンプル数の中で Precision を算出する．閾値を高く設定するほど，評価値が高くなり，その閾値を超える確率を出力する回数が減少する傾向が見られる．

投資テストを行うにあたり，収益の高さによって 2 クラス分類問題と 3 クラス分類問題の比較を行う．モデルは，A-CNN と A-CNN+のうち検証用データで最も F1 値が高かったものを用いる．

コストは，取引を行う度にかかるものとする．簡単のためスプレッドを均一の 1 円とし，取引一回にかかるコストとして，ハーフスレッドの 0.5 円とする．さ

Figure 6: 出力に閾値を設けた場合の Precision と予測機会の割合 (2 クラス分類)．棒グラフ: Precision(左軸)．折れ線グラフ: 出力の確率が閾値を超えたサンプル数の割合 (右軸)．閾値は (0,50, 0.55, 0.60) とした．サンプル数の割合は評価用データのサンプル数を 1 としたときの割合を示している．閾値が 0.5 の場合は 1 となる．

らに，すべての銘柄で運用を行ったとして，このポートフォリオから得られる利益を算出する．

結果を Figure 7 に示す．閾値を変更しつつ投資テストを収益を計算した．それぞれの分類問題いずれもの特徴として，閾値を大きくしていくとある点を境に収益がプラスに転ずることがわかる．さらに大きくしていくと，収益が最大となる点を通り，減衰し 0 に近く．収益がプラスとなっている領域は，一回の取引でハーフスプレッドのコストを上回る領域である．閾値を大きくしすぎると一回の予測の精度を高めることができるものの，取引の回数が小さくなるため収益の和は減衰する．

これから，本研究で提案した注文の符号化手法と，埋め込み行列の平均化を用いた提案手法が実用的であることがわかる．そして両者を比較すると，より大きな収益を上げることができ多くの閾値でプラスの収益を上げることのできる 2 クラス分類の方が良い結果だと言える．これは 2 クラス分類問題の方が，3 クラス分類問題より学習が容易であったためだと考えられる．

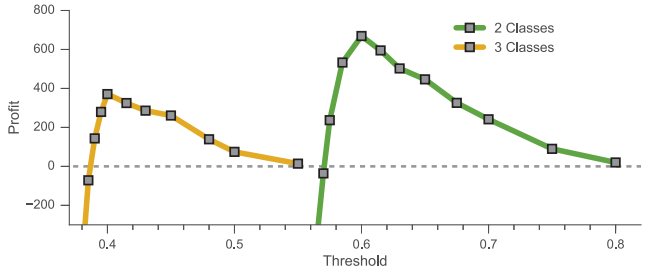


Figure 7: 閾値を変化させテストを行ったときの収益の推移．コストを設定した上で，すべての銘柄に対して得られる収益の合計．

5　まとめと今後の課題

本研究では，株価の短期動向予測のため，注文の符号化手法を提案し，価格動向予測に有用な特徴的な注文を捉えるのに適するように CNN を改良した．

大規模な高頻度データから複数のタスクを設定し実験を行い，提案した手法が他の手法より優れていることを示した．さらに，埋め込み層による分析を行い，その意図が機能していることを確認した．また，すべてのタスクで成行注文に強く反応するモデルであることを，仲値の予測問題では，仲値に近い指値注文とキャンセル注文も成行注文と同等に注目していることを発見した．

最後に，提案手法の実務的な有用性を示すために投資テストを行った．成行注文による仲値の売買を行い，プラスの成績を上げることで，実用面でも耐えうるモデルであることを示した．さらに，出力の確率に閾値を設定し，2 クラス分類のモデルを 3 クラス分類に適用することで，3 クラス分類のモデルをパフォーマンスの面で上回った．

今後の課題として，板の情報との組み合わせが挙げられる．板の情報は流動性という面で，注文系列に不足する情報を補足することができる点で，さらなる精度向上が見込まれる．

参考文献

[1] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, Using Deep Learning to Detect Price Change Indications in Financial Markets, *European Signal Processing Conference* (2017).

[2] M. F. Dixon, Sequence Classification of the Limit Order Book using Recurrent Neural Networks, *Journal of Computational Science* (2017).

[3] Z. Eisler, J. P. Bouchaud, J. Kockelkoren, The price impact of order book events: market orders, limit orders and cancellations, *Quantitative Finance*, vol. 12, pp. 1395-1419 (2012).

[4] R. Cont, A. Kukanov, and S. Stoikov, Price impact of order book events, *Journal of Financial Econometrics*, vol. 12, pp. 47-88 (2014).

[5] Y. Kim, Convolutional Neural Networks for Sentence Classification, *Empirical Methods in Natural Language Processing*, pp. 1746-1751 (2014).

[6] R. Johnson, and T. Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, *North American Chapter of the Association for Computational Linguistics* (2014).

ボラティリティ・クラスタリングが観測される時系列のローソク足同時分布モデル

Candlestick Joint-Distribution Models for Time Series with Clustered Volatility

内木　正隆[1*]　DE BRECHT Matthew[1]　櫻川　貴司[1]
Masataka Naiki[1]　DE BRECHT Matthew[1]　Takashi Sakuragawa[1]

[1] 京都大学 人間・環境学研究科 数理科学講座
[1] Graduate School of Human and Environmental Studies, Kyoto University

Abstract: *Open-high-low-close price* (also OHLC) series have been widely used for the price-movement analysis of financial time series, including to draw *candlestick charts*. Modeling these data is complicated by the fact that such data are often unlikely to be samples of stationary stochastic processes, as can be seen in the well-known phenomenon of *volatility clustering*. In this research, first we try to remedy this matter by using the sequences of differences between high and low prices, which are pointed out to often have higher autocorrelations than the absolute returns of close-price series, and normalize the scales of OHLC by their exponential moving averages. Under our experimental conditions, the *Earth Mover's Distance* (EMD) between normalized S&P500 training and test data is about one-seventh of the EMD between the unnormalized data. Second, we try to model the normalized data by introducing 6 generative models for them. The EMDs between data generated by our learned models and the normalized test data are about one-sixth of the EMD between the normalized test data and the delta distribution located at the barycenter of the normalized training data. However, they are about 5 times larger than the EMD between the normalized test and training data.

1　はじめに

本論文では、金融時系列データ等でよく用いられる、一定期間ごとのいわゆる四本値の一部の情報について、確率過程としての定常性を強める変換の方法と、変換後のデータを学習する生成モデルを提案し、評価を行う。

金融時系列データは多くの回数の売買によって動いた価格変動の履歴全てから構成される。しかしこれらの値動き全体のデータ量は非常に膨大なため、実際には一定期間ごとの最初の値 (始値)、最後の値 (終値)、その期間内での最大値 (高値)、最小値 (安値) の**四本値**、あるいは取引された量 (出来高) を加えた 5 つの値のみがその期間の値動きデータとして利用できる場合も多い。これらは終値のみに比べて与える情報がより大きく、これらを用いると終値のみを用いた場合に比べて良い精度で予測可能になる場合もある [2]。しかし金融関連の時系列分析を行う場合、四本値全てを同時に扱うモデルや研究はあまり多くなく、終値のみのスカラー時系列の予測を行う場合がかなりある。本論文では逆に終値の時系列の情報の部分を除いた他の情報を扱うことを考える。高値や安値の値動きを調べるための方法の一つとして、その期間内の始値に対する終値・高値・安値の相対値を扱うことができる。これら相対値は四本値を同時に時系列として可視化する手法の一つであるローソク足チャートの個々のローソクの形を表す 3 次元データである。本論文ではこれら相対値データのモデル化を目指す。関連研究としては相対的な終値の変化率の予測を試みた例は [10][11] など多くあるものの、高値や安値の動きを終値も含めた同時分布として予測・評価する例は今回見つけられなかった。

モデル化の際に出てきた問題点とそれらの解決を目指す提案を以下に挙げる。まず扱うデータが定常過程のサンプルとは思えず、時期により標準偏差が異なるため定常過程でのモデル化が難しかった。そのため指数移動平均を用いて時系列データの変換をするという方法を提案して非定常性を抑えた。また分布のモデル化の際に、実データがルベーグ絶対連続な同時分布では表現できない分布のサンプルに見えるという問題が

*連絡先：京都市左京区吉田二本松町
E-mail: naiki.masataka.44z@kyoto-u.jp
sakura@i.h.kyoto-u.ac.jp

あった。つまり密度関数で表される分布ではうまく表現できないのである。これについては実際に最大値や最小値を求めるような生成モデルを 3 種類提案し、それでも表現できない部分をそれぞれ別に扱う (後述する間引き処理) ことで合計 6 種類のモデルを提案することで対応した。実際のデータ集合と、それを推定した生成モデルからの推定値集合の類似度を表す指標としては Earth Mover's Distance (EMD) を用いた。

評価時の対象データとしては、アメリカ株の代表的なインデックス指数の一つである S&P500[1]、期間は 2001 年から 2016 年までの期間を対象とした。2013 年までを訓練データ Train、2014 年以降をテストデータ Test として分割して、指数移動平均のパラメータの学習は標準偏差のばらつきを最小化することで行った。結果は指数移動平均によるある種の正規化を行った後の訓練データとテストデータの EMD が 0.0571 で、変換前の訓練データとテストデータの EMD が 0.4236 だった。

変換後の訓練データの 6 種の生成モデルでの学習は、訓練データとモデルが生成するサンプルの EMD を最小にするパラメータを求めることで行った。変換後のテストデータと学習後の生成モデルが生成するサンプルの EMD は 0.3 程度、変換後の訓練データと生成サンプルの EMD は 0.18-0.23 程度であった。

論文の構成は以下の通りである。2 節で本論文で考察対象とする事象について述べ、関連用語を説明する。3 節で時系列の定常性の定義、ボラティリティ・クラスタリングが起きている場合に用いる既存のデータ処理方法と、データの定常性を増加させる為に本論文で導入した手法を述べる。これはある種の正規化を行なっていることになる。4 節では分布間の類似性を表す指標として今回採用した Earth Mover's Distance について説明する。5 節では正規化後の分布を学習する為導入した 6 種の生成モデルと、それらのパラメータの学習 (最適化) 方法、S&P500 のデータへの適用結果について述べ、6 節では考察を行う。

2 時系列とローソク足

時系列の期間毎の四本値を図 1,2 のように表示したものは**ローソク足**と呼ばれる。これは高値と安値の幅を棒の長さ、終値と始値の幅を箱の高さ、終値と始値の大小を箱の色で示したもので、こういった情報を同時に表現することができる。

以降全取引期間中の (通常一定幅の) 各期間を $I_t, t = [0, 1, .., T]$ で指し、I_t における始値:$(Open)_t$・高値:$(High)_t$・安値:$(Low)_t$・終値: $(Close)_t$ について考える。例えば 2017 年 3 月の市場が開場している日の日足 (各期

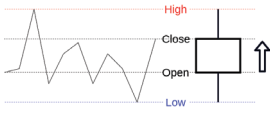


図 1: 始値 < 終値の場合

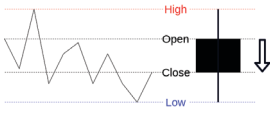


図 2: 始値 > 終値の場合

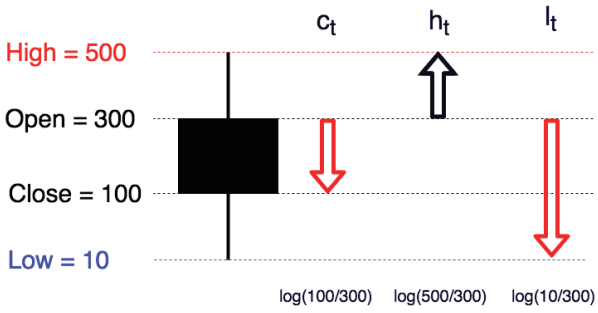


図 3: ローソク足と c_t, h_t, l_t の関係性

間を各営業日の取引時間としたもの) のインデックスは、各日付 (3/1,3/2,3/3,3/6,...) を各インターバル区間 $(I_0, I_1, I_2, I_3, ...)$ に対応させる。

今回は終値の相対値だけではなく、始値と比較した高値・安値の相対的な値についても関心があるため、それらの同時分布を扱い、(1) 式で c_t, h_t, l_t を定義する。図 3 でもわかるように、これらから成る 3 つ組はローソク足の形に対応している。これらの値は常に $l_t \leq c_t \leq h_t, l_t \leq 0 \leq h_t$ を満たす。

$$\begin{aligned} c_t &\stackrel{\text{def}}{=} \log((Close)_t) - \log((Open)_t) \\ h_t &\stackrel{\text{def}}{=} \log((High)_t) - \log((Open)_t) \quad \geq 0 \\ l_t &\stackrel{\text{def}}{=} \log((Low)_t) - \log((Open)_t) \quad \leq 0 \end{aligned} \tag{1}$$

金融時系列の時間あたり変化率の分布は、裾の重い分布とよばれる性質を持つことが多いことが知られている [4]。ここで裾の重い分布とは、確率分布の裾がガウス分布のように指数減衰的ではなく、それよりも緩やかな分布のこととする。

以上の 3 次元データを発生させる生成モデルを考える場合、例えば $h_t = 0$ や $c_t = h_t$ となる割合が実データで 0 でない場合がかなりあることが問題となる。これは密度関数等で表されるような**ルベーグ絶対連続**な確率分布ではうまく表現できないデータであることを意味しており、それを考慮に入れた生成モデルを考える必要がある。

[1] YahooFinance (https://finance.yahoo.com/quote/%5EGSPC/) のデータを用いた。

3 変動率の時系列モデル

確率過程の性質として**定常性**がある。これは基本的な統計量が時刻に関わらず一定である性質のことである。

定義 3.1 確率過程 $\{Y_t\}_{t\in\mathbb{Z}}$ が (弱) 定常性をもつとは任意の $t, h \in \mathbb{Z}$ において期待値・共分散が定義できて、かつ以下が成立することをいう[2]。このとき $\{Y_t\}$ は定常過程とよばれる。

$$\mathbb{E}[Y_t] = \mu(\text{定数})$$
$$\mathrm{Cov}[Y_t, Y_{t-h}] = \gamma_h(h \text{ の関数})$$

モデルを作る対象の時系列がこうした定常性をもった過程から生成されていると仮定できればモデル化が容易なことが多い。しかし一般に金融時系列では一定期間ごとの変化率の絶対値の分布が必ずしも時刻に対して一定ではなく、しかも連続して似た値になりやすい傾向が知られている。こういった現象はボラティリティ・クラスタリング、あるいは分散不均一性と呼ばれる [7]。図 4 の左側を見ると、今回用いた S&P500 のデータでも経済が順調だった 2006 年の $\mathrm{std}[c_t]$ [3] と比べて、リーマン・ショックが起きた 2008 年の $\mathrm{std}[c_t]$ の方が約 4 倍である。この場合に 2006 年と 2008 年を同じ過程から発生した値と仮定してモデル化するのは適当ではないと思われる。

こうしたデータの変化率の絶対値のばらつきを揃えるため、本論文では $\{u_t \stackrel{\text{def}}{=} \log((High)_t/(Low)_t)\}$ のパラメータ α の指数移動平均 $\bar{u}_t$ を用いることを提案する。実際の変換は $c_{t+1}, l_{t+1}, h_{t+1}$ を、パラメータ $0 < \alpha < 1$ による指数移動平均 $\bar{u}_t$ で割ることで行う。$\bar{u}_t$ は (2) で定義される。

$$\bar{u}_t = \begin{cases} \alpha u_t & (t = 0) \\ \alpha u_t + (1-\alpha)\bar{u}_{t-1} & \text{otherwise.} \end{cases} \quad (2)$$

元の変化率データである c_t, h_t, l_t を $\bar{u}_t$ を用いて $c'_t \stackrel{\text{def}}{=} c_t/\bar{u}_{t-1}, h'_t \stackrel{\text{def}}{=} h_t/\bar{u}_{t-1}, l'_t \stackrel{\text{def}}{=} l_t/\bar{u}_{t-1}$ と変換する。指数移動平均のパラメータ α は訓練データの期間ごとのばらつきをできるだけ一定とするような、つまり短い期間毎の標準偏差集合の標準偏差最小化によって求める。訓練データ学習期間中の i 番目の月 (2001 年 1 月,2001 年 2 月,...,2013 年 12 月) を m_i として、(3) 式で与えられるデータのばらつき具合を最小化する α を選択する。

$$\sum_{r\in\{c',h',l'\}} \mathrm{std}[\{\mathrm{std}[\{r_t\}_{t\in m_i}]\}_i] \quad (3)$$

S&P500 からの訓練データ Train で α の最適化を行った結果 $\alpha = 0.29072$ となった。またこのとき $\sigma_{c'} = 0.77795$ だった。このように時刻 u_t の指数移動平均を時刻 u_{t+1} のある種の予測値的な値として用いることは、金融時系列で変動の幅を推定するモデルとして一般的な **GARCH モデル** [2] と類似している。ただし GARCH モデルでは予測値と実際の値との誤差などによって係数を最適化するのに対し、提案手法では (3) の値を最小化しているところに新規性があると考えられる。

図 4 の右側は指数移動平均による変換後の c'_t, h'_t, l'_t のグラフで、左側の変換前に比べ各年ごとの標準偏差の大きさのばらつきが抑えられているのが判る。さらに訓練期間中の t に関して $c''_t \stackrel{\text{def}}{=} c'_t/\sigma_{c'}, h''_t \stackrel{\text{def}}{=} h'_t/\sigma_{c'}, l''_t \stackrel{\text{def}}{=} l'_t/\sigma_{c'}$ とする。ただし $\sigma_{c'} \stackrel{\text{def}}{=} \mathrm{std}[c'_t]$。

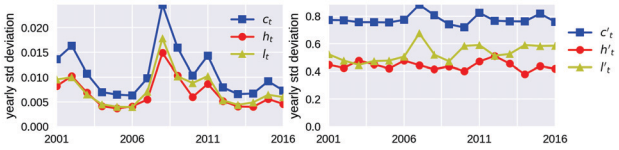


図 4: 2001 年から 2016 年まで各年内での S&P500 についての標準偏差。左は $\mathrm{std}[\{c_t\}_{t\in m_i}], \mathrm{std}[\{h_t\}_{t\in m_i}], \mathrm{std}[\{l_t\}_{t\in m_i}]$、右は指数移動平均による変換を行った $\mathrm{std}[\{c'_t\}_{t\in m_i}], \mathrm{std}[\{h'_t\}_{t\in m_i}], \mathrm{std}[\{l'_t\}_{t\in m_i}]$, のグラフ。

4 Earth Mover's Distance

本論文では正規化の結果の評価と生成モデルの学習のために、**Earth Mover's Distance** (EMD) を用いて分布間の類似性を測る手法を提案する。学習の場合には実際の同時分布と、それを推定した生成モデルからの出力の分布間の類似度を最大化することになる。EMD は一般に 2 つの分布の類似度を表す指標で、値が少ないほど類似性が高く、値の計算の基礎となる後述の $d(-,-)$ について $d(x,x) = 0$ であれば同じ分布同士の EMD は 0 である。EMD を直感的に説明するなら、分布を砂山と考えて片方の砂山からもう片方の砂山へ砂を移動させるための最小コストと説明できる。

計算機関連ではもともと色・テクスチャなどを考慮して定義される画像間の距離を与えるために使われ、画像修復や画風変換に用いられる [5]。定義の形式こそ異なるものの、EMD に相当する距離が Wasserstein Distance や最適輸送距離という名前でも知られており [9]、近年では深層学習における生成モデルで最適化の目的関数としても注目されるようになっている [3]。

以下の有限の場合の説明は [8] を参考にしている。このとき EMD は以下の LP 問題の最小値によって与えられる。各点に重み付けを持つ分布 $P \stackrel{\text{def}}{=} \{(p_i, w_{p_i})\}_{i=1}^m$,

[2] $\mathbb{E}$ は期待値、Cov は共分散をとる記号

[3] $\mathrm{std}[\{r_t\}_{t\in X}] \stackrel{\text{def}}{=} \sqrt{\frac{1}{|X|}\sum_{t\in X}(r_t - \bar{r})^2}$.

$Q \stackrel{\text{def}}{=} \{(q_j, w_{q_j})\}_{j=1}^{n}$ があるとする。距離行列 $D \in \mathbb{R}^{m\times n}$ を、p_i と q_j の間の輸送コストである $d_{ij} = d(p_i, q_j) \geq 0$ を要素に持つ行列として定義する。ただし $d(-,-)$ はユークリッド距離などベクトル同士の何らかの性質の差を表す指標である。d が距離の公理を満たせば EMD も距離の公理を満たす。$F \in \mathbb{R}^{m\times n}$ は輸送フローを表し、p_i から q_j へのフローをあらわす $f_{ij} \geq 0$ を要素に持つ行列である。最適化の目的関数はコストとフローの対応する要素同士の積の和で、これが分布間の移動の仕事量をあらわす。また制約条件によって、フローは 0 以上で p_i, q_j へのフロー流入量の最大値は決まっており、全フローの合計フローを一定に定めている。

$$\begin{aligned}&\underset{F}{\text{minimize}} \sum_i \sum_j d_{ij} f_{ij} \\ &\text{subject to } f_{ij} \geq 0, \sum_j f_{ij} \leq w_{p_i}, \sum_i f_{ij} \leq w_{q_j}, \\ &\qquad \sum_i \sum_j f_{ij} = \min\left(\sum_i w_{p_i}, \sum_j w_{q_j}\right). \quad (4)\end{aligned}$$

これによって求められた最小値をそのときのパラメータ $\hat{F}$ で正規化したのが P, Q の間の EMD である (5)。以上の定義は P, Q のそれぞれの重みの合計値が 1 であることを仮定していない。それぞれが分布、つまり全ての i, j において $w_{p_i} = 1/m, w_{q_j} = 1/n$ である場合も多く、こういった場合には正規化しなくても分子がそのまま $\text{EMD}(P, Q)$ となる。本論文ではこの意味で EMD を用いる。

$$\text{EMD}(P, Q) \stackrel{\text{def}}{=} \frac{\sum_i \sum_j d_{ij} \hat{f}_{ij}}{\sum_i \sum_j \hat{f}_{ij}} \quad (5)$$

(4) の最適化問題は線形計画問題の中の最適輸送問題のソルバーを用いて解かれる。これの計算コストが $\mathcal{O}(n^3 \log n)$ あるため大きな要素数に対しては計算コストが高くなる [1]。また (5) の値は一意に定まるが $\hat{F}$ は一意とは限らない。こういった問題の解決目的で様々な EMD の変種が存在するがここでは割愛する。

5 生成モデルの構成と結果

本論文で導入した計 6 つの生成モデルの具体的構成方法・評価方法・実験結果についてまとめる。

5.1 モデルの構成

本論文では実際に価格の変動率を左右する何らかの定常な分布が存在すると仮定して、全ての各インターバル区間 I_t の間に一定回数 (K 回) の価格の変動が、ある分布に従って独立に発生するモデルを考える。また、ルベーグ絶対連続でない分布をうまく表す方法として、後述の「間引き」処理を試みる。各生成モデルは分布モデル、同時分布計算モデル、(もしあれば) 間引き処理、スケールの調整からこの順で構成される。間引きを行う方法を (b)、行わない方法を (a) とする。今回は分布の形について (i, ii, iii) の 3 通り、間引きを行うかどうかで (a,b) の 2 通り、これらを組み合わせて (i)-(a),(i)-(b), (ii)-(a), (ii)-(b), (iii)-(a), (iii)-(b) の計 6 通りのモデルを実験対象とする。

分布モデル　変化率の元となる分布モデル p_θ の候補は多い。本論文では p_θ をガウス分布とする方法 (i)、ジョンソン S_U 分布システム [6] とする方法 (ii)、分散が或るパレート分布に従うガウス分布とする方法 (iii)、の 3 通りを試みた。

(i) は (ii), (iii) など p_θ が裾の重い分布 (確率分布の裾がガウス分布のように指数減衰的ではなく、それよりも緩やかな分布 [4]) を表現可能な場合と比較するために採用する。スケールに関しては学習パラメータに含めないため $\theta = (\mu \in \mathbb{R})$ となる。$r_{t,k} = q_{t,k} + \mu, q_{t,k} \sim N(0, 1)$ として $r_{t,k}$ を計算する。独立なガウス分布からのサンプルの和はガウス分布に従うことが知られており、p_θ が $N(\mu, \sigma^2)$ であるとき $\hat{c}_t = \sum_{k=1}^{K} r_{t,k} \sim N(K\mu, K\sigma^2)$ となる。

(ii) は $\theta = (\mu' \in \mathbb{R}, \sigma' \in \mathbb{R}_{>0}, \mu \in \mathbb{R})$ で $r_{t,k} = \mu + \sinh(\sigma' q_{t,k} + \mu'), q_{t,k} \sim N(0, 1)$ とする。σ' によって裾の重さを調節可能で、μ, μ' の組み合わせによって左右非対称な分布も表現可能となっている。

(iii) は $\theta = (\mu \in \mathbb{R}, c \in \mathbb{R}> 0, d \in \mathbb{R}_{>0})$ で $r_{t,k} = \mu + \sigma_{t,k} q_{t,k}, q_{t,k} \sim N(0, 1), \sigma_{t,k} \sim \text{pareto}(c, d)$ となる。c は $\sigma_{t,i}$ の下限を意味するパラメータ、d は裾の重さを調節するパラメータとなっている。

同時分布計算モデル　$k = 1, 2, .., K$ として各変動率 $r_{t,k} \in \mathbb{R}$ を分布モデルから独立同分布にサンプリングする。$r_{t,k}$ から長さ $K+1$ の時系列 $\{s_{t,k} \stackrel{\text{def}}{=} \sum_{k'=1}^{k} r_{t,k'}\}_{k=0}^{K}$ を計算する。$s_{t,0} = 0$ である。今回のモデルでは $s_{t,k}$ はインターバル区間 t の k 回目の変動後の価格と見なす。よってインターバル区間 t の始値を 0 としたときの最大値・最小値・終値の値を計算することで、サンプル $\hat{c}_t, \hat{h}_t, \hat{l}_t$ の同時分布を得ることができる (図 5)。つまり $\hat{c}_t \stackrel{\text{def}}{=} s_{t,K}, \hat{h}_t \stackrel{\text{def}}{=} \max_k s_{t,k}, \hat{l}_t \stackrel{\text{def}}{=} \min_k s_{t,k}$ とする。常に $\hat{l}_t \leq 0 \leq \hat{h}_t, \hat{l}_t \leq \hat{c}_t \leq \hat{h}_t$ である。

このようにモデル化することで、直感的には変動の積み重ねから結果が生じている実際の取引データの発生の仕組みをある程度近似しているのみならず、有限回の累積で高値・安値を発生させているため、ルベー

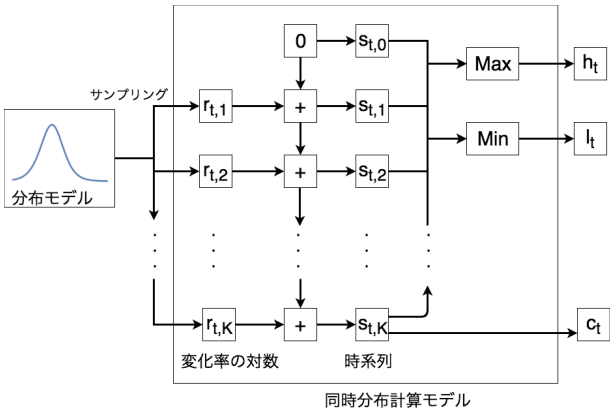


図 5: 同時分布計算モデル

グ絶対連続でない 3 次元分布を表すモデルとなっている。ただし少なくとも後述のように分布の比率を合わせるための処理が必要と考えられる。

間引き処理　今回用いるデータでは分布中のルベーグ絶対連続ではないと思われる部分の比率が同時分布生成モデルのそれらの比率と一致しない場合が多い。例えば今回用いた S&P500 のデータの場合、高値と始値が等しいデータが 10 %以上もあり、それを表せる間引きなしのモデルは限られてしまう。訓練データと、モデルが生成するサンプルの高値＝終値、高値＝始値、安値＝終値、安値＝始値の 4 つの場合の組み合わせの比率を一致させるように生成データを独立な乱数により一定の割合で間引く。各分布モデルについて、間引きする場合・しない場合の両方の生成モデルを考える。

5.2　モデルのパラメータ最適化手法と結果

生成モデルのパラメータ最適化の際には 4 節で説明した EMD を用いる。各要素間の指標 $d(-,-)$ としてはユークリッド距離の 2 乗を用いた。今回は毎回生成する (間引き後の) サンプル数も訓練データの個数と同じとした。前節であげた 6 つのモデルから成る集合を M、各モデルを $m \in M$、モデル m のパラメータを $\theta^{(m)}$ と表記する。以降各インターバル内の変動回数が K 回で、パラメータが $\theta^{(m)}$ の、モデル m から発生させた N 個の値を、$\mathrm{Predicted}(\langle m, \theta^{(m)}, K\rangle, N)$ と表記する。

$\theta^{(m)}, K$ の探索についてはある種のランダムサーチを用いた。ある分布に従って実際に $\theta^{(m)}, K$ をサンプリングした中で、最も $\mathrm{EMD}(\mathrm{Train}'', \mathrm{Predicted}(\langle m, \theta^{(m)}, K\rangle, N))$ の小さかった $\theta^{(m)}, K$ を選択する。ここで Train'' は指数移動平均による変換後に標準偏差を 1 に正規化したデータである。なお、指数移動平均の係数 α と標準偏差正規化の係数 $\sigma_{c'}$ は Test データに対しても同じ値:$\alpha = 0.29072, \sigma_{c'} = 0.77795$ を用いた[4]。結果のデータを Test'' と表記する。各モデル m ごとに K は 3 から 200 の整数の中から 100 個サンプリング、各ペア $\langle m, K\rangle$ ごとに $\theta^{(m)}$ のサンプリングを行う。$\theta^{(m)}$ のサンプリングの設定は表 1 の通り。

モデル	個数	パラメータ	下限	上限
(i)	30	μ	$-0.02/K$	$0.02/K$
(ii)	200	μ	$-0.02/K$	$0.02/K$
		μ'	$-0.02/K$	$0.02/K$
		σ'	0	3.0
(iii)	200	μ	$-0.02/K$	$0.02/K$
		c	0	0.01
		d	0	5.0

表 1: 各モデルにおけるランダムサンプリングの対象とする分布とサンプリング個数

6 通りの各モデルの生成値に対する $\mathrm{Train}'', \mathrm{Test}'', \mathrm{Mean}''$ との EMD の値を表 2 に記載した。ここで Mean'' は Train'' の重心に位置する δ 分布である。

EMD	Train''	Test''	Mean''
(i)-(a)	0.2317	0.2966	1.768
(ii)-(a)	0.2324	0.2938	1.781
(iii)-(a)	0.2314	0.2953	1.780
(i)-(b)	0.2318	0.2984	1.768
(ii)-(b)	0.1769	0.2957	1.774
(iii)-(b)	0.1797	0.2935	1.770
Train''		0.05712	1.803

表 2: 生成サンプル,$\mathrm{Train}'', \mathrm{Test}'', \mathrm{Mean}''$ 同士の EMD

6　まとめ・考察

本論文では従来それほど研究されていないと考えられる、時系列データの各インターバルでの高値・安値・終値の始値に対する相対値の生成モデルを研究対象とした。まず、このような場合に分布の近似度を測る方法として EMD を用いることを提案した。また、定常性を強める変換と、生成モデルを提案し、評価実験を行った。

具体的には、定常性が必ずしもなく、ボラティリティ・クラスタリングが生じるような確率過程が発生する時系列について、高値と安値の差の指数移動平均の値によりデータを変換することで定常性を増やすことを意図した方法を提案し、S&P500 のデータを使用した評価を行った。標準偏差を正規化した Train と Test の EMD

[4] それゆえ Test'' の標準偏差は必ずしも 1 にならない。

に比べて、変換後に標準偏差を正規化した Train″ と Test″ の EMD は約 1/7 であった。

また、変換後の Train″ を学習する生成モデルとして、累積和を取ってから最大値・最小値を求め、間引きを行うことでルベーグ絶対連続でない同時分布を表せるモデルを 6 つ提案し、やはり S&P500 のデータにより評価実験を行った。しかしながらこれらの異なるモデルを考えたものの、各生成モデルが発生したサンプルと Test″ の EMD は全て 0.3 程度であった。一方、Train″ の重心に位置する δ 分布 Mean″ と Test″ の EMD が約 1.8 であった。ただし 0.3 という値は Train″, Test″ の EMD の値約 0.06 に比べると大きいため、今回採用した生成モデルの表現力は実験データにはそれほどマッチしていなかったと考えられる。また、裾の重い分布への対応を意図した (ii),(iii) の分布モデルや間引きを行うモデル (b) が優れていることを期待していたものの、今回の実験結果ではそういった傾向は認められなかった。(i)-(iii) にそれほど違いが見られなかった理由としては、元の分布モデルのこの程度の違いでは、累積和を取るとどれも正規分布に近くなってしまい、結果に影響を及ぼさなかった、ということが考えられる。

発展としては異なる性質を持つ為替・債券といった金融時系列を用いた実験や生成モデルの表現能力を上げることが考えられる。例えば同時分布計算モデルの出力に対してジョンソン S_U 分布システムのような変換を行うことで Test″ が裾の広い分布となっている場合にさらに適応することを目指すことができる。分布モデルが発生するサンプルは今回各回で独立としたが、独立でない場合をモデル化する方向性も考えられる。また EMD を用いた非定常性を抑えるための変換として、指数移動平均以外の方法も考えられるし、今回のように標準偏差の標準偏差を最小化するのではなく、訓練データの範囲で一定期間ごとに分割されたデータペアの EMD の平均を最小化することで、最適なパラメータ (今回の場合には α) を求める方法もある。売買アルゴリズムによっては生成サンプルに基づく売買シミュレーションを行うことが可能なので、それを行うことが考えられる。今回の研究は終値の時系列そのものを予測対象としたものでないが、それを行った研究があるので、それらと組み合わせて四本値の確率予測モデル的なものを作ることや、それらを用いた売買シミュレーションを行ってみることも考えられる。

謝辞

京都大学人間・環境学研究科数理科学教室の渡部崇氏との討論は大変有益でした。また、日頃お世話になっている同教室の先生方、学生の方々にも感謝します。

参考文献

[1] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Chapter iv network flows. *Handbooks in operations research and management science*, Vol. 1, pp. 211–369, 1989.

[2] Sassan Alizadeh, Michael W Brandt, and Francis X Diebold. Range-based estimation of stochastic volatility models. *The Journal of Finance*, Vol. 57, No. 3, pp. 1047–1091, 2002.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Søren Asmussen. *Applied probability and queues*, Vol. 51. Springer Science & Business Media, 2008.

[5] Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference on*, Vol. 30, p. 158, 2011.

[6] N. L. Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, Vol. 36, No. 1/2, pp. 149–176, 1949.

[7] Thomas Lux and Michele Marchesi. Volatility clustering in financial markets: a microsimulation of interacting agents. *International journal of theoretical and applied finance*, Vol. 3, No. 04, pp. 675–702, 2000.

[8] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, Vol. 40, No. 2, pp. 99–121, 2000.

[9] Cédric Villani. *Optimal transport: old and new*, Vol. 338. Springer Science & Business Media, 2008.

[10] Halbert White. Economic prediction using neural networks: The case of ibm daily stock returns. 1988.

[11] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, Vol. 50, pp. 159–175, 2003.