

金融情報学研究会 (第19回)

日時 2017年 10月 14日 (土)

会場 東京大学

SIG-FiN
JSAI Special Interest Group on
Financial Informatics

人工知能学会
金融情報学研究会

第19回 人工知能学会 金融情報学研究会 (SIG-FIN)

2017年10月14日(土) 東京大学

01. 忍耐強い(Patient)アクティブ投資は市場を効率的にするのか? 1
---人工市場によるシミュレーション分析---
水田孝信(スパークス・アセット・マネジメント), 堀江貞之(野村総合研究所)
02. 約定価格の量子化ボラティリティモデル 9
森谷博之(Quasars22)
03. 能動的キャンセルを含む人工市場によるキャンセル寿命の冪分布の再現 13
吉村勇志(東京大学), 陳昱(東京大学)
04. 機械学習による中小企業の信用スコアリングモデルの構築 20
澤木太郎(リコー), 笠原亮介(リコー), 田中拓哉(リコー)
06. 北朝鮮リスクの日本株式市場への影響-AI ベースのリスクモデルは何を語るのか? 24
西山昇(Dragons' Desk Ltd., 千葉商科大学)
07. 多層ニューラルネットワークと GA を用いた TOPIX 運用 AI 29
石原龍太(かんぽ生命保険)
08. 潜在特徴関係モデルを用いた時系列金融ネットワークの解析と予測 35
伊藤翔太郎(神戸大学), 江口浩二(神戸大学)
09. LSTM-RNN を用いたイベント考慮・時系列予測の試み 42
南正太郎(あすかアセットマネジメント)
10. Prediction of Bitcoin Price Movements with Machine Learning Algorithms 45
Takuya Shintate(International Christian University), Lukas Pichl(International Christian University),
Taisei Kaizoji(International Christian University)
11. 機械学習を用いた共和分ペア・トレード戦略 47
今村光良(日興グローバルラップ, 筑波大学), 中川慧(日興グローバルラップ, 筑波大学),
吉田健一(筑波大学)
12. 新しいポートフォリオ構築アルゴリズムに向けたモーメンタム効果のモデル化 51
海野一則(東京工業大学), 菊地剛正(慶應義塾大学), 國上真章(東京工業大学),
山田隆志(山口大学), 寺野隆雄(東京工業大学)

13. 決算短信から抽出した業績要因文の事業セグメントに基づく分類と業績文の抽出 ……………	59
村野壮人(成蹊大学), 酒井浩之(成蹊大学), 坂地泰紀(東京大学), 江口潤一(大和証券投資信託委託)	
14. 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 ……………	65
小林和正(成蹊大学), 酒井浩之(成蹊大学), 坂地泰紀(東京大学), 平松賢士(アイフィスジャパン)	
15. 業種別企業業績要因を含む新聞記事の抽出 ……………	71
丸澤英将(東京大学), 和泉潔(東京大学), 坂地泰紀(東京大学), 田村浩道(野村證券)	
16. 語の類義性・対義性を考慮したドメイン特化型辞書構築手法の提案と評価 ……………	78
伊藤諒(東京大学), 坂地泰紀(東京大学), 和泉潔(東京大学), 須田真太郎(三菱 UFJ トラスト投資工学研究所)	
17. 日足・4 時間・15 分のチャートの CNN による最適化 ……………	86
河合継(クリスタルメソッド), 小澤昂(東京工業大学), 大川堯郁(東京大学)	
18. 外国為替市場の個々のトレーダのトラッキング解析: ミクロモデルの提案とその平均場理論 ……	92
金澤輝代士(東京工業大学), 末重拓己(東京工業大学), 高安秀樹(Sony Computer Science Laboratories), 高安美佐子(東京工業大学)	
19. ファンダメンタルファクターモデル(リターンモデル)における機械学習手法の応用可能性検証 …	95
杉友盛佑(エピックパートナーズインベストメンツ), 南正太郎(あすかアセットマネジメント)	
20. 深層学習と高頻度注文情報による株価動向推定 ……………	98
田代大悟(東京大学), 和泉潔(東京大学)	
21. 外国為替取引におけるプロスペクト理論 ……………	104
落合友四郎(大妻女子大学), Jose C. Nacher(東邦大学)	
22. 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 ……………	109
水門善之(野村證券), 勇大地(マイクロソフト コーポレーション)	
23. 敵対的学習を応用したエージェントベース・モデルのベイズ推定 ……………	112
塩野剛志(クレディ・スイス証券株式会社)	

忍耐強い (Patient) アクティブ投資は市場を効率的にするのか？ —人工市場によるシミュレーション分析—

水田 孝信*¹ 堀江 貞之*²
Takanobu Mizuta Sadayuki Horie

*¹ スパークス・アセット・マネジメント株式会社 SPARX Asset Management Co., Ltd.
*² 株式会社 野村総合研究所 Nomura Research Institute, Ltd.

値上がり期待できる株式を選別して投資するアクティブ投資は、投資先企業に本源的に存在する価値に基づいて投資するため、企業価値に即した適正な価格を発見し、市場価格をその価格に近づける（市場を効率的にする）という主張がある。しかしながら、実際のアクティブ投資が市場を効率的にしているかどうかは分かっていない。近年の実証分析は、比較対象となるインデックス（日経平均株価などの指数）から大きく乖離した割合で銘柄を保有しインデックスに対して大きく異なる利益となるリスクをとったうえで売買量が少ない、いわゆる“忍耐強い (Patient) アクティブ投資”が利益を得ていることを明らかにした。今後このような投資が増えることが予想されるが、売買量が少ないにも関わらず市場価格に影響を与え市場を効率的にするのかどうかは重要な論点である。そこで本研究では、忍耐強いアクティブ投資の特徴を反映した投資家を導入した人工市場モデルを構築し、このような売買量が少ない投資家が市場価格に影響を与え市場を効率的にするのかを議論した。その結果、忍耐強いアクティブ投資はまれに起こる、市場価格が企業価値に即した適正な価格から大きく乖離して市場が不安定になり、市場がさらに非効率になりそうなどきみに多く売買を行い、市場を効率化することに寄与していることが示された。市場価格の変動が大きくなると、投機的な注文はその変動をさらに大きくすることがあるが、忍耐強いアクティブ投資の注文は、このような増幅を防いでいる可能性も示された。本研究ではさらに、“忍耐強くない (Impatient) アクティブ投資”もモデル化し、アクティブ投資家の数を一定に保ちその構成を変化させた場合も分析した。その結果、忍耐強くない投資家が減り、忍耐強い投資家が増えると、アクティブ投資の総利益のみならず、投資家1人あたりの利益も増えることが分かった。つまり、忍耐強い投資家が増えることにより、競争が増えて利益を奪い合うのではなく、仲間が増えることにより売却したい価格に到達しやすくなり投資機会が増え利益を獲得しやすくなることが分かった。このことは、忍耐強い投資は市場が効率的になると利益が減るといよりは、市場が非効率すぎて利益が減るといふ側面もあることを示している。

1. はじめに

株式や債券などに投資するファンドの投資手法には大きく分けて、値上がり期待できる銘柄を選別しそれらに投資する“アクティブ”投資と、日経平均株価などの指数（インデックス）と同じ収益を得られるようにインデックスを構成する銘柄と同じ銘柄を同じ割合で保有する“パッシブ”投資がある。近年、アクティブファンドの収益の平均がパッシブファンドより少ないことを主張する実証研究 [French 08, Bogle 14] があることや、特に米国において手数料が高いファンド*¹を販売する場合に販売員が説明責任を負う法改正により、パッシブファンドの投資残高が増える一方、アクティブファンドの投資残高は減っている [A.T.Kearney 16, 神山 17]。

アクティブ投資は投資先企業に本源的に存在する価値（ファンダメンタル価値）に基づいて取引を行うため、企業価値に即した適正な価格を発見し、その価格付近に市場価格を近づける（市場を効率的にする）という、資本主義の重要な機能である投資資本の適切な配分を担っているという主張がある（優れたレビューとして [Wurgler 10]）。そのため、アクティブ投資が減りパッシブファンドが増えることは、市場価格が適切に形成されず資本の適切な配分機能が破壊され、社会にとって好ましくないという主張すらある [Fraser-Jenkins 16]。

しかしながら、実際のアクティブ投資が市場を効率的にするかどうかや、どれくらいアクティブ投資が存在すれば市場が効

率的になるかなどは分かっていない。また、パッシブファンドは売買をほとんど行わないため直接市場価格をゆがめているとはいえ、実際、どれくらいパッシブファンドが増えれば市場効率性を損なうのかなどは全く分かっていない [Goodman 16]。

また、アクティブ投資といってもさまざまである。近年の実証分析 [Cremers 16] は、比較対象となるインデックスから大きく乖離した割合で銘柄を保有しインデックスに対して大きく異なる利益となるリスクをとったうえで売買量が少ない、いわゆる“忍耐強い (Patient) アクティブ投資”が利益を得ていることを明らかにした。忍耐強いアクティブ投資では、利益が実現するまで長い期間を必要とし、その間に生じる企業のファンダメンタル価値と市場価格が乖離するために損益が悪化する期間に耐えなければならない。それに耐えられない、忍耐強くない (Impatient) アクティブ投資は、損切などをしてしまい利益を逃してしまう [Cremers 16]。

この利益の差により、今後、忍耐強くないアクティブ投資が減り、忍耐強いアクティブ投資が増えることが予想されるため、売買量が少ないにも関わらず市場価格に影響を与え市場を効率的にするのかどうかはますます重要な論点となる。実際、売買量が少ないアクティブ投資の方がそうでないアクティブ投資よりファンダメンタルをより正確に測定できているので、売買が少なかったとしても、市場を効率化する [Albagli 15, Cremers 15] という主張と、売買が少くないアクティブ投資は市場を効率化しない [Suominen 11] という両方の主張がある。いずれにせよ、どのようなメカニズムで少ない売買で市場が効率化するのかは分かっていない。[Pastor 16] はアクティブ投資の売買量は時系列で大きく変動しており売買量が多いときに利益を得ていることを示したが、これは上記のメカニズムに大きな示唆を与えていると考えられる。

実証研究では上記のような、どのように特定種類の投資家が市場価格に影響を与えるかといった、ミクロ・マクロ相互作用

連絡先: 水田 孝信, mizutata@gmail.com

<http://www.mizutatakanobu.com/>

当日の発表スライドは以下にあります

<http://www.mizutatakanobu.com/20171014x.pdf>

*¹ パッシブファンドは銘柄を選別するための調査を行わないため、アクティブファンドに比べファンドを購入する投資家が負担する手数料が少ないことがほとんどである [French 08, Bogle 14, 杉田 16]。

用を含むメカニズムを分析することは困難である。また、取引参加者に占めるパッシブファンドの割合が現在ほど多かったことは過去ないため、これ以上パッシブファンドが増えた場合の議論を実証研究だけで行うのは困難である。そもそも、どのような投資家がどのくらい存在するのかを測定すること自体容易ではない。価格形成や流動性にはさまざまな要因が複雑に関わっているため、実証分析では投資家の構成割合の変化が与える影響だけを取り出すことは困難である。

このような実社会におけるマイクロ・マクロ相互作用を含むメカニズムや、実社会でまだおきていない状況、および状況の変化の純粋な影響を議論するのにすぐれた手法として、コンピュータ上で仮想的にその状況を作り出し検証する、社会シミュレーションがある。社会シミュレーションは、例えば、自動車道の整備が交通渋滞へ与える影響分析や、テロや火災、伝染病が発生した場合の避難の方法やあるべき対策の分析などで、大きな成果をあげている*2。

金融市場に関する社会シミュレーションは、エージェントベースドモデルの一種である人工市場モデルを用いて行われる*3。人工市場モデルは、架空の投資家であるエージェントと、架空の取引所である価格決定メカニズムから構成され、コンピュータ上で仮想的に金融市場をシミュレーションする。人工市場モデルを用いたシミュレーションでは、これまでにない投資家の分布が与える影響やまったく新しい規制の効果を議論できるうえ、その純粋な影響を抽出できる。これが人工市場シミュレーション研究の強みである。これらの強みはこれまでの伝統的な経済学で使われてきた手法にはないものであるとして、Nature と Science に人工市場モデルに期待を寄せる論者が掲載されている [Farmer 09, Battiston 16]。そして人工市場モデルを用いたシミュレーション研究は、金融市場の規制・制度の議論に貢献*4したり、バブルや金融危機の発生メカニズムの解明に貢献したりした。

しかしながら、売買量の少ない忍耐強いアクティブ投資が、売買量が少ないにも関わらず市場価格に影響を与え市場を効率的にするのかどうかや、このような投資がさらに広がるのが市場を効率的にしていくのかどうかを、人工市場シミュレーションを使って議論した研究はない*5。そもそも、このような忍耐強いアクティブ投資の特徴を反映したエージェントがモデル化されたこともない。

そこで本研究では、実証研究 [Cremers 16] が提唱した忍耐強い (Patient) および忍耐強くない (Impatient) アクティブ投資の特徴を反映したエージェントを導入した人工市場モデルを構築し、売買量が少ない忍耐強いアクティブ投資家が市場価格に影響を与え市場を効率的にするのかどうかや、忍耐強くないアクティブ投資家が減り忍耐強いアクティブ投資家が増えた場合、各々損益がどのようになるかを議論した。

2. 人工市場モデル

実証研究 [Cremers 16] が提唱した忍耐強い (Patient) および忍耐強くない (Impatient) アクティブ投資の特徴を反映した

エージェントをモデル化する。その特徴を表現できる範囲内において、モデルはできる限りシンプルであることは重要である。モデル構築の基本理念は本稿の付録“モデル構築の基本理念”にて説明した。

2.1 エージェント

全エージェント数を N とする。初め、半数のエージェント $N/2$ は株式 1 株を、もう半数のエージェント $N/2$ はキャッシュ C_0 を保有している。 C_0 はエージェントによらず定数である。1 株保有しているエージェントは常に 1 株の売り注文を、株式を保有していないエージェントは常に 1 株の買い注文を出す。そのため、2 株以上の保有、空売り (マイナスの保有株数) は発生しない。

これにより、各エージェントは毎期、注文価格だけを決定すれば注文が一意に定まる。そのため、投資家の行動の違いを注文価格の差異だけに焦点をあてられ結果の解釈がしやすく、保有株数の細かい調整が発生しないため売買量が少ない投資家も表現できる。以下、エージェントの種類ごとの注文価格の決定方法を説明する。

2.1.1 Patient エージェント

実証研究 [Cremers 16] が提唱した、忍耐強い (Patient) アクティブ投資の特徴を反映した Patient エージェントをモデル化する。アクティブ投資では、利益が実現するまで長い期間を必要とし、その間に生じる企業のファンダメンタル価値と市場価格が乖離するために損益が悪化する期間に耐えなければならないが、それに耐えられるのが Patient エージェントである。

Patient エージェントは N_P 体存在する。時刻 t 、エージェント j の注文価格 $P_o^{t,j}$ は、

$$P_o^{t,j} = P_f \exp(d\sigma^j \pm m(\mu^j + 1)). \quad (1)$$

ここで、 d および m は定数であり、 σ^j は j ごとに異なる実数を出力する正規分布乱数、 μ^j は j ごとに異なる実数を出力する 0 から 1 までの一様乱数である。 \pm は買い注文のときは $-$ 、売り注文のときは $+$ とする。

直近 (時刻 $t-1$) の市場価格 P^{t-1} に依存せず、株式がもつ本源的な価値 (ファンダメンタル価格 P_f) を元に注文価格 $P_o^{t,j}$ を決める。各エージェントはファンダメンタル価格 P_f を知らないが推定を試みている。 $d\sigma^j$ は、その推定したファンダメンタル価格と P_f の差の P_f に対する比率を示している。また、ファンダメンタル投資家は推定ファンダメンタル価格から十分安い価格で買おうとし、十分高い価格で売ろうとする傾向があるといわれ、この十分な価格差のことを安全マージン (Margin of Safety) とよぶ [Graham 03]。 $m(\mu^j + 1)$ は安全マージンの推定ファンダメンタル価格に対する比である。

注文価格 $P_o^{t,j}$ は直近の市場価格 P^{t-1} にまったく依存せず、エージェントの損益がどのようになっても注文価格などの注文行動に変化がない。そのため、損益の悪化に耐え、目標となる市場価格になるまで売買しないため売買が少なくなるという特徴を持っている。

2.1.2 Impatient エージェント

実証研究 [Cremers 16] が提唱した忍耐強くない (Impatient) アクティブ投資の特徴を反映した Impatient エージェントをモデル化する。アクティブ投資では、利益が実現するまで長い期間を必要とし、その間に生じる企業のファンダメンタル価値と市場価格が乖離するために損益が悪化する期間に耐えなければならないが、それに耐えられない Impatient エージェントは、損切などをしてしまい利益を逃してしまふ [Cremers 16]。

*2 例えば、[出口 09, 和泉 12a] などが詳しい。

*3 優れたレビューとして、[和泉 03, LeBaron 06, Chen 12, 和泉 12b, 水田 14, Mizuta 16, Todd 16, 和泉 17a, 和泉 17b] がある。

*4 人工市場モデルの金融市場の規制・制度の議論への貢献は [水田 14, Mizuta 16, 和泉 17a] が詳しい。また、東京証券取引所の持ち株会社、日本取引所グループは人工市場モデルによる研究を“JPX ワーキングペーパー” (<http://www.jpex.co.jp/corporate/research-study/working-paper/>) として多く公表している。

*5 人工市場モデルを用いてパッシブファンドが価格へ与える影響を議論した研究ならある [高橋 11, Braun-Munzinger 16]。

Impatient エージェントは N_{Imp} 体存在する。通常時の注文価格の決定は、Patient エージェント同様に式 (1) で決定する。ただし、時刻 t 、エージェント j が、株式を保有しているときに、損切価格 $P_{\text{lc}}^{t,j}$,

$$P_{\text{lc}}^{t,j} = P_b \exp(-m(\mu^j + 1)), \quad (2)$$

を下回った場合、株式を売却する *6。ここで、 P_b はすでに保有している株式を購入した価格 (簿価) である。損切を実行した後は、エージェント j が推定するファンダメンタル価格 $P_f \exp(d\sigma^j)$ を市場価格 P^t が上回ったら式 (1) の通常の注文価格での注文を行い、それまでは注文を出さない。

2.1.3 テクニカルエージェント

テクニカルエージェントは N_{T} 体存在する。そのうち半数の $N_{\text{T}}/2$ 体が順張り、残りの半数が逆張りである。順張りの時刻 t 、エージェント j の注文価格 $P_o^{t,j}$ は、

$$P_o^{t,j} = P^t \frac{P^t}{P^{t-tm^j}}. \quad (3)$$

テクニカルエージェントは予想価格と同じ価格で注文をする。そのため、予想リターンは $\log(P_o^{t,j}/P^t)$ となる。順張りの場合、予想リターン $\log(P_o^{t,j}/P^t)$ が過去のリターン $\log(P^t/P^{t-tm^j})$ と一致するとするため、式 (2) が求められる。

逆張りの時刻 t 、エージェント j の注文価格 $P_o^{t,j}$ は、

$$P_o^{t,j} = P^{t-tm^j}. \quad (4)$$

ここで、 tm^j は j ごとに異なる自然数を出力する 1 から tm_{max} までの一様乱数で、 tm_{max} は定数である。逆張りの場合、予想リターン $\log(P_o^{t,j}/P^t)$ が過去のリターンの逆 (逆符号) $-\log(P^t/P^{t-tm^j})$ と一致するとするため、式 (3) が求められる。

現実の金融市場の価格変動を再現するためにテクニカルエージェントが必要であることが知られている *7。そのため、本モデルでもなるべくシンプルなテクニカルエージェントを導入した。

2.1.4 ノイズエージェント

ノイズエージェントは N_{N} 体存在し、時刻 t 、エージェント j の注文価格 $P_o^{t,j}$ は、

$$P_o^{t,j} = P^t \exp(\eta\sigma^{t,j}). \quad (5)$$

ここで η は定数、 $\sigma^{t,j}$ は t および j ごとに異なる実数を出力する正規分布乱数である。本研究では常に十分な量の取引が行われている株式を取り扱う。これまでに述べたエージェントだけでなく注文価格が特定の価格付近に偏り売買があまり成立しないことが多く発生するので、ノイズエージェントを導入した。なお、実際の金融市場においてもこのような、流動性 *8 を供給する市場参加者が多く存在する。

2.1.5 パッシブファンドのモデル化について

本研究では、パッシブエージェントは導入しない。日経平均株価などの指数 (インデックス) と同じ収益を得られるようにインデックスを構成する銘柄と同じ銘柄を同じ割合で保有するパッシブファンドは、設定解約がある場合やインデックスの組

*6 具体的には価格 1 の売り注文を出す。

*7 例えば、[Chen 12]。本稿の付録“モデルの妥当性”も参照。

*8 多くの待機している注文がもたらす売買の成立のしやすさ。例えば、[黒崎 15a, Kurosaki 15b] が詳しい。

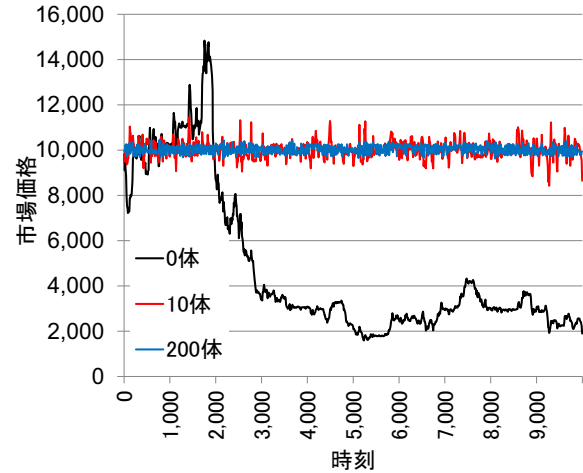


図 1: Patient エージェントの体数 $N_{\text{P}} = 0, 10, 500$ のときの市場価格 P^t の時系列。

み入れ銘柄変更時以外は取引を行わない。本研究では両者ともに考慮していない効果であり、パッシブエージェントをモデル化すると、売買を一切行わないエージェントとなる。そのため、パッシブファンドのみが増加する場合は取り扱えないことに注意が必要である。また、Impatient エージェントがいないときの Patient エージェント数 N_{P} が減少することは、アクティブ投資からパッシブファンドへ資金が移っていることをモデル化していると解釈できる。よって、3.1 節では、アクティブ投資からパッシブファンドへ資金が移っている場合、価格形成がどのようになるかを調べているともいえる。

2.2 価格決定メカニズム

時刻 t のすべてのエージェントの注文価格が決定されたのち、板寄せ方式 (call auction) [東証 15, TSE 15] で取引を成立させ市場価格 P^t を決定する。すなわち、買い注文は高い注文から、売り注文は安い注文から順番につき合わせていき、売買の注文価格が同じになったところを P^t とする。板寄せ方式では、連続的に 1 注文ごとに取引を成立させるのではなく、売りと買いの注文を各々集めておき、ある特定の時刻になったら一括で取引を成立させる。取引価格は需要・供給曲線の交差する価格に決まる。供給 (需要) 曲線は、売り手 (買い手) がこの価格より高く売りたい (安く買いたい) 注文数を累積して描いた曲線である。

3. シミュレーション結果

各種パラメータは、 $N_{\text{T}} = 100, N_{\text{N}} = 1000, C_0 = P_f = 10000, d = 0.05, m = 0.02, tm_{\text{max}} = 100, \eta = 0.5$ とした。また各シミュレーションは $t = t_e = 10000$ まで行った。本モデルおよびこれらのパラメータの妥当性については本稿の付録“モデルの妥当性”で説明した。また、本モデルは他のモデルに比べパラメータが少なく、恣意性が入りにくいのが特徴である。

3.1 Patient エージェントが存在する場合

ここでは、 $N_{\text{Imp}} = 0$ とし、 $N_{\text{P}} = 0, 10, 20, 50, 100, 200, 500$ に対して、その他の条件を乱数表も含め全く同じにして、各種統計値を算出する。これを乱数表を変更して 100 回行い、その平均値を以後用いる (ただし、図 1 のみはある 1 回の試行の

表 1: Patient エージェントの体数ごとの市場非効率性と各エージェント種類別の最終損益および売買数量。

Patient の 体数	市場 非効率性	1 体あたり最終損益			最終損益合計		1 体あたり売買数量			売買数量合計	
		ノイズ	テクニカル	Patient	Patient		ノイズ	テクニカル	Patient	Patient	
0	109.59%	-0.03	0.31	—	—	—	5,000	1,344	—	—	
10	3.11%	-0.04	0.04	3.15	32	32	5,000	1,551	88.08	881	
20	2.04%	-0.05	-0.07	2.66	53	53	5,000	1,674	77.04	1,541	
50	1.59%	-0.07	-0.23	1.90	95	95	5,000	1,744	58.29	2,914	
100	1.47%	-0.10	-0.37	1.35	135	135	5,000	1,772	44.20	4,420	
200	1.35%	-0.12	-0.51	0.86	172	172	5,000	1,805	30.35	6,069	
500	1.24%	-0.14	-0.66	0.42	208	208	5,000	1,840	16.26	8,131	

表 2: Patient エージェント (P) および Impatient エージェント (ImP) の体数ごとの市場非効率性と各エージェント種類別の最終損益および売買数量。

体数		市場 非効率性	1 体あたり最終損益		最終損益合計			1 体あたり売買数量		売買数量合計		
P	ImP		P	ImP	P	ImP	P+ImP	P	ImP	P	ImP	P+ImP
100	100	8.65%	0.49	0.35	49	35	84	17.45	28.34	1,745	2,834	4,579
120	80	2.20%	0.64	0.56	77	45	122	23.13	36.30	2,776	2,904	5,680
140	60	1.67%	0.71	0.62	99	37	136	25.47	37.09	3,565	2,225	5,791
160	40	1.42%	0.77	0.67	124	27	150	27.57	38.22	4,411	1,529	5,939
180	20	1.35%	0.82	0.71	148	14	162	29.09	39.79	5,235	796	6,031
200	0	1.35%	0.86	—	172	—	172	30.35	—	6,069	—	6,069

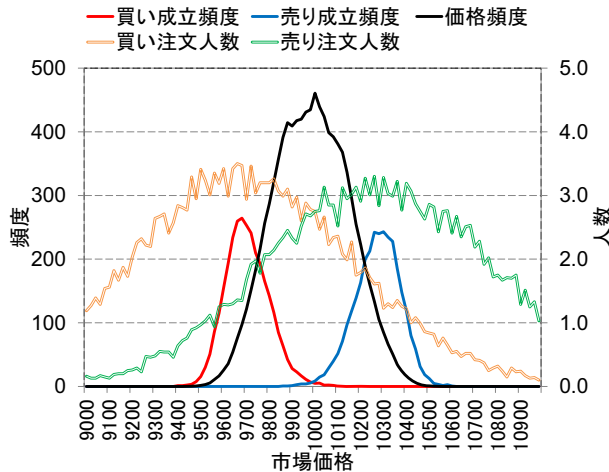


図 2: 市場価格帯別 (20 刻み) の, 市場価格がその価格帯になった頻度 (価格頻度) と Patient エージェントの取引数量および式 (1) の売り買い両方の注文価格 $P^{t,j}$.

結果を用いた。). 2.1.5 節で述べたように, Patient エージェント数 N_P が減少することは, パッシブファンドが増加していることをモデル化していると解釈できるため, 本節ではアクティブ投資が減少 (N_P が減少) しパッシブファンドが増加した場合を調べているといえる。

市場の効率性を測定する指標として, 市場非効率性 M_{ie} ,

$$M_{ie} = \frac{1}{t_e} \sum_{t=1}^{t_e} \frac{|P^t - P_f|}{P_f}, \quad (6)$$

を定義した*9。ここで $||$ は絶対値を示す。 M_{ie} は 0 以上の値をとり, 0 なら完全に効率的, 大きくなればなるほど非効率であることを示す。

図 1 は Patient エージェントの体数 $N_P = 0, 10, 200$ のときの市場価格 P^t の時系列である。 Patient エージェントが増えるほど $P_f = 10000$ 付近に収まっていることが分かり, 市場が効率的になっていることが伺える。 実際, 表 1 は N_P ごとの市場非効率性 M_{ie} を示しているが, N_P が増えるほど, 市場が効率化しているのが分かる。つまり, アクティブ投資が減少すると市場が非効率になる可能性を示している。これは, アクティブ投資からパッシブファンドへ資金が移っていくと, 市場が非効率になる可能性も示している。ただし 2.1.5 節で述べたように, アクティブ投資が減少せず, パッシブファンドが増加する場合は本研究のモデルで調べることはできないことに注意。

表 1 はさらに, N_P ごとの各エージェント種類別の売買数量 (1 体あたりおよびそのエージェントでの合計) も示している。 Patient エージェントの売買数量は他のエージェントと比べて非常に少ないにも関わらず, 市場を効率化している。

これらの理由を調べるため, 図 2 は, 市場価格帯別 (20 刻み) の, 市場価格がその価格帯になった頻度 (価格頻度) と Patient エージェントの取引数量および式 (1) の売り買い両方

*9 市場の効率性を示す指標は多く提案されている [伊藤 07, Verheyden 13] が, 本研究で用いる市場非効率性は, 通常は観測できないファンダメンタル価格 P_f を直接使用しており, 人工市場シミュレーションでしか用いることができない定義である。人工市場シミュレーションでは P_f が明確であるため, 推定ではない理想的な市場の効率性を測定できる市場非効率性の使用が可能である。

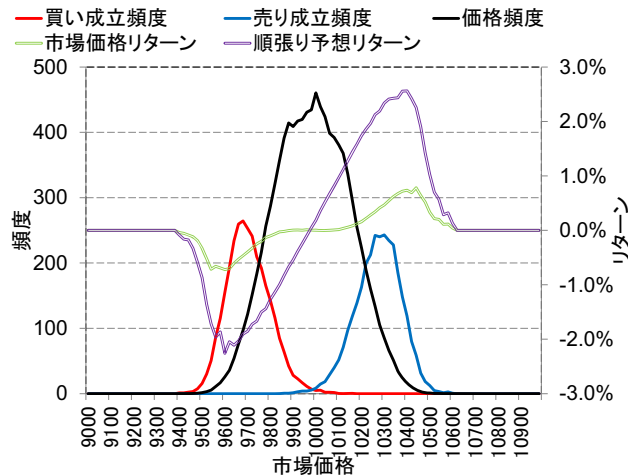


図 3: 市場価格帯別 (20 刻み) の, 市場価格がその価格帯になった頻度 (価格頻度), 平均市場価格リターン $\log(P^t/P^{t-1})$ および順張りのテクニカルエージェントの予想リターン ($\log(P_o^{t,j}/P^t)$).

の注文価格 $P_o^{t,j}$ を示した。注文価格は広く分布しているにも関わらず、実際に成立した買い (売り) 注文は、真のファンダメンタル価格 $P_f = 10000$ から離れた $P^t = 9800$ ($P^t = 10300$) 付近であった。これは真のファンダメンタル価格に近い価格を推定できた (σ^j が小さい) Patient エージェントのみが、買いと売りの両方の注文価格が実現した市場価格に含まれたことを示している。例えば、 σ^j がプラスに大きい場合、買い注文 $P_f \exp(d\sigma^j - m(\mu^j + 1))$ は P_f 付近の多くの実現した市場価格であったとしても、売り注文 $P_f \exp(d\sigma^j + m(\mu^j + 1))$ が高すぎて市場価格が到達せず売却できない。このように市場価格が高頻度になる価格帯である P_f 付近で買い (売り) を行う Patient エージェントは売り (買い) 注文が市場価格が実現しない価格となっているため取引が実現しない。そのため、市場価格が頻度は低い実現する価格帯であり、かつ、注文価格が多く存在する価格帯で、Patient エージェントの注文が成立することとなる。逆に言えば、Patient エージェントの注文は、買い (売り) 注文が多く成立する価格帯よりも安くなる (高く) なるのを防いでおり、その結果、注文が成立する価格帯が市場価格が実現する下限 (上限) となっているともいえる。このような価格帯は P_f から離れ市場が非効率になっているときである。つまり、Patient エージェントは市場が非効率になったときのみ売買を行っているといえる。

図 3 は、市場価格帯別 (20 刻み) の、平均市場価格リターン $\log(P^t/P^{t-1})$ および順張りのテクニカルエージェントの予想リターン $\log(P_o^{t,j}/P^t)$ を図示した。Patient エージェントの買い (売り) 注文が成立している価格帯は、市場価格リターンおよび順張りのテクニカルエージェントの予想リターンが共にマイナス (プラス) になっている。すなわち、市場価格が P_f より低く (高く) なりさらに低く (高く) なりそうなとき、つまり市場価格の変動がさらに大きくなりそうなときに、これを防ぐ方向の Patient エージェントの買い (売り) 注文が成立している。この売買行動は [Pastor 16] の実証研究の結果と整合的である。また、この市場効率化のメカニズムは、[Albagli 15, Cremers 15] が主張するメカニズムとは異なるものである。

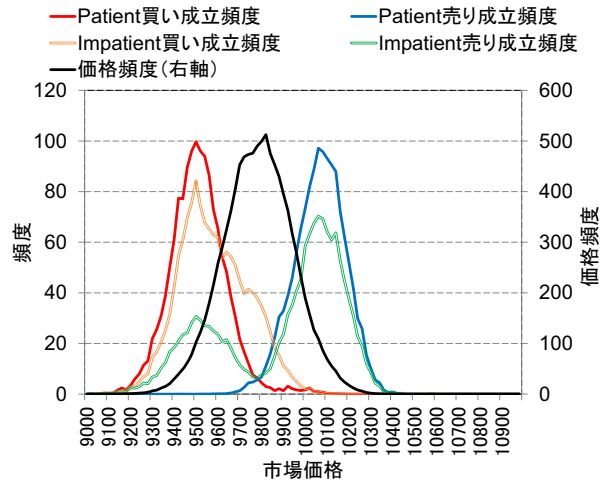


図 4: $N_P = 120, N_{Imp} = 80$ のときの、市場価格帯別 (20 刻み) の、市場価格がその価格帯になった頻度 (価格頻度)、Patient と Impatient エージェントの売買注文の成立頻度。

このように、Patient エージェントはまれに起こる P^t が P_f から大きく乖離して市場が不安定になり、市場がさらに非効率になりそうなときのみによく売買を行い、市場価格に影響を与え、市場を効率化することが示された。さらに、順張りのテクニカルエージェントの注文価格は市場価格の変動が大きくなるとさらにその変動をさらに大きくするが、Patient エージェントの注文は、このような増幅を防いでいることも示された。つまり、売買数量は少なくとも、市場が非効率化しそうなときのみ売買を行い、市場効率化に大きく関与している可能性が示された。このことは、アクティブ投資からパッシブファンドへ資金が移ると市場を非効率にする可能性を示したと考えられる。

表 1 は、 N_P ごとの各エージェント種類別の平均の最終損益も示している。最後まで保有した株式の価格は P_f として、損益を計算した。Patient エージェントは他の種類のエージェントより高い収益となっている。また、 N_P が多くなるにつれて、1 体あたり最終損益が少なくなっているもののマイナスにまではなっていないことが分かる。一方、最終損益合計は増えている。1 体あたりおよび合計のいずれも、売買数量と同じ傾向となっている。すなわち、Patient エージェントの増加により売買数量合計は増加し損益合計も増加するが、その増加量はエージェント数の増加ほどではないため、1 体あたりの売買数量、最終損益は減少してしまうと考えられる。

3.2 Impatient および Patient エージェントが存在する場合

ここでは、 $\{(N_P, N_{Imp})\} = \{(100, 100), (120, 80), (140, 60), (160, 40), (180, 20), (200, 0)\}$ の場合を調べた。 $N_P + N_{Imp} = 200$ で一定であることに注意。

表 2 は、 N_P, N_{Imp} ごとの市場非効率性 M_{ie} , Patient と Impatient エージェントの 1 体あたりおよび合計の最終損益、売買数量を示した。 N_P が増えて N_{Imp} 減るにつれて、 M_{ie} は減少し市場は効率的になっている。合計および 1 体あたりのいずれも、売買数量は Impatient エージェントの方が多くても関わらず利益は低くなっている。そして最も注目すべきは、前節 (3.2 節) と異なり、Patient エージェントの最終損益

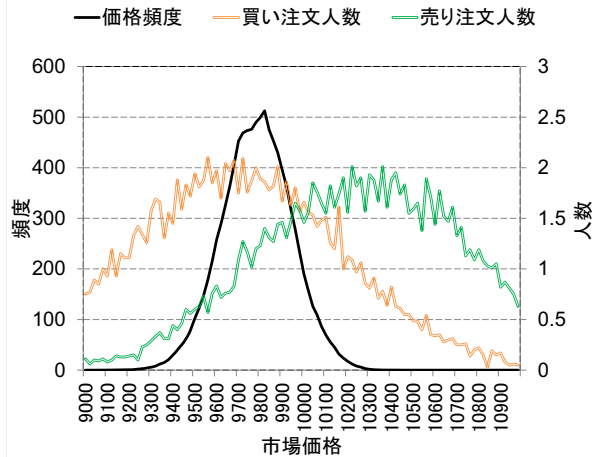


図 5: $N_P = 120, N_{Imp} = 80$ のときの、市場価格帯別 (20 刻み) の、市場価格がその価格帯になった頻度 (価格頻度) および Patient エージェントの式 (1) の売り買い両方の注文価格 $P_o^{t,j}$.

が合計のみならず 1 体あたりでも増加していることである。Impatient エージェントが減り、Patient エージェントが増えると、Patient エージェントにとって競合が増えて利益を奪い合うということはおきず、むしろ、仲間が増え売却したい価格に到達しやすくなり投資機会が増えて利益を獲得しやすくなっていることが考えられる。このことは、忍耐強い投資は市場が効率的になると利益が減るというよりは、市場が非効率すぎて利益が減るという側面もあることを示している。

これらを調べるため、図 4 は $N_P = 120, N_{Imp} = 80$ のときの、市場価格帯別 (20 刻み) の、市場価格がその価格帯になった頻度 (価格頻度) と、Patient と Impatient エージェントの売買注文の成立頻度を示した。Impatient エージェントがない場合の図 2 と比べ、市場価格が全体的に低くなっている。また、Impatient エージェントの売り成立頻度は $P^t = 9500$ あたりにも山があり、買うべきときに損切を行っていることが分かる。この損切が、市場価格が全体的に低くなっている原因と考えられる。また、Impatient エージェントの方が、買い注文が市場価格の頻度が高い $P^t = 9800$ あたりまで延びてきている。つまり、割安でないときの購入が増えてしまっている。これは推定ファンダメンタル価格が高めの Impatient エージェントが、 $P^t = 9800$ 付近で購入し、売り注文価格には市場価格が到達しない一方、損切価格には到達してしまい、割安でない購入と損切を繰り返していることが考えられる。これが Patient エージェントであれば、前節 (3.2 節) で述べたように、売却の機会が訪れずずっと保有したままとなるので売買は発生しない。この差が、Impatient エージェントの売買が多いにも関わらず利益が少ない理由と考えられる。

図 5 は、 $N_P = 120, N_{Imp} = 80$ のときの、市場価格帯別 (20 刻み) の、Patient エージェントの式 (1) の売り買い両方の注文価格 $P_o^{t,j}$ を示した。図 4 で説明したように、Impatient エージェントの損切によって市場価格が全体的に低くなったため、Patient エージェントの売り注文がもっとも多い価格帯 ($P^t = 10200 \sim 10500$) に市場価格が全く到達しなくなっている。つまり、Impatient エージェントが存在することによる価格形成の変化が、Patient エージェントの売却機会を奪い利益

を少なくしている。すなわち、Patient エージェントにとっては、同じ Patient エージェントが利益を奪い合うというよりは、Impatient エージェントが価格形成に与える影響が利益を得る機会を奪っているといえる。

4. まとめと今後の課題

本研究では、実証研究 [Cremers 16] が提唱した忍耐強い (Patient) および忍耐強くない (Impatient) アクティブ投資の特徴を反映したエージェントを導入した人工市場モデルを構築し、売買量が少ない忍耐強いアクティブ投資家が市場価格に影響を与え市場を効率的にするのかどうかや、忍耐強くないアクティブ投資家が減り忍耐強いアクティブ投資家が増えた場合、各々損益がどのようになるかを議論した。

その結果、Patient エージェントはまれに起こる市場価格がファンダメンタル価値から大きく乖離して市場が不安定になり、市場がさらに非効率になりそうなきのみに多く売買を行い、市場価格に影響を与え、市場を効率化することが示された。さらに、順張りのテクニカルエージェントの注文は市場価格の変動が大きくなるとさらにその変動を大きくするが、Patient エージェントの注文は、このような増幅を防いでいることも示された。つまり、売買数量は少なくても、市場が非効率化しそうなきのみ売買を行い、市場効率化に大きく関与している可能性が示された。この売買行動は [Pastor 16] の実証研究の結果と整合的である。また、この市場効率化のメカニズムは、[Albagli 15, Cremers 15] が主張するメカニズムとは異なるものである。

このことは、アクティブ投資が減少すると市場が非効率になる可能性を示している。これは、アクティブ投資からパッシブファンドへ資金が移っている場合、市場が非効率になる可能性も示している。ただし、アクティブ投資が減少せず、パッシブファンドが増加する場合は本研究のモデルで調べることはできないことに注意が必要である。

Impatient エージェントが減り、Patient エージェントが増えると、Patient エージェントにとって競合が増えて利益を奪い合うということはおきず、むしろ、仲間が増え売却したい価格に到達しやすくなり投資機会が増えて利益を獲得しやすくなっていることが分かった。

Impatient エージェントの損切によって市場価格の頻度分布が全体的に低くなったため、Patient エージェントの売り注文がもっとも多い価格帯に市場価格が全く到達しなくなっていた。つまり、Impatient エージェントが存在することによる価格形成の変化が、Patient エージェントの売却機会を奪い利益を少なくしていた。すなわち、Patient エージェントにとっては、同じ Patient エージェントが利益を奪い合うというよりは、Impatient エージェントが価格形成に与える影響が利益を得る機会を奪っていたといえる。このことは、忍耐強い投資は市場が効率的になると利益が減るというよりは、市場が非効率すぎて利益が減るという側面もあることを示している。

今後の課題は、外部環境の変化によって明らかにファンダメンタル価値が変化する銘柄 (シクリカル銘柄) や、ファンダメンタル価値より低い価格で取引され続けた銘柄 (万年割安銘柄) など、さまざまな特徴をもつ銘柄について調べることである。本研究では、ファンダメンタル価値が安定しており、しかもその周りで売買されている銘柄を対象とした。しかし、他の特徴をもつ銘柄では、忍耐強いアクティブ投資の売買行動や、市場価格に与える影響は異なる可能性がある。

また、このような特徴が異なる銘柄が複数存在する場合のシ

表 3: $N_P = 0, N_{ImP} = 0$ の場合の各種統計量

リターンの標準偏差	1.12%
リターンの尖度	2.38
	ラグ
	1 0.18
二乗リターンの	2 0.16
自己相関係数	3 0.15
	4 0.14
	5 0.14

シミュレーションも今後の課題である。アクティブ投資が参照するインデックスには、さまざまな特徴をもった銘柄が含まれているが、[Cremers 16] はこれらの平均損益（インデックスの損益）に比べて負けていることに、アクティブ投資が耐えられるかどうかを議論していた。本研究では 1 銘柄の単純な損益で忍耐強さを議論し、忍耐強くない場合は損切をしてしまうというモデル化を行った。しかしより厳密には、さまざま特徴をもった銘柄の集まりの平均に対する相対的な損益が悪化したときに、インデックスの保有割合に近づけてしまうという売買が行われている。インデックスにさまざまな特徴の銘柄が含まれている場合は、両者は異なるメカニズムを持つ可能性がある。これら調べるためにはモデルの大規模・複雑化が必要であり、今後の課題である。

さらに、アクティブ投資は減少せず、パッシブファンドが増加した場合の分析も今後の課題である。本研究ではパッシブファンドを全く売買しないものとしたため、この場合を扱うことができなかった。しかし、実際のパッシブファンドに資金が入り出した場合はその分の売買を行うので、この売買の影響を分析することは重要であると考えられる。

さらに、既に述べたように、人工市場シミュレーションは投資家の分布の変化の純粋な効果を見ることができるといえる。実現したことがない投資家の分布での分析もすることができる。ただその効果は確実な予想ではない。さまざまなケースでのシミュレーションを行い、これまで予想されていなかった、“あり得る”メカニズムでの現象を見つけておくことが、人工市場シミュレーションの大きな役割であり、人工市場シミュレーションの限界である。そのため、さらなる詳細な議論では、実証分析など他の手法の結果と比較検討する必要がある。

付録

モデル構築の基本理念

人工市場シミュレーションを用いれば、これまでにない投資家の分布が与える影響やまったく新しい規制の効果を議論できるうえ、その純粋な影響を抽出できる。これが人工市場シミュレーション研究の強みである。そして、多くの人工市場シミュレーション研究がこれまでにない投資家の分布が与える影響や、規制・制度の変更を分析してきた [和泉 03, LeBaron 06, Chen 12, 和泉 12b, 水田 14, Mizuta 16, Todd 16, 和泉 17a, 和泉 17b]。

ただその効果は確実な予想ではない。さまざまなケースでのシミュレーションを行い、これまで予想されていなかった、“あり得る”メカニズムでの現象を見つけておくことが、人工市場シミュレーションの大きな役割となる。金融市場でこれから実際におこる現象を定量的にも忠実に再現することが目的ではなく、規制や制度の変更が、どのようなメカニズムで価格形成に影響を与え、どのようなことが起こり得るのかという知識獲得が目的である。これは例えば実証分析など他の手法ではできないことである。

人工市場モデルは普遍的に存在するマクロ現象を再現すべきであると考えられる。人工市場シミュレーションでは、マクロ現象である市場価格のリターンや売買数量をモデル化しない。あくまで、投資家を模した“エージェント”と取引所を模した“価格決定メカニズム”といったマイクロメカニズムをモデル化し、そのマイクロメカニズムの相互作用の積み上げとしてマクロ現象が出力される。そのため、マイクロメカニズムのモデル化は現実の市場に即したものとし、結果と

して出力されるマクロ現象は、現実の市場で普遍的に存在するマクロ的性質を再現されるように作る必要がある。

しかし、普遍的ではなく特定の時期や資産、地域で出現するマクロ的性質すべてを再現することは本研究の目的ではない。必要以上に多くのマクロ的性質一つのモデルで再現しようとすると、過剰に複雑なモデルをもたらす、関連する要素が多くなりすぎて、発生メカニズムの理解を妨げてしまう。

実際、複雑な人工市場モデルに対して、モデルが複雑になるとパラメータが増えモデルの評価が困難になるという批判がある [Chen 12]。モデルが複雑すぎると関連する要素が多くなりすぎて、発生メカニズムの理解を妨げてしまう。また、パラメータが増えるほどさまざまな出力がだせるようになり、モデルを作った人が導きたい結果へ恣意的に導くためのパラメータ設定が行われる恐れがある。シンプルでパラメータが少ないモデルほど、パラメータ調整によって特定の結果に導くことが困難であるため評価が容易となる。

以上により、本研究では、分析目的を果たせる範囲内なるべくシンプルなモデルの構築を行っている。実際の市場を完全に再現することを目的としておらず、普遍的ではなく特定の時期や資産、地域で出現するマクロ的性質すべてを再現することや、実際には存在するであろう投資家をすべて網羅することはあえて行っていない。

[Weisberg 12] が述べているように、よいシミュレーションモデルとはその研究目的によって異なる。そのため、本研究のモデルは本研究の目的にのみおいてよいモデルであり、他の研究目的においてはよいモデルではない。また [Weisberg 12] が述べているように、数理モデルと異なり、シミュレーションモデルは投資家などのミクロの行動やその行動の理由と、市場価格などのマクロ現象との相互作用のメカニズムの解明ができることが強みである。数理モデルが強みとするマクロ現象の特徴分析や予測といったことは本研究の目的とせず、メカニズムの解明に焦点をあてている。

モデルの妥当性

人工市場モデルの妥当性は実証分析で得られている fat-tail や volatility-clustering といった代表的な stylized fact が再現できるかどうかで評価される [LeBaron 06, Chen 12, 和泉 12b, 水田 14, Mizuta 16]。ファット・テールは、市場価格のリターンの分布が正規分布ではなく裾が厚い、すなわち、尖度が正であることである。ボラティリティ・クラスタリングは市場価格のリターンの 2 乗が、大きなラグでも自己相関係数が有意に正であることである。

[Sewell 11] など多くの研究で述べられているように、金融市場は不安定であり、安定的に、どのような時期にも有意に観測されるスタイライズド・ファクトはファット・テールとボラティリティ・クラスタリングの 2 つしかない。

しかも、これらは統計量の有意に正であることだけが安定して観測され、値そのものは、時期によって異なる。ファット・テールについては、実証分析でよく観測されるリターンの分布の尖度は 1 ~ 100 程度であり、ボラティリティ・クラスタリングについては、実証分析でよく観測されるリターンの自己相関係数は 0.01 ~ 0.2 程度と、かなりばらつきがある [Sewell 11]。

本研究のように、金融市場に共通する性質を分析対象とする人工市場が再現すべきは、これらの統計量が有意に正であり、問題ない範囲に値が収まっていることであって、特定の値に近づけることは本質的ではない。

表 3 は、 $N_P = 0, N_{ImP} = 0$ のときの毎期のリターン $\log(P^t/P^{t-1})$ の標準偏差と尖度、リターンの 2 乗の自己相関である。リターンの尖度がプラスで、実証分析でよく観測される 1 ~ 100 程度の範囲に収まっている。ゆえに、ファット・テールが再現されている。また、リターンの 2 乗の自己相関もラグがあってもプラスで実証分析でよく観測される 0.01 ~ 0.2 程度に収まっている。ゆえに、ボラティリティ・クラスタリングが再現されていると考えられる。

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および株式会社野村総合研究所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あります。

参考文献

- [Albagli 15] Albagli, E.: Investment horizons and asset prices under asymmetric information, *Journal of Economic Theory*, Vol. 158, Part B, pp. 787 – 837 (2015), <http://dx.doi.org/10.1016/j.jet.2014.12.008>
- [A.T.Kearney 16] A.T.Kearney, : The \$20 billion impact of the new fiduciary rule on the U.S. wealth management industry, *A.T. Kearney study, Perspective for Discussion*, A.T. Kearney, No. October (2016), <https://goo.gl/SA2EM9>
- [Battiston 16] Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., Hommes, C., Jaeger, C., May, R., and Scheffer, M.: Complexity theory and financial regulation, *Science*, Vol. 351, No. 6275, pp. 818–819 (2016), <http://science.sciencemag.org/content/351/6275/818>
- [Bogle 14] Bogle, J. C.: The arithmetic of “all-in” investment expenses, *Financial Analysts Journal*, Vol. 70, No. 1, pp. 13–21 (2014), <http://www.cfapubs.org/doi/pdf/10.2469/faj.v70.n1.1>

- [Braun-Munzinger 16] Braun-Munzinger, K., Liu, Z., and Turrell, A.: Staff Working Paper No. 592 An agent-based model of dynamics in corporate bond trading, *Bank of England, Staff Working Papers* (2016), <http://www.bankofengland.co.uk/research/Pages/workingpapers/2016/swp592.aspx>
- [Chen 12] Chen, S.-H., Chang, C.-L., and Du, Y.-R.: Agent-based economic models and econometrics, *Knowledge Engineering Review*, Vol. 27, No. 2, pp. 187–219 (2012), <http://dx.doi.org/10.1017/S0269888912000136>
- [Cremers 15] Cremers, M. and Pareek, A.: Short-Term Trading and Stock Return Anomalies: Momentum, Reversal, and Share Issuance, *Review of Finance*, Vol. 19, No. 4, p. 1649 (2015), <https://doi.org/10.1093/rof/rfu029>
- [Cremers 16] Cremers, M. and Pareek, A.: Patient capital outperformance: The investment skill of high active share managers who trade infrequently, *Journal of Financial Economics*, Vol. 122, No. 2, pp. 288–306 (2016), <http://dx.doi.org/10.1016/j.jfineco.2016.08.003>
- [出口 09] 出口 弘, 木嶋 恭一: エージェントベースの社会システム科学宣言—地球社会のリベラルアーツめざして, 勁草書房 (2009), <http://www.keisoshobo.co.jp/book/b26210.html>
- [Farmer 09] Farmer, J. D. and Foley, D.: The economy needs agent-based modelling, *Nature*, Vol. 460, No. 7256, pp. 685–686 (2009), <https://www.nature.com/articles/460685a>
- [Fraser-Jenkins 16] Fraser-Jenkins, I.: The Silent Road to Serfdom: Why Passive Investing is Worse Than Marxism, *Sanford C. Bernstein research report* (2016)
- [French 08] French, K. R.: Presidential Address: The Cost of Active Investing, *The Journal of Finance*, Vol. 63, No. 4, pp. 1537–1573 (2008), <http://dx.doi.org/10.1111/j.1540-6261.2008.01368.x>
- [Goodman 16] Goodman, B.: Is Passive Investing Really Worse Than Marxism? Despite doing everything in its power to demonize index funds, an Alliance Bernstein team fails to prove they actually harm markets., *Barron's on The Wall Street Journal, Dow Jones & Company* (2016), <https://goo.gl/imMA2I>
- [Graham 03] Graham, B. and Zweig, J.: *The Intelligent Investor: The Definitive Book on Value Investing*, HarperCollins (2003)
- [伊藤 07] 伊藤 幹夫, 杉山 俊輔: 市場効率性の時変構造, in *KEIO Economic Society Discussion Paper*, No. 5, 慶応義塾大学 (2007), http://web.econ.keio.ac.jp/staff/ito/pdf06/KESDP_06-6.pdf
- [和泉 03] 和泉 潔: 人工市場, 森北出版 (2003), <https://www.morikita.co.jp/books/book/2239>
- [和泉 12a] 和泉 潔, 実世界とエージェントシミュレーション協同研究委員会: 実世界とエージェントシミュレーション, 電気学会 (2012), http://www.bookpark.ne.jp/cm/ieej/detail.asp?content_id=IEEJ-GH1262-PRT
- [和泉 12b] 和泉 潔: 第3章 金融市場—人工市場の観点から, 杉原 正顕 (編), 計算と社会 (岩波講座 計算科学 第6巻), 岩波書店 (2012), <http://iss.ndl.go.jp/books/R100000002-I023728505-00>
- [和泉 17a] 和泉 潔, 川久保 佐記, 米納 弘渡: 第5章 強靱な金融システム, 古田 一雄 (編), レジリエンス工学入門, 日科技連出版社 (2017), <http://www.juse-p.co.jp/cgi-bin/html.pl5?i=ISBN978-4-8171-9624-8>
- [和泉 17b] 和泉 潔: 第6章 可能世界ブラウザとしてのエージェントシミュレーション, マルチエージェントのためのデータ解析 (マルチエージェントシリーズ), コロナ社 (2017), <http://www.coronasha.co.jp/np/isbn/9784339028126/>
- [神山 17] 神山 哲也, 岡田 功太: アクティブ運用の苦境と資産運用業界再編の可能性—英ヘンダーソンと米ジャナス合併の事例—, 資本市場クォーターリ, 野村資本市場研究所, 冬号 (2017), <http://www.nicmr.com/nicmr/report/repo/2017/2017win10.html>
- [黒崎 15a] 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シリーズ, No. 15-J-2 (2015), https://www.boj.or.jp/research/wps_rev/wps_2015/wp15j02.htm/
- [Kurosaki 15b] Kurosaki, T., Kumano, Y., Okabe, K., and Nagano, T.: Liquidity in JGB Markets: An Evaluation from Transaction Data, *Bank of Japan Working Paper Series*, No. 15-E-2 (2015), https://www.boj.or.jp/en/research/wps_rev/wps_2015/wp15e02.htm/
- [LeBaron 06] LeBaron, B.: Agent-based computational finance, *Handbook of computational economics*, Vol. 2, pp. 1187–1233 (2006), [http://dx.doi.org/10.1016/S1574-0021\(05\)02024-1](http://dx.doi.org/10.1016/S1574-0021(05)02024-1)
- [水田 14] 水田 孝信: 人工市場シミュレーションを用いた金融市場の規制・制度の分析, 博士論文, 東京大学大学院工学系研究科 (2014), <http://hdl.handle.net/2261/59875>
- [Mizuta 16] Mizuta, T.: A Brief Review of Recent Artificial Market Simulation (Multi-Agent Simulation) Studies for Financial Market Regulations and/or Rules, *SSRN Working Paper Series* (2016), <http://ssrn.com/abstract=2710495>
- [Pastor 16] Pastor, L., Stambaugh, R. F., and Taylor, L. A.: Do Funds Make More When They Trade More?, *SSRN Working Paper Series* (2016), <http://ssrn.com/abstract=2524397>
- [Sewell 11] Sewell, M.: Characterization of financial time series, *Research Note, University College London, Department of Computer Science*, No. RN/11/01 (2011), <http://finance.martinsewell.com/stylized-facts/>
- [杉田 16] 杉田 浩治: 発足から40年を迎えるインデックスファンド—その軌跡と今後の展開—, 証券レビュー, 日本証券経済研究所, Vol. 56, No. 1 (2016), <http://www.jsri.or.jp/publish/review/pdf/5601/04.pdf>
- [Suominen 11] Suominen, M. and Rinne, K.: A Structural Model of Short-Term Reversals, *SSRN Working Paper Series* (2011), <http://ssrn.com/abstract=1787270>
- [高橋 11] 高橋 大志: 社会シミュレーションによる金融市場分析, 横幹連合コンファレンス予稿集, Vol. 2011, pp. 69–69 (2011), <http://doi.org/10.11487/oukan.2011.0.69.0>
- [Todd 16] Todd, A., Beling, P., Scherer, W., and Yang, S. Y.: Agent-based financial markets: A review of the methodology and domain, in *Computational Intelligence for Financial Engineering Economics (CIFER), 2016 IEEE Symposium Series on Computational Intelligence on* (2016), <https://doi.org/10.1109/SSCI.2016.7850016>
- [東証 15] 東証: 東証公式 株式サポーター 株式取引編, 東京証券取引所 (2015), <http://www.jpx.co.jp/learning/tour/books-brochures/detail/08.html>
- [TSE 15] TSE, : *Guide to TSE Trading Methodology*, Tokyo Stock Exchange (2015), <http://www.jpx.co.jp/english/equities/trading/domestic/>
- [Verheyden 13] Verheyden, T., De Moor, L., and Bossche, Van den F.: A Tale of Market Efficiency, *Review of Business and Economic Literature*, Vol. 58, No. 2, pp. 140–158 (2013), http://www.rebel-journal.org/table_of_content.aspx?sy=2013&pn=2
- [Weisberg 12] Weisberg, M.: *Simulation and Similarity: Using Models to Understand the World*, Oxford Studies in the Philosophy of Science (2012), (邦訳: 松王政浩: 科学とモデル-シミュレーションの哲学 入門-, 名古屋大学出版会 (2017)) <http://www.unp.or.jp/ISBN/ISBN978-4-8158-0872-3.html>
- [Wurgler 10] Wurgler, J.: On the Economic Consequences of Index-Linked Investing, Working Paper 16376, National Bureau of Economic Research (2010), <http://www.nber.org/papers/w16376>

約定価格の量子化ボラティリティモデル

Stochastic Volatility Model for Transaction data

森谷博之¹

¹ Quasars22 Private Limited

Abstract:

A financial market is a diversified dynamical system with many constraints, and price movements are modeled in terms of the micro properties of each transactions and the macro properties of dynamical systems. These two properties must be bridged by the multiplicity. The model focuses on the size of tick and the number of transactions with the price movement compared with the previous transactions and explains stochastic nature of short-term volatilities and persistently stable long-term volatilities as a results of unique behaviour of heterogeneous market participants.

1 はじめに

日経225miniの先物市場では、取引が成立するとき、取引の最小単位、または呼び値の単位は5円と大阪取引所により規定されている。したがって、価格の動きは呼び値の単位の整数倍となる。また、実際の約定価格は、約定の度に価格が動く場合もあれば、価格が一定のまま約定が連続する場合もある。本稿では前者を緊急性取引、後者を非緊急性取引として分類する。多くの場合、緊急性の取引の数は、非緊急性の取引の数に比べて大幅に少ない。その前の約定価格と同じ価格で取引が成立する非緊急性取引の数は全体の8割近くに達する。マーケットメイカー(MM)が作り出すビット・アスク・バウンスの影響を受けた値動きも非緊急性取引と判断するとその数は9割を大きく超える。したがって、値動きを伴う緊急性を要する取引の数は、全体の数%にとどまる。

そこで本稿ではまずは緊急性の取引から生じる価格の動きに焦点を当てた。全ての動きを情報として維持するために、連続した緊急性取引の約定価格の差を2乗し、その累積和(sum of squared price increments: sspi)を計算したところ、安定性していることが分かった。この累積和の特徴は特定の期間のリターンから算出したボラティリティに比べると、ランダムな価格の動きによる、価格変動の相殺という現象を排除できるところにある。

つぎに、市場に流動性を供給するMMは常に逆選択リスクに曝され、安定した収益の確保が難しく、かつ、投資家は、投資の開始時と終了時の取引費用の最小化に困難を見出している。そこで、緊急性取引から生じる価格の動きの特徴を調べてみたところ、将来の価格の動きはおおむね予測不能であることが分かった。ゆ

えに、このような状態に置かれたMMと投資家は、それぞれの投資行動を最適化するために活動し、その活動を通して、緊急性取引の数が極端に多くなり、またsspiが安定するのではないかという仮説を考察した。

2 量子化ボラティリティモデル

日経225miniの先物価格(以後先物価格)の動きは

$$S(N) = S(0) + \sum_{n=1}^N x(n),$$

と書くことができる。ここで $x(n) = S(n) - S(n-1)$ 。 $S(0)$ は初期値、 $S(n)$ は n 番目($n = 1, \dots, N$)の約定価格である。 x はある確率過程に従うが、i.i.d. 確率変数であるとは限らない。

$x(n)$ は呼び値の単位 ξ_0 に正の整数 i を掛けた数値であり、その方向性はプラスにもマイナスにも動く。 ξ は $i, i = 1, 2, \dots, I$ によりラベル付けされている。そして、

$$\xi_{\delta, i} = \delta \cdot \xi_0 \cdot i,$$

と表される。ここで δ が+1のときに、上昇を、-1のときに下落を表す。 $\xi_{\delta, i}$ を約定価格差と呼ぶ。 $i = 0$ のとき、その取引は非緊急性取引と分類される。また、緊急性取引は $i \neq 0$ と定義される。従って、緊急性取引の価格差の集合は

$$\Xi = \{\xi_I\} = (-\xi_I, \dots, -\xi_i, \dots, -\xi_1, \xi_1, \dots, \xi_i, \dots, \xi_I),$$

で与えられる。ゆえに $x(n) \in \Xi$ 。

ここで平方価格差をつぎのように定義する。

$$e_i = \xi_{\delta, i}^2 = (\delta \xi_0 \cdot i)^2,$$

そして、その集合は

$$\Psi = \{e_I\} = (e_1, \dots, e_i, \dots, e_I).$$

で与えられる。

N 個の取引のあるときの、平方価格差の累積和 (the sum of squared price increment:sspi) を

$$E = \sum_{n=1}^N x(n)^2,$$

と書く。 $x(n)^2$ は e_i の中のひとつをとる、または $x(n)^2 \in \Psi$ である。

2.1 配位、緊急取引の数、そして平方価格差の和の分布

配位は、連続した N 個の緊急取引の i でラベル付けされた平方価格差の数を頻度として表現していて、それは確率論でいう古典的占拠問題 (occupancy problem) と同じである。

配位は

$$\{N_I\} = (N_1, \dots, N_i, \dots, N_I),$$

と表現される。ここで、 N_i は i でラベル付けされた緊急取引の数であり、 N_I は最も大きな平方価格差の数である。従って、 $N = \sum_{i=1}^I N_i$ が成り立つ。 N の数が十分に大きく、 E が一定であるという前提の下では、 N_i の分布は

$$\frac{N_i}{N} = \frac{e^{-be_i}}{Z(I, b)},$$

で与えられる [?]。ここで、 $Z(I, b)$ は分配関数、または状態和と呼ばれるもので、 $\sum_{i=1}^I e^{-be_i}$ で与えられる。 b はラグランジュ乗数であり、 e_i の平均値の逆数に相当する。

つぎに、この配位の分布から無作為に $N^{(k)}$ 個取り出し、その和 E_j を求めてみる。 e_i は指数分布にしたがうのであるから、その和である $E_j, j = 1, \dots, J$ はガンマ分布に従う。 E_j は

$$P(E_j, N^{(k)}, b) = \frac{b^{N^{(k)}} E_j^{N^{(k)}-1} \exp(-bE_j)}{\Gamma(N^{(k)})}$$

で与えられる。ここで、 $\Gamma()$ ガンマ関数である。また、 $N^{(k)} \in (0, \dots, k, \dots, N)$ である。それぞれの $N^{(k)}$ 個の取引は、 E_j の分布の形状を特定し、 b はスケールパラメータ、 $N^{(k)}$ は形状のパラメータである。

$N^{(k)}$ が十分に大きく成れば、 E_j の分布は正規分布で近似できる。

2.2 時間間隔、緊急取引の数、そして平方価格差の和の分布

つぎに、平方価格差の和を緊急取引の数で特定するのではなく、時間間隔で特定してみよう。その際には、時間間隔当たりの緊急取引の数は変動する。緊急取引の数と平方価格差の和の関係は

$$P(E_j, J, \beta) = \frac{M_j}{M} = \frac{\exp(-\beta E_j)}{Z(J, \beta)}, \quad (1)$$

で与えられる [?]。ここで、 $Z(J, \beta) = \sum_{j=1}^J \exp(-\beta E_j)$ であり、 β はラグランジュ乗数であり、 $1/\bar{E} = M/E$ で与えられる。

3 データ分析

分析データは、2016年1月から2017年4月までの日経225miniの約定価格データを用いた。それぞれの期間と分析に用いた限月は表??の通りである。

表 1: 日経 225 ミニ 分析期間と限月

年月	限月
201601	2016年3月限
201602	2016年3月限
201603	2016年6月限
201604	2016年6月限
201605	2016年6月限
201606	2016年9月限
201607	2016年9月限
201608	2016年9月限
201609	2016年12月限
201610	2016年12月限
201611	2016年12月限
201612	2017年3月限
201701	2017年3月限
201702	2017年3月限
201703	2017年6月限
201704	2017年6月限

3.1 ビッド・アスク・バウンス効果と緊急性取引

マーケットメイカーの役割は市場に流動性を供給することである。しかし、マーケットメイカーは、情報の非対称性の問題を抱えていて、時として逆選択の状態

に陥ってしまう[?].このような問題の解決策の1つとして、マーケットメイカーは取引の高速化により、極力短い期間で確実に収益を上げる取引を繰り返し、そこから得られる収益で、逆選択で生じた損失を補うという戦略を採用している[?].しかし、このような取引戦略では、最良買値と最良売値の間で多くの取引が約定し、それが価格の動きに負の自己相関をもたらしてしまう。このような現象をビッド・アスク・バウンス (BAB) という。このような取引から生じる負の自己相関は、緊急性取引の分析に偏りをもたせる可能性がある。従って本稿の分析ではこのような値動きも非緊急性取引と分類し、緊急性取引とは別の取引とした。表??は各取引の数を示している。非緊急性取引が全体の取引の9割5分を超えている。また、マーケットメイクに伴う、またはビッド・アスク・バウンス効果を生み出す取引は緊急性取引の5倍前後となる。

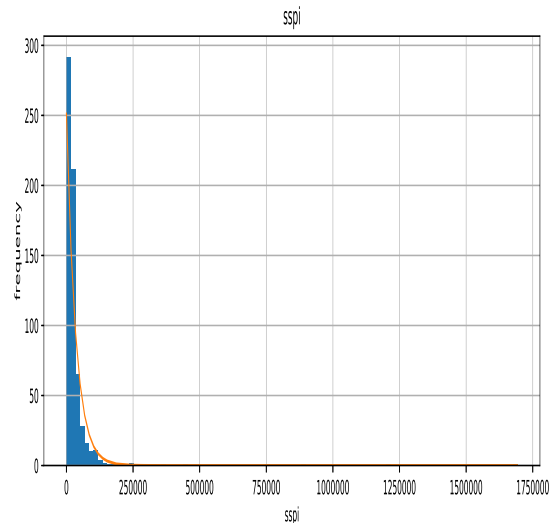
表 2: 各種取引の数

年月	全取引	非緊急性	緊急性	BAB
201601	3,137,278	3,012,221	124,102	568,114
201602	3,189,848	3,076,035	112,821	557,071
201603	1,408,362	1,362,491	45,115	228,067
201604	2,255,747	2,195,408	59,325	369,885
201605	1,575,588	1,539,075	35,576	245,469
201606	1,746,743	1,679,812	66,123	291,488
201607	2,973,356	2,898,793	57,089	292,719
201608	3,523,295	3,453,095	27,757	219,453
201609	2,512,357	2,463,202	22,897	151,000
201610	2,751,476	2,694,886	17,218	157,523
201611	4,275,559	4,196,766	42,161	258,711
201612	2,638,139	2,586,687	21,722	150,247
201701	3,727,576	3,662,157	29,855	217,559
201702	3,611,064	3,547,905	26,376	201,783
201703	2,727,687	2,680,420	19,544	151,320
201704	3,806,572	3,748,768	23,908	215,889

3.2 平方価格差の和の安定性

日経 225mini の取引は日中立会 (8:45-15:15) と、ナイトセッション (16:30-翌日 5:30) の2つの立会で行われる。それぞれの立会について平方価格差の和を算出した。緊急性取引の数が十分にあり、累積平方価格差 (E) が大きな N に対して一定であれば、各立会で算出された平方価格差の和 (E_j) の分布は、式 (??) で与えられる分布と整合性があるはずである。得られた結果をヒストグラムに表わしたのが図??である。

図 1: 平方価格差の和の分布



非常に良い当てはまりを見せている。この期間には、BREXIT とトランプの大統領選挙の勝利、そして、9回の日銀政策決定会合があり、そのつど平方価格差の和には大きな影響を与えた。しかし、その動きは式 (??) で与えられる範囲を超えるものではなかった。

3.3 緊急性取引に伴う値動きのランダム性

??節の結果から予想されるように、緊急性取引の価格の動きはランダムである可能性が高い。そこで、連検定、ダービン・ワトソンの検定を用いて、緊急性取引の価格の動きのランダム性を調べてみた。

日中立会の取引時間を 8:45-9:00, 9:00-10:00, 10:00-11:00, 11:00-12:00, 12:00-13:00, 13:00-14:00, 14:00-15:00, 15:00-15:10 に分け、またナイトセッションの立会時間を 6:30-17:00, 17:00-18:00, 18:00-19:00, 19:00-20:00, 20:00-21:00, 21:00-22:00, 22:00-23:00, 23:00-24:00, 0:00-1:00, 1:00-2:00, 2:00-3:00, 3:00-4:00, 4:00-5:00, 5:00-5:25 に分割し、22の時間帯に分け、それぞれの時間帯で30個以上の緊急性取引が観測されたときに、これら2つの検定を施した。

3.3.1 連検定

連検定はノンパラメトリックな無規則性の検定 [?] である。

H_0 : 緊急性取引の価格の動きには規則性はない。 H_1 : 緊急性取引の価格の動きには規則性がある。として検定を行い、p-値が 0.1 より大きければ帰無仮説を棄却できないと判断した。結果は表??に示した。数値は帰無仮説を棄却できなかった割合である。全体では 6 割弱の時間枠で無規則性が観測された。

3.3.2 ダービン・ワトソン (DW) の検定

ダービン・ワトソンの検定は回帰分析の残差の系列相関の分析に用いられる [?]。ここでは緊急性取引の価格の動きには系列相関がないを帰無仮説とする。ダービン・ワトソン比が 1.5 から 2.5 の間にあるとき、帰無仮説を棄却できないと判断する。表??の結果は帰無仮説を棄却できない割合である。多くの時間枠で系列相関は見られなかった。

表 3: 無規則性の検定結果

年月	30 以上	30 未満	連検定	DW 検定
201601	320	1	0.54	0.99
201602	327	1	0.45	0.98
201603	326	42	0.59	0.88
201604	342	7	0.57	0.92
201605	283	27	0.56	0.86
201606	324	46	0.71	0.98
201607	322	44	0.48	0.77
201608	268	159	0.54	0.77
201609	223	177	0.75	0.91
201610	199	201	0.60	0.67
201611	302	98	0.55	0.67
201612	245	188	0.73	0.93
201701	284	83	0.65	0.92
201702	283	117	0.61	0.83
201703	210	230	0.53	0.66
201704	247	166	0.55	0.75

連検定、ダービン・ワトソン検定の結果、緊急性取引の値動きには多くの場合に規則性が発見できなかった。

4 観察結果の考察

??、??節の結果から緊急性取引の多くの値動きから規則性は発見されなかった。従って、マーケット・メイカーにとっても安定的に利益を上げることは難しく、また投資家がポジションを取る場合、またはポジションを閉じる場合にも、それに伴う費用の削減は易しくなことが分かる。このような状況の下で、市場では多

くの市場参加者が、それぞれの目的を達成するために、それぞれの行動を切磋琢磨し、多重度の上昇を接点として市場に安定をもたらしていると考えられる。その結果として累積平方価格差と平方価格差の和は長期的には安定し、短い時間間隔では、その時々起きた予期せぬ事象を吸収するために、平方価格差の和は、急激な上昇を示すと考えられる。

参考文献

- [1] Moriya,H.:Quantized price volatility model for transaction data,' *Evolutionary and Institutional Economics Review* (2017) <https://doi.org/10.1007/s40844-017-0078-1>
- [2] Cartea,A., and Penalva,J.: Where is the value of high frequency trading?, *Quarterly Journal of Finance* 2(3) September(2012)
- [3] Bouchard,B.,Dang,N-M., and Lehalle,C-A. :Optimal control of trading algorithms- a general impulse control approach, *SIAM Journal on Financial Mathematics* 2(1)403-438(2011)
- [4] 武藤真介:統計解析ハンドブック、朝倉書店 (1995)

能動的キャンセルを含む人工市場によるキャンセル寿命の 冪分布の再現

Reproduction of Power-law Distribution of Cancelled Order Lifetime using Artificial Market that Agents Submit Cancellation Orders Actively

吉村 勇志¹ 陳 昱¹

Yushi Yoshimura¹, Yu Chen¹,

¹ 東京大学大学院新領域創成科学研究科人間環境学専攻

¹Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo

Abstract: 多くの人工市場モデルにおいて、エージェントが注文を出す順序や頻度に対して十分な考察が行われておらず、エージェントはランダムもしくは固定の順序に応じて注文を出し、各エージェントの注文頻度はほぼ同一であることが多い。ところが、このようなモデルにおいては市場の状態に応じてエージェントが自分の注文を能動的にキャンセルすることが出来ない。株式市場においてキャンセルは注文総数の40%以上を占め、市場に大きな影響を及ぼしており、これを適切にモデル化することが求められている。本研究はその第一歩として、エージェントがキャンセル注文を出す確率が板に指値注文を出してから経過時間に依存するモデルを構築し、キャンセルされた注文の寿命(キャンセルされるまでの経過時間)の分布の冪乗則を再現した。

1 序論

連続ダブルオークションを採用している金融市場において、売買差益を目的とする市場参加者は執行コスト(execution cost)と遅延リスク(delay risk)の兼ね合いから注文の価格や数量、形式を決定している。即ち、より有利な価格に注文を出す程、それが一定期間内に約定する確率は下がり、また他の市場参加者に意図を読まれる恐れがある。例えば Avellaneda[1]においては板の形状や流入する注文の分布に一定の関数形を仮定することで、特に HFT の代表であるマーケットメイカーを念頭に置きながら、売買差益を目的とするトレーダーの期待効用最大化する注文の価格決定を解析的に導出している。ところが、このモデルは執行コストと遅延リスクの問題を完全に解決しているとは言い難い。何故ならば、市場の挙動を観測し、自分の出した注文が望ましいものでなくなった場合にはキャンセル(し、必要に応じて新たな注文を提出)するという要素が欠けているからである。この問題を解決するには、人工市場を用いて動的環境の中で市場参加者の意思決定を議論する必要がある。

しかしながら、現在の人工市場の多くはエージェントの意思決定が単純であり、執行コストと遅延リスク、及びそこに含まれるキャンセルの問題を十分

に取り扱っているとは言えない。例えば Chiarella[2]においてはエージェントが注文を出す価格は過去のリターンやファンダメンタル価格の重み付け和から導出されるのみであり、注文価格から約定確率を見積もったりはしていない。また、キャンセルに関しても一定時間以上板に存在した注文が自動で削除される仕組みがあるのみであり、エージェントが市場の状態を見て能動的にキャンセルを入れることはない。キャンセルに関してはより現実的な要素を加えたモデルも存在するものの、不十分である。例えば Leal[3]においてはトレーダーが2種類、即ち HFT(High Frequency Trader)と LFT(Low Frequency Trader)に分けられ、前者にはより短いキャンセルまでの猶予時間を与えることにより HFT が高頻度でキャンセルを行うという性質を模擬しているが、一定時間以上約定されなかった注文のみが自動的にキャンセルされるという点には変わらない。Leal[4]においては HFT が能動的キャンセルを行うようにはなったものの、注文の全てが約定した場合に利益が出るかどうかを基準としており、未だに単純に過ぎる。

そこで、動的環境における執行コストと遅延リスクの問題を考察する第一歩として、人工市場におけるエージェントのキャンセルに関する意思決定のモデル化を行う。ここで、注文の価格や数量の決定で

はなくキャンセルを最初に扱う理由は、そのどちらを扱うにしても注文の順序や頻度の問題を扱わなければならない(注文の価格や数量はエージェントがどれくらいの時間待てるかに影響されるが、その忍耐力の強さは注文の頻度にも影響を与える。キャンセルは一度指値注文を出したエージェントが短い時間内にもう一度注文(キャンセル)を出すことであるから、エージェントの注文する順序が単純なランダムや固定順序では上手くいかない)が、実市場の統計データからの検証がキャンセルの場合の方がより容易だからである。というのも、一人の市場参加者が複数の注文を同時に板に置かず、板に自分の注文を置いている状態ではキャンセル注文しか出さないと仮定すると、キャンセルによる注文の寿命(板に注文が出されてから、板上から消えるまでの経過時間)分布を用いて、市場参加者の注文頻度とその前に注文を出してから経過時間との関係を導出することが出来るからである。

本研究の具体的目的は、各エージェントの注文頻度をエージェント及び市場の状態に応じて変化させることで、キャンセルによる注文の寿命分布 τ_c の冪乗則を再現することである。

2 人工市場モデル

2.1 モデル構築背景

図1に示すように、キャンセル及び約定による注文の寿命分布は冪分布に従い、かつその指数は異なっている(前者が $\alpha_c = 2.1$ 、後者が $\alpha_M = 1.5$)ことが知られている。各エージェントの注文頻度をエージェント及び市場の状態に応じて変化させることで、キャンセルによる注文の寿命分布の冪乗則を再現する。

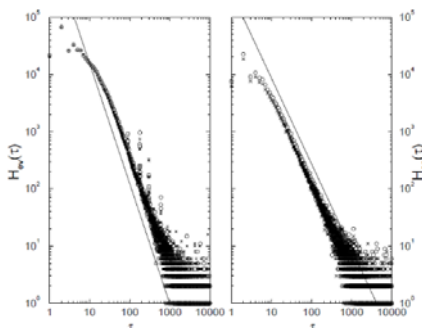


図1 注文の寿命(注文が出されてから板上から消えるまでの時間)分布(左がキャンセル、右が約定、実線は冪関数、指数は本文参照)(Challet[5]より)

執行コストと遅延リスクの問題のうちキャンセル寿命に関してのみ考察するのであるから、関係のな

い要素に関しては簡略化を行う。注文の数量は常に1とし、指値注文と成行注文は確率的に選択される。指値注文は最終約定価格(以下価格)より一定の範囲から確率的に選択される。その確率分布は、以下必要に応じて述べる。また、板上に自分の未約定の注文を持つエージェントは指値・成行注文を行わないものとする。これにより、一人のエージェントが板上の複数の注文をどのような順序でキャンセルするかや、古い注文のキャンセルと新しい注文の提出のどちらを優先するか等、多数の意思決定の問題を除き、考察を単純化することが出来る。

2.2 モデル内容

モデルは離散的な時間ステップ毎に進み、各時間ステップに1人のエージェントのみが市場に注文を出す。1つの時間ステップが次の3つのステップに分かれている。また、エージェントは板上に自分の未約定の注文が存在するかどうかで挙動が異なり、存在するエージェントをL(on LOB)エージェント、しないエージェントをN(Not on LOB)エージェントと以下呼称する。

1)緊急性 u_i の計算

各エージェントは自身の市場参加の緊急性 u_i を計算する。N エージェントの緊急性は $u_{LO,MO}$ とし、これは単純の為定数であるとする。L エージェントの緊急性は $u_c(\tau)$ により計算する。 τ は板上の注文がそのエージェントによって出されてからの経過時間であり、時間ステップ数によりカウントされる。

2)注文エージェントの決定

この時間ステップに注文を出すエージェントを、 u_i に応じてルーレット選択により1人選ぶ。即ち、エージェント j の選択される確率は式(1)で表される。

$$\frac{u_j}{\sum_{i=1}^N u_i} \quad (1)$$

3)エージェントの注文

選択されたエージェントが注文を出す。N エージェントの場合、指値注文と成行注文を確率的に選択する。売買方向はランダムであり、注文の数量は1のみを取るとする。指値注文の価格は最終約定価格から一定の範囲内で、買い(売り)注文の場合には低く(高く)なるという条件下で、確率的に決める。最終約定価格からの距離の従う確率分布は、以下必要に応じて述べる。L エージェントの場合、自分が板上に出している唯一の注文に対しキャンセル注文を出す。時間ステップが1進み、経過時間 τ が τ_{max} に達した注文は削除される。

2.3 モデルの目的と調整パラメータ

キャンセルされた注文の寿命分布が $\tau_c^{-\alpha_c}$ に従うような $u_c(\tau)$ の関数形を見つける。 $(\alpha_c = 2.1)$

3 サブモデルによる $u_c(\tau)$ の考察

エージェントの意思決定を簡略化して依存する変数を注文の経過時間 τ のみとしたが、これでもまだ $u_c(\tau)$ の関数形を見つけるのは困難である。そこで、人工市場モデルの要素の一部を抜き出し、簡略化したサブモデルを構築し、その考察を行う。その後、サブモデルの考察を基に人工市場モデルの $u_c(\tau)$ の関数形を1つ仮定し、人工市場モデルのテストを行う。

3.1 サブモデル1：単一注文モデル

板上に注文が1つのみ存在する場合に、キャンセル率 f_c の経過時間 τ 依存性を調べる。キャンセル率 f_c はルーレット選択における相対的な被選率である $u_c(\tau)$ とは当然異なるが、エージェント数 N が大きければL エージェントの人数や τ の分布は時間的にほぼ一定と見做せるであろうから、定数倍の関係にあると仮定出来、恐らくキャンセル寿命の分布も同じく冪関数になるだろうと予想し、単一注文キャンセルモデルの考察を始めた。

3.1.1 モデル内容

注文が常に1つだけ存在する市場を考える。その仮定により、注文がキャンセルされると同時に新しい注文が入るものとする。キャンセル寿命の分布のみを考察するモデルである為、約定は無視する。注文は経過時間 τ というパラメータを持つ。

各時間ステップ、先ず経過時間 τ を1増やす。その後、注文は確率 $f_c(\tau)$ でキャンセルされ、 $1 - f_c(\tau)$ でキャンセルされない。キャンセルされた場合、その時の経過時間 τ をキャンセル寿命 τ_c として記録し、また注文が新しい注文に置き換わったと考え経過時間 τ を0にする。キャンセルされなかった場合、それ以上特に何も起こらない。

3.1.2 解析解

注文の経過時間が τ に存在する確率分布 $P(\tau)$ と、 τ におけるキャンセル確率 $f_c(\tau)$ を求める。 τ は1つの時間ステップに必ず1つの値のみを取ることより式(2)が成り立つ。

$$\sum_{\tau=1}^{\tau_{max}} P(\tau) = 1 \quad (2)$$

尚、 τ_{max} を ∞ とすると左辺が発散し解を持たないことは証明可能である。そして、キャンセル寿命 τ_c が観測される確率は、経過時間が τ_c でありかつそれがキャンセルされる確率であり、それが冪分布に従う。これは数式(3)として表現される。

$$P(\tau)f_c(\tau) \propto \tau^{-\alpha_c} \quad (3)$$

これを解くと、解析解として式(4-5)を得る。

$$P(\tau) = \frac{1}{\tau_{max}} \left\{ a \sum_{k=1}^{\tau_{max}-1} (\tau_{max} - k) k^{-\alpha_c} + 1 \right\} - a \sum_{k=1}^{\tau} k^{-\alpha_c} \quad (4)$$

$$f_c(\tau) = a\tau^{-\alpha_c}/P(\tau) \quad (5)$$

$$a = 1/\sum_{k=1}^{\tau_{max}} k^{-\alpha_c+1} \quad (6)$$

3.1.3 人工市場モデルへの適用

2節の人工市場モデルにおいて $u_c(\tau) = f_c(\tau)$ としてシミュレーションを行ったところ、図2に示すように、エージェント数 N が特定の値においてのみ望ましい指数の冪分布が得られた。N エージェントの出す指値注文と成行注文の比率を変えてもこの結果は特に変わらなかった。

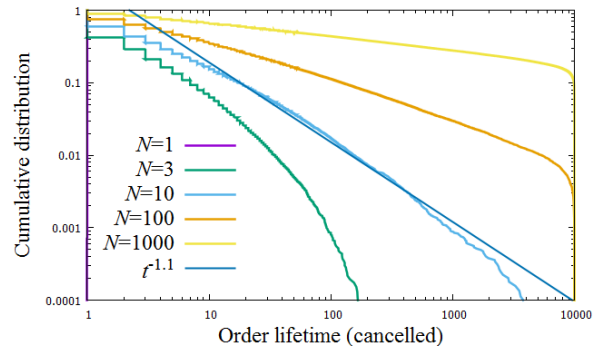


図2 キャンセル寿命の累積分布($u_{LO,MO} = 0.9$, $\tau_{max} = 10000$)

$N < 10$ においてはキャンセル寿命が $\tau_c^{-2.1}$ より短く、 $N > 10$ においては長くなるのがグラフより読み取れる。尚、 $N=1$ は特殊な場合であり、1人のエージェントが指値注文とキャンセルを交互に出しており、キャンセル寿命は全て1である。

3.2 サブモデル2：時間保存則モデル

先述したように、単一注文モデルの結果を人工市場モデルに適用した場合、エージェント数 N が大きい程キャンセル寿命も長くなるのが分かっている。即ち、系全体で見た場合に、板上の注文の経過時間

τ の合計値はキャンセルまたは約定によって注文が減少することによって減るが、エージェント数 N が大きい程それが激しいということである。逆に、経過時間 τ の合計を増加させる要素という、毎時間ステップに板上の注文それぞれの経過時間 τ が 1 増加することであるから、板上の注文数 N_{LOB} だけ系全体の τ は増加することになる。 N の増加に伴ってキャンセル寿命が伸びるという観測結果は、当然 N_{LOB} は N の増加に伴って増加すると考えられるので、系全体の τ 合計値が時間的に安定 (τ 合計値が 0 または ∞ に発散するという状況は N_{LOB} が 0 または N に収斂していることを示し、これは恐らくあり得ないと考えられる) だという考えに整合的である。そこで、3.1 のサブモデルと異なり、単一の注文のキャンセルの過程ではなく、系全体における注文の経過時間 τ の合計及び注文の個数 N_{LOB} の増加・減少速度の釣り合いを考察することによりキャンセル寿命の冪分布を再現出来るのではないかと予想する。また、キャンセル寿命と約定寿命の分布を比較すると後者の方が冪指数が小さく、より大きな τ の注文の消滅が多い為、 τ 合計への影響力が大きい。従って、 τ 合計の均衡を考える上で約定寿命についても考慮する必要がある。

3.2.1 モデル内容

2 節の人工市場モデルと基本的には同じであり、エージェントが L エージェントと N エージェントという 2 つの状態を取るが、板を考えず、約定に関しては特殊な処理を行う。注文エージェントとして選ばれた N エージェントが成行注文を行った場合、L エージェントの中から 1 人を、そのエージェントの注文の経過時間 τ の冪乗、 $\tau^{-\alpha_M}$ によるルーレット選択により選び、そのエージェントの注文と約定したものととして、そのエージェントの状態を L から N エージェントに変える。このようにする理由としては、キャンセル寿命の冪分布の再現が現在の目標であり、約定寿命の冪分布の発生メカニズム解明までは今回は扱わないからである。

3.2.2 解析解

板上の注文の個数の均衡値を N_{LOB}^* とする。また、均衡点における、指値注文、成行注文、キャンセルの比率をそれぞれ p_L 、 p_M 、 p_C とする。板上に経過時間 τ の注文が存在する確率は $P(\tau)$ と記述する。ここで、均衡状態を考えており、かつ現在は指値、成行、キャンセルの数量は何れも必ず 1 であるので、 N_{LOB}^* が存在するという仮定から式(7)が成り立つ。

$$p_L = p_M + p_C = 0.5 \quad (7)$$

注文の存在確率と総数の関係から、式(8)が成立。

$$\sum_{\tau=1}^{\tau_{max}} P(\tau) = N_{LOB}^* \quad (8)$$

また、 $P(1)$ は直前の時間ステップに指値注文が出される確率に等しいので $P(1) = p_L = 0.5$ 、キャンセルや成行注文の発生確率の合計は分かっており、各 τ 毎の比率も求めたい冪分布として分かっているの、 $P(\tau)$ の解析解が式(9)のように定まる。

$$P(\tau) = 0.5 - p_C \frac{\sum_{j=1}^{\tau-1} j^{-\alpha_C}}{\sum_{j=1}^{\tau_{max}} j^{-\alpha_C}} - p_M \frac{\sum_{k=1}^{\tau-1} k^{-\alpha_M}}{\sum_{k=1}^{\tau_{max}} k^{-\alpha_M}} \quad (9)$$

N_{LOB}^* は現実の市場の板上の注文数を適用すれば良いので、式(8-9)は τ_{max} を現実のパラメータから導出する為の式でもある。 N_{LOB}^* と τ_{max} の関係の数値計算例を表 1 に示す。ここから、現実的な N_{LOB}^* の値を得る為には本モデルのようにキャンセルのみならず約定も考慮しなければならないことが分かる。尚、式(8-9)より式(8)が導出出来るが、これが $N_{LOB} = N_{LOB}^*$ における τ 合計値の増加と減少の釣り合いの式である。

$$N_{LOB}^* - p_C \frac{\sum_{j=1}^{\tau_{max}} j^{-\alpha_C+1}}{\sum_{j=1}^{\tau_{max}} j^{-\alpha_C}} - p_M \frac{\sum_{k=1}^{\tau_{max}} k^{-\alpha_M+1}}{\sum_{k=1}^{\tau_{max}} k^{-\alpha_M}} = 0 \quad (8')$$

表 1 N_{LOB}^* と τ_{max} の関係の数値計算
キャンセル寄与、約定寄与は N_{LOB}^* の値のうちキャンセル、約定が寄与する量を表す

τ_{max}	N_{LOB}^*	キャンセル寄与	約定寄与
10^4	5.37	1.93	3.45
10^5	13.1	2.26	10.9
10^6	36.8	2.35	34.5
10^7	111	2.50	109
10^8	347	2.62	345

また、このモデルにおいて、指値注文と成行注文の相対的頻度は外生的に与えるが、キャンセルの頻度は $u_C(\tau)$ の関数形や N_{LOB} 、 N に依存する為、 $N_{LOB} = N_{LOB}^*$ において式(7)が確かに成り立つという保証は無い。そこで、式(7)が成り立つ条件式が必要となるが、それが式(10)である。

$$(N - N_{LOB}^*) \frac{p_L - p_M}{p_L + p_M} - \sum_{\tau=1}^{\tau_{max}} P(\tau) u_C(\tau) = 0 \quad (10)$$

左辺第1項はN エージェントの指値注文による N_{LOB} が増える効果と成行注文による減る効果を表し、第2項はキャンセルにより減る効果を表している。第1項は既知なのでこれにより $P(\tau)u_C(\tau)$ の τ に関する総和が求まる。 $P(\tau)u_C(\tau)$ はある時間ステップにキャンセル寿命 τ が観測される確率であるから異なる τ 間での数値の大きさの比はキャンセル寿命分布と同じ冪関数になる。よって、式(11)のように $P(\tau)u_C(\tau)$ が求まる。

$$P(\tau)u_C(\tau) = \frac{\tau^{-\alpha_C}}{\sum_{j=1}^{\tau_{max}} j^{-\alpha_C}} \sum_{\tau=1}^{\tau_{max}} P(\tau)u_C(\tau) \quad (11)$$

$P(\tau)$ は式(9)で既に明らかとなっているので、以上より $u_C(\tau)$ が式(12)のように求まる。

$$u_C(\tau) = \frac{1}{P(\tau)} (N - N_{LOB}^*) \frac{p_L - p_M}{p_L + p_M} \frac{\tau^{-\alpha_C}}{\sum_{j=1}^{\tau_{max}} j^{-\alpha_C}} \quad (12)$$

3.2.3 人工市場モデルへの適用

第2節の人工市場モデルにおいて、 $u_C(\tau)$ を式(12)によって定めた場合のシミュレーション結果について述べる。 $\tau_{max} = 10^7$ 、エージェント数は $N = 200$ 、その他のパラメータについては以下の通りとする。N エージェントの指値注文と成行注文の比率は $p_L : p_M = 0.5 : 0.045$ となるように定める。この数値は東証一部市場の4銘柄における注文の比率(表2)を参考にした。ティックサイズは1、初期価格は10000とする。

表2 注文タイプの選択比率(宮崎[6]より)

銘柄コード	4523	4661	6273	6954
成行注文	4.61%	4.48%	3.49%	4.59%
指値注文	52.43%	52.33%	51.63%	52.05%
キャンセル	42.97%	43.19%	44.88%	43.36%

指値注文の価格は、 ΔP を $[1, 1000]$ の範囲から $\Delta P^{-1.5}$ に従ってランダムに選び、買い注文の場合には最終約定価格から ΔP を引いた値を、売り注文の場合には加えた値に決定されるものとする。

この条件でシミュレーションを行ったところ、キャンセル寿命の冪分布を再現することが出来た。それを図3に示す。

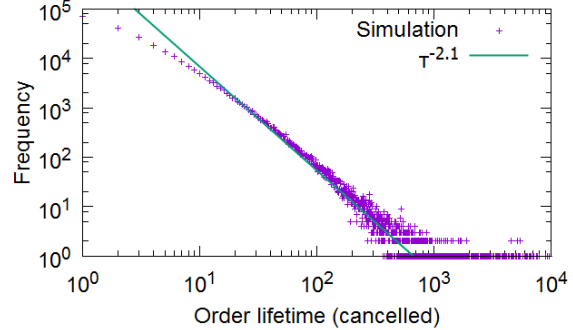


図3 人工市場シミュレーションのキャンセル寿命の分布

尚、約定寿命に関しては、冪分布に従うとは見做せなかった。これを図4に示す。

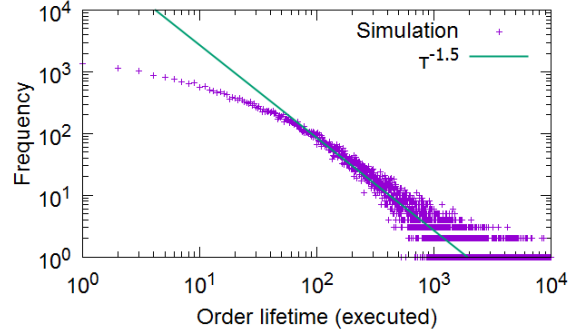


図4 人工市場シミュレーションの約定寿命の分布

4 結果と考察、今後の展望

3.2.3の人工市場モデルについて、その統計性質をいくつか調査し、考察を行う。まず、市場モデルとの妥当性検証として著名な stylized facts である volatility clustering と fat tail について調べた。その結果、前者は再現された(図5)が、後者に関してはリターンの累積分布の裾野が冪分布に従わず、現実以上に厚いことが分かった(図6)。このことは価格の時系列(図7)に垂直な振動が多々見られることにも表れている。

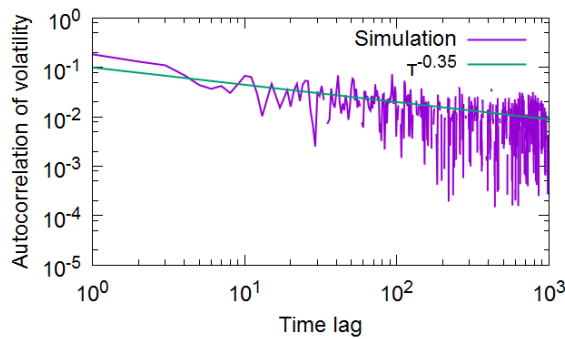


図5 ボラティリティの自己相関

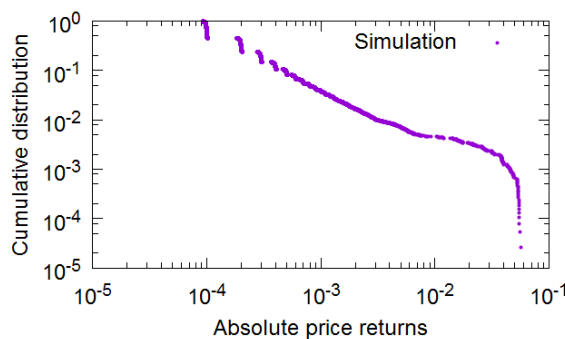


図6 リターンの累積分布

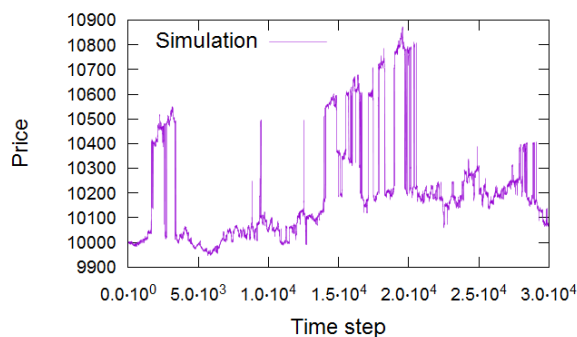


図7 価格時系列

垂直な振動という極めて大きな価格変動の連続が生じるのは、スプレッドが大きいにも関わらず指値注文がスプレッドを小さくする前に成行注文が出され約定することが多発しているということであり、N エージェントの指値・成行の使い分けや指値注文の価格の決定が適切でないということである。エージェントの指値・成行注文をランダムに決めるのではなく、執行コストと遅延リスクの問題から考察される必要がある。また指値・成行注文を執行コストと遅延リスクの問題から考察するのであれば、エー

ジェントがキャンセルする確率も注文がの経過時間のみではなく注文の価格等にも依存しなければならない。即ち、最良気配値に近い価格に置かれている注文は単位時間あたりのキャンセルされる確率が高く、最良気配値から遠い価格に置かれた注文はキャンセル率が低いということが知られている(図8)が、これは前者は遅延リスクを重く見て早期の約定を狙っているのに対し、後者は執行コストを重く見て有利な価格で取引したいと考えていると解釈出来る。従って、執行コストと遅延リスクから指値・成行注文の意思決定を行う際には、キャンセルの意思決定に関しても、自分が行った指値の意思決定に応じたものにならない。現状では図9に示す通り、キャンセル率が最良気配値からの距離に応じて一旦減少するも、約100ティックを超えると逆に上昇してしまっている。

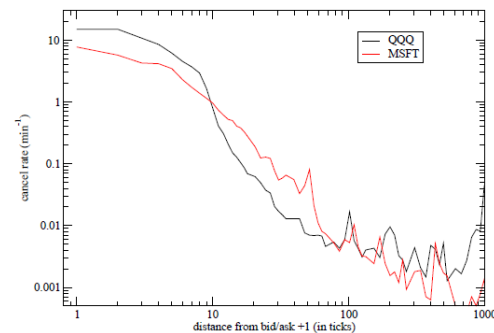


図8 価格とキャンセル率の関係(Potters[7]より) 横軸は注文の置かれている価格の最良気配値からの距離をティック数で表したもので、縦軸は10分間にその価格でキャンセルが行われた回数の平均値を、その価格の板の厚みの平均値で割ったもので、10分間のうちにその価格に存在する注文がどれ程キャンセルされるかを表す

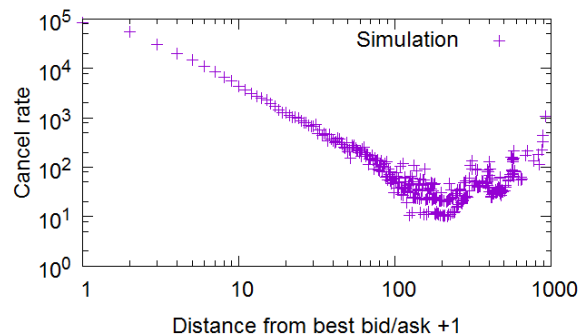


図9 価格とキャンセル率の関係 縦軸はキャンセル注文の総数を板の厚みの平均で割ったもの

また、現状ではモデルのエージェントの意思決定過程が一様で、多様な投資戦略を持つ現実の市場を模倣し切れていない。以上の問題を解決する為、 $u_c(\tau)$ を指値注文の価格にも依存する $u_c(\tau, \Delta P)$ に拡張し、その後Nエージェントの指値・成行注文の意思決定のモデルを構築する予定である。 u_c の拡張を先に行うのは、キャンセル率と価格の関係という、モデル拡張の妥当性を検証する統計則があるからである。 u_c の拡張はキャンセル率と価格の問題の解決のみならず、エージェントの多様性の導入にもなると考えている。というのも、エージェント自体に短期的投資を好むか長期的投資を好むか選好を与えなくとも、偶然最良気配値に近い価格に注文を出したエージェントには以後それが約定またはキャンセルされるまで短期的投資をするトレーダーであるかのように振る舞い、最良気配値から遠い価格に注文を出したエージェントは長期的投資をするトレーダーであるかのように振る舞わせることが出来るからである。エージェント自体に多様な選好は無くとも、様々な選好のエージェントが常に生成と消滅を繰り返す形になる。Nエージェントの指値・成行注文の意思決定のモデルに関しては、妥当性検証に直接的に用いることが出来る統計性質を見つけるのが難しい。しかし、影響を与えるものは約定寿命の分布や板形状等多く存在する為、同じくそれらに影響を与えるキャンセルの問題を進展させた後に、それでも再現されなかった統計性質の再現という形で妥当性検証が行えるだろうと予想する。

加えて、現在の人工市場モデルでは $p_L = 0.5$ という束縛条件が必要であるが、表2を見れば分かる通り $p_L > 0.5$ に拡張される必要がある。これに関しては、指値・成行注文の数量を1に限定するという条件を緩和することで実現可能であると考えている。実際、少なくとも Hautsch[8]が分析したNASDAQの10銘柄については指値注文より成行注文の方が平均数量が大きく、その場合 $p_L > 0.5$ へ人工市場を拡張しても板上の注文数の均衡を保つことは可能な筈である。Nエージェントの意思決定に指値・成行注文の数量を加えることにより、現実の市場の観測値から人工市場における指値注文、成行注文、キャンセルの割合を決定し、現実と同じパラメータの下で市場の性質を考察することが出来るようになると思われる。

5 結論

キャンセル寿命の冪分布を再現する人工市場を、人工市場モデル全体からを一部の要素を取り出して簡略化したサブモデルの考察を基に構築した。その際、単一の注文のキャンセルまでの挙動を考察する

のは有効ではなかったが、市場全体が均衡点において安定的に振る舞う条件を考察することは有効であった。市場の均衡を考える為、キャンセル寿命のみの統計性質を再現するにあたっては、約定寿命についても考慮する必要があることが明らかとなった。

このようにして出来上がった人工市場モデルであるが、価格変動の統計性質を適切に再現していないことが分かり、指値注文、成行注文、キャンセルの何れにおいても意思決定モデルの改良が必要である。

参考文献

- [1] M. Avellaneda and S. Stoikov: High-frequency trading in a limit order book, *Quantitative Finance*, Vol. 8, No. 3, pp. 217–224, (2008)
- [2] C. Chiarella, G. Iori, and J. Perelló: The impact of heterogeneous trading rules on the limit order book and order flows, *Journal of Economic Dynamics and Control*, Vol. 33, No. 3, pp. 525–537, (2009)
- [3] Leal, S. J., Napoletano, M., Roventini, A., & Fagiolo, G.: Rock around the clock: an agent-based model of low-and high-frequency trading. *Journal of Evolutionary Economics*, Vol. 26, No. 1, pp. 49–76, (2016)
- [4] Leal, S. J., & Napoletano, M.: Market stability vs. market resilience: Regulatory policies experiments in an agent-based model with low-and high-frequency trading. *Journal of Economic Behavior & Organization* (in press), (2017)
- [5] D. Challet and R. Stinchcombe: Limit order market analysis and modelling: On a universal cause for over-diffusive prices, *Physica A: Statistical Mechanics and its Application*, Vol. 324, No. 1, pp. 141–145, (2003).
- [6] 宮崎文吾, 和泉潔, 山田健太: スプレッドに着目した確率的板モデルによる株価変動のハースト指数の再現, 第14回金融情報学研究会, Vol.14, (2015)
- [7] M. Potters and J. P. Bouchaud: More statistical properties of order books and price impact, *Physica A: Statistical Mechanics and its Application*, Vol. 324, No. 1–2, pp. 133–140, (2003)
- [8] N. Hautsch and R. Huang: Limit Order Flow, Market Impact, and Optimal Order Sizes: Evidence from NASDAQ TotalView-ITCH Data, (recorded in *Market Microstructure: Confronting Many Viewpoints*, pp. 137–161), (2012)

機械学習による中小企業の信用スコアリングモデルの構築

Credit Scoring for SMEs Using Machine Learning Techniques

澤木 太郎^{1*} 田中 拓哉¹ 笠原 亮介¹

Taro Sawaki¹, Takuya Tanaka¹, and Ryosuke Kasahara¹

¹株式会社リコー

研究開発本部 リコーICT研究所 AI応用研究センター

¹Ricoh Company, LTD.

Research and Development Division

Ricoh Institute of Information and Communication Technology

Applied AI Research & Development Center

Abstract: A credit scoring model is a useful tool for Small and Medium-sized Enterprises (SMEs) lending. In this study, we investigated methods to improve the accuracy of the scoring model using machine learning. As a result, we have shown that Gradient Boosting Decision Tree (GBDT) can obtain the highest accuracy. We found out that GBDT shows better performance than other methods especially when we use more than 10000 learning data. In addition, we demonstrated that ensemble learning can further improve accuracy. According to our simplified estimation, it was suggested that the ensemble learning can reduce the default rate by 16% compared with the conventional method.

1. はじめに

企業のデフォルトリスクを推定する信用スコアリングモデルは融資を行う際の与信業務に極めて有用なツールである。特に近年は金利が低下している影響で、ますます与信コストを下げるのが求められる。そのため、信用スコアリングモデルの重要性は今後さらに高まっていくと考えられる。

従来、信用スコアリングモデルは Logistic Regression(LR)のような統計的な手法が用いられてきた。金融機関は実務の信用スコアリングモデルを公表しないため、具体的にどういった手法が多く使われているかは不明であるが、Logistic Regression、もしくはLogistic RegressionとDecision Treeを組み合わせたハイブリッドモデルが多いと考えられる。

一方で近年はディープラーニングを中心として、様々な機械学習の手法が提案されており、著しく精度が向上している。それにともなって、機械学習手法を使ったスコアリングモデル構築に関する報告が増えている。[1]

これらの報告の多くは、データ件数が1000件前後の小規模なデータセットによって検証が行われている。しかし、一般的に機械学習で高い精度を出すためには、多数のデータが必要である。また、ほとんどが個人の信用情報をもとに構築されたコンシュー

マ向けのスコアリングモデルであり、法人向けのスコアリングモデルに関する検証はあまり進んでいないのが現状である。海外ではわずかに報告があるものの[2][3]、国内については我々の調査した範囲ではそのような検証を行ったという報告は確認できていない。

本研究では、国内の中小企業データを用いて様々な機械学習手法によりスコアリングモデルを構築し、どのような手法が高い精度を出せるのかを調べることを目的としている。

近年注目されている代表的な機械学習手法として、Neural Network、Gradient Boosting Decision Tree、Random Forest、Support Vector Machineがある。それぞれに特徴があり、どの手法が最適なのかは、適用する分野やデータセットの特性に依存する。そこで、本研究ではまずそれぞれの手法を適用して、どの手法が高い性能を出せるのかを調査した。

機械学習では一般的に単一のモデルのみを学習するよりも、複数のモデルを組み合わせるアンサンブル学習の方が高い精度が得られる場合が多いことが知られている。もっとも単純な方法としては、各学習器の出力を平均する方法がある。他に Bagging[6]や Boosting、Stacked Generalization[7]などの手法も知られている。本研究では Stacked Generalizationにより複数の機械学習モデルを組み合わせることでさら

*連絡先：澤木 太郎, taroh.sawaki@nts.ricoh.co.jp

に精度を向上させることができるかどうかを検討した。

2. 方法

2.1. 使用データ

本研究で検証に用いたデータは国内の法人企業約10万件のデータである。本データはリコーリース株式会社から提供を受けた実務データである。同社は小口かつ大量のリースが特徴であり、データの構成は中小企業が中心になっている。

特徴量は売上高などの数値変数と業種などのカテゴリカル変数を含んでおり、合計119種類ある。また、各法人に対して必ずしもすべての特徴量が取得できるわけではないため欠損値を含んでいる。

2.2. 機械学習手法

本研究で検証を行った機械学習手法は以下の3つである。

- Gradient Boosting Decision Tree (GBDT)
- Random Forest (RF)
- Neural Network (NN)

また、ベンチマークとして Logistic Regression によるモデル構築も合わせて行った。

データは、6:3:1の割合でトレーニング用、バリデーション用、テスト用データに分割した。トレーニング用データは学習に使用し、バリデーション用データは後述のパラメータ最適化に使用した。テスト用データは性能の評価のみに用いた。

各機械学習手法にはハイパーパラメータが複数存在する。最適なパラメータはデータセットによって異なるため、パラメータの最適化を行う必要がある。パラメータ探索の手法としてはいくつかの手法が知られているが、本研究では Bayesian Optimization [4] を採用した。

学習は企業が融資実行後にデフォルトしたかどうかを示すフラグを教師データとして学習を行った。

アンサンブル学習には Stacked Generalization を用いた。Stacked Generalization では異なる機械学習モデルを多層にしてアンサンブルする手法である。学習に使用するデータを n -fold に分割し、各機械学習手法によって、out-of-fold データに対する予測値を算出する。こうして得られた予測値を1層目の出力として、次の層の特徴量として入力する。本研究では、図1に示す層構成で学習を行なった。2層目の出力の平均値を最終的な出力とした。

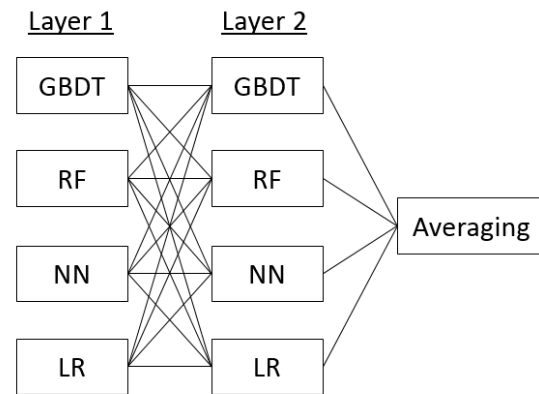


図1 Stacked Generalization の構成

2.3. 評価方法

テストデータに学習で生成したスコアリングモデルを適用して得られた予測結果の評価は CAP 曲線と AR(Accuracy Ratio)値を使用する[5]。CAP 曲線は横軸に推定デフォルト確率の上位 x 件の全体に占める割合 x/N を、縦軸に推計デフォルト確率の高い上位 x 件のうち実際にデフォルトした件数 N_x の割合 N_x/N_d をプロットしたものである。ここで、 N は評価に使用したデータの総数、 N_d は評価データの中のデフォルトした件数の総数である。CAP 曲線の例を図2に示す。B のような曲線が典型的な例である。モデルの説明力が全くない場合は C の直線を描き、予測が完全に正解していれば A のような形を描く。

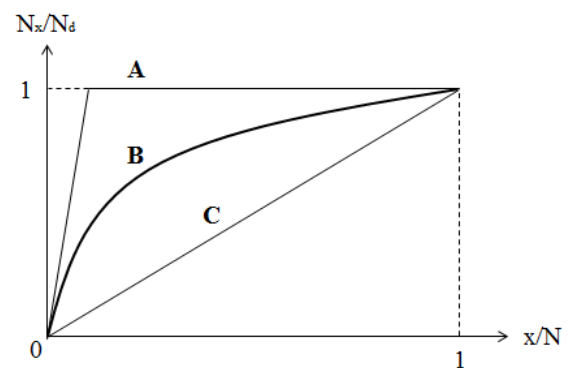


図2 CAP 曲線

ここで、CAP 曲線が B の曲線であるとしたとき、 $(B$ と C で囲まれた部分の面積) / (A と C で囲まれた部分の面積) の比が AR 値である。

3. 結果と考察

各手法で学習した結果を表 1 に示す。これらの結果は各手法をアンサンブルせずに単独で用いた結果である。

表 1 単モデルの性能

手法	AR 値	
	Validation	Test
GBDT	0.621	0.619
Random forest	0.535	0.539
Neural network	0.528	0.530
Logistic regression	0.545	0.523

他の手法と比較して、GBDT が著しく高い精度となっており、テストデータに対して AR 値が 0.619 だった。それに対して Random Forest と Neural Network は既存手法である Logistic Regression と比べてほとんど精度に差が出ない結果となった。GBDT はその他の手法と比較して 0.1 程度上回っているため、信用スコアリングモデルの構築に適したアルゴリズムだと考えられる。

一般的に機械学習はデータ数が多いほど精度が向上する。特に Boosting や Neural Network は十分な精度を得るために大量のデータが必要となることも多い。今回は約 9 万件という大量のデータを学習に使用したが、データ数が少ない場合には精度が下がり、異なる結果になる可能性がある。そこで、学習に使用するデータ数が精度にどのような影響を及ぼすかを調べるため、学習データを変えて学習・予測を行った結果が図 3 である。

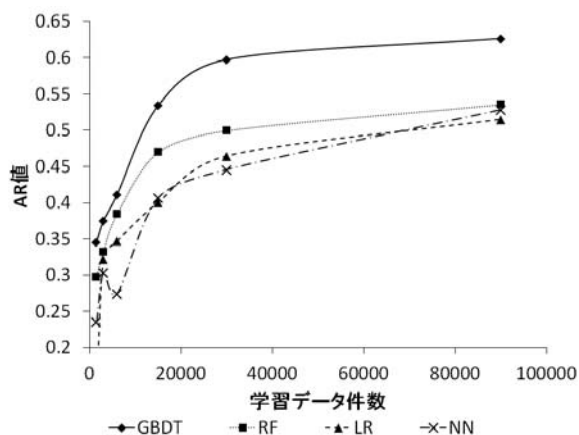


図 3 学習データ件数と精度の関係

横軸がトレーニング用データの件数、縦軸が AR 値である。いずれの手法も学習データ数が増えるほど精度が高くなる。データ数が少ない場合も多い場合も、総じて GBDT の精度が高い。特にデータ数が 1 万件を超えたあたりから他の手法との差が広がっている。一方で学習データが少ない場合には、GBDT と Random Forest はあまり精度に差がない。GBDT の優位性を活かすためには、一定以上のデータ数が必要だと考えられる。

Stacked Generalization を用いたアンサンブル学習の結果を表 2 に示す。

表 2 アンサンブル学習の結果

手法	AR 値	
	Validation	Test
GBDT	0.621	0.619
アンサンブル学習	0.638	0.630

アンサンブル学習の方が GBDT 単体と比較してテストデータにおける AR 値が 0.011 高く、アンサンブル学習の有効性が示された。Stacked Generalization は層数や学習器の種類などがハイパーパラメータになっており、多数のバリエーションが考えられる。今回結果を示したのはあくまでもその一例であるため、さらに精度を高めることができる可能性がある。

従来の手法である Logistic Regression をアンサンブル学習に置き換えることで得られる経済効果を考察するため、それぞれの CAP 曲線を図 4、図 5 に示す。

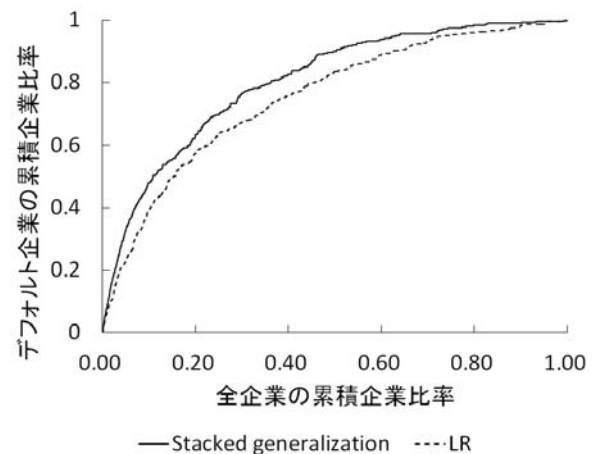


図 4 アンサンブル学習と LR の CAP 曲線

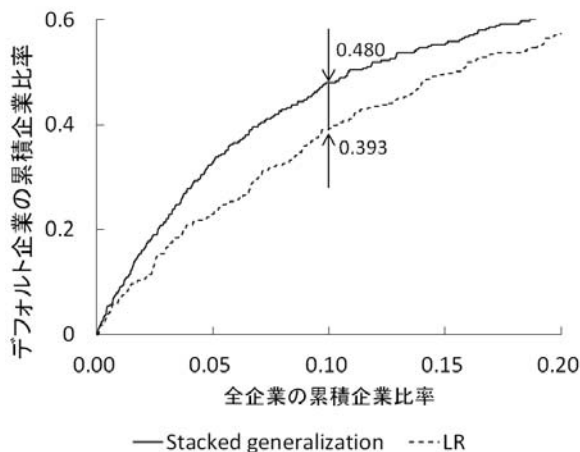


図 5 アンサンブル学習と LR の CAP 曲線 (拡大)

仮にデフォルト率が上位 10%の法人企業を謝絶したケースを考えると、横軸が 0.1 のときの CAP 曲線の縦軸の値が全デフォルト企業のうち何%を謝絶できたかを示す値となる。Logistic Regression によりスコアリングモデルを構築した場合、横軸が 0.1 のときの縦軸の値は 0.393 なので、全デフォルト企業のうち 39.3%を謝絶したことになる。逆に言えば残りの 61.7%の法人については実際にデフォルトしてしまうことになる。一方で、アンサンブル学習を行なってスコアリングモデルを構築した場合には、全デフォルト企業のうちの 48%を謝絶できるためデフォルト率を低く抑えることが可能である。定量的に比較すると、アンサンブル学習の方が従来手法である Logistic Regression よりも、約 16%デフォルトを少なく抑えられる。

4. まとめ

本研究では、国内の中小企業を中心とした法人への与信データ約 10 万件を使用して、機械学習による信用スコアリングモデルの構築を行なった。GBDT、Random Forest、Neural Network の 3 種類の手法を用いて精度を比較したところ、GBDT がもっとも高い精度を得られることが分かった。GBDT は従来手法である Logistic Regression と比べると、AR 値が約 0.1 高く、信用スコアリングモデルの構築において極めて有効な手法であることが示された。

学習に使用するデータ件数と精度の関係についても調べたところ、学習データが少ない場合には GBDT と Random Forest は同程度の精度だが、学習データが 1 万件以上の場合に GBDT のほうが大きく精度が高くなることがわかった。このことから、

GBDT の優位性を活かすためには、大量のデータが必要である。

最後にアンサンブル学習の一つの例として、Stacked Generalization を用いて、複数の機械学習モデルをアンサンブルしたところ、単体の GBDT よりも AR 値でさらに 0.01 程度高い精度を得られた。

本研究では、法人企業の信用スコアリングモデル構築において、GBDT を含めた機械学習モデルをアンサンブルすることによって、Logistic Regression により構築された従来のモデルと比較して、大きく精度を向上できることを示した。簡易的な試算では、約 16%デフォルトを低減する効果があることが示された。

謝辞

本研究では、リコーリース株式会社に提供していただいたデータセットを利用した。非常に貴重なデータを提供していただいたリコーリース株式会社に感謝の意を表す。

参考文献

- [1] S. Lessmanna, B. Baesensb, H. Seowd, and L. C. Thomas: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research*, vol. 247, No.1, pp. 124-136, (2015)
- [2] D. Fantazzini, and S. Figini: Random Survival Forests Models for SME Credit Risk Measurement, *Methodology and Computing in Applied Probability*, Vol. 11, No. 1, pp. 29-45, (2009)
- [3] H. S. Kim, and S. Y. Sohn: Support vector machines for default prediction of SMEs based on technology credit, *European Journal of Operational Research*, Vol. 201, No.3, pp. 838-846, (2009)
- [4] J.Snek, H. Larochelle, and R. P. Adams: Practical Bayesian Optimization of Machine Learning Algorithms, *Advances in Neural Information Processing Systems*, Vol. 25, (2012)
- [5] 山下智志, 川口昇, 敦賀智裕: 信用リスクモデルの評価方法に関する考察と比較, *Financial Research and Training Center discussion paper series*, 11, (2003)
- [6] L. Breiman: Bagging predictors, *Machine Learning*, Vol.24, No. 2, pp. 123-140, (1996)
- [7] D. H. Wolpert, Stacked generalization, *Neural Networks*, Vol. 5, No. 2, pp. 241-259, (1992)

北朝鮮リスクの日本株式市場への影響

-AI ベースのリスクモデルは何を語るのか？

The impact of North Korea risk on the Japan equity market
: what do AI based risk models tell us?

西山 昇

Noboru Nishiyama

Dragons' Desk Ltd. / 千葉商科大学（会計専門職大学院）

Abstract: We analyze Japan equity portfolios for their risk exposures to North Korea geopolitical destabilization effects over the past few months, by using an implementation of the EM algorithm integrated with a GARCH process. The model identifies which latent factors in the Japan equity market relate to North Korea and predicts their risk impact for the near future.

Key words: EM algorithm, GARCH process, machine learning, decision making.

1. はじめに

地政学的リスクの定量的なモデル化は資産運用の実務家（特にファンドマネージャ）にとって重要なテーマの一つである。特に地政学的リスクに対応するためのファクターは過去に経験したことがない、あるいは、経験したとしても事前に決めることが有効であるとは限らない場合がある。その上、状況が経時的に変化していくことがロバストネス（頑健性）とセンシティブティ（柔軟性）の両者のバランスをとるモデル化を困難にしている。

今回の論考では、EM アルゴリズムがもつ AI 的な側面にフォーカスする。ここでの AI 的な側面とは次の 3 つの特徴をさす。①意思決定のための学習アルゴリズム、②人的なインプットなしのファクター選択、③意思決定に有効な分散・共分散の予測

EM algorithm を活用した AI ベースのリスクモデルの特徴は次の 3 点にある。

(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。

本稿では、過去数ヶ月間の北朝鮮のニュースに関連した地政学的に不安定な影響に対する日本株ポートフォリオのリスクエクスポージャーに注目して分析する。

日本株ポートフォリオは、東証上場の日本株全体をユニバースとして最適化ポートフォリオとして構成する。

分析には GARCH プロセスを統合した EM アルゴリズムを使用する。モデルは日本株式市場の潜在ファクターのうちのどれが北朝鮮リスクと関連しているのかを識別し、保有ポートフォリオに与える可能性のある近未来のリスクインパクトを予測する。

地政学的イベントは、事後的に分析することができても、事前に要因（ファクター）を特定することができないためリスクを推定することは難しい。

実際の投資運用における地政学的リスクのようなイベントリスクに対応できる定量的なモデル化には何が必要なのか。その問いに対する一つの考え方を検討する。

本稿では、説明変数をあらかじめ特定しない統計的マルチファクターモデルを利用して、過去に発生したイベント前後のリスク値の推移を確認することでモデルの地政学的リスクへの適用可能性を探る。

検討作業の詳細は後述するが、最適化ポートフォリオを2パターン用意する。一つは北朝鮮リスクのような地政学的リスクに敏感に反応することを想定したポートフォリオ、別の一つは逆に地政学的リスクに敏感に反応しないポートフォリオである。

北朝鮮のリスク指標として韓国(South Korea)株式インデックスを代替指標として日本株式の個別銘柄の対市場ベータを計測する。

北朝鮮の地政学的なイベントが発生していると考えられる過去数カ月の期間に最適化ポートフォリオを構築して一定期間のリスク量の変動を確認する。

推定にあたっては、すでに発生したイベントであることからインサンプルのデータを使用するものの、モデルによる推定を実施する前日までの利用可能なデータを想定して、その先を予測するアウトサンプルの最適化を実行する方法をとっている。

2. 地政学的リスクとは

2.1 地政学的リスクの分析

2017年1月、イアン・ブレマー氏(Eurasia Group)は2017年に予想される地政学的リスクとしてGeopolitical Recession(地政学的後退)とのタイトルで次の10個をあげていた。[1]

- (1) 独立する米国 (INDEPENDENT AMERIKA) — トランプ大統領が述べる“America first”の哲学、そして彼の公約である“make America great again”が米国のコアバリューを組み立てる。それが米国のIndependentに向かう原動力となる。
- (2) 中国の過剰反応 (CHINA OVERREACTS) — 2017年の秋に開催される全国人民代表大会議のリーダーの権力の変更がさまざまな影響をもたらす。
- (3) 弱体化するメルケル (A WEAKER MERKEL) — このことが欧州全体の政治的リスクを高める可能性がある。
- (4) 改革なし (NO REFORM) — 先進国と新興国の両方におけるリーダーシップの欠如から構造改革が進まない状況となる。それは投資家もつ成長への期待と新しい投資機会を低下させる。
- (5) テクノロジーと中東 (TECHNOLOGY AND THE

MIDDLE EAST) — 中東では独裁制が存続しているものの、ここ数十年の間に政府は弱体化してきている。

- (6) 中央銀行が政治的な攻撃対象になる (CENTRAL BANKS GET POLITICAL) — ここ数十年で初めて新興国だけでなく、米国、欧州圏、英国の中央銀行が攻撃にさらされる。
- (7) ホワイトハウス対シリコンバレー (THE WHITE HOUSE VS SILICON VALLEY) — 大統領選挙ではトランプ氏を指示していなかったカリフォルニア州のITリーダーたちと関係の先行き。
- (8) トルコ (TURKEY) — エルドガン大統領の権限の拡大に対する投票において賛成、反対が拮抗している中、大統領が賛成支持の拡大に向け積極的なキャンペーンを展開する。
- (9) 北朝鮮 (NORTH KOREA) — 2017年は北朝鮮にとって大きな年となる。しかしそれはよいことではない。
- (10) 南アフリカ (SOUTH AFRICA) — 現政権と対立する勢力による政治的リスクが高まる。政治的不安定は地域経済への悪影響となる。

これらのことから、地政学的リスクの重要な一面には、国内の政治状況が与える地域を含めた対外的な影響がある。北朝鮮の国内状況と対外政策が連動しているようにはみえないが、実際にミサイル発射実験を繰り返し、核開発を進め極東地域の政治状況を不安定化させていることはリスクとして認識できる。

ここで資産運用者(ファンドマネージャ)の対応策として何があるのか。仮にある程度、事前に地政学的リスクの発生を予測できたとして、投資運用しているファンドにおいて何らかの対応策を事前にとることが可能なかがポイントになる。

2.2 地政学的リスクと投資運用

地政学的リスクは、グローバルに活動する投資運用会社のCRO(Chief Risk Officer: 最高リスク責任者)にとって投資運用リスクにおけるエマージングリスク(Emerging risk)の一つとして取り扱われている。[2]

エマージングリスク(Emerging risk)とは、それまで無関係とされていたファクター(要因)に相関が突然発生することにより生じる投資運用リスクを指す。

各投資運用会社がエマージングリスク(Emerging risk)を予測するために具体的な方法で計測しているとは限らないものの、直観的な方法を含め、資産運用リスク管理プロセスの一部として、さまざまな

方法でモニターしているとされる。

Market Risks (マーケットリスク) 管理の観点から CRO がエマージングリスク (Emerging risk) として認識しているものは次のとおりである。([2] p30)

(i)市場流動性リスク：グローバルのルールと規制が大きく変化した結果として、厳しいマーケットストレスが発生したケースでは、金融仲介者は過去に実行可能だったマーケットメーカー、市場仲介者としての役割をもった行動が十分にできない。

これはマーケットメイクによって作り出される、流動性に大きく依存した市場だけでなく、(相対取引である)債券市場にもあてはまる

(ii)金利リスク：グローバル金融危機 GFC (Global Financial Crisis) 以降、グローバル市場と米国中央銀行の政策アクションとの相関が上昇してきている。債券、株式、商品と通貨などを含む金融市場は、米国の金利と連動して変化する指標として繰り返し反応してきた。

(iii)マクロ (経済) リスク：投資運用会社は、原油価格の下落、デフレリスク、通貨変動の関係のみならず米国、欧州、日本と中国のマクロ経済の展開をウォッチし続けている。

(iv)地政学的リスク：すべての投資運用会社が言及したが、地政学的リスクが増加している。これは原油価格の下落に関連する地政学的リスクのみならず、ロシア、ウクライナ、テロ組織 IS に関する緊張状態を含んでいる。

3. 地政学的リスク計測のモデル化

3.1 統計的マルチファクターモデル

地政学的リスクの要因として考えられるファクターは常に変化するものである。過去の経験のみでは事前に特定することはむづかしい。

そのため最初から説明変数を特定して感応度を推定するのではなく、ファクターと感応度を個別資産のリターンを被説明変数として同時に推定できる APT 型の統計的ファクターモデルを適用する。[3]

$$R : \text{return}, F : \text{risk factors}, \beta : \text{sensitivity}$$

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

(式 3-1) 統計的マルチファクターモデルのリターン表現

Σ : variance-covariance (分散・共分散)

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

(式 3-2) 統計的マルチファクターモデルのリスク表現

(式 3-2) 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_ε はアンシステムティック (非組織的) リスク、それ以外の右辺の項全体をシステムティック (組織的) リスクと呼ぶ。

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

システムティック (組織的) リスクにある統計的ファクター σ_n^2 (n は統計的ファクター数) がポートフォリオ特性 (国、業種、スタイル等) とどのように関連しているかを確認することで運用するポートフォリオがどの特性からどれくらい影響を受けているかを調べられる。

先に説明したとおりポートフォリオ特性の一つとして韓国 (South Korea) 株式インデックス (以下、SK インデックス) を代替指標とする。北朝鮮のリスク指標としてとして日本株式の個別銘柄の対市場ベータを計測する。

実務的なリスク属性分析としては、ベータの他に共通のリスクファクターの中からトータルリスクに対するリスクファクターの限界的寄与の計測、トラッキングエラー、投資スタイルのトレンド分析などの推定がある。

3.2 仮想ポートフォリオの特性

東証上場銘柄をユニバースとして最適化ポートフォリオを組む。SK インデックスに対するベータをテイルトさせる 2 種類 (高リスク、低リスク) の仮想日本株式ポートフォリオを 2017 年 6 月末の時点で作成する。

仮想ポートフォリオは、SK インデックスに対する感応度が高いポートフォリオ (高リスク) と SK インデックスに対する感応度が低い (低リスク) ポートフォリオを構成する最適化をかける。

最大組み入れ銘柄数の上限を 200 銘柄、個別銘柄の最大組み入れ比率を 1% とする。

次に北朝鮮リスクが高い最適化ポートフォリオと北朝鮮リスクが低い最適化ポートフォリオの特性を概観する。ⁱⁱ

(組み入れ銘柄数) は、高リスクポートフォリオは 177 銘柄に対して、低リスクポートフォリオは 124

銘柄となる。

次にポートフォリオリスクを高める業種別リスクを高い順番に 10 個あげる。東証の業種分類による。

(高リスクポートフォリオ：業種リスク TOP10)

Top 10 Style Risks
Style
TOPIX Machinery
TOPIX Electric Appliances
TOPIX Transportation Equipment
TOPIX Secs & Commodity Futures
TOPIX Banks
TOPIX Chemicals
TOPIX Glass & Ceramics Products
TOPIX Other Financing Business
TOPIX Iron & Steel
TOPIX Nonferrous Metals

(低リスクポートフォリオ：業種リスク TOP10)

Top 10 Style Risks
Style
TOPIX Retail Trade
TOPIX Construction
TOPIX Services
TOPIX Foods
TOPIX Chemicals
TOPIX Wholesale Trade
TOPIX Textiles & Apparels
TOPIX Information & Communication
TOPIX Pharmaceutical
TOPIX Metal Products

次にポートフォリオリスクを高める個別銘柄をリスクの高い順番に 10 個あげる。リスク下げするためには、これらの銘柄の保有比率を低下させる調整を実施することになる。

(高リスクポートフォリオ：銘柄リスク TOP10)

Top 10 Riskiest Assets
Name
NTN
Yamaichi Electronics
Mazda Motor
DMG MORI
JTEKT
NSK
Shindengen Electric Mfg
Kobe Steel
SUMCO
Mitsuba

(低リスクポートフォリオ：銘柄リスク TOP10)

Top 10 Riskiest Assets

Name
Genky stores
Morinaga
Kusuri No Aoki
M3
Okamoto Industries
MEGMILK Snow Brand
Kewpie
AIN HOLDINGS
TSURUHA Holdings
Fuji Kyuko

4. 結果の考察

4.1 リスクファクターの意味

本稿では、北朝鮮リスクを EM algorithm を使った統計的マルチファクターモデルでとらえることを試みた。フォワードルッキングなリスク計測では、トータルリスクの短期変動（分散共分散行列）を予測する必要がある。

地政学的リスクという市場環境の変化を定量的にとらえるための一つの枠組みとして統計的マルチファクターモデルが適用できることを確認した。

(トータルリスク) = (システムティックリスク) + (アンシステムティックリスク) のシステムティックリスク部分を個別のポートフォリオ特性（国、業種、スタイル）に結びつけた結果である。

4.2 実際の運用のために必要なこと

今回のモデルでは、各ファクターに対する S K インデックスのベータ（感応度）を計測して、日本株の北朝鮮リスクの大きさを S K インデックスに対するベータの大きさにより調整した。ポートフォリオ全体のリスクを低下させるためには、スタイル、個別銘柄の保有比率の高さを調整する方法が適用できることがわかる。

実際には、最適化を使って S K インデックスへのリスク（感応度）を最小化することと最大化することにより 2 種類の最適化ポートフォリオを構築した。

またトータルリスクに対するリスクファクターの限界の寄与を計測することで定量的にエクスポージャーをコントロールすることができる。

5. おわりに

本稿では、北朝鮮リスクという地政学的な事象を事例に、統計的マルチファクターモデルの適用可能性を検討した。

統計的には、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、その波及効果を予測に反映することが求められる。

統計的マルチファクターモデルの枠組みにより、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。またファンドマネージャは、ポートフォリオへの事前には把握できないとされた地政学的な影響を早めに調整することができる。

さらに最適化によりその時点の地政学的リスクに対する耐性のあるポートフォリオを構築することによるバックテストによる最適化シナリオの有効性を確認できる。

発表では、AI ベースのリスクモデルによる最適化ポートフォリオのバックテストの経過を説明する。

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。

参考文献

- [1] TOP RISKS 2017: THE GEOPOLITICAL RECESSION, Ian Bremmer(Eurasia group), January, 2017.
(<https://www.eurasiagroup.net/issues/top-risks-2017>)
- [2] Securities and Futures Commission (HK), Risk-focused Industry Meeting Series: Asset Management: Looking Forward January 2015,参考資料として香港の金融規制当局である SFC (Securities and futures commission of Hong-Kong、以下 SFCHK) が公開した小冊子”Risk-focused Industry Meeting Series: Asset Management: Looking Forward, January 2015”。この小冊子は SFCHK が 2014 年 3 月から同年 11 月までの間、同規制当局が管轄している投資運用会社のリスク管理部門を含む各部門の責任者と資産運用の現状と問題意識についてオープンな対話 (Risk-focused Industry Meeting Series) を重ね、その内容をまとめたものである。小冊子では、資産運用ビジネス全体について幅広く網羅している。
(<http://www.sfc.hk/web/EN/published-resources/industry-related-publications/risk-focused-industry-meeting-series-report.html>)

[3] EM Applications, Ltd.,

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

ⁱ 森金融庁長官インタビュー2016年8月2日 18:45 JST【東京】「...金融危機は繰り返し起こるものだが、その都度別の姿をしてやってくる。ほぼ確実にいえるのは、全く同じ危機は二度は起こらないということだ。なぜなら、我々は常に過去の危機に対応して、その防止に努めるからである。従って、次の危機は、予期しない所からやってくるようになる。

本年 (2016 年) 4 月の ISDA (国際スワップデリバティブ協会) 年次総会における基調講演で、私は第二次世界大戦の戦艦大和の例をお話した。大和は海からの攻撃に耐えうる分厚い装甲を備えていたが、空からの攻撃には脆弱であった。過去の危機は海からやってきたかもしれないが、次の危機は空からやって来るかもしれない。...」

ⁱⁱ Analytics provided by EM Applications

多層ニューラルネットワークと GA を用いた TOPIX 運用 AI

“TOPIX trading AI” using Multi-layer Neural Networks and GA

石原 龍太

Ryota Ishihara

株式会社かんぽ生命保険 運用企画部

Investment Planning Department, JAPAN POST INSURANCE Co., Ltd.

Abstract: In This paper, we propose a construction method of an artificial intelligence for TOPIX trading. We apply a Multi-layer Neural Networks as a prediction method, and optimize to maximize the Information Ratio using GA. We also conduct a simulation using TOPIX market data from January 2014 to December 2016, showed the effectiveness of the proposed construction method from the information ratio level.

1 はじめに

現在、幅広い領域で注目を集めている人工知能(以下、AI)は、金融の世界にもブレークスルーをもたらすことが期待されている。本稿のテーマとする AI を用いた資産運用の研究とは、「過去に市場で発生した事象を学習し、適切な投資判断を行う AI」を構築し、資産運用に活用しようとする取り組みである。近年、米国のヘッジファンドが市場に影響を与える多種多量のデータ(ビッグデータ)を学習させた AI を資産運用に活用したり、国内の機関投資家が AI を用いて運用するファンドの組成を行う等、AI は資産運用の現場に広まってきている。

本稿のテーマに関するこれまでの研究成果を概観すると、絶対リターンを追求した運用を行う AI¹の構築を目指したものが多い、というのが筆者の印象である(運用成果の表現が分かりやすい、というのも一因であろう)。一方、機関投資家の資産運用の現場では、リターンだけでなく、リスクにも配慮が求められることが一般的である。特に、株式のようなベンチマークとなる市場インデックスが存在する資産では、市場インデックスからの乖離をリスク、市場インデックスに対する超過収益をリターンとして捉え、リスク調整後リターン²を追求した運用がなさ

¹ 平林/伊庭(2008)の研究では収益額、山口/坂井/田中(2003)の研究では資産残高の最大化を追求している。なお、辻岡/山本(2011)の研究では、シャープレシオの最大化を追求する AI も提案している。

² 投資におけるリスクとリターンの両方を考慮して運用の巧拙を表す指標をリスク調整後リターンと呼ぶ。その代表的なものが、シャープレシオやインフ

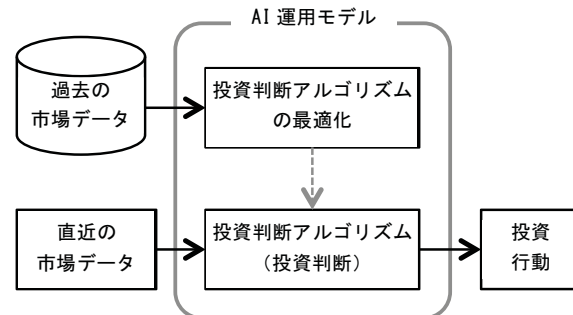
れることが多い。

そこで本稿では、東証株価指数(TOPIX)を運用対象資産として取り上げ、多層ニューラルネットワークと GA を用いて「リスク調整後リターンを追求した資産運用を行う AI (以下、AI 運用モデル)」の構築を試みる。

2 AI 運用モデルの構築

本稿で構築する AI 運用モデルの概念図を図 1 に示す。AI 運用モデルは、直近の市場データを知覚し、投資判断アルゴリズムで導出した投資判断から投資行動を決定するものとする。また、AI 運用モデルは過去の市場データを学習し、人間によるチューニングを必要とせず、自ら投資判断アルゴリズムを最適化することができるものとする。

図 1 AI 運用モデルの概念図



(出所) 筆者作成。以下同じ。

オメーションレシオである。

2.1 AI 運用モデルの概要

2.1.1 運用コンセプト

AI 運用モデルは、市場環境に応じて機動的にリバランスを行い、リスク調整後リターンの最大化を目指すものとする。

2.1.2 投資対象資産

AI 運用モデルが投資する資産は、TOPIX と無リスク資産（収益率は常に 0%とする）の 2 種類とする。また、資産の売買（TOPIX と無リスク資産の交換）は当日終値で行うものとし、取引コスト（売買手数料等）は発生しないものとする。

2.1.3 資産アロケーション

AI 運用モデルは、機動的なリバランスにより資産価格の下落による損失を回避するため、市場環境に応じて日々保有資産の組入比率を変更（リバランス）するものとする。また、選択する保有資産の組入比率は「TOPIX100%, 無リスク資産 0%」又は「TOPIX0%, 無リスク資産 100%」のいずれかとする。

2.1.4 市場データ

AI 運用モデルが知覚する市場データ³は、表 1 で定義する TOPIX の日次の四本値と配当データ (t=1, ..., T) とする。ここで、t 日の TOPIX の配当 $TpxDIV_t$ は以下のように定義⁴する。

$$TpxDIV_t = \left(\frac{t \text{ 日の TOPIX 配当込み指数}^5}{t-1 \text{ 日の TOPIX 配当込み指数}} - \frac{TpxCL_t}{TpxCL_{t-1}} \right) \times TpxCL_{t-1}$$

表 1 市場データ (TOPIX 四本値と配当)

日付	始値	高値	安値	終値	配当
T	$TpxOP_T$	$TpxHI_T$	$TpxLO_T$	$TpxCL_T$	$TpxDIV_T$
:	:	:	:	:	:
t	$TpxOP_t$	$TpxHI_t$	$TpxLO_t$	$TpxCL_t$	$TpxDIV_t$
:	:	:	:	:	:
1	$TpxOP_1$	$TpxHI_1$	$TpxLO_1$	$TpxCL_1$	$TpxDIV_1$

³ 先行研究では「終値（一本値）」を知覚データに用いているものが多いが、筆者は適切な投資判断には情報の深さよりも厚みが重要であると考え、本稿では「TOPIX 四本値と配当」を用いる。

⁴ 有効桁数は小数点第 2 位までとし、小数点第 3 位以下は切り捨てる。

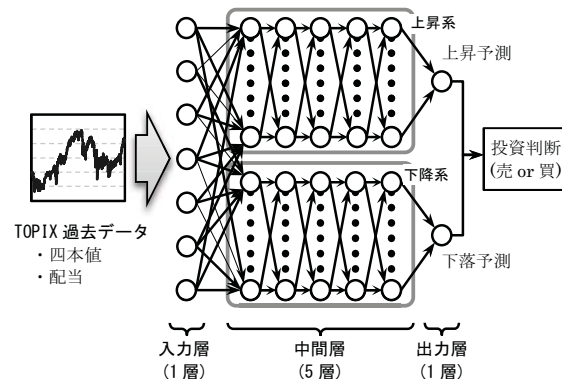
⁵ TOPIX 配当込み指数は、(株)東京証券取引所が終値のみを算出している指数である。

2.2 投資判断アルゴリズムの作成

AI 運用モデルの投資判断アルゴリズムは、入力層（1 層）、中間層（5 層）、出力層（1 層）の 7 層からなる多層ニューラルネットワーク⁶を用いて作成する。多層ニューラルネットワークとは、入力層と出力層の間に中間層をもつニューラルネットワークであり、さまざまな情報が混在する雑多なデータから、本質的な情報を抽出することに適した数理モデルとされる。

本稿で用いる多層ニューラルネットワークの構造を図 2 に示す。投資判断アルゴリズムは、入力層で市場データを知覚し、2 系（上昇系、下降系）の中間層と出力層で TOPIX の価格変動（上昇または下落）を予測し、当該予測に基づく投資判断を導出する。

図 2 多層ニューラルネットワークを用いた投資判断アルゴリズム



(注) 丸はニューロン、矢印はニューロン間の結合と処理の流れを示す。

2.2.1 入力層

入力層のニューロン y_{kt}^{in} は市場データを知覚し、表 2 で定義した出力値を中間層 1 のニューロンに受け渡す。ここで k を入力層を構成するニューロンの識別番号 ($k=1, \dots, 7$) とする。

⁶ ニューラルネットワークとは「人間の脳内の神経細胞（ニューロン）が、別の神経細胞から受け取った信号を次の神経細胞へと伝達していく流れの中で、様々な情報処理を行う仕組み」を模倣した数理モデルである。層（並列に並んだニューロンで形成されるユニット）の数が 3 桁以上となるような大規模な多層ニューラルネットワークはディープニューラルネットワークと呼ばれる。

表2 入力層ニューロンの出力値

ニューロン	出力値
y_{1t}^{in}	$\log\{(TpxCL_t + TpxDIV_t)/TpxOP_{t-1}\}$
y_{2t}^{in}	$\log\{(TpxCL_t + TpxDIV_t)/TpxHI_{t-1}\}$
y_{3t}^{in}	$\log\{(TpxCL_t + TpxDIV_t)/TpxLO_{t-1}\}$
y_{4t}^{in}	$\log\{(TpxCL_t + TpxDIV_t)/TpxCL_{t-1}\}$
y_{5t}^{in}	$\log(TpxCL_t/TpxOP_t)$
y_{6t}^{in}	$\log(TpxCL_t/TpxHI_t)$
y_{7t}^{in}	$\log(TpxCL_t/TpxLO_t)$

2.2.2 中間層

中間層は、上昇系 (up) と下降系 (down) の並列構造をもつ5層構成とし、両系とも1つの層あたり10個のニューロンで構成する。

$\omega_{jkk}^{s-1(s)}$ を j 系における s 層のニューロン k と $s-1$ 層のニューロン k' の間の結合ウェイト (k' が11のものは定数項を意味する), $f(x)$ を0または1の値をとるステップ関数 (x が0より大きい場合は1, そうでなければ0を返す) としたとき, 中間層1のニューロン y_{jkt}^{mid1} の出力値は以下のように表現する。

$$y_{jkt}^{mid1} = f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$x = \omega_{j1k}^{in(mid1)} y_{1t}^{in} + \dots + \omega_{j7k}^{in(mid1)} y_{7t}^{in} + \omega_{j8k}^{in(mid1)} \times 5\%$$

また, 中間層2~5のニューロン $y_{jkt}^{mid2}, y_{jkt}^{mid3}, y_{jkt}^{mid4}, y_{jkt}^{mid5}$ の出力値は以下のように表現する。

$$y_{jkt}^s = f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$x = \omega_{j1k}^{s-1(s)} y_{j1t}^{s-1} + \dots + \omega_{j10k}^{s-1(s)} y_{j10t}^{s-1} + \omega_{j11k}^{s-1(s)}$$

ここで j を系 ($j=up, down$), s を層 ($s=in, mid1, mid2, mid3, mid4, mid5$), $s-1(s)$ を層間の結合 ($s-1(s)=in(mid1), mid1(mid2), mid2(mid3), mid3(mid4), mid4(mid5)$), k を j 系の s 層を構成するニューロンの識別番号 ($k=1, \dots, 10$) とする。 j 系の中間層 s のニューロン y_{jkt}^s は, 出力値を j 系の中間層 $s+1$ 又は j 系の出力層のニューロンに受け渡す。

2.2.3 出力層

出力層は、上昇系のニューロン $z_{up t}$ と下降系のニューロン $z_{down t}$ で構成する。 $\omega_{jkk}^{mid5(out)}$ を j 系における出力層のニューロンと中間層5のニューロン k' の間の結合ウェイト, y_{jkt}^{mid5} を中間層5のニューロン k' から受け取る値, $f(x)$ を0または1の値をとるステップ関数としたとき, 出力層のニューロン z_{jt} の出力値は以下のように表現する。

$$z_{jt} = f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$x = \omega_{j1}^{mid5(out)} y_{j1t}^{mid5} + \dots + \omega_{j10}^{mid5(out)} y_{j10t}^{mid5} + \omega_{j11}^{mid5(out)}$$

出力層において, $z_{up t}$ は TOPIX 上昇の兆候, $z_{down t}$

は下落の兆候を検知するものとし, $z_{up t}$ が“1”の場合, 投資判断アルゴリズムが TOPIX の上昇を予測し, $z_{down t}$ が“1”の場合, 投資判断アルゴリズムが TOPIX の下落を予測していることを意味する。

2.3 投資行動の決定

AI 運用モデルは, 多層ニューラルネットワークの上昇系のニューロン $z_{up t}$ と下降系のニューロン $z_{down t}$ の出力値から, 表3のとおり当日(時点 t)の投資判断を行う。AI 運用モデルは, この投資判断に基づき当日終値で売買を行う。なお, 前日の保有資産の TOPIX 組入比率 x_{t-1} が100%の場合における“買い判断”, x_{t-1} が0%の場合における“売り判断”については, 当日の売買を行わない ($x_t = x_{t-1}$) ものとする。

表3 出力層ニューロンの出力値と投資判断

		上昇系 $z_{up t}$	
		0	1
下降系 $z_{down t}$	0	現状維持 ($x_t = x_{t-1}$)	買い判断 ($x_t = 100\%$)
	1	売り判断 ($x_t = 0\%$)	現状維持 $x_t = x_{t-1}$

(注) x_t は時点 t における AI 運用モデルの保有資産の TOPIX 組入比率とする。

3 投資判断アルゴリズムの最適化

AI 運用モデルの投資判断の巧拙は, 投資判断アルゴリズムのニューラルネットワークを構成するニューロン間の結合ウェイト $\omega_{jkk}^{s-1(s)}$ の組み合わせ (以下, ネットワーク構造) によって決まる。つまり, 運用コンセプト (リスク調整後リターン最大化) を具現化するためには, 適切なネットワーク構造を探索することが重要となる。

3.1 基本設定

3.1.1 リターンの表現

本稿では, TOPIX の日次収益率 $Tpxr_t$ と $Cashr_t$ を以下のように表現する。

$$Tpxr_t = \log\left(\frac{TpxCL_t + TpxDIV_t}{TpxCL_{t-1}}\right)$$

$$Cashr_t = 0\%$$

このとき, AI 運用モデルの日次収益率 AIr_t は, TOPIX と無リスク資産の日次収益率を前日の組入比率で加重平均した値として, 以下のように表現され

る。

$$AIr_t = Tpxr_t x_{t-1} + Cashr_t (1 - x_{t-1})$$

また、N を m 月に含まれる日次データの件数としたとき、m 月における AI 運用モデルの月次収益率 AIr_m と TOPIX の月次収益率 $Tpxr_m$ は以下のように表現される。

$$AIr_m = \sum_{t=1}^N AIr_t$$

$$Tpxr_m = \sum_{t=1}^N Tpxr_t$$

3.1.2 リスクの表現

本稿のリスク指標には、トラッキングエラーを用いる。AI 運用モデルのトラッキングエラー $\sigma(e)_{AI}$ は以下のように表現する。

$$\sigma(e)_{AI} = \sqrt{\frac{\sum_{m=1}^M \{(AIr_m - Tpxr_m) - (\overline{AIr} - \overline{Tpxr})\}^2}{M - 1}}$$

$$\overline{AIr} = \frac{\sum_{m=1}^M AIr_m}{M}$$

$$\overline{Tpxr} = \frac{\sum_{m=1}^M Tpxr_m}{M}$$

ここで M を計測期間の月数とする。

3.1.3 リスク調整後リターン表現

本稿のリスク調整後リターンには、インフォメーションレシオ⁷を用いる。AI 運用モデルのインフォメーションレシオ IR_{AI} は以下のように表現する。

$$IR_{AI} = \frac{(\overline{AIr} - \overline{Tpxr}) \times 12}{\sigma(e)_{AI} \times \sqrt{12}}$$

3.2 投資判断アルゴリズムの最適化

本稿では、「学習」、「検証」の2ステップにより最適なネットワーク構造を探索し、AI 運用モデルの投資判断アルゴリズムを決定する。

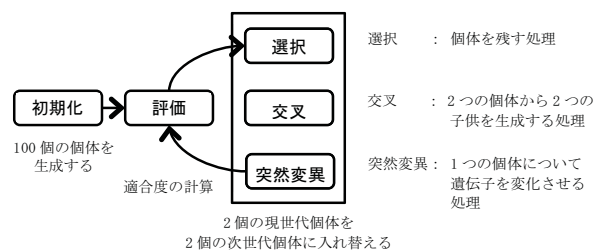
3.2.1 学習

2002年1月～2011年12月の10年間を学習期間とし、当該期間の市場データをAI 運用モデルに学習させる。具体的には、当該期間のインフォメーションレシオを最大化する投資判断アルゴリズムのネットワーク構造を、遺伝的アルゴリズム⁸（以下、GA）を用いて探索することで、AI 運用モデルに運用コンセプトを具現化する投資判断アルゴリズムを適応的に獲得させる。本稿における GA の操作は平林/伊庭(2008)を参考にした。

図3にGAの概念図を示す。GAの探索の流れとしては、まず異なる遺伝子型を持つ個体を多数生成させて初期個体とする。ここで、各個体は投資判断アルゴリズムのネットワーク構造を表す遺伝子型（遺伝子の集まり。各遺伝子の値は、ニューラルネットワークのニューロン間の結合ウェイト $\omega_{jkk}^{s-1(s)}$ を表す）として、図4のように定義される。各遺伝子の値は区間[-1, +1]の実数値とし、初期個体の遺伝子及び突然変異で置換する遺伝子の値は区間[-1, +1]の乱数で決定する。次に各個体について学習期間の売買シミュレーション⁹を行い、各個体の適合度を評価する。次に適合度の高い個体が生き残るような方法で選択を行い、一定数の個体に対して交叉や突然変異の処理（遺伝的操作）を行う。こうして作られた次世代についても適合度の評価を行い、以下同じ処理を繰り返す。

GAのパラメータは、適合度を学習期間のインフォメーションレシオ IR_{AI} 、初期個体数を100個体、遺伝的操作回数を500～10000回、選択の方法をエリート選択、交叉の方法を1点交叉、突然変異の方法を置換、突然変異率を3%とする。

図3 遺伝的アルゴリズムの概念図



(出所) 多田(2016) (筆者により一部修正)

図4 遺伝子型

$\omega_{j1k}^{in(mid1)}$...	$\omega_{j1k}^{s-1(s)}$...	$\omega_{j1k}^{mid5(out)}$...
---------------------------	-----	-------------------------	-----	----------------------------	-----

然変異を繰り返しながら環境に適合していく仕組み」に基づく、数理モデルの最適化の手法を遺伝的アルゴリズムという。

⁹ 各個体の保有資産の組入比率の初期値は、パフォーマンス計測期間の1営業日前の投資判断（買い：TOPIX100%，それ以外：TOPIX0%）により決定する。

⁷ 本稿のインフォメーションレシオは、年率換算したリスクとリターンを用いて算出する。

⁸ 「生物が生存していくなかで、選択、交叉、突

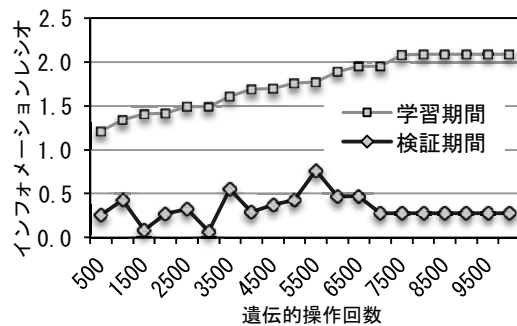
3.2.2 検証

投資判断アルゴリズムが学習期間のデータに過剰適合（過学習）してしまうことを避けるため、学習期間の直後の2年間（2012年1月～2013年12月）を検証期間とし、当該期間のデータを用いて、学習ステップで最適化した個体の検証を行う。

具体的には、学習ステップのGAにおける遺伝的操作の回数が500, 1000, …, 10000回の各時点で適合度が最も高い個体を抽出し、抽出した各個体について検証期間のインフォメーションレシオを計測し、その水準から個体の過剰適合を判断する。

図5は、抽出した各個体の遺伝的操作回数と学習期間及び検証期間におけるインフォメーションレシオの関係を示す。図5より、遺伝的操作回数が5500回を超える辺りから検証期間のインフォメーションレシオが小さくなる傾向を示したため、遺伝的操作を6000回以上行った個体を過剰適合していると判断する。

図5 遺伝的操作回数とインフォメーションレシオ



3.2.3 投資判断アルゴリズムの決定

検証の結果、学習ステップでGAの遺伝的操作を5500回行った個体を、AI運用モデルの投資判断アルゴリズムに決定する。

4 実証分析

本稿で構築したAIの有効性を確認するため、検証期間の直後の3年間（2014年1月～2016年12月）を実証分析期間とし、当該期間の市場データを用いて売買シミュレーションを行う。

4.1 有効性の確認

表4に売買シミュレーションの結果を示す。表4より、実証分析期間におけるAI運用モデルのインフォメーションレシオは0.4946と、一般に良いとされる

目安¹⁰である0.5に近い値となった。これは、AI運用モデルが、運用コンセプトを具現化する投資判断アルゴリズムを適応的に獲得したことを示唆するものと考え、本稿で構築したAIが、実際の資産運用においても有効であるものと解釈する。

表4 売買シミュレーション結果（実証期間）

	収益率 (年率)	トラッキング エラー(年率)	インフォメーションレシオ
AI運用モデル	9.55%	4.96%	0.4946
TOPIX	7.10%	-	-

4.2 売買シミュレーション結果の考察

図6に実証分析期間におけるAI運用モデルとTOPIXの累積収益率の推移、図7に学習期間と実証分析期間における月次のTOPIX収益率とAI運用モデルの超過収益率の関係を示す。

図6 実証分析期間における累積収益率の推移

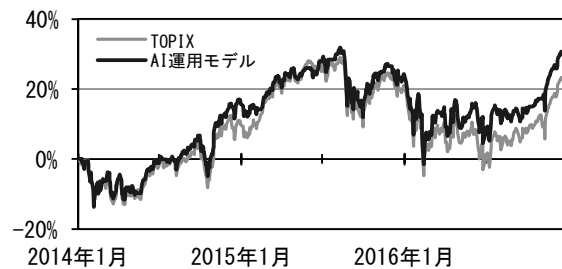
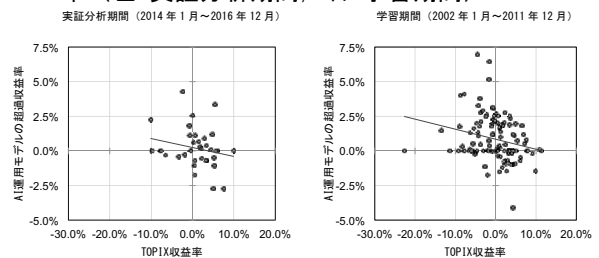


図7 TOPIX月次収益率とAI運用モデルの月次超過収益率（左：実証分析期間，右：学習期間）



(注) グラフ内の直線は回帰直線を表す。

¹⁰ 一般に、インフォメーションレシオが0.5以上であればGood（良い）、1以上であればExcellent（優秀）、2以上であればExceptional（例外的）と言われる（三好(2007)）。但し、インフォメーションレシオの水準は測定するデータ数等によって変わってくるため、いくつ以上でどの程度優秀かを一概に言い表すことは難しい。

図 6 より、AI 運用モデルが TOPIX に比べて高い収益率を獲得していることがわかる。これは、AI 運用モデルが市場環境に応じた機動的なリバランスにより、TOPIX の下落による資産価格の減少を回避したことを示唆するものと考ええる。

また図 7 より、AI 運用モデルが高い超過収益率を獲得した月は、TOPIX が小幅に下落した月に集中しており、TOPIX が大幅に下落した月は超過収益率を相対的に獲得できていない(AI 運用モデルは TOPIX の大幅な下落の兆候を検知できない)という傾向が見られる。これは、AI 運用モデルの知覚データの数が少ないことやニューラルネットワークの階層が浅いことが一因ではないかと考える。

知覚データの拡大やニューラルネットワークの多層化等により、TOPIX が大幅に下落する局面における、AI の投資判断の精度を改善することが今後の課題である。

5 おわりに

本稿では、多層ニューラルネットワークで作成した投資判断アルゴリズムを GA で最適化する手法により「TOPIX に対するリスク調整後リターンを最大化する資産運用 AI」を構築した。また、インフォメーションレシオの水準から、構築した AI が実際の資産運用においても有効であることを示した。

なお、筆者としては現在のところ、資産運用の現場において AI が人間に置き換わるような劇的な変化が起こるとは考えていない。当面の間、既存の運用手法を AI で補完するような形(例えば、業種や個別銘柄の選択眼に長けたファンドマネージャーが、本稿で構築したような機動的なリバランスを行う AI を併用し、運用パフォーマンスを改善させるような方法)で、AI は資産運用の現場に広まっていくのではないかと考える。

最後に、本稿が資産運用の分野における AI に対する理解の広まりや活用の検討につながる一助となることを期待したい。

本稿の内容は、筆者個人に属するものであり、筆者の所属組織の公式見解を示すものではありません。また、本稿に含まれる誤りは全て筆者の責に帰するものです。

参考文献

- [1] 岡本訓幸, 野島哲, 岩田雄一郎, 鴻丸靖弘, 大塚征輝, “新潮流で広がるクオンツ運用のフロンティア”, 三菱UFJ信託資産運用情報 2016年6月号, (2016)
- [2] 柏木亮二, “金融領域での人工知能の活用”, 金融ITフォーカス 2015年11月号, (2015)
- [3] 小牧昇平, 白山晋, “Deep Belief Network を用いた日経平均株価の予測に関する研究”, 第12回人工知能学会 金融情報学研究会 SIG-FIN-012-08, (2014)
- [4] 多田智史, “あたらしい人工知能の教科書”, 翔泳社, (2016)
- [5] 辻岡卓, 山本耕司, “取引戦略生成に適した評価指標の検討”, 第7回人工知能学会 ファイナンスにおける人工知能応用研究会 SIG-FIN-007-05, (2011)
- [6] 砺波元, “資産運用のパフォーマンス測定”, 金融財政事情研究会, (2000)
- [7] 平林明憲, 伊庭斉志, “遺伝的アルゴリズムによる外国為替取引手法の最適化”, 2008年度人工知能学会全国大会(第22回) 論文集 3H1-2, (2008)
- [8] 堀江貞之, 金子久, 川橋仁美, 富永洋子, 浦壁厚郎, “日本の資産運用ビジネス 2016/2017”, 野村総合研究所, (2016)
- [9] 三好秀和, “ファンドマネジメントのすべて”, 東京書籍, (2007)
- [10] 山口和孝, 坂井修一, 田中英彦, “ニューラルネットとGAを用いた株式売買支援システム”, 情報処理学会第65回全国大会講演論文集 2003(1), pp. 215-216, (2003)

潜在特徴関係モデルを用いた時系列金融ネットワークの解析と予測

Time dependent analysis and prediction of financial networks using supervised latent feature relational models

伊藤翔太郎^{1*} 江口浩二¹
Shotaro Ito¹ Koji Eguchi¹

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: In recent years, many researchers have taken keen interest in analyzing various kinds of relational data, such as social networks and financial networks. These data can be expressed as a graph or network where each vertex or node is an entity and each edge or link is a relation between a pair of entities. Moreover, each link is often associated with continuous and/or discrete relational attributes, such as in financial networks, the interest rate for a transaction and whether the transaction is international or intranational. In this paper we focus on max-margin latent feature relational models (called Med-LFRM) that are based on Indian buffet process (IBP) and maximum entropy discrimination (MED). For the estimation of model parameters, the Bayesian estimation is deemed equivalent to minimizing an objective function, which involves misclassification errors. We focus on link prediction problem for the networks with continuous and discrete relational attributes. We also focused on the time dependent analysis for the networks, and therefore, we estimated the model parameters considering the observations in the previous time interval. We demonstrate, through experiments with inter-bank financial networks, the effectiveness of the above model in terms of the link prediction performance.

1 はじめに

近年、社会的ネットワークや金融ネットワークなどの関係データの可用性が増加しており、それらのデータを統計解析に用いることで有用な知見を得ることが課題となっている。このようなデータは、エンティティをノードで、それらの間に存在する関係をリンクで表すようなグラフ構造として表現することができる。既観測のリンクから未観測のリンクを予測するリンク予測は、このようなデータの分析における基本的な問題の一つである [1]。この問題を考えるとき、各エンティティまたはリンクの持つ属性などの情報を利用し予測を行うこともある [2][3]。

確率モデルに基づいた様々な手法を用いたリンク予測に関する研究は発展を続けている。本稿において着目するのは、リンク構造の確率分布を定義するために各ノードが潜在特徴を持つと仮定し、それと共にシグモイド関数などのリンク尤度関数を利用するようなモ

デルである [3][4]。

しかし、潜在特徴の未知の次元数を決定するために、ほとんどの手法においては交差検定などによってモデルを選択する必要がある。この時、多くの異なる訓練データによる結果を比較する必要があり、それにより大きなコストがかかってしまう。そこで、Miller らはインド料理過程 (Indian Buffet Process: IBP) [5] に基づいたノンパラメトリックベイズ法を用いて、データから自動的に未知の潜在特徴の次元数を推定することを提案した [3]。これが潜在特徴関係モデル (Latent Feature Relational Model: LFRM) である。そして LFRM を発展させ、最大エントロピー識別 (Maximum Entropy Discrimination: MED) [6][7] の枠組みに基づき、リンク予測の精度を測るヒンジ損失などの目的関数を直接最小化することによりリンク予測を行うモデルが Zhu によって提案された [8]。これがマージン最大化潜在特徴関係モデル (Max-Margin Latent Feature Relational Model: MedLFRM) である。このモデルは、それぞれ独立に研究されてきたノンパラメトリックベイズ法とマージン最大化法を統合したモデルである。このモデ

*連絡先：神戸大学大学院システム情報学研究科
〒0657-0013 兵庫県神戸市灘区六甲台町 1-1
E-mail: shotaro@cs25.scitec.kobe-u.ac.jp

ルにおいては、バイズ推定の計算が目的関数を最小化することと等価になり、それによって教師ラベルを用いることが可能となる。また、ソフトマージンにより誤分類を許容し、より柔軟なモデルを実現することができる。ソフトマージンの最大化に関する部分問題は、既存の高性能な求解法によって解くことができる。

MedLFRM が、LFRM よりもリンク予測の精度が優れていることが知られている [8]。しかしながら、それは連続値のみで表現された関係属性について結果が示されたものであり、連続値と離散値の関係属性が混在する場合については検討されていない。また、ネットワークの時系列解析についても検討されていない。本稿ではこの二点に着目して評価を行う。後者に関しては、前時区間のデータから学習したパラメータを当時区間のパラメータの初期状態に設定して学習することによって時間依存性を考慮し、リンク予測の精度の向上を図る。

本稿の実験は、LFRM や MedLFRM では連続値表現が仮定されていた関係属性を連続値と離散値が混合したものと仮定する。これは実際の金融ネットワークに対して柔軟に対応するためである。これを踏まえて、2009 年から 2012 年の欧州の銀行間での取引をまとめたデータセットを用いてリンク予測問題に対して実験を行った。

本稿の構成は以下ようになる。第 2 章では、LFRM、MED、MedLFRM、などの既存の手法を紹介する。第 3 章では、MedLFRM による連続値・離散関係属性付きネットワークの時系列解析について述べ、その条件下におけるパラメータの推定方法について述べる。連続値・離散関係属性を考慮したリンク予測の実験結果を第 4 章で示し、第 5 章で結論および今後の課題について述べる。

2 関連研究

2.1 無限潜在特徴関係モデル

潜在特徴関係モデル (LFRM) は Miller らによって提案されたモデルである [3]。このモデルは、各ノードが二値の値をとる潜在特徴ベクトルを持つと仮定し、それらの未知の次元を自動的に推定すると共に、ネットワーク間のリンクが生成される尤度を推定するモデルである。尤度は各ノードの持つ潜在特徴とノード間のリンクに付与される関係属性、そしてそれらの重みを用いて算出される。

ネットワーク内のノード数を N とし、 $N \times N$ の二値隣接行列を Y とする。この時、ノード i とノード j の間にリンクが存在する場合は $Y_{ij} = +1$ とし、リンクが存在しない場合は $Y_{ij} = -1$ とする。 Y は完全には

観測されておらず、既観測のリンクから未観測のリンクの有無を予測できるモデルを学習することが目的となっている。また、ノード i とノード j の間のリンクに作用する関係属性 $X_{ij} \in \mathbb{R}^D$ が観測されている場合もある。

各ノードの持つ潜在特徴の数を K とすると、各ノードは二値潜在特徴ベクトル $\mu_i \in \mathbb{R}^K$ の集合とみなすことができる。ここで Z を $N \times K$ の二値潜在特徴行列とすると、 $Z = [\mu_1^\top; \dots; \mu_N^\top]$ となる。 Z_i はノード i の二値潜在特徴ベクトルを表し、ノード i が潜在特徴 k を持つとき、 $Z_{ik} = 1$ となり、そうでない場合は $Z_{ik} = 0$ となる。また、 W を $K \times K$ の実数値重み行列とし、 $W_{kk'}$ は、ノード i が潜在特徴 k を持ち、ノード j が潜在特徴 k' を持つとき、その二つのノード間のリンクの生成に影響を与える重みであるとする。以上より、リンク尤度は以下のように定義される。

$$p(Y_{ij} = 1 | X_{ij}, Z_i, Z_j) = \Phi(\mu + \eta^\top X_{ij} + Z_i^\top W Z_j) \quad (1)$$

ここで、 Φ はシグモイド関数である。そして、 μ は尤度に影響を与える大域的バイアス値であり、 η は関係属性の実数値重みベクトルである。最適な事前分布を得るために、インド料理過程 (IBP) [5] を Z の事前分布として用いる。これによって、 Z を推定すると同時に、潜在特徴数 K も推定することができる。 W は、各成分において独立して事前分布 $\mathcal{N}(0, \sigma_w^2)$ を仮定する。

IBP を確率が 1 となる無限二値行列の事前分布と仮定する。これにより生成される行列は、潜在特徴をいくつか持っていたとしても各成分は必ず正の値をとる。行列の成分のサンプリングは以下のように行われる。1 番目のノードに対応する行のうち、 $\text{Poisson}(\beta)$ の数だけの成分を 1 とする。ここで、 β はハイパーパラメータである。次に、 i 番目のノードに対応する行に属する成分のうち、既に他の行で 1 となっている成分は、その 1 となっている成分の数に比例した確立で 1 となる。また、 $\text{Poisson}(\beta/i)$ の数だけの成分を新しく 1 にする。これを有限個のノードの数だけ繰り返すことで潜在特徴行列の事前分布を得る。この過程は交換可能であるため、選択される行の順番には影響されない。

2.2 最大エントロピー識別

最大エントロピー識別 (MED) [6][7] は、事前分布を用いて目的関数である正則化項付き相対エントロピー最小化問題を解くことにより事後分布を学習する手法である。

応答変数 Y が $\{+1, -1\}$ を取るような二値分類問題を考える。 X を入力特徴ベクトルとし、 $F(X; \eta) = \eta^\top X_n$ を η によってパラメータ化された識別関数とする。また、 ℓ を正の損失パラメータとし、ヒンジ損失関数を

$h_\ell(x) = \max(0, \ell - x)$ と定義する. η の事前分布を $p_0(\eta)$, 事後分布を $p(\eta)$ とすると, 単一の η を推定する通常の SVM とは異なり, MED は $p_0(\eta)$ を用いて以下の正則化項付き相対エントロピー最小化問題を解くことにより, $p(\eta)$ を学習するものである.

$$\min_{p(\eta)} \text{KL}(p(\eta)||p_0(\eta)) + C\mathcal{R}(p(\eta)) \quad (2)$$

ここで, C は正の定数である. $\text{KL}(p(\eta)||p_0(\eta))$ は KL ダイバージェンス, すなわち相対エントロピーであり, $\mathcal{R}(p(\eta)) = \sum_n h_\ell(Y_n \mathbb{E}_{p(\eta)}[F(X_n; \eta)])$ はヒンジ損失である.

応答変数 Y の予測値は以下ようになる.

$$\hat{Y} = \text{sign} \mathbb{E}_{p(\eta)}[F(X; \eta)] \quad (3)$$

2.3 マージン最大化潜在特徴関係モデル

マージン最大化潜在特徴関係モデル (MedLFRM) は Zhu によって提案されたモデルである [8]. このモデルは, MED における識別関数 F を LFRM におけるリンク尤度として定義することにより, より効果的にリンク尤度を推定できるモデルである.

LFRM と同様に, 二値潜在特徴行列 Z と, 実数値重み行列 W , 関係属性 X_{ij} が与えられると, 識別関数は以下ようになる.

$$\begin{aligned} f(Z_i, Z_j; X_{ij}, W, \eta) &= Z_i W Z_j^\top + \eta^\top X_{ij} \\ &= \text{Tr}(W Z_j^\top Z_i) + \eta^\top X_{ij} \end{aligned} \quad (4)$$

η は関係属性に対する実数値重みベクトルである. ここで, $\Theta = \{W, \eta\}$ をすべてのパラメータとし, Θ と事前分布 $p_0(\Theta)$ は確率変数とする. 予測を行うためには, 潜在変数の不確実性を取り除く必要があるため, より効果的な識別関数を $p(Z, \Theta)$ に関する期待値として以下のように定義する.

$$f(X_{ij}) = \mathbb{E}_{p(Z, \Theta)}[f(Z_i, Z_j; X_{ij}, \Theta)] \quad (5)$$

したがって, 応答変数 Y の予測値は $\hat{Y}_{ij} = \text{sign} f(X_{ij})$ となる. \mathcal{I} を観測されたリンクの組の集合とし, ヒンジ損失関数を 2.2 節と同様に定義すると, ヒンジ損失は以下ようになる.

$$\mathcal{R}_\ell(p(Z, \Theta)) = \sum_{(i,j) \in \mathcal{I}} h_\ell(Y_{ij} f(X_{ij})) \quad (6)$$

ここで, $p_0(Z)$ を潜在特徴行列の事前分布とすると, 以上より MedLFRM を以下の問題を解くことと定義できる.

$$\min_{p(Z, \Theta) \in P} \text{KL}(p(Z, \Theta)||p_0(Z, \Theta)) + C\mathcal{R}_\ell(p(Z, \Theta)) \quad (7)$$

一般的に, 補助変数を導入することによりマージンの依存性を条件付き独立に変換し, 推定を簡単化できるということが知られており, これを変分近似と呼ぶ. これを行うために, Teh らによって提案された IBP の棒折り過程 (Stick Breaking Prior: SBP) [9] を用いる. $\pi_k \in (0, 1)$ を行列 Z の列 k と対応するパラメータとし, このパラメータ π は棒折り過程によって生成される. ここで, $\pi_1 = \nu_1, \pi_k = \nu_k \pi_{k-1} = \prod_{i=1}^k \nu_i$ であり, ν_i は α をハイパーパラメータとするベータ分布 $\text{Beta}(\alpha, 1)$ からサンプリングされるとする. ある π_k について, 列 k 内の各 Z_{nk} は, 独立にベルヌーイ分布 $\text{Bernoulli}(\pi_k)$ からサンプリングされるものとする. この過程を経て, 連続確率 π_k は減少し, データセット上の特徴 k を観測する確率は指数関数的に減少する.

MedLFRM におけるリンクの生成過程を以下に示す.

1. 潜在特徴行列の 1 行目である Z_1 に対して,
 - (a) ポアソン分布 $\text{Poisson}(\beta)$ から成分数 M を選択
 - (b) ベータ分布 $\text{Beta}(\alpha, 1)$ から, 各成分 $i \in \{1, \dots, M\}$ に対してパラメータ ν_i を選択
 - (c) 棒折り過程 SBP(ν) から, 各成分 $k \in \{1, \dots, M\}$ に対してパラメータ π_k を生成
 - (d) ベルヌーイ分布 $\text{Bernoulli}(\pi_k)$ から, 各成分 $Z_{1k} \in \{Z_{11}, \dots, Z_{1M}\}$ を選択
2. 潜在特徴行列の n 行目である Z_n に対して,
 - (a) これまで選択されていない成分に対してはポアソン分布 $\text{Poisson}(\beta/n)$ から, 既に選択されたことのある成分に対してはその時用いられた確率から成分数 M を選択
 - (b) ベータ分布 $\text{Beta}(\alpha, 1)$ から, 各成分 $i \in \{1, \dots, M\}$ に対してパラメータ ν_i を選択
 - (c) 棒折り過程 SBP(ν) から, 各成分 $k \in \{1, \dots, M\}$ に対してパラメータ π_k を生成
 - (d) ベルヌーイ分布 $\text{Bernoulli}(\pi_k)$ から, 各成分 $Z_{nk} \in \{Z_{n1}, \dots, Z_{nM}\}$ を選択
3. 正規分布 $\mathcal{N}(0, \sigma_w^2)$ から, 重み行列 W の各成分を選択
4. 正規分布 $\mathcal{N}(0, \sigma_\eta^2)$ から, 重みベクトル η の各成分を選択
5. リンク評価関数 $\Phi(Z_i W Z_j^\top + \eta^\top X_{ij})$ を用いて, 各ノード対 $(i, j) \in \mathcal{I}$ に対して応答変数 Y_{ij} を生成

補助変数を導入することによって拡張された問題を以下に示す.

$$\min_{p(\nu, Z, \Theta)} \text{KL}(p(\nu, Z, \Theta)||p_0(\nu, Z, \Theta)) + C\mathcal{R}_\ell(p(Z, \Theta)) \quad (8)$$

ここで, $p_0(\nu, Z, \Theta) = p_0(\nu)p(Z|\nu)p_0(\Theta)$

3 提案手法

前章ではネットワークデータに対しての MedLFRM によるリンク予測問題について述べた。この章では、連続値で表現された関係属性に加え、離散関係属性がリンクに付与されている場合を想定したリンク予測問題について述べる。またそのようなモデルを用いたネットワークの時系列解析についても述べる。

3.1 MedLFRM による連続値・離散関係属性付きネットワークの解析

前章で述べていたモデル [8] では関係属性は連続値のみで表現されていたが、実データセットではしばしば関係属性の中に離散表現されているものが存在することもある。したがって、そのようなデータセットに対するリンク予測問題を扱うとき、関係属性を連続値表現とも離散表現ともする必要がある。したがって、 X_{ij} はこれまで $X_{ij} \in \mathbb{R}^D$ と定義されていたが、 X_{ij}^c を離散表現された関係属性、 X_{ij}^r を連続値表現された関係属性とすると、 $X_{ij}^c \in \mathbb{I}^D, X_{ij}^r \in \mathbb{R}^{D'}$ といったように再定義し議論を行う。この時、識別関数内の関係属性の重みベクトル η は連続値関係属性と離散関係属性それぞれに定義する。よって、識別関数は以下ようになる。

$$\begin{aligned} f(Z_i, Z_j; X_{ij}^c, X_{ij}^r, W, \eta^c, \eta^r) \\ &= Z_i W Z_j^\top + \eta^{c\top} X_{ij}^c + \eta^{r\top} X_{ij}^r \\ &= \text{Tr}(W Z_j^\top Z_i) + \eta^{c\top} X_{ij}^c + \eta^{r\top} X_{ij}^r \quad (9) \end{aligned}$$

ここで、 $\Theta = \{W, \eta^c, \eta^r\}$ をすべてのパラメータとし、 Θ と事前分布 $p_0(\Theta)$ は確率変数とする。2.3 節と同様に、ここからより効果的な識別関数を $p(Z, \Theta)$ に関する期待値として以下のように定義する。

$$f(X_{ij}^c, X_{ij}^r) = \mathbb{E}_{p(Z, \Theta)} [f(Z_i, Z_j; X_{ij}^c, X_{ij}^r, \Theta)] \quad (10)$$

したがって、応答変数 Y の予測値は $\hat{Y}_{ij} = \text{sign} f(X_{ij}^c, X_{ij}^r)$ となる。 \mathcal{I} を観測されたリンクの組の集合とし、ヒンジ損失関数を 2.2 節と同様に定義すると、ヒンジ損失は以下ようになる。

$$\mathcal{R}_\ell(p(Z, \Theta)) = \sum_{(i,j) \in \mathcal{I}} h_\ell(Y_{ij} f(X_{ij}^c, X_{ij}^r)) \quad (11)$$

3.2 MedLFRM によるパラメータの推定

次に MedLFRM によるパラメータの推定方法について述べる。MedLFRM を提案した Zhu の手法 [8] に従った推定方法を以下に述べる。

切断平均場近似 (truncated mean field approximation) [10] によって、 $p(\nu, Z, \Theta)$ を次のように表す。

$$p(\nu, Z, \Theta) = p(\Theta) \prod_{k=1}^K p(\nu_k | \gamma_k) \left(\prod_{i=1}^N p(Z_{ik} | \psi_{ik}) \right) \quad (12)$$

ここで、 $p(\nu_k | \gamma_k)$ はベータ分布 $\text{Beta}(\gamma_{k1}, \gamma_{k2})$ からサンプリングされたもの、 $p(Z_{ik} | \psi_{ik})$ は $\text{Bernoulli}(\psi_{ik})$ からサンプリングされたものである。 K は切断レベルである。これらを踏まえて、MedLFRM の問題は以下の手順を反復することで解くことができる。

1. $p(\Theta)$ の推定

$p(\nu, Z)$ が与えられたとき、部分問題を以下の制約の形で書くことができる。

$$\begin{aligned} \min_{p(\Theta), \xi} \text{KL}(p(\Theta) || p_0(\Theta)) + C \sum_{(i,j) \in \mathcal{I}} \xi_{ij} \\ \forall (i,j) \in \mathcal{I}, \text{ s.t. : } Y_{ij} (\text{Tr}(W \bar{Z}_{ij}) + \mathbb{E}[\eta^{c\top} X_{ij}^c + \mathbb{E}[\eta^{r\top} X_{ij}^r]) \\ \geq \ell - \xi_{ij} \quad (13) \end{aligned}$$

ここで、 $\bar{Z}_{ij} = \mathbb{E}_p[Z_j^\top Z_i]$ は潜在特徴ベクトルの内積の期待値であり、 $\xi = \{\xi_{ij}\}$ はソフトマージンを実現するためのスラック変数である。ラグランジュの双対理論を用いることで、 $p(\Theta)$ の最適解を得ることができる。 $p(\Theta)$ は以下のように表すことができる。

$$p(\Theta) \propto p_0(\Theta) \exp \left\{ \sum_{(i,j) \in \mathcal{I}} \omega_{ij} Y_{ij} (\text{Tr}(W \bar{Z}_{ij}) + \eta^{c\top} X_{ij}^c + \eta^{r\top} X_{ij}^r) \right\}$$

$\omega = \{\omega_{ij}\}$ はラグランジュ乗数である。

ここで、 η を η^c と η^r を接続して得たベクトルであるとする。一般に使用される標準正規事前分布 $p_0(\Theta)$ により、 $p(\Theta)$ の最適解を得ると以下のように表すことができる。

$$p(\Theta) = p(W) p(\eta) = \left(\prod_{kk'} \mathcal{N}(\Lambda_{kk'}, 1) \right) \left(\prod_d \mathcal{N}(\kappa_d, 1) \right)$$

ここで、 $\mathcal{N}(\Lambda_{kk'}, 1), \mathcal{N}(\kappa_d, 1)$ のそれぞれの期待値は、 $\Lambda_{kk'} = \sum_{(i,j) \in \mathcal{I}} \omega_{ij} Y_{ij} \mathbb{E}[Z_{ik} Z_{jk}']$ 、 $\kappa_d = \sum_{(i,j) \in \mathcal{I}} \omega_{ij} Y_{ij} (X_{ijd}^c + X_{ijd}^r)$ とする。双対問題は以下ようになる。

$$\begin{aligned} \max_{\omega} \sum_{(i,j)} \ell \omega_{ij} - \frac{1}{2} (\|\Lambda\|_2^2 + \|\kappa\|_2^2) \\ \text{s.t. : } 0 \leq \omega_{ij} \leq C, \forall (i,j) \in \mathcal{I} \end{aligned}$$

この時、部分問題は等価的に以下のように書き換えることができ、これを解くことでパラメータ Λ

と κ を直接求めることができる.

$$\min_{\Lambda, \kappa, \xi} \frac{1}{2} (\|\Lambda\|_2^2 + \|\kappa\|_2^2) + C \sum_{(i,j) \in \mathcal{I}} \xi_{ij}$$

$$\forall (i, j) \in \mathcal{I}, \text{ s.t. } : Y_{ij} (\text{Tr}(\Lambda \bar{\mathbf{Z}}_{ij}) + \kappa^\top (X_{ij}^c + X_{ij}^r)) \geq \ell - \xi_{ij} \quad (14)$$

これは, SVM (Support Vector Machine) の二値分類問題の形式と一致しているため, SVMLight や LIBSVM などの既存の高性能なソルバーによって解くことができる.

2. $p(\nu, Z)$ の推定

$p(\Theta)$ が与えられると, 部分問題は以下のようになる.

$$\min_{p(\nu, Z)} \text{KL}(p(\nu, Z) \| p_0(\nu, Z)) + C \mathcal{R}_\ell(p(Z, \Theta))$$

切断平均場近似より, 以下の式が得られる.

$$\text{Tr}(\Lambda \bar{\mathbf{Z}}_{ij}) = \begin{cases} \psi_i \Lambda \psi_j^\top & \text{if } i \neq j \\ \psi_i \Lambda \psi_i^\top + \sum_k \Lambda_{kk} \psi_{ik} (1 - \psi_{ik}) & \text{if } i = j \end{cases}$$

マージンの制約は ν に依存しないので, $p(\nu)$ は Doshi-Velez ら [11] と同様にして解を得ることができる.

また, 劣勾配法を用いることで $p(Z)$ を解くことができる. ここで, 観測されたリンクの組の集合 \mathcal{I} を以下のように定義する.

$$\mathcal{I}_i = \{j : j \neq i, (i, j) \in \mathcal{I} \text{ and } Y_{ij} f(X_{ij}^c, X_{ij}^r) \leq \ell\}$$

$$\mathcal{I}'_i = \{j : j \neq i, (j, i) \in \mathcal{I} \text{ and } Y_{ji} f(X_{ji}^c, X_{ji}^r) \leq \ell\}$$

また, $g(x) \leq \ell$ であるとき, $\partial_x h_\ell(g(x))$ は $-\partial_x g(x)$ に等しく, そうでなければ 0 となる. これより, \mathcal{R}_ℓ の劣勾配は以下ようになる.

$$\partial_{\psi_{ik}} \mathcal{R}_\ell = - \sum_{j \in \mathcal{I}_i} Y_{ij} \Lambda_k \cdot \psi_j^\top - \sum_{j \in \mathcal{I}'_i} Y_{ji} \psi_j \Lambda_{\cdot k}$$

$$- \mathbb{I}(Y_{ii} f(X_{ii}^c, X_{ii}^r) \leq \ell) Y_{ii} (\Lambda_k \cdot \psi_i^\top + \psi_i \Lambda_{\cdot k} + \Lambda_{kk} (2 - \psi_{ik}))$$

ここで, Λ_k は Λ の k 番目の行を表し, $\Lambda_{\cdot k}$ は Λ の k 番目の列を表す. また, $\mathbb{I}(\cdot)$ は指示関数であり, 括弧内に示された命題が真であるとき 1, そうでないとき 0 を返す. 部分問題の劣勾配を 0 とすると, 以下の ψ_{ik} の更新式を得る.

$$\psi_{ik} = \Phi \left(\sum_{j=1}^k \mathbb{E}_p[\log \nu_j] - \mathcal{L}_k^\nu - C \partial_{\psi_{ik}} \mathcal{R}_\ell \right) \quad (15)$$

ここで, \mathcal{L}_k^ν は $\mathbb{E}_p[\log(1 - \prod_{j=1}^k \nu_j)]$ の下限である.

3.3 MedLFRM による時系列ネットワークの解析

ここまでは一つのネットワークにおけるリンク予測問題について着目してきた. この節では複数のネットワークを時系列的に捉えたときの予測について述べる.

これまで述べられてきていたモデルにおいては, 潜在特徴行列 Z は ψ をハイパーパラメータとする Bernoulli(ψ) からサンプリングされたものであり, この ψ をモデルの中でランダムに初期化, 学習し予測を行っていた. しかし, 複数のネットワークを時系列的に捉えたとき, 前時区間のデータの ψ の学習結果を当時区間においての初期状態として据えることで時間依存性を考慮する.

また, 識別関数 F 内の実数値重み行列 W , 実数値重みベクトル η もモデル内でランダムに初期化, 学習されるものであるが, この二つについても先ほど述べたように前時区間のデータの学習結果を当時区間においての初期状態に据えることで時間依存性を考慮することができ, 予測精度の向上につながると考える.

4 実験

この章では実データセットを用いて, 時間 $t-1$ までの学習結果を用いて時間 t のデータを学習し, 時間 $t+1$ におけるリンクを予測する実験を行い, その結果について考察する.

4.1 データセット

実験には, 欧州債務危機が起こった 2009 年 7 月から 2012 年 12 月における欧州銀行間での取引を記録したデータセットを用いる. 通年のデータを月ごとに集計し, 12 個のデータセットとして扱う. データセットには 153 の銀行と 14 の国が含まれる. データセットに含まれる各銀行をノード, 当月にある銀行と別の銀行との間に現金のやり取りがあった場合, そこにリンクがあると見なす. この時, 現金のやり取りというのは銀行 A から銀行 B へと x ユーロの現金が送られたというような状態を指す. したがって, このネットワークは有向グラフとなり, 隣接行列は非対称行列となる. リンクに付与される関係属性は, 取引された現金の量, その取引が行われた際の金利, そして取引が同じ国内の銀行間で行われたのであれば 1, 異国の銀行間で行われたのであれば -1 をとるような二値変数の 3 つとする.

4.2 実験設定

次にパラメータの設定について述べる. 切断レベル K は 40 よりも大きくすれば十分に良い結果を得られ

表 1: 提案手法と従来手法による翌月の取引有無に関する予測結果

	MAP	average AUC
time-dependent model	0.387637	0.919695
time-independent model	0.346412	0.905515

ることが知られている [8] ことから, 本実験においても $K = 50$ と設定する. 損失パラメータ l , ハイパーパラメータ α は全ての実験において $l = 1, \alpha = 0.1$ と設定した. また, このデータセットは不均衡である. (つまり, 正例よりも負例の方が多く存在する.) したがって, 正のデータに対しては C^+ , 負のデータに対しては C^- という異なった正則化定数を用いることにし, $C^- = 0.1$ とした上で $C^+ = 10C^-$ とした. 実験を行う時, 先に述べたように前月のパラメータの学習結果を当月の初期状態として設定する. ただし, 1 回目の実験を行う時, つまり 1 月のデータセットに対して実験を行う際は, W は $[0, 0.1]$ の区間で一様に初期化, ψ は 0.5 に $[0, 0.001]$ の区間で一様に分布したランダムノイズを加えたものとなるように初期化, η は平均が 0 となるように初期化する.

4.3 評価方法と結果

先に述べたパラメータを用いて実験を行う. また, それとは別に前月のパラメータの学習結果を用いず, ランダムに初期化した上で学習していく従来手法でも実験を行う. この二つの結果の MAP (mean average precision) を比較する. MAP は予測精度評価法の 1 つであり, $[0, 1]$ の範囲の値を取りうるもので, 1 に近いほど精度がよいということになる. また同様にして得られた AUC (Area Under the Curve) の平均 (12ヶ月分) も評価指標として用いる. AUC も予測精度評価法の一つであり, ROC (Receiver Operating Characteristic) 曲線の積分値を表すものである. 完全に理想的な予測をした場合に 1 を, 完全にランダムな予測をした場合に 0.5 をとる. このときの評価結果を表 1 に示す.

また, 4.1 節で述べた 3 つの関係属性のうち, どれか 1 つを考慮せずにリンク予測を行うという実験を全ての関係属性に対して行い, その結果を比較する. この実験を行う際も前月のパラメータの学習結果を用いる場合と用いない場合に分けて実験を行う. このときの評価結果を表 2~表 4 に示す.

4.4 考察

表 1 から, 前月のパラメータの学習結果を当月の初期状態に据えることで, ランダムに初期化を行って予測

表 2: 国際取引, 国内取引の別を考慮しない場合の予測結果

	MAP	average AUC
time-dependent model	0.375433	0.915590
time-independent model	0.347404	0.906652

表 3: 取引量を考慮しない場合の予測結果

	MAP	average AUC
time-dependent model	0.380175	0.918378
time-independent model	0.346403	0.905512

表 4: 金利を考慮しない場合の予測結果

	MAP	average AUC
time-dependent model	0.362422	0.911382
time-independent model	0.347434	0.905525

する場合よりも精度が改善されていることが確認できる. 次に, AUC よりも MAP の方が比較的向上している点に関して考察する. まず AUC はリンクが有るノード対にリンクが有ると予測できた場合とリンクが無いノード対にリンクが無いと予測できた程度が全ノード対を占める割合を表す指標であり, MAP に関してはリンクが有るところを有ると予測できた割合だけに注目した指標である. そして 4.2 節で述べたようにこのデータセットは不均衡なものであるため, リンクが有るということを予測できたことにより価値があると考えられる. よって, 今回の実験においてより評価指標として意味を持つのは MAP の方であるといえる. その MAP により大きな改善が見られたことから, 前月のパラメータの学習結果を当月のパラメータの初期状態として用いることで効果的に時間依存性を反映させることができたと言える. さらに, 各関係属性を考慮せずに行った実験結果について考察する. 金利を考慮せずに行った場合に比較的大きく予測精度が損なわれていることから, 取引の有無を予測するという観点においては, 金利がどの程度であったかという情報がより重要なだと予想することができる. 逆に, 取引量を考慮せずに行った場合には比較的予測精度に差は見られないため, 過去の取引量は将来の取引の有無を予測するにあたってはさほど重要ではないと解釈できる.

5 おわりに

本稿では, 実問題において想定される連続値と離散値で表現される関係属性が混在する場合のネットワークに対するリンク予測問題について, マージン最大化潜在特徴関係モデル (MedLFRM) を拡張し, 予測精度の評価を行った. そしてそのようなネットワークの時系列解析についても検討した.

銀行間の取引データを用いて、前時区間において学習したパラメータを当時区間のパラメータの初期状態として据えながらリンク予測問題の実験を行い、各月のAUCの平均と、MAPを用いて評価を行った。比較対象として、従来の手法と同じく時系列を考慮せず、パラメータをランダムに初期化するリンク予測問題の実験も行った。その結果、前時区間のパラメータの学習結果を用いた方が予測精度が良くなることが確認された。これより、ネットワークを時系列的に捉えて未知パラメータや潜在変数を学習することで、リンク予測の性能を改善できることを示した。

今後の展望として、損失パラメータ ℓ や正則化定数 C の適切な値を交差検定などの手法によって決定し、リンク予測問題において精度の更なる向上を試みる事が挙げられる。また、本稿ではリンク予測問題について議論を行ったが、リンクに付与された関係属性の予測を行うといったことも考えられる。この場合、関係属性が連続値で表現されているのか離散値で表現されているのかによって、問題設定を回帰または分類に区別して考える必要がある。関係属性の予測を行うことによりノード間にあるリンクがどのようなものであるのかを予測でき、さらに本稿において用いた時系列的な考えを導入することで、より詳細で広範囲なネットワーク解析が可能になると考えられる。

謝辞

本研究を行うにあたり、助言と協力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、同大学大学院システム情報学研究科谷口隆晴准教授に感謝する。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 Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, Vol. 58, No. 7, pp. 1019.1031, 2007.
- [2] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 635.644. ACM, 2011.
- [3] Kurt T Miller, Thomas L Griffiths, and Michael I Jordan. Nonparametric latent feature models for

link prediction. In *Advances in Neural Information Processing Systems*, Vol. 22, pp. 1276.1284, 2009.

- [4] Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, Vol. 20, pp.737.744, 2007.
- [5] Thomas L Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, Vol. 18, pp. 475.482, 2005.
- [6] Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems*, Vol. 12, pp.470.476, 1999.
- [7] Tony Jebara. *Machine learning: discriminative and generative*. Springer, 2004.
- [8] Jun Zhu. Max-margin nonparametric latent feature models for link prediction. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [9] Yee W Teh, Dilan Gorrur, and Zoubin Ghahramani. Stick-breaking construction for the indian buffet process. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pp. 556.563, 2007.
- [10] Christopher M Bishop, et al. *Pattern recognition and machine learning*. springer New York, 2006.
- [11] Finale Doshi, Kurt Tadayuki Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the indian buffet process. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 137.144,2009.

LSTM-RNN を用いたイベント考慮後の株価時系列予測

Predicting Equity Price With Corporate Action Events Using LSTM-RNN

南 正太郎*¹
Shotaro Minami*¹あすかアセットマネジメント株式会社
Asuka Asset Management Co., Ltd

The forecasting the stock price of a particular has been a difficult task for many of analysts and researchers. In fact, investors are highly interested in the research area of stock price prediction. However, to improve the accuracy of forecasting a single stock price is a really challenging task, therefore in this paper, I propose a sequential learning model for prediction of a single stock price with corporate action event information and Macro-Economic indices using LSTM-RNN method. The results show the proposed model is expected to be a promising method in the stock price prediction of a single stock with variables like corporate action and corporate publishings.

1. はじめに

株式資産運用の実務においては、株価を予測することは重要な仕事の一つである。最終的には株価は需給によって決定するものの、企業の将来の業績を調べることで売買を判断する投資家もいれば、株価の動きだけを見て判断する投資家もいる。また、PER や PBR といった株式指標を参考に割安割高を判断する投資家もいるだろう。好きな企業だから買うという投資家もいるかもしれない。このようにさまざまな見通しや考え方をを持った市場参加者によって株価は決まるため、将来の株価を予測するのは簡単ではない。

企業の将来の業績から株価を予測する手法には、例えば将来の株式配当を現在価値に割引くことで求める配当割引モデルや企業が事業を行った結果生み出される将来キャッシュフローの現在価値を求める割引キャッシュフローモデルなどがある。また、株価は EPS (一株当たり純利益) \times PER (株価収益率) によって構成されていると考える倍率法では、 PER は変わらないと想定した場合、将来の業績が好調のため EPS が倍になるのであれば、株価も単純に倍になると考えられる。少なくともその株価水準に長期的には収斂されていくだろうと期待される。

個別企業を分析する人が、多くの企業で四半期ごとに公表される企業の決算発表や日々のプレスリリースに着目するのは、業績の進捗を知ること、これまでの見通しが正しかったかどうかを判断したり、今後の企業の成長がどの程度期待できるのかを見極める意味がある。将来にわたって企業価値を高めると投資家の多くが考える情報があれば株価は上昇するからである。

このように、将来の株価を予測する上では、企業の決算発表やプレスリリースといった情報はとても重要な意味を持つ。また、過去の株価はこのような情報に基づいて判断してきた投資家の思惑が含まれているといえるだろう。

本稿で提案する LSTM-RNN (Long Short Term Memory Recurrent Neural Network) は、再帰的ニューラルネットワーク (Recurrent Neural Network) の一種であり、通常の RNN において長期の系列データを学習する際に発生する勾配問題を回避することから、長期の時間依存も短期の時間依存も学習

できる特性がある。従来の RNN では考慮できない、過去の情報をネットワーク内に保持し、必要なタイミングでのみ取得・置換したいとの課題から登場してきた背景がある。

発表された情報がどの時点の株価に影響しているのかは知ることは難しいものの、そのような状況を織り込んだ株価予測を行う上では、LSTM-RNN は有効な手段になるものと思われる。

2. 関連研究と本稿について

LSTM-RNN を用いた株価予測を行った関連研究について説明する。

株価だけに応用した研究としては、Murtaza らは、インドの取引所の NIFTY50 銘柄の株価について、Open、Close、High、Low それぞれのデータを組み合わせたときの LSTM-RNN による株価の予測精度について比較している。結果としては4つのデータを用いて予測を行った場合が最も望ましいことを示した。また、Qun らは、個別銘柄の寄り付きの株価について、上海総合指数とフォーラム内の情報から取得した感情データを変数にして予測することを行っている。結論としては従来型の RNN を適用するよりも良好な結果が得られたようだ。

LSTM-RNN は時系列データ分析において適用する場合、効果が高いことが知られているが、株価を予測するという点においてもモデルの設定の仕方によっては先の論文のように有効な予測モデルの手段になりえる可能性がある。

そこで本稿では、LSTM-RNN を用いることで企業のイベントを織り込んだ株式予測の手法を提案する。具体的には、特定の企業を取り上げ、決算発表日、コーポレートアクションのプレスリリース、企業が投資家向けに公表している受注残高を変数に用いて、将来株価を予測する。

3. LSTM-RNN を用いた株価時系列予測

3.1 モデル

x_t は t 時点の入力値 (ベクトル)、 h_t は出力値とするとき、LSTM-RNN は以下に定式化される。

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t \odot \sigma_h(c_t) \quad (5)$$

c_t はメモリセルの状態ベクトル、 W 、 U 、 b はパラメータ行列とベクトルである。また、 f_t は古い情報を記憶する重みベクトルである忘却ゲート、 i_t は新しい情報を得るための重みベクトルである入力ゲート、 o_t は出力候補のベクトルを出す出力ゲートを表す。活性化関数のうち、 σ_g はシグモイド関数、 σ_c, h は双曲線正接関数である。

3.2 データについて

分析対象として、株式会社ツガミ（証券コード：6101）の株価を取り上げ分析を行った。工作機械メーカーであるツガミは、自動旋盤機を作っている世界的ニッチ企業である。生産と売り上げの大半は中国であることから、企業業績は国内の受注だけでなく海外の受注に大きく依存している。また、国内市場が縮小する中、海外へ事業の基盤を移す政策を進めており、また同時に明確な資本政策も発表していることから、生産工場の移管や自己株式の取得、ストックオプションなど積極的なコーポレートアクションを公表している。シクリカル銘柄にも分類されることがあるツガミは、中小型株であることに加えて、マクロ指標（工作機械受注高）や企業が月中に発表する前月の国内と海外の受注高推移、コーポレートアクションのプレスリリースに株価が反応しやすい特徴がみられる。

今回分析に用いる変数は、株価に加えて、企業が HP 上に公表するイベント情報（プレスリリース）と企業の国内および海外の受注残高、内閣府が提供している工作機械受注高である。イベント情報は、決算発表日などの法定開示に関するものと、自己株式取得、ストックオプションの発行などコーポレートアクションに関するものを分けて変数にしている。

なお、日次で分析するために月次および四半期のマクロ指標については、多項式スプライン補完を行った。使用したデータの期間はデータ間で共通に取得できた 2013 年 4 月 15 日から 9 月 11 日までである。イベント情報については、データクリーニングの過程で ID として数値にエンコーディングしている。データが公表された翌日から予測に使えるようにラグを考慮している。全期間 1082 日のうち 700 日をトレーニング期間、3 割にあたる残りの期間をテスト期間とした。

3.3 モデルの検証

実運用を想定するならば実績値と乖離がないほうが望ましい。そこで、RMSE(Root Mean Squared Error) を使ってモデルの評価をすることにした。

$$RMSE = \sqrt{\frac{1}{N} \sum (\hat{y}_t - y_t)^2} \quad (6)$$

\hat{y}_t は予測値、 y_t は実績値。

3.4 実装

1999 年に Hochreiter らによって提唱された LSTM-RNN に基づいている。計算するにあたって、学習時には、Epoch 数を 50、100、500 の 3 つのパターンで計算。隠れ層のユニット数は 50、LSTM 層は 1 とした。学習のフレームワークとしては、フロントエンドに Keras、バックエンドには Tensorflow を使用して実装している。確率的勾配降下法の最適化を行う上で、Adam(Adaptive Moment Estimation) により学習率の設

定を行っている。また、GPU は GeForce GTX 1060 3GB を使用した。

4. 検証結果

表 1 は、説明する変数に株価のみを用いた場合、イベントのみを用いた場合、企業が公表する受注残高とマクロ指標を用いた場合、イベント及び受注残高、マクロ指標のすべてを用いた場合について、それぞれの RMSE につき Epoch 数別にまとめたものである。イベントを考慮した場合は、株価のみを用いた予測よりも総じて予測誤差が小さい。マクロ指標を用いた場合は、Epoch 数が 500 の時のみ予測誤差が小さくなっていく。すべてを変数にした場合は、むしろ精度が低くなってしまいう結果となった。運用実務においては、実績値の上昇・下降をどのくらい予測できているかが重要になることから、一致率も表 2 に掲載している。どの分析においても 50 % に満たないことからあまり実用的ではないものの、株価のみを使用した結果よりも一致率が高くなるパターンもあるようだ。ただし、使用したデータが少ないことから、改善する可能性はある。

RMSE	株価のみ	イベント	マクロ指標	すべて
Epoch = 50	19.5	18.9	19.9	20.4
Epoch = 100	17.1	16.9	17.6	18.1
Epoch = 500	15.0	14.9	14.8	15.3

表 1: RMSE

一致率	株価のみ	イベント	マクロ指標	すべて
Epoch = 50	0.4842	0.4868	0.4711	0.4737
Epoch = 100	0.4895	0.4789	0.4921	0.4947
Epoch = 500	0.4579	0.4526	0.4553	0.4711

表 2: 一致率

5. 結論

本研究は、LSTM-RNN を用いて特定の企業が出すイベント情報や受注残高を織り込んだ上での将来の株価を予測するというを試みた。説明する変数に株価のみを用いた場合、イベントのみを用いた場合、企業が公表する受注残高とマクロ指標である工作機械受注高を用いた場合、イベント及び受注残高、マクロ指標のすべてを用いた場合を比較して株価の予測値と実績値の差がどの程度異なるのかについて比較検討した。結果としてはイベントのみを用いた予測が最も実績値との誤差が小さいことが分かった。Epoch 数を変えて計算してもその傾向は変わらないことから、当該企業についてはイベント情報が将来株価を予測する上で有効かもしれない。ただし、本稿は、このようなイベントを考慮する時系列予測の提案が目的であり、分析の精度を高めるためにはより長期にわたる分析を行うことが望ましいだろう。

本研究で提案したモデルは金融の世界では幅広い応用の可能性を持っている。株価に影響を与える変数は企業によって異なるであろうから、個別企業ごとに分析してみるのも興味深い。また、例えば、国内に限らず上場企業は法定開示資料や決算報告など定期的な情報の開示が必須であることから海外にも

同様の分析を行うことができる。今回の研究においては単にイベントがあったという情報だけであるが、その情報にポジティブ・ネガティブの属性を持たせたものでもよい。むしろそのほうが株価を説明する材料としては適切なのではないだろうか。株価を予測するだけでなく、企業が開示する受注状況からマクロ指標を予測することにも応用できるであろう。最適な自己株式の取得のタイミングを評価することに応用できれば企業にとって役に立つ可能性もある。今後の課題である。

参考文献

- [南 17] 南正太郎, 光定洋介: AIの資産運用への応用の可能性と限界, 証券アナリストジャーナル, Vol.55, 8月号 (2017).
- [Jason 17] Jason Brownlee: Long Short-Term Memory Networks With Python Develop Sequence Prediction Models With Deep Learning, MACHINE LEARNING MASTERY (2017).
- [Argimiro 14] Argimiro Arratia: Computational Finance An Introductory Course with R, ATLANTIS PRESS (2014).
- [Akhter 14] Akhter Mohiuddin Rather: A Hybrid Intelligent Method of Predicting Stock Returns, Hindawi Publishing Corporation, Advances in Artificial Neural Systems Volume 2014, Article ID 246484, 7 pages (2014).
- [Murtaza 2015] Murtaza Roondiwala, Harshal Patel, Shraddha Varma: Predicting Prices Using LSTM, International Journal of Science and Research, 2319-7064 (2015).
- [Qun 2017] Qun Zhuge, Lingyu Xu and Gaowei Zhang: LSTM Neural Network with Emotional Analysis for Prediction of Stock Price, Engineering Letters, 25:2, EL_25.2.09 (2017).

機械学習アルゴリズムによるビットコイン価格変動の予測

Prediction of Bitcoin Price Movements with Machine Learning Algorithms

新立 拓也¹ ピフル ルカーシュ¹ 海蔵寺 大成¹

Takuya Shintate¹, Lukáš Pichl¹, and Taisei Kaizoji¹

¹ 国際基督教大学

¹International Christian University

Abstract: We study the limits of prediction accuracy of Bitcoin price data in CNY currency using tick data from the OKCoin Bitcoin exchange (source: Kaiko data). The tick data contain the price, volume, and trade direction, and are transformed to the OHLCV format using standard methods. In this report, we deploy the Support Vector Machine algorithm by Vapnik to estimate the sign of the hour-to-hour transaction return using a sampling moving window of varying size on the past data. Several kernel functions are validated. Our first results for all months of the year 2015 show that the hit ratio accuracy level (the fraction of correctly predicted upward or downward events) does not exceed 60%. It remains to be established whether this low result corresponds to the causal extraction limit inherent in the data, or whether it can be improved by deploying other methods, such as LSTM networks in deep learning.

Introduction

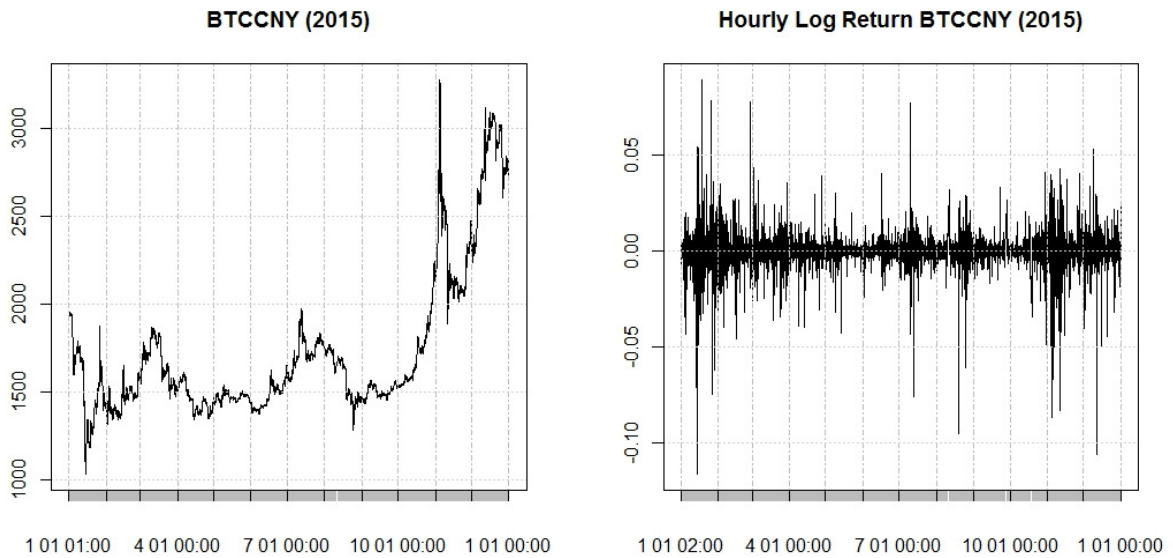
Bitcoin price data time series denominated in standard currencies are subject to many extreme events, since bubbles and crashes in the cryptocurrency markets are quite common. For instance, at the beginning of 2017, the Bitcoin price was about 1000 USD, and since then it has seen the rise to almost 5000 USD on September 1st of the same year, followed by a sharp drop in just a two-week period to about 3000 USD-level on September 14th, when China (accounting for major source of Bitcoin demand) announced suspension of Bitcoin trades at domestic exchanges. Therefore the BTC (XBT) price series are a serious benchmark both for econometric and machine learning algorithms.

The available literature related to Bitcoin time series prediction is relatively limited, with the exception of the burst of mostly economic publications this year [1-3]. The insight from the field of computational intelligence still remains rather limited. We have analyzed the BTCEUR price times series using recurrent neural network in Elman topology in a previous work [4], finding out that that on the daily scale the MSE of the predicted data (5-day ahead prediction) scales as the Realized Volatility computed from intraday trading data.

Support Vector Machine

When predicting the logarithmic returns on equidistant time grid, the MSE is typically used as the accuracy criterion for training neural networks. It is, nevertheless, more intuitive and an easier subject to interpretation to focus just on the hit ratios, i.e. the trend data (class 1 for logarithmic return $R > 0$, class -1 for $R < 0$; no-tick data excluded). Then we have a binary classifier problem, for which, according to the standard findings in the literature [5], the Support Vector Machine method is considered to be superior. As quoted in [5], “The underlying motivation for using SVMs is the ability of this methodology to accurately forecast time series data when the underlying system processes are typically nonlinear, non-stationary and not defined a-priori.”

The present dataset of hourly data for BTCCNY prices is broken to 12 monthly segments, and for each of them, 2/3 of the data is used for SVM training, and the remaining part for hit ratio validation. We adopt the R-package kernlab [6] implementing the original method of Vapnik [7]. The relative constant for penalizing mis-predictions is set as $C=10$. Figure 1 depicts the BTCCNY time series in 2015 and shows the clustering behavior of the logarithmic return on the 1 hour scale.



Results and Discussions

We have experimented with various polynomial kernels of orders 1, 2, 3, and 4, tuning the scale and offset parameters, and the Radial Basis function kernel, with manually tried and automatically determined Gaussian function width. The highest validation scores were obtained for the automatically set RBF kernel. These results are summarized in Table 1 for the input space dimension (size of the predictor sampling window) $N=5$.

Table 1 SVM-prediction hit ratio for BTCCNY 2015

JAN	FEB	MAR	APR	MAY	JUN
52.08	52.53	50.85	51.80	53.11	50.86
JUL	AUG	SEP	OCT	NOV	DEC
51.45	50.22	56.87	56.88	53.04	58.75

As the above table indicates, the SVM-predicted results are rather in accord with the efficient market hypothesis, suggesting a low level of causal content in the time series data. We have been currently implementing the LSTM deep-learning networks in an attempt to confirm whether the data in Table 1 are an upper bound or not. At the conference, we will also briefly review our previous work on BTC time series analysis [4].

References

- [1] Bariviera, A. F., Basgall, M. J., Hasperué, W., and Naiouf, M.: Some stylized facts of the Bitcoin market, *Physica A: Statistical Mechanics and its Applications* Vol. 484, pp. 82-90, (2017)
- [2] Blau, B. M.: Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance* Vol. 41, pp. 493-499, (2017)
- [3] Katsiampa, P.: Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, Vol. 158, pp. 3-6, (2017)
- [4] Kodama, O., Pichl, L., Kaizoji, T.: Regime change and trend prediction for Bitcoin times series data, in *Proceedings of CBU International Conference on Innovations in Science and Engineering*, Prague, 2017, DOI: <http://dx.doi.org/10.12955/cbup.v5.954>
- [5] Sapankevych, N. I., Sankar, R.: Time Series Prediction Using Support Vector Machines: A Survey, *IEEE Computational Intelligence Magazine*, Vol. 4, No. 2, (2009)
- [6] Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A., kernlab -- An {S4} Package for Kernel Methods in {R}, *Journal of Statistical Software*, Vol. 11, No. 9, pp. 1-20, (2004).
- [7] Cortes, C., Vapnik, V.: Support Vector Networks, *Machine Learning* Vol. 20, pp. 273-297, (1995)

機械学習を用いた共和分ペア・トレード戦略

A machine-learning approach to pairs trading strategy

今村 光良^{1,3*} 中川 慧^{1,2} 吉田 健一²

Mitusyoshi Imamura^{1,3}, Kei Nakagawa^{1,2}, Kenichi Yoshida²

¹ 日興グローバルラップ株式会社

¹ Nikko Global Wrap 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 Business Sciences

Abstract: 機械学習を用いた株価予測については近年多く研究されている。しかしながら、その多くが株価ないし株価指数そのものをそのまま予測対象としている。一般に株価は非常に複雑な振る舞いを示し、予測が難しい。一方で、同業種同規模などの似通った銘柄間では価格差（スプレッド）が平均回帰すること（所謂共和分性）が知られている。そこで本研究では、株価そのものではなく、共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を提案する。具体的には、LSTMを用いて、定常性を満足するモデルで最も代表的なAR(1)過程に従う人工的な時系列データを事前学習する。学習したLSTMを用いてペア・トレード戦略に適用した結果、単純なAR(1)過程や実データを直接学習させたLSTMよりも良好な結果が得られた。

1 はじめに

機械学習を用いた株価予測については近年多く研究されている。しかしながら、その多くが株価ないし株価指数そのものをそのまま予測対象としている。一般に株価は非常に複雑な振る舞いを示し、予測が難しい。一方で、同業種同規模などの似通った銘柄間では価格差（スプレッド）が平均回帰することが知られている。計量経済学の文脈ではこれを共和分性（co-integration）として様々な研究が行われている [2]。

ペア・トレード戦略はこのような価格変動が似通った銘柄を見つけ、当該ペアの価格差が均衡水準の周りを推移すると仮定する。共和分性を満たすペアのスプレッドは定常過程となるため、平均が時点に依らず一定、すなわちある均衡水準への平均回帰性を持つ。そして、スプレッドが均衡水準から乖離したとき、将来その乖離が修正されるだろうという平均回帰に賭けて、相対的に割高な方を売り、割安な方を買うことで収益獲得を狙う戦略である。

共和分性を利用したペアトレードの実証研究として、[3]の研究がある。彼らは1962年から2002年までの米国株式市場において共和分関係にあるペアに注目し、

ボリンジャーバンドを用いた実証分析を行った。具体的には、ペアのスプレッドが均衡水準から ± 2 標準偏差以上乖離したときをシグナルとしてポジションを構築し、均衡水準に平均回帰したときにポジションを解消するという方法を用いた。このような共和分ペアに対してボリンジャーバンドを用いてポジションの構築を行う手法は、ペアトレード戦略の有効性を検証する上で実務及び実証分析のスタンダードとなっている。また、先行研究 [7] では、スプレッドの均衡水準からの乖離が小幅である場合と大幅である場合での回帰スピードの非線形性を指摘し、スプレッドがTARモデルに基づく非線形共和分関係 [1] を満たすと仮定した。そして回帰スピードが変化する閾値をトレードのシグナルとして使用し、その有意性を確認した。上述の通り、共和分性を満たすスプレッドは定常過程となり、株価そのものよりも予測しやすいことが想定される。

そこで本研究でも、株価そのものではなく、共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を検討する。従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事をLSTMを使ってモデル化する事が提案手法の特徴である。

*連絡先：日興グローバルラップ株式会社
〒103-0016 東京都中央区日本橋小網町 9-2
E-mail: ic140tg528@gmail.com

2 提案手法

本研究では、共和分性を満たす株価ペアは定常性を満たすという特性を利用したペア・トレード戦略のためのスプレッド予測方法を提案する。まず共和分性を満たす株価ペアを見つけ、スプレッドを作成する。次に作成したスプレッドに適合すると推定される共和分の性質を決めるパラメータを使い AR(1) 過程に従う人工データを生成し、LSTM を用いて学習を行う。

Step 1:

共和分性を満たす株価ペアからスプレッドを作成し、AR(1) 過程を特徴づけるパラメータを推定する。

Step 2:

推定したパラメータを使い AR(1) 過程に従う人工データを生成し、LSTM で学習する。

Step 3:

学習した LSTM で実際のスプレッドの予測を行う。

実際の株価データから作成したスプレッドを直接学習するのではなく、一旦共和分の性質を決めるパラメータ (詳細は後述する) を推定し、推定したパラメータで生成した人工データを LSTM で学習させる事が提案手法の特徴である。

以下、提案手法の設計意図について述べる。

2.1 共和分性

共和分性は非定常な時系列データの線形結合が定常過程となる時系列的性質であり、[2] によって提唱され、長期的な均衡関係を記述するものとして経済、ファイナンスの様々な実証分析において利用されてきた。定常過程は、時系列の平均が時間に依らず一定であるため、平均回帰という扱いやすい性質を持っている¹。

一方で通常、株価はランダムウォークであると言われている。ランダムウォークとは非定常な時系列の代表例であり、単位根過程とも言われる。一般にランダムウォークのような非定常な時系列を線形に組み合わせても、同様に非定常である。しかし、組み合わせ方をうまく選べば、定常過程となる場合がある。このとき共和分の関係があるという。具体的にはランダムウォークする 2 つの株価のペア $\{X_t, Y_t\}$ に対して、ある定数 β が存在し、以下の (見せかけの) 回帰式 $Y_t = \beta X_t + \epsilon_t$ における誤差項 $\{\epsilon_t\}$ が定常となることが共和分性の満たす条件である。

¹正確には過程の期待値と自己共分散が時間を通じて一定である確率過程を指す。

2.2 LSTM

LSTM は、データ間の依存関係を学習できるニューラル・ネットワークであるリカレント・ニューラル・ネットワーク (RNN) の一つである [4]。RNN は、AR 過程と同様に前の出力を次の入力に追加するモデルであり、時間方向に展開すると静的なニューラル・ネットワークと見ることができる。RNN は、時間方向に展開したネットワーク上で誤差逆伝播を用いて学習を行うが、系列が長くなると、勾配が消失してしまう。したがって、長期依存を学習できないという問題が生じる。これに対し、LSTM では、重みを掛けずに誤差を逆伝播させることによって、長期依存を学習できなくなる問題を解消している。

ここで、本研究では従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事を LSTM を使ってモデル化する。単純な AR 過程のモデル化では難しかったパラメータの時間変化を扱う事が LSTM を用いる事の意図である。

なお、本研究では、オープンソースの深層学習フレームワークである Chainer[6] のライブラリにて提供されている LSTM によりこれを実装した。また、ネットワークの構成は、同ライブラリにて提供されている LSTM のサンプルコード² と同様の構成とした。ただし、入力層は全結合層に変更し、ユニット数を 1、LSTM 層についてはユニット数を 200、出力層はユニット数を 1 と変更している。学習時のパラメータ最適化には Adam [5] を用いた。

2.3 AR 過程の学習

定常性を満たすモデルで最も代表的な AR(1) 過程 $S_t = \alpha S_{t-1} + \epsilon_t$ ここで ϵ_t は正規分布 $N(0, \sigma)$ に従う誤差項、 $S_t = Y_t - \beta X_t$ を LSTM に学習させる事を考える。

ここで、[7] でも指摘されている通り、株価ペアが共和分性を満たしても、スプレッドの定常性を表現するパラメータ α が時期によって変化することには注意を要する。従来研究では α を時期によらず一定として扱い、予測精度の低下をもたらしていたと考えられる。

そこで、実際のスプレッドのデータから α と σ を推定する。TOPIX 指数および日経平均株価のペアに対して、1985 年 1 月から 2006 年 12 月末までの 20 日間における AR(1) の α の推移は図 1 に示す通りである。

当該期間における α の中央値は 0.7552 であることから、 α が変化し、 $\{0.55, 0.65, 0.75, 0.85, 0.95\}$ の 5 つの値を取っていると推定した。またこの時の σ についても同一期間の実データからフィッティングし推定した。次に、1 バッチあたり各 α を持つ AR(1) に従う時系列

²<https://github.com/chainer/chainer/blob/master/examples/>

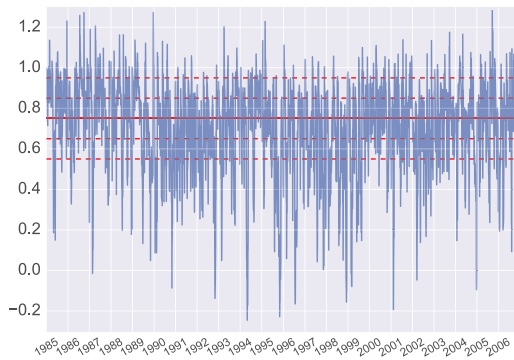


図 1: 1985 年 1 月から 2006 年 12 月末までの 20 日間における AR(1) の α の推移

長 20 のサンプルを 50 生成し，入力値として，0 から 1 の範囲に基準化し，バッチサイズ 250 として，損失関数の値が十分に低下する 600 回まで学習させた。

α については次章で報告する株価のペア以外にも適用できる一般的な値と考えており，提案手法は実質 σ をデータから推定するだけで他の対象にも応用できる，即ち分散を計測するのに十分な短期間のデータがあれば適用できる，と考えているが，この点については他の株価データにより今後確認していく必要がある。

3 実証分析

提案手法の有効性を評価するために，TOPIX 指数と日経平均株価のペアを対象に分析を行った。これらは統計的に共和分性が担保されているだけでなく，経験的にも当該ペアは共和分性を満たしていると考えられている。データは Bloomberg から取得した。検証期間は 2007 年 1 月から 2017 年 7 月末までとし，20 営業日後のスプレッドを予測対象とした。スプレッドを作成する際に使用するパラメータ Y は日経平均株価， X は TOPIX 指数とし，定数 β については，年始最初の営業日のタイミングで推定を行い，毎年更新する。また，AR 過程学習時に使用する σ は 1985 年 1 月から 2006 年 12 月末までの値を用いた。なお，簡単のため収益率の計算はスプレッドの予測が上昇なら式 (1) 下落なら式 (2) を 20 で割った値とした。

$$(S_{t+20} - S_t)/(Y + \text{abs}(\beta X)) \quad (1)$$

$$-(S_{t+20} - S_t)/(Y + \text{abs}(\beta X)) \quad (2)$$

また，比較対象として，実際のスプレッドを学習した LSTM(raw) を加える。学習に用いたスプレッドの期間は 1985 年 1 月から 2006 年 12 月末までであり，当該期間における時系列長 20 のサンプルを 250 生成し，

表 1: 各モデルにおける正答率および収益率

手法	正答率	収益率
AR(1)	56.01 %	13.45 %
LSTM(raw)	56.59 %	14.06 %
LSTM(AR)	58.33 %	17.26 %

入力値として，0 から 1 の範囲に基準化し，AR 過程を学習させた LSTM と学習パラメータを一致させ，バッチサイズ 250 として，損失関数の値が十分に低下する 600 回まで学習させた。

前述の AR 過程を学習した LSTM を用いる提案手法の結果 LSTM(AR) と，標準的な AR の予測結果 AR(1)，および，実際のスプレッドで学習した LSTM で予測した結果 LSTM(raw) の比較を表 1 に示す。

また各収益率の推移を図 2 に，F 値を表 2 に示す。表 2 においては，価格が上昇した場合 1，下落した場合 0，平均 ave 毎に結果をわけて示す。

正答率・収益率ともに，LSTM(raw) は標準的な方法である AR(1) の結果を再現できており，LSTM がスプレッドの時系列をの学習する能力を持つ事を示している。また提案手法による結果 LSTM(AR) は正答率・収益率ともに AR(1) および LSTM(raw) を上回っている。特に収益率の変化 (図 2) は LSTM(raw) を安定して上回っている。

実際のスプレッドを直接学習した LSTM(raw) より一旦パラメータ α, σ を取り出した結果を使って生成した人工データを学習した LSTM(AR) が精度良く予想できる結果は興味深い。AR(1) を用いた共和分を満たす時系列データ分析の妥当性，言い替えると共和分に関連する既存研究が妥当であった事を示唆していると考ええる。

4 まとめ

本研究では，株価そのものではなく，共和分性を満たす株価ペアのスプレッドを機械学習で予測する手法を提案した。具体的には，LSTM を用いて，定常性を満足するモデルで最も代表的な AR(1) 過程に従う人工的な時系列データを事前学習する。従来研究で一定とされていた共和分の性質を決めるパラメータが時間により変化する事を LSTM を使ってモデル化する事が提案手法の特徴である。

学習した LSTM を用いてペア・トレード戦略に適用した結果，単純な AR(1) 過程や実データを直接学習させた LSTM よりも良好な結果が得られた。

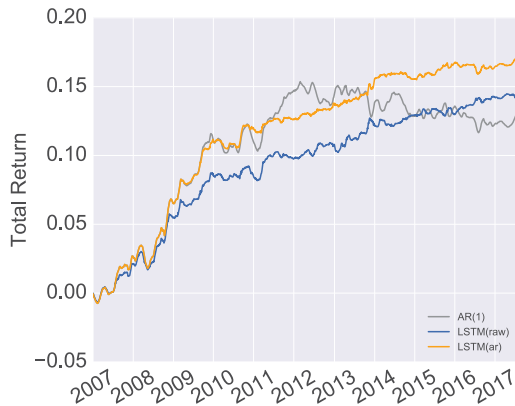


図 2: 各モデルの収益率

表 2: 各モデルの F 値

method		precision	recall	f1-score
AR(1)	0	50 %	63%	56 %
	1	63 %	51%	56 %
	ave	57 %	56%	56 %
LSTM(raw)	0	52 %	25%	34 %
	1	58 %	81%	68 %
	ave	55 %	57%	53 %
LSTM(AR)	0	53 %	46%	50 %
	1	62 %	68%	65 %
	avg	58 %	58%	58 %

参考文献

- [1] Walter Enders and Pierre L Siklos. Cointegration and threshold adjustment. *Journal of Business & Economic Statistics*, Vol. 19, No. 2, pp. 166–176, 2001.
- [2] Robert F Engle and Clive WJ Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp. 251–276, 1987.
- [3] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, Vol. 19, No. 3, pp. 797–827, 2006.
- [4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.
- [5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The International Conference on Learning Representations (ICLR)*, 2015.
- [6] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, Vol. 5, 2015.
- [7] 中川慧. 非線形共和分関係に基づくペアトレード戦略. *テクニカルアナリストジャーナル*, Vol. 3, pp. 1–8, 2016.

新たなポートフォリオ再構成アルゴリズムに向けての 株式市場におけるモーメンタム効果のモデル化

Modeling the Momentum Effect in Financial Stock Markets

Toward New Portfolio Rebalance Algorithm

海野 一則¹ 菊地 剛正² 國上 真章¹ 山田 隆志³ 寺野 隆雄¹
Kazunori Umino¹, Takamasa Kikuchi², Masaaki Kunigami¹, Takashi Yamada³, Takao Terano¹

¹ 東京工業大学

¹ Tokyo Institute of Technology

² 慶應義塾大学

² Keio University

³ 山口大学

² Yamaguchi University

Abstract: This research has two objectives: (1) to model and analyze the momentum effect, (2) to propose a portfolio reconstruction algorithm that can use the momentum effect to obtain excess profit. The momentum effect tends to be present in the stock market, and describes the phenomenon whereby rising (declining) stocks tend to continue to rise (decline). However, because existing research does not separate momentum effects from stock price fluctuations it is not always possible to obtain excess return when working with an unknown data set that contains a momentum effect. In this research, we define a new External Force Momentum Effect (EFME) model based on bias in stock price rises (declines). We prepared an artificial data set that contained this momentum effect and constructed a portfolio with the proposed algorithm. The relationship between the EFME model and excess return is then analyzed to verify that excess profit can be obtained. Additionally, we confirmed that the proposed method can obtain higher excess return than the existing method when applied to artificial and real stock data sets.

Keywords: momentum effect, anomalies, trading algorithm, on-line portfolio selection strategy

1. はじめに

本論文の目的は以下の2点である: (1) 株式市場におけるモーメンタム効果をモデル化し, その効果により超過収益が得られることを示す; (2) ポートフォリオ再構成アルゴリズム(以後, PRA と略す)を提案し, モーメンタム効果により超過収益を得られることを示す。

既存研究において, 株式市場において”モーメンタム効果”により超過収益が得られることが報告されている[1,2]. ”モーメンタム効果”とは, 過去に上昇(下降)傾向にある株式は, 将来もその傾向が続く現象である。上記の研究は, 株式市場にモーメンタム効果が

存在していることを示しており, 過去6ヶ月間に高い(低い)リターンを示した株式は, その後の1~6ヶ月に高い(低い)リターンを示すと報告している。さらに, それらの結果を追認する研究結果も報告されている[3,4,5].

しかしながら, それらの研究に異議を唱える研究結果も報告されている。 ”モーメンタム効果”により超過収益が得られるという検証結果は, データの最適化によるものであり, 実際に超過リターンが得られるわけでは無いという主張である。つまり, 様々な条件から有効性を示す組み合わせを発見しただけで, 流動性を考慮した現実的な条件下で検証を行ったケースでは, 統計的に有意なリターンは認められなかったと述べている[6].

このような論争が生じる原因は, モーメンタム効果

を明確に定義・モデル化できておらず、実際の株価データから”モメンタム効果”のみを取り出して検証することが困難だからである。本研究では、モメンタム効果をモデル化した人工データセットを作成し、そのデータセットを用いて再現性のある形で投資手法の有効性を分析する枠組みを提案する。

モメンタムは、式(1)により定義される。モメンタムは一定期間のリターンであることが示されており、”モメンタム効果”とは、株式のリターンが高い(低い)株式は、将来もその傾向が続く現象であるとされている[1]。 P_t は時点 t における株価を示し、 $R_t^{MOM_period}$ は期間 MOM_period における時点 t のモメンタムを示している。もし、MOM_period が 1 の時はリターンに等しくなる。

$$R_t^{MOM_period} = \frac{P_t}{P_{t-MOM_period}}, \quad (1)$$

本研究では、”モメンタム効果”をモデル化し、モメンタム効果を持つ複数の株式で構成された人工データセットを生成した。External Force Momentum Effect (以後、EFME と略す) モデルは、外的要因によってバイアスが発生し、その影響により株価変動においてモメンタム効果が発生すると仮定したモデルである。

本稿は、以下の構成を取る。2章でEFMEモデルを説明するとともに、モメンタム・バイアスの生成手順について説明し、3章では本研究で使用した検出方法と、ポートフォリオ再構成手法への応用について述べ、4章ではデータセットと検証方法について説明する。5章ではEFMEモデルにおけるモメンタム効果を分析するとともに、本研究で提案するポートフォリオ再構築手法(PRA)を用いた検出結果を分析し、6章でまとめる。

2. モメンタム効果のモデリング

2.1 External Force Momentum Effect モデル

本研究では、幾何ブラウン運動(Geometry Brownian Motion: GBM)を株価変動の基本モデルとし[11,12]、価格の変動特性は外的要因によるバイアスで決定されると考える。本研究で用いる EFME モデルを以下の手順により生成する。

- (1) GBMに基づき生成したデータセット(複数の株式データ)に対して、各株式に複数個の上昇・下落バイアスをランダムかつ等しい確率で発生させる(図1)。
- (2) バイアスの開始日および継続期間は一様乱数により決定され、各バイアスの情報は、タイプ(上

昇/下落)および、開始日、終了日、発生確率、強さで決定される(図1)。

- (3) バイアスが価格変動に与える影響は、(2)より各株式において付加されたバイアスに従い、設定期間において、”発生確率”に応じて上昇(下落)バイアスは、上昇(下落)日からランダムに選択し、選択した日に対して、”強さ”に応じてリターンを拡大(縮小)することで価格変動に特性を与える(図2)。

本モデルは、同一の株式に複数のバイアスが影響を与え、それらの期間が重なることを認めたモデルである(図1,2)。GBMデータセットに対して複数のバイアスによる効果が反映されたものがモメンタム効果を持つデータセットとなる(図3)。

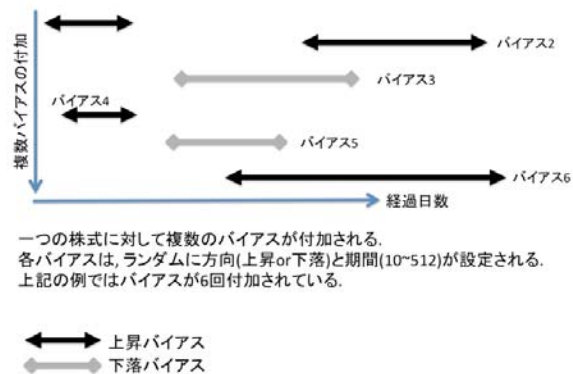


図1 EFMEモデルによる1株式に対するバイアスの付加:

上昇・下落バイアスはGBMデータに対して付加される。複数の上昇・下落バイアスがランダムに開始時期と継続期間が与えられる。この事例では、6回バイアスが付加されており、その順序に従い該当日のリターンを拡大(縮小)する。

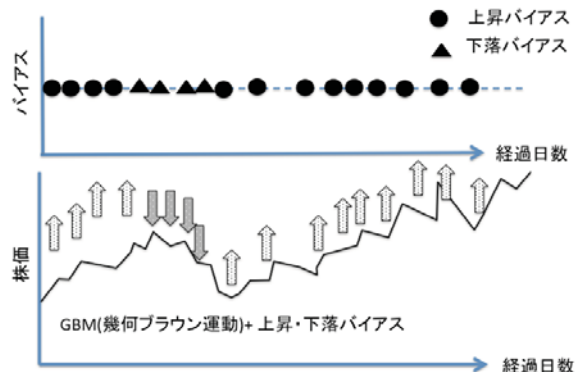


図2 EFMEモデルによる付加されたバイアスと価格変動の概念図 (バイアスの重ならないケース)

上図: 上昇・下落バイアスを示している。

下図: バイアスの影響を受けたGBMデータの価格変動

FEMEにより上昇・下落バイアスが付加され、その影響により価格変動が起こる。上昇バイアス(●)および下落バイアス(▲)の影響により、下図の価格変動が起こる。

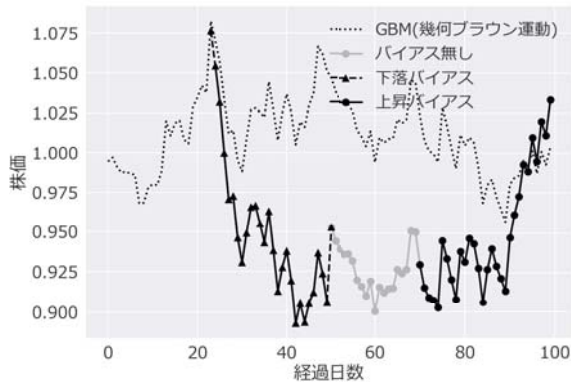


図3 上昇・下落バイアスによりGBMにより生成されたデータが影響を受け異なる軌跡となる

EFME(External Force Momentum Effect)モデルにより、GBMにより生成されたデータが異なる変動特性(モメンタム効果)を示す。上昇(下落)バイアスは、選択された期間の中で価格の上昇(下落)日の中から、指定された発生比率(prob)に基づきランダムに作用し、リターンを一定比率(strength)で拡大(縮小)させる(2~7)(Algorithm1)。

本モデルでは GBM データセットに対して複数のバイアスによる効果が反映されたものがモメンタム効果を持つデータセットとなる(図1)。外部バイアス(External Force Bias: EB) は、上昇バイアスおよび下落バイアス、バイアス無しの3種類のタイプで表現される。データセットの株式 $i = [1, \dots, M]$, 時点 $t = [1, \dots, N]$ とし、バイアスが加えられた数を $k = [1, 2, \dots, J]$ とする時に、GBM と EB、資産価格 S は以下の式で表される(2,3,4,5,6,7)。GBM_{*i*} は、株式 *i* の日次のリターンベクトルを示しており、EB_{*i*} は、株式 *i* のバイアスのリストを示しており、図1の例では *N* 個のリストで構成され、リストの各要素(例えば、EB_{*i,1*})は6個のバイアスを持つことになる。

$R_i^{GBM,1}$ は、GBM により生成された株式 *i* の1日のリターンを表す。また、式(2,3,4,5,6,7)の EB および Bias 関数によりバイアスが価格変動に特性を与える。

$$GBM_i = [R_{i,1}^{GBM,1}, R_{i,2}^{GBM,1}, \dots, R_{i,N}^{GBM,1}] \quad (2)$$

$$EB_i = [EB_{i,1}, EB_{i,2}, \dots, EB_{i,N}] \quad (3)$$

$$EB_{i,t} = [bias_1, bias_2, \dots, bias_J] \quad (4)$$

$$bias_k = [t1, t2, bias_type] \quad (5)$$

$$bias_type \in \{non_bias, rising_bias, decline_bias\}$$

$$S_i^{EFME} = f_{EFME}(GBM_i, EB_i), \quad (6)$$

=

$$cumprod\left(\sum_{k=1}^J Bias\left(R_i^{GBM,1}, EB_i[k], prob, strength\right)\right),$$

(7)

※ *J*: バイアスの数

prob: バイアスの発生確率

strength: バイアスがリターンに与える効果の大きさ

cumprod: 累積積 cumulative product

2.2 GBMにバイアスを付加するアルゴリズム

本提案手法では、幾何ブラウン運動(Geometry Brownian Motion: GBM)に外因によるバイアス(EB)をデータセットの全株式に付加する手順をAlgorithm1 に示す。

本研究では、2種類のデータセット GBM-EB-Small およびGBM-EB-Big が用いられている。GBM-EB-Big では、prob=0.2 と strength=1.017 が設定されており、これは、ランダムに選択された期間 $t1 \sim t2$ において上昇(下落)バイアスであれば、選択された期間の上昇(下落)日のリターンが 1.017 (1/1.017)倍になる。

今回用いるデータセットは、GBM および Algorithm1 に基づきデータを生成し、200組のデータセットを作成した。各データセットの大きさは、オリジナルの株価データセットに合わせて、株式数を391とし、データ長を4362とした。

Algorithm1 AddExternalBiasToGBM

$non_bias, rising_bias, decline_bias := 0, 1, 2$

// GBM-EB-Big 設定の場合

$prob, strength := 0.2, 1.07$

// 生成された GBM データ

$GBM_i = [R_{i,1}^{GBM,1}, R_{i,2}^{GBM,1}, \dots, R_{i,N}^{GBM,1}]$

// EB データの生成と初期化

$EB[M, J] := initialize([0, 0, non_bias])$

for $i=1, \dots, M$ do

for $k=1, \dots, J$ do

$t1 := UniformRandomNumber(1, N)$

$period := UniformRandomNumber(10, 512)$

$t2 := t1 + period$

if $t2 > N$ then $t2 := N$

$bias_type :=$


```
RandomSelect({rising_bias,decline_bias})
EB[j,m]:= (t1,t2, bias_type)
```

```
// adding bias to GBM
```

```
for i=1,...,M do
  for k=1,...,J do
    t1,t2, bias_type:=EB[i,m]
    Bias(GBMi, t1,t2, bias_type, prob,strength)
```

3. 株式のモメンタム評価スコアに基づく ポートフォリオ再構成アルゴリズム

本節では、モメンタム評価スコア(以後、MES と略す) を説明するとともに、ポートフォリオ再構成アルゴリズム(PRA)に応用した MES-PRA の説明を行う。

MES は、今までの研究とは異なる視点による評価スコアである。MES は複数の期間の新高値(新安値)を記録した回数を基にした定性的な評価指標である。本論文では、ポートフォリオの構成株式数 U を与えることで、MES の上位 $1 \sim U$ の株式を保持する PRA が MES-PRA である。本章では、MES の説明を行うとともに、取引コストを含めたパフォーマンスを評価する。

3.1 モメンタム評価スコア (MES) の概要

MES は、モメンタム効果の高い株式を選択するための指標であり、新高値(新安値)の発生回数という定性的な評価に基づくものである。本論文では 6 種類の定量的評価指標を用いているが(3.4 節)、それらの指標を用いてモメンタム効果の高い株式を選択することは難しいため、新たな定性的評価指標である MES を提案する。

MES の考え方とアルゴリズムを以下に述べる。図 3 に示したように、上昇・下落バイアスにより価格変動の特性が異なる。そこで、変動の軌跡の違いが新高値や新安値の発生回数が異なることに注目した評価手法を図 4 に示す。

n 日間に発生する新高値や新安値の発生回数が、株式の上昇(下落)トレンドに応じて、スコアが高(低)くなっている。上昇(下落)トレンドを外部からのバイアスによる現象と考えれば、バイアスの大きさを間接的に計測することになり、時点 t において複数の評価期間のスコアの平均値を求めたものが MES、であり、これが評価値となる。

さらに、ノイズの影響を小さくするために、過去 L_{MES} 日間の指数移動平均により平滑化を行う。

高いMESを選択することは、強い上昇バイアスの影響を受けていることを示している。これは、高い上昇率を持つ株式を選択しているのではなく、強い

上昇バイアス(新高値を記録する回数が多く、新安値を記録する回数の少ない)を持つ株式を選択している。

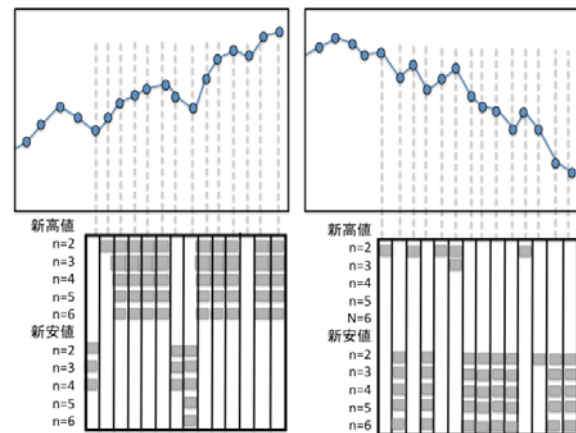


図4 MESの計算方法の実例

n 日間に発生する新高値(新安値)の発生頻度をカウントすることでスコアを計算する。この図では、 $n=2,3,\dots,6$ における新高値(新安値)の例を示している。灰色の部分には新高値または新安値を記録したことを示している。

新高値は1.0を与え、新安値は0.0を、それ以外(白い部分)は0.5を与え、その合計値をスコアとする。

左の図:1日目スコア= $1.0 \times 0 + 0.0 \times 3 + 0.5 \times 7 = 3.5$

2日目スコア= $1.0 \times 1 + 0.0 \times 0 + 0.5 \times 9 = 5.5$

3日目スコア= $1.0 \times 5 + 0.0 \times 0 + 0.5 \times 5 = 7.5$

右の図:1日目スコア= $1.0 \times 1 + 0.0 \times 0 + 0.5 \times 9 = 5.5$

2日目スコア= $1.0 \times 0 + 0.0 \times 5 + 0.5 \times 5 = 2.5$

3日目スコア= $1.0 \times 1 + 0.0 \times 0 + 0.5 \times 9 = 5.5$

これらの値は1日ごとの計算値であり、これを(10)に従い指数移動平均を求めたものが各日のMESとなる。

3.2 モメンタム評価スコア (MES) の計算方法

本節では、MES を実際に計算するアルゴリズムの詳細を示す。実際のスコアは、 n 日間で $n=[10,20,\dots,250]$ の各ケースについてスコアを求め(8,9),それを過去 L_{MES} 日の指数移動平均値を MES とする(10).また、MES は全て株価データのみから得られるものであり、ファンダメンタルな指標は用いない。

$$MES_t^{Day} = \sum_{i=1}^{25} f_{MES}(t, price_vec, i * 10), \quad (8)$$

$$f_{MES}(t, v, n) =$$

$$\begin{cases} 1 & \text{if } v[t] > \max(v[t-n+1:t-1]) \\ 0 & \text{if } v[t] < \min(v[t-n+1:t-1]), \\ 0.5 & \text{otherwise} \end{cases}, \quad (9)$$

$$MES_t = EMA(MES_{t-L_{MES}+1}^{Day}, \dots, MES_t^{Day}), \quad (10)$$

* 1. The exponential moving average is

$$L_t = \alpha Y_t + (1 - \alpha)L_{t-1}$$

$$\alpha = (1 - W), \quad 0 <= W < 1,$$

52 週の高値(1 年間の最高値)を記録していることが有益な情報であると報告されており [10], 本手法は, 複数の期間における高値(安値)情報を取り込むことで株式に働いているバイアスの検出精度を上げている.

3. 3 MESを用いたポートフォリオ再構成 アルゴリズム (MES-PRA)

MES-PRA は, 構成株式数 U が与えられた時, MES の高い株式を降順に $1 \sim U$ を等比率で選択することで, ポートフォリオ構成するアルゴリズム (Algorithm2) を以下に示す.

Algorithm2 MES-PRA (MES, S, M, N, u, PR)

Input: MES score array MES ; // $MES_{t,i}$: stock i on day t
 Stock set S ;
 Number of stocks M ;
 Number of days from the start N ;
 Number of portfolio component stocks u ;
 Portfolio reconstruction system PR ;

Output: portfolio P_{all} ;

$t := 0$; $M = (MES_{k,1}, \dots, MES_{k,N})$

Let s_t be all the stocks selected

By constructing a portfolio with s_t stocks selected at the same ratio by PR , we make P_t .

let $P_t = PR(s_t)$ # Construct a uniform weight portfolio

while $t > N$ do

$t := t + 1$;

$M = (MES_{t,1}, \dots, MES_{t,N})$

// M より, 上位 $1 \sim u$ までの株式を選択し,

// その株式セットを s_k とする.

$s_k = \text{get_MES_rank_topK}(S, M, u)$

let $P_t = PR(s_t)$

$P_{all} = (P_1, \dots, P_N)$

return P_{all}

3. 4 パフォーマンス評価指標

本節では, この研究で用いられているリターンおよびリスク, リスク調整済みリターンについて述べる. 本研究で用いる指標は, パフォーマンスの指標

として最終資産倍率および年平均リターン(%)を用い, リスクとして年平均ボラティリティ(%)および最大ドローダウン(%)を利用する. さらに, リスク調整済みリターンとして, シャープレシオ [7] とカルマーレシオ [8] を用いる.

- (1) 最終資産倍率: Final Asset Multiplier (AM)
- (2) 年平均リターン: Annual average Return (AR)
- (3) 年平均ボラティリティ: Annual Volatility (AV)
- (4) 最大ドローダウン: Maximum Draw Down (MDD)
- (5) シャープレシオ: Sharpe Ratio (SR)
- (6) カルマーレシオ: Calmar Ratio (CR)

3. 5 取引コスト (資産の入替時に発生するコスト)

取引コストは, 取引手数料とマーケットインパクトを合算したコストとし, 米国市場において発生する取引コストは 0.08~0.14% と報告されており [9], 本研究では 0.14% を用いることとした. なお, 取引コストは, 売りまたは買いの片道のコストであり, 売買する金額に対して 0.14% が一回の取引で減少する.

4. データセットと評価方法

4. 1 株価データと人工データ

本研究では, 実際の株価データセットと, 同じデータサイズ(銘柄数と日数)を持つ 4 種類の人工データセットを用いる(表 1). 人工データセットは, 幾何ブラウン運動のみによるデータセットである GBM および, GBM に上昇バイアスを 1 回のみ付加した(バイアスの重なりのない)GBM-EB-UP, 複数の上昇・下落のバイアスを付加した GBM-EB-Small, より大きなバイアスを付加した GBM-EB-Big を用意した.

表 1 人工データセットと実際の株価データセット

名前	データセットの内容	
	基本情報 (日数および構成 株式数 は全て同じ. また人 工データセットは 200 サンプ ルで検証) * 1	バイアスの パラメータ
GBM	4362 日, 391 株式 サンプル数: 200 0.00%/年 (AR:平均) -3.0%/年 (AR:中央値)	無し
GBM-EB- UP	4362 日, 391 株式 サンプル数: 200 8.95%/年 (AR:平均) 4.22%/年 (AR:中央値)	prob=0.10 effect=1.010 付加バイアス数(J) = 1
GBM-EB-	4362 日, 391 株式	prob=0.15

名前	データセットの内容	
	基本情報 (日数および構成 株式数は全て同じ。また人 工データセットは200サンプ ルで検証) *1	バイアスの パラメータ
Small	サンプル数: 200 4.81%/年 (AR:平均) -3.0%/年 (AR:中央値)	effect=1.013 付加バイアス数(J) = 20
GBM-EB- Big	4362 日, 391 株式 サンプル数: 200 15.6%/年 (AR:平均)- -3.0%/年 (AR:中央値)	prob=0.20 effect=1.017 付加バイアス数(J) = 20
名前	基本情報 *2	米国 S&P500 インデ ックスのリターン [備考]
SP	米国 S&P500 に属する株式 1999.01.01~2016.4.29 4362 日, 391 株式 15.7%/年 (AR:平均) 9.22%/年 (AR:中央値) ※ 2016.4.29 時点で S&P500 に属しており, 1999.01 より 上場している全株式を対象 としたため, 生存者バイアス が存在している.	7.60%/年(S&P500 指 数) データセットは, 十 分な流動性を持つ 株式で構成されて いる[6].

*1. ここでの平均と中央値は, 200サンプルの平均値と中央値を示している.

$ave_r = \text{average}(\text{AR}(\text{stock}_1), \dots, \text{AR}(\text{stock}_{391}))$

AR:平均値: $\text{average}(ave_1, \dots, ave_{200})$

AR:中央値: $\text{median}(ave_1, \dots, ave_{200})$

*2. データセットSPにおける平均と中央値は

AR:平均値: $\text{average}(\text{AR}(\text{stock}_1), \dots, \text{AR}(\text{stock}_{391}))$

AR:中央値: $\text{median}(\text{AR}(\text{stock}_1), \dots, \text{AR}(\text{stock}_{391}))$

4.2 比較手法

比較手法1: Random-PRA

MES-PRAの優位性を分析するために, 同じ取引回数と構成株式数であり, 株式をランダム選択するポートフォリオ構築手法であるRandom-PRAを用意し, パフォーマンスの比較対象とした.

比較手法2: WML+

WML+は, WML手法[1,2]のロング・ポジションのみを用いた手法である. WML手法は, ロングおよびショートの方のポジションを持つ手法である. 代表的な戦略であるJegadeesh (6,1)[2]は, 過去6ヶ月の

最もリターンの高い銘柄群(上位10%)のロング・ポジションおよび, リターンの低い銘柄群(下位10%)のショート・ポジションを1ヶ月保持し, これを1ヶ月毎に繰り返す手法である. SPデータセットにおいてWMLのショート・ポジションでは損失が発生しており[2]とは異なる結果となったため, リターンが得られたロング・ポジションのみを評価対象とするWML+を設定し本検証に用いた.

4.2 分析手順

分析は以下3種類の実験結果に対して行われる.

- (1) GBMおよびGBM-EB-UPに, MES-PRAおよびRandom-PRAを適用し, その結果を分析する(実験1).
- (2) GBMおよびGBM-EB-Small, GBM-EB-BigにMES-PRAおよびRandom-PRAを適用し, その結果を分析する(実験2).
- (3) 実際の株価データセットであるSPに対して, ポートフォリオを構成株式数の異なるMES-PRAおよびインデックス, 既存手法を利用したWML+を比較分析する(実験3).

5. 検証結果と考察

5.1 データセットの分析

本節では, 人工データセットと実際の株価データセットとの特徴について分析する. SPデータセットは, 資産倍率は0.06~434倍の範囲となっている. 200セットの全データにおいて, GBMでは0.01~40.0倍であるが, GBM-EB-Smallでは0.0002~2810倍, GBM-EB-Bigでは 2.0×10^{-5} ~58900倍と株式の最終資産倍率の範囲が拡大している. ただし, GBM-EB-BigとSPの年率リターンの平均値は同等である.

5.2 検証結果の分析

5.2.1 人工データによる検証: GBMおよびGBM-EB-UPの分析(実験1)

本項では, 各200セットのGBMとGBM-EB-UPの分析を行った. MES-PRAおよびRandom-PRAは, ポートフォリオを10銘柄の均等比率で構成し, 取引コストを0.14%とした結果である(図5).

本実験により, MESを用いて高いモメンタム効果を持つ株式を選別可能であることが示された. このような予測可能性が認められる原因は, Algorithm1において, ランダムに選択されたバイアスの継続期間が長期に持続ケースが選択されることによるものと考えられる.

また、長期の持続的バイアスが付加されていないGBMでは、MES-PRAを適用しても超過リターンは得られていない。

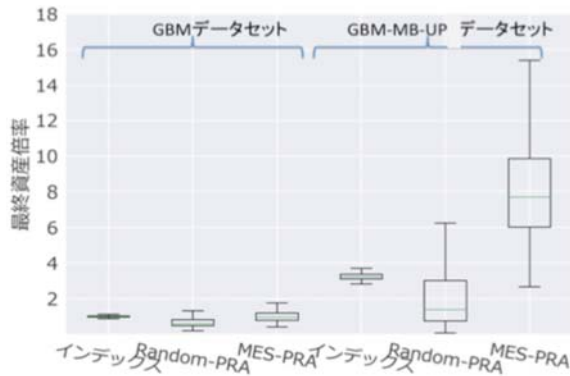


図 5 GBM および上昇バイアスを付加したケースによる検証結果:

各 200 サンプルの GBM と GBM-MB-UP のインデックスおよびMES-PRA, Random-PRA の比較を行なった. GBMでは、資産倍率において全ての結果は、ほぼ同等となった. GBM-MB-UP では、インデックスの資産倍率(AM)の平均は 3.51 倍であり、MES-PRA では 7.91 倍であった. 一方、Random-PRA は取引コストにより 1.34 倍とインデックスより低くなっている. ポートフォリオは 10 株式の等比率であり、片道取引コストは 0.14% で L_{MES} は 450 に設定した結果である.

5. 2. 2 人工データによる検証: GBMおよびGBM-MB-Small, GBM-MB-Bigの分析(実験2)

GBMおよび、GBM-MB-Small, GBM-MB-Bigの3種類のデータセット(各200セット)における、MES-PRAによるシミュレーション結果を図6,7に示す. なお、その際の条件として、10株式の等比率によるポートフォリオであり、片道取引コストは0.14%であり、 L_{MES} は450に設定されている.

MES-PRAでは、GBM-MB-SmallおよびGBM-MB-Bigにおいては超過収益が得られており(図6)、Random-PRAではどのデータセットにおいても超過リターンは得られていない(図7).

データセット(各200個) GBM,GBM-MB-Small, GBM-MB-Bigの資産倍率の平均は、1.00, 2.25, 12.3倍であり、MES-PRAを用いた結果は、1.34, 9.71, 94.2倍である. バイアスが加えられた場合に超過リターンが得られており、バイアスの大きさに応じてリターンが高くなっている. また、Random-PRAを用いた場合は、0.69, 0.74, 0.94倍であった. Random-PRAを用いた結果は、データセットの中央値に近いものとなった.

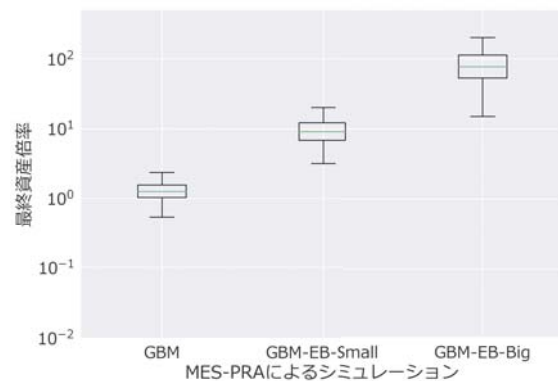


図6. MES-PRAによる人工データセットによるシミュレーション結果(GBMおよび上昇・下落バイアスを付加したデータセット)

GBMデータセットでは超過リターンは得られず、上昇・下落バイアスが付加されたデータセットでは超過リターンが得られており、付加されたバイアスに応じてリターンが高くなっている.

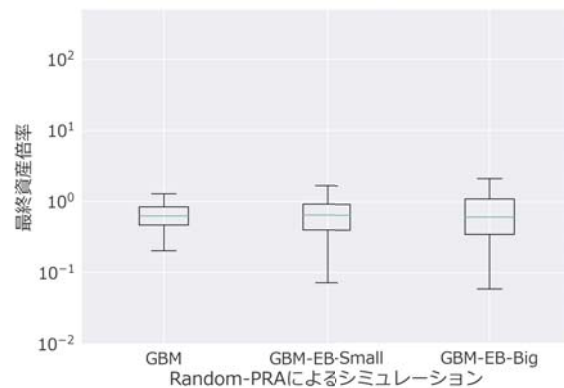


図7. Random-PRAによる人工データセットによるシミュレーション結果(GBMおよび上昇・下落バイアスを付加したデータセット)

全てのデータセットにおいて、Random-PRAでは、超過リターンは得られていない. GBM, GBM-MB-Small, GBM-MB-Bigにおいてデータセット(各200個)の平均資産倍率は1.00,2.25,12.3倍であり、中央値は全て0.69であった. 少数の株式が平均リターンを押し上げているため、ランダムに選択するRandom-PRAでは、中央値に近いリターンとなった.

5. 2. 3 実データでの検証: SPの分析(実験3)

SPデータセットを用いて、MES-PRAおよび、Random-PRA, WML+によるシミュレーション結果($L_{MES}=512$)を分析した(図8.9). MES-PRA(3,10,20銘柄)による均等ポートフォリオがWML+やインデックスより優れたARおよびSRを示した.

ただし、MES-PRAの欠点として、資産価格の一時
的な下落リスク(MDD)が高く、CRにおいては
MES-PRAにおける10銘柄の均等ポートフォリオが
最も優れていた(図8,9).

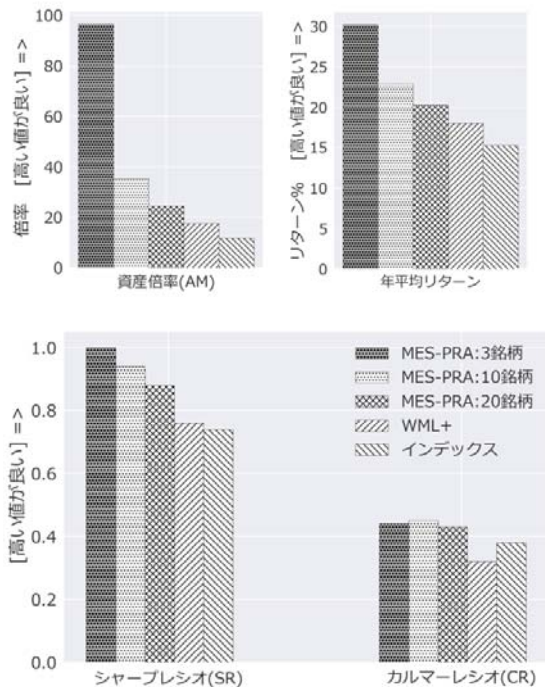


図8 パフォーマンス評価

MES-PRA(3,10,20銘柄の均等ポートフォリオ) および
WML+, インデックスの比較では、MES-PRAが、AMおよ
びAR, SRにおいて最も高いパフォーマンスを示し、構成
銘柄数が少ないケースの方が、高いパフォーマンスを示
した。

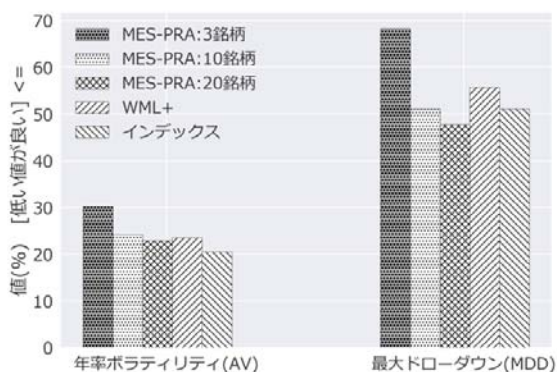


図9 リスク評価

リスクのみで評価した場合には、MES-PRAの最も高いリ
ターンを示すMES-PRAの3銘柄選択ケースが、VOLおよび
MDDにおいて最も高いリスクを示した。特に、MDDは
70% に近く、資産価格の減少リスクが高いことを示して
いる。

6. まとめ

本研究では、以下の結論を得た。

- 1) 幾何ブラウン運動(GBM)と外部バイアスにより
定義されるEFMEモデルを提案した。
- 2) 本提案手法により生成された人工データセット
と、MES-PRAを適用することにより、EFMEモデ
ルと超過収益との関係を分析することが可能と
なった。
- 3) MES-PRAにより、バイアスを付加した人工デー
タセットと実際の株価データから超過リターン
を得られることを確認した。
- 4) MES-PRAにより、バイアスが付加された人工デー
タセットのみで超過収益が得られることがわか
かった。よって、付加されたバイアスが超過収
益に寄与していると考えられる。
- 5) 上昇・下落バイアスを付加された人工データセ
ットの検証結果が、モメンタム効果を有すると
される米国のS&P500株式のデータセットにお
いて再現された。

参考文献

- [1] Fama, Eugene F., and Kenneth R. French. "The cross - section of expected stock returns." *the Journal of Finance* 47.2 (1992): 427-465.
- [2] Jegadeesh, Narasimhan, and Sheridan Titman. "Returns to buying winners and selling losers: Implications for stock market efficiency." *The Journal of finance* 48.1 (1993): 65-91.
- [3] Fama, Eugene F., and Kenneth R. French. "Size, value, and momentum in international stock returns." *Journal of financial economics* 105.3 (2012): 457-472.
- [4] Hong, Harrison, and Jeremy C. Stein. "A unified theory of underreaction, momentum trading, and overreaction in asset markets." *The Journal of finance* 54.6 (1999): 2143-2184.
- [5] Novy-Marx, Robert. "Is momentum really momentum?." *Journal of Financial Economics* 103.3 (2012): 429-453.
- [6] Hou, Kewei, Chen Xue, and Lu Zhang. *Replicating Anomalies*. No. w23394. National Bureau of Economic Research, 2017.
- [7] Sharpe, William F. "Capital asset prices: A theory of market equilibrium under conditions of risk." *The journal of finance* 19.3 (1964): 425-442.
- [8] Eling, Martin, and Frank Schuhmacher. "Does the choice of performance measure influence the evaluation of hedge funds?." *Journal of Banking & Finance* 31.9 (2007): 2632-2647.
- [9] Robert, Engle, Ferstenberg Robert, and Russell Jeffrey. "Measuring and modeling execution cost and risk." *The Journal of Portfolio Management* 38.2 (2012): 14-28.
- [10] Low, Rand Kwong Yew, and Enoch Tan. "The role of analyst forecasts in the momentum effect." *International Review of Financial Analysis* 48 (2016): 67-84.
- [11] Postali, Fernando AS, and Paulo Picchetti. "Geometric Brownian motion and structural breaks in oil prices: a quantitative analysis." *Energy Economics* 28.4 (2006): 506-522.
- [12] Marathe, Rahul R., and Sarah M. Ryan. "On the validity of the geometric Brownian motion assumption." *The Engineering Economist* 50.2 (2005): 159-192.
- [13] Bier, Martin, and R. Dean Astumian. "Biasing Brownian motion in different directions in a 3-state fluctuating potential and an application for the separation of small particles." *Physical review letters* 76.22 (1996): 4277.

決算短信から抽出した業績要因文の事業セグメントに基づく分類と業績文の抽出

Classification based on business segments of causal information and extraction of performance sentences from summary of financial statements

村野壮人¹ 酒井浩之¹ 坂地泰紀² 江口潤一³

Taketo Murano¹, Hiroyuki Sakai¹, Hiroki Sakaji², and Junichi Eguchi³

¹成蹊大学 理工学部 情報科学科

¹ Department of Computer and Information Science, Faculty of Science and Technology,
Seikei University

² 東京大学

² The University of Tokyo

³ 大和証券投資信託委託株式会社

³ Daiwa Asset Management

Abstract: In this research, we propose a method to automatically classify sentences including causal information concerning business performance (e.g. “Orders of semiconductor manufacturing equipment were good.”) extracted from summary of financial statements of companies based on business segments of the companies. Moreover, we propose a method to extract performance sentences from summary of financial statements. For example, the sentences including causal information extracted from summaries of financial statements of SUBARU Co., Ltd. are classified to either “automobile” segment or “aerospace” segment. In addition, our method extracts performance sentences, e.g. “Sales were ¥3,262.0 billion, an increase of ¥93.7 billion (2.9%) compared with the previous fiscal year.”, by deep learning and automatically generates training data.

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が求められており、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、企業の発行している「決算短信」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行われている[1][2][3][4][5]。

投資家が投資活動を行うにあたり、上場企業の業績情報の収集は必要不可欠である。また、業績情報の中でも特に業績要因が投資判断において重要である。なぜなら、業績回復の要因が、その企業の主力事業が好調であることであったならば株価への影響は大きい、株式売却益の計上などの特別利益の計上が要因であるならば株価への影響は軽微であるからである。

関連研究として、酒井らは企業の決算短信 PDF か

ら業績要因文を自動抽出する研究を行っている[1]が、抽出された業績要因がその会社の中でどの程度の重要性かを判断するのは難しい。例えば、株式会社 SUBARU では事業セグメントが「自動車」と「航空宇宙」の2つあり、それぞれのセグメントごとの2016年3月31日から2017年4月1日までの利益は自動車セグメントで397,657百万円、航空宇宙セグメントで9,102百万円となっており、大きな差がある。そのため、SUBARUにとって自動車セグメントの業績要因は航空宇宙セグメントの業績要因よりも重要であると考えられる。そのため、酒井らの研究によって抽出された業績要因がどの事業セグメントに属するかを分類し、さらに、事業セグメントごと業績によって重要度を付与できれば、投資判断を行ううえで重要な情報源となると考える。

そこで本研究では、例えば、SUBARUの決算短信から抽出された業績要因文である「国内の登録車につきましては、全面改良を行った「インプレッサ」

に加え、「レヴォーグ」および「フォレスター」の販売が好調に推移したことにより、売上台数は12.6万台と前期に比べ1.5万台(13.3%)の増加となりました。」に「自動車」等の事業セグメントに自動で分類する手法を提案する。また、それに付随し、分類されたセグメント毎の重要度の業績による付与を目標とし、業績文の自動抽出についても行う。例えば、「売上高は、自動車売上台数の増加などにより、為替変動に伴う売上高の減少を吸収し、過去最高となる3兆3,260億円と前期に比べ937億円(2.9%)の増収となりました。」といった業績文を深層学習によって抽出する。

2. 業績要因文の事業セグメントに基づく分類

業績要因文の事業セグメントに基づく分類をするにあたって、各企業ごとの事業セグメント名が必要になる。その事業セグメント名を自動で収集できることが望ましいが、今回の研究では人手で事業セグメント名を収集し、その事業セグメントに基づいた分類を行うものとする。また、業績要因文を分類するにあたって、学習データを人手にて作成し、その学習データによる機械学習手法による方法が考えられる。しかし、本タスクにおいては企業ごとに分類に必要な学習データが必要になるため、全ての企業に対応できる学習データの作成には多大な労力を要する。そのため、実際に業務に利用する際のことを考慮し、機械学習に用いる学習データをもなるべく自動的に生成し、業績要因文を事業セグメントに分類することを目標とする。以上の説明をふまえ、本手法の概要を以下に示す。

Step 1: 決算短信 PDF や企業 Web ページから各企業の事業セグメント名を人手で収集する。

Step 2: 決算短信 PDF から抽出された業績要因文に直前に出現した事業セグメント名を付与したものを教師データとして K 近傍法で分類する。

Step 3: 自動で直前に出現した事業セグメント名を付与した業績要因文を、業績要因文に含まれる企業キーワード(後述)の合計スコアを元に閾値で絞り、それを教師データとして K 近傍法によって分類する。

Step 2 のみでは精度が低すぎるため、Step 3 の処理を実行した。

2.1. 各企業の事業セグメント名の収集

企業の多くは、事業セグメントという、企業の構成単位によって、業務内容を分類している。本研究では事業セグメント名の自動抽出はせず、人手によって企業ごとに収集したセグメント名を業績要因文に付与していく。表1に人手で収集した企業と事業セグメントの例を記す。

表1: 企業毎の事業セグメント

企業名	事業セグメント名
日本電信電話	地域通信事業, 長距離・国際通信事業, 移動通信事業, データ通信事業
SUBARU	自動車, 航空宇宙
三菱地所	ビル事業, 生活産業不動産事業, 住宅事業, 海外事業, 投資マネジメント事業, 設計監理事業, ホテル事業, 不動産サービス事業
花王	ビューティケア事業, ヒューマンヘルスケア事業, ファブリック&ホームケア事業, ケミカル事業
日立製作所	情報・通信システム, 社会・産業システム, 電子装置・システム, 建設機械, 高機能材料, オートモティブシステム, 生活・エコシステム

2.2. 業績要因文への事業セグメント名の付与

業績要因文へ事業セグメント名を付与するにあたって、業績要因文の直前、または文中に出現していた事業セグメント名を付与する。この時、同じ文中に複数の事業セグメント名が出現していた場合、その文へ付与するセグメント名は「無し」とした。

例えば、SUBARUで「一方、軽自動車につきましては、4月に「ルクラ」、「プレオ」、「プレオバン」を投入したことや、「サンバー」シリーズが前年同期を上回る台数で推移したことにより、売上台数は51千台と前年同期比5千台(10.6%)の増加となりました。」といった業績要因文が抽出された場合、文中に事業セグメントの一つである「自動車」が含まれているため、「自動車」セグメントに分類される。

2.3. k 近傍法による業績要因文の分類

2.2 節の手法によって生成されたデータを学習データとして、セグメント名の分類を、K 近傍法を用いて行う。また、文中に複数の事業セグメント名が出

現していた場合、その業績要因は特定の事業セグメントに属しないと判断し、セグメント名を「無し」とした。ただし、2.2節の手法によって作成した学習データによる分類は精度が低いため、学習データの絞り込みを行った。具体的には、学習データとする業績要因文に含まれる企業キーワードのスコアの合計値を算出し、企業毎のその平均値に 0.8 をかけたものを閾値とし、閾値未満となった業績要因文を教師データから除外した。ここで、企業キーワードとは酒井らの手法[1]によって決算短信から抽出された、企業ごとの重要なキーワードである。企業キーワードは、企業 t の決算短信 PDF 集合 $S(t)$ に含まれる名詞 n に対して、以下の式でスコア $W(n, S(t))$ を計算し、スコアが大きい名詞を企業キーワードとして抽出する。

$$W(n_i, S(t)) = \left(0.5 + 0.5 \frac{tf(n_i, S(t))}{\max_{j=1, \dots, m} TF(n_j, S(t))} \right) \times H(n_i, S(t)) \times \log_2 \frac{N}{df(n_i)}$$

ここで、

$S(t)$: ある企業 t の決算短信の集合。

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

$H(n, S(t))$: $S(t)$ の各決算短信である d に名詞 n が出現する確率に基づくエントロピー。

$df(n)$: 名詞 n を含む決算短信をもつ企業の数。

N : 決算短信を収集した企業の数。

企業キーワードとスコア $W(n, S(t))$ の例を表 2 に示す。

表 2: 企業キーワードの例

企業名	企業キーワード	$W(n, S(t))$
花王	コンシューマーブ ロダクツ事業	7.65
	コンシューマーブ ロダクツ	7.15
	ヒューマンヘルス ケア事業	6.26
	ヒューマンヘルス ケア	5.74
SUBARU	レガシィ	1.50
	インプレッサ	1.48
	フォレスター	1.44
	宇宙事業部門	1.05
	航空宇宙事業部門	1.05

生成した学習データ D と、学習データ以外の決算短信から新たに抽出した業績要因文 T の文書間類似度

を(sim)を以下の式によって求め、学習データにおける事業セグメントが付与された文集合の近傍となる業績要因文を求める。

$$sim(V_d, V_t) = \frac{V_d \cdot V_t}{\|V_d\| \cdot \|V_t\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

$$V_d = \{x_1, x_2, \dots, x_n\}, V_t = \{y_1, y_2, \dots, y_n\}$$

ここで、

V_d : 学習データに含まれる業績要因文 D の企業キーワードを要素、そのスコアを要素値としたベクトル

V_t : テストデータとなる業績要因文 T の企業キーワードを要素、そのスコアを要素値としたベクトル

求められた近傍となる文に付与された事業セグメント名による投票で、テストデータの業績要因文に付与する事業セグメント名を決定する。

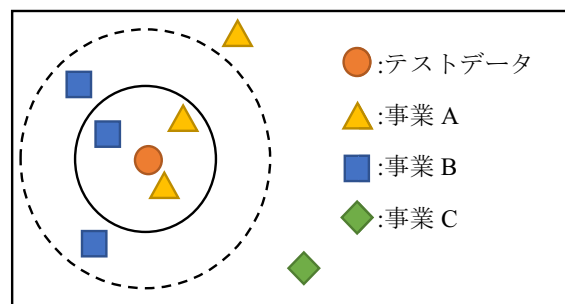


図 1: k 近傍法による分類

図 1 の例では最近傍法ならテストデータに最も近い事業 A に分類され、 $k=3$ でやった場合実線の円の中に入っている 3 つのセグメントで投票が行われる。この場合、事業 A が 2 つ、事業 B が 1 つなので事業 A に分類される。 $k=5$ の場合、点線の円の中に含まれる 5 つのセグメントで投票が行われ、事業 A が 2 つ、事業 B が 3 つとなるので、テストデータは事業 B に分類される。

3. 業績文の自動抽出

事業セグメントごとの業績を業績要因文に付与し事業セグメントと結びつけ、重要度とするため業績文を抽出する。なお、事業セグメントごとの業績は 1 つの表にまとまっていることが多いが、その表は企業ごとにフォーマットが異なっている。そのため、表から事業セグメントごとの業績を直接、抽出することは困難であり、決算短信に含まれる文から業績文を抽出する必要があった。

業績文の抽出には深層学習を用い、そのための学習データも自動生成する。学習データはランダムに

選んだ 1,000 社の過去の決算短信 PDF から以下の条件に合致する文をそれぞれ正例、負例とする。

正例は、「売上」「億円となりました」のどちらも含まれている文、例えば、「当社の当第3 四半期累計売上高は、主に前年同期比出荷ビットの増加により、32.2%増の 4,222 億円となりました。」や、「HE&S 分野の売上高は、液晶テレビの販売台数が減少しましたが、主に為替の好影響により、前年同期比 11.8%増加し、2,638 億円となりました。」である。

負例は、文中に算用数字、漢数字、共に含まない文、例えば「品質管理及びコンプライアンスに関する教育の強化につきましても継続的に推進しております。」や、「以下、前年同期比については、当該変更を反映した前年同期の数値を用いております。」のようなものである。

以上の条件において抽出された、それぞれ正例 1273 文、負例 328911 文を学習データとし、学習データの正例、負例のどちらにも出現している内容語素性 395 個とした深層学習を用いて分類する。

深層学習のモデルの入力層のノード数は学習データから抽出された素性 (395 語) 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。エポック数は 50、活性化関数として、ReLU を使用した。

本モデルによって業績文 (正例) として識別された文の誤分類を調べたところ、「これらの結果、売上台数の合計は、90.6 万台と前期に比べ 9.3 万台 (11.4%) の増加となりました。」のような、車の台数に対する文などで、「円」という言葉が入っていないことがわかった。そのため、「円」を含んでいない文を除去した。本手法により抽出できた業績文の例を以下に示す。

- 当第 1 四半期連結累計期間の営業収益は、海外ビジネスにおける為替影響があったものの、国内ビジネスの規模拡大などにより、3,735 億円 (前年同期比 3.7%増) となりました。
- 一方、負債は前期と比較して 4 億 11 百万円 (6.0%) 減少し、65 億 8 百万円となりました。
- また、自営店と加盟店の売上を合計したチェーン全店売上は 4 兆 5,156 億 5 百万円 (前年同期比 5.2%増) となりました。

4. 業績文と事業セグメント名の出現位置による分類

業績要因文への事業セグメントの分類を、3 章で抽出した業績文の出現位置を利用して行う手法を試みた。事業セグメント名が出現する文と業績文の間を同一セグメントに対する文集合と仮定し、その中に業績要因文が出現したときのみ、業績要因文に事業セグメント付与を行う。しかし、決算短信には企業毎に書き方の特徴があり、事業セグメント名が出現した直後に業績文が出現し、そのあとに業績要因文が出現する場合があったため、他の事業セグメント名が出現するまで、業績文が複数回出現した時、最後の業績文までをひとまとまりとして、そのまとまりの中に出現した業績要因文を分類した。2 節と 4 節での、業績要因文の事業セグメント分類した結果を表 3 に示す。

表 3: 業績要因文のセグメント分類

企業名	事業セグメント名	業績要因文
花王	ビューティケア事業	日本では、「ビオレ」の洗顔料や日焼け止め、乾燥性敏感肌ケア「キユレル」の売り上げが伸長し、前期を上回りました。
	ファブリック & ホームケア事業	アジアでは、インドネシア、タイで衣料用洗剤「アタック」が好調に推移し、台湾、香港で、抗菌機能を高めた衣料用液体洗剤を発売して市場を活性化し、売り上げが伸長しました。
SUBARU	自動車	また、軽自動車につきましては、新型車「シフォン」が販売に寄与したものの、その他車種が減少したことにより、売上台数は 3.3 万台と前期に比べ 0.1 万台 (3.4%) の減少となりました。
	航空宇宙	(航空宇宙事業部門)防衛省向け製品では、新多用途ヘリコプター「UH-X」の契約に基づく開発本格化などにより、売上高は前期を上回りました。

5. 評価

5.1 業績要因文の分類の評価

業績要因文のセグメント分類の評価は、事業セグメント名を収集した上場企業の中から選んだ 11 社の企業の決算短信において、自動生成したままの学習データを用いた最近傍法（自動生成）、閾値によって学習データの再生成を行ったものを用いた最近傍法（最近傍法）、 $k=3$ の k 近傍法 ($k=3$)、業績文の位置を利用して分類したもの（業績文利用）の場合の 4 パターンで自動的に業績要因文への事業セグメント名の付与をし、精度を求めた。評価結果を表 4 に示す。また、このとき、どのセグメントにも属しないと分類された業績要因文は全体の数から除外して計算をする。表中の分子は正しく分類された業績要因文の数、分母は抽出されたのち、セグメント名が付与された業績要因文の数である。

また、近傍法での分類が有効かを調べるため、学習データを自動生成したのち、手動で修正を加えた。「花王」、「SUBARU」の 2 社で最近傍法を実行した場合の精度を表 5 に示す。

表 5: 手動学習データでの最近傍法

企業名	精度
花王	0.9 (9/10)
SUBARU	1 (10/10)

5.2 業績文の抽出の評価

業績文の抽出は、業績要因文の事業セグメントの分類に使用したものと同一 11 社で行った。11 社の最新の決算短信 PDF から抽出した。深層学習による分類と、分類の後に、「円」を含まないものを除去した二通りの精度を表 6 に示す。

表 4: 業績要因文のセグメント分類数と精度

企業名	自動生成	最近傍法	$k=3$	業績文利用
日本たばこ産業	6/8	6/7	6/7	0/0
セブン&アイ・ホールディングス	3/6	2/4	1/3	1/1
花王	7/8	3/3	5/5	6/6
ヤフー株式会社	2/8	2/5	1/5	2/2
中国工業	5/9	0/1	0/0	6/7
村田製作所	0/4	0/1	1/1	1/1
トヨタ自動車	5/10	0/0	0/0	2/9
SUBARU	8/10	6/8	6/8	8/9
三菱地所	2/9	0/0	0/0	0/0
日本電信電話	4/6	1/1	0/0	0/0
NTT ドコモ	6/6	2/2	8/9	5/7
合計 (精度)	48/84 (0.571)	22/32 (0.688)	28/38 (0.737)	31/42 (0.738)

表 6: 業績文の抽出精度

深層学習による抽出	0.928 (65/70)
「円」を含まないものを除去	0.984 (62/63)

6. 考察

業績要因文のセグメント分類は、企業ごとの決算短信の書き方に大きく左右されてしまうため、各企業での精度の差が大きくなってしまっている。例として、「花王」、「NTT ドコモ」の 2 社はどの分類法を用いたとしても、多くの業績要因文に対し、セグメント分類が行われ、その精度も良いが、「トヨタ自動車」、「三菱地所」は業績要因文の抽出の時点で抽出結果に誤ったものも多いため、分類ができなくなっている。「ヤフー株式会社」では、明確にセグメント分類できる業績要因文が 2 文しかないにもかかわらず、無理に分類しようとしてしまうため、精度が下がってしまっていたが、業績文を利用した分類の場合は 2 文のみを適切に分類することができている。しかし、「日本たばこ産業」の場合、他の手法ではうまく分類できていたが、業績文の抽出の再現度が低いことと、業績文の出現する位置の関係上、業績文を利用した場合、全てが未分類となってしまった。

このように、企業によって決算短信の書き方の特徴に多く作用されてしまう。そのため、複数の分類法を企業ごとに使い分けることが良いと考える。しかし、実際に運用するにあたって同じシステムで多くの企業に対応できることが求められるため、2.3 節で行った企業キーワードのスコアを用いた学習データの絞り込みによる誤分類の除去や、 k 近傍法と業績文利用の 2 つの和集合を取ることで、精度・再現率の向上をしたい。

また、「NTT ドコモ」では「通信事業」という事業名に対する文を「通信関連」などと表記されるなど

の表記揺れが起きていることが確認できた。このような表記揺れは他の企業の決算資料でも見られるので、人手で収集した事業セグメント名を元に表記揺れに対応できるようにするか、自動で事業セグメント名を収集できるようにする技術が求められる。

7. まとめ

本研究では、企業の決算短信 PDF から抽出した業績要因文に対する事業セグメントへの分類手法を提案した。業績要因文の抽出では、人手で収集した事業セグメント名を元に学習データを自動生成し、生成された学習データを用いた k 近傍法によって、業績要因文を分類した。評価の結果、k=3 での k 近傍法で精度 73%、また、業績文の位置を利用した分類でも精度 73%となり、比較的、良好な精度を得ることができた。今後は再現率の向上と共に、業績文中に出現する業績情報による業績要因文への重み付けも行う予定である。

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] 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.
- [3] 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.
- [4] 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.
- [5] 室野莉沙, 酒井浩之, 坂地泰紀, ベネット ジェイソン, “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.

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

Extraction of Basis Information on Analyst's Forecasts and Assigning Polarity to Analyst Reports

小林和正¹ 酒井浩之¹ 坂地泰紀² 平松賢士³

Kazumasa Kobayashi¹, Hiroyuki Sakai¹, Hiroki Sakaji², Kenji Hiramatsu³

¹成蹊大学 理工学部 情報科学科

¹Department of Computer and Information Science, Faculty of Science and Technology,
Seikei University

²東京大学

²The University of Tokyo

³株式会社アイフィスジャパン

³IFIS JAPAN LTD

Abstract: In this paper, we proposed a method for supporting investors. Our method extracts basis information on analyst's forecasts from analyst reports and assigns polarity to the analyst reports. Analyst reports which are written about a company's performance or profitability by securities analysts are useful for investment but investors can only read it a little because many reports are published. Therefore, a system which judges investing by an artificial intelligence technique is required. By giving polarity to the analyst reports, the proposed method catches a slight change in performance. This ability of method is useful to judge whether investors need to read analyst reports carefully.

1. はじめに

近年、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。業績予測や株価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証券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、全てに目を通し、内容を把握することは困難であり、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いる。

そこで本研究は、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からアナリスト予想根拠情報を抽出する手法と、レーティングが変動しないアナリストレポートに対して、深層学習を用いて極性を付与する手法を提案する。ここで、アナリスト予想根拠情報とは、例えば、「世界経済の回復に加え、米ドル安や中国の供給削減期待などから市況は急回復し、同社の輸出も今後大きく改善する可能性が高い。」といった、投資判断やアナリストレポートの内容を把握するうえで重要な、アナリスト予想の根拠情報を含む文（以降、アナリスト予想根拠文とする）と定義される[1]。アナリスト予想根拠文の抽出や極性の付与により、アナリストレポートの内容把握に要する時間の削減や、レーティングが変動しない程度のわずかな業績変化を捉え、そのアナリストレポートを熟読するか判断するための情報となることが期待できる。

アナリストレポートからのアナリスト予想根拠文の抽出は、酒井らの先行研究が存在している[1]。しかし、酒井らの手法におけるアナリスト予想根拠文の再現率は60%程度であり、それほど高いわけではない。酒井らはアナリスト予想根拠文を抽出するた

めの手がかりとなる表現を「手がかり表現」と定義し、そのような表現をブートストラップ的に獲得する。そして、獲得された手がかり表現を使用してアナリスト予想根拠文を抽出している。しかし、ブートストラップ的に手がかり表現を獲得する過程において、手がかり表現として不適切な表現を削除する必要があり、そのため、適切な手がかり表現であるにもかかわらず、獲得できない場合がある。そのため、酒井らの手法によるアナリスト予想根拠文の抽出は、比較的高い精度（75%程度）を達成しているものの、再現率は低いという結果となっている。

そのため、本研究では、深層学習を使用してアナリストレポートからアナリスト予想根拠文を抽出し、精度を落とさずに酒井らの手法よりも高い再現率を達成する手法を提案する。具体的には、酒井らの手法は高い精度を達成していることに着目し、酒井らの手法による抽出結果をさらに絞り込むことで、より高い精度のアナリスト予想根拠文の集合を作成する。そして、作成された高精度のアナリスト予想根拠文を深層学習の学習データとすることで学習データを自動生成し、その自動生成された学習データを使用して深層学習を行い、アナリスト予想根拠文を抽出する。

2. アナリスト予想根拠情報の抽出

本研究は、アナリストレポートからアナリスト根拠情報を抽出する酒井らの手法[1]によって抽出したアナリスト予想根拠文を深層学習の学習データとして使用する。以下、酒井らの手法[1]について簡単に述べる。

2.1 アナリストレポートからの手がかり表現の自動獲得

アナリスト予想根拠情報の抽出には、業績発表記事や決算短信から業績要因文を抽出した酒井らの手法[2][3]を適用し、アナリストレポートからアナリスト予想根拠情報を抽出する際に有効な手がかりとなる表現（以降、「手がかり表現」と定義）を獲得する。この手がかり表現からアナリスト予想根拠情報を含むアナリスト予想根拠文を抽出する。

まず、「予想する」、「考える」、「高い」の3つの手がかり表現を手で与え、それに係る節を取得する。取得された節の集合において、頻繁に出現する表現を共通頻出表現として抽出する。ここで、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1で求める。この値が、式2で求めた閾値 T_e 以上の共通頻出表現

を選別する。

$$H(e) = -\sum_{s \in S(e)} P(e, s) \log_2 P(e, s) \quad (1)$$

ここで、アナリストレポートの集合において、 $S(e)$: 共通頻出表現 e に係る手がかり表現の集合。
 $P(e, s)$: 共通頻出表現 e が手がかり表現 s に係る確率。

$$P(e, s) = \frac{f(e, s)}{\sum_{s \in S(e)} f(e, s)}$$

ここで、 $f(e, s)$ は共通頻出表現 e が手がかり表現 s に係る回数である。

$$T_e = \alpha \log_2 |N_s| \quad (2)$$

選別した共通頻出表現から新たな手がかり表現を獲得する。先ほどと同様に、様々な手がかり表現に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に係る確率に基づくエントロピーを求め、手がかり表現を選別する。

2.2 手がかり表現と共通頻出表現を使用したアナリスト予想根拠文の抽出

アナリストレポートから獲得した手がかり表現と共通頻出表現を用いて、アナリストレポートからアナリスト予想根拠文を抽出する。まず、アナリストレポートから手がかり表現を含む文を抽出して係り受け解析し、手がかり表現を含む文節を取得する。この文節にかかっている文節を取得して連結する。連結された文節列に共通頻出表現が含まれていれば、その文をアナリスト予想根拠文として抽出する。

2.3 文末手がかり表現の自動獲得

2.1節で取得された手がかり表現は1文節のみだが、多分節で構成されている手がかり表現も多く存在する。アナリストレポートには文末に特徴的な表現が多く出現している傾向があることに着目し、2.2節で抽出されたアナリスト予想根拠文の文末に出現しており、多分節で構成される手がかり表現（以降、文末手がかり表現と定義）を取得する。まず、2.2節で抽出されたアナリスト予想根拠文の文末に出現する1文節（以降、文末文節と定義）を取得する。これらの文末文節は有効な手がかり表現ではないが、文末文節に係っている文節列を取得して文末文節と

組み合わせると、有効な文末手がかり表現となる可能性がある。そこで、文末手がかり表現を獲得するにあたり、文末文節に係っている文節列を取得するが、文末文節とそれに係る文節列の組み合わせは膨大な数になるため、組み合わせを絞り込む必要がある。文末文節 c に係る文節列 p に対して式3でスコアを求め、平均値を上回る文節列のみを抽出する。そして、抽出した文節列と文末文節とを連結した表現を、文末手がかり表現として獲得する。

$$Score(p, c) = -f(p, c) \sqrt{fp(p)} \log_2 P(p, c) \quad (3)$$

$$P(p, c) = \frac{f(p, c)}{N(c)}$$

ただし、アナリストレポートから取得したアナリスト予想根拠文の集合において、

$P(p, c)$: 文末文節 c から取得される文字列 p の出現確率。

$f(p, c)$: 文末文節 c から取得される文字列 p の取得回数。

$N(c)$: 文末文節 c から取得される文字列の総数。

文末手がかり表現を使用したアナリスト予想根拠情報の抽出は、2.2節で示した手法と同様である。

3. 深層学習によるアナリスト予想根拠文の抽出

3.1 学習データの自動生成

酒井らの手法[1]により抽出されたアナリスト予想根拠文から学習データを生成する。しかし、抽出されたアナリスト予想根拠文の精度は75%程度であり、それをそのまま学習データとすることはできない。そこで、以下の処理を行うことで、抽出されたアナリスト予想根拠文を絞り込み、その結果を学習データとする。

Step 1: 手がかり表現を3つ以上、または、1つの文末手がかり表現を含んでおり、かつ、共通頻出表現を3つ以上含む文において、後述のキーワードに基づくスコアが30以上のアナリスト予想根拠文を正例とする。

Step 2: 共通頻出表現、手がかり表現、文末手がかり表現を含まない20文字以上の文を負例とする。

Step 3: Step 1 および Step 2 で抽出した学習データを利用し、深層学習でテストデータのアナリスト

レポートからアナリスト予想根拠文を抽出する。

Step 1 で使用するスコアはアナリスト予想根拠文に含まれる企業にとって重要なキーワードのスコアの和である。このキーワードは対象企業の決算短信PDFから抽出している。決算短信PDFを採用した理由としては、上場企業が年4回、決算短信を発表しているのに対し、アナリストレポートは企業によって数にばらつきがあるためである。決算短信PDFを使用したキーワードのスコア付与は、企業 t の決算短信PDFに含まれる名詞 n に対して以下の式4で重み $W(s, S(t))$ を算出する。

$$W(n_i, S(t)) = \left(0.5 + 0.5 \frac{TF(n_i, S(t))}{\max_{j=1, \dots, m} TF(n_j, S(t))} \right) \times H(n_i, S(t)) \times \log_2 \frac{N}{df(n_i)} \quad (4)$$

ここで、

$S(t)$: ある企業 t の決算PDFの集合。

$TF(n, S(t))$: $S(t)$ において、名詞 n が出現する頻度。

$H(n, S(t))$: $S(t)$ の各決算短信PDF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。

Step1で抽出された正例と、それに含まれる手がかり表現、共通頻出表現、文末手がかり表現をいくつか示す。

しかし、消費材は耐久財に比べて利用頻度が高いことからサイトのPVの向上等のメディア価値の向上に繋がると考える。

- 手がかり表現：高い、繋がると、比べて、考える
- 共通頻出表現：向上、利用頻度、価値
- 文末手がかり表現：繋がると考える。

国内トラック需要は強く、新興国市場に強い同社は、本来は成長余地を大きく有している。

- 手がかり表現：強く、大きく、強い
- 共通頻出表現：市場、国内、需要
- 文末手がかり表現：なし

Step2で抽出された負例をいくつか示す。手がかり表現および共通頻出表現を含まず、文末手がかり表現

もないことがわかる。

ただし、いずれのニュースも非常に大きなインパクトを持つというわけではない。

事業内容Flashコンテンツを中心としたオリジナル動画キャラクターの開発等を展開。

上記の手法により、正例として20,824文、負例として正例と同数の20,824文の学習データを自動的に生成した。

3.2 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5で重みを計算する。

$$W_p(t, S_p) = TF(t, S_p)H(t, S_p) \quad (5)$$

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6で求める。

$$H(t, S_p) = -\sum_{s \in S_p} P(t, s) \log_2 P(t, s) \quad (6)$$

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

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7で重みを計算する。

$$W_n(t, S_n) = TF(t, S_n)H(t, S_n) \quad (7)$$

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。

$$W_p(t, S_p) > 2W_n(t, S_n)$$

$$W_n(t, S_n) > 2W_p(t, S_p)$$

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。

事業、拡大、販売、収益、株価、可能、向け、改善、影響、増加、需要、増益、国内、製品

3.3 モデル

深層学習のモデルについて以下に述べる。入力層は41,648文の学習データから抽出された5,699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5,699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。

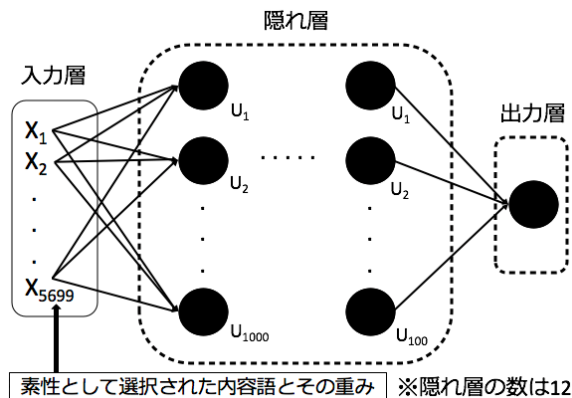


図1: ニューラルネットワークのモデル

4 アナリストレポートに対する極性付与

アナリスト予想根拠情報の抽出により、アナリストレポートの内容把握に要する時間の削減が期待できるが、レーティングが変動しない程度のわずかな業績変化を捉え、そのアナリストレポートを熟読するか判断するための情報としては、アナリストレポートに対して極性を付与するほうが有効である。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。レーティングが下がっていればネガティブである。しかし、多くのアナリストレポートではレーティングの変更はないものの、記述されている内容をポジティブ、ネガティブに分類できる。

そこで、本研究では、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し、その学習データによる深層学習にて、レーティングの変更がないアナリストレポートに対する極性（ポジティブ、ネガティブ）を試みた。

ここで、学習データからの素性選択は 3.2 節で述べた手法と同じであり、深層学習のモデルは 3.3 節のモデルと同じである。

5 評価

5.1 アナリスト予想根拠文抽出の評価

本手法を実装し、自動生成した学習データを使用して図 1 のモデルを用いて深層学習を行い、アナリストレポートからアナリスト予想根拠文を抽出した。抽出された例を以下に示す。

文： 4月からの薬価改定影響、ジェネリック促進策による市場拡大効果を踏まえ業績見直しを見直した。

文： 政府の薬価改定強化がG x 市場にも及んでいること、ブランド医薬品メーカーがG x 事業でのプレゼンスを高めていること、この2点が最大の懸念事項だ。

評価のための正解データは、手がかり表現と共通頻出表現を獲得したアナリストレポートを含まない、

すなわち深層学習の学習データとして使用していないアナリストレポート集合から 12 個のアナリストレポートを無作為に選択し、その中の 468 文から人手でアナリスト予想根拠文を抽出して作成した。次に、選択したアナリストレポートから本手法にて抽出したアナリスト予想根拠文が正解データの文と一致すれば正解とし、精度、再現率、F 値を算出した。

本手法 1: 深層学習による手法の抽出結果と、手がかり表現と文末手がかり表現を使用した既提案手法[1]の抽出結果との和集合をとる手法

本手法 2: 深層学習による手法のみ

本手法 3: 深層学習による抽出結果と、既提案手法による抽出結果との積集合をとる手法

比較手法: 手がかり表現と文末手がかり表現を使用する既提案手法

評価結果を表 2 に示す。

表 2: アナリスト予想根拠文抽出の評価結果

手法	精度 (%)	再現率 (%)	F 値
本手法 1	75.13	69.60	72.25
本手法 2	83.33	44.11	57.68
本手法 3	83.90	35.78	50.16
比較手法	75.00	61.76	67.73

5.2 アナリストレポートへの極性付与の評価

4 章の手法を実装し、7,454 個のアナリストレポートが学習データとして自動生成された。そして、この学習データによる深層学習を用いて、評価用のレーティングが変動しなかった 155 個のアナリストレポートに対して極性を付与した。正解データを評価用と同じアナリストレポートを人手にて極性を付与することで作成し、本手法の精度を求めた。また、学習手法として SVM を使用した場合を比較手法とした。SVM の場合も、学習データ、および、素性は深層学習と同じである。評価結果を表 3 に示す。

表 3: 極性付与の評価結果

手法	Positive 精度 (%)	Negative 精度 (%)	全体 精度 (%)
深層学習	75.0 (84/112)	76.7 (33/43)	75.5 (117/155)
SVM	79.5 (70/88)	64.2 (43/67)	72.9 (113/155)

6 考察

本手法 1 は比較手法と比較して、精度、再現率、ともに高くなっており、良好な結果と言える。深層学習と手がかり表現と文末手がかり表現を使用する手法を組み合わせて和集合をとった結果が、積集合をとった結果や深層学習の結果より向上している。これは、手がかり表現と文末手がかり表現を使用する手法で抽出できなかったアナリスト予想根拠文は深層学習で抽出できており、さらに、深層学習で抽出できなかったアナリスト予想根拠文は手がかり表現と文末手がかり表現を使用する手法で抽出できていることを示している。

深層学習のみを使用した本手法 2 において、精度が高く、再現率が低い原因としては、負例に正例に分類されるべき文が含まれていたからであると考えられる。手がかり表現、共通頻出表現、文末手がかり表現を含んでいないとしても、アナリスト予想根拠文である可能性があり、負例の抽出条件を再考する必要がある。

7 まとめ

本研究では、アナリストレポートの内容を把握するうえで重要な、アナリスト予想の根拠情報を含む文（アナリスト予想根拠文）を自動的に抽出する手法を提案した。アナリストレポートから抽出されたアナリスト予想根拠文のみを提示することで、アナリストレポートの内容把握に必要な時間を削減できたり、アナリスト予想根拠文の中からその企業の事業に関連する根拠情報のみを選別することで、その企業への投資判断やアナリストレポートを熟読するかどうかを判断するための情報となることが期待できる。

本手法は酒井らの手法[1]から学習データを自動生成し、深層学習を行うことで、酒井らの手法[1]で抽出できなかったアナリスト予想根拠文を獲得できた。評価の結果、精度 75.13%、再現率 69.60%であり、酒井らの手法[1]と比較しても良好な精度、再現率を達成した。

参考文献

- [1] 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.
- [2] Hiroyuki Sakai, Shigeru Masuyama, “Cause Information Extraction from Financial Articles Concerning Business Performance”, IEICE Trans. Information and Systems,

vol.ED, no.4, pp.959-968, 2008.

- [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 ジェイソン, “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.

業種別企業業績要因を含む新聞記事の抽出

Extraction of Newspaper Articles including Cause Informations Concerning Business Performances for each Industry

丸澤 英将^{1*} 和泉 潔¹ 坂地 泰紀¹ 田村 浩道²

Hidemasa Maruzawa¹ Kiyoshi Izumi¹ Hiroki Sakaji¹ Hiromichi Tamura²

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: These days, a growing number of individual investors is attracting public attentions even in Japan, and securities companies are actively providing them with investment informations. Especially, analyst reports written by professional security analysts are important investment judgment materials, but their timing of publication varies by brands. In this paper, using the structures of causal relationships of sentences in analyst reports, the ways of security analysts paying attention to cause informations concerning business performances were learned, and newspaper articles including similar causal relationships were extracted. We aim to realize a real-time investment supporting system.

1 はじめに

近年、わが国でも個人投資家の増加が目立っており、証券会社も積極的に投資情報の提供を行っている。中でも、市場分析の専門家である証券アナリストが企業業績を予想するアナリストレポートは、重要な投資判断材料だが、発行時期は銘柄ごとにまちまちである。この間欠性を補うシステムとして、日々発行される新聞の記事などのデータから企業業績を変動させる要因になりうる経済イベント（業績要因）を即時に獲得してデータベースに蓄積し、顧客からの問い合わせに自然言語で回答する対話型投資支援システムが考えられる。例えば、ある銘柄の株価の変動要因を尋ねる質問に回答するなどの使い方が想定できる。

上記の目的のため、アナリストレポート中でどのような経済イベントが目立つか、企業業績の予想の根拠として用いられているかという因果関係の特徴を学習することで、新聞など別の媒体で報じられている経済イベントから、アナリストの行う企業業績予想を推測することを考える。特定の文の特徴を学習して、別の文章から類似の文を獲得する手法として、単に文全体に含まれる単語の組で一致度を測る bag-of-words 法よ

るものがあるが、その手法では背後にある因果関係を把握できていないという問題がある。

2 提案手法

対話型投資支援システムのための業績要因データベース構築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、因果関係のうちの原因を示す文の部分が指す内容を、企業業績予想の根拠情報として獲得する。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、根拠となりうる経済イベントを取得する。同様の経済イベントから、過去にアナリストがどのような予想を導いたかを参照し、新聞記事中の経済イベントによる企業業績の変化を予想する。このようにして生成した根拠情報、業績予想のリストを、企業業績要因データベースとする。

本稿では、この流れのうち、新聞記事中からアナリストレポートの根拠となりうる経済イベントを取得する段階までを論じる。

*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: m2016hmaruzawa@socsim.org

2.1 アナリストレポートからの根拠部、予想部の抽出

初めに、アナリストレポート中の因果関係を抽出するために、酒井らのブートストラップ法による手法 [?] を用いた。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、本研究では、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とを分離して抽出する工夫をした。前者を予想部、後者を根拠部と呼ぶこととする。また、各根拠部が指す経済イベントを、根拠情報と呼ぶこととする。アナリストレポート中の文の例を示す。

原油安及び探鉱費の増加を主因に、YY.M 期の純利益予想を下方修正した。

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「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一方、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。

この工夫により、以降に記す根拠部の特徴量抽出の際に、アナリストの予想を示す表現を排除し、予想の根拠として用いられた文の部分のみを対象とすることができる。また、ある根拠からどのような予想が導かれたという因果関係の対応が得られるため、新聞記事中の経済イベントによる企業業績の変化を予想するための学習データとすることができる。なお、企業業績に関する記述を原因・結果表現それぞれに分けて抽出する手法には、坂地ら [?] によるものがあるが、文章パターンの認識が決算短信に特化されている。アナリストレポートではより多様な表現が用いられていたため、本研究では、基礎手法として汎用性の高い酒井ら [?] の手法をもとにした。

2.2 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、根拠部の特徴を学習する。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、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以下の値を計算し、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。

$$\frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \left(\log \frac{|D|}{|d: t_i \in d|} + 1 \right) \quad (1)$$

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。業種とは、事業内容の銘柄をまとめた分類で、業種別に根拠部の特徴を学習することで、意味のある特徴を獲得できると期待される。

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。例えば、「鉄鋼・非鉄」業種に属する銘柄の根拠部には、鉄や銅の資源価格に関する根拠情報を指すものが多く含まれると推測される。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。例えば、ベンゼンの資源価格に関する根拠情報を指す根拠部は、「化学」業種に偏って存在するであろうが、為替に関する根拠情報を指す根拠部は、幅広い業種に満遍なく存在すると推測される。

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計算する。根拠部の特徴量には名詞の組を用いているため、この多クラス分類は、根拠部が指す根拠情報をおおよそ多クラス分類していると見なすこともできる。

$$f_{n,m} = |v: v \in (C_n \cap I_m)| \quad (2)$$

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 ($n = 1, 2, \dots, N_C, N_C$: クラスの総数)、 I は根拠部の属する業種を表す集合 ($m = 1, 2, \dots, N_I, N_I$: 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、平均情報量を用いて定量化する。

$$-\sum_m f_{n,m} \log_2 f_{n,m} \quad (3)$$

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。

2.3 新聞記事からの業種別根拠情報の獲得

前節までで、アナリストレポートから抽出される根拠情報を多クラス分類し、各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うるかの頻度分布を得た。以降、これらの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、新聞記事の文章を表す特徴量を設計する。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問題が生じる。同じ根拠情報を指す場合でも、アナリストレポートと新聞記事ではわずかに異なる表現を用いていることや、学習に用いたアナリストレポートの発行された期間では特定の経済イベントのみが注目され、逆の経済イベントを指す表現が獲得できないことが想定されるからである。前者の例として、株価が上がり続けていたという根拠情報を指すために、アナリストレポートでは「持続的な株価の上昇」、新聞記事では「株価続騰」と表現されることがある。後者の例としては、学習に用いたアナリストレポートの発行された期間では「原油価格の下落」という経済イベントのみが注目されていた場合、同じような銘柄の業績に影響を与えるであろう「原油価格の高騰」という表現が無視されてしまうことがある。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 [?] 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以下の値を計算し、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。

$$\frac{n_{i,l}}{\sum_k n_{k,l}} \cdot \left(\log \frac{|D|}{|d: t_i \in d|} + 1 \right) \quad (4)$$

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。

ここで、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一部の記事を除外する。経済記事の中には、新聞社が独自に企業の決算の内容を予想する「観測記事」と呼ばれる記事がある。観測記事は過去の

経済イベントを事実として報じるものではなく、将来の予想を伝えるものであるため、マーケットレポートに似た性質の文章である。そのため、観測記事を抽出する正規表現を用いて、対象から除外する。また、実際の企業の決算内容を報じる「決算記事」も、マーケットレポートでの予想部にあたる文章となるため、同じく正規表現を用いて、対象から除外する。

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。

$$s_{l,n} = v_l \cdot g_n \quad (5)$$

ここで、 $s_{l,n}$ は新聞記事の文章 a_l とクラス 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 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。

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うるかの頻度分布との加重平均をとることで、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うるかの指標とする。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。

$$c_{l,m} = \sum_n s_{l,n} f_{n,m} \quad (6)$$

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。

新聞記事をその文の各業種への業績寄与度を用いることで、特定の期間中の新聞記事のうち、各業種に属する銘柄の業績予想の根拠となりうる重要記事を一覧できることが期待される。ただし、単に業績寄与度の降順で並べると、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がある。例えば、日経平均株価の動向などがこれに該当する。この影響を取り除くため、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求める。

$$\text{dev}(c_{l,m}) = \frac{c_{l,m} - \mu_l}{\sigma_l} \cdot 10 + 50 \quad (7)$$

$$\mu_l = \frac{1}{N_l} \sum_m c_{l,m}$$

$$\sigma_l = \frac{1}{N_l} \sum_m (c_{l,m} - \mu_l)^2$$

ここで、 $\text{dev}(c_{l,m})$ 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である。

特定の業種についてこの偏差値が高い新聞記事の文章は、その業種への業績寄与度が特徴的に高いことを意味する。一方、この偏差値のみの降順で新聞記事を並び替えると、業績寄与度のみで並べた場合と逆に、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がある。例えば、企業業績とは関わりの薄いスポーツ記事などがこれに該当する。したがって、各業種への業績寄与度とその値の偏差値の調和平均をとったものを、新たに業績関連度指数と定義する。ただし、調和平均を求める際には、各業種への業績寄与度の値は平均 50、標準偏差 10 に正規化する。

$$r_{l,m} = \frac{2 c'_{l,m} \text{dev}(c'_{l,m})}{c'_{l,m} + \text{dev}(c'_{l,m})} \quad (8)$$

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。

新聞記事をその文の各業種への業績関連度指数の降順で並べることで、特定の期間中の新聞記事のうち、各業種に属する銘柄の業績予想の根拠となりうる重要記事を一覧できる。

3 実験設定

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からの根拠部、予想部の抽出に当たって、初期表現を以下のように選定した。初期の手がかり表現には、名詞の後に動詞「する」が続く熟語動詞の出現頻度上位 20 位の中から、根拠・予想を示す箇所に高確率で用いられる表現を選んだ。また、初期の共通頻出表現には、熟語の出現頻度上位 20 位の中から、根拠・予想を示す箇所に高確率で用いられる表現を選んだ。特に、根拠部の初期の共通頻出表現には、ポジティブ・ネガティブの判断を含む表現を用いるようにした。

選定の結果、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。学習データには、野村証券株式会社の Global Markets Research レポート（2013 年下半期発行分、日本株 216 銘柄の表紙部分）を用いた。

3.2 根拠情報の業種別特徴の学習

係り受け解析器として CaboCha [?], 形態素解析器に MeCab¹ を使用した。多クラス分類には、k-means

¹<http://taku910.github.io/mecab/>

法を用い、 $k=100$ とした。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。

3.3 新聞記事からの業種別根拠情報の獲得

新聞記事には、日経新聞の 2014 年の記事 (スポーツ記事など、経済記事以外も含む) 119,767 件を用いた。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らず類似度の閾値には、0.7 を使用した。観測記事を抽出する正規表現には、見出しと最初の 1 文について、1ヶ月分を手で仕分けることで得た以下の表現を用いた。

「*決算予想.*」または
「(利益 | 損益).*(る見通しだ | たようだ | たもようだ | になりそうだ | になった公算が大きい).」

同じく、決算記事を抽出する正規表現には、手手で得た以下の表現を用いた。

「決算(を | で).*発表.*た。」 または
「(利益 | 損益).*(円 | ドル).*(発表 | だった).*」
または
「(利益 | 損益).*(% | 倍に)(増えた | 減った | 増加した | 減少した | 増となった | 減となった | 上回った | 下回った).」 または
「(利益 | 損益).*(最高 | 最低 | 黒字 | 赤字).*た。」
または
「(円 | ドル)(から | に).*(上げた | 下げた | 修正した).」

抽出した重要記事について、トップ 30 における精度を求めた。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特徴量となる名詞を抜き出して特徴量に用いた bag-of-words 法による、トップ 30 における精度を用いた。精度の算出に当たっては、重要記事抽出対象の時期のアナリストレポート中で、5 つ以上の銘柄においてアナリストによる業績予想の根拠とされていた概念を正解とし、手手で評価した。

4 結果・考察

4.1 アナリストレポートからの根拠部、予想部の抽出

根拠部、予想部それぞれで新たに抽出された共通頻出表現、手がかり表現の一部を示す。

根拠部共通頻出表現

下落、上振れ、好調さ、減速、悪化

根拠部手がかり表現

織り込んで、踏まえ、主に

予想部共通頻出表現

成長、増収、増益、採算改善、コスト削減

予想部手がかり表現

見込まれる、期待される、続こう

根拠部、予想部ともに、それぞれの特徴を捉えた共通頻出表現、手がかり表現が新たに抽出できている。特に、根拠部の共通頻出表現は、初期表現にはポジティブな意味を持つ単語のみを与えたにも関わらず、対応するネガティブな表現も抽出できている。一方、予想部の共通頻出表現は、ポジティブな表現が大半を占める結果となった。これは、学習の対象としたアナリストレポートの発行時期が、いわゆるアベノミクスによる市場全体の回復期にあたるため、アナリストの予想がポジティブな方向に偏っていたためと考えられる。

4.2 根拠情報の業種別特徴の学習

根拠部の手がかり表現をもとに、6,655文の根拠部を得、この根拠部から、特徴量に用いる名詞を2,968個得た。根拠部を多クラス分類した結果、1つのクラスに属する根拠部の数は最低12個、最大1,492個となった。平均情報量上位10クラス、すなわち、様々な業種に満遍なく存在する根拠部が属するクラスについて、その重心ベクトルを組成する代表的な名詞の組を記す。

海外の事業環境の好調さ、人件費の増加など、幅広い銘柄において業績に影響を与える根拠情報を示唆しているクラスが並んでいる。

平均情報量下位10クラス、すなわち、特定の業種に偏って存在する根拠部が属するクラスについて、その重心ベクトルを組成する代表的な名詞の組と、そのクラスが偏在する業種を記す。

資源価格や為替前提の変更などは、特定の業種の銘柄の業績にのみ影響を与えやすい根拠情報といえる。一方、決算に関する根拠情報などは、後述するように、学習データに偏りがあるために不適切な特徴が獲得されてしまっている。また、金融業種が特徴的なクラスを構成する傾向があることが分かる。

表 1: 平均情報量上位 10 クラス。

クラスの重心ベクトルの組成
販売 法人税 低下 月 減収
事業 環境 航空 海外 好調
業績 好調 予想 足元 月
増加 費用 人件費 経費 研究開発費
継続 判断 投資 レーティング 決算
効果 これら 買収 税 連結
可能 業績 開発 達成 良
影響 ジェネリック 円高 為替 薬価改定
成長 中期 利益 期待 事業
計画 会社 中期 営業利益 営業増益

表 2: 平均情報量下位 10 クラス。

クラスの重心ベクトルの組成	偏在する業種
発表 取材 決算 材料 期待	金融
前提 為替 変更 下期 判断	機械
効率 系列 資本 高水準 比較	小売り
シナジー 統合 経費削減 システム	金融
銅 鉱山 金属 価格 下落	鉄鋼・非鉄
上昇 低下 基準 予想 不具合	公益
セクターバリュエーション 低下 上昇	機械
ヒアリング 決算 継続 月 急激	化学
投信 解約 外債 販売 回り	金融
自動車保険 収支 改善 等級 制度	金融

4.3 新聞記事からの業種別根拠情報の獲得

新聞記事をその文の各業種への業績関連度指数の降順で並べた重要記事のうち、化学業種についての2014年1月～3月の重要記事の見出しを記す。

大王紙——高値圏，海外事業が拡大
ナイロン原料が下落，アジア価格，中国で荷余り感，
苦戦の中国，15年黒字へ，花王，現地大手の販売網活用，設備投資，アジア中心
ベンゼン，3月4%安，アジア向け，中国で需給緩和。
ベンゼン上昇，10ヵ月ぶり高値，1月アジア向け。
ベンゼン価格，3ヵ月ぶり下落，アジア向け。
配合飼料3期ぶり上げ，4～6月価格，農家向け，原料高で。
古紙，輸出価格が下落，4月積み，需要鈍り9～10%安。
鉄スクラップ値下がり，市中価格，アジア需要伸びず，鋼材の値上げに弱材料。
価格差を読む(1) 発電燃料と電力—安い石炭利幅大きく，重油は採算割れ水準。

繊維，化学原料の資源価格の動向など，化学業種に属する銘柄の業績に関係があると考えられる根拠情報を含む記事が並んでいる。本手法によるトップ30における精度は70%，bag-of-words法によるトップ30における精度は57%だった。この業種は，因果関係の構造に注目して根拠部を分離したことで，特徴量の学習が適切に行われたといえる。

次に，自動車業種についての2014年1月～3月の重要記事の見出しを記す。

円高，一時100円台，NY市場。
為替市場，円高に備え，円買いオプション，半年ぶり高水準。
円相場の膠着続く，米景気期待・中国不安が綱引き。
円売り持ち高解消—服，投機筋，円高圧力緩和。
円高，じわり天井感，海外投機筋に実需の壁
円高103円台，東京市場。
円横ばい，102円33～34銭
投機筋，円売り膨らむ，6年半ぶり水準，日米の金利差拡大。
NY市場，円一時103円台。
為替——円安基調が一服か

主に為替に関する根拠情報が並んでいる。これは，自動車業種に属する銘柄のアナリストレポートでは，為替が主要な根拠情報として用いられているためと考えられる。特に，今回学習の対象としたアナリストレポートの発行時期が，いわゆるアベノミクスによる急激な円安の進行が注目されていた時期にあたるため，為替要因に偏って言及されており，その特徴を強くしたものと推測される。このように，新聞記事中に直接自動車に関係する名詞が無い場合でも根拠情報として獲得

できることが，因果関係を考慮した本手法の強みと言える。本手法によるトップ30における精度は100%，bag-of-words法によるトップ30における精度は83%だった。一方，為替のみで業績予想が行われるのはこの時期に特異な現象であることも想定される。すなわち，精度が高くても再現率が低い可能性がある。時期によらない根拠情報を獲得するためには，学習対象の期間を拡大する，アナリストレポート以外の時期によらない根拠情報を含む学習データと組み合わせるなどの工夫が必要である。

最後に，家庭用品業種についての2014年1月～3月の重要記事の見出しを記す。

デル，新興国で積極投資，中国など，ソフト販売を拡大。
ピジョン，ブラジル進出，育児用品，新興国投資を倍増，品質・ブランド生かす。
香港・味千中国——高値圏，和風ラーメン回復（アジア新興国NOW）
新興国——中国の景気動向に注目
中国，世界から投資1月16%増。
新興国——中国の製造業景況感に注目
欧州企業，鈍い収益回復，10～12月，新興国景気の減速響く。
中国蒙牛乳業——成長期待で戻り歩調に
G20閉幕，成長持続へ危機感共有——新興国の不満，抑え込む，先進国，中国は改革
香港株——改革にらみ神経質な展開

新興国に関する雑多な記事が混在しており，十分に適切な根拠情報が取得できていない。本手法によるトップ30における精度は53%，bag-of-words法によるトップ30における精度は63%だった。（例に挙げた3業種についてのトップ30における精度を表??にまとめる。）これは，学習の対象としたアナリストレポートにおいて，家庭用品業種に属する銘柄の業績予想には，主に新興国における当該銘柄の商品の販売動向が言及されており，このうち新興国を表す名詞がこの業種の特徴として過学習されてしまったためと推測される。また，一部の業種では，そもそも学習データが少ないために意味上のまとまりのある特徴を学習できていない場合もあった。このような学習の偏りや不足を防ぐためには，根拠情報として用いられなかった過去の新聞記事を負例として学習に用いる，決算短信など根拠情報を得られる他のデータを学習の対象に組み入れるなどの工夫が必要である。

また，本稿の段階では，膨大な数の新聞記事が各業種の根拠情報として適切であるかという正解データが未整備のため，得られた結果に対して十分定量的な評価ができていない。専門家の監修による正解データ

表 3: 業種別重要記事のトップ 30 における精度 (%).

手法	化学	自動車	家庭用品
本手法	70	100	53
bag-of-words 法	57	83	63

の作成とそれを用いた結果の再現率, F 値などによるさらなる定量評価は, 今後の課題とする.

5 まとめ

アナリストレポートの文章の因果関係の構造から, 証券アナリストが企業業績要因として注目する情報を業種別に学習し, 新聞から同様の因果関係を含む記事を抽出することで, 即時性の高い投資支援の実現を目指す構想を提示した. その初段として, 業種別に抽出した重要記事に対する簡単な評価・考察を行った.

結果, 一部の業種では因果関係を学習していないと抽出できない記事を適切に獲得できたが, 学習データの期間的な偏りや内容面での偏り, 量の不足により, 不適切な結果となった業種もあった. 学習データの質・量の向上, 結果のさらなる定量評価が今後の主な課題である.

参考文献

- [1] 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, 人工知能学会第 17 回金融情報学研究会, pp. 25–30 (2016)
- [2] 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)
- [3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean: Distributed Representations of Words and Phrases and their Compositionality, *NIPS 2013*, pp. 3111–3119 (2013)
- [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)

語の類義性・対義性を考慮した ドメイン特化型辞書構築手法の提案と評価

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 Ito¹ Sakaaji Hiroki¹ Kiyoshi Izumi¹ Shintaro Suda²

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

金融・経済分野における新たな分析データとして、非構造化データであるテキスト情報が注目を集めており、テキストマイニングを金融・経済分野に応用した研究が盛んに行われている。テキストマイニングには、従来では指標化されていなかった、市場や企業に関する情報をテキスト情報から抽出することが期待されているが、そのような抽出対象の情報の一つとして、ある事象に対してポジティブもしくはネガティブかを評価するセンチメント指数が挙げられる。そして、Bollen and Huina (2011) の研究にみられるように、テキストから抽出したセンチメント指数と、市場変動との関係性を検証する研究が多く行われている。

テキストからセンチメントを定量化する手法は、センチメント分析 (Sentiment Analysis) において、これまでに数多くの研究がなされている。センチメント分析には様々な手法が存在するが、大きく分けて Pang

et al. (2002) を始めとする機械学習のアプローチと、Turney (2002) を始めとする語彙ベースのアプローチに大別される。機械学習のアプローチでは、テキストの特徴量とポジティブやネガティブなどの極性ラベルとの関係性を、機械学習モデルによって学習し、未ラベルテキストに対して学習済みモデルを適用することで極性を付与する。一方、語彙ベースのアプローチでは、テキスト中に出現するポジティブな単語の出現比率とネガティブな単語の出現比率の差を以って、対象となるテキスト全体の極性を算出する方法が用いられる。センチメント分析に関するさらなる詳細は、Ravi and Ravi (2015) を参照されたい。

ここにおいて、語彙ベースによるアプローチを用いた場合、極性語とその極性値が組となった極性辞書が必要となるが、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。一例として、Loughran and McDonald (2011) は、語彙ベースによるアプローチにおいて広く用いられている H4N (Harvard-IV-4 TagNeg) に含まれるネガティブな単語が、ファイナンスの文脈においてネガティブな極性を有するとは限らず、ファイ

*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。

ナンス文書のセンチメント分析において、ファイナンス専用の辞書を用いる事の重要性を指摘している。以上のように、語彙ベースによるセンチメント分析のアプローチを用いた場合、網羅的にかつ解析対象に適した辞書を構築する事が必要であり、かつ辞書構築が自動でなされる事が期待される。

さて、解析対象となるドメインに特化した辞書自動構築を考えた際、対象ドメインのコーパスの知識を利用することが重要であると考えられる。また、一般にコーパス中に含まれる単語数の膨大さ故、シードとして与えるデータセットを構築するコストを削減することが重要であり、半教師あり学習による辞書構築が望ましいと考えられる。さらに、単語の類義・対義関係を列挙した、既存の知識ベースの情報を用いることで、より低コストで精度の高い辞書構築が行えると考える。

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、既存の辞書に対する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。

本研究の貢献は、以下の二点である。まず第一に、提案手法のモデルは先行研究の手法と比較して、より良い辞書構築精度を得る事が出来た点である。次に、提案手法のモデルを用いる事で、既存の辞書を拡張する事の有用性を示した点である。

2 関連研究

本章では、極性語を対象にした辞書構築に関する関連研究について述べる。

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。Turney(2002)は、2単語間の共起度合いを定量的に表す PMI (Pointwise Mutual Information) を用い、ある単語が種表現として与えられたポジティブな単語とネガティブな単語のどちらと共起しやすいかを以って、単語の極性を付与する方法を提案している。また、Hamilton et al. (2016)は、対象ドメインのコーパスから作成した、PPMI 行列の次元削減による単語

ベクトルの獲得と、ラベル拡散法によるドメイン特化の極性辞書構築手法を提案している。さらに、伊藤ら(2017)は単語の係り受け関係からなるネットワークの表現学習とブートストラップ法を用いた、極性語の抽出方法を提案している。しかしながら、伊藤ら(2017)の手法においては、極性語候補を抽出するのみで、ポジティブ・ネガティブに分離した辞書を構築出来ていない課題や、ブートストラップ法において生じる意味ドリフト問題を解消できていないという課題が存在する。

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。Kamps et al. (2004)は、英語概念辞書の WordNet¹を用いて種表現から極性値を伝搬させる方法を提案している。さらに、Rao and Ravichandran (2009)は、シソーラスから構成されたネットワーク上に対して、半教師あり学習のラベル伝播法を用いることで、極性語を獲得する手法を提案している。

シソーラスベースのアプローチでは、人手によって構造化された情報を入力とするため、一般にシソーラスベースのアプローチは、コーパスベースのアプローチと比較して、辞書構築の精度が優るという利点があるが、入力となるシソーラスがドメインに特化したものでなければ、ドメインに特化した極性辞書を構築する事が出来ないという欠点が存在する。

そこで、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する事で、この欠点を補った研究として、Alhothai and Hoey (2017)の研究がある。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。SNWELP モデルはシソーラスの知識を用いつつも、コーパスのドメイン知識を利用するため、精度良くドメイン特化型の辞書を構築出来るという利点がある。

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

¹<https://wordnet.princeton.edu/wordnet/>

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。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 \vec{u}_c , \vec{u}_d とした時、類似度を以下の式で表される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。

$$\text{sim}(\vec{u}_c, \vec{u}_d) = \frac{\vec{u}_c \cdot \vec{u}_d}{\|\vec{u}_c\| \|\vec{u}_d\|}$$

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 \mathbf{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。図 1 はラベル拡散法の概要を示したものである。ラベル拡散法は、少数のノードにシードとして付与された

ラベルの情報と、ネットワークの構造情報を用いて、未ラベルノードのラベルを推定する半教師あり学習の手法である。

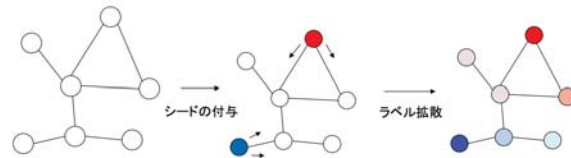


図 1: ラベル拡散法の概要

さて、 V を行列 \mathbf{E} の行数とした時、 \mathbf{p} を $\mathbf{p} \in \mathbb{R}^{|V|}$ 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、 $\mathbf{s} \in \mathbb{R}^{|V|}$ 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 $\bar{\mathbf{P}}^P(w_i)$ を求める。

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

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

Althouthai and Hoey (2017) の SNWELP モデルとの相違点は、行列 \mathbf{E} において、SNWELP モデルでは分散表現から得られた隣接行列 \mathbf{M} と、類義語グラフから得られた隣接行列 \mathbf{S} の平均を用いて、行列 \mathbf{E} を算出しているが、SMLS モデルでは、類義語グラフのみならず対義語グラフを考慮した上で、行列 \mathbf{E} を算出している点である。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、精度が向上することが期待される。

3.2 DLS モデル

次に、DLS モデルを提案する。SMLS モデルでは、コーパスから Skip-gram モデルを用いることで各単語の分散表現を学習し、各単語の分散表現を用いて k-近傍グラフを作成した後に、類義語ネットワーク・対義語ネットワークと合わせてラベル拡散法によって、単語の推定極性値を得ているが、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という違いがある。

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。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。

$$\begin{aligned} & \sum_{w \in V} \sum_{c \in V} \{ (\#(w, c) \log \sigma(\text{sim}(w, c)) \\ & + k \#(w) P_0(c) \log \sigma(-\text{sim}(w, c))) \\ & + \left(\frac{1}{\#(w, u)} \sum_{u \in W(c) \cap S(w)} \text{sim}(w, u) \right. \\ & \left. - \frac{1}{\#(w, v)} \sum_{v \in W(c) \cap A(w)} \text{sim}(w, v) \right) \} \end{aligned}$$

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、 $\text{sim}(w_1, w_2)$ 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、 $S(w) \cdot A(w)$ 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。

一般に、類義語・対義語同士はセンテンスにおいて置換可能な関係であることが多く、そのため対象単語から、周囲に出現する単語を予測する中で分散表現を学習する Skip-gram モデルでは、対義語同士の分散表現が類似する可能性が高いが、d-LCE 法では、シソーラスから抽出した単語の類義性・対義性に関する制約を、分散表現学習時に加えた上で学習を行うため、対義語同士の分散表現が類似する問題を解消することが可能となる。

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 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。

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²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。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

²<https://nlp.stanford.edu/pubs/stock-event.html>

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)³によって公開される議事録を用い、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。なお、FOMCの議事録は経済環境や金融市場の振り返りを行うレビューに相当するセクションと、将来の展望について書かれたフォワードガイダンスに相当するセクションに大別されるが、本実験ではこの内レビューに相当する部分を対象テキストとした。

そして、FOMC 議事録のレビューテキストをコーパスとして、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。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

図 3 は、各モデルにおける ROC 曲線を比較したものである。図から分かるように、偽陽性率の値が小さい部分において、SMLS モデルは真陽性率が相対的に高い結果となった。

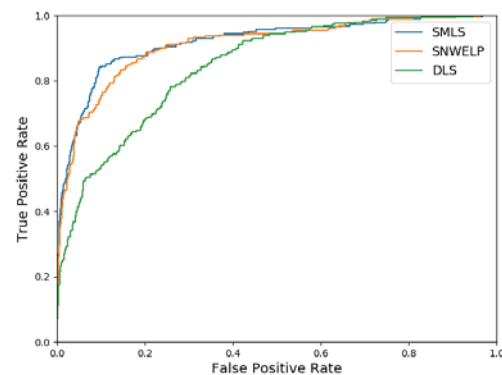


図 2: 各モデルにおける ROC 曲線の比較

³<https://www.federalreserve.gov>

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

表 5 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。

以下のセンテンスは、2004 年 12 月において発表された議事録レビュー部分のテキストを引用したものである。

Spending on services continued to register solid gains.

消費トピックには、solid gains のような言い回しが良く用いられるが、既存辞書には含まれなかった gain を、辞書拡張した結果ポジティブな極性語として含む事が出来たため、消費トピックに対する説明力が大きく向上したと考えられる。

このように、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

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、未拡張の場合の辞書と比較して、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、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。

本研究の課題として、辞書構築精度向上のために、コーパスの知識、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。極性語は特定の概念を評価するという点において、文法的に類似しているという特徴が存在する。よって、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。Omer Levy and Yoav Goldberg (2014) は、単語間の係り受け関係を用いて、単語の分散表現を得る Dependency Based Word Embeddings を提案しているが、本研究において、Dependency Based Word Embeddings を拡張し、先述の三種類の情報を同時に埋め込んだ分散表現を用い、辞書構築を行うことの有効性の検証は、今後の課題としたい。

参考文献

- [1] Alhothali, A., and Hoey, J. (2017), Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations, *arXiv preprint arXiv:1703.09825*.
- [2] Bollen, J., and Huina, M. (2011) Twitter mood as a stock market predictor, *Computer* **44**: 91-94.
- [3] Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora, *arXiv preprint arXiv:1606.02820*.
- [4] Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (2004) Using WordNet to Measure Semantic Orientations of Adjectives, *In Proceedings of the 4th International Conference on Language Resources and Evaluation*.
- [5] Levy, O., and Goldberg, Y. (2014) Dependency-Based Word Embeddings, *In ACL (2)*: 302-308.
- [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)**: 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*: 1170-1175.
- [8] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.
- [9] 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*.
- [10] Pang, B., Lee, L., and Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques, *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*: 79-86.
- [11] Rao, D., and Ravichandran, D. (2009) Semi-supervised polarity lexicon induction, *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.
- [12] Ravi, K., and Ravi, V. (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications, *Knowledge-Based Systems* **89**: 14-46.
- [13] Turney, P. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL'02, Association for Computational Linguistics, Stroudsburg*: 417-424.
- [14] 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*: 321-328.
- [15] 伊藤諒, 和泉潔, 須田真太郎 (2017) ネットワークの表現学習による金融専門極性辞書の構築, 第 31 回人工知能学会全国大会
- [16] 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: 60-71.

7 付録

表 5: 辞書拡張の有無によるトピック別極性値のマクロ指標に対する説明力の検証結果

	マクロ統計	GDP 成長率	鉱工業 生産指数	PCE	非農業部門 雇用者数	インフレ率	経常収支
辞書拡張無し	切片	2.7×10^0	1.9×10^0	4.6×10^0	1.5×10^6	2.2×10^0	-4.1×10^{11}
	t 値	26.05	8.72	38.17	16.46	31.42	-28.39
	係数	1.1×10^0	4.3×10^0	1.7×10^0	2.4×10^6	1.1×10^0	-2.0×10^{11}
	t 値	6.33	14.80	8.81	20.36	9.37	-5.53
	R^2	0.180	0.545	0.298	0.694	0.324	0.143
	切片	2.7×10^0	1.9×10^0	4.6×10^0	1.5×10^6	2.2×10^0	-4.2×10^{11}
辞書拡張有り	t 値	26.43	8.39	42.04	17.28	32.25	-29.01
	係数	1.2×10^0	4.1×10^0	1.9×10^0	2.4×10^6	1.1×10^0	-2.1×10^{11}
	t 値	6.91	13.71	11.57	21.86	10.17	-6.27
	R^2	0.207	0.507	0.422	0.723	0.361	0.177

FXにおける日足・4時間・15分のチャートのCNNによる最適化 Optimization of Daily, Weekly, and 15 Minute FX Chart by Using CNN

河合 継^{1*} 小澤 昂^{1,2} 大川 堯郁^{1,3}
Kei Kawai¹, Takashi Ozawa^{1,2}, Takafumi Okawa^{1,3}

¹ クリスタルメソッド株式会社

¹ Crystal method co.ltd

² 東京工業大学理学院数学系数学コース数理ファイナンス専攻

² Department of Mathematics, Tokyo Institute of Technology

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

Abstract: In recent years, many people attempt to apply machine learning to a broad range of fields. At the same time, thanks to improvements of the information technology, we can use the enormous number of the past financial data as a learning data. In this paper, we describe prediction results of the foreign exchange price. The prediction model was made by using Convolutional Neural Network(CNN). Then, we compare the results of the model which learned numeric data with the model which learned image data.

1 はじめに

近年では、人工知能分野の技術を金融市場の分析や予測に応用する動きが活発になっている。同時に、情報技術の発達により、これまでの膨大な金融市場のデータをインターネットを通じ、学習データとして用いることが出来るようになった。

前回の研究では、為替レート (USD/JPY) のチャート画像による 15 分後の値動き予測を行った。15 分足のチャート画像を用いて 2006 年のデータで学習を行い、2007 年から 2016 年のデータで検証を行った。検証では、年率 2% のリターンとなった。今回の研究では、より予測の精度を向上させるため、日足/4 時間足/15 分足のチャート画像を入力データとした。また、為替レートの時系列データ (数値データ) を入力データとした CNN による学習・検証も行い、画像データによる結果との比較・考察を行った。

2 手法

2.1 前提

実際にトレードを行う際には、日足、4 時間足、15 分足を情報として用いる事が多い。そのため、今回の研究では、それらの三つの情報を一つのデータにまとめ、学習・検証を行った。過去データの取り込みにおいては、学習データに未来のデータが混入しないように、日足は前日分、4 時間足は予測する 15 分足が含まれない時刻までの 4 時間足を学習データとして取り込んだ。また、現在の日付の途中までの未確定足の情報は、実際のトレードでは重要なシグナルとなるが、今回は 4 時間、15 分足にその情報が含まれていると考え、前日の日足を用いてデータ構築を行った。

今回の研究で用いたデータなどについては、下記の通り。

通貨ペア：USD/JPY、日足データ数：3115、4 時間足データ数：18956、15 分足データ数：304567、画像サイズ：892 × 500、チャート画像一つあたりの足本数：211、チャート数値データファイル数：199、利用フレームワーク：chainer、使用言語：python。

*連絡先：クリスタルメソッド株式会社
〒162-0844 東京都新宿区市谷八幡町 1 新高ビル 904
E-mail:kawai@crystal-method.com

2.2 準備

2.2.1 画像データ・数値データ

今回の検証においては、画像データ量が30万枚程度となるため、機械学習時のマシンパワーが必要になると考えられた。そのため、GPUを用いた学習環境として Nvidia GTX 1080 Ti, Nvidia GTX 1060 Ti を使用して計算を行った。学習用ハードウェアの構成は、それぞれ数値学習マシンでは GTX 1080 Ti 1 枚, GTX 1060 Ti 1 枚, 画像学習マシンでは GTX 1080 Ti 2 または 3 枚とした。¹

また、画像データ、数値データの共通の準備として、価格データを SQL データベース化し、検証時に問い合わせを行い、売買タイミングでの利益などのデータを抽出し、データベースには 15 分足のデータを入力した。また、当該 15 分足の終値で約定し、次足終値でクローズする仕組みとした。値幅に関しても同様に、当該足終値と次足の終値の差を用いて検証を行った。

今回の検証では、ラベルの振り方を 3 種類用意し、それぞれの振り分け方に対して学習と検証を行った。ラベルの振り方は、表 1 の様である。

表 1: ラベルの振り分け

ラベル	1	2	3
振り分け方 1	-10pip 10pip	10pip 以上	-10pip 以下
振り分け方 2	-15pip 15pip	15pip 以上	-15pip 以下
振り分け方 3	-20pip 20pip	20pip 以上	-20pip 以下

2.2.2 画像データ

- チャート画像の取得・加工プログラム

学習データに用いるためのチャート画像の取得には、MQL4, C++ によって作成したプログラムを用い、MetaTrader² のチャートをコマ送りにし、USD/JPY の 15 分足チャート画像をスクリーンショットとして 30 万枚分を png 形式で保存した。また、同様にして 15 分足の日付、時間に対応する日足、4 時間足のチャート画像を読み込んだ。その後、png 形式で保存した画像データに対し、ビット深度やサイズ変更を行う加工プログラムを作成、実行した。

日足、4 時間足、15 分足の取得データのサンプルは、それぞれ以下の図 1, 2, 3 に示す通りであ

る。また、画像はそれぞれ 8bit 画像として生成し、同時刻の日足、4 時間足、15 分足のサンプルを RGB 画像として統合する事で 1 枚の画像にまとめ、その画像を CNN で学習させた。(図 4)

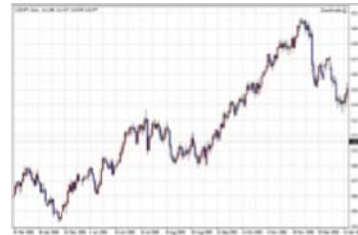


図 1: 日足



図 2: 4 時間



図 3: 15 分足

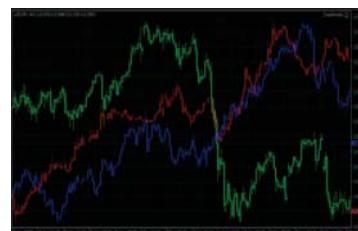


図 4: 日足 (青)・4 時間足 (緑)・15 分足 (赤) 統合データ

¹近年の GPU においては、PCI バスが 16 レーンの帯域となっているが、現実的には 8 レーン× 4 のマザーボードと速度があまり変わらないため、安価な 8 レーン× 4 のボードを選択した。この学習については CPU 性能もかかかってきており、コア数の多いものの学習がより早く進んだ。2008 年の 8 コアマシンと 2017 年の 8 コアマシンでは、計算時間に大きな差は見られなかった。

²外国為替証拠金取引プラットフォーム
<https://www.metatrader4.com/ja>

- Chainer CNN プログラム畳み込みニューラルネットワークを利用し、上記画像を学習するプログラムを Python により作成した。
- Chainer 学習結果返却プログラム
上記で学習した学習済みデータセットに、学習させていないデータを入力として与えることで売り、買い、または何もしないかを返却するプログラムを Python により作成した。
- チャート価格取得プログラム
あるチャート画像を対象として、チャートの開始時刻 (チャート左端)、チャートの現在時刻 (チャートの右端)、次の足の時刻 (チャートの右端+1)、及び右端の終値、右端+1 の終値、両者の価格差を全ての 15 分足で取得した。
- 結果検証プログラム
学習済みデータセットが返却した売却情報と上記のデータベースに対して SQL を発行し、Excel にまとめるプログラムを Python, Postgres により作成した。

2.2.3 数値データ

数値データに関しては、USD/JPY の FX チャートにおいて、日足・4 時間足・15 分足データをメタトレーダーにより CSV 形式で取得した。その後、CSV データをプログラムを用いて対象とする 15 分足、4 時間足、日足データをそれぞれ 200 本分をファイルにまとめた。出力データは、高安/始終ごとに差分と平均の計算を行った。数値データの例は、表 2 に示す通りである。

表 2: 数値データの例

	日付	時刻	終値 - 始値
日足	2016.03.21	00:00	40.756
4 時間足	2016.11.25	04:00	0.608
15 分足	2016.12.30	22:15	0.095
	高値 - 安値	(始値 + 終値)/2	(高値 + 安値)/2
日足	0.437	111.587	111.7165
4 時間足	0.398	113.555	113.555
15 分足	0.065	116.7725	116.7835

2.2.4 検証実施時のモデル

検証時のモデルには、数値・画像ともに CNN(畳み込みニューラルネット) を用いた。画像畳み込みニューラルネットに関しては、日足、4 時間足、15 分足をそれぞれ RGB にマッピングし、入力層の大きさは使用する画像データ数によって変更を加えた。数値データ、及び画像データを学習させた際のモデルは、それぞれ図 5、図 6 の通りである。³ 図 5 において、①を 4 回実行し、②を 2 回実行した。また、図 6 においては、①を 3 回実行し、②を 2 回実行した。

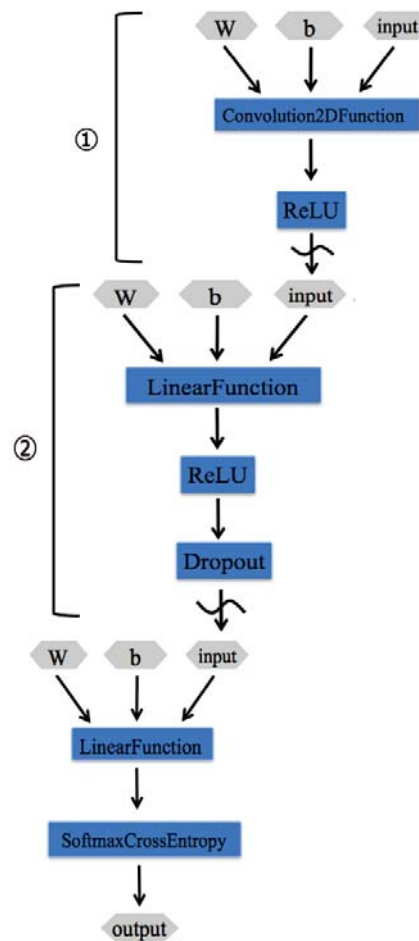


図 5: 数値データ学習モデル

³学習結果に偏りが生じてしまう現象が起こってしまった際には、学習する時の一つ目の中間層のフィルター枚数を入力する画像データの量に比例して変化させることによって、軽減することが出来た。

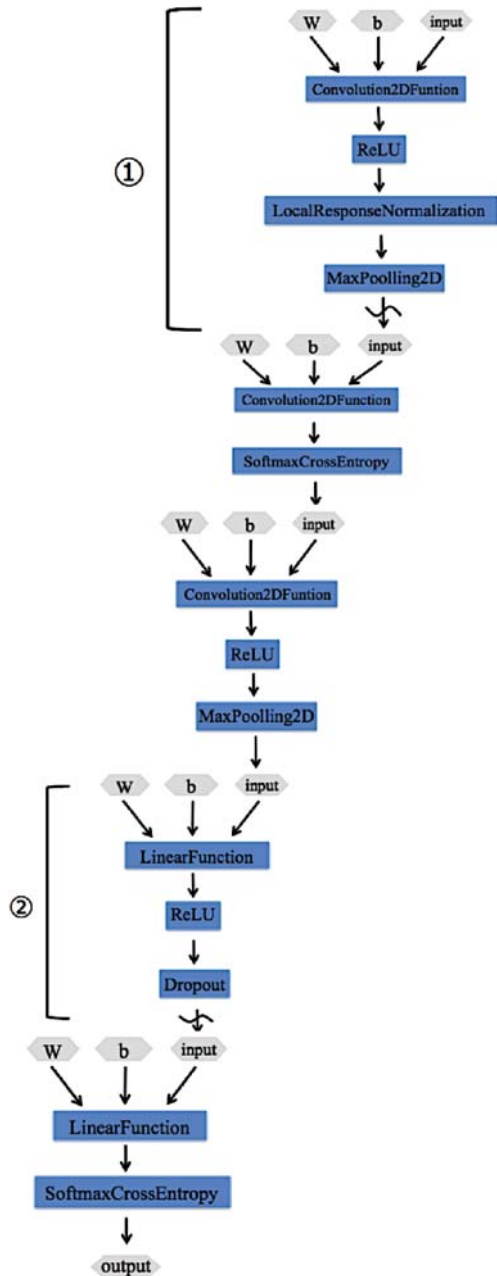


図 6: 画像データ学習モデル

2.3 検証方法

検証方法は、下表 3 に示すように、2006 年～2008 年を学習させたモデルについて学習させた年度以降の検証を行った。また、値動きの幅に応じて、3 つのパターンに対して検証を行った。

表 3: 学習・検証年度

学習年度	検証年度	勝敗の値幅	比較する点
2006～2008	2009～2015	10pip	勝敗の値幅 数値と画像の比較
2006～2008	2009～2015	15pip	勝敗の値幅 数値と画像の比較
2006～2008	2009～2015	20pip	勝敗の値幅 数値と画像の比較

3 結果

3.1 数値データによる結果

数値データを用いて、2006 年～2008 年までを学習させたモデルの結果を表 4 に示す。また、このモデルにおける損益グラフを、図 7～9 に示した。(縦軸：円)

表 4: 検証結果 (2006～2008 年を学習)

指標 / 2009-2016 を検証	10pip	15pip	20pip
Retuen(単利)	1.63%	-0.54%	1.84%
Retuen(複利)	1.56%	-0.60%	1.84%
Sharp Ratio(単利)	0.391	-0.164	0.872
Sharp Ratio(複利)	0.373	-0.18	0.869
勝率	50.5%	50.2%	50.8%
取引回数	26941	15507	4276

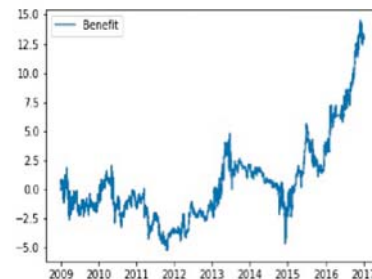


図 7: 数値データ損益 (10pips)

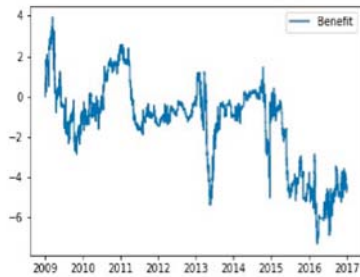


図 8: 数値データ損益 (15pips)

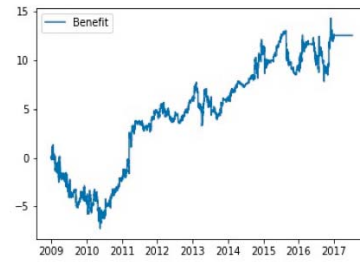


図 10: 画像データ損益 (10pips)

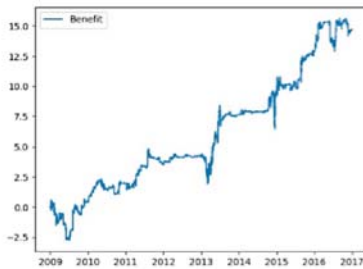


図 9: 数値データ損益 (20pips)

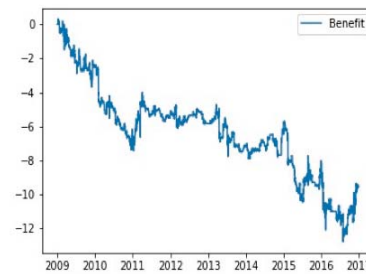


図 11: 画像データ損益 (15pips)

3.2 画像データによる結果

また、画像データを用いて、2006年～2008年までを学習させたモデルの結果を表5に示す。この時の損益グラフは図10～12の様であった。(縦軸：円)

画像データを用いた学習においては、2006年度の15分足のみを学習させた前回研究と比較するため、2006年の15分足、4時間足、日足を学習させた結果を表6、図13に示す。(縦軸：円)

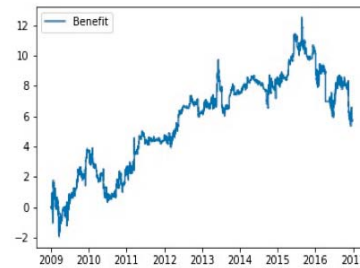


図 12: 画像データ損益 (20pips)

表 5: 検証結果 (2006～2008 年を学習)

指標 / 2009-2016 を検証	10pip	15pip	20pip
Retuen(単利)	1.63%	-0.54%	1.84%
Retuen(複利)	1.56%	-0.60%	1.84%
Sharp Ratio(単利)	0.391	-0.164	0.872
Sharp Ratio(複利)	0.373	-0.18	0.869
勝率	0.505	0.502	0.508
取引回数	26941	15507	4276

表 6: 検証結果 (2006 年を学習)

指標 / 検証年度	2007-2016
Retuen(単利)	1.016%
Retuen(複利)	1.076%
Sharp Ratio(単利)	0.283
Sharp Ratio(複利)	0.299
勝率	51.0%
取引回数	16879

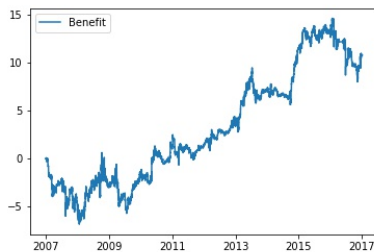


図 13: 画像データ損益 (2006 年を学習)

4 考察

前回の研究発表後に取得した、2006 年の 15 分足のみを学習データとして用いたモデルの損益は 1000pip, 年利 1%ほどの利益であったが、今回の日足情報と 4 時間足を 15 分足に加えて学習を行ったモデルの損益は、1500pip, 1.5%程度であり、改善が見られた。

また、画像データでの結果と数値データでの比較では、若干数値の結果が上回った。数値データ・画像データ共に、15pip のラベル付けにおいてマイナスとなった。原因は特定できておらず、どの場合においても勝率の有意性は見られなかった。これらを踏まえ、画像データについて、今後改善すべき点としては日足、4 時間足、15 分足を一つの画像としたが、それぞれのコンテキストを踏まえた学習となっていない可能性があるため、別画像として実施したい。また、数値データについては、一部複数回実行した場合、結果が全く同じにならなかったことがあったため、これらの原因特定はこれから実施したいと考えている。更に、経済指標をモデルを用いることによって、より良いモデルを構築出来る可能性があるため、今後研究を進めたいと考えている。

また、データ量などは数値データの方が少なく学習を行うことができた。また、それぞれの学習データを用いた際のメリット・デメリットは以下のようなものが考えられた。

数値データを使用するメリット

- より正確な情報を用いることができる。
- 計算量が格段に減る。

数値データを使用するデメリット

- CNN は画像データについて解析を行うために開発されている歴史がありの本来の目的を果たせない。

画像データを使用するメリット

- 人間が目で見たま取引を行うという目的を果たすためには CNN モデルが必要で、画像データは CNN モデルの目的に合致している。

画像データを使用するデメリット

- 入力の次元が多次元になり、計算に負担がかかる。

謝辞

本研究は人工知能学会金融情報学研究会の先生方に多大なご指導、ご鞭撻を頂き完成したもので、心より感謝申し上げます。また、共同研究をしている学生のお二人をご指導されている先生方、大学関係者の方々もこのように優秀な人材を育てられていることに対して、非常に強い感謝の念を抱いております。

参考文献

- [1] ジョン・コーツ, 訳 小野木明恵:トレーダーの生理学, 早川書房, (2013)
- [2] 新納 浩幸: Chainer による実践深層学習, オーム社, (2016)
- [3] ラリー・ペサベント: フィボナッチ逆張り売買法, パンローリング, (2008)
- [4] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner: Gradient-Based Learning Applied to Document Recognition, *Proceedings of the IEEE*, 86(11):2278-2324(1998)

外国為替市場の個々のトレーダのトラッキング解析： ミクロモデルの提案とその平均場理論

Analysis of individual traders' strategies in a foreign exchange market: microscopic modeling and corresponding mean-field theory

金澤輝代士^{1,2*} 末重拓己² 高安秀樹^{1,3} 高安美佐子^{1,2}

¹ 東京工業大学 科学技術創成研究院

² 東京工業大学 情報理工学院

³ Sony Computer Science Laboratories

Abstract: Recent technological breakthrough has enabled us to study the microstructure of financial markets using the high-frequency trading data. In this presentation, we review our recent preprint (arXiv: 1703.06739), in which individual traders's strategies are analyzed on the basis of informative order book data with anonymized trader identifications. We empirically study the trend-following behavior of individual traders on the basis of conditional statistical analysis. We then propose a microscopic model of financial markets on the basis of the empirical finding of trend-following of individual traders. We further develop a systematic theory to our microscopic model paralleling to the mathematical formulation of kinetic theory. Finally, the agreement between empirical results and our theoretical predictions are shown in terms of the order book profile and the price movement distribution.

1 導入：外国為替市場のデータ解析

近年社会現象の高頻度データが豊富に集まるようになり、社会科学分野で注目を浴びている。例えば、金融市場では価格時系列のみならず、オーダーブックデータと呼ばれる高頻度データが入手可能になった。オーダーブックデータは最終的な取引に至るまでのトレーダー達のやり取り（指値注文）を記載したデータであり、価格時系列よりもはるかに多い情報が含まれている。事実、オーダーブックの一部には将来の価格形成と強相関をもつ部分が存在することがデータからも示されている [1, 2]。また、オーダーブックを理論的にモデル化することも盛んに行われており、単なる時系列モデルを超えたより精密なモデル化の方向として着目を浴びている。

一方、より細かいスケールとしては個々のトレーダーの戦略をモデル化する方向性も考えられる。事実エージェントベースモデルの文脈では、Lux-Marchesi モデルのように個々のトレーダーの戦略の多様性を反映させたミクロモデルなどが提案されており [3]、価格時系列レベルのマクロな性質と整合的な結果を再現するこ

とに成功している。しかし、既存のミクロモデルの殆どは純理論的に構築されており、そのモデルの仮定は実証的に検証されていなかった。これはトレーダーのアカウントレベルのデータはなかなか入手が困難であり、長らく実証が困難だったことに起因する。ところが近年、高速トレーダー（HFT）が台頭することに伴い、トレーダーのミクロな性質の実証研究が少しずつ行われ始めている [4]。本講演では特に、外国為替市場における個々のトレーダーの行動分析・その理論的枠組みに関する研究結果を文献 [5] に基づいて報告する。特に、単なる個々のトレーダーの戦略分析のみにとどまらず、そういったミクロ構造からよりマクロな市場の構造（e.g., オーダーブックの形状・価格変動の分布）をも体系的に数理的手法を用いて説明することを試みた。

2 研究結果

我々は NEX 社が提供するインターバンク市場 Electronic Broking Services (EBS) のデータ解析を行い、個々のトレーダーの行動を直接トラッキングしながら戦略分析を行った。今回解析に使用したデータは、2016 年 6 月 5 日から 6 月 10 日の週についての円ドル市場のデータであり、全てのトレーダーの指値注文（新規注

*連絡先：東京工業大学科学技術創成研究院 ビッグデータ数理科学研究ユニット
〒 226-8502 神奈川県横浜市緑区長津田町 4259 S1-3
E-mail: kanazawa.k.ae@m.titech.ac.jp

文・キャンセル注文)・成行注文の行動記録が匿名化されたトレーダー ID とともに記載されている。特に我々は HFT の指値注文に着目を行い、過去の価格時系列(約定値)に応じて個々の HFT の指値注文がどのように変化するかを条件付き統計の手法を用いて解析した。その結果、市場でも特に高速に反応する HFT は、過去の短時間での価格変化と正の相関を持ちながら指値注文を動かす傾向があることがわかった。更に、その移動量の平均値 $\langle \Delta z_i \rangle$ と直近の市場の価格差 Δp との間には普遍的な関係があり、ある単一のスケーリング関数(双曲線正接関数)でフィットできることがわかった。この結果は、HFT は概ねトレンド・フォローの戦略を採用しており、弱いトレンドには線形にトレンド・フォローするが、強いトレンドにはあまり反応しない戦略を採用していると解釈できる。

次に我々は、観測されたトレンド・フォローに関する経験則を取り入れることによって新たなマイクロモデルを理論的に提案した。このモデルは文献 [6, 7] を実証データに基づいて改良したモデルである。更にこのモデルを分子運動論 [8] で発達してきた平均場近似の手法を用いて、解析的に解くことに成功した。分子運動論では多体系の運動方程式を出発点に、解析力学の Liouville 方程式を経由して、少数分布関数についての Bogoliubov-Born-Green-Kirkwood-Yvon (BBGKY) 階層方程式を導く。更に BBGKY 階層方程式に平均場近似 (i.e., 分子カオスの仮定) を施すことにより、Boltzmann 方程式に到達する。我々は同種の数理テクニックを用いることで、外国為替市場のオーダーブックを記述する Boltzmann-like な方程式を導出した。更にトレーダーの人数 N が大きい極限での漸近展開を行い、解析解を導出した。その結果オーダーブックの平均的形状をフィッティングパラメータなしで説明することに成功した。更に、市場での約定価格の変化が短期では指数分布・長期ではべき分布に従うことを定性的に説明した。

外国為替市場での価格変化は長期スケールではべき分布に従うことが知られているが、我々はデータ解析を通じて、少なくとも円ドル市場では短期スケールでは指数分布に従っており、長期スケールではその重ね合わせとしてべき分布に見えることを実証的に示した。特にべき分布の裾野となるイベントが発生する時間帯は、市場が不活発でトレーダーが殆どいない時間帯とオーバーラップすることを示した。以上を通じて、個々のトレーダーの注文行動を出発点に、オーダーブックの形状・価格変動の分布といったよりマクロな性質を体系的に説明した。

3 今後の展望

現在新たに進めている研究結果・今後の展望についても、講演の時間があれば報告する予定である。具体的には次のような内容を考えている：文献 [5] では主に HFT の指値注文の超短期の時間スケールに焦点を当てて調査を行ったが、どの程度の時間スケールでトレーダーが反応を起こすかなどの、時間スケールの多様性はまだ調べられていない。今後の課題として個々のトレーダーが反応する典型的な時間スケールの多様性を研究する予定である。また、文献 [5] では指値注文に焦点を当てて研究を行ったが、次は成行注文についての統計的性質などを調べる必要があると考えている。例えば、文献 [9] では成行注文に逆張りのトレーダーが存在していることが示唆されているが、この市場のデータセットでも同様のトレーダーが存在するか、または違う傾向が見られるのか、などを体系的に調査することも課題である。

謝辞

本研究を遂行するにあたり、香取眞理氏、早川尚男氏、一木信吾氏、柳澤大地氏、山田健太氏、小川駿氏、F. van Wijland 氏、D. Sornette 氏、佐野雅己氏、佐野友彦氏、伊藤隆敏氏との議論が参考になりました。ここで謝辞させていただきます。また、本研究は科学研究費補助金若手研究 B (No. 16K16016)、学術振興会特別研究員研究奨励費 (No. 17J10781)、戦略的国際共同研究プログラム (SICORP) からサポートを受けていることを明記させていただきます。

参考文献

- [1] Y. Yura, H. Takayasu, D. Sornette, and M. Takayasu, *Phys. Rev. Lett.* **112**, 098703 (2014).
- [2] Y. Yura, H. Takayasu, D. Sornette, and M. Takayasu, *Phys. Rev. E* **92**, 042811 (2015).
- [3] F. Slanina, *Essentials of Econophysics Modelling* (Oxford, 2014).
- [4] A.J. Menkveld, *J. Financ. Mark.* **16**, 712 (2013); *Annu. Rev. Financ. Econ.* **8**, 1 (2016).
- [5] K. Kanazawa, T. Sueshige, M. Takayasu, H. Takayasu, ArXiv: 1703.06739 (2017).
- [6] H. Takayasu, H. Miura, T. Hirabayashi, and K. Hamada, *Physica A* **184**, 127 (1992).

- [7] K. Yamada, H. Takayasu, T. Ito, and M. Takayasu, *Phys. Rev. E* **79**, 051120 (2009).
- [8] S. Chapman and T.G. Cowling, *The Mathematical Theory of Non-Uniform Gases* (Cambridge University Press, Cambridge, 1970).
- [9] J.A. Brogaard, *High Frequency Trading and Its Impact on Market Quality* (Northwestern University, 2010).

ファンダメンタルファクターモデル（リターンモデル）における機械学習手法の応用可能性検証

Fundamental Factor Models using Machine Learning

杉友 盛佑*¹ 南 正太郎*²
Seisuke Sugitomo Shotaro Minami

*¹エピックパートナーズインベストメント株式会社
Epic Partners Investments Co., Ltd

*²あすかアセットマネジメント株式会社
Asuka Asset Management Co., Ltd

Fundamental factor models are one of the important methods for the quantitative active investors (Quants), so many investors and researchers use fundamental factor models in their work. But often we come up against the problem that highly effective factors do not aid in our portfolio performance. We think one of the reasons why is that the traditional method is based on multiple linear regression. Therefore in this paper, we tried to apply our machine learning methods to fundamental factor models as the return model. The results show that applying machine learning methods yield good portfolio performance and effectiveness more than the traditional methods.

1. はじめに

計量的アクティブ運用（クオンツ）の代表的なツールにマルチファクターモデルがある。これは、株式の投資収益率を複数のファクターで説明するものである。ファイナンス分野における一般的なマルチファクターモデルは、1976年にステファン・ロスにより提唱された裁定価格理論（APT: Arbitrage Pricing Theory）と同義に使用されることがあるが、運用実務においては、CAPMをベースとしたBARRA型アプローチやFama-French型アプローチのマルチファクターモデルも広く用いられている。マクロ経済変数を先験的に与えて、個別企業の株式収益率を求める方法や、過去の株式収益率から因子分析によって因子を引き出す方法などは、APT型のマルチファクターモデルに分類される。一方、PERやPBRに代表される投資指標のような個別企業の株式が持つ銘柄属性を用いて個別企業の株式収益率を求める方法は、BARRA型またはFama-French型のマルチファクターモデルに分類される。Barra型アプローチはBar Rosenbergにより紹介され、Grinold and Kahn (2000)、Conner et al (2010)により拡張されたものである。ある1時点における株式銘柄群の株式収益率は共通のファクターによって説明されるとすることから、クロスセクション回帰分析によって計算される。またFama-French型は、Eugene Fama and Kenneth French (1992)によって初めて紹介された。本稿はFama-French型のマルチファクターモデルに分類されるものである。

もう一つ、マルチファクターモデルを用いるに当たって明確にしておかなければならない事項がある。

個別企業の株式収益率を銘柄属性で説明するマルチファクターモデルの利用方法として通常2種類考えられる。1つは、銘柄属性の寄与度から相場の傾向を判断して、将来の株式収益率を計算するために利用する方法（リターンモデル）。もう一つは、銘柄属性の寄与度をマーケット・リターンで時系列回帰してファンダメンタル・ベータを求め、ポートフォリオの属性を分解する方法（リスクモデル）である。本稿においては、リ

ターンモデルを想定している。

以上が分析するうえで押さえておかなければならない前提である。これらは、ファイナンス分野でも混同して議論されるケースがある。本稿の目的は、計量的アクティブ運用における機械学習手法の応用可能性を探ることであり、今後の計量アクティブ運用の分野への発展に寄与することであることから前提を明確にしておく。

本稿におけるマルチファクターモデル（以下ファンダメンタルファクターモデルと呼ぶ）を以下で定義する。

$$R_{it} = \sum_{j=1}^k X_{ijt} f_{jt} + \epsilon_{it} \quad (1)$$

ただし、 R_{it} は t 期における企業 i の株式収益率、 X_{ijt} は t 期における企業 i の第 j ファクターのファクター・エクスポージャー、 f_{jt} は t 期における第 j ファクターのファクター・リターン、 ϵ_{it} はファクターでは説明できない誤差

個別企業の将来の株式収益率を複数の銘柄属性（ファクター）による重回帰分析により算出するモデルである。このモデルの下では株式収益率とファクターの間における関係は線形であるが、金融市場の複雑性を考えると非線形を仮定するほうがより適切に関係性を表現できると考えられる。そこで、本稿では、非線形関係を表現できる機械学習の代表的な手法（サポートベクターマシン、勾配ブースティング、ニューラルネットワーク）を用いて、従来の線形モデルと比較し、非線形手法の実務運用における有効性と応用可能性を検証する。

2. 関連研究と基礎的概念

2.1 関連研究

[Joseph 14] は、PBRのファクター・リターンを予測するために、VIX、VIXの1か月変化率、PBRの分布、PBRの分布の変化率、1か月前のPBRのファクターリターンを変数とし、パラメータの推定に縮小推定法（Shrinkage Method）を用いたロジスティック回帰分析を用いた検証を行っている。また、プライスマンタムについても同様の分析を行っている。

連絡先: 杉友 盛佑 エピックパートナーズインベストメント株式会社 sugitomo@epicgroup.jp

ともに、翌期の予測精度が大幅に高くなっている結果を得ている。他の機械学習手法として、CART (Classification and Regression tree) を使った検証もしており、時期によってはロジスティック回帰分析よりも有効性が見られることを示した。

上記の分析は、ファクター・リターンの時系列予測に他の変数を用いて非線形の機械学習手法の適用を試みたものであるが、リターン・モデルについて適用したものではない。リターン・モデルでは、ファクター・リターンの予測値 (ファクター・ウェイトと呼ぶ) と直近のファクター・エクスポージャーから、将来の予測リターンが計算され、これを予測アルファと呼ぶが、上記の研究は予測モデルのマルチファクターに言及したものではない点で我々の研究とは異なる。

以降では、分析に用いた機械学習分析手法の基礎的概念について整理する。

2.2 サポートベクター回帰

サポートベクター回帰は非線形回帰の一手法であり、説明変数の特徴空間への非線形写像を考え、特徴空間において線形回帰を行うという手法である。本研究で用いる ϵ -SVR は、線形回帰の時、 ϵ -insentive 損失関数 $L^\epsilon(x, y, f)$ を使って線形関数 $y = f(x) = \langle w \cdot x \rangle + b$ を推定する。

$$L^\epsilon(x, y, f) = \max(0, |y - f(x)| - \epsilon) \quad (2)$$

なお予測値が実測値を上回る場合は $L^\epsilon(x, y, f) = \xi$ とし、予測値が実測値を下回っている場合は $L^\epsilon(x, y, f) = \hat{\xi}$ で表現する。最終的には以下の主問題を解くことになる。

$$\min_{w, L, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l L^\epsilon(x_i, y_i, f), \quad (3)$$

$$s.t. (\langle w \cdot x_i \rangle + b) - y_i \leq \epsilon + \xi_i \quad (4)$$

$$y_i - (\langle w \cdot x_i \rangle + b) \leq \epsilon + \hat{\xi}_i \quad (5)$$

$$\xi_i, \hat{\xi}_i \geq 0, i = 1, 2, \dots, l \quad (6)$$

ここで、上記の式における元データ x を非線形関数 $\phi(x)$ によって高次元空間に写像することにより、非線形回帰へと拡張できる。上記の主問題は実際には双対定理を使い双対問題として解くことができる。この時双対問題中に $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ と内積計算が現れる。一般的に高次元空間で内積計算を直接行うことはとても煩雑であり、この内積計算を行うことができるカーネル関数 $K(x_i, x_j)$ を適用する。本論文では、カーネル関数として、ガウシアンカーネル $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2})$ を用いる。

2.3 決定木、勾配ブースティング

取り得る選択肢や起こり得るシナリオを樹形図の形で洗い出し、それぞれの選択肢の期待値を比較した上で意思決定を行う手法が決定木であり、金融業界やコンサルティング業界においてよく用いられる手法である。本研究では、より精緻なモデルを構築するため、アンサンブル学習の一つである勾配ブースティングという手法と決定木を組み合わせる。勾配ブースティングとは、学習データから復元抽出を繰り返し複数のデータセットを作成し、それぞれに対して弱学習器を作り、全ての弱学習器の解で多数決を取り最終解を求める方法である。弱学習器を作成する際に、前回作成した弱学習器の結果を利用して、誤分類された値の重みを大きくするように更新す

る。この重みづけの際に勾配降下法を用いるのが、勾配ブースティングである。本研究では、この弱学習器に決定木を用いた、GBDT(Gradient Boosting Decision Tree) をモデルとして用いる。

2.4 ニューラルネットワーク

ニューラルネットワークには、階層型ニューラルネットワークと非階層型と呼ばれるものがあるが、今回用いる手法は、階層型ニューラルネットワークである。階層型ニューラルネットワークとは、入力層、中間層、出力層をもつネットワークである。入力層に説明変数を取り、中間層でそれらの説明変数に対してランダムに重みづけを与え、その結果が出力層である目的変数に近づくように、最適な重みを計算していく。あたかも人間の脳の中の神経細胞であるニューロンが、多数の他の神経細胞から信号を受け取り、意思決定を行っていることと類似していることからニューラルネットワークと呼ばれている。

2.5 実装

筆者らが構築した株式分析システムに基づいている。データは FactSet より取得、基本となるデータベース・計算システムおよびファンダメンタルファクターモデルについては Python、機械学習の計算については R と機械学習パッケージである 'nnet' を用いた。

3. 検証

3.1 検証における考え方

通常、ファンダメンタル・ファクターモデルにて算出した将来の株式収益率が実運用で活用できるかどうかについては、ファクター・リターンの有効性と安定性を検証しなければならない。しかしながら、機械学習手法 (サポートベクターマシン、勾配ブースティング、ニューラルネットワーク) についてはファクター・リターンに該当する係数を特定するのは困難である。そのため次のような検証を行う。まず、機械学習手法の応用の結果求められる将来の株式収益率を個別銘柄対比で大きい順にソートし 5 分位のグループに分ける。最も大きいグループをロングポートフォリオ、最も小さいグループをショートポートフォリオとし、それぞれ翌期のポートフォリオリターンを計測する。その後、ロングポートフォリオとショートポートフォリオの差分 (スプレッド・リターン) を計算。分析期間にわたって繰り返し計算を行い、従来の線形モデルと機械学習手法の結果を比較する。もし、将来の収益率の予測力が高いほど、実現した将来の株式収益率と予測した収益率の差は小さくなり、ロング・ショートポートフォリオの累積リターンは大きくなるはずである。なお、実証するにあたり、実現値と予測値を用いて RMSE および MAE も併せて計算している。

3.2 検証手順

ユニバースは、TOPIX500 構成銘柄とした。使用するファクターについては一般的によく使われる投資指標である PER、PBR、ROE、対数時価総額、3ヵ月 β の 5 つとする。ただし、PER と PBR については逆数に変換している。分析期間は 2000 年 1 月末から 2017 年 6 月までである。ポートフォリオについては毎月リバランスによりリターンを計測している。なお、個別銘柄の所属する業種によるファクター値の水準の違いによ

る影響を除くため、まず東証 33 業種内で基準化 (z-score) し、その後、再度 TOPIX500 全体で基準化した。

比較基準となる重回帰モデルについては、過去 1 年間 (t 期よりも前) の各銘柄の月末ファクター値を説明変数、翌月末 (t 期) リターンを目的変数として回帰係数 (ファクター・リターン) を算出。得られた回帰係数 (ファクター・リターン) をテスト期間における月末の各銘柄のファクター値 (ファクター・ウェイト) にかけることで、各銘柄の将来の株式収益率 (期待リターン) を算出している。先の検証についての考え方に従って、得られた期待リターンをもとに作った 5 分位ポートフォリオの 1 分位をロング、5 分位をショートし、翌月まで保有した時のロングポートフォリオとショートポートフォリオのリターン差 (スプレッド・リターン) を求めている。これを月次でローリングし得られたリターンを累積した。比較対象である機械学習手法 (サポートベクターマシン、勾配ブースティング、ニューラルネットワーク) についても分析のプロセスは同様である。期待リターンを計算する際に、機械学習手法を用いている。

有効性を検証するために、ポートフォリオのパフォーマンスの比較および予測精度の 2 つのパターンを示す。なお、ポートフォリオパフォーマンスについては、月次平均リターン、月次標準偏差、およびシャープレシオ。予測精度として、計算期間に渡り累積した RMSE と MAE である。

3.3 検証結果

	重回帰分析	GBDT	SVM	NN
平均リターン	0.10	0.09	0.12	0.18
標準偏差	2.83	2.86	2.89	2.72
シャープレシオ	0.037	0.034	0.042	0.068
累積 RMSE	1832.17	1831.61	1823.53	1832.92
累積 MAE	1428.71	1428.14	1418.28	1428.39

表 1: 実験結果

ポートフォリオのパフォーマンスの比較では、月次平均リターンはニューラルネットワークによるモデルが最も高い結果となった。ボラティリティを考慮したシャープレシオについても同様の結果である。予測精度の比較では、実績値と予測値から求めた RMSE と MAE の累積値はともに SVM が最も良い結果となった。GBDT についてはポートフォリオのパフォーマンスについては重回帰分析に劣るものの、累積 RMSE、累積 MAE については精度が高い結果となった。

4. 結論

本研究では、ファンダメンタルファクターモデルに着目しリターン予測の精度について、従来の重回帰分析に加え、GBDT、SVM、ニューラルネットワークを適用し比較した。結果として、当てはまりの精度である累積 MAE は全ての非線形モデルで改善が見られ、累積 RMSE、月次平均リターンや月次シャープレシオにおいても、多くのモデルで改善が見られた。このことは、金融市場における銘柄のリターンとファクター値間の関係は従来の線形関係ではなく、非線形関係が存在すると考えられ、そのような非線形関係を捉えられるモデルが、従来のモデルに対し優位性を持つと考えられる。

本研究の今後の展望として、非線形分析は実運用においても重要な意味がある。例えば、ファクターにベットしている

大規模ファンドやスマートベータと呼ばれるファンドは、ティルトしているファクターの有効性は見られるもののファンドパフォーマンスがそぐわないことが良くある。これは、従来のマルチファクターモデルは線形モデルであるために、本来非線形で評価すべきものを線形評価することによるズレが起因している可能性がある。また、クロスセクション回帰分析への応用 (BARRA 型) も期待されるだろう。もしくはクロスセクションと時系列回帰分析の両方を考慮した多変量回帰分析への応用も興味深い結果が得られるかもしれない。計量アクティブ運用の分野はそれだけでも奥深い研究対象であることから、基本的な分析手法や考え方を押さえた上で、機械学習手法を応用するのは簡単ではない。しかしながら、実運用において広く使われる考え方でもあることに加え、計量アクティブ運用と機械学習手法の相性はよい側面もあることを考えれば、今後あらゆる角度からの応用が求められる。課題にしたい。

5. 留意事項

本論文は筆者らが所属するエピックパートナーズインベストメンツ株式会社及びあすか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あります。

参考文献

- [南 17] 南正太郎, 光定洋介: AI の資産運用への応用の可能性と限界, 証券アナリストジャーナル, Vol.55, 8 月号 (2017).
- [吉野 03] 吉野貴晶: 株式投資のための定量分析入門, 日本経済新聞社, (2013).
- [T.DANIEL 95] T.DANIEL COGGIN AND FRANK J. FABOZZI: The Handbook of Equity Style Management, Addison-Wesley, 2nd edition (1998) (邦訳: 資産運用時代新時代の株式投資スタイル, 野村総合研究所 訳, 野村総合研究所発刊, (1996)).
- [Eric 11] Eric Zivot: Factor Models for Asset Returns, University of Washington (2011).
- [Argimiro 14] Argimiro Arratia: Computational Finance An Intoductory Course with R, ATLANTIS PRESS (2014).
- [Joseph 14] Joseph Mezrich: Factor forecasting with machine learning, Nomura Equity Research Report (2014).
- [Ludwig 06] Ludwig B Chincarini, Daehwan Kim: Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management, McGraw-Hill Library of Investment and Finance (2006).
- [Richard 99] Richard C. Grinold, Ronald N. Kahn: Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk, McGraw-Hill Library of Investment and Finance (1999).

深層学習と高頻度注文情報による株価動向推定

田代 大悟^{1*} 和泉 潔¹
Daigo Tashiro¹ Kiyoshi Izumi¹

¹ 東京大学大学院工学系研究科
¹ Graduate School of Engineering, The University of Tokyo

Abstract:

In this paper, we propose order-based approach to predict future movements of a stock price. Our models employ a convolutional neural network(CNN) over embedded orders that have quantitative and qualitative variables. For each dataset of stock codes, the models outperform traditional feature-based approaches. Furthermore, we show that training under less influence of noise can be performed by applying an averaging filter to embedded feature space. Analysis of the embedding layer reveals that the models put emphasis on the features of market orders that are correlated with price return.

1 はじめに

アルゴリズム・トレード(以下アルゴ)とは、機関投資家などから注文を委託された証券会社が、小口化し売買を行う機械的取引方法である。アルゴはコンピュータを利用して自動的に発注と売買を高速に行う点で High Frequency Trading(HFT) と共通するが、HFT がマイクロ秒オーダーで高速かつ高頻度で小口売買で利鞘を稼ぐ一方、アルゴは比較的低頻度にかつ取引コストの低減することに主眼を置いて執行を行う。取引コストをいかに低減するかは、アルゴにおいて重大な課題であり、高度なアルゴリズム開発のために、多様な分野の学術研究や先端技術が応用されている [2]。なかでも、アルゴの主な執行戦略である VWAP (volume-weighted average price) 戦略 [1] と深層学習による価格予測を併用することにより、パフォーマンスを向上したとの金融機関からの報告があり、本研究でも VWAP の支援となるような価格予測モデルの構築を目指す。

前述の通り機械的な売買方法の台頭や金融市場の電子化と高速化に伴い、蓄積される注文データはサンプリング頻度が極めて高く、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。市場価格は、トレーダーが出した注文を取引所が集計、約定処理の後に形成される。そのため、注文データには価格時系列データよりも多くの情報が含まれていると考えられる。しかし、この高頻度データを、タイムドリブンで執行を行うアルゴへの応用を考えた場合、不等間隔に観測される注文、取引をうまく

扱う必要がある。

ここでは、そのような高頻度データに機械学習を応用した例を紹介する。Alec らは、板情報のある時点での仲値やビッドアスクスプレッドなどの特徴量を人手によって設計し、SVM(Support Vector Machine) による仲値の変動の予測を行っている [3]。また、特徴選択をしていない生の板情報の入力と、ニューラルネットワークを用いた複雑な関数近似による予測も行われている。Tsantekidis らは、板に累積した注文の価格と量を入力として、LSTM(Long Short-Term Memory) による予測を行い、SVM を大きく上回る結果を得ている [4]。

板情報だけを説明変数に用いる場合、時点々々でのアスクとビッドの強度や均衡を動的に辿ることができる。しかし、ベストアスクまたはベストビッドの量が減少した場合、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判別がつかないというデメリットがある。成行注文とは一般的に即時約定する注文でトレーダーの強い意思を表したものである。さらに、成行注文とリターンに相関がある [5][6] ため、キャンセル注文のもつ意味、情報とは異なると考えられる。

そこで本稿では、前述の課題を克服するため、注文ベースでの株価の動向予測について説明する。2 章では、提案手法となる、注文を記号列に変換する前処理、CNN(Convolutional Neural-Network)、埋め込み特徴行列の平均化を用いた CNN について説明し、3,4 章で具体的な実験と結果考察を行う。

*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: m2016dtashiro@socsim.org

2 CNNによる短期価格動向予測

ここでは初めに、注文の記号化について説明する。2.2では、注文系列から、タイムドリブンに対応する特徴へ変換することを意識し、問題を設定する。最後に、CNNによる価格動向予測モデル、埋め込み特徴行列の平均化を用いたCNNによる価格動向予測モデルの構築を行う。

2.1 注文の符号化

注文には、価格、時刻といった数値情報(量的変数)と、種類といったカテゴリ情報(質的変数)を持つ。ここではこのような注文を記号に変換し、その時系列を記号列で表現することを目的とする。注文の特徴には、売成行(Ask Market Order)／買成行(Bid Market Order)／売指値(Ask Limit Order)／買指値(Bid Limit Order)／売キャンセル(Ask Cancel)／買キャンセル(Bid Cancel)といった注文の種類、他、価格、注文量、時刻などがある。ここでは種類と価格、そして前の注文からの時刻差のみに限定する。まず、各指値注文と各キャンセル注文に対しては、注文の入った時点で板の最良売り気配と最良買い気配の平均を仲値として求め、価格と仲値との差の絶対値を保持する。時刻差は一つ前の注文との時刻差である。これは、注文の特徴として凡その発注者の情報を加えることを意図している。例えば、マイクロ秒オーダーでの注文は機械的なトレーダーによる注文だと識別できるようにしている。

次に、価格差と時間差をそれぞれ、ある分類規則に従いカテゴリ化する。その後、各注文に識別符号を与える。成行注文2種に関しては、時刻差のみを考慮し識別符号を付与し、指値注文とキャンセル注文に関しては、仲値からの価格差と時刻差によって分類後、符号を付与する。

注文と識別符号間の変換は一意であり、符号から注文へは復元可能となっている。なお、記号に変換することで価格といった質的変数の情報を落とすことになるが、大量のデータの中からパターンを識別することで、タスクと注文の意味の関係を自ら獲得することを期待している。

2.2 問題設定

本節では、価格動向予測の問題設定を行う。まず不平等間隔のタイムスタンプを持つ注文に対して、任意の一定間隔 T で区間を設ける。ここで、各区間内での注文の系列の集合を $\mathcal{C} = \{S^1, S^2, \dots\}$ と表し、各系列 S^i は $S^i = (x_1^i, x_2^i, \dots)$ と書く。 x_τ^i は、2.1で定義された、系列 S^i 内の τ 番目の注文を表現する記号であり、系列

S^i は、 i 番目の区間における注文の可変長の記号列を表す。各注文 x_τ^i は、前処理にて分類される識別符号の集合で大きさ I の集合 \mathcal{I} の要素である。

目的は、各系列 $S^i = (x_1^i, x_2^i, \dots)$ を説明変数として、 S^i に対応する時間領域の終点から T' 後の価格の動向ラベル t^i の予測である。ニューラルネットワークで表現される関数のパラメータを θ とし、出力となる条件付き確率を $p(t^i | x_1^i, x_2^i, \dots; \theta)$ で表す。教師あり学習によって、この条件付き確率が最大となる最適なパラメータ θ を探索する。

2.3 CNNを用いたモデル

モデルにはCNNを用いる。CNNは、入力の領域内での不変性を仮定し、畳み込みフィルタというパラメータを各窓において共有することによりパラメータの学習を容易にする。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 [7][8]。本研究でCNNを用いる理由として、可変長系列の学習に対してパディングを行うことで容易にミニバッチ学習を行える点、RNN(Recurrent Neural Network)であれば予測のタスクが系列の後方のデータに依存するのに対して、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、価格動向に相関が比較的小さいであろう指値注文やキャンセルが系列の後方に集中してもうまく学習することができる。と期待した点が挙げられる。

ここではCNNモデルの定式化を行う。まず、系列 S 内の注文 x_τ の埋め込みベクトル \mathbf{x}_τ を次のように得る。

$$\mathbf{x}_\tau = \mathbf{w}_{\text{embed}} \mathbf{x}_\tau^{\text{onehot}} \quad (1)$$

ここで、 $\mathbf{w}_{\text{embed}} \in \mathbb{R}^{e \times I}$ は埋め込み行列、 e は埋め込みベクトル \mathbf{x}_τ の次元数である。 $\mathbf{x}_\tau^{\text{onehot}}$ は注文 x_τ を表す onehot ベクトルである。系列 S はパディングによって長さ n で統一後、次のように表現する。

$$\mathbf{x}_{1:n} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \quad (2)$$

$\mathbf{x}_{i:i+j}$ は注文 $(x_i, x_{i+1}, \dots, x_{i+j})$ を連結したものを表す。窓幅 h に対して畳み込みを行うフィルタの重み行列を $\mathbf{w}_{\text{conv}} \in \mathbb{R}^{e \times h}$ とすると、次のような式を得る。

$$c_i = \tanh(\mathbf{w}_{\text{conv}} \cdot \mathbf{x}_{i:i+h} + b) \quad (3)$$

c_i は埋め込み後の注文の局所的な小行列 $\mathbf{x}_{i:i+h}$ を畳み込みこむことによって得られる新たな特徴。 \cdot はドット積、 $b \in \mathbb{R}$ はバイアスである。ストライド幅 1 とし、 $(\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \dots, \mathbf{x}_{n-h+1:n})$ 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を得る。

$$\mathbf{c} = [c_1, c_2, \dots, c_{n-h+1}]^T \in \mathbb{R}^{n-h+1} \quad (4)$$

次に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 $\hat{c} = \max(\mathbf{c})$ を得る。 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 $\hat{\mathbf{c}} = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_{k_{\text{conv}}}]^T$ を全結合層の重み $\mathbf{w}_{\text{fully}} \in \mathbb{R}^{l \times k_{\text{conv}}}$ を用いて次のように変換する。

$$\mathbf{o} = \mathbf{w}_{\text{fully}} \hat{\mathbf{c}} \quad (5)$$

l は動向ラベルのサイズである。 p_j を j 番目のラベルの確率とすると、得られる確率は以下ようになる。

$$p_j = \frac{\exp o_j}{\sum_{i=1}^l \exp o_i} \quad (6)$$

2.4 学習

(7) 式に示す目的関数を最小にするようパラメータ $\theta = (\mathbf{w}_{\text{embed}}, \mathbf{w}_{\text{conv}}, \mathbf{w}_{\text{fully}})$ を探索し、最適パラメータ $\theta = \theta^{\text{opt}}$ を得る。

$$L = - \sum_{i=1}^C \sum_{j=1}^o t_{i,j} \log p_j \quad (7)$$

ただし、 $C = |C|$ かつ $t_{i,j}$ は系列 S_i に対応するラベルを onehot ベクトルで表現した時の j 番目の要素である。

2.5 埋め込み行列の平均化を用いた CNN

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、注文単位などのマイ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。平均化を行うのは、埋め込み行列の局所的な範囲を対象とする。

ニューラルネットワーク構造を Figure 1 に示す。2.3 での (2) 式に示す $\mathbf{x}_{1:n}$ に対して平均プーリングを、窓幅 $1 \times k_{\text{pool}}$ 、ストライド幅 $1 \times k_{\text{pool}}$ 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のパディングサイズ m_{pad} は以下ようになる。

$$m_{\text{pad}} = \begin{cases} 0 & (n \equiv 0 \pmod{k_{\text{pool}}}) \\ \lceil \frac{k_{\text{pool}} - (n \bmod k_{\text{pool}})}{2} \rceil & (\text{otherwise}) \end{cases}$$

ただし、 $\lceil x \rceil$ は天井関数であり、 x 以上の最小の整数を表す。床関数を $\lfloor x \rfloor$ とすると、プーリング後の特徴行列は、

$$\mathbf{x}_{\text{pool}} \in \mathbb{R}^{e \times \lfloor \frac{m_{\text{pad}} + n}{k_{\text{pool}}} \rfloor}$$

のように表すことができる。その後は 2.3, 2.4 と同様に順伝搬および学習を行う。

3 実験

3.1 銘柄と期間

データは FLEX-FULL ヒストリカルデータを用いた。このデータは東証上場現物株式の各銘柄のティックデータを蓄積したものである。ティックデータには取引情報だけでなく、注文、取引の時刻、銘柄情報、価格、数量といった情報が含まれる。データは大和ハウス工業 (銘柄コード 1925)、積水ハウス (1928)、キリン (2503)、味の素 (2802)、東レ (3402)、旭化成 (3407)、三菱ケミカルホールディングス (4188)、第一三共 (4568)、JXTG ホールディングス (5020)、新日鐵住金 (5401)、東芝 (6502)、パナソニック (6752) の各々 2013 年 7 月 1 日から 2014 年 6 月 30 日までの、全 12 銘柄、245 日のデータを用いた。

3.2 注文の記号列変換

2.1 で述べた手法で、注文の前処理を行う。本実験では、それぞれの特徴量に対して、以下のようなカテゴリサイズを行った。価格に関しては、注文の出現数の多い、仲値に近い注文ほど分類の粒度を細かく、出現数の少ない、仲値に遠い注文ほど粒度を粗く設定した。時刻差に関しては、500ms 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 20ms を境に高速な注文か比較的低速な注文なのか、識別できることを期待して設定した。このようにして得られる識別符号の集合 I の大きさは $I = 90$ となった。

$$\begin{pmatrix} \text{MarketOrder}^{\text{ask}} \\ \text{MarketOrder}^{\text{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 20\text{ms} \\ 20\text{ms} \sim 500\text{ms} \\ 500\text{ms} \sim \end{pmatrix} \\ \begin{pmatrix} \text{LimitOrder}^{\text{ask}} \\ \text{LimitOrder}^{\text{bid}} \\ \text{Cancel}^{\text{ask}} \\ \text{Cancel}^{\text{bid}} \end{pmatrix} \times \begin{pmatrix} \sim 1 \\ 1 \sim 2 \\ 2 \sim 3 \\ 3 \sim 5 \\ 5 \sim 7 \\ 7 \sim 10 \\ 10 \sim \end{pmatrix} \times \begin{pmatrix} \sim 20\text{ms} \\ 20\text{ms} \sim 500\text{ms} \\ 500\text{ms} \sim \end{pmatrix}$$

種類 価格 時間差

例を挙げると、買成行注文の、前の注文から 30ms 後の注文であれば 4、売キャンセル注文のうち仲値から 2 の、前の注文から 5ms 後の注文であれば 75 の符号を与えている。

3.3 データセット

本節では、一つの銘柄において、1 日単位で連続する注文の記号列からデータセットを作成する手法を述べる。ある銘柄の 1 年分 245 日のうち、前 8 割の 196

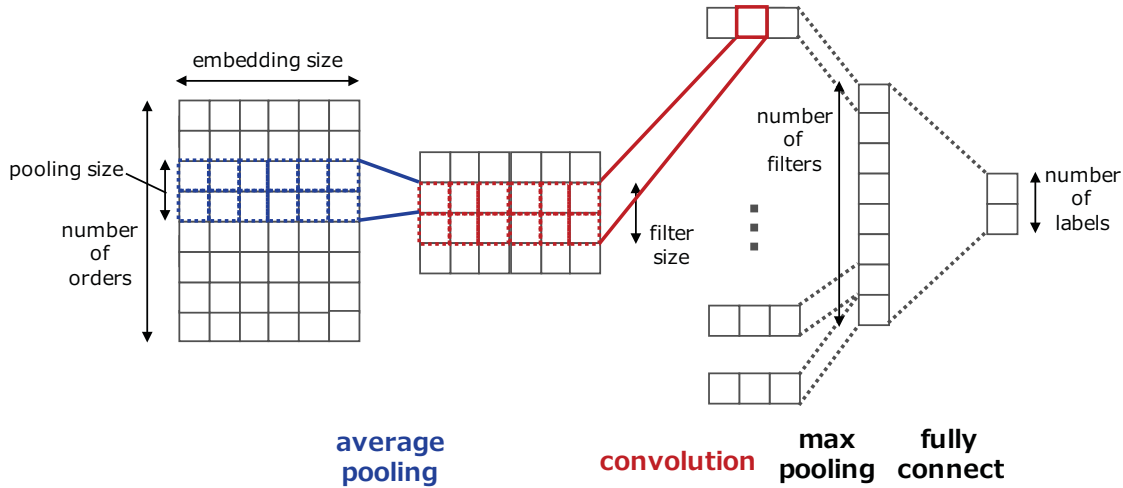


Figure 1: 埋め込み行列の平均化を用いた CNN モデルの構造.

日分を訓練用生データ，後 2 割の 49 日分を検証用生データとして分割する．訓練用，検証用それぞれに対して，以下の方法でデータセットを作成する (Figure2)．各日毎のザラ場内 (9:00~11:30,12:30~15:00) の注文に対して，9:00:00 を起点として 11:30:00 までの注文を 30 秒間隔で逐次サンプリングを行い，複数の可変長記号列を生成する．後場に対してもこれを同様にを行う． i 日の系列集合のうち j 番目の注文記号列を S_i^j とする． S_i^j は区間 $[\tau^{j-1}, \tau^j)$ から得られたとする．ただし， $i \in (1, 2, \dots, 245)$ ，かつ $j \in (1, 2, \dots, 600)$ である．なお， j の属する集合のサイズ 600 はザラ場中 5 時間を 30 秒で割って求める 1 日の区間の数である．これで，訓練用 $196 \times 600 = 117600$ ，検証用 $49 \times 600 = 29400$ の系列数となる．訓練用に対しては，区画を決定する起点を 10 秒ずつずらして，これまでの処理を行う．つまり，9:00:10 と 9:00:20 を起点としてそれぞれ 30 秒間隔でサンプリングを行っていく．これは Data Augmentation のようなもので，注文の系列の一部に重複を認めるものの，訓練用のデータセット数を大きくするために行う．このようにして，訓練用の系列数を 3 倍とした．次に，ある S^j に対して，対応する区間の終点の株価 p_{τ^j} と終点から 30 秒後の株価 $p_{\tau^{j+1}}$ を比較して次のようにラベリングを行う．

$$t^j = \begin{cases} 0 & (p_{\tau^j} > p_{\tau^{j+1}}) \\ 1 & (p_{\tau^j} < p_{\tau^{j+1}}) \end{cases} \quad (8)$$

なお，価格に変動がない場合，つまり $p_{\tau^j} = p_{\tau^{j+1}}$ となった系列と，注文数が 20 以下の系列に関しては，本研究での予測の対象外としてデータセットから除外した．さらに，訓練用に関しては，正値ラベルと負値ラベルの系列数を同一とした．これらの処理をすべての銘柄に対して行い，得られる訓練用データセットは，最

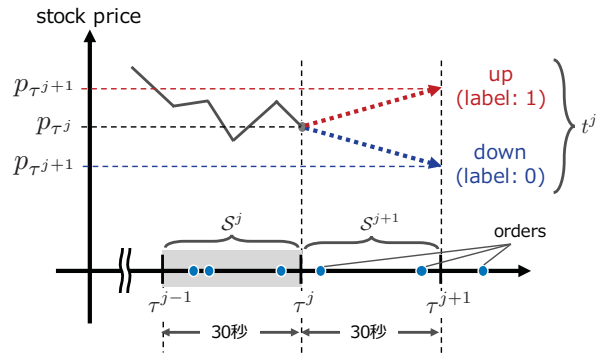


Figure 2: 系列 S^j のサンプリングとラベリング.

大で銘柄コード 6752 の 155202 件，最小で銘柄コード 4188 の 36364 件となった．

3.4 実装

各銘柄で用意したデータセットに対して，それぞれのモデルを用意する．(7) 式が最小となるように，最適化を行う．埋め込みサイズ $e = 5$ ，畳み込みフィルタの窓幅 $h = \{1, 3, 5, 7\}$ とし，各窓幅に対して 5 つのフィルタ，全 20 フィルタを用意した．埋め込み行列の平均化を用いたモデルでは，平均プーリングの窓幅 $k_{\text{pool}} = 10$ とする．勾配降下法の最適化アルゴリズムには Adam を， $\lambda = 0.0005$ とした重みへの L2 正則化を，全結合層では比率 0.5 でのドロップアウトを使用し，バッチサイズ 100 としたミニバッチでの学習を 50 エポック行った．深層学習フレームワークとして chainer を使用し実装した．

3.5 ベースライン

ベースラインとして、ロジスティック回帰、非線形 SVM, MLP(Multi Layer Perceptron)を用いた。提案手法と同様に、各銘柄に対してそれぞれのモデルを学習する。入力としては、系列 S 内の注文の頻度を各注文の種類 I の長さのベクトルで表現したものをを用いる。可変長の系列 S の注文に対して、これらの手法では入力次元を一定にする必要があるからであり、このようなナイーブな手法をとることとする。MLP の中間層 2 層の次元はそれぞれ 8,4 とした。

4 結果と考察

4.1 モデルの評価

12 銘柄の各 5 手法、全 60 の学習済みモデルを、検証用データを用いて予測と評価を行う。3.3 で付与した、上昇したか下落したかというラベルに対してそれぞれ F 値を求め、サンプル数で加重平均をとったもので評価する。検証データのサンプル数は、最大で銘柄コード 6752 の 12835 件、最小で銘柄コード 4188 の 3122 件となった。結果を Table1 に示す。CNN では、銘柄コード 2503, 2802 の以外で、ベースラインすべてのモデルを上回るまたは等しい結果を残した。A-CNN モデルでは、すべての銘柄におけるベースラインおよび CNN モデルを上回る結果となった。CNN と A-CNN の提案手法と、ベースラインとを比較すると、ベースラインの入力が時系列情報を落としているため、提案手法が上回ったと考えられる。しかしそれは、可変長時系列をそのまま扱うことのできる提案手法の利点によるものである。

4.2 埋め込み層の分析

CNN と A-CNN の埋め込み層の分析を行った。各モデルの埋め込み行列 $w_{\text{embed}} = [w_1, w_2, \dots, w_{90}]$ の各列 w_i は識別符号 i の注文の埋め込みベクトルに対応する。これ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られることできる。Table 2 には、各モデルにおける埋め込みベクトルのノルムを注文の種類毎に集計平均したものを、さらに銘柄毎に平均をとったものである。CNN, A-CNN とともに、成行注文のノルムが、指値注文、キャンセル注文のそれに比べて大きい。さらに A-CNN のそのの方がより大きいノルムを持つ。これから、すべての注文の中から成行注文の特徴を強く捉えようと、モデルが学習していることがわかる。本研究でのタスクはリターンではなく価格動向であるが、成行注文とリ

ターンに相関があるという主張 [5][6] を支持しているものだと考えられる。

4.3 成行注文比率と評価値

次に、なぜ銘柄間で評価値に差異があるのかを調査した。ベースラインで最もスコアの高かった MLP と CNN, A-CNN の 3 手法のそれぞれ 12 個のモデルの評価値と成行注文比率との関係を Figure 3 に示す。成行注文比率とは、検証データ内における、すべての注文に対する成行注文の割合である。それぞれのモデルに対して線形フィッティングを施すと、決定係数は MLP, CNN, A-CNN の順で高くなる。相関係数も CNN で 0.909 (p 値 4.29×10^{-5}), A-CNN で 0.940 (p 値 5.67×10^{-6}) と非常に高い。提案手法は、成行注文の密度が高ければ予測精度が高くなることを示している。

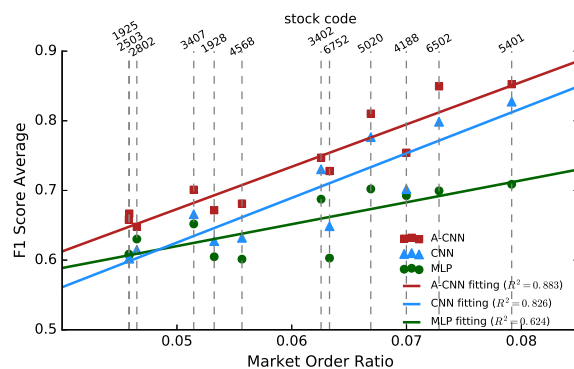


Figure 3: 各銘柄 × 評価値上位 3 手法のモデルにおける成行注文比率と F1-average との関係。 R^2 は決定係数。

逆に成行注文以外の指値注文やキャンセルの比率が大きくなると精度が低くなることから、指値注文やキャンセルは価格動向との相関が低くノイズとなっていたと考えられる。CNN と A-CNN を比較すると、注文の意味空間で平均とることが、順伝搬、逆伝搬でのノイズを除去し、成行注文に反応するように訓練する一助になったと考えられる。

5 まとめ

本稿では、注文の符号化と、一定時間内でのサンプリングにより得られる可変長注文記号列を、CNN によって固定長の特徴に変換後、予測を行うモデルを提案した。分析の結果、価格変動と相関のある成行注文をとらえるような訓練をすることわかった。今回のタスクにおける精度の問題を改善するためには、成行注文のみを入力とし、シンプルな CNN や RNN を用いた場合の方が精度が高くなると考えられる。しかし、他の問

Table 1: 各手法 × 各銘柄のモデルの F1-Average による評価. A-CNN は 2.5 での埋め込み行列に平均化を適用したモデル.

手法	銘柄コード											
	1925	1928	2503	2802	3402	3407	4188	4568	5020	5401	6502	6752
Logistic	0.602	0.602	0.596	0.614	0.662	0.631	0.661	0.600	0.666	0.691	0.674	0.585
SVM	0.588	0.594	0.593	0.613	0.674	0.646	0.662	0.592	0.689	0.701	0.692	0.589
MLP	0.610	0.607	0.608	0.630	0.689	0.652	0.693	0.608	0.713	0.709	0.701	0.605
CNN	0.610	0.627	0.605	0.616	0.732	0.667	0.703	0.633	0.777	0.828	0.799	0.649
A-CNN	0.659	0.672	0.667	0.649	0.747	0.701	0.754	0.683	0.810	0.853	0.850	0.728

Table 2: 埋め込みベクトルのノルム平均.

注文の種類	手法	
	CNN	A-CNN
MarketOrder ^{ask}	1.31	2.22
MarketOrder ^{bid}	1.32	2.17
LimitOrder ^{ask}	0.382	0.323
LimitOrder ^{bid}	0.402	0.307
Cancel ^{ask}	0.373	0.279
Cancel ^{bid}	0.373	0.286

題設定下での注文の意味や特徴を探るには、本研究で行ったようなすべての注文を用い、平均化を行う手法は活用できると考えられる。今後、関連研究で行われているような、指値、キャンセルの相関が比較的大きいであろう仲値の予測を行い、埋め込みベクトルの関係がどのように変化するかを調査したいと考えている。また本研究では調査できていないが、予測の結果が単に価格のトレンドフォローであるか否かの検証、注文のその他の特徴である、価格や時間差に意味があるかなどの分析を今後検討している。

参考文献

- [1] R. Almgren and J. Lorenz, "Adaptive Arrival Price," *Algorithmic Trading III, Institutional Investor*, pp. 59–66, 2007.
- [2] 杉原慶彦 「取引コストの削減を巡る市場参加者の取組み：アルゴリズム取引と代替市場の活用」『金融研究』第 30 巻第 2 号, 29–88 頁, 2011 年.
- [3] A. N. Kercheval and Y. Zhang, "Modelling high-frequency limit order book dynamics with support vector machines," *Quantitative Finance*, vol. 15, no. 8, pp. 1315-1329, 2015.
- [4] A. Tsantekidis, N. Passalis, A. Tefas, J. Kannianen, M. Gabbouj, and A. Iosifidis, "Using Deep Learning to Detect Price Change Indications in Financial Markets," *European Signal Processing Conference, Greece*, 2017.
- [5] P. Weber and B. Rosenow, "Order book approach to price impact," *Quantitative Finance*, vol. 5, no. 4, pp. 357-364, 2005.
- [6] F. Abergel, M. Anane, A. Chakraborti, A. Jedidi, and I. Toke, "Limit Order Books," *Cambridge University Press*, 2016.
- [7] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Conference on Empirical Methods in Natural Language Processing*, pp. 1746-1751, 2014.
- [8] Y. Zhang and B. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," *arXiv 1510.03820*, 2015.

外国為替取引におけるプロスペクト理論

Prospect theory on Foreign Exchange Transaction

落合友四郎¹、ホセ ナチエル²

Tomoshiro Ochiai¹ and Jose Nacher²

¹大妻女子大学社会情報学部

¹Faculty of Social Information Studies, Otsuma Women's University,

²東邦大学理学部

² Department of Information Science, Faculty of Science, Toho University,

Abstract:

人間の経済行動には非合理的な行動傾向があることが、行動経済学におけるプロスペクト理論において知られている。この理論では、人は利得している状態の時は損失回避行動をとるのに対して、損失を出している状態の時はリスク選好行動を取ることを示している。今回、外国為替証拠金取引のデータとして、28ヶ月にわたる8万1300人による2800万回の取引データを解析した。そして、この取引の参加者達にもプロスペクト理論から示唆される不合理な投資行動パターンが存在することを示した。さらに、この取引の参加者達を、利益を得ているグループと損失を出しているグループに分けた時、損失を出しているグループの方がより強くプロスペクト理論から示唆される非合理的な取引行動をとっていることがわかった。

はじめに

インターネットの発達に伴い、ECサイトなどの売買履歴やポイントカードなどで記録される客の行動履歴、またSNSなどの交流履歴などの多くのデータがビッグデータとして蓄えられてきている。これらのデータの中には人間の行動パターンに関する重要な知見が含まれていると考えられる。これらの売買履歴、通信履歴、交流履歴などのデータから、将来の人間の行動予測などをして、ビジネスなどに応用しようとする研究が盛んになってきている。

行動経済学で知られているように、人間の行動には人間特有の行動バイアスがあることが知られている[1-8]。その一つとして行動経済学で知られているプロスペクト理論がある[1]。この理論は、1978年にダニエル・カーネマンとエイモス・トベルスキーによって発表され、2002年にのノーベル経済学賞を受賞した。

プロスペクト理論は、人間は利得場面と損失場面では非対称的な行動(感情)をするということを明らかにした。一つの特徴は、人間は同じ金額であれ

ば損失のほうが、利益よりも感情的に大きな痛みを感じるというものである。もう一つの特徴は、人間は利得場面ではリスク回避的な行動をとるのに対して、損失場面ではリスク選好的な行動をとる。これを言い換えると、自分が得している場面では防御的な行動をとるのに対して、損している場面ではギャンブル的な行動をとるということである。

今回、外国為替証拠金取引の海外の会社の顧客の取引データをもちいて、投資家の投資行動パターンを研究した[9]。

プロスペクト理論

プロスペクト理論は価値関数によって特徴づけられる。価値関数とは、何かしらの経済的な損得を x としたときに、それを得た時の感情的な価値を $v(x)$ として表す。価値関数として用いられる代表的な関数の形は以下で与えられる[8]。

$$v(x) = \begin{cases} -\lambda(-x)^\beta, & x < 0 \\ x^\alpha, & x \geq 0 \end{cases}$$

ここで、 $0 < \alpha, \beta < 1$ で、 $\lambda > 1$ である。この価値関数の様子は図1に示した。

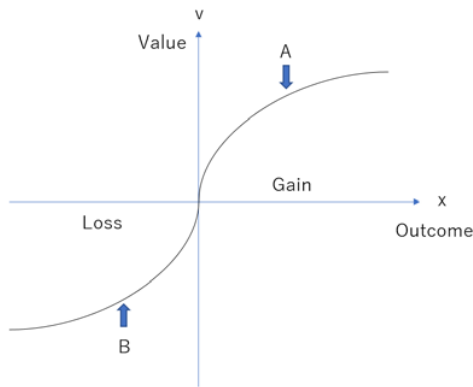


図1：プロスペクト理論の価値関数

この価値関数の要点は2つある。まず第一に、原点における曲線の上昇率、下降率が利得領域 ($x > 0$) と損失 ($x < 0$) 領域で違うということである。これは同じ額の損失と利得であれば、損失のほうがより大きな感情的な痛みと感じる。

第2の要点は、利得領域 (図の点A) では、リスク回避的な行動をとるのに対して、損失領域 (図の点B) ではリスク選好的な行動をとるということである。これは、図1の価値関数では次のように説明される。点Aでは、さらに利益を得た時の価値関数の増分と、利得状態から利益を減少させた場合の価値関数の減少分を比べると、後者の減少分の方が大きいので、人はリスクを回避するように行動する。一方、点Bでは、損失領域からさらに損失を出したときの価値関数の減少幅と、損失領域から損失が回復したときの価値関数の増大分を比べると後者のほうが大きいので、人はリスク選好的な行動をとる。

これをまとめると次のように言うことができる。人間のおこなう投資では、勝つときは利幅が小さくなり、負けた時は損失幅が大きくなる傾向があるといえる。

解析に用いたデータ

今回海外にある外国為替証拠金取引業者から提供を受けた外国為替証拠金取引のデータとして、28ヶ月にわたる8万1300人による2800万回の取引データを解析した。各取引データには、個人ID、通貨ペア、取引のエントリー時刻、エグジット時刻、利益(損失)金額などが記載されている。

解析方法と結果

それぞれの取引のポジションにエントリーした時

刻を τ_{opened} 、エグジットした時刻を τ_{closed} とすると、その取引のポジション保持時間は次で定義される。

$$\tau = \tau_{closed} - \tau_{opened}$$

利益の出た取引と損失が出た取引に分けて、それぞれのポジション保持時間の分布を図2に示した。横軸がポジションの保持時間で単位はミリ秒である。縦軸が頻度で、利益の出た取引と損失が出た取引のそれぞれで正規化してある。

図2をみてまず気づくことは、まず両方の分布ともに、べき乗分布に近い形をしていることである。これは、ポジション保持時間には秒単位から月単位までの取引があり、特に典型的なスケールがないことを示している。

また、利益の出た取引(緑色)のポジション保持時間をみてもと、一秒以下のところに分布がない(ゼロ)ことがわかる。これは、いわゆる買値と売値の間のスプレッドが原因である。ポジション保持時間が1秒以下では、スプレッド分を超えるように利益をだすことが不可能(に近い)ということである。

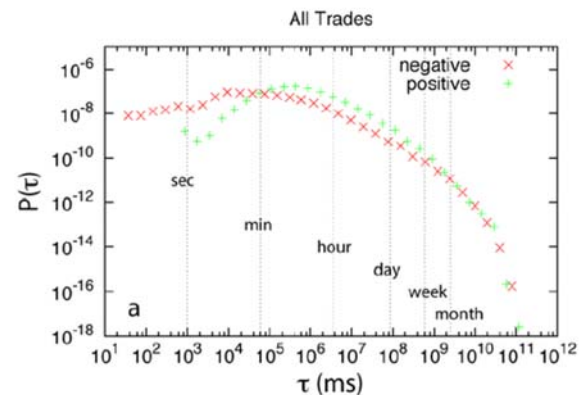


図2：ポジション保持時間の分布(利益の出た取引(緑色)と損失が出た取引(赤色)を別々に集計)

さて、次に各投資家ごとの投資行動の傾向を見ていく。各投資家の勝率を次の式で定義する。

$$w = \frac{N_+}{N_+ + N_-}$$

ここで、 N_+ は投資家ごとの利益の上がった取引数、 N_- は投資家ごとの損失の出た取引数である。この各投資家ごとの勝率の分布を図3に示した。分布は右のほうに偏っており、多くの投資家の勝率は50パーセントを超えていることがわかる。利益幅、損失

幅を無視して、勝率だけを考えると多くの投資家は「勝っている」ことになる。

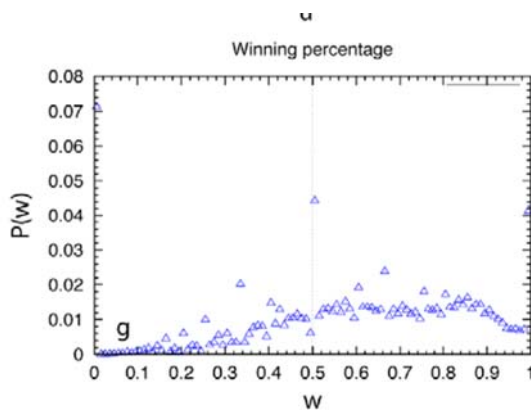


図3：各投資家ごとの勝率の分布

もちろん、最終的な損益は、勝率のみならず利益幅と損失幅にも寄るので、その様子を次に調べてみる。各投資家に対して、リスクリワード比率 (Risk Reward Ratio) を次の式で定義する。

$$r = \frac{\langle p_+ \rangle}{\langle |p_-| \rangle}$$

ここで、 $\langle p_+ \rangle$ は、その投資家の利益がでた取引の平均利益額である。 $\langle |p_-| \rangle$ は、その投資家の損失がでた取引の平均損失額である。ここで絶対値をとっているので、 $\langle |p_-| \rangle$ はプラスの値になる

図4に、各投資家ごとのリスクリワード比率 r の分布を表示した。この図4をみて気づくことは、 $r = 1$ を軸に左右非対称になっていることである。もし、トレードが特に心理的なバイアスのない機械トレードや、猿がダーツを投げて取引をするようなランダムトレードであれば左右対称になるはずである。ところが、驚くべきことに図4をみると、 $r = 1$ の左と右では全く異なる様相を示している。右側のリスクリワードが良い投資家（利幅のほうが損失幅より平均的に大きい投資家）は、べき乗則に似た曲線で減少していく。このことはリスクリワードの良い投資家は圧倒的に少ないが、まったく存在しないわけではなく、まれには存在するということである。

一方、図4の $r = 1$ の左側のリスクリワード比が悪い投資家（平均利益幅より平均損失額のほうが大きい投資家）の分布をみると、図4の右側とは全くことなる様相を示す。 $r = 1$ の左側ではリスクリワードが悪くなくても、頻度はかわらずに一様分布のようになっている。そして、 $r = 1$ の右側に比べて圧倒的に

高い頻度がみとれる。これは、大部分の投資家のリスクリワード比率が1以下で、平均損失額が平均利益額を上回っていることを示す。

このことが、勝率の高い投資家が多いのにも関わらず、最終的に利益の出している投資家はわずか14.7%しかおらず、85.2%もの投資家は最終的には損失を出している理由になっている。リスクリワード比が悪いために、勝率が高くても負け幅が大きく最終的に利益を残すことができない人がほとんど（8割以上）となっている。

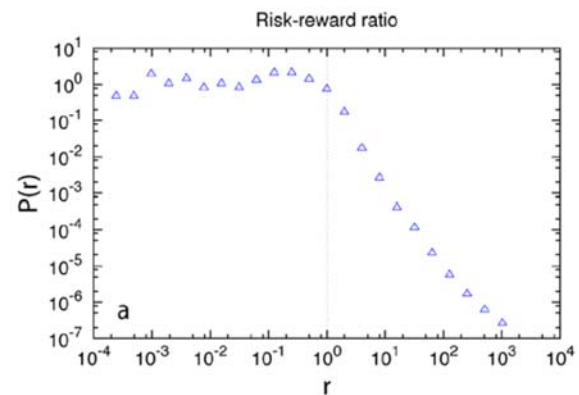


図4：各投資家に対するリスクリワード比率の分布。

次に、各投資家ごとに利得取引と損失取引のポジション平均保持時間比率を次の式で定義する。

$$s = \frac{\langle \tau_+ \rangle}{\langle \tau_- \rangle}$$

ここで、 $\langle \tau_+ \rangle$ は各投資家の利益のでた取引のポジションの平均保持時間である。同様に、 $\langle \tau_- \rangle$ は損失のでた取引のポジションの平均保持時間である。図5はこの各投資家ごとに利得取引と損失取引のポジション平均保持時間比率の分布を示している。

リスクリワード比の分布と似たような非対称性が見られる。多くの投資家が利得のあるポジションを早く解消する一方、損失の出ているポジションのクローズを先延ばしにしていることがわかる。

利益・損失グループに分類した場合

次に、これまでの解析を最終的に利益を出している投資家のグループと、損失を出しているグループに分けて解析してみる。

図6は、投資家ごとのリスクリワード比を最終的に利益を出した投資家のグループと、損失の出した投資家のグループに分けて表示した。図6をみると、利益を出しているグループは、損失の出したグループに比べて、リスクリワード比の分布が右にシフトしていることがわかる。

まず図6の左側（リスクリワード比が1以下）をみると、損失を出しているグループ（赤色）は、高い水準で一様分布をしています。一方、利益を出しているグループ（緑色）は頻度がだいぶ下の水準である。次に、図6の右側（リスクリワード比が1以上）を見てみる。どちらのグループもべき乗則に従って減少していくが、利益を出しているグループのほうが裾が厚くなっている。

これらの2つのグループの差は次のことを意味している。利益を出しているグループはプロスペクト理論による心理的影響を受けにくい投資家が多いのに対して、損失を出しているグループはプロスペクト理論の影響を大きく受けているということである。

次に図7は、最終的に利益を出した投資家のグループと損失の出した投資家のグループに分けて、各投資家ごとに利得取引と損失取引のポジション平均保持時間比率の分布を図示した。図6のリスクリワード比の分布と同様に、利益を出しているグループのほうが、損失を出しているグループに比べて、右半分では上のほうに位置して、左半分では逆に下回っている。これは、最終的に損失を出しているグループは、損失の出ているポジションをなかなか「損切」することができずにポジションをだらだらと持ち続けるのに対して、利益の出ているポジションは手早く手仕舞ってしまう傾向があることを示す。

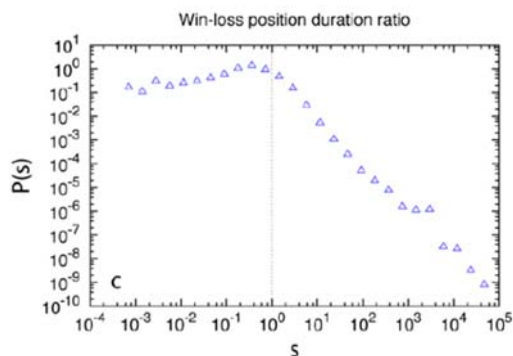


図5：投資家ごとに利得取引と損失取引のポジション平均保持時間の比率の分布

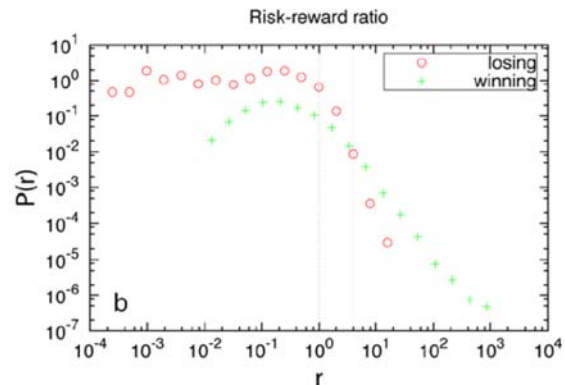


図6：各投資家に対するリスクリワード比率の分布（利益を出したグループ（緑色）と損失を出したグループ（赤色）を別々に表示した。）

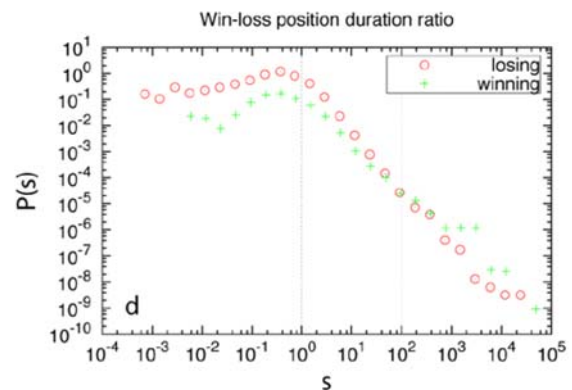


図7：投資家ごとに利得取引と損失取引のポジション平均保持時間の比率の分布（利益を出したグループ（緑色）と損失を出したグループ（赤色）を別々に表示した。）

終わりに

外国為替証拠金取引の莫大なデータを調べることにより、プロスペクト理論から予想される投資家行動パターンがあることを示した。また、利益を出しているグループは、そうでないグループに比べてプロスペクト理論から予想される心理的影響を受けにくいことが分かった。

参考文献

- [1] Diamond P, Vartiainen H, Behavioral Economics and Its Applications Hardcover. Princeton University Press,

(2007)

- [2] Kahneman D, Tversky A, Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–291, (1979)
- [3] Tversky A, Kahneman D, The Framing of Decisions and the Psychology of Choice. *Science* 211: 453–458, (1981)
- [4] Tversky A, Kahneman D, Advances in prospect theory: Cumulative representation of uncertainty. *J Risk Uncertain* 5: 297–323, (1992)
- [5] Wakker P, Tversky A, An Axiomatization of Cumulative Prospect Theory. *J Risk Uncertain* 7: 147–75, (1993)
- [6] Schmidt U, Starmer C, Sugden R, Third-generation prospect theory. *J Risk Uncertain* 36: 203–223, (2008)
- [7] Barberis N, Thirty years of prospect theory in economics: A review and assessment. *J Econ Perspect* 27: 173–196, (2013)
- [8] Yan Li and Liyan Yang, Prospect theory, the disposition effect, and asset prices, *Journal of Financial Economics*, 107(3), 715-739, (2013).
- [9] Liu Y-Y, Nacher JC, Ochiai T, Martino M, Altshuler Y, Prospect Theory for Online Financial Trading. *PLoS ONE* 9(10): e109458, (2014)

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

BOJ monetary policy analysis based on
the facial expression analysis of the BOJ governor

水門善之¹ 勇大地²

Yoshiyuki Suimon¹, Daichi Isami²

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: After the Bank of Japan (BOJ)'s monetary policy meeting, the BOJ governor makes a press conference to explain the monetary policy. In this research, using facial expression recognition algorithms based on deep learning, we analyzed the governor's facial expressions in the press conference and estimated the emotional indexes such as "Happiness", "Anger", "Sadness", "Surprise". As a result, we found that the indexes of "Anger" and "Disgust" increased much just before making major policy changes. On the other hand, the index of "Sadness" tended to decline after the monetary policy changes. This suggests that the information based on the facial expression analysis can be a useful material for forecasting the future monetary policy.

はじめに

日本銀行（以下、日銀）は、金融政策に関する方針を、金融政策決定会合にて審議・決定している。決定会合の後には、日銀総裁が記者会見を行い、金融政策運営に関する説明を行っている。

日銀は、総裁記者会見の内容を、後日文書化して公表している。ただし、これらはいくまでテキストデータであり、それらがどのようなトーンや表情で発言されたかについての情報は含まれていない。

しかし2014年度以降、日銀は決定会合後の記者会見の様子をメディアを通じて動画配信することを解禁した。これにより、総裁の会見内容について、テキスト情報だけではなく、表情やトーンも含めた解釈が可能になった。

情報理論的な定義における情報量でみた場合、文章よりも動画・音声の方が、情報量は遥かに大きい。本研究では、これらの点に注目し、会見動画の解析を行うことで、テキストデータに含まれない情報の抽出を試みた。具体的には、深層学習等を用いた表情認識アルゴリズムを用いることで、会見における総裁の表情を解析し、「喜び」・「怒り」・「悲しみ」・「驚き」等の感情の変化を指数化した。その結果、重大

な金融政策変更を行う直前の回の会見では、「怒り」や「嫌悪」の値が高くなる一方、金融政策変更後の会見では、「悲しみ」の数値が低下する傾向が確認された。

このことは、政策変更前の金融政策に対する問題意識の高まりと、金融政策変更によってそれが緩和されることによる安堵が、表情に表れている可能性を示していると考えられる。同時に、表情解析に基づく情報が、金融政策の先行きを考える上で有用な材料となり得ることも示唆していると言えよう。

日銀の金融政策運営

本研究では、日銀の金融政策決定会合後の総裁記者会見動画の解析を行ったが、分析内容の紹介を行う前に、以下では最近の日銀の金融政策運営を振り返りたい。

日銀は、2013年4月に「量的・質的金融緩和」政策を導入して以降、長期国債の買入れ額を大幅に増額するなどの金融緩和政策を実施してきた。そのような中、日銀は2016年に従来の金融政策のフレームワークを大きく転換させる二つの意思決定を行った。一つ目は、2016年1月に決定された「マイナス金利付き量的・質的金融緩和」の導入である。具体的

には、金融機関が保有する日銀当座預金の一部に-0.1%というマイナスの金利を適用させるものであり、金融市場に大きなサプライズをもたらした。

二つ目は、2016年9月の「長短金利操作（イールドカーブ・コントロール）付き量的・質的金融緩和」の導入決定が挙げられる。特に、長期金利を含む金利操作目標を提示するイールドカーブ・コントロールは、これまでの金融政策のフレームワークを大きく拡張したものであり、画期的な政策変更であったと言える。

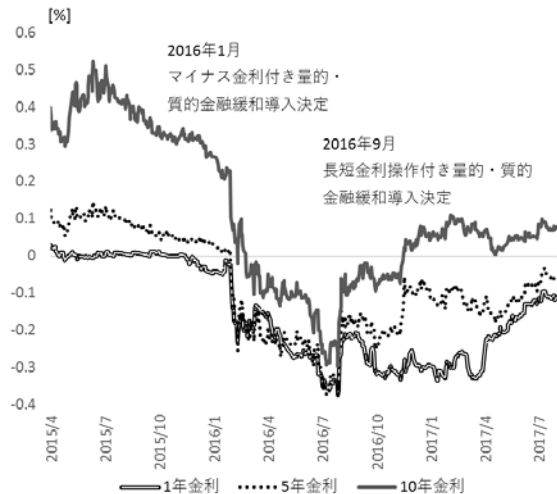


図1: 日本国債の年限別金利変化

総裁会見動画の解析

次に、本研究で行った、日銀の金融政策決定会後の総裁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の表情の変化を解析するため、まず、日経チャンネル[2]にて公表されている会見動画を約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。なお、今回分析対象としたデジタルな動画データ自体、アナログな観測情報に離散化処理を施したものであるが、ここでは人間の表情の変化を捉える為には、約0.5秒毎間隔の画像データで十分と考え、再度の離散化を行った。

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 Serviceにおける表情認識アルゴリズムを感情値の計測に用いた[4]。そ

のため、総裁記者会見における表情データ自体を顔認識アルゴリズム作成の為の学習には用いていない点には注意が必要である。

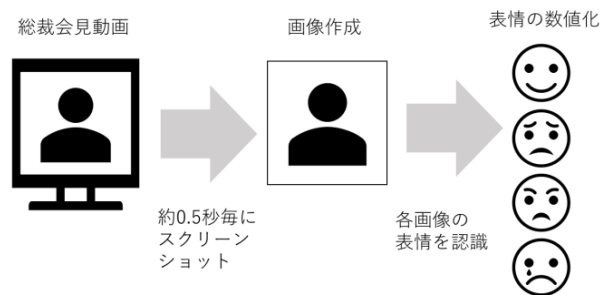


図2: 会見動画の感情値の計測手順

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」等の感情値の計測を行った。そして、各会見において、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。

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

更に、会見毎に算出した各感情値の割合を検証したところ、重大な金融政策変更の前後の記者会見では、総裁の表情に関して特徴的な変化が確認された。

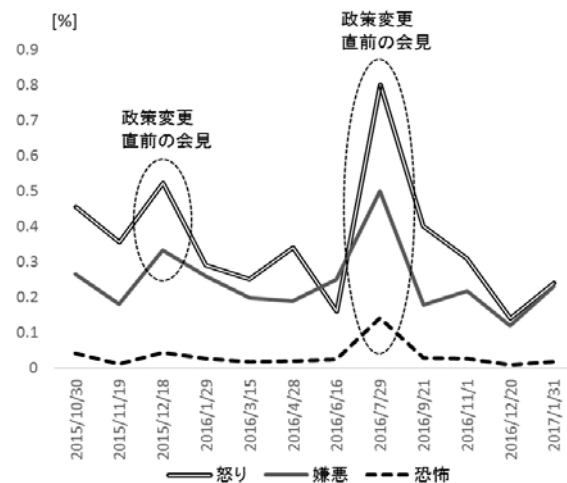


図3: 政策変更前に見られる感情値の変化

例えば、前述したマイナス金利政策の導入や、イールドカーブ・コントロール政策の導入といった、それまでの金融政策のフレームワークを変更するような、重大な金融政策変更を行う直前の記者会見で

は、「怒り」や「嫌悪」の割合が高くなる傾向が見られた(図3)。また、それらの金融政策変更に向けては、「悲しみ」の割合が上昇を続け、金融政策変更の決定を行った後の記者会見では、「悲しみ」の割合が低下する傾向も見られた(図4)。

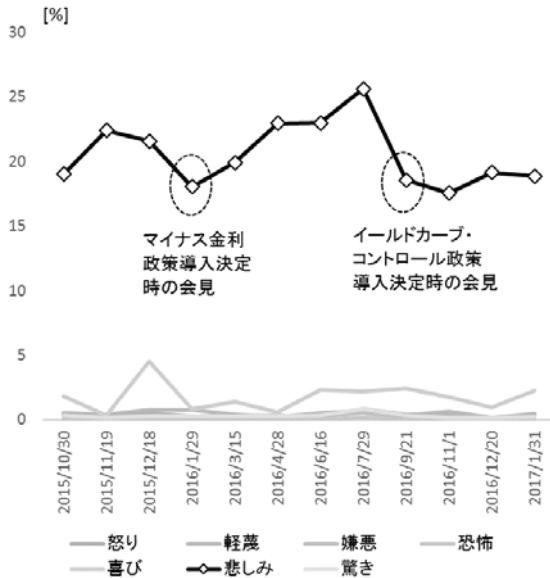


図4: 政策変更後に落ち着きを見せる感情値

このことは、金融政策の変更前に既存の金融政策に対する問題意識の高まりがネガティブな表情として表れていた可能性と、また、金融政策変更によって、それが緩和されたことによる安堵が、表情に表れていた可能性を示していると言えよう。

もちろん、記者会見における総裁の表情の変化は、その発言の内容によるところも大きい点には注意が必要だ。ただし、通常、日銀総裁は公の場において、「怒り」や「嫌悪」といったネガティブな感情を露わにする場面は少ない。本研究では、人工知能技術を用いることで、表情の機微な変化を計量的に測定したことで、総裁の表情に表れたネガティブな感情の高まりを時系列で捉えることができたと考える。

実際、既存の金融政策に対する問題意識が高まることで、新たな金融政策への変更が施される点を踏まえると、政策変更前に、ネガティブな感情が高まることは、理にかなっている。

例えば、2016年1月に決定されたマイナス金利政策の導入だが、当時は、原油価格の下落等を背景に、日銀が目指す2%の物価安定目標が遠ざかる中、日銀は既に大幅な国債買入れ等を中心とした金融緩和政策を実施しており、追加的に取り得る金融政策手段が限られてきた面は否めなかった。そのような中で

の幅を広げた大きな一手であった。

また、2016年9月にはイールドカーブ・コントロール政策の導入が決定されたが、政策変更以前は、金融市場において、マイナス金利政策導入後に進行したイールドカーブの過度な平坦化を背景とした、金融システム機能の圧迫に関する批判が散見されていた。そのような中、日銀は、長期金利を適切な水準に誘導させ、イールドカーブの過度な平坦化を抑える仕組みを導入した。従来であれば、金融政策における金利の操作目標は短期金利であったが、長期金利までも誘導の対象とする金融政策の導入は、これまでの概念を取り払う画期的な政策変更であった。

しかし、既存の金融政策に対して、総裁自身がどの程度問題意識を高めているかを、政策変更以前の発言内容や公式な文書から読み解くのは容易ではない。本研究では、公式な文書等には記録されない、総裁の表情の機微な変化を、人工知能技術を用いて計測することを試みた。そして、そのようにして得られた感情変化の時系列データと、金融政策変更との関係を確認したことで、新たな政策分析アプローチの可能性を示すことができたと考える。

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] 財務省国債金利情報
https://www.mof.go.jp/jgbs/reference/interest_rate/
- [2] 日経チャンネル「Live! 日銀総裁会見」
<http://channel.nikkei.co.jp/>
- [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/>

敵対的学習を応用したエージェントベース・モデルのベイズ推定

Estimation of Agent Based Models by Approximate Bayesian Computation with Adversarial Training

塩野 剛志^{1*}

Takashi Shiono¹

¹クレディ・スイス証券株式会社 経済調査部

¹ Economic Research, Credit Suisse Securities (Japan) Limited

Abstract: 本研究では、Agent Based Model (ABM) の未知パラメータを、ニューラルネットワークの Adversarial Training を応用したベイズ的アプローチによって推定する方法を提案する。ABM のベイズ推定では、(1) パラメータの探索値毎のシミュレーション、(2) シミュレーション生成データと実データを突き合わせた尤度の計算、(3) 事後分布のサンプリング、という計算負荷の大きいプロセスをくり返す。そのため、特に、推定するパラメータ数が複数になる場合、(2) と (3) における効率化が必須となってくる。(2) の尤度計算について、パラメトリックな分布を仮定して省力化する手法もあるが、特定化の誤りによる大きな精度損失が懸念されることから、より柔軟な近似手法への需要がある。本研究では、こうした問題意識から、Adversarial Training によって、ABM の推定において尤度を近似するニューラルネット (判別器) を、現実データから適合的に獲得するアプローチを検討した。

1.はじめに

1.1. DSGE の問題点と ABM への期待

Agent Based Model (ABM) の応用は金融市場のミクロ的構造をシミュレートする人工市場の分野において頻繁に活用されてきたが (例えば、[和泉 03])、近年ではマクロ経済分析の分野でも、その導入が進みつつある。

その背景には、現代のマクロ経済学の主流モデルである Dynamic Stochastic General Equilibrium (DSGE、確率的動学一般均衡) モデルへの批判がある。ABM を用いることで、標準的な DSGE が過度に単純化した現実経済の特徴について、柔軟に取り込むことが期待できるというのだ。

DSGE に対しては、様々な観点からの批判が存在する (例えば、[Blanchard 16a]、[Blanchard 16b]、[Gali 16]、[Keen 16]、[Kocherlakota 16]、[Muellbauer 16])。頻繁に指摘される点として、1) 永遠に生存する代表的家計の仮定²、2) 単一の定常

状態の周りでの線形近似 (複数定常点やストレンジ・アトラクタの無視)、3) 金融システムの不十分な取り扱い、などへの批判がある。

もっとも、最近ではそれぞれの批判を考慮した個別の DSGE が開発されている³ (例えば、[Arouba 13]、[Arouba 17]、[Carroll 17]、[Coimbra 17]、[Gelfer 16]、[Kaplan 16])。どのようなモデルも現実を単純化することで、より重要と考えられるメカニズムの理解に焦点化するものだが、リーマン・ショック以前に隆盛したニュー・ケインジアン型 DSGE (例えば、[Christiano 05]) では、上述のような、ミクロ家計の異質性 (例えば、所得・資産格差) やその相互作用、経済変数間の非線形関係、金融部門と実体経済の関わり、について過度な単純化を行ってしまった。その結果、2008 年の金融危機に際して、ケインズ的な財政出動と金融緩和以外に、特段、有意義な政策対応を示すことができなかった、という

に当てはめてしまったため、ミクロの家計間の相互作用やネットワークの影響を無視しているという批判。なお、多数の家計が平均で代表できるのは、個別家計の idiosyncratic risk がアローの状態依存証券 (保険) 市場における取引で完全にヘッジされていることに等しい。

3様々な批判を同時に回避するような一般的なモデルが定型化されているわけではない。

^{1*} 連絡先: クレディ・スイス証券株式会社 経済調査部, 〒106-6024 東京都港区六本木 1-6-1 泉ガーデンタワー 26 階, Takashi.shiono@credit-suisse.com
2個別家計のミクロ経済学的な行動理論をマクロの家計部門全体

問題意識である。こうした批判は、DSGE が焦点化すべきテーマの選択に失敗したということであり、必ずしも DSGE の代わりに ABM を用いることを動機付けるようなものではない。

これに対して、より本質的な論点は、「均衡」を事前に要請する枠組みへの批判である ([Blanchard 16a]、[Keen 16])。これが DSGE と ABM の根本的な違いでもある ([Grazzini 15])。すなわち、DSGE においては、経済の状態を記述するすべての時点 t において、全ての市場の需要と供給が均衡に達していると考えられる。モデルにおける外生変数（構造ショックと呼ばれる）が確率的に時間変化しており、その実現値を前提に、経済主体は、合理的期待形成の下、每期毎期、それぞれの最適化問題の解に従い、相互無矛盾の均衡点に到達し直している⁴。DSGE では、そうした均衡のスナップショットの連続が、マクロ経済変動の正体である、というのである。

重要なことに、ここでは、市場均衡価格の発見に要する「緩和時間」が、モデルが経済を記述するタイムフレーム（通常、四半期）とは別に存在し、その時間が十分に短いことが、暗黙に想定されている。Dynamic と Equilibrium という一見して二律背反な概念が結び付けられているのは、こうした考え方によるものである。

しかし、現実の経済活動においては、価格調整によって超過需要／供給が解消するには、短くはない時間を要する、と考えるのが自然である。多くの取引において「意図せざる在庫」が確認されているのはその証左の一つであろう⁵。全ての市場が常に需給均衡した状態で存在していると考えられるのは無理がある。主流派マクロ経済学者の DSGE への評価は、

⁴ 各経済主体の合理的期待形成の下での最適化行動が、価格調整による全市場の需給均衡（一般均衡と呼ばれる）を導くのは、各市場において中央集権的な「ワルラスの競り人」が存在し、市場参加者の需要・供給の情報を集約して、市場価格を上げたり下げたりするプロセスを経るからである。もしくは、マーケット・メーカーが存在し、裁定取引によってその役割を代替することが想定される。

⁵ シンプルな DSGE では実証的に観察されるインフレ率の自己回帰性の高さ（粘着性）が十分に再現できない等の問題が報告されている。こうした問題に対して DSGE は通常、合理的期待形成と每期均衡という考え方は維持したまま、取引コストや情報取得の制約を導入するなどして、ヒューリスティックな修正を施している。しかし、こうした齟齬は事前に均衡を要請するという設定に起因しているかもしれない。また、在庫の保有に関して言えば、与えられた情報とコスト条件の下、最適化された判断として合理的・意図的に在庫を持つと考え、意図せざる在庫はないことになる。

多かれ少なかれ、こうした問題意識を共有している。例えば、[Blanchard 16b]は、DSGE をマクロ経済分析における唯一の正当なモデルと看做す気運を止め、個別市場の分析（部分均衡）から得られる知見をより評価することを主張している。また、毎期の全市場の均衡成立（一般均衡）と、それを支えるミクロ的基礎付けに拘る余り、創発的なマクロ現象の説明に無理やりミクロ現象の論理を持ち込むこと ([Korinek 15]) や、経済主体の多様な期待形成や意思決定の可能性を排除すること [Kocherlakota 2016]) は見直すべき、等との指摘がみられる。

1.2. マクロ経済モデルとしての ABM

こうした、より根本的な DSGE への問題意識に対して、ABM はひとつの代替モデルとしての有用性を持つと言えよう。というのも、ABM ではエージェントの異質性を考慮した上で、それぞれの行動パターンを極めて柔軟に設定できる。そのため、毎期の市場均衡や合理的期待形成を事前に要請する必要はない。

実際、マクロ経済を ABM で表現した [Caiani 16] や [Assenza 15] では、超過需要や超過供給に直面したエージェントの調整行動を、モデルの時間ステップに沿った適応的ヒューリスティックを充てることで陽表的に考慮している。

また、これらの研究は、従来のマクロ経済研究において観測されてきた、フィリップス・カーブやオーケン法則のような定型化されたマクロ現象 (Stylized Facts) を、適応的エージェントによる ABM の枠組みから創発的に再現している。さらに、経済主体毎のバランスシートを明示的にモデル化することで、金融部門と実体経済の関わりをより綿密に表現しており、従来の DSGE が見落としてきた、金融機関に対するブレード政策の金融危機や実体経済への影響を議論する際などにも有益である。

この他、ABM では、経済主体に合理的期待形成を要請する必要がないため、行動経済学が示唆するようなエージェントの多様な期待形成や戦略的（ゲーム理論的）ネットワーキングなどをモデル化する可能性も開けてくる。

このように、ABM は、その柔軟性の高さから、DSGE を補完する立場として、有用性が理解されつつあり、今後も様々な経済現象のモデル化に応用されていく可能性がある。

しかし、しばしば、ABM は実証的検証を経ていない点が弱点だと指摘されている（例えば、[Gallegati 09]）。ABM のパラメータは、再現したい特定の経済現象とモデルのシミュレーション結果がフィットするようにカリブレードするのが基本とな

っており、データによる統計的なパラメータ推定が行われることはまだ少ない⁶。これに回答し、最近では ABM のパラメータを統計的に推定する研究が登場し始めた。

以下では、こうした ABM の統計的推定における先行研究を概説したあと、その改良に繋がり得るひとつの手法を提案する。さらに、最小規模の市場価格形成モデルを例として、シミュレーション・データによる実験的な推定結果を報告する。その上で、最後に、提案手法や実験方法の課題を議論する。

2. 先行研究：ABM の統計的推定

[Ghoshadze 16]は、エージェントのマイクロ状態分布の動的発展（誘導形）がフォッカー・プランク方程式として解析的に記述できる類の小規模 ABM について、Generalized Method of Moment (GMM) を用いて推定した。こうしたアプローチは、GMM という確立した回帰分析の体系をそのまま適応するため信頼性は高いものの、当然ながら、解析的な誘導形が導出できる ABM にしか使用できない。そうしたモデルは小規模にならざるを得ず、また、ABM の利点である柔軟で多様なエージェント行動も設定しづらくなる。

これに対して、[Grazzini 15]は、ABM のマクロ集計値がエルゴード性を持つ（マイクロ主体の初期条件に依存しない）ことを要件とした上で、Simulated Minimum Distance (SMD) の方法を用いてパラメータ推定する方法を提案した。これは、ABM のシミュレーションで生成されたマクロ集計データと、現実のデータとの距離を最小化するようにパラメータを決める方法である。距離を測るには、特定のモーメント値や、それぞれのデータを使って推定したメタ・モデル（例えば VAR）のパラメータ値が用いられる。こうした方法は、ABM の生成するシミュレーション・データのエルゴード性、並びに、現実のデータの定常性が担保できる限り、比較的広いクラスの ABM に適用できる可能性がある。

さらに、[Grazzini 17]は、SMD において、特定のモーメントやメタ・モデルを分析者が事前に選択する必要がある点を問題視し、代替的手法を提案した。具体的には、パラメータの事前分布をモデルのデータへの尤度を用いて更新し、事後分布を得るベイズ

⁶こうした状況は成立当初の DSGE と酷似している。DSGE の文脈では、その後、ベイズ統計を駆使した推定の方法論が整備され、実証上のパフォーマンスも向上していった。現在では、データ・フィッティングのなされた DSGE は、NYFED に代表されるように、中央銀行などで定期的な予測に用いられるに至った ([Gelfer 16])。

推定の使用を提案した。この時、尤度はパラメータ値毎のシミュレーション・データの密度分布を用いて評価される (Non-parametric Bayesian)。この方法であれば、推定に使用するモーメントを事前に選ばずに、現実のデータ全体の情報を用いることができる。しかし、モデルのパラメータ数が増えるにつれ、ノンパラメトリックな尤度計算が困難となる。同論文では、計算コスト節約のために、シミュレーション・データの密度分布を正規分布などでパラメタライズする方法 (Parametric Bayesian)、並びに、尤度関数を 1-0 の Indicator 関数で近似する方法 (Approximate Bayesian Computation) も提案されている。しかし、どちらも SMD と同様に、事前の裁量の設計が必要となってくる。

本研究では、この[Grazzini 17]による Approximate Bayesian Computation (ABC) を土台としつつ、比較的柔軟に尤度関数を近似することを目指した新しい推定手法を模索する。詳細は次章で述べるが、尤度関数の近似に使用する Indicator 関数を一種の判別器と看做した上で、Adversarial Training の手法⁷をヒントに、データから適応的に決めていくアプローチを提案する。

3. 提案手法

3.1. ABM の状態空間表現

以下では、[Grazzini 17]に倣って、ABM をベイズ推定する際の枠組みを示す。

まず、ABM をマイクロ・エージェントの状態変数の組 $x_{i,t}$ が以下のように遷移するマルコフ・チェーンと看做す。

$$(式 1) \quad \begin{aligned} x_{i,t+1} &= f_i(x_t, \Xi_t, \theta), \text{ with } x_t \equiv \{x_{i,t}\}, \\ &\text{for } i = 1 \dots N, t = 1 \dots T. \end{aligned}$$

ここで、 f_i はマイクロ状態の実数値 \mathbb{R}^k を取る関数、 Ξ_t は確率的要素をまとめたベクトル、 $\theta \in \Theta$ は ABM のパラメータベクトルである。なお、 Θ は θ の取りうる値を意味するパラメータ空間であり、実数空間 \mathbb{R}^q のコンパクトな部分集合と想定する。また、使用する ABM モデルはエルゴード性を満たすものと想定し、したがって、確率的要素 Ξ_t の影響は時間とともに減衰する。

この表現において、DSGE との違いは関数 f に集約されている。その構造は複雑で非線形性や if-then 分岐などの不連続を含むものと想定される。合理的

⁷例えば、機械学習による画像生成の分野で注目されている GAN (Generative Adversarial Net) が挙げられる。

期待などは要請されず、したがって、エージェントが相互に最適化行動の解に従うという意味での毎期の「均衡」は担保されていない。

なお、ABM のエージェントのミクロ状態 X_t を集計することで K 個のマクロ変数 $y_t \equiv \{y_{k,t}\}, k = 1 \dots K$ が定義されるとする。なお、このマクロ変数は観測可能なデータと対応することが想定される。

(式 2)

$$y_t = m(X_t).$$

式 1 と式 2 から以下のように、 $t+1$ 期のマクロ変数に対する関係を導くことができる。

(式 3)

$$y_{t+1} = g(X_t, \Xi_t, \theta).$$

エルゴード性を持つマルコフ連鎖としての ABM が十分な期間 $t > T^*$ を経た長期では、一種の定常状態に到達しており、システムの状態は状態変数の初期条件 X_0 、並びに、確率要素のランダム・シード s の影響から独立となる。この時、マクロ変数の定常状態が以下のように定義される。

(式 4)

$$y^* = E[y_t | t > T^*] = g^*(\theta).$$

また、現実に観察されるデータと ABM との関係は以下のように定式化される。

(式 5)

$$y_{t+1}^R = g^*(\theta^R, u_t).$$

ここで、 u_t は、誤差ベクトルであり、実データと ABM シミュレーションによるマクロ変数値との差を生む要素（観測誤差 measurement error、モデル定式化誤差 specification error など）をすべてまとめたものとする。

3.2. ABM のベイズ推定

以上の表現体系の下で、パラメータの事前分布 $p(\theta)$ をモデルのデータへの尤度を用いて更新し、事後分布 $p(\theta | Y^R)$ を得るベイズ推定の方法は以下のように示される。

(式 6)

$$p(\theta | Y^R) \propto \mathcal{L}(\theta; Y^R) p(\theta).$$

ここで、パラメータの実データに対する尤度は $\mathcal{L}(\theta; Y^R) \equiv p(Y^R | \theta)$ with $Y^R \equiv \{y_t^R\}, t = 1 \dots T$ となる。特に、ABM の定常状態においては、現実のデータセットの値を観測する確率が、順序不問となり、以下のようにあらわされる。

(式 7)

$$\mathcal{L}(\theta; Y^R) \propto \prod_{t=1}^T f(y_t^R | \theta).$$

実際のベイズ推定では、パラメータ値ごとに実データに対する尤度を計算する。その上で、尤度に応じてパラメータの事後分布を更新して収束させる、というステップが必要となる。

[Grazzini 17]は、ABM の各パラメータごとに尤度を計算する方法について、以下の 3 つの方法を提案している。

- 1) ノンパラメトリックな方法
- 2) パラメトリックな方法
- 3) Approximate Bayesian Computation (ABC)

1) のノンパラメトリックな方法は、各パラメータ値 θ の下での ABM シミュレーションから、マクロ変数 $y_t(\theta)$ の経験密度分布を得てカーネル密度推定 (KDE) 等のスムージングを行い、その密度分布における実データ値 y_t^R の出現確率によって尤度を評価する、というものである。

2) のパラメトリックな方法では、マクロ変数 $y_t(\theta)$ の定常状態 $y^*(\theta)$ 周りのゆらぎ e_t が既知の分布関数（正規分布など）に従うと仮定し、実データとシミュレーション・データの誤差をその分布関数で評価して尤度を求める。

最後に、3) の ABC は、集団遺伝学に端を発した Likelihood-Free と呼ばれる折衷的方法であり、本研究の提案手法はこれを土台としている。パラメトリックな方法よりも柔軟な尤度関数の近似が可能であり、かつ、ノンパラメトリックな方法よりも計算コスト節約をできるという利点がある。具体的には、尤度関数を 1-0 の Indicator 関数で置き換える。Indicator 関数は、ABM が出力するシミュレーション・データと実データの距離が十分に近いと判断されれば 1、そうでなければ 0 を返すもので、以下の 3 段階で構成される。

- (i) 集約統計量 $\mu(\cdot)$ (summary statistics) の決定
- (ii) 距離尺度 $d(\mu, \mu^R)$ (distance measure) の定義
- (iii) 許容閾値 h (tolerance threshold) の設定

その上で、ABC のアルゴリズムは次の 3 ステップの繰り返しとなる。

1. 事前分布から 1 組のパラメータベクトル θ^i を抽出
2. パラメータ θ^i の下で、ABM シミュレーションを実施し、生成データ $Y(\theta^i) \equiv \{y_t(\theta^i)\}, t = 1 \dots T$ を得る
3. 集約統計量を生成データ $\mu(Y(\theta^i))$ と実データ

$\mu(Y^R)$ それぞれについて計算。距離が許容閾値の範囲内 $d(\mu, \mu^R) \leq h$ であれば θ^i を採用、さもなければ棄却する

ステップ3では、尤度の代わりとなる Indicator 関数によってパラメータの事後分布を更新しているが、その際の方法論（所謂、事後分布からのサンプリング）としては、Rejection sampling, Importance sampling, Particle Filter, MCMC などが提案されている。

実務的には、集約統計量の選定が最も困難であり、本質的に、SMD と同様の任意性が問題となる。例えば、平均にしか影響しないパラメータと分散にしか影響しないパラメータが同時に含まれるような場合、どちらのモーメントを集約等計量に選定したかで、パラメータの推定結果が大きく異なるだろう。

3.3. 判別器の敵対的学習による尤度近似

これに対して、提案手法の発想は、この ABC における Indicator 関数に相当する機構を、ニューラルネットを用いた判別器の Adversarial Training（敵対的学習）によって、データ適合的に構成しようということである。それによって、分析者が事前に (i) ~ (iii) の指定をすることなく、ABC 推定することが期待できる。

なお、Adversarial Training は、GAN ([Goodfellow 14]) や DC-GAN ([Radford 16]) など画像生成分野での成功により、近年注目を集めているが、時系列データに対する適用も進みつつある。

例えば、LSTM を Adversarial Training した研究として[Lamb 16]がある。その手法では、時系列データを生成する Generator と、その生成データと実データを比較して判別する Discriminator という2つのニューラルネットを用意する。そして、Generator は、Discriminator が実データと見間違えるような精度の高い理論値を生成するように学習され、Discriminator は実データと生成データをより正確に見分けるように学習される。こうした Adversarial Training のプロセスは、しばしば、きわめて精巧な偽物をつくる贋作家と、それを見破ることが使命の鑑定士との競争にも例えられる。

提案手法の具体的なプロセスは以下の通りである。すなわち、事前分布から抽出したパラメータ θ^i のうち、その下での ABM の生成データ $Y(\theta^i)$ が、判別器 (Discriminator) によって実データ Y^R と見分けにくいと評価されたものほど重点的にリサンプルする（重いウェイトを置く）ことで、事後分布を更新する。

これに対して、判別器は、入力されたデータが、

実データのものか (1)、生成データのもの (0) かを、より正確に分類するように訓練されていく。

つまり、ABM のベイズ推定の側から見れば、ニューラルネットによる判別器を尤度関数の代わりとしており、また同時に、Adversarial Training の側から見れば、ABM を Generator の代わりとしている。

形式的には、以下のような目的関数を最大化するように判別器 $D(\cdot|\xi)$ のパラメータ ξ が調整される。

(式 8)

$$\mathcal{C}(\xi|\theta) = \mathbb{E}_{Y^R} [-\log D(Y^R|\xi)] + \mathbb{E}_{Y \sim g(\theta)} [-\log(1 - D(Y(\theta)|\xi))].$$

こうして学習された判別器 $D(\cdot|\xi, Y^R)$ を Indicator 関数として用いた ABC によって、AMB パラメータ θ の事後分布を更新する。

(式 9)

$$p(\theta|Y^R) \propto \mathcal{L}(\theta; Y^R)p(\theta) \text{ with } \mathcal{L}(\theta; Y^R) = \mathbb{E}_{Y \sim g(\theta)} [-\log D(Y(\theta)|\xi, Y^R)]$$

なお、具体的な判別器のニューラルネット構造には、様々なオプションがあり得る。本研究では、極力単純に、入力されるマクロ変数データの全体 (K 次元 T サンプル) を CNN で縮約した後、線形分類する判別器を用いた。

より複雑なアイデアでは、[Lamb 16]のように、事前に最尤法で実データを学習した LSTM に、実データと生成データの両方を入力し、その隠れ層の時系列を集約されたファクターと看做して判別器に入力する、というような方法も考えられる。ABM が生成するマクロ変数の数が大きく、また、時系列の順序情報も活用したい場合には、こうした隠れ層（潜在ファクター）に対する判別を行う利点も大きいと思われる。

4. 検証実験

4.1. 最小限の市場価格形成モデル

実験では、[Grazzini 17]に倣って、小規模の市場価格形成モデルを使用した。これは、[Cliff 97]によって提案された ABM であり、単一の商品について板情報（市場価格）を参照しながら自身の利ザヤ $\mu_{i,t}$ を調整して指値注文する売り手と買い手で構成される。詳しくは元論文を参照されたいが、簡潔にはエージェント i の指値 $p_{i,t}$ が以下のように更新される。

(式 9)

$$p_{i,t+1} = (1 + \mu_{i,t+1})v_i,$$

$$\mu_{i,t+1} = \frac{p_{i,t} + \Delta_{i,t}}{v_i} - 1,$$

$$\Delta_{i,t} = \beta(\tau_{i,t} - p_{i,t}).$$

ここで、 v_i はエージェント固有で時間不変の取引サイズ。また、 $\tau_{i,t}$ は各エージェントの公開されない目標価格で、直近の市場価格よりも高い場合には引き下げられ、低い場合には引き上げられる。各エージェントに共通のパラメータ β は、目標価格と直近の発注価格の乖離にどの程度敏感に反応するかを決める感応度である。この値が高いほど、エージェントは他のエージェントの価格に敏感であり（相互作用が高い）、市場価格は伸縮的となる。

4.2. シミュレーション・データによる推定法検証

実験では、この感応度パラメータ β の推定を試みる。重要なことに、この ABM において、マクロ変数である市場価格の時間平均はパラメータ β と単調な関係を持たない⁸。他方、その標準偏差には単調増加な関係がみられる（図 1、図 2）。よって、SMD や ABC で推定を行う場合には、市場価格の標準偏差を集約統計量とすべきある。逆に言えば、集約統計量として市場価格の平均値を用いると、パラメータ空間に対して目的関数（尤度）がフラットとなって正確な推定は困難となる。

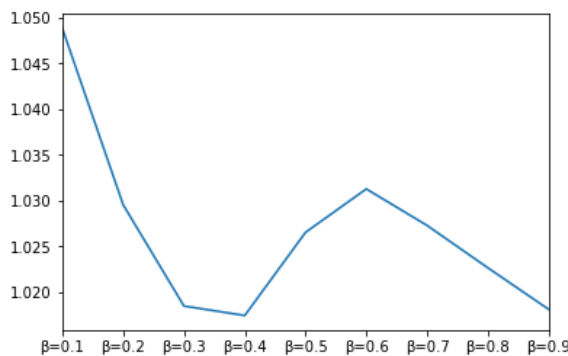


図 1：市場価格の平均と β の関係

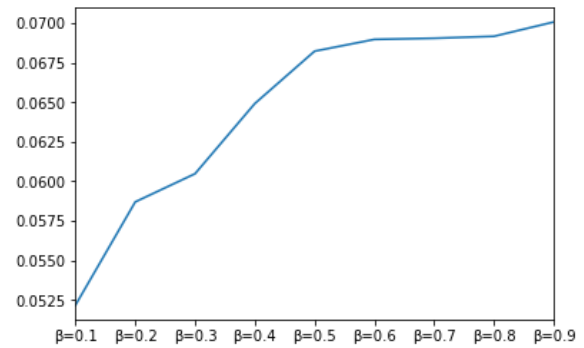


図 2：市場価格の標準偏差と β の関係

今回の小規模 ABM では推定すべきパラメータがひとつであり、マクロ変数のモーメントとの対応を事前に確認することができるので、集約統計量の選定に困難はない。しかし、前述の通り、推定したいパラメータが複数ある場合、どのような集約統計量を用いるのが最も識別に有利なのか、事前に明らかではない。

それに対して、本研究の提案手法は、そうした集約統計量を指定しないデータ適合的なアプローチである。したがって、提案手法を用いた以下の検証実験において、ある程度正確な推定が確認できれば、中規模のマクロ経済 ABM のような、未知パラメータが複数あるモデルに対しても応用可能性が開けるだろう。

検証実験では、事前に指定した β の値に基づき ABM をシミュレートして得た市場価格を、疑似的な実データ (Y^R) とする⁹。その上で、提案手法により推定した $\hat{\beta}$ が、疑似データ生成に使用した値 β に十分近いかを確認する。

5. 結果の解釈と課題

実際に、 $\beta = \{0.1, 0.9\}$ の 2 つの指定値について、提案手法によって推定した結果が図 3、図 4 である。なお、事後分布からのサンプリングには Particle Filter を用いており、事前分布は一様分布 $U(0.0, 1.0)$ である。

⁸区間 $\beta \in (0.0, 1.0)$ について 0.1 刻み。パラメータ値 1 つについて ABM シミュレーションを 1000 回実行し、その平均値を図示。

⁹ エージェントの数は売り手・買い手ともに 20、取引サイズは $v_i \sim \mathcal{N}(1, 0.1)$ で指定。シミュレーションの繰り返し期間は 800 期であり、その内、はじめの 300 期分は初期値の影響が残ると看做し破棄。

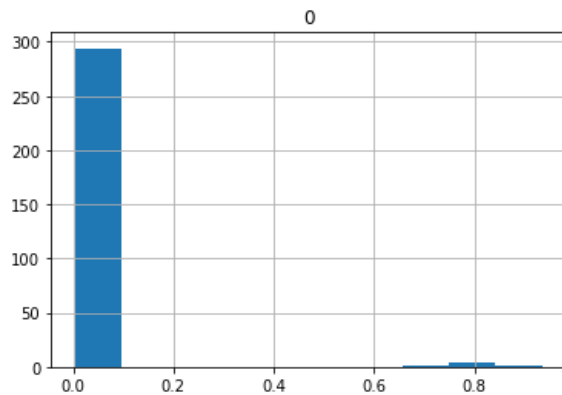


図 3：推定パラメータ $\hat{\beta}$ の事後分布
($\beta = 0.1$ の場合)

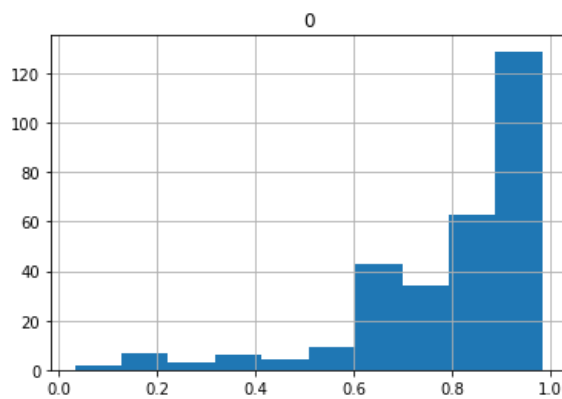


図 4：推定パラメータ $\hat{\beta}$ の事後分布
($\beta = 0.9$ の場合)

上図から分かるように、推定値 $\hat{\beta}$ の事後分布は、それぞれ 0.1 と 0.9 の近傍にピークをもっている。すなわち、SMD や ABC のように集約統計量を指定することなく、ニューラルネットの判別器が尤度の近似を成し、最低限必要な推定は行われたと言える。

しかし、事後分布の形状には縮退が生じており、不安定な結果であると言わざるを得ない。今回は、パラメータ数が一つであることと、計算リソースの都合上、Particle の数を 300 と小さく抑えていたため、縮退が強くなった可能性が高い。今後、推定パラメータ数が十数個に及ぶ中規模マクロ経済 ABM などの推定を行う場合には、10,000 程度のかかなり大きなパーティクル数が望ましいだろう。

また、判別器の Adversarial Training における反復 (Epoch) 数は、計算環境の都合上 100 回に抑えているが、望ましくは、Epoch 数の大小による事後分布の差を確認する必要があるだろう。

判別器の学習と、ベイズ推定による ABM パラメータの事後分布更新を同時進行するため、当然ながら、収束には相応の繰り返しが必要と考えられる。

仮説的な推論に過ぎないが、十分な収束を目指した場合、ノンパラメトリックなベイズ推定と比較した計算コストの削減効果は、パラメータが多い ABM でしか有益なものにならない可能性があるだろう。

総じて、本研究では、ABM のパラメータ推定における、Adversarial Training を応用したベイズ推定のアプローチに、一定の可能性が示唆されたと言える。

しかし、今回の実験では、計算リソースの限界から十分な検証が出来たとは言い難く、結果の解釈には注意が必要だと考える。今後の課題として、計算環境の改善や、より時間をかけた追検証を継続していきたい。また、主要な問題意識である中規模マクロ経済 ABM の実際の経済データによる推定についても、同時に取り組んでいきたいと考えている。

参考文献

- [Arouba 13] Arouba, S.B., and Schorfheide, F.: Macroeconomic Dynamics Near the ZLB: A Tale of Two Equilibria, *FRB of Philadelphia Working Paper*, No. 13-29 (2013), <http://dx.doi.org/10.2139/ssrn.2294960>
- [Arouba 17] Arouba, S.B., Bocola, L., and Schorfheide, F.: Assessing DSGE Model Nonlinearities, *Journal of Economic Dynamics and Control*, Forthcoming (2017), <https://doi.org/10.1016/j.jedc.2017.07.006>
- [Assenza 15] Assenza, T., Gatti, D.D., and Jakob, G.: Emergent dynamics of a macroeconomic agent based model with capital and credit, *Journal of Economic Dynamics and Control*, Vol.50, pp.5-28 (2017), <https://doi.org/10.1016/j.jedc.2014.07.001>
- [Blanchard 16a] Blanchard, O.: Do DSGE Models Have a Future?, *Peterson Institute for International Economics POLICY BRIEF*, PB 16-11 (2016), <https://piie.com/system/files/documents/pb16-11.pdf>
- [Blanchard 16b] Blanchard, O.: Further Thoughts on DSGE Models, *Peterson Institute for International Economics REALTIME ECONOMIC ISSUES WATCH*, (2016), <https://piie.com/blogs/realtime-economic-issues-watch/further-thoughts-dsge-models>
- [Caiani 16] Caiani, A., Godin, A., Caverzasi, E., Gallegati, M., Kinsella, S., and Stiglitz, J.E.: Agent based-stock flow consistent macroeconomics: Towards a benchmark model, *Journal of Economic Dynamics and Control*, Vol.69, pp.375-408 (2016), <https://doi.org/10.1016/j.jedc.2016.06.001>
- [Carroll 17] Carroll, C., Slacalek, J., Tokunaka, K., and Whilte, M.N.: The Distribution of Wealth and the Marginal Propensity to Consume, *Quantitative Economics*,

- Forthcoming (2017),
<http://www.econ2.jhu.edu/people/ccarroll/papers/cstwMPC/>
- [Christiano 05] Christiano, L.J., Eichenbaum, M., and Evans, C.L.: Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy, *Journal of Political Economy*, Vol. 113, No.1 (2005),
<http://www.journals.uchicago.edu/doi/abs/10.1086/426038>
- [Cliff 97] Cliff, D., and Bruton, J.: Minimal-intelligence Agents for Bargaining Behaviors in Market Based Environments, *HP Laboratories Bristol*, HPL-97-91 (1997)
- [Coimbra 17] Coimbra, N., Rey, H.: Financial Cycles with Heterogeneous Intermediaries, *NBER Working Paper*, No. 23245 (2017),
<http://www.nber.org/papers/w23245>
- [Gali 16] Gali, J.: Some Scattered Thoughts on DSGE Models, *mimeo*, (2016),
http://www.crei.cat/wp-content/uploads/2016/07/dsge_ebook.pdf
- [Gallegati 09] Gallegati, M., and Richiardi, M.: Agent Based Models in Economics and Complexity, *Complex Systems in Finance and Econometrics*, (2009),
https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-7701-4_3
- [Gelfer 16] Gelfer, S.: Financial Crises and Labor Market Dynamics: Evidence from a Data-Rich DSGE Model, *mimeo*, (2016),
https://docs.wixstatic.com/ugd/a9dfbd_d7c79b0040e44efbdeb17cdd47580c8.pdf
- [Ghoshadze 16] Ghoshadze, J., and Lux, T.: Bringing an Elementary Agent-Based Model to the Data: Estimation via GMM and an Application to Forecasting of Asset Price Volatility, *Journal of Empirical Finance*, Vol.37, pp.1-19, (2016),
<https://doi.org/10.1016/j.jempfin.2016.02.002>
- [Goodfellow 14] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks, *mimeo*, (2014),
<https://arxiv.org/abs/1406.2661>
- [Grazzini 15] Grazzini, J., and Richiardi, M.: Estimation of Ergodic Agent-Based Models by Simulated Minimum Distance, *Journal of Economic Dynamics and Control*, Vol.51, pp. 148-165 (2015),
<https://doi.org/10.1016/j.jedc.2014.10.006>
- [Grazzini 17] Grazzini, J., and Richiardi, M.: Bayesian Estimation of Agent-based Models, *Journal of Economic Dynamics and Control*, Vol.77, pp. 26-47 (2017),
<https://doi.org/10.1016/j.jedc.2017.01.014>
- [Kaplan 16] Kaplan, G., Moll, B., Violante, G.L.: Monetary Policy According to HANK, *NBER Working Paper*, No. 21897 (2016),
<http://www.nber.org/papers/w21897>
- [Keen 16] Keen, S.: Olivier Blanchard, Equilibrium, Complexity, And the Future of Macroeconomics, *Forbes* Oct 4 (2016),
<https://www.forbes.com/sites/stevekeen/2016/10/04/olivier-blanchard-equilibrium-complexity-and-the-future-of-macroeconomics/3/#1aeab5b43e5f>
- [Kocherlakota 16] Kocherlakota N.: Toy Models, *mimeo*, (2016),
<https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbmxb2NoZXJsYWtvdGEwMDI8Z3g6MTAyZmZlODcxNGZiOGY4Yg>
- [Korinek 15] Korinek A.: Thoughts on DSGE Macroeconomics: Matching the Moment, But Missing the Point?, *mimeo*, (2015),
<https://www.ineteconomics.org/uploads/downloads/Korinek-DSGE-Macro-Essay.pdf>
- [Lamb 16] Lamb, A., Goyal, A., Zhang, Y., Zhang, S., Courville, S., Bengio, Y.: Professor Forcing: A New Algorithm for Training Recurrent Networks, *NIPS 2016 Accepted Papers*, (2016),
<https://arxiv.org/abs/1610.09038v1>
- [Muellbauer 16] Muellbauer, J.: Macroeconomics and consumption: Why central bank models failed and how to repair them, *VOX Columns*, (2016),
<http://voxeu.org/article/why-central-bank-models-failed-and-how-repair-them>
- [Radford 16] Radford, A., Metz, L., and Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, *ICLR2016 Accepted Papers*, (2016),
<https://arxiv.org/abs/1511.06434>
- [和泉 03] 和泉 潔: 人工市場—市場分析の複雑系アプローチ, 森北出版 (2003),
<https://www.morikita.co.jp/books/book/2239>