

金融情報学研究会 (第18回)

日時 2017年3月10日(金)

会場 FinGate (東京茅場町)

SIG-FIN
JSAI Special Interest Group on
Financial Informatics

人工知能学会
金融情報学研究会

第18回 人工知能学会 金融情報学研究会 (SIG-FIN)

- 1、人工市場シミュレーションを用いたバッチオークションの分析.....1
水田孝信(スパークス・アセット・マネジメント株式会社), 和泉潔(東京大学大学院 工学系研究科)
- 2、人工市場シミュレーションを用いたレバレッジド ETF が原資産価格変動に与える影響分析.....9
八木勲(神奈川工科大学 情報学部), 水田孝信(スパークス・アセット・マネジメント株式会社)
- 3、人工市場を用いた分散投資規制が資産価格急落時の市場に与える影響分析.....16
野崎淳(神奈川工科大学大学院 情報工学専攻), 水田孝信(スパークス・アセット・マネジメント株式会社), 八木勲(神奈川工科大学 情報学部)
- 4、心理的要素を考慮した投資行動モデル.....25
宮坂純也(東京都市大学大学院 工学研究科), 穴田一(東京都市大学 知識工学部)
- 5、Analysis of group behavior bias in Financial Markets using artificial market.....32
Yating Wang (The University of Tokyo), Fujio Toriumi (The University of Tokyo)
- 6、深層学習と拡張手がかり表現による業績要因文への極性付与.....38
酒井浩之(成蹊大学 理工学部), 坂地泰紀(成蹊大学 理工学部), 山内浩嗣(三菱 UFJ トラスト投資工学研究所), 町田亮介(三菱 UFJ トラスト投資工学研究所), 阿部一也(三菱 UFJ トラスト投資工学研究所)
- 7、経済テキストデータを用いた極性概念辞書構築とその応用.....44
伊藤友貴(東京大学大学院 工学系研究科), 坪内孝太(ヤフー株式会社), 山下達雄(ヤフー株式会社), 和泉潔(東京大学大学院 工学系研究科)
- 8、Cross-lingual news article comparison using bi-graph clustering and Siamese-LSTM.....52
Enda Liu (The University of Tokyo), Kiyoshi Izumi (The University of Tokyo), Kota Tsubouchi (Yahoo! JAPAN), Tatsuo Yamashita (Yahoo! JAPAN)
- 9、銘柄固有の金融極性辞書の構築.....58
関和広(甲南大学), 柴本昌彦(神戸大学)

10、株主招集通知における議案別の開始ページの推定.....	65
高野海斗(成蹊大学), 酒井浩之(成蹊大学), 坂地泰紀(成蹊大学), 和泉潔(東京大学), 岡田奈奈(株式会社日経リサーチ), 水内利和(株式会社日経リサーチ)	
11、ネットワークの表現学習を用いた金融専門極性辞書の構築.....	70
伊藤諒(東京大学大学院 工学系研究科), 和泉潔(東京大学大学院 工学系研究科), 須田真太郎(三菱UFJトラスト投資工学研究所)	
12、人工知能を用いた金融政策予想と市場予測分布に基づく為替の投資戦略.....	77
上田翼(三井住友アセットマネジメント株式会社), 東出卓朗(三井住友アセットマネジメント株式会社(出向元:三井住友銀行), 一橋大学大学院)	
13、金融レポート、およびマクロ経済指数によるリアルタイム日銀センチメントの予測.....	81
余野京登(東京大学大学院 工学系研究科), 和泉潔(東京大学大学院 工学系研究科)	
14、深層学習と高頻度データを用いた株式注文状況の推定.....	84
田代大悟(東京大学大学院 工学系研究科), 和泉潔(東京大学大学院 工学系研究科)	
15、Predicting stock fluctuations using Two-level Mapping and SCW.....	89
Muhtar Fukuda (Faculty of Environmental and Information Studies, Nagoya Sangyo University)	
16、Improvement of Support Vector Machine Applied to Limit Order Book.....	97
Hayato Kijima(東邦大学大学院 理学研究科), Hideyuki Takada(東邦大学 理学部)	
17、マルチタスク最大マージントピックモデルによる収益予測.....	102
中川雄太(神戸大学大学院 システム情報学研究科), 上野良輔(神戸大学大学院 システム情報学研究科), 江口浩二(神戸大学大学院 システム情報学研究科)	
18、A discussion on how to measure and deal with some global geopolitical risk for investment risk management.....	110
Noboru Nishiyama (Dragons' Desk Limited, 千葉商科大学)	
19、AIトレーダー コンピュータービジョン編.....	114
河合継(クリスタルメソッド株式会社), 小澤昂(東京工業大学 数学科)	

- 20、深層生成モデルによる時系列ネットワークの低次元埋め込み.....120
 円道滉一郎(神戸大学大学院 システム情報学研究科), 江口浩二(神戸大学大学院 システム情報学研究科), 羽森茂之(神戸大学大学院 経済学研究科), 金京拓司(神戸大学大学院 経済学研究科)

人工市場シミュレーションを用いたバッチオークションの分析

水田 孝信*1 和泉 潔*2
Takanobu Mizuta Kiyoshi Izumi

*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 東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

近年、取引市場同士の競争や投資家の要望などにより、取引市場のシステムの高速化が進んだ。取引システムの高速化により、流動性を供給して利益を得るマーケットメーカー戦略の注文量が以前より増え流動性が向上したという評価がある一方、投資家同士の取引のスピード競争を招き、その競争のために費やされたシステムコストは他の投資家に転化されるという批判がある。高速な取引の優位性を無効にする取引方式としてバッチオークション方式が提案されているが、流動性を供給するマーケットメーカー戦略は損益のリスクが高くなるため継続することが難しくなり、流動性の供給が減ってしまい、むしろ投資家の取引コストが上昇する可能性があるという批判もある。本研究では、ザラバ方式とバッチオークション方式を比較可能な人工市場シミュレーションを用いて、マーケットメーカー戦略の損益のリスクを分析することによりその存続可能性を議論し、マーケットメーカー戦略がバッチオークション方式においても流動性を供給し続けることができるかどうかを調べた。その結果、板寄せ間隔が大きくなると、取引成立率が減少し流動性供給が減少する可能性が示された。さらに、板寄せ間隔が大きくなると、オーバーナイトのポジションをゼロにすることが難しく上にポジションも大きく価格変動リスクが大きくなり、リスクをおさえてマーケットメーカー戦略を継続することが困難になる可能性が示された。さらに、ザラバ方式のときのみに、リスクに見合った収益を得る可能性がある可能性が示され、バッチオークション方式のときは、少なくともザラバ方式では機能したマーケットメーカー戦略では、リスクに見合った収益をあげるのは難しくなる可能性が示された。これらの示唆は、ザラバ方式では流動性を供給していたマーケットメーカー戦略が、バッチオークション方式になるとその供給を維持できなくなり、それらが撤退することにより、流動性が低下する可能性を暗示しているとも考えることができる。

1. はじめに

2000年代から2010年代初頭にかけて、取引市場同士の競争や投資家の要望などにより、取引市場のシステムの高速化が進んだ*1。取引システムの高速化により、売り買いともに注文を常に提示して他の投資家に取引機会(流動性)を提供することで提示した売り買いの価格差(注文スプレッド)分の利益を得るマーケットメーカーという投資戦略の注文量が以前より増え、流動性が向上したという評価がある。一方で、過度な高速化は取引市場のシステムコストを増大させるだけでなく、投資家同士の取引のスピード競争を招き、その競争のために費やされたシステムコストは取引コストとして他の投資家に転化されているという批判がある [Farmer 12, Budish 15]。

[Budish 15] は、そのような投資家の取引の高速化競争を終わらせる方法として、高速な取引の優位性を無効にする取引方式としてバッチオークション方式を提案した。現在多くの取引市場で、売り手と買い手の双方が価格を提示し売り手と買い手の提示価格が合致するとその価格で直ちに取引が成立する、ザラバ方式(連続ダブルオークション方式)が採用されている。一方、バッチオークション方式では、ある程度の期間、例えば数秒といった期間、注文を成立させず注文の受付のみを行い、その期間が終わると集まった注文をつき合わせて取引価格の決定を行う。こうすることにより、高速な取引の優位性がなくなるため、投資家同士のスピード競争をしなくてよくなるという主張である。

[Budish 15] はさらに、実証分析と簡単なモデルによる分析により、ザラバ方式では非常に短い期間のみ裁定機会が存在するが、バッチオークション方式ではそれがなくなり、高速化競

争をしなくてすむと主張した。また [Fricke 15] は簡単なモデルを用いて、ボラティリティ(価格変動の大きさ)がもっとも小さくなるなどの条件から、バッチオークション方式の最適な注文受付期間を議論した。さらに [Manahov 16] は簡単なモデルを用いた実証分析からレイテンシーアービトラージ*2が可能な時間スケールを算出し、それを防ぐ観点からバッチオークション方式の最適な注文受付期間を議論した。

一方、バッチオークション方式には批判も多い。[大墳 14] は、流動性を供給して利益を得るマーケットメーカー戦略において、バッチオークション方式だといくらで取引が成立するか見込みづらくなるためこの戦略の損益のリスクが高くなり、継続することが難しくなり、流動性の供給が減ってしまう可能性を指摘している。流動性の供給が減ると、最良の買い注文価格と売り注文価格の差が大きくなるなどして、むしろ他の投資家の取引コストが上昇することや、そもそも取引できる機会が減ってしまうことが考えられる*3。

実際、[Budish 15, Fricke 15] のモデルでは、流動性を供給する投資家の損益は考慮しておらず、[大墳 14] が指摘する流動性を供給する投資家の存続可能性を取り扱っていない。

流動性を供給する投資家の存続可能性を取り扱った実証研究として [Bellia 15] がある。[Bellia 15] は参考になる事例として、東京証券取引所において、取引開始・終了時に行われる板寄せ方式のときとその間に行われるザラバ方式のときの注文データを比較する実証分析を行い、マーケットメーカー戦略

連絡先: 水田 孝信, mizutata@gmail.com

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

† 本稿は JPX ワーキングペーパー [水田 16a] を再構成したものである。

*1 優れたレビューとして [清水 13] がある。また、取引市場間の競争における高速化の重要性についての解説に [水田 12] がある。

*2 大口注文を複数の市場に分割して出したとき、各注文が各市場に到達するまでにかかるわずかな時間差を利用して、他の投資家が裁定取引を行うこと。特に米国では、他の市場の方が有利な価格で取引できる注文を受けた市場は、その有利な価格で取引できる市場にその注文を回送する義務があり、この回送の速度が遅いために裁定機会が発生することがあり、その是非が大きな議論となっている。日本では米国ほど市場は分断されておらずこのような回送義務もないため、このような裁定機会ほとんどないと考えられている。詳しくは [大墳 14]。

*3 [大墳 14] はさらに、バッチオークション方式では注文受付期間の最後の方に注文が集中して従来以上の高速化競争になる可能性も指摘している。

を行っていると考えられる投資家は板寄せ方式のときよりもザラバ方式のときのほうが取引への参加率が高いことを示した。この戦略は取引の成立が即座に行われることが重要であり、仮にバッチオークション方式になった場合、流動性を供給する取引参加が減る可能性を指摘している。

ザラバ方式からバッチオークション方式への変更という事例はないため実証研究は不可能である。また、先に述べたような参考になる事例はあるものの、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、実証研究では制度の違いの効果だけを取り出すことは困難である。

このような実社会で導入されたことがないシステムを検証する方法として、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という手法がある。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^{*4}。

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。人工市場モデルを用いたシミュレーション(人工市場シミュレーション)を用いれば、これまでに導入されたことがない金融市場の規制・制度も議論することができるうえ、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。

近年、学术界のみならず金融の規制当局や取引所関係者も、金融市場の規制や取引所の制度などを分析する人工市場シミュレーションに興味を示し始めている。実際、NATURE 誌と並んで最も権威がある学術雑誌である SCIENCE 誌に掲載された [Battiston 16] では、“2008 年の金融危機以降、経済や金融市場を理解するために、ネットワーク理論や(人工市場モデルを含む) エージェントベースドモデルといった複雑系理論を用いた手法に関心が集まってきている”と述べ、さらに“金融の複雑系に潜む正のフィードバック現象を弱くし金融システムの安定化させるような、政策や規制はどのようなものかの知見を得られる。”と述べている。

そして、多くの人工市場シミュレーション研究が新しい金融市場の規制や制度、新しい方式の市場を分析し、どのような規制や制度が良いかという議論に貢献した^{*5}。

JPX ワーキングペーパーにおいても人工市場シミュレーションを用いた金融市場の制度変更の影響を分析したものがいくつかある。[水田 13] は呼び値の刻みの変更が価格形成に与える影響を、[草田 15a, 草田 15b] はマーケットメーカーの有無が流動性に与える影響を、[水田 15, Mizuta 16c] は取引所システムの高速化が市場効率性に与える影響を調べた。

しかしながら、人工市場シミュレーションを用いて、バッチオークション方式の市場においてマーケットメーカー戦略の損益のリスクを分析し、その存続可能性を議論した研究はない。

そこで本研究では、[草田 15a, 草田 15b] の人工市場モデルをベースに、板寄せ間隔 δt というパラメーターを新たに導入することにより、ザラバ方式 ($\delta t = 1$) とバッチオークション方式 ($\delta t > 1$) を連続的に変更できる価格決定メカニズムを実装し、マーケットメーカー戦略の損益のリスクを分析することによりその存続可能性を議論し、マーケットメーカー戦略が流動性を供給し続けることができるかどうかを調べる。

2. 人工市場モデル

[Chiarella 02] では、シンプルでありながら、実証分析で得られた長期間に存在する価格変動の統計的性質 (stylized fact) を再現できるエージェントモデルの構築に成功している。[水田 13] では、[Chiarella 02] のモデルをベースにモデルを構築し、[Chiarella 02] のモデルでは再現されていなかった約定件数やキャンセル率、1 ティックごとの騰落率の標準偏差など、高頻度な時間スケールでの性質 (マーケット・マイクロ・ストラクチャー) も再現した。[草田 15a, 草田 15b] では、[水田 13] のモデルに、[Nakajima 04] をベースにモデル化したマーケットメーカー戦略を行うエージェントを追加した。

本研究では [草田 15a, 草田 15b] の人工市場モデルをベースに、板寄せ間隔 δt というパラメーターを新たに導入することにより、ザラバ方式 ($\delta t = 1$) とバッチオークション方式 ($\delta t > 1$) を連続的に変更できる価格決定メカニズムを実装した。本研究の目的にはモデルがシンプルであることはとても重要である。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 価格決定メカニズム

本モデルでは、板寄せ間隔 δt を導入することにより、ザラバ方式 ($\delta t = 1$) とバッチオークション方式 ($\delta t > 1$) を連続的に変更できる価格決定メカニズムを構築した。

ザラバ方式は、売り手と買い手の双方が価格を提示し、売り手と買い手の提示価格が合致するとその価格で直ちに取引が成立する方式である。バッチオークション方式では、ある程度の期間注文を成立させず注文の受付のみ行う。その期間が終わると、買い手は高い注文価格の注文から売り手は安い注文価格の注文から順に取引を成立させていき、付き合わせる注文がなくなるまで繰り返したところで取引価格の決定を行う。この注文の付け合せと取引価格の決定を行う作業を板寄せとよぶ。

本研究ではノーマルエージェント (NA) が注文を出すごとに時刻 t が 1 増える。板寄せの時間間隔を δt とし、 $\delta t = 1$ のとき、注文が入るたびに板寄せが行われてザラバ方式と一致するようにモデル化した。時刻 t に板寄せが行われたとき、残った注文の仲値 (最も高い買い注文の価格と最も安い売り注文の価格の平均) を取引価格 P^t とする。また、板寄せが行われない時刻 t においても、仮にその時刻に板寄せをした場合に決定される取引価格 (取引見込み価格) を求めそれを P^t とする。これにより、 δt がいかなる値であっても、すべての時刻 t において、取引価格 (取引見込み価格) P^t が連続的に算出される。

図 1 に本研究の価格決定メカニズムの例を示す。図 1 の上の段は $\delta t = 1$ (ザラバ方式) の場合、下の段は $\delta t = 4$ (バッチオークション方式) であり、左から時刻 $t = 0, 1, 2, 3, 4$ のときの注文状況を示している。 $t = 0$ では、 $\delta t = 4$ でも板寄せが起きた直後であり、同じ注文状況であったとする。 $t = 1$ において、価格 99 の売り注文 (新規注文) があった。 $\delta t = 1$ では、この新規注文は価格 99 の買い注文と取引が成立しこの買い注文は消える。一方、 $\delta t = 4$ では板寄せを行う時刻ではないため、新規注文を残す。同様に、 $t = 2, 3$ において、新規注文として、価格 100 の買い、価格 101 の買いの注文が来るが、 $\delta t = 1$ の場合はいずれも取引が成立し、 $\delta t = 4$ の場合は板寄せを行う時刻ではないため、新規注文を残す。 $t = 4$ において、新規注文として価格 98 の売り注文が来る。 $\delta t = 1$ の場合は、やはり取引が成立する。 $\delta t = 4$ においては、 $t = 4$ は板寄せを行う時刻のため、買い手は高い注文価格の注文から売り手は安い注文価格の注文から順に取引を成立させていき、付き合わせる注文がなくなるまで繰り返す。その結果、価格 101 の買いと価格

*4 例えば、[出口 09, 実世 12] などが詳しい。

*5 優れたレビューとして、[LeBaron 06, Chen 12, 和泉 12, Cristelli 14, Mizuta 16b]。また、人工市場シミュレーションの長所および限界については、本稿の付録“モデル構築の基本理念”も参照。

	新規注文 時刻 $t=0$	売り 99 $t=1$	買い 100 $t=2$	買い 101 $t=3$	売り 98 $t=4$
ザラバ方式 $\delta t=1$	売り 価格 買い	売り 価格 買い	売り 価格 買い	売り 価格 買い	売り 価格 買い
	1 101	1 101	1 101	1 101 1	1 101
	1 100	1 100	1 100 1	100	100
	99 1	1 99 1	99	99	99
	98 1	98 1	98 1	98 1	1 98 1
		即座に 取引成立	即座に 取引成立	即座に 取引成立	即座に 取引成立
バッチオーク ション方式 $\delta t=4$	売り 価格 買い	売り 価格 買い	売り 価格 買い	売り 価格 買い	売り 価格 買い
	1 101	1 101	1 101	1 101 1	1 101 1
	1 100	1 100	1 100 1	1 100 1	1 100 1
	99 1	1 99 1	99 1	99 1	1 99 1
	98 1	98 1	98 1	98 1	1 98 1
		取引成立 させない	取引成立 させない	取引成立 させない	特定の時刻が 来ると取引成立

図 1: 価格決定メカニズムの例

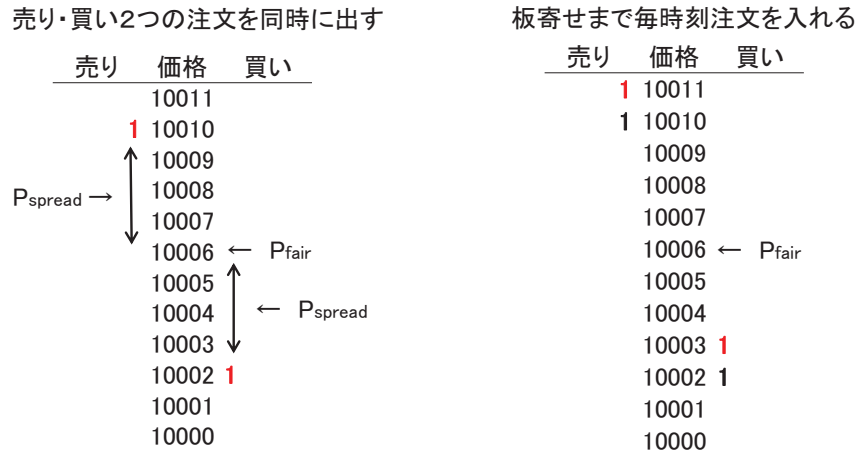


図 2: マーケットメーカーエージェント (MM) の注文

98 の売り, 価格 100 の買いと価格 99 の売りがそれぞれ付け合わされ, 取引は価格はそれらの注文を削除した後の仲値である 99.5 で決定する。

ここで $t = 4$ において残った注文が異なることに注目すべきである。 $\delta t = 1$ の場合は, すべての注文がなくなっており, $\delta t = 4$ の場合は, 4 つの注文が残っている。そして, $\delta t = 1$ の場合は, 数量 4 の買い注文が成立しているのに対して, $\delta t = 4$ の場合は, 数量 2 の買い注文のみ成立している。このように δt に応じて, 価格形成や注文状況は異なるものとなる。

2.2 エージェント

本モデルは 1 つの証券のみを取引対象とする。 n 体のノーマルエージェント (NA) と 1 体のマーケットメーカーエージェント (MM) がある。 NA はエージェント番号 $j = 1$ から順番に $j = 2, 3, 4, \dots$ と注文を出す。 MM は NA が注文を出す直前に買いと売りの 2 つの注文を出す。 図 2 に示すように, MM は板寄せが行われるまで毎時刻注文を入れ, 板寄せ直後にすべての注文をキャンセルする。 NA の注文は, 注文を行ってからキャンセル時間 t_c だけ経過した場合キャンセルされる。最後の NA, $j = n$ が注文を出すと, 次の時刻にはまた初めの NA,

$j = 1$ から注文を出し繰り返される。時刻 t は 1 体の NA が注文を出すごとに 1 増える。

注文数量は常に 1 と一定とする。また, いずれのエージェントも保有する資産の数量に制限はなく (キャッシュが無量大), マイナスの保有数量 (空売り) にも制限はない。価格の変化幅の最小単位 (呼値の刻み) は δP とし, それより小さい端数は, 買い注文の場合は切り捨て, 売り注文の場合は切り上げる。

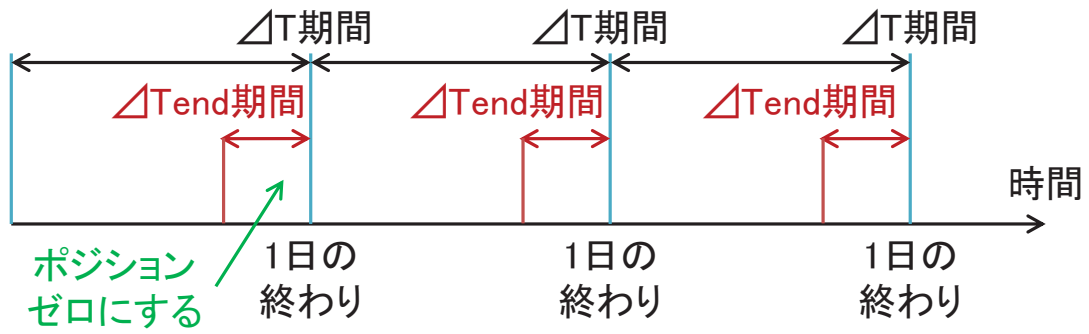
2.2.1 ノーマルエージェント (NA)

NA は, 実際の市場の価格形成の性質を再現するために導入するものであり, stylized fact や高頻度取引にかかわる統計量を再現するなるべくシンプルな, ごく一般的な投資家をモデル化したものとした。

NA は注文価格 $P_{o,j}^t$, 売り買いの別を以下のように決める。時刻 t に NA, j が予想する価格の変化率 (予想リターン) $r_{e,j}^t$ は,

$$r_{e,j}^t = \frac{1}{w_{1,j} + w_{2,j} + u_j} \left(w_{1,j} \log \frac{P_f}{P^t} + w_{2,j} r_{h,j}^t + u_j \epsilon_j^t \right). \quad (1)$$

ここで, $w_{i,j}$ は時刻 t , NA, j の i 項目の重みであり, シミュ



PMM3

ポジションが増加する注文を出さない

最後 ΔT 期間、売り持ちの場合

売り	価格	買い
×	10011	
出さない ↑	10010	
	10009	← P _{fair}
	10008	
	10007	↓

PMM4

ポジションが減少する注文を反対側の価格で出す

最後 ΔT 期間、売り持ちの場合

売り	価格	買い
	10011	↓ ←ここに出す
	10010	
	10009	← P _{fair}
	10008	
	10007	

図 4: マーケットメーカーエージェント (MM) がポジションを閉じる時間帯に出す注文

レーション開始時に、それぞれ 0 から $w_{i,max}$ まで一様乱数で決める。 u_j は NA, j の 3 項目の重みであり、シミュレーション開始時に 0 から u_{max} まで一様乱数で決める。 \log は自然対数である。 P_f は時間によらず一定のファンダメンタル価格、 P^t は前節で定義された取引価格 (取引見込み価格)、 ϵ_j^t は時刻 t , エージェント j の乱数項であり、平均 0, 標準偏差 σ_ϵ の正規分布乱数である。 $r_{h,j}^t$ は時刻 t に NA, j が計測した過去リターンであり、 $r_{h,j}^t = \log(P^t/P^{t-\tau_j})$ である *6。ここで τ_j はシミュレーション開始時に 1 から τ_{max} までの一様乱数で NA ごとに決める。

式 (1) の第 1 項目はファンダメンタル価格と比較して安ければプラスの予想リターンを高めればマイナスの予想リターンを示す、ファンダメンタル価値を参照して投資判断を行うファンダメンタル投資家の成分である。第 2 項目は過去のリターンがプラス (マイナス) ならプラス (マイナス) の予想リターンを示す、過去の価格推移を参照して投資判断を行うテクニカル投資家の成分であり、第 3 項目はノイズを表している。

予想リターン $r_{e,j}^t$ より予想価格 $P_{e,j}^t$ は、

$$P_{e,j}^t = P^t \exp(r_{e,j}^t) \quad (2)$$

で求まる。注文価格 $P_{o,j}^t$ は平均 $P_{e,j}^t$, 標準偏差 P_σ の正規分布乱数で決める。ここで、 P_σ は定数である。そして、売り買いの別は予想価格 $P_{e,j}^t$ と注文価格 $P_{o,j}^t$ の大小関係で決める。すなわち、

$$\begin{aligned} P_{e,j}^t > P_{o,j}^t &\text{なら数量1の買い} \\ P_{e,j}^t < P_{o,j}^t &\text{なら数量1の売り} \end{aligned} \quad (3)$$

とする *7。

*6 ただし、 $t < \tau_j$ のときは、 $r_{h,j}^t = 0$ とした。

*7 ただし、 $t < t_c$ のときは十分な板の厚さを確保するため、 $P_f >$

2.2.2 マーケットメーカーエージェント (MM)

本研究では [草田 15a, 草田 15b] をベースにマーケットメーカーエージェント (MM) をモデル化した。図 2 に示すように MM は、フェアバリュー P_{fair} より注文スプレッド P_{spread} だけ高い値段 ($P_{fair} + P_{spread}$) に数量 1 の売り注文、 P_{spread} だけ安い値段 ($P_{fair} - P_{spread}$) に数量 1 の買い注文を入れる。MM は NA が注文を出す直前にこれらの注文を出し、板寄せが行われるまで毎時刻注文を入れる。こうすることにより、 δt の値によらず MM の総注文数量は同じとなる。そして板寄せ直後にすべての注文をキャンセルする。

MM には以下の 4 種類を実装した。1 つめはシンプル MM(SMM) であり、 $P_{fair} = P^t$ とする。2 つめはポジション MM(PMM) であり、

$$P_{fair} = (1 - kS^3)P^t \quad (4)$$

とする。ここで、 k は定数、 S は MM の保有数量 (ポジション) である。ポジションが正 ($S > 0$) の場合、 P_{fair} は P^t より小さくなり、買い注文も売り注文も安くなる。買い注文は取引成立しづらくなり、売り注文は取引成立しやすくなるため、ポジションの絶対値 $\|S\|$ が減少しやすくなる。ポジションが負 ($S < 0$) の場合、 P_{fair} は P^t より大きくなり、買い注文も売り注文も高くなる。買い注文は取引成立しやすくなり、売り注文は取引成立しづらくなり、ポジションの絶対値 $\|S\|$ は減少しやすくなる。つまり、PMM は SMM に比べ、ポジションをおさえ、ポジションが大きくなることによる価格変動のリスクをおさえることができる。

実際の市場においても、マーケットメーカー戦略をとる投資家は価格変動のリスクをおさえるために、不要なポジション

$P_{o,j}^t$ なら数量 1 の買い、 $P_f < P_{o,j}^t$ なら数量 1 の売りとする。また、 $P^t = P_f$ とする

をとらないように工夫している。特に、1日の取引時間が終了してから翌営業日の取引時間が開始するまでの時間(オーバーナイト)にポジションがあると、オーバーナイトに大きな出来事があった場合などに大きな価格変動リスクが発生するため、オーバーナイトはポジションをゼロにすることがほとんどである。そのため本研究では、[草田 15a, 草田 15b]では考慮されていなかった、オーバーナイトのポジションをゼロにするPMMを2種類実装した。

図3は1日の定義とポジションをゼロにする期間を示した。1日の長さは ΔT とし、この期間のうち最後の ΔT_{end} 期間に、PMMはポジションをゼロにするための工夫を行う。その1つめはPMM3であり、図4の左に示すように、ポジションが正($S > 0$)のときは買い注文を出さない、ポジションが負($S < 0$)のときは売り注文を出さないというものである。もうひとつはPMM4であり、図4の右に示すように、ポジションが正($S > 0$)のときは買い注文を出さないのみならず、売り注文の価格を $P_{fair} - P_{spread}$ (買い注文で使用する予定だった価格)に変更し、ポジションが負($S < 0$)のときは売り注文を出さないのみならず、買い注文の価格を $P_{fair} + P_{spread}$ (売り注文で使用する予定だった価格)に変更する。

PMM4はPMM3よりも、より取引が成立しやすい価格で注文を出すことにより、より積極的にポジションを0にすることを目指す。しかしながら、NAが供給した注文に自身の注文を対当させる、すなわち流動性を奪う注文を出すため、 $\|S\|$ が大きかった場合、一方向に価格を変動させてしまいながら $\|S\|$ を減らすことになり、価格変動リスクが高くなってしまう。そのため、もっとも現実的なモデルはPMM3であるが、次章で述べるように、 δt が大きい場合はPMM3では1日の終わりにポジションを0にできないため、PMM4を用意した。

3. シミュレーション結果

本研究では[水田 13]および[草田 15a, 草田 15b]と同じである以下のパラメータを用いた。具体的には、 $n = 1,000$, $w_{1,max} = 1$, $w_{2,max} = 10$, $u_{max} = 1$, $\tau_{max} = 10,000$, $\sigma_\epsilon = 0.06$, $P_\sigma = 30$, $t_c = 20,000$, $\delta P = 0.02$, $P_f = 10,000$, $k = 0.00000005$, $\Delta T = 20,000$, $\Delta T_{end} = 2,000$ とした。またシミュレーションは時刻 $t = t_e = 10,000,000$ まで行った*8。

また、4種類のMM、 $P_{spread}/P_f = 0.03\%, 0.1\%, 0.3\%, 1\%$ および $\delta t = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000$ に対して*9、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。

3.1 注文スプレッド P_{spread} と取引成立率

表1は、板寄せ間隔 δt 、MMの注文スプレッド P_{spread}/P_f ごとのMMの取引成立率を示した。取引成立率は、取引成立数量/総注文数量とした。MMはPMM4を用いた。

δt が大きくなると、MMの取引成立率が減少しているのが分かる。MMは注文を常に提示してNAに取引機会(流動性)を提供しているため、MMの取引成立率の減少は流動性の減少を示唆している。よって、 δt は小さい方がMMにより流動性が供給され、ザラバ方式が最もMMによる流動性供給が多い可能性を示している。

*8 これらのパラメータの妥当性検証については本稿の付録“モデルの妥当性検証”で説明した。さらなる詳細は[水田 14]に書かれている。

*9 つまり、4種類のMM、4通りの P_{spread} 、10通りの δt の計、160(=4×4×10)通りのケースを試した。

P_{spread}/P_f が大きくなると当然、MMの取引成立率は小さくなる。 $P_{spread}/P_f = 0.1\%$ の場合、 $\delta t = 1$ ならまだ現実的な取引成立率であり分析可能であるが、 $\delta t \geq 50$ で取引がほとんどなくなり、以下で調べるMMのポジションや損益の分析が意味をなさなくなる。そのため以後は、 $\delta t \geq 50$ でも意味のある分析できるように、 $P_{spread}/P_f = 0.03\%$ を用いる。

3.2 MMの種類ごとのポジション

表2は、板寄せ間隔 δt 、MMの種類ごとの、全期間または1日の終わりのみで計測したポジションの絶対値 $\|S\|$ の平均を示した。 $P_{spread}/P_f = 0.03\%$ とした。

実際の市場においては、マーケットメーカー戦略をとる投資家は価格変動のリスクをおさえるために、不要なポジションをとらないように工夫している。特に、1日の取引時間が終了してから翌営業日の取引時間が開始するまでの時間(オーバーナイト)にポジションがあると、オーバーナイトに大きな出来事があった場合などに大きな価格変動リスクが発生するため、オーバーナイトはポジションをゼロにすることがほとんどである。

そのため、少なくとも1日の終わりのみで計測した $\|S\|$ の平均が大きいと、リスクが高すぎて現実的でないといえる。1日の終わりに $\|S\|$ を小さくすることをモデル化していないSMMとPMMは当然、これらの値が0となっておらず現実的でない。

PMM3においては、 δt が小さい場合は、これらの値が0となっており、オーバーナイトのリスクをとらずにすんでいる。一方、 $\delta t \geq 20$ の場合は、 $\|S\| = 0$ に出来ておらず、オーバーナイトのリスクにさらされており、現実的でない。さらに、前節で述べたように δt が大きいほど取引成立が少ないにもかかわらず、全期間での $\|S\|$ が大きくなっており価格変動リスクが高まっている。すなわち、 δt が大きい場合、オーバーナイトの $\|S\| = 0$ を実現することが難しい上にポジションも大きく価格変動リスクが大きくなり、リスクをおさえてマーケットメーカー戦略を継続することが困難になる可能性を示している。

PMM4のように積極的な価格で $\|S\|$ を減らす注文を出して始めて、 $\delta t \geq 20$ の場合でも $\|S\|$ を0近くに減らすことができている。PMM4では、NAが供給した注文に自身の注文を対当させる、すなわち流動性を奪う注文を出すため、 $\|S\|$ が大きかった場合、一方向に価格を変動させてしまいながら $\|S\|$ を減らすことになり、価格変動リスクが高くなってしまう。そのような取引をしなければ $\delta t \geq 20$ の場合は、 $\|S\|$ を0近くに減らすことができない。このことから δt が大きい場合、オーバーナイトの $\|S\| = 0$ を実現するにはまた別のリスクをとる必要があり、リスクをおさえてマーケットメーカー戦略を継続することが困難になる可能性を示している。 δt が大きい場合は、リスクがあるとはいえ、PMM4が唯一、オーバーナイトの $\|S\|$ を0近くにできるので、以後、PMM4を用いる。

3.3 最終損益

表3は、板寄せ間隔 δt ごとの、MMの最終損益/ P_f 、ポジションの絶対値 $\|S\|$ の平均、MMとNAの取引成立率を示した。MMはPMM4用い、 $P_{spread}/P_f = 0.03\%$ とした。

$\delta t = 2, 5, 10$ の場合、MMの損益は負となっており損をしている。本研究では、市場が非常に安定しているとき、すなわちMMにとっては最も収益があげやすい環境のみを取り扱っている。それにもかかわらず収益をあげられていないということは、現実の市場ではさらに収益を上げるのは厳しいと考えられる。 $\delta t = 20, 50, 100, 200$ の場合でも収益はあるとはいえ、 $\delta t = 1$ の場合に比べ小さく、前節までに述べたように δt が大きいほど価格変動リスクが大きいことを考えると、 $\delta t = 1$

表 1: 板寄せ間隔 δt , MM の注文スプレッド P_{spread}/P_f ごとの MM の取引成立率 (PMM4)

MM の取引成立率	P_{spread}/P_f			
	0.03%	0.10%	0.30%	1.00%
1(ザラバ方式)	8.06%	1.53%	0.00%	0.00%
2	6.30%	0.88%	0.00%	0.00%
5	3.93%	0.37%	0.00%	0.00%
10	2.47%	0.14%	0.00%	0.00%
20	1.49%	0.02%	0.00%	0.00%
50	0.77%	0.00%	0.00%	0.00%
100	0.48%	0.00%	0.00%	0.00%
200	0.32%	0.00%	0.00%	0.00%
500	0.21%	0.00%	0.00%	0.00%
1000	0.22%	0.00%	0.00%	0.00%

表 2: 板寄せ間隔 δt , MM の種類ごとの, 全期間または 1 日の終わりのみで計測したポジションの絶対値 $\|S\|$ の平均 ($P_{spread}/P_f = 0.03\%$)

$\ S\ $ の平均	SMM		PMM		PMM3		PMM4	
	全期間	1 日の 終わりのみ	全期間	1 日の 終わりのみ	全期間	1 日の 終わりのみ	全期間	1 日の 終わりのみ
1(ザラバ方式)	12,357	12,371	3.18	3.08	2.90	0.00	2.89	0.00
2	17,42	17,441	3.10	3.25	2.79	0.00	2.79	0.00
5	4,409	4,414	3.87	3.95	3.48	0.00	3.48	0.00
10	1,744	1,744	4.44	4.34	4.01	0.02	3.96	0.00
20	548	548	4.84	4.71	4.52	0.78	4.35	0.00
50	384	385	5.27	5.14	5.02	2.63	4.63	0.00
100	369	370	5.57	5.51	5.56	4.26	4.80	0.00
200	174	174	5.91	6.11	5.92	5.69	4.38	0.00
500	72	71	5.75	6.06	5.70	5.81	2.32	0.03
1000	290	290	5.94	6.11	5.61	5.80	1.76	0.06

の場合に比べマーケットメーカー戦略を継続することが困難になっているといえる。少なくとも, $\delta t = 1$ の場合に比べ, リスクに見合った収益をあげづらくなる可能性が示されている。

4. まとめと今後の課題

本研究では [草田 15a, 草田 15b] の人工市場モデルをベースに, 板寄せ間隔 δt というパラメーターを新たに導入することにより, ザラバ方式 ($\delta t = 1$) とバッチオークション方式 ($\delta t > 1$) を連続的に変更できる価格決定メカニズムを実装し, マーケットメーカー戦略の損益のリスクを分析することによりその存続可能性を議論し, マーケットメーカー戦略が流動性を供給し続けることができるかどうかを調べた。

その結果, δt が大きくなると, MM の取引成立率が減少し MM による流動性供給が減少する可能性が示された。さらに, δt が大きくなると, オーバーナイトのポジションをゼロにすることが難しい上にポジションも大きく価格変動リスクが大きくなり, リスクをおさえてマーケットメーカー戦略を継続することが困難になる可能性が示された。さらに, $\delta t = 1$ のとき, すなわちザラバ方式のときのみ, リスクに見合った収益を得る可能性がある可能性が示され, $\delta t > 1$ のとき, すなわちバッチオークション方式のときは, 少なくともザラバ方式では機能

したマーケットメーカー戦略では, リスクに見合った収益をあげるのは難しくなる可能性が示された。

これらの示唆は, ザラバ方式では流動性を供給していたマーケットメーカー戦略が, バッチオークション方式になるとその供給を維持できなくなり, それらが撤退することにより, 流動性が低下する可能性を暗示しているとも考えることができる。これは, [大塚 14] の指摘と整合的である。

今後の課題は, バッチオークション方式に適したマーケットメーカー戦略があるか検討することである。本研究では, ザラバ方式で機能したマーケットメーカー戦略がバッチオークション方式では機能しない可能性を示した。しかし, バッチオークション方式でのみ機能するマーケットメーカー戦略が存在することは否定できず, 今後の課題である。また, 本研究では待機している注文が非常に多い (板が厚い) 状況のみを取り扱った。板が薄い銘柄では, バッチオークション方式の方が注文を出しやすく流動性が高まることも考えられ, 今後の課題である。さらに, そもそも流動性を供給するマーケットメーカー戦略があまり参加していない銘柄では, 本研究の議論は成り立たず今後の課題である。

さらに, 既に述べたように, 人工市場シミュレーションはその導入の純粋な効果を見ることができうるうえ, 過去に導入されたことがないものも分析することができる。ただその効果は

表 3: 板寄せ間隔 δt ごとの, MM の最終損益, ポジションの絶対値 $\|S\|$ の平均, MM と NA の取引成立率 (PMM4, $P_{spread}/P_f = 0.03\%$)

	MM の 最終損益 / P_f	$\ S\ $ の平均			NA の 取引成立率
		全期間	1 日の 終わり のみ	MM の 取引成立率	
1(ザラバ方式)	51.98	2.89	0.00	8.06%	39.1%
2	-29.42	2.79	0.00	6.30%	39.1%
5	-14.90	3.48	0.00	3.93%	37.6%
10	-4.08	3.96	0.00	2.47%	36.3%
δt	20	1.51	4.35	0.00	34.9%
	50	3.68	4.63	0.00	33.4%
	100	2.53	4.80	0.00	32.5%
	200	0.93	4.38	0.00	31.8%
	500	-0.06	2.32	0.03	0.21%
1000	-0.10	1.76	0.06	0.22%	30.5%

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、これまでに導入されることがない金融市場の規制・制度も議論することができるうえ、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規制や制度の変更や、新しいタイプの市場を分析してきた [LeBaron 06, Chen 12, 和泉 12, Cristelli 14, Mizuta 16b]。

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騰落率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”といったミクロメカニズムをモデル化し、そのミ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、ミ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す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を一つのモデルで再現しようとする、過剰に複雑なモデルをもち、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せず、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。

モデルの妥当性検証

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

表 4: ザラバ方式 ($\delta t = 1$), MM なしの場合の各種統計量

取引	約定率	32.3%
	キャンセル率	26.1%
	注文件数 / 1 日	6467
標準 偏差	1 期間	0.0512%
	1 日 ($\Delta T = 20000$)	0.562%
	尖度	1.42
	ラグ	
二乗リターンの 自己相関係数	1	0.225
	2	0.138
	3	0.106
	4	0.087
	5	0.075

[LeBaron 06, Chen 12, 和泉 12, Cristelli 14, Mizuta 16b]. ファット・テールは、市場価格の騰落率の分布が正規分布ではなく裾が厚い、すなわち、尖度が正であることである。ボラティリティ・クラスターリングは市場価格の騰落率の 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・クラスターリングの 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騰落率の分布の尖度は 1 ~ 100 程度であり、ボラティリティ・クラスターリングについては、実証分析でよく観測される騰落率の自己相関は 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 4 は、ザラバ方式 ($\delta t = 1$), MM なしの場合の統計量を示した。統計量は 100 回の試行の平均値を用いている。約定率、キャンセル率ともにいずれのティックサイズにおいても実際の市場の値に近い値を得ており、モデルの妥当性を示唆している *10。1 日 (ΔT) ごとの騰落率の標準偏差 *11 も、概ね実際の市場に近い値が得られ、この側面からもモデルの妥当性を示唆している。

表 4 は、10 期間ごと *12 の騰落率を用いて計算した尖度と騰落率の 2 乗の

*10 約定率 = 約定件数/注文件数, キャンセル率=キャンセル件数/(注文件数+キャンセル件数)。

*11 本研究のシミュレーションではオーバーナイトの価格変動が存在しないので、ここでの 1 日ごとの騰落率の標準偏差は、実際の市場のイントラデイのボラティリティに相当する。

*12 本モデルの時刻は注文をただけで取引が成立しない場合も時刻が進むため、時刻 1 ごと全ての価格を用いたスタイライズド・ファクトは多くの価格変動が無いデータによりバイアスがかかってしまうため、10 ステップ間隔

自己相関も示している。騰落率の尖度がプラスで、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、騰落率の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・クラスターリングが再現されていると考えられる。

以上により、本研究のモデルが、約定件数やキャンセル率、1日の騰落率の標準偏差など短期のマイクロ・ストラクチャーの性質および、長期に観測される価格変動の統計的な性質も再現していることが示された。それゆえ本研究の目的である、バッチオークション方式においてマーケットメーカー戦略の存続可能性を議論するのに妥当なモデルであることが示された。

留意事項と謝辞

本論文はスパークス・アセット・マネジメント株式会社の公式見解を表すものではありません。すべては個人的見解であります。本研究の一部は、JST, CREST および JSPS 科研費 15H02745 の助成を受けたものです。

参考文献

- [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>
- [Bellia 15] Bellia, M., Pelizzon, L., Subrahmanyam, M. G., Uno, J., and Yuferova, D.: Low-Latency Trading and Price Discovery Without Trading: Evidence from The Tokyo Stock Exchange Pre-Opening Period, *SSRN Working Paper Series* (2015), <http://ssrn.com/abstract=2705962>
- [Budish 15] Budish, E., Cramton, P., and Shim, J.: The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response, *The Quarterly Journal of Economics*, Vol. 130, No. 4, pp. 1547–1621 (2015), <http://qje.oxfordjournals.org/content/130/4/1547.abstract>
- [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>
- [Chiarella 02] Chiarella, C. and Iori, G.: A simulation analysis of the microstructure of double auction markets, *Quantitative Finance*, Vol. 2, No. 5, pp. 346–353 (2002), <http://www.tandfonline.com/doi/abs/10.1088/1469-7688/2/5/303>
- [Cristelli 14] Cristelli, M.: *Complexity in Financial Markets, Modeling Psychological Behavior in Agent-Based Models and Order Book Models*, Springer (2014), <https://dx.doi.org/10.1007/978-3-319-00723-6>
- [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 12] Farmer, D. and Skouras, S.: Review of the Benefits of a Continuous Market Vs. Randomised Stop Auctions and of Alternative Priority Rules (policy Options 7 and 12), *Manuscript, Foresight, Government Office for Science, UK* (2012), <https://www.gov.uk/government/publications/computer-trading-continuous-market-vs-randomised-stop-auctions-and-alternative-priority-rules>
- [Fricke 15] Fricke, D. and Gerig, A.: Too fast or too slow? Determining the optimal speed of financial markets, *SSRN Working Paper Series* (2015), <http://ssrn.com/abstract=2363114>
- [和泉 12] 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://www.iwanami.co.jp/moreinfo/0113060/>
- [実世 12] 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
- [草田 15a] 草田 裕紀, 水田 孝信, 早川 聡, 和泉 潔: 保有資産を考慮したマーケットメイク戦略が市場間競争に与える影響: 人工市場アプローチによる分析, JPX ワーキング・ペーパー, No. 8, 日本取引所グループ (2015), <http://www.jpx.co.jp/corporate/research-study/working-paper/>
- [草田 15b] 草田 裕紀, 水田 孝信, 早川 聡, 和泉 潔: 保有資産を考慮したマーケットメイク戦略が取引所間競争に与える影響: 人工市場アプローチによる分析, 人工知能学会論文誌, Vol. 30, No. 5, pp. 675–682 (2015), <http://doi.org/10.1527/tjsai.30.675>
- [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)
- [Manahov 16] Manahov, V.: A note on the relationship between high-frequency trading and latency arbitrage, *International Review of Financial Analysis* (2016), <http://dx.doi.org/10.1016/j.irfa.2016.06.014>
- [水田 12] 水田 孝信: 金融市場における最新情報技術: 1. 金融の役割と情報化の進展-市場の高速化と課題-, 情報処理, Vol. 53, No. 9, pp. 892–897 (2012), <http://id.nii.ac.jp/1001/00083434/>
- [水田 13] 水田 孝信, 早川 聡, 和泉 潔, 吉村 忍: 人工市場シミュレーションを用いた取引市場間におけるティックサイズと取引量の関係性分析, JPX ワーキング・ペーパー, No. 2, 日本取引所グループ (2013), <http://www.jpx.co.jp/corporate/research-study/working-paper/>
- [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, PhD thesis, 東京大学大学院工学系研究科 (2014), <http://hdl.handle.net/2261/59875>
- [水田 15] 水田 孝信, 則武 誉人, 早川 聡, 和泉 潔: 人工市場シミュレーションを用いた取引システムの高速度化が価格形成に与える影響の分析, JPX ワーキング・ペーパー, No. 9, 日本取引所グループ (2015), <http://www.jpx.co.jp/corporate/research-study/working-paper/>
- [水田 16a] 水田 孝信, 和泉 潔: 人工市場シミュレーションを用いたバッチオークションの分析, JPX ワーキング・ペーパー, No. 17, 日本取引所グループ (2016), <http://www.jpx.co.jp/corporate/research-study/working-paper/>
- [Mizuta 16b] 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>
- [Mizuta 16c] Mizuta, T., Noritake, Y., Hayakawa, S., and Izumi, K.: Affecting Market Efficiency by Increasing Speed of Order Matching Systems on Financial Exchanges – Investigation using Agent Based Model, in *Computational Intelligence for Financial Engineering Economics (CIFER), 2016 IEEE Conference on, 2016 IEEE Symposium Series on Computational Intelligence on* (2016), <http://ssci2016.cs.surrey.ac.uk/>
- [Nakajima 04] Nakajima, Y. and Shiozawa, Y.: Usefulness and feasibility of market maker in a thin market, in *ICBEES (International Conference Experiments in Economic Sciences)*, Vol. 47 (2004), <http://www.cc.kyoto-su.ac.jp/project/orc/execo/EES2004/PROG/47.pdf>
- [大埴 14] 大埴 剛士: 米国市場の複雑性と HFT を巡る議論, JPX ワーキング・ペーパー, 特別レポート, 日本取引所グループ (2014), <http://www.jpx.co.jp/corporate/research-study/working-paper/>
- [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/>
- [清水 13] 清水 葉子: HFT, PTS, ダークプールの諸外国における動向～欧米での証券市場間の競争や技術革新に関する考察～, 金融庁金融研究センター ディスカッションペーパー (2013), <http://www.fsa.go.jp/frtc/seika/discussion/2013/01.pdf>

人工市場シミュレーションを用いたレバレッジドETFが原資産 価格変動に与える影響分析

Analysis of the Impact of Leveraged ETF Rebalancing Trades on the Underlying Asset Market Using Artificial Market Simulation

八木 勲^{1*} 水田 孝信²
Isao Yagi¹ Takanobu Mizuta²

¹ 神奈川工科大学情報学部

¹ Faculty of Information Technology, Kanagawa Institute of Technology

² スパークス・アセット・マネジメント株式会社

² SPARX Asset Management Co., Ltd.

Abstract: Financial markets occasionally become highly volatile, as a result of a financial crisis or other factors. Previously, index futures trading and program trading have been singled out as direct causes of market destabilization, but more recently it has been suggested that leveraged ETFs (funds aimed at amplifying several-fold the movement of a price index such as the Nikkei Stock Average or underlying assets) rebalancing trades may also be a factor. This study uses a financial market simulation (artificial market) constructed virtually on a computer to assess the impact of leveraged ETF rebalancing trades on the underlying assets market. Analysis results showed that a larger amount of the managed assets of leveraged ETFs corresponds to a higher volatility of the underlying securities market. They also demonstrated that leveraged ETF trading can destroy the underlying assets market, if the leveraged ETF trading impact on the market is greater than that of ordinary volatility of the underlying assets.

1 はじめに

近年リーマンショックやギリシャ危機、トランプショックなどにより金融市場のボラティリティが大きくなるのが度々起こっている。このとき市場を不安定にする直接的な原因として指数先物（日経平均 225 先物、S&P500 先物など）取引やプログラムによる自動売買などが挙げられてきたが、最近ではレバレッジドETF（Exchange Traded Funds）のリバランス取引も一因ではないかと言われている。

レバレッジドETFは原資産や原指数（例えば、日経平均 225 など）の収益率の何倍かの収益率を目指すETFである。これらのETFは、金融市場指数に投資するため個別株式銘柄よりもローリスクでありながら、レバレッジの恩恵により比較的高いリターンが望めるので、運用残高が飛躍的に伸びてきている。レバレッジドETFは、レバレッジを維持する（保有する原資産の純資産総額を、レバレッジドETFの純資産総額の決

められた倍数に維持する）よう、原資産が上昇すれば原資産を買い、原資産が下落すれば原資産を売るというリバランス取引を日々行わなければならない。そのため、これらの売買が原因で原資産の価格を不安定にするのではないかとされている [2]。

一例を挙げると、日経平均 225 を原指標とするレバレッジドETFでは、日経平均先物を運用することで日経平均の収益率の何倍かの収益率を得ることを目指している。このレバレッジドETFのリバランス取引は、株式市場の取引が終了した後から日経平均先物市場の取引が終了するまでの間に行われる可能性が高く、その期間ボラティリティが高くなり市場が不安定になっているのではないかと考えられている。また、日経平均 225 は日経平均先物と価格の連動性が高いため、日経平均先物が大量に売買されボラティリティが上昇すると、日経平均 225 のボラティリティも同じように上昇し市場が不安定になるのではないかと危惧されている。

レバレッジドETFが原指標の与える影響についてはこれまでにさまざまな実証研究がなされてきた。

Cheng ら [1] は、ETF の原指標が 1 日に 1% 動くと、レバレッジドETFのリバランス取引の大きさは、そ

*連絡先：神奈川工科大学情報学部情報工学科
〒 243-0292 神奈川県厚木市下荻野 1030
E-mail: iyagi2005@gmail.com

の日の市場取引終了時点の原指標の出来高の16.8%を占める可能性があることを示唆している。また、従来より景気後退時や株価が大暴落した後は原指標のボラティリティは大きくなる傾向があり、レバレッジ取引が多いほどその傾向が強くなることが報告されている[8]。レバレッジドETFは近年盛んに取引されてきた商品であるので、原指標のボラティリティの上昇の原因がレバレッジドETFにあるとは一概に言い難い。また、Deshpandeら[3]は、レバレッジドETFによって取引される売買量は、S&P500のそれに対してわずか0.0079%であるため、市場に与えるインパクトは限定的だと結論づけている。さらに、Trainor Jr.[9]は、レバレッジドETFのリバランス取引がS&P500のボラティリティに与える影響を、前日の取引終了前のボラティリティと翌日の取引開始後のボラティリティを比較することで検証したが、明確な結果は得ることができなかったと結論づけている。

このように実証分析を主とした先行研究では、レバレッジドETFが市場に与える影響は千差万別である。また、金融市場の形成にはさまざまな事象と要因が複雑に関わってくるため、実証分析では一部の要因だけを抽出して分析することが困難である。

金融市場の分析法の1つとして人工市場を用いる方法がある。人工市場とは計算機上に仮想的に構築した金融市場マルチエージェントシステムである[4, 5, 7]。エージェントにそれぞれ独自の売買手法を与え、それらに投資家として金融資産の取引をさせ、市場がどのような振る舞いをするかを確認することができる。その一方で、市場側に何らかの制約（市場安定化や効率化のための制度等）をモデル化し組み込むことで、投資家がどのように振る舞うか、さらに、彼らの振る舞いによって市場にどのような影響が現れるかを検討することも可能である。しかし、これまでに人工市場を用いたレバレッジドETFに関する研究はあまり行われていない[10]。

そこで、本研究では、八木ら[11]が構築した人工市場モデルにレバレッジドETF取引モデルを実装し、レバレッジドETFのリバランス取引が原資産価格変動に与える影響について検証を行った。その結果、レバレッジドETFの運用資産が大きくなるほど、原資産のリバランス取引量が増大し、ボラティリティが大きくなることがわかった。そして、レバレッジドETFの売買が、少なくともボラティリティより大きなマーケットインパクトを市場に与えると、原資産市場を破壊してしまう可能性があることも判明した。

2 人工市場モデル

本研究では、八木ら[11]の人工市場モデルを基にモデルを構築した。本モデルでは原資産を想定したリスク資産1つのみを売買対象とする。以下、本モデルにおける取引価格決定法と、本モデルに参加する投資家エージェントについて述べる。

2.1 取引価格決定法

2.2節にて述べる一般投資家エージェントは、それぞれの手法で発注価格と発注資産数を決め注文を出す。市場では、第 t 期の全てのエージェントの売り注文と買い注文をつき合わせて売買を成立させる。買い手側は高い発注価格のエージェントから、売り手側は安い発注価格のエージェントから優先的に取引に参加する。買い手側の発注価格が売り手側の発注価格を上回る、もしくは一致するとき売買が成立する。この決定法のことを一般に板寄せ方式という。

一方、レバレッジドETFを構成するために原資産を売買するレバレッジドエージェントは、上述したように、今期の収益率が原資産の収益率の何倍かになるように市場取引終了間際にリバランス取引を行う。このレバレッジドエージェントの取引法を模倣するため、上記板寄せ方式を以下のように拡張する。まず、一般投資家エージェントの注文を仮につき合わせて、今期の原資産の仮価格を求める。レバレッジドエージェントはこの仮価格に基づいて自分自身の資産保有ポジションを計算し、そのポジションと現在保有している資産数の差分を注文する。最後に、先ほどの一般投資家エージェントの注文にレバレッジドエージェントの注文をつき合わせて、今期の原資産価格を求める。

2.2 エージェント

取引に参加するエージェントタイプは、ファンダメンタリスト、チャートリスト、ノイズトレーダー、レバレッジドエージェントの4種類である。なお、レバレッジドエージェントを除くエージェントのことを総じて一般投資家エージェントと呼ぶ。

2.2.1 ファンダメンタリスト

本市場におけるファンダメンタリストは、リスク資産の理論価格に基づいてその資産の当期価格を予想し、その予想価格において当期資産価値が最大になるようリスク資産保有数を調整する。第 t 期の理論価格 P_t は外部より与えられ、エージェント i の第 t 期の予想価格 $\tilde{P}_{i,t}$ は、平均が $(1 + \epsilon_{i,t})P_t$ で、分散が $(\alpha(1 + \epsilon_{i,t})P_t)^2$

の正規分布に従うものとする。ただし、 P_t は理論価格、 $\epsilon_{i,t}$ はエージェント i の第 t 期における強気度¹、 α はエージェント i の予想価格のばらつきを表す係数である。なお強気度は平均 0、分散 σ_ϵ^2 の正規分布に従うものとする。その他、第 t 期の取引価格を P_t 、第 t 期の取引前のキャッシュを $C_{i,t-1}$ 、第 t 期の取引前の資産保有数を $q_{i,t-1}$ とすると、第 t 期の資産価格決定前のエージェント i の総資産 $W_{i,t-1}$ は次のように表される。

$$W_{i,t-1} = C_{i,t-1} + P_{t-1} \cdot q_{i,t-1} \quad (1)$$

その結果、条件式 (1) の下で、価格決定後の総資産量から計算される効用の主観的期待値を最大にするリスク資産保有数（最適リスク資産保有数） $\tilde{q}_{i,t}$ は、

$$\tilde{q}_{i,t} = \frac{(1 + \epsilon_{i,t})P_t - P_{t-1}}{a(\alpha(1 + \epsilon_{i,t})P_t)^2}$$

と表すことができる。ただし、 $a(> 0)$ はリスク回避係数で、この値が大きいほど、リスクを回避するためファンダメンタリストは保有資産数を小さくする。

ファンダメンタリスト i は $\tilde{q}_{i,t}$ を基に売買方針を決定する。第 t 期において、 $\tilde{q}_{i,t}$ が $q_{i,t-1}$ より大きいとき、エージェント i は、価格 $\tilde{P}_{i,t}$ で資産数 $\tilde{q}_{i,t} - q_{i,t-1}$ の買い注文を出す。一方、 $\tilde{q}_{i,t}$ が $q_{i,t-1}$ より小さいとき、エージェント i は価格 $\tilde{P}_{i,t}$ で資産数 $q_{i,t-1} - \tilde{q}_{i,t}$ の売り注文を出す。第 t 期において、 $\tilde{q}_{i,t}$ と $q_{i,t-1}$ が同じとき、エージェント i は売買せずに待機する。

2.2.2 チャーティスト

本市場のチャーティストは移動平均に基づいた売買を行う。チャーティストには順張りとは逆張りが存在する。エージェント i が利用する、第 t 期における $n_{i,t}$ 期間移動平均を $MA_{t,n_{i,t}} = \sum_{j=1}^{n_{i,t}} P_{t-j}/n_{i,t}$ とし、 $\Delta MA_{t,n_{i,t}} = MA_{t,n_{i,t}} - MA_{t-1,n_{i,t}}$ とする。

エージェント i が順張り（逆張り）のとき、次の方針で売買を行う。 $\Delta MA_{t,n_{i,t}}$ が 0 より大きいとき、価格 $(1 + \alpha_T)P_{t-1}$ で資産数 $q_{i,t}^T$ の買い（売り）注文を出す。 $\Delta MA_{t,n_{i,t}}$ が 0 より小さいとき、価格 $(1 + \alpha_T)P_{t-1}$ で資産数 $q_{i,t}^T$ の売り（買い）注文を出す。 $\Delta MA_{t,n_{i,t}}$ が 0 のときは売買せずに待機する。

なお、 $n_{i,t}$ の初期値は、 $[1, 25]$ の一様分布に従う乱数とし、 α_T は平均 μ_P^T 、分散 $(\sigma_P^T)^2$ の正規分布に従う乱数とする。 $q_{i,t}^T$ は平均 μ_V^T 、分散 $(\sigma_V^T)^2$ の正規分布に従う乱数とする。

¹強気派ほど予想価格を理論価格より高く設定するので、正方向に大きな値を設定し、弱気派ほど予想価格を理論価格より低く設定するので、負方向に大きな値を設定する。

2.2.3 ノイズエージェント

ノイズエージェント i はそれぞれ 1/3 の確率で、買い、売り、待機を選択する。エージェント i は、売り買いともに、価格 $(1 + \alpha_N)P_{t-1}$ で資産数 $q_{i,t}^N$ の注文を出す。ただし、 α_N は平均 μ_P^N 、分散 $(\sigma_P^N)^2$ の正規分布に従う乱数とし、 $q_{i,t}^N$ は平均 μ_V^N 、分散 $(\sigma_V^N)^2$ の正規分布に従う乱数とする。

2.2.4 レバレッジドエージェント

レバレッジドエージェントは、原資産の収益率の L 倍を目指すレバレッジド資産を運用するエージェントである。原資産価格が上昇もしくは下降すると、各エージェントのレバレッジド資産のレバレッジ²も変化する。そこでレバレッジドエージェントは、保有純資産総額の収益率が原資産の収益率の L ($-L$) 倍になるよう原資産のリバランス取引を行う³。

例えば、キャッシュを 1,000 のみ保有するレバレッジドエージェントが、価格 100 の原資産をレバレッジ 2 倍で運用するとする（表 1）。はじめにレバレッジドエージェントは原資産を 20 購入する（キャッシュの不足分 1000 は借金するものとする）。このときのレバレッジドエージェントの保有純資産総額は 1,000 ($= 100 \times 20 - 1,000$) で、エクスポージャーは 2,000 となり、レバレッジは 2 ($= (100 \times 20) / (100 \times 20 - 1,000)$) である。第 1 期の原資産価格が 110 になると、レバレッジドエージェントの保有純資産総額は 1,200 ($= 1,000 \times 1.2$)、エクスポージャーは 2,200 ($= 2,000 \times 1.1$) となる。このとき、必要とされるエクスポージャーは 2,400（レバレッジは 1.83 ($= 2,200 / 1,200$)) であるので、原資産を 1.81 ($= (2,400 - 2,200) / 110$) だけ追加購入して、レバレッジを 2 倍に維持しなければならない。次に第 2 期の原資産価格が 99 になると、レバレッジドエージェントの保有純資産総額は 960 ($= 1,200 \times 0.8$)、エクスポージャーは 2,160 ($= 2,400 \times 0.9$) となり、レバレッジは 2.25 ($= 2,160 / 960$) となる。必要とされるエクスポージャーは 1,920 であるので、原資産を 2.42 ($= |1,960 - 2,160| / 99$) だけ売却して、レバレッジを 2 倍に維持しなければならない。

以上のような売買を行うレバレッジドエージェントを次のようにモデル化する。

第 t 期の一般投資家エージェントのみの取引によって決まる原資産の仮価格を P_t' とし、第 t 期のエージェント i の資産保有数を $S_{i,t}$ とする。そして、目標となるレバ

²レバレッジ = 保有原資産総額 / 保有純資産総額。ただし、保有原資産総額 = 原資産価格 × 原資産保有数、保有純資産総額 = 原資産価格 × 原資産保有数 + キャッシュ。

³原資産価格が変動するごとに、レバレッジが L 倍になるよう修正していかないと、保有原資産総額の収益率が保有純資産総額の収益率の L 倍にならなくなるため。

表 1: 原資産の収益率に対する 2 倍のレバレッジド ETF のリバランス取引例

期	原資産		2x 原資産		2x レバレッジド ETF			
	(1) 価格	(2) 収益率	(3) 収益率	(4) エクスポージャー (前期の (6)×(2))	(5) 純資産総額 (レバレッジ) (前期の (5)×(3) (= (4)/(5)))	(6) 必要な エクスポージャー (= (5)×2)	(7) 必要な 売買 (= (6)-(4))	(8) 必要な 原資産取引数 (= (7)/(1))
0	100			2,000	1,000 (2)	2,000		
1	110	10%	20%	2,200	1,200 (1.83)	2,400	200	1.81
2	99	-10%	-20%	2,160	960 (2.25)	1,920	-240	-2.42

レレッジ L が達成されている状態で取引を開始する。すなわち、初期資産保有数 $S_{i,0}$ は、 $S_{i,0} = L \cdot C_{i,0} / P_0$ とし、それに伴い、初期キャッシュ $C_{i,0}$ は、 $C_{i,0} = C_{i,0} - P_0 \cdot S_{i,0}$ となる（レバレッジ分、借金をすることになる）。第 t 期のエージェント i の保有純資産総額 $NAV_{i,t}$ 、レバレッジ $L_{i,t}$ はそれぞれ次のように表される。

$$NAV_{i,t} = P'_t \cdot S_{i,t} + C_{i,t}$$

$$L_{i,t} = \frac{P'_t \cdot S_{i,t}}{NAV_{i,t}}$$

第 t 期のエージェント i の発注数 $V_{i,t+1}$ は次のように表される⁴。ただし、上記のように発注数に端数が出る場合は切り捨てる。

$$V_{i,t+1} = \left\lfloor \frac{(L-1)P'_t \cdot S_{i,t} + L \cdot C_{i,t}}{P'_t} \right\rfloor$$

本エージェントの注文は可能な限り約定させないといけないので、買い注文の際の発注価格はどの一般投資家エージェントよりも高く、売り注文のときの発注価格はどの一般投資家エージェントよりも安く設定する。

$NAV_{i,t}$ が 0 以下となったとき、本エージェントは破産したものとし、一旦初期状態に戻した後、破産した次の期から再度市場に参加させる。

2.3 周囲の好成績投資家の投資法に影響される投資家のモデル化

エージェントの評価学習機能についても既述の人工市場モデル [11] に従う。すなわち、取引終了後、各エージェントは運用成績の評価を行い、他のエージェントと比較して相対的に成績が悪いエージェントは、成績のよいエージェントの売買ルールの模倣を行う。ファンダメンタリストは強気度を、チャーティストは移動平均の期間を模倣する。さらに、エージェントが試行錯誤的に新しい売買ルールを求める姿を客観的に表す

⁴ $V_{i,t+1}$ は次の式から求められる。

$$L = \frac{P'_t(S_{i,t} + V_{i,t+1})}{P'_t(S_{i,t} + V_{t,i}) + (C_{i,t} - P'_t \cdot V_{i,t+1})}$$

ため、一部のエージェントの売買ルール変更 randomness をもたせた。

本市場ではエージェントの売買タイプの割合を固定しているため、エージェントが他の売買タイプに移ることはない。すなわち、成績の悪いファンダメンタリストは成績のよいファンダメンタリストの模倣を試みるが、運用成績のよいチャーティストやノイズエージェントを模倣することはない。なおノイズエージェントは、運用成績の評価と模倣は行わない。以下、各タイプにおける運用成績評価と模倣の手順について述べる。ここで、あるエージェント i の第 $t-1$ 期から第 t 期の資産変化率を $R_{i,t}$ ($= W_{i,t} / W_{i,t-1}$)、過去 N 期の資産変化率の平均 $\bar{R}_{i,t}$ ($= \sum_{j=1}^N R_{i,t-(j-1)} / N$) とする。

2.3.1 ファンダメンタリスト

エージェント i の $\bar{R}_{i,t}$ が、全てのファンダメンタリスト中の下位 $N_L\%$ に属するとき、確率 p_i ($=$ エージェント i の順位 / 全ファンダメンタリスト数) でエージェント i の強気度を変更する。すなわち、資産変化率の平均がすべてのファンダメンタリスト中の上位 $N_H\%$ に属するエージェントを 1 つランダムに抽出し、そのエージェントの強気度を i の新しい強気度 $\epsilon_{i,t+1}$ とする。

さらに、より多くの利益を期待するエージェントの売買ルール変更を実現するために、資産変化率の平均が、全てのファンダメンタリストのその上位 $N_H\%$ 以外のエージェントに対し、それぞれ P_c の確率で強気度をランダムに変更する。

2.3.2 チャーティスト

エージェント i の $\bar{R}_{i,t}$ が、全てのチャーティスト中の下位 $N_L\%$ に属するとき、確率 p_i ($=$ エージェント i の順位 / 全チャーティスト数) でエージェント i の移動平均の期間を変更する。すなわち、資産変化率の平均が全てのチャーティスト中の上位 $N_H\%$ に属するチャーティストを 1 つランダムに抽出し、そのエージェントの移動平均期間を i の新しい移動平均期間 $n_{i,t+1}$ とする。

さらに、より多くの利益を期待するエージェントの売買ルール変更を実現するために、資産変化率の平均が、全てのチャートリストの上位 $N_H\%$ 以外のエージェントに対し、それぞれ P_n の確率で売買方針（順張りりと逆張りり）および移動平均期間 $n_{i,t+1}$ を変更する。ただし、 $1 \leq n_{i,t+1} \leq 25$ とする。

3 実験

本研究では、レバレッジドエージェントの初期キャッシュ量を、一般投資家エージェントの1倍、10倍、100倍、500倍、1,000倍、1,500倍、1,600倍、1,700倍、1,800倍、1,900倍、2,000倍、5,000倍、10,000倍として実験を試みた。実験環境は以下のとおりである。資産価格と資産の仮価格の初期値はともに10,000とする。資産の理論価格は10,000とし試行期間中固定値とする。一般投資家エージェントの個体数は10,000とする。内訳は、ファンダメンタリストが4,500、チャートリストが4,500、ノイズトレーダーが1,000とする。これらの初期資産は1,000,000とする。レバレッジドエージェントの個体数は1とする。各エージェントタイプに関するパラメータは次のとおりである。 $\alpha = 0.02$, $a = 0.001$, $\sigma_\epsilon = 0.1$, $\mu_P^T = 0$, $\sigma_P^T = \sqrt{0.02}$, $\mu_V^T = 2$, $\sigma_V^T = 0.1$, $\mu_P^N = 0$, $\sigma_P^N = \sqrt{0.02}$, $\mu_V^N = 2$, $\sigma_V^N = \sqrt{0.1}$, $N = 5$, $N_H = N_L = 20\%$, $P_\epsilon = P_n = 0.05$ 。

実験期間は10,000期だが、市場が安定した5,001期から10,000期までの5,000期を市場のボラティリティのサンプリング期間とした。なお、一般投資家エージェントの取引が安定してからレバレッジドエージェントを参入させるため、レバレッジドエージェントの取引は3,001期から開始している。

3.1 モデルの妥当性

実験を始めるにあたって、本研究にて提案した人工市場モデルの妥当性を検証する。検証法として、既の実証分析 [6] で得られている代表的な統計的性質 (Stylized Facts) のうち、下記の2つを満たしているかどうかで判断した。

- 価格収益率の尖度が正である。
- 価格収益率の自己相関の2乗が正の相関関係である。

表2は、レバレッジドエージェントが未参加のとき（左から2列目）と、初期キャッシュ量が一般投資家エージェントの1倍から10,000倍（左から3列目以降）までのレバレッジドエージェントが参加したときの、5,000期から10,000期までの価格収益率に基づいた Stylized

Facts の平均（試行回数は各20回）である。表2から、レバレッジドエージェントが未参加のときと、初期キャッシュ量が一般投資家エージェントのそれに対してそれほど大きくない間は上記の2つの条件を満たしている。これらの結果から本提案モデルは Stylized Facts を満たしており、実験環境として妥当であると判断できる。

4 実験結果と考察

表3からわかるように、初期キャッシュ量が大きくなるほど、価格収益率のボラティリティも大きくなっている。このような結果になる理由は以下のとおりである。

レバレッジドエージェントは、仮価格 P_t^i が直前の取引価格 P_{t-1} より高ければ原資産保有数を増やすために買い注文を出し、低ければ原資産保有数を減らすために売り注文を出す。このときレバレッジドエージェントは目標とする保有数を確保するために、原資産を非常に高い価格で買ったり、非常に低い価格で売ったりすることがある。初期キャッシュ量が大きくなればなるほど、リバランス取引が増すため、思わぬ価格での売買を行う可能性が高くなり、マーケットへのインパクトも大きくなる。よって、初期キャッシュ量が大きくなると、一取引あたりの発注量が増大し、マーケットへのインパクトも大きくなるため、ボラティリティも大きくなると考えられる。

表3からも、レバレッジドエージェントの初期キャッシュ量が大きくなるほど、1取引あたりの平均発注数も、マーケットインパクト MI （ここでは第 t 期のマーケットインパクト MI_t を $MI_t = |P_t - P_t^i|/P_t$ とし、1試行あたりの平均を MI としている）も、ボラティリティもすべて大きくなっていることが見てとれる。

表2からは、初期キャッシュ量が大きくなりすぎると、市場収益率の2乗の自己相関が負になることが見てとれる（本現象は、表2からは初期キャッシュ量が一般投資家の1,700倍以上から発生しているように見えるが、正確には1,600倍から発生している。これは、本結果は試行回数20回の平均で示しているからであり、ごく数回しか本現象が発生していない1,600倍の場合は表2に顕在化していないからである）。これは現実市場では見受けられない現象であり、市場が破壊されたことを示唆している。注目すべき点は、表3において、レバレッジドエージェントの初期キャッシュ量が1,600倍以上の市場の MI が、レバレッジドエージェントが未参加の市場のボラティリティより大きくなっている点である。これは、初期キャッシュ量が大きくなり、レバレッジドエージェントの取引がボラティリティより大きなマーケットインパクトを与えるようになると、市場が破壊されてしまうことを示唆しているのではない

表 2: 各試行におけるボラティリティと Stylized Facts

	一般投資家エージェントの初期キャッシュ量に対する レバレッジドエージェントの初期キャッシュ量の倍率													
	未参加	×1	×10	×100	×500	×1,000	×1,500	×1,600	×1,700	×1,800	×1,900	×2000	×5000	×10000
ボラティリティ ($\times 10^{-2}$)	1.16	1.17	1.19	1.38	1.74	1.46	1.38	14.4	25.9	35.6	20.7	23.9	39.5	39.0
尖度	4.32	3.99	3.82	2.83	0.69	6.46	22.4	14.88	4.05	-1.31	3.88	3.06	-1.87	-1.87
ラグ														
1	0.497	0.488	0.491	0.392	0.175	0.391	0.605	0.386	-0.338	-0.769	0.089	-0.110	-0.997	-0.997
2	0.701	0.713	0.719	0.719	0.789	0.609	0.497	0.534	0.751	0.982	0.450	0.340	0.999	0.999
3	0.503	0.499	0.503	0.423	0.233	0.330	0.365	0.231	-0.421	-0.775	0.270	0.186	-0.997	-0.997
4	0.642	0.654	0.660	0.653	0.706	0.482	0.267	0.413	0.670	0.976	0.287	0.076	0.999	0.999
5	0.513	0.513	0.515	0.447	0.275	0.290	0.184	0.114	-0.479	-0.779	0.238	0.313	-0.997	-0.997

表 3: レバレッジドエージェントの初期キャッシュ量とマーケットインパクトの関連性

	一般投資家エージェントの初期キャッシュ量に対する レバレッジドエージェントの初期キャッシュ量の倍率													
	未参加	×1	×10	×100	×500	×1,000	×1,500	×1,600	×1,700	×1,800	×1,900	×2,000	×5,000	×10,000
(1) レバレッジド ETF エージェント の発注数	-	0.775	7.64	70.7	186	84.9	91.0	1166	2561	2745	1149	2001	2858	2860
(2) MI($\times 10^2$)	-	0.000489	0.00499	0.0392	0.0842	0.0585	0.0798	4.90	10.5	14.2	7.11	8.60	16.5	16.5
(3) ボラティリティ ($\times 10^2$)	1.16*1	1.17	1.19	1.38	1.74	1.46	1.38	14.4	25.9	35.6	20.7	23.9	39.1	39.0
(4) レート (= (2)/(3) の*1)	-	0.000421	0.00430	0.0337	0.0725	0.0503	0.0687	4.22	9.07	12.2	6.13	7.41	14.2	14.2

かと考えられる。

個人的見解であります。

5 まとめ

本研究では、人工市場を用いてレバレッジド ETF のリバランス取引が原資産の価格変動に与える影響について検証した。その結果、レバレッジド ETF の初期キャッシュ量が大きくなるほど、原資産のリバランス取引が増大するとともにマーケットインパクトも大きくなり、原資産市場のボラティリティが大きくなることわかった。そして、レバレッジド ETF の売買が、少なくともボラティリティより大きなマーケットインパクトを与えると、市場を破壊してしまう可能性があることも判明した。今後の課題としては、レバレッジド ETF の売買がどの程度のマーケットインパクトを与えると市場が破壊されてしまうのか、詳細に調査することが挙げられる。

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の公式見解を表すものではありません。すべては

謝辞

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

参考文献

- [1] M. Cheng and A. Madhavan. The Dynamics of Leveraged and Inverse Exchange-Traded Funds. *Journal of Investment Management*, 7(4), 2009.
- [2] M. Childs. Fink Says Leveraged ETFs May 'Blow Up' Industry. *Bloomberg*, May 2014.
- [3] M. Deshpande, D. Mallick, and R. Bhatia. Understanding Ultrashort ETFs. *Barclays Capital Special Report*, January 2009.
- [4] S. Kobayashi and T Hashimoto. Analysis of institutional evolution in circuit breakers using the concepts of replicator and interactor. In *The 9th*

Asia-Pacific Complex Systems Conference, pages 80–86, 2009.

- [5] B. LeBaron. Agent-based financial markets: Matching stylized facts with style. In *Post Walrasian Macroeconomics Beyond the Dynamic Stochastic General Equilibrium Model*, chapter 3, pages 221–238. Cambridge University Press, 2006.
- [6] S. Martinez-Jaramillo and E. P. K. Tsang. An heterogeneous, endogenous and co-evolutionary gp-based financial market. *IEEE Transactions on Evolutionary Computation*, 13(1):33–55, 2009.
- [7] T. Mizuta, K. Izumi, I. Yagi, and S. Yoshimura. Regulations’ effectiveness for market turbulence by large erroneous orders using multi agent simulation. In *Proc. IEEE Computational Intelligence for Financial Engineering and Economics (CIFEr2014)*, pages 138–143, March 2014.
- [8] G. W. Schwert. Stock Volatility and the Crash of 87. *The Review of Financial Studies*, 3(1), 1990.
- [9] W. J. Trainor Jr. Do Leveraged ETFs Increase Volatility. *Technology and Investment*, 1, 2010.
- [10] I. Yagi and T. Mizuta. Analysis of the impact of leveraged etf rebalancing trades on the underlying asset market using artificial market simulation. In *Social Simulation Conference 2016(SSC2016)*, September 2016.
- [11] I. Yagi, T. Mizuta, and K. Izumi. A study on the effectiveness of short-selling regulation in view of regulation period using artificial markets. *Evolutionary and Institutional Economics Review*, 7(1):113–132, 2010.

人工市場を用いた分散投資規制が資産価格急落時の市場に与える影響分析

A Study on the Market Impact of the Rule for Investment Diversification at the Time of a Market Crash using an Multi-Agent Simulation

野崎 淳^{1*} 水田 孝信² 八木 勲³
Atsushi Nozaki¹ Takanobu Mizuta² Isao Yagi³

¹ 神奈川工科大学大学院情報工学専攻

¹ Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology

² スパークス・アセット・マネジメント株式会社

² SPARX Asset Management Co. Ltd

³ 神奈川工科大学情報学部

³ Faculty of Information Technology, Kanagawa Institute of Technology

Abstract: 投資信託の取引において、投資家が想定外のリスクを負うのを未然に防ぐため、分散投資規制が設けられた。分散投資規制は、1つの発行体への投資が過度に集中しないよう、1つの発行体への投資額を投資信託純資産の一定割合以下にするものである。これまでに、2資産市場の人工市場において分散投資規制が資産価格急落時の市場に与える影響について検証されており、資産価格が急落したとき全ての資産の保有額が純資産に対して非常に少なくなってしまうことが確認されている。一方、実際の投資信託運用者は、資産価格の変動などによって予め設定している保有額の下限を下回らないよう、保有額の少ない資産を買い増ししたりして純資産に対する全資産保有額を調整している。そこで本研究では、2資産市場の人工市場にまだ分散投資していない資産（資産3）を追加し、ある資産の価格が急落したときに分散投資規制が市場にどのような影響を与えるかを、純資産に対する全資産保有額の下限がある場合とない場合に分けて調査した。その結果、資産1のファンダメンタル価格が急落し、分散投資規制が導入されている市場において、純資産に対する資産の保有額に下限を設けた場合は、資産1と資産2の市場価格は下落する一方で、初期状態では分散投資されていなかった資産3が購入され、市場価格が上昇することが確認できた。

1 はじめに

近年金融商品の複雑化やリスクの複合化により、投資家がリスクを把握しにくくなっている。例えば、投資信託において、投資先資産のバリエーションは広がってきたが、ある特定の資産にリスクが集中しているかもしれない。また、ある投資信託の運用成績は、保有している特定の資産（例えばデリバティブ）の運用成績に大きく依存しているかもしれない。このような状況を投資家がすべて把握することは不可能である。

投資家が想定外のリスクを負うのを未然に防ぐため、

投資信託の運用会社側でリスクをコントロールすることが求められている。2013年には投資信託法制が改正され、2014年12月より分散投資規制という規制が設けられることとなった[1][2]。

分散投資規制は、1つの発行体への投資が過度に集中しないよう、1つの発行体への投資額を投資信託純資産の一定割合以下にするもので、欧州では既に同様の規制が適用されている。例えば、ある投資信託が特定の発行体への、1) 株式、2) 株式を原資産とするオプション、3) 社債に投資していた場合、それぞれの投資額はファンド純資産の10%以下、かつ、これらの合計投資額が投資信託純資産の20%以下とすることが求められている。ここで言う発行体とは、債券の発行元のことであり、株式や債券などの有価証券を投資家に売

*連絡先：神奈川工科大学大学院情報工学専攻
神奈川県厚木市下荻野 1030
E-mail:s1585027@cco.kanagawa-it.ac.jp

る企業、政府、金融機関などがそれに当たる。

分散投資そのものの是非についてはこれまでに多くの実証研究がなされている。例えば、Cremersらは分散投資しない方がポートフォリオの運用成績はよいことを示している [3]。一方、分散投資に制約を設けることによって市場の価格形成にどのような影響を与えるかは、これまで議論されたことはない。なぜなら、現実の市場には、さまざまな投資家が取引に参加しているため、分散投資の制約による取引のみが価格形成に与える影響を抽出することは困難であるからである。

このように特定の取引が市場に与える影響を分析する方法の1つとして、人工市場を用いる方法がある。人工市場とは、コンピュータ上に仮想的に構築されたマルチエージェントシステムの金融市場である [4][5][6]。エージェントにそれぞれ独自の売買手法を与え、それらを投資家として金融資産の取引をさせ、市場がどのような振る舞いをするかを確認することができる。その一方で、市場側に何らかの制約(市場安定化や効率化のための制度等)をモデル化し組み込むことで、投資家がどのように振る舞うか、さらに、彼らの振る舞いによって市場にどのような影響が現れるかを検討することも可能である。

人工市場を用いて金融規制を分析した研究はいくつかあるが、野崎らの研究を除いて分散投資規制を取り扱った研究は確認できなかった。これまでに野崎ら [7]は、2資産市場の人工市場において分散投資規制が資産価格急落時の市場に与える影響について検証した。その結果、分散投資規制が導入された市場では、2つの資産のうち、片方の資産のファンダメンタル価格が急落すると、もう片方の資産価格も連動して下落することが判明した。しかし、野崎らの人工市場モデルでは、資産価格が急落したとき全ての資産の保有額が純資産に対して非常に少なくなってしまう。一方、実際の投資信託運用者は、資産価格の変動などによって予め設定している保有額の下限を下回らないよう、保有額の少ない資産を買い増したりして純資産に対する全資産保有額を調整している。

そこで本研究では、2資産市場の人工市場にまだ分散投資していない資産(資産3)を追加し、ある資産の価格が急落したときに分散投資規制が市場にどのような影響を与えるかを、純資産に対する全資産保有額の下限がある場合とない場合に分けて調査した。その結果、全ての資産のファンダメンタル価格が一定のときは、分散投資規制の有無に関わらず、どの資産の市場価格も安定して推移することが確認できた。次に、資産1のファンダメンタル価格が急落するとき、分散投資規制が導入されている市場では、資産1の市場価格下落と連動して資産2の市場価格も下落することが確認できたが、導入されていない市場では資産1の市場価格のみが下落した。さらに、純資産に対する資産の

保有額に下限を設けた場合は、資産1と資産2の市場価格は下落する一方で、初期状態では分散投資されていなかった資産3が購入され、市場価格が上昇することが確認できた。

本論文の構成は以下の通りである。まず、2章では、本研究で使用した人工市場モデルについて説明する。2.1節では注文プロセスを、2.2節では学習プロセスを、2.3節では分散投資規制のモデルを記す。そして、3章では、前章で提案した人工市場モデルを用いたシミュレーション結果と考察を述べる。3.1節では、2章で提案した人工市場モデルの妥当性を検証した。3.2節では全ての資産のファンダメンタル価格が一定の場合、3.3節では、資産1のファンダメンタル価格が急落した場合に、分散投資規制が市場の価格形成に与える影響を調査した。3.4節では、一方の資産のファンダメンタル価格が急落したときのそれぞれの資産価格の変動を、分散投資規制が適用された市場とそうでない市場にわけて観察し、なぜそのような価格推移になったのか理由を調査する。最後に、4章ではまとめと今後の課題を記す。

2 人工市場モデル

2.1 注文プロセス

本モデルは2つのリスク資産のみを取引対象とし、価格決定メカニズムは連続ダブルオークション方式(ザラバ方式)とした。 n 体のエージェントがおり、エージェント番号 $j = 1$ から順番に $j = 2, 3, 4, \dots$ と注文を出す。最後のエージェント $j = n$ が注文を出すと、次の時刻にはまた初めのエージェント $j = 1$ から注文を出し繰り返される。時刻 t は1体のエージェントが注文を出すごとに1増える。つまり、注文をただけで取引が成立しない場合も1ステップ進む。エージェント j は注文価格、売り買いの別を以下のように決める。時刻 t にエージェント j が予想する価格の変化率(予想リターン) $r_{e,j,k}^t$ は、

$$r_{e,j,k}^t = \frac{1}{w_{1,j,k} + w_{2,j,k} + u_{j,k}} (w_{1,j,k} \log \frac{P_{f,k}^t}{P_k^{t-1}} + w_{2,j,k} r_{h,j,k}^t + u_{j,k} \epsilon_{j,k}^t) \quad (1)$$

ここで、 $w_{i,j,k}$ は時刻 t 、エージェント j 、リスク資産 k の i 項目の重みであり、シミュレーション開始時に、それぞれ0から $w_{i,max}$ まで一様乱数で決める。また、後述で述べる学習過程により変化する。 $u_{j,k}$ は時刻 t 、エージェント j 、資産 k の3項目の重みであり、シミュレーション開始時に、それぞれ0から u_{max} まで一様乱数で決め、その後も一定である。 $P_{f,k}^t$ は時間によらず一定の

資産 k のファンダメンタル価格, P_k^{t-1} は資産 k の時刻 $t-1$ での取引価格 (取引されなかった時刻では一番最近に取引された価格であり, 時刻 $t=0$ では $P_k^0 = P_{f_k}^0$ とする), $\epsilon_{j,k}^t$ は時刻 t , エージェント j , 資産 k の乱数項で, 平均 0, 標準偏差 σ_e の正規分布乱数である. $r_{h_{j,k}}^t$ は資産 k , 時刻 t にエージェント j が計測した過去リターンで, $r_{h_{j,k}}^t = \log(P_k^{t-1}/P_k^{t-\tau_j})$ である. ここで, τ_j は 1 から τ_{max} まで一様乱数でエージェントごとに決める.

式 (1) の第 1 項目はファンダメンタル価格と比較して安ければプラスの予想リターンを高めればマイナスの予想リターンを示す, ファンダメンタル価値を参照して投資判断を行うファンダメンタル投資家の成分である. 第 2 項目は過去のリターンがプラス (マイナス) ならプラス (マイナス) の予想リターンを示す, 過去の価格推移を参照して投資判断を行うテクニカル投資家の成分であり, 第 3 項目はノイズを表している.

予想リターン $r_{e_{j,k}}^t$ より予想価格 $P_{e_{j,k}}^t$ は,

$$P_{e_{j,k}}^t = P_k^{t-1} \exp(r_{e_{j,k}}^t) \quad (2)$$

で求まる. 注文価格 $P_{o_{j,k}}^t$ は $P_{e_{j,k}}^t - P_d$ から $P_{e_{j,k}}^t + P_d$ までの一様乱数で決める. ここで, P_d は定数である. 価格の変化幅の最小単位は δP とし, 売り買いの別は予想価格 $P_{e_{j,k}}^t$ と注文価格 $P_{o_{j,k}}^t$ の大小関係で決まる.

$$\begin{aligned} P_{e_{j,k}}^t > P_{o_{j,k}}^t &\text{なら 1 単位の買い} \\ P_{e_{j,k}}^t < P_{o_{j,k}}^t &\text{なら 1 単位の売り} \end{aligned} \quad (3)$$

注文数は常に 1 と一定とする.

各エージェントは, 注文価格や注文数を決定後, 分散投資規制や純資産に対する全資産保有額に下限を導入している場合はそれらに違反していないか判定を行う. 分散投資規制や純資産に対する全資産保有額の下限の詳細は 2.3 節で説明する. 規制を満たしている場合のみ注文を出し, 満たしていない場合は注文を出さない.

本モデルの価格決定メカニズムはザラバ方式であるため, 買い (売り) 注文は, 注文価格より安い (高い) 注文が既に存在すれば, 最も安い (高い) 売り (買い) 注文と即座に取引が成立する. また, 約定できる注文が市場に無ければ今回の注文を残す. 市場に残した注文がキャンセル時間 t_c だけ経過しても取引が成立しなかった場合は, キャンセルされる.

2.2 学習プロセス

戦略を状況に応じて切り替えるという学習プロセスを, 以下のようにモデル化した. 学習はエージェントごとに, 各エージェントが注文を出す直前に行われる.

ファンダメンタルな投資家成分だけの場合の予想リターンを $r_{e_{1,j,k}}^t = \log(P_{f_k}^t/P_k^t)$, テクニカルな投資家成分だけの場合の予想リターンを $r_{e_{2,j,k}}^t = r_{h_{j,k}}^t$ とする ($k = 1, 2, 3$). これら $r_{e_{i,j,k}}^t$ が資産 k の学習期間のリターン $r_{i_k}^t = \log(P_k^{t-1}/P_k^{t-t_i})$ と比べ,

$$\begin{aligned} \text{同符号なら, } w_{i,j,k} &\leftarrow w_{i,j,k} + k_l r_{i_k}^t p_j^t (w_{i,max} - w_{i,j,k}) \\ \text{異符号なら, } w_{i,j,k} &\leftarrow w_{i,j,k} - k_l r_{i_k}^t p_j^t w_{i,j,k} \end{aligned} \quad (4)$$

のように, $w_{i,j,k}$ を書き換える. ここで, k_l は定数, p_j^t は時刻 t , エージェント j に与えられる 0 から 1 までの一様乱数である.

このような過去の実績から学習するプロセスとは別に, 小さい確率 m で $w_{i,j,k}$ を再設定する. つまり, 0 から $w_{i,max}$ までの一様乱数で決めなおす. これはランダム学習を意味しており, 実績からの学習と組み合わせることにより, エージェントが試行錯誤的により良い戦略のウェイトを求める姿を客観的にモデル化している.

2.3 規制のモデル化

本モデルでは, 分散投資規制と純資産に対する資産保有額の上限・下限を実装している. 純資産に対する資産保有額の上限を次のようにモデル化する.

$$\sum_{k=1}^3 |P_k^t S_{j,k}^t| \leq v_{upper} NAV_j^t \quad (5)$$

ここで, $S_{j,k}^t$ は, 時刻 t におけるエージェント i , 資産 j の保有量であり, $S_{j,k}^t > 0$ なら買い保有している状態, $S_{j,k}^t < 0$ なら空売りしている状態を表している. v_{upper} は純資産に対する全資産保有額の上限比率を表し, 今回は $v_{upper} = 1$ とする. C_j^t は時刻 t におけるエージェント j のキャッシュの保有量を表している. NAV_j^t は純資産であり,

$$NAV_j^t = \sum_{k=1}^3 (P_k^t S_{j,k}^t + C_j^t) \quad (6)$$

である.

次に, 純資産に対する資産保有額の下限を次のようにモデル化する.

$$v_{lower} NAV_j^t \leq \sum_{k=1}^3 (P_k^t S_{j,k}^t) \quad (7)$$

ここで, v_{lower} は純資産に対する全資産保有額の下限比率のことを示し, 今回は 0.8 とする. 各エージェン

表 1: パラメータ設定

parameters	value
n	1,000
$w_{1,max}$	1
$w_{2,max}$	10
u_{max}	1
τ_{max}	10,000
σ_e	0.03
P_d	1,000
t_c	13,000
t_l	10,000
k_l	4
m	0.01

トは純資産の v_{lower} 倍が全資産保有額の下限になるように設定し、資産保有額が下限を下回った場合、その状態を解消するために1単位の成行買い注文を出す(これ以降、上記の注文を下限違反買い注文と呼ぶ)。そして、下限違反買い注文でどの資産を買うのかは、分散投資規制に違反していない資産で最も予想リターンが高い資産を購入する。なお、純資産に対する資産保有額の下限を設けて実験を行うのは、3.4節だけである。

次に、分散投資規制を次のようにモデル化する。

$$\frac{|P_k^t S_{j,k}^t|}{NAV_j^t} \leq w_{dir} \quad (8)$$

w_{dir} は分散投資規制の規制比率であり、今回は $w_{dir} = 0.5$ とした。各エージェントは、注文を発注する際、その注文が規制に違反する、つまり、上記の式(8)を満たさない場合は、その注文をキャンセルする。また、すでに式(8)に違反している場合は、規制に違反した状態を解消するために1単位の成行注文を出す。これ以降上記の注文を規制違反注文(買い注文の場合は規制違反買い注文、売り注文の場合は規制違反売り注文)とする。

3 シミュレーション結果

本研究では、パラメータを表1のように設定し、シミュレーションは $t = 1,000,000$ まで行った。また、各実験におけるエージェントの初期キャッシュ量 (C_j^0) を 100,000、資産1と資産2の初期保有量 ($S_{j,1}^0, S_{j,2}^0$) を 48、資産3の初期保有量 ($S_{j,3}^0$) を 0 とした。

3.1 モデルの妥当性

シミュレーション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

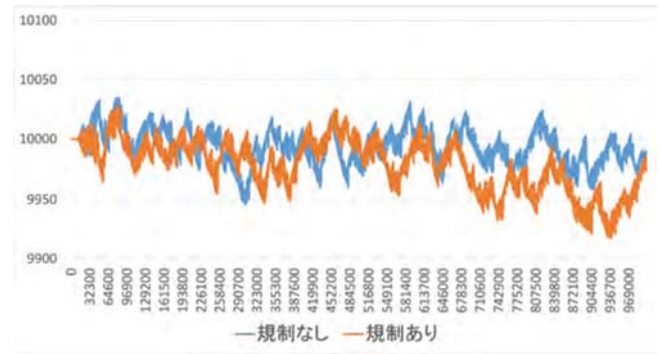


図 1: ファンダメンタル価格一定時の資産1の価格推移

的な stylized fact が再現できるかどうかで評価される [8][9]. fat-tail は、市場価格の騰落率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。また、volatility-clustering は市場価格の騰落率の2乗が、ラグが増えても自己相関係数が有意に正であり、長期記憶性を持つことである。本研究では、実証研究で得られた fat-tail と volatility-clustering が再現できるようにパラメータを調整した。

表2は各試行での stylized fact を示している。各々の場合について100回の試行を行い平均値をとった。Stylized fact はすべて100ステップ間隔での価格を用いて計算した。というのも、本モデルの時刻は注文をしただけで取引が成立しない場合も時刻が進むため、時刻1ごとの全ての価格を用いた stylized fact は多くの価格変動がないデータによりバイアスがかかってしまうため、100ステップ間隔での測定とした。表2の結果より、いずれのシミュレーションでも尖度がプラスの値で fat-tail となっており、騰落率の2乗の自己相関もラグが増えてもプラスで volatility-clustering も再現されている。これらにより長期に観測される価格変動の統計的な性質を再現していることが示された。

3.2 全ての資産のファンダメンタル価格が一定の場合

この節では、全ての資産のファンダメンタル価格が一定の場合 ($P_{f_1}^t = P_{f_2}^t = P_{f_3}^t = 10,000$) の実験とその結果を示す。図1には資産1、図2には資産2、図3には資産3の分散投資規制が導入されている市場と導入されていない市場の価格推移を示す。

図1、図2、図3から、全ての資産のファンダメンタル価格が一定のときは、分散投資規制の有無にかかわらず、どの資産の市場価格も安定して推移していることが確認できた。

表 2: 各試行における Stylized Fact

分散投資規制の有無	資産 1		資産 2		資産 3		
	規制なし	規制あり	規制なし	規制あり	規制なし	規制あり	
尖度	4.18	3.89	4.09	4.04	4.16	4.22	
価格騰落率の 2乗の 自己相関	ラグ						
	1	0.12	0.15	0.12	0.13	0.11	0.11
	2	0.09	0.11	0.09	0.10	0.10	0.10
	3	0.08	0.08	0.07	0.07	0.08	0.08
	4	0.06	0.07	0.06	0.06	0.06	0.06
	5	0.06	0.06	0.05	0.06	0.06	0.05
	6	0.05	0.05	0.05	0.04	0.05	0.04

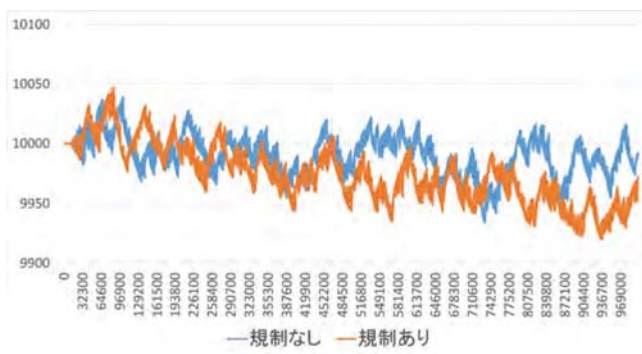


図 2: ファンダメンタル価格一定時の資産 2 の価格推移

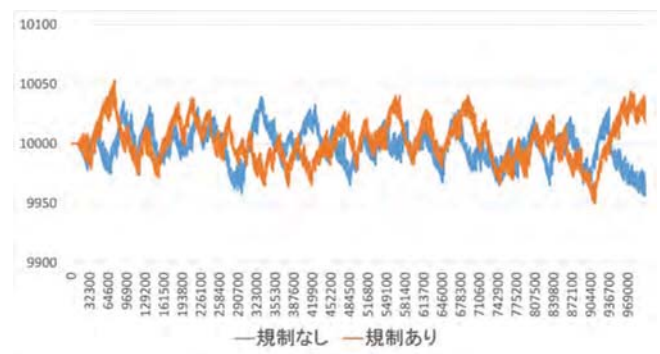


図 3: ファンダメンタル価格一定時の資産 3 の価格推移

3.3 資産 1 のファンダメンタル価格が急落する場合

本節では、資産 1 のファンダメンタル価格が急落する場合 ($P_{f_2}^t = P_{f_3}^t = 10,000$ 、時刻 $t = 0$ から $t = 100,000$ までは $P_{f_1}^t = 10,000$ 、 $t = 100,001$ 以降は $P_{f_1}^t = 7,000$) の実験とその結果を示す。図 4 には分散投資規制が導入されていない市場、図 5 には分散投資規制が導入されている市場の価格推移を示す。

図 4 と図 5 より、資産 1 と資産 3 の市場価格には大きな違いは見られない。一方、資産 2 の市場価格は、資産 1 の市場価格の下落より少し遅れて下落し、ファンダメンタル価格よりも低い価格 (今回は約 9,000) に収束している。

資産 2 の市場価格が資産 1 の市場価格に連動して下落した理由は次のように考えられる。まず、資産 1 の市場価格が急落し始めるとエージェントの純資産は小さくなるため、多くのエージェントの資産 2 の保有額が規制に違反した状態となる。すると、規制違反状態を解消するために規制違反売り注文が発生するため、資産 2 の市場価格も下落するものと考えられる。

具体例として、図 6 のような場合を考える。資産 1

の市場価格急落前の純資産は 1,000,000、資産 1 と資産 2 の保有額はそれぞれ 500,000 で、 w_{dir} は 0.5 である (図 6 左)。仮に資産 1 の市場価格が 30% 下落したとすると、純資産は 850,000 に、資産 1 の保有額は 350,000 に減少する (図 6 右)。しかし、資産 2 の保有額は変わらないため、純資産に対する資産 2 の保有額が 50% を超えてしまい、分散投資規制に違反する。よって規制違反状態を解消するため、75,000 の規制違反売り注文が発生する。これらの行動を多くのエージェントがとることにより、資産 2 の取引価格が下落すると考えられる。

資産 2 の市場価格が下落後上昇する理由は次のように考えられる。資産 1 の市場価格の下落が収まると、資産 2 の規制違反状態も次第に解消される。それに伴い、規制違反売り注文が発生しなくなり、資産 2 の価格の下落が止まる。このとき資産 2 の市場価格がファンダメンタル価格に比べて非常に低くなっているため、エージェントの多くが資産 2 を割安だと判断し、プラスの予想リターンを出すことから成行買い注文が増加する。その結果、資産 2 の市場価格が上昇する。

資産 2 の市場価格がファンダメンタル価格よりも低い価格に収束する理由は、次のように考えられる。資

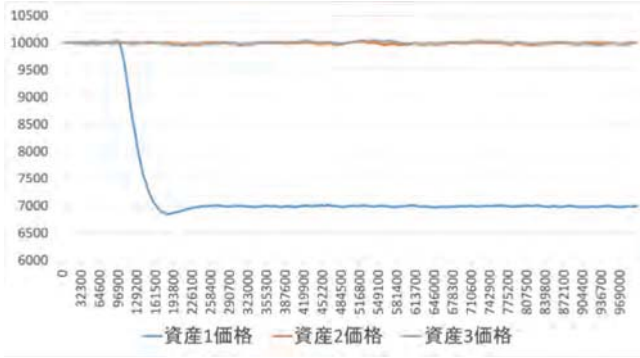


図 4: 資産 1 のファンダメンタル価格急落時における分散投資規制が導入されていない場合の価格推移

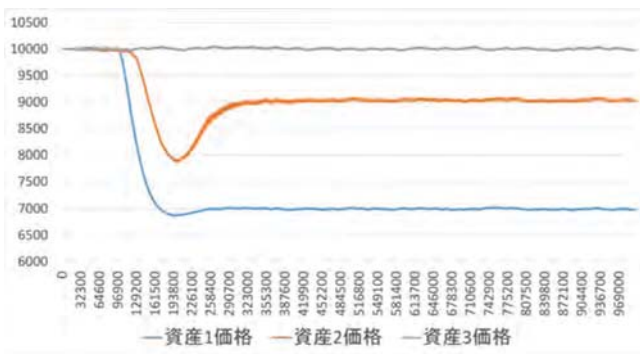


図 5: 資産 1 のファンダメンタル価格急落時における分散投資規制が導入されている場合の価格推移

資産 2 の市場価格の下落が収束後上昇することで、再び分散投資規制に違反するエージェントが増加し、規制違反売り注文と成行買い注文がほぼ同数になることで、価格が収束していると考えられる。

以上の資産 2 の価格推移についての仮説のうち、まず資産 2 の市場価格が資産 1 の市場価格に連動して下落する理由が正しいか検証する。今回は、資産 1 のファンダメンタル価格急落直後から資産 2 の価格の下落が収まるまでの期間の 3 種類の注文（規制違反売り注文、成行売り注文、成行買い注文）を調査した（図 7）。図 7 より、ファンダメンタル価格が急落した後しばらくしてから、大量の資産 2 の規制違反売り注文が発生していることが分かる。ただし、資産 2 の市場価格の下落は約 200,000 期まで続くが、規制違反売り注文は約 140,000 期をピークに減っていることも確認できる。この理由は、多くのエージェントの資産 2 の保有額の規制違反状態が、取引が進む中で解消されていくからである。規制違反売り注文が減った後も資産 2 の市場価格が下落を続けている理由は、市場価格トレンドが一定期間一方向に続くと、多くのエージェントが注文を

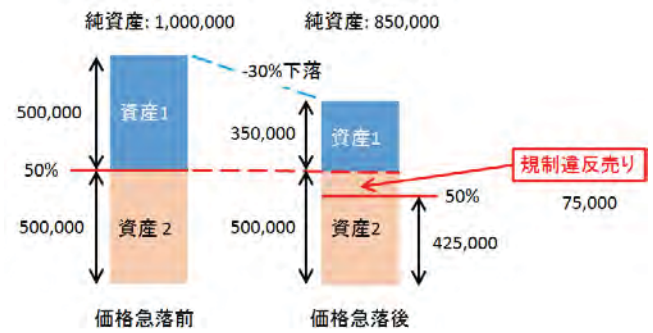


図 6: 価格急落時の規制違反注文を出すまでのメカニズム

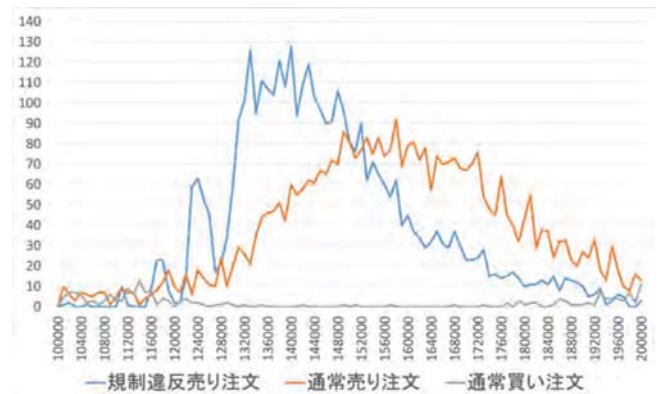


図 7: 100,000 期から 200,000 期における資産 2 の約定注文数

決める際、テクニカル要素の影響を強く受けてマイナスの予想リターンを出すことから成行売り注文が増加するからである [10]。

次に、資産 2 の市場価格が下落後上昇し、ファンダメンタル価格よりも低い価格に収束する理由が正しいか検証するため、資産 2 の市場価格が最低になったときから価格が収束するまでの期間の 3 種類の注文（規制違反売り注文、成行売り注文、成行買い注文）を調査した（図 8）。図 8 より、資産 2 の市場価格が最低値になった時から、市場価格が安定するまでの期間（約 200,000 期～300,000 期にかけて）と、成行買い注文が売り注文と比較して多い期間が、ほぼ一致していることが確認できる。また、成行買い注文が売り注文とほぼ同数になる時期が資産 2 の市場価格が収束し始める時期とほぼ一致していることも確認できる。以上の結果より、資産 2 の価格推移についての仮説は正しいと考えられる。ただし、資産 2 の市場価格がファンダメンタル価格よりも安い価格に収束するという結果は現実では見られない可能性が高い。なぜなら、今回の実験ですべてのエージェントが分散投資規制を適用さ



図 8: 200,000 期から 500,000 期における資産 2 の約定注文数

れているが、現実の市場では分散投資規制が適用されない投資家も数多く存在するからである。

3.4 純資産に対する全資産保有額に下限を設けたとき

本人工市場モデルでは、各エージェントは保有する資産価格が最大になるよう取引を行うが、現実の市場で投資信託を運用する機関投資家が必ずしも同じような取引を行うとは限らない。なぜなら、多くの投資信託では、預かった資金のほとんどを資産の購入にあてる (full investment) ことを課せられ、キャッシュは少しの保有しか許されていないので、通常純資産に対して一定比率以上の資産を保有するような取引を行うからである。よって本節では、3.3 節の実験環境に、純資産に対する資産保有額に下限を設け、実験を行う。

図 9 に純資産に対する資産保有額に下限を設けた場合 ($v_{lower} = 0.8$) の価格推移を示す。図 5 と図 9 を比較して、図 9 の資産 1 の市場価格が図 5 の市場価格よりもファンダメンタル価格到達後大きくオーバーシュートしていることがわかる。また、図 9 における資産 2 の市場価格は、図 5 の市場価格よりも下落の幅が大きく、その後上昇して収束する価格は図 5 よりも低いことが見てとれる。そして、図 9 における資産 3 の市場価格は、図 5 とは異なり資産 2 の市場価格下落とほぼ同時に上昇している。

資産 1 の市場価格のオーバーシュート量が大きくなった理由は次のように考えられる。エージェントの資産 3 の初期保有株数は 0 であるため、資産 3 を空売りするエージェントも存在し、資産 3 の価格が上昇することでそのエージェントの純資産は小さくなる。その結果資産 1 が規制違反状態となり、規制違反売り注文が発生することで、資産 1 の市場価格がさらに下落する

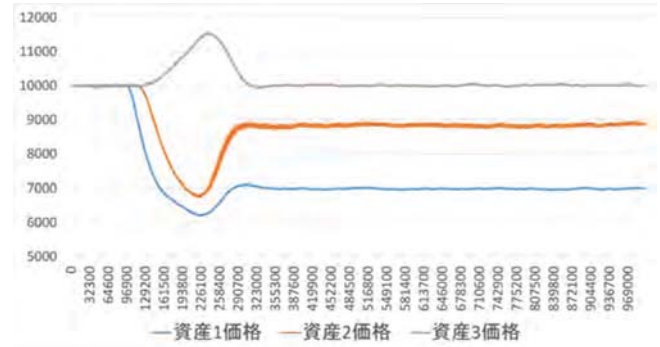


図 9: 純資産に対する資産保有額に下限を設けた場合の価格推移

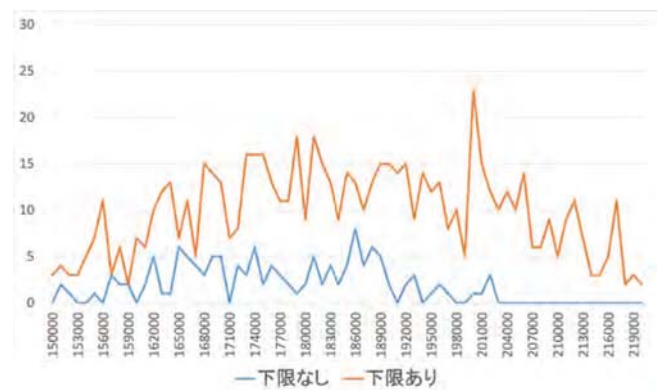


図 10: 150,000 期から 220,000 期における資産 1 の約定注文数

と考えられる。また、資産 2 の市場価格の下落幅が大きくなるのも同様の理由であると考えられる。

上記の仮説が正しいか検証するため、資産 1 のファンダメンタル価格急落直後から、資産 1 および資産 2 の市場価格の下落が収まるまでの期間の 4 種類の注文 (規制違反売り注文, 通常売り注文, 通常買い注文, 下限違反買い注文) を調査した (図 10, 図 11)。図 10 より、資産 1 の市場価格がファンダメンタル価格に到達してから価格の下落が収束するまでの期間と、規制違反売り注文が増える期間が一致していることが確認できる。図 11 より、図 7 と比較して規制違反売り注文が多く、それによりテクニカル要素の影響を強く受けて成行売り注文も増えている。以上より、資産 1 と資産 2 の市場価格の下落幅が大きくなる仮説は正しいと考えられる。

次に、資産 3 の市場価格が上昇した理由は、以下のように考えられる。資産 1 のファンダメンタル価格急落に伴い、資産 1 と資産 2 の市場価格が下落し、純資産に対する両資産の保有額の割合が低下する。すると、



図 11: 100,000 期から 230,000 期における資産 2 の約定注文数

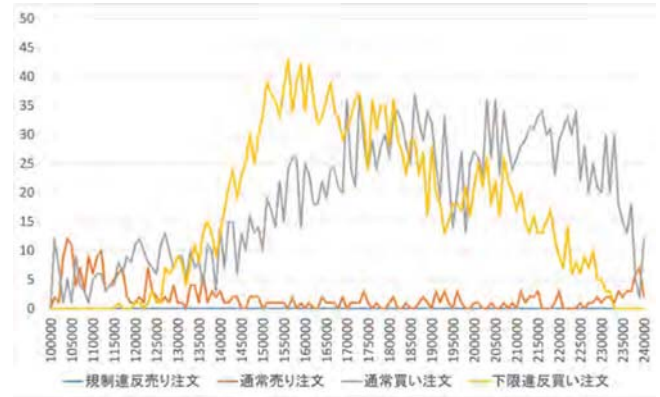


図 13: 100,000 期から 240,000 期における資産 3 の約定注文数



図 12: 価格急落時の下限違反注文を出すまでのメカニズム

純資産に対する資産保有額の下限に違反するので、下限違反買い注文が発生するが、資産 1 と資産 2 は分散投資規制に違反するため、これ以上買うことができない。よって、資産 3 の下限違反買い注文が発生するため、資産 3 の市場価格が上昇したと考えられる。

具体例として、図 12 のような場合を考える。純資産に対する資産保有額の下限比率を 0.95 とする。図 6 右の状態から規制違反売り注文が発生して資産 2 の規制違反状態が解消された場合を考える (図 12 左)。このときのエージェントの純資産は 850,000、保有キャッシュは 75,000、資産 1 の保有額は 350,000、資産 2 の保有額は 425,000 である。現在の純資産に対する資産保有額の比率は 0.91 であり、純資産に対する資産保有比率の下限に違反する。そのため、32,500 の資産 3 の下限違反買い注文が発生するので、資産 3 の市場価格が上昇すると考えられる。資産 3 の市場価格の上昇が資産 1 の市場価格の下落よりも遅れているのは、本モデルにおいて注文量は常に 1 であるため、ファンダメンタル価格急落に伴う資産の売却に時間がかかるからであると考えられる。

上記の仮説が正しいか検証するため、資産 1 のファンダメンタル価格急落直後から、資産 3 の市場価格の上昇が収まるまでの期間の 4 種類の注文 (規制違反売り注文、成行売り注文、成行買い注文、下限違反買い注文) を調査した (図 13)。図 13 より、資産 3 の市場価格が上昇し始める期間 (約 110,000 期) から下限違反買い注文が発生していることが確認できる。よって、この仮説は正しいと考えられる。下限違反買い注文は約 160,000 期をピークに減っているが資産 3 の市場価格が上昇は約 240,000 期まで続いている。この理由は、市場価格トレンドが一定期間一方方向に続くと、多くのエージェントが、テクニカル要素の影響を強く受けてプラスの予想リターンを出し、成行買い注文が発生するようになるからである [10]。

4 まとめと今後の課題

本研究では、人工市場にまだ分散投資していない資産 (資産 3) を追加設定し、全ての資産のファンダメンタル価格が一定のときと、ある資産のファンダメンタル価格が急落したときに、分散投資規制が市場にどのような影響を与えるかを調査した。さらに、分散投資規制下において、ある資産の価格が急落純資産に対する全資産保有額の下限がある場合とない場合に分けて調査した。

その結果、全ての資産のファンダメンタル価格が一定のときは、分散投資規制の有無にかかわらず、どの資産の市場価格も安定して推移することが確認できた。次に、資産 1 のファンダメンタル価格が急落するとき、分散投資規制が導入されている市場では、資産 2 の市場価格は資産 1 の市場価格に連動して下落することが確認できたが、導入されていない市場では資産 1 の市場価格のみが下落した。さらに、純資産に対する資産

の保有額に下限を設けた場合は、資産 1 と資産 2 の市場価格は下落する一方で、資産 3 の市場価格が上昇することが確認できた。

今後の課題としては、以下の通りである。本人工市場モデルは 3 資産市場であるが、現実の金融市場では取引可能な資産は数多く存在するため、さらに資産数を増やした人工市場で実験を行うことが挙げられる。さらに、エージェントの 1 回あたりの注文数は 1 に固定されているため、市場の動きが現実市場より緩やかになっている可能性がある。よって各エージェントの 1 回あたりの発注数を 1 以上にすることが挙げられる。

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。

謝辞

本研究は公益財団法人全国銀行学術研究振興財団と JSPS 科研費 15K01211 の助成を受けたものです。この場を借りてお礼申し上げます。

参考文献

- [1] Sectional Committee of the Financial System of the Financial System Council, “Final Report by ”Working Group on Review of Investment Trust and Investment Corporation Regulation”, Dec. 2009.
- [2] J. The Investment Trusts Association, “Investment Trusts in Japan 2014,” 2014.
- [3] M. Cremers and A. Petajisto, “How Active Is Your Fund Manager? A New Measure That Predicts Performance,” *Review of Financial Studies*, vol.22, no.9, 2009.
- [4] C. Chiarella, G. Iori, and J. Perelló, “The Impact of Heterogeneous Trading Rules on the Limit Order Flows,” *Journal of Economic Dynamics and Control*, vol.33, no.3, 2009.
- [5] S.H. Chen, C.L. Chang, and Y.R. Du, “Agent-based Economic Models and Econometrics,” *Knowledge Engineering Review*, vol.27, no.2, 2012.
- [6] T. Mizuta, K. Izumi, I. Yagi, and S. Yoshimura, “Regulations’ effectiveness for market turbulence by large erroneous orders using multi agent simulation,” *Proc. IEEE Computational Intelligence for Financial Engineering and Economics (CIFEr2014)*, pp.138–143, March 2014.
- [7] A. Nozaki, T. Mizuta, and I. Yagi, “Investigation of the rule for investment diversification using an artificial market,” *The 16th Workshop on JSAI Special Interest Group on Financial Informatics*, 2016. (in Japanese).
- [8] M. Sewell, “Characterization of financial time series,” 2006.
- [9] R. Cont, “Empirical properties of asset returns: stylized facts and statistical issues,” *Quantitative Finance*, vol.1, pp.223–236, 2001.
- [10] T. Mizuta, K. Izumi, I. Yagi, and S. Yoshimura, “Investigation of price variation limits, short selling regulation, and uptick rules and their optimal design by artificial market simulations,” *Electronics and Communications in Japan*, vol.98, no.7, pp.13–21, 2015.

心理的要素を考慮した投資行動モデル

The Investment Decision Making Model

Considering Psychological Factor

宮坂 純也¹ 穴田 一²

Junya Miyasaka¹, Hajime Anada²

¹ 東京都市大学大学院工学研究科

¹ Tokyo City University Graduate Division, Graduate School Of Engineering

² 東京都市大学知識工学部

² Tokyo City University Under Graduate Division, Faculty Of Knowledge Engineering

Abstract: In recent years, many researchers pay attention to behavioral finance. Behavioral finance focuses on investor's irrationality. In the field of behavioral finance, researchers are trying to elucidate cause of investor's decision making by observation. Some researchers constructed mathematical model of investor's decision making considering investor's psychological factor using an artificial market. However, these researches didn't fully express the psychological factors and real investor's decision making. Therefore, we construct investor's decision making model depending on investor's loss and profit using an artificial market. And we tested efficiency of our model.

1. はじめに

従来の伝統的な経済学では効率的市場仮説に基づいて議論がなされている。効率的市場仮説では投資家は常に合理的な投資行動を取るとされており、株価に影響を与えるような情報は即座に株価に反映されるとしている。そのため、投資家は継続的に収益を上げることは出来ないと考えられている。しかし、現実の市場において継続的に収益を上げ続ける投資家が存在する。このような現象は投資家の非合理的な行動によって引き起こされていると考えられている。しかし、株式市場に対して直接的に実験を行い、検証することは不可能である。そこで、コンピュータ上に仮想的な人工市場を構築し、様々な実験及び検証が行われている[1][2][3]。

現在、常に合理的な投資行動を行う投資家を前提とせず、現実の投資家の行動を観察し、その意思決定を行う原因を心理学の面からアプローチし、解明しようとしている行動経済学に注目が集まっている[4][5][6]。また、投資家の行動の心理的な要素に着目し、人工市場を用いてアプローチした先行研究として並河らの研究がある[7]。並河らは不確実性下の実

証的意思決定論であるプロスペクト理論を投資家の行動に取り入れたモデルを構築した。プロスペクト理論とは利益と損失が同額であった場合には損失を過大に解釈する、という価値の評価における非対称性を表した理論である[8][9]。並河らのモデルでは、投資行動を購入、売却、何もしないをランダムに決定後、株価移動平均と投資家の予想価格を比較する。そして、予想価格が株価移動平均を下回って損すると予想した際には、損益を利益より過大評価するため、得をすると予想したときよりも購入量を減らし、売却量を増やすようなバイアスがかかるようにしている。このモデルではランダムに決定する投資行動の投資量を変えることでプロスペクト理論を表現している。しかし、このモデルは、投資家の重要な要素である投資行動選択を全てランダムに決定することや、投資家心理が損益予測にのみ影響を受けていると仮定していることが問題として挙げられる。

そこで、本研究では実際の市場に存在し、株価トレンドを基に投資判断を行う順張り投資家及び逆張り投資家、株価トレンドに関係なく投資行動を行うランダム投資家を導入した。また、投資家の株式投資における投資収益が、投資家の投資判断に最も大きく影響すると考え、本研究では投資家の収益性に基づいて意思決定が変化するモデルを構築した。そして、このモデルの有効性について検証した。

連絡先：宮坂純也，東京都市大学大学院工学研究科

〒158-8557 東京都世田谷区玉堤 1-28-1

g1581820@tcu.ac.jp

2. 提案手法

本研究では各投資家は現金と株式のみを保有し、取引可能な銘柄は1つとする。また、市場内には順張り投資家、逆張り投資家、ランダム投資家の3種類のみが存在するものとする。順張り投資家及び逆張り投資家はトレンドの向き、株価と株価移動平均との乖離度合いを基に投資行動を決定する。ランダム投資家は投資行動をランダムに決定する。本研究では実市場の25期間の終値を取得し、続く T 期間実験を行う。その際、初期資産として25期目の株価で100株相当分の現金及び株式100株分を投資家に持たせる。また、約定方法については取引が1期1回の板寄せ方式を用いる。これは、現実の株式市場では板寄せ方式とザラバ方式の2つが用いられているが、本研究では各期の終値を基に予想価格及び注文量を決定しているためである。

2.1. 投資行動確率の決定

順張り投資家及び逆張り投資家は始めにトレンドの向きにより購入、売却を選択する。しかし、実際に投資行動を起こすかどうかについては株価と株価移動平均との乖離度合いにより決定する。 t 期における株価と株価移動平均との乖離度合いを表す $S(t)$ は次式で表される。

$$S(t) = \frac{p(t) - MA_{25}(t)}{\sigma(t)} \quad (1)$$

ここで、 $p(t)$ は t 期における株価、 $MA_{25}(t)$ は t 期における過去25期間の株価移動平均、 $\sigma(t)$ は t 期における過去25期間の株価の標準偏差を表す。

2.1.1. 順張り投資家

順張り投資家は株価トレンドに沿った取引を行う投資家である。そのため、上昇トレンド時には購入、下落トレンド時には売却を行う。順張り投資家 i は t 期における過去 q_i 期間(q_i は順張り投資家 i のトレンド考慮期間)の株価に対してあてた回帰直線の傾きである $trend_{q_i}(t)$ が正である場合に上昇トレンドと判断し、次式で表される購入確率 $P_i^{fb}(t)$ で株を購入する。

$$P_i^{fb}(t) = \frac{1}{1 + \exp\{-a \times (S(t) - j_{buy} + b \times n_i(t))\}} \quad (2)$$

$$n_i(t) = \frac{Vsum_i^d(t) - Vsum_i(0)}{Vsum_i(0)}$$

ここで、 a は $S(t)$ が変化した場合の購入確率の立ち上がりやすさを決めるためのパラメータ、 j_{buy} は $P_i^{fb}(t)$ が0.5となる $S(t)$ の値を決めるためのパラメータ、 b は

資産が変化した場合に投資家が感じる心理的インパクトの強さを決めるためのパラメータ、 $Vsum(0)$ は投資家の初期保有総資産、 $Vsum_i^d(t)$ は t 期における投資家 i の過去 d 期間の総資産移動平均を表す。また、この時 $1 - P_i^{fb}(t)$ の確率で何もしないを選択する。

図1に t 期における、 $j_{buy} = -1.0$ の際の順張り投資家 i の購入時の行動確率 $P_i^{fb}(t)$ を掲載する。

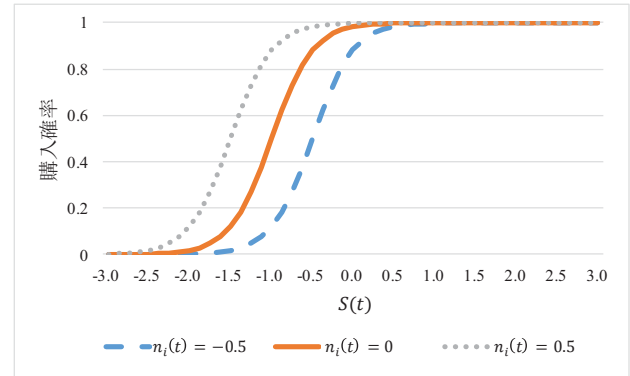


図1 順張り投資家の購入確率

順張り投資家は、上昇トレンドと判断した際に、 $S(t)$ が負から正に変化するにつれて、購入しやすくなる。そして、投資により利益を上げている投資家は自信があるため、 $n_i(t) = 0$ 時と比べて積極的に購入するようになる。損失を出している投資家は自信がなくなり慎重になっているため、 $n_i(t) = 0$ 時と比べて消極的に購入するようになる。

次に、下落トレンド時における順張り投資家の行動について記す。順張り投資家 i は $trend_{q_i}(t)$ が0以下の際に、下落トレンドと判断し、次式で表される売却確率 $P_i^{fs}(t)$ で株を売却する。

$$P_i^{fs}(t) = \frac{-1}{1 + \exp\{-a \times (S(t) - j_{sell} + b \times n_i(t))\}} + 1 \quad (3)$$

ここで、 j_{sell} は $P_i^{fs}(t)$ が0.5となる $S(t)$ の値を決めるためのパラメータを表す。また、この時 $1 - P_i^{fs}(t)$ の確率で何もしないを選択する。

図2に t 期における、 $j_{sell} = 1.0$ の際の順張り投資家 i の売却時の行動確率 $P_i^{fs}(t)$ を掲載する。

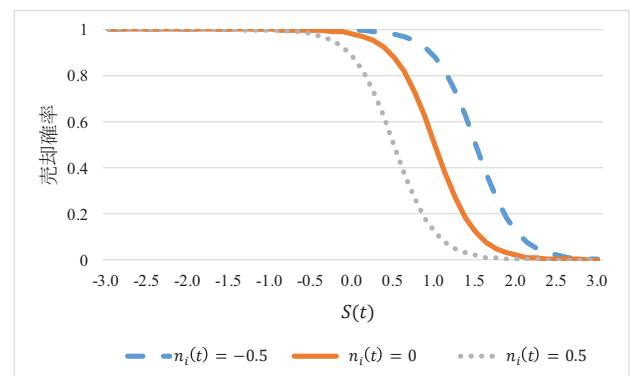


図2 順張り投資家の売却確率

順張り投資家は下落トレンドと判断した際に、 $S(t)$ が正から負に変化するにつれて、売却しやすくなる。そして、投資により利益を上げている投資家は自信があるため、 $n_i(t) = 0$ 時と比べて積極的に売却を行うようになる。損失を出している投資家は自信がなくなり慎重になっているため、 $n_i(t) = 0$ 時と比べて消極的に売却するようになる。また、 $S(t) < 0$ の際には下落トレンドであると判断し、100%に近い確率で売却を選択する。

2.1.2. 逆張り投資家

逆張り投資家とは、下落トレンド時に底値で買い、上昇トレンド時に天井値で売ることによって利益をあげようとする投資家である。逆張り投資家 i は次式で表される $P_i^c(t)$ で株を売買する。

$$P_i^c(t) = \frac{1}{1 + \exp\{-a \times (|S(t)| - g + b \times n_i(t))\}} \quad (4)$$

ここで、 g は $P_i^c(t)$ が 0.5 となる $S(t)$ の値を決めるためのパラメータを表す。逆張り投資家は $S(t) > 0$ を満たす際に上昇トレンドと判断し、 $P_i^c(t)$ の確率で売却を選択する。 $S(t) \leq 0$ を満たす際には下落トレンドと判断し、 $P_i^c(t)$ の確率で購入を選択する。また、この時逆張り投資家は購入及び売却時に $1 - P_i^c(t)$ の確率で何もしないを選択する。

図 3 に t 期における、 $g = 1.0$ の際の逆張り投資家 i の行動確率 $P_i^c(t)$ を掲載する。

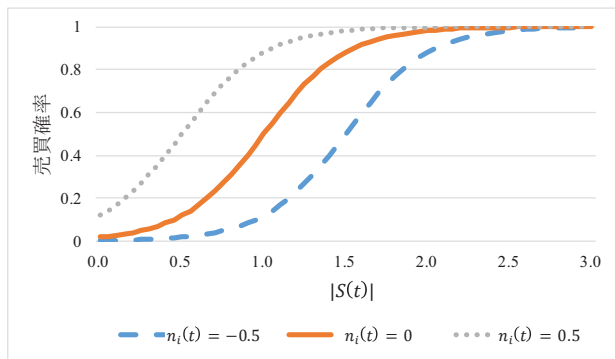


図 3 逆張り投資家の売買確率

逆張り投資家は $|S(t)|$ の値が大きくなるほど売却を行いやすくなる。そして、投資により利益を上げている投資家は自信があるため $n_i(t) = 0$ 時と比べて積極的に株の購入、売却を行うようになる。損失を出している投資家は自信がなくなり慎重になっているため、 $n_i(t) = 0$ 時と比べて消極的に株の購入、売却をするようになる。

2.1.3. ランダム投資家

ランダム投資家は株価トレンド及び損益状態に影響を受けることなく、購入、売却、そして何もしない、の 3 択を等確率でランダムに選択する。

2.2. 取引の流れ

以下に本モデルでの取引の流れを示す。Step.1 で投資家の初期設定を行い、step.2 から step.6 を T 期間繰り返す。

Step.1 初期設定

導入期間である 25 期の終値を取得する。そして、25 期の株価で全ての投資家に 100 株相当分の現金及び株式 100 株を持たせる。そして、順張り投資家のトレンド考慮期間 $q_i(t)$ を $[5, q]$ の一様乱数で決定する。

Step.2 全投資家の総資産計算

全投資家の総資産を計算する。 t 期における投資家 i の総資産 $Vsum_i(t)$ は次式で表される。

$$Vsum_i(t) = Vm_i(t) + Vs_i(t) \times p(t) \quad (5)$$

ここで、 $Vm_i(t)$ は t 期における投資家 i の保有現金、 $Vs_i(t)$ は t 期における投資家 i の保有株式を表す。

Step.3 全投資家の予想価格計算

全投資家の予想価格を計算する。 t 期における投資家 i の予想価格 $Ex_i(t)$ は次式で表される。

$$Ex_i(t+1) = p(t)(1 + \mu + \sigma_{rate}\epsilon) \quad (6)$$

ここで μ は t 期における過去 25 期の株価変化率の移動平均、 σ_{rate} は導入期間である 25 期の株価変化率の標準偏差、 ϵ は標準正規乱数を表す。

Step.4 投資行動決定

各投資家は 2.1 節に従って投資行動を決定し、予想価格で注文を行う。

Step.5 注文量計算

全投資家の注文量を計算する。 t 期における投資家 i の売却量 $Vsell$ 及び購入量 $Vbuy_i(t)$ は次式で表される。

$$Vsell = A \times Vs(0) \quad (7)$$

$$Vbuy_i(t) = A \times \frac{Vm(0)}{Ex_i(t)}$$

ここで、 A は資産の投資割合を表すパラメータ、 $Vm(0)$ は投資家の初期保有現金、 $Vs(0)$ は投資家の初期保有株式を表す。

Step.6 約定

約定は板寄方式を採用し、取引は 1 期 1 回である。

3. モデルの評価方法と結果

本研究では計算機実験を行うにあたり、東京証券取引所 1 部に上場している銘柄を使用した。その際、大きな変動が見られる銘柄では、外部要因により変動している可能性があるため、株価変化率の絶対値

の平均が 0.01 未満の銘柄を抽出し、計算機実験を行った。実験に用いた銘柄は東証 1 部において食料品に分類される銘柄でブルボン(株), 山崎製パン(株), 日東富士製粉(株), イフジ産業(株), 東洋精糖(株), 森永乳業(株)である。そして各実験では乱数の種を変えて 30 試行ずつ行った。

本実験では 2014 年 1 月 8 日から 2014 年 2 月 13 日までの 25 期間の株価の終値データを取得し、25 期間の株価移動平均, 株価変化率の標準偏差及び移動平均を求め、2014 年 2 月 14 日から 2014 年 7 月 9 日までの 100 期間で株価の生成を行った。

また、計算機実験を行う際に、投資家の人数比を決定する必要がある。そこで、投資家の人数比に関しては現実の投資家の投資行動傾向を参考にして決定した[10]。図 4 は日経平均株価と個人投資家の売買動向に関してまとめたものである。

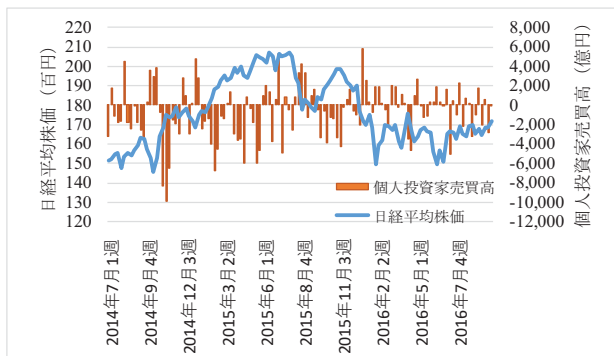


図 4 日経平均株価と個人投資家の売買動向

この図では個人投資家売買高が正であれば株式の買い越し, 負であれば売り越しを表す。個人投資家は日経平均株価が上昇している際に売り越し, 下落している際に買い越しをしており, 逆張りの投資行動をしていることが分かる。そこで, 本研究では逆張りの割合が順張りの割合に比べ, 多く設定する。そこで, 投資家 100 人のうちランダム投資家を引いた人数で順張り 4 割, 逆張り 6 割(ランダム 10:順張り 36:逆張り 54), 順張り 3 割, 逆張り 7 割(ランダム 10:順張り 27:逆張り 63), $T=100$, a を 2.0 刻みで[4.0, 8.0]の範囲, d を 5.0, 10.0, q を 15.0, 25.0, 投資家が感じる心理的インパクトの強さ b を 0.05 刻みで[1.0, 1.2]の範囲, A を 0.1, j_{buy} を 0.25 刻みで[-1.0, 1.0]の範囲, j_{sell} を 0.5, 1.0, g を 0.25 刻みで[0.5, 1.5]の範囲で実験を行った。

本研究では提案手法が実市場を再現できたかどうかを検証するために実市場と提案手法における株価変化率の自己相関, 株価変化率の 2 乗の自己相関, 株価変化率が平均 $\pm 1\sigma$ 内に収まっている割合, 株価変化率が平均 $\pm 3\sigma$ 外に外れている割合を用いた。そして, 各指標に対して現実の市場の値と, 本モデルにて計算した 30 試行の平均値を用いて平均の差の

検定(有意水準 5%)を行った。検定を行う際に, 株価変化率の自己相関及び株価変化率の 2 乗の自己相関ではラグ 1 の値のみを用いた。各統計量において 30 試行の平均値が実市場の統計量と最も差がなかった銘柄は東洋精糖, パラメータの組み合わせは $a:4.0, d:10.0, q:25.0, b:1.15, A:0.1, j_{buy}:0.0, j_{sell}:0.5, g:1.5$, 順張り 3 割, 逆張り 7 割である。

図 5 に実市場の株価変動, 上述のパラメータの組み合わせで行った計算の中の 1 試行である計算結果, 計算結果で収益により行動が変化しない場合の結果である計算結果-心理, 収益により行動が変化しない条件で行った計算の中の 1 試行である計算結果(心理なし)の結果を掲載する。また, 計算結果(心理なし)において 30 試行の平均値が実市場の統計量と最も差がなかったパラメータの組み合わせは $a:4.0, d:10.0, q:15.0, A:0.1, j_{buy}:0.0, j_{sell}:0.5, g:1.25$ である。また, -25 期から -1 期が導入期間, 0 期以降が計算期間である。

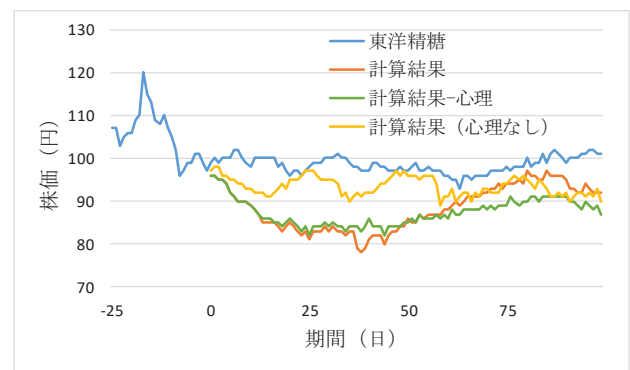


図 5 東洋精糖の株価変動

この図より, 計算結果と計算結果-心理において, 株価変動に大きな差がないことがわかる。計算結果(心理なし)では計算結果に比べて, 価格の下落が小さいことが分かる。

次に, 図 6 に株価変化率の自己相関, 図 7 に株価変化率の 2 乗の自己相関の結果を掲載する。また, 最良群は 30 試行の平均である。

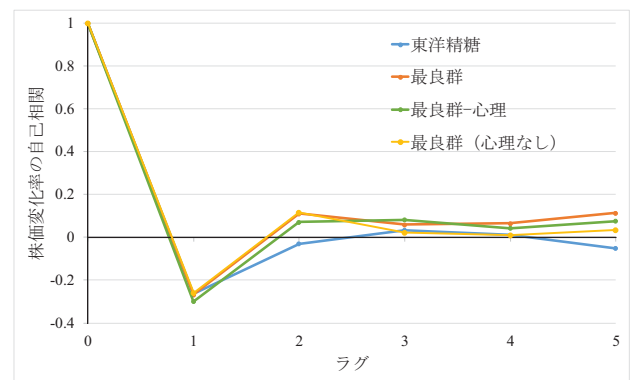


図 6 株価変化率の自己相関

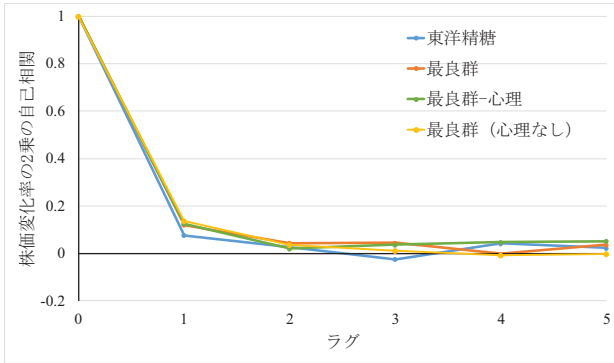


図 7 株価変化率の 2 乗の自己相関

次に、図 8 に株価変化率が平均±1σ内、平均±3σ外の割合となる割合の度数分布、表 1 に株価変化率が平均±1σ内、平均±3σ外の割合の値を掲載する。

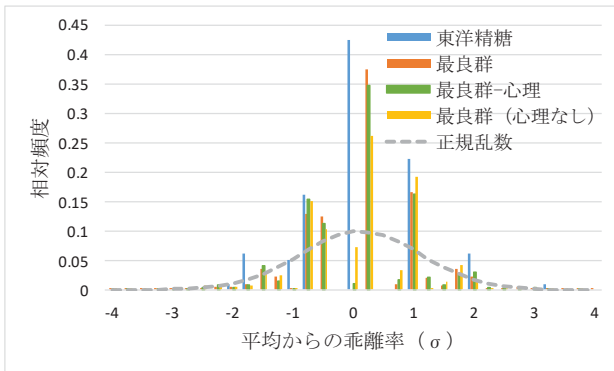


図 8 株価変化率が平均±1σ内、平均±3σの割合

表 1 株価変化率の変動幅の大きさの割合

	平均±1σ内	平均±3σ外
東洋精糖	85.9%	1.0%
最良群	80.4%	1.1%
最良群-心理	80.6%	0.7%
最良群 (心理なし)	81.5%	0.8%
正規乱数	68.0%	0.3%

図 6、図 7 より、最良群では、実市場と近い相関関係を確認することができた。図 8、表 1 より、最良群において、現実の市場の特徴である、平均±1σ内、平均±3σ外が正規乱数より高い頻度で観測されており、現実の市場の特徴を再現できていることを確認した。

次に、投資家の割合が異なる場合の株価変動を掲載する。図 9 に順張り 4 割、逆張り 6 割、図 10 に順張り 3 割、逆張り 7 割の際の株価変動の結果を掲載する。また、掲載する結果は実市場、乱数の種が異なる 5 試行の株価変動である。

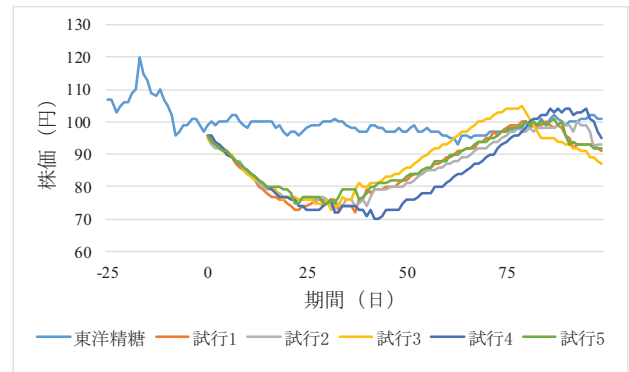


図 9 順張り 4 割、逆張り 6 割時の株価変動

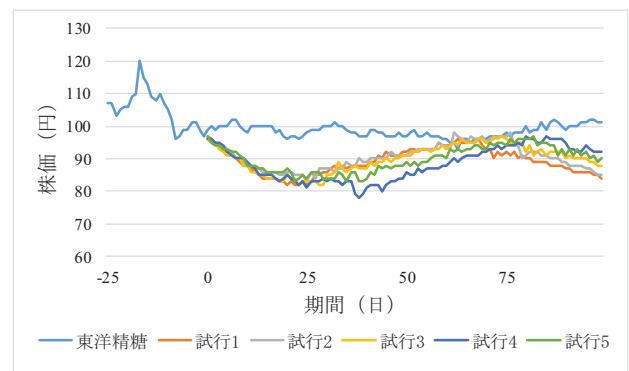


図 10 順張り 3 割、逆張り 7 割時の株価変動

図 9、図 10 より、順張り 4 割、逆張り 6 割時の株価変動の割合の方が、順張り 3 割、逆張り 7 割時の株価変動の割合に比べて株価が下落していることが確認できる。

次に、導入する銘柄の株価トレンドが異なる場合の株価変動の結果を掲載する。図 11 に導入期間が上昇トレンドであるブルボン、図 12 に導入期間が均衡気味である森永乳業、図 13 に導入期間が下落トレンドである東洋精糖の株価変動を掲載する。

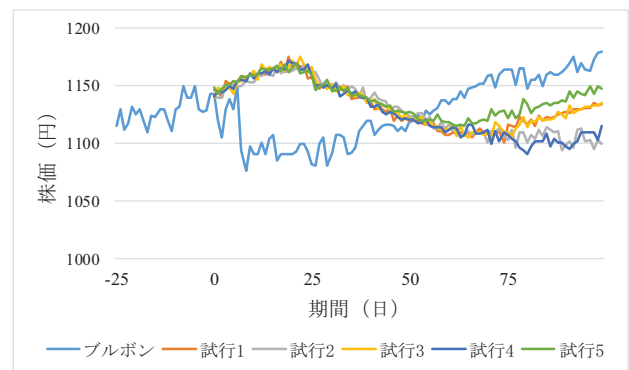


図 11 ブルボンの株価変動

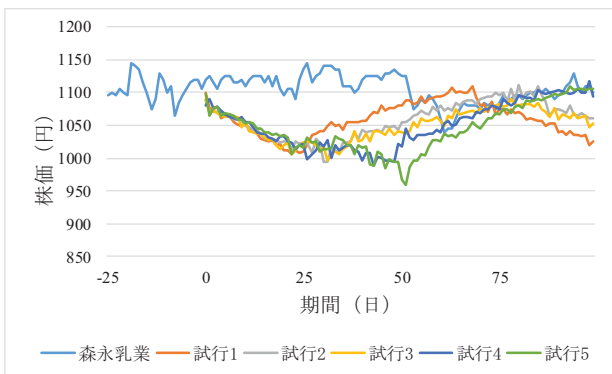


図 12 森永乳業の株価変動

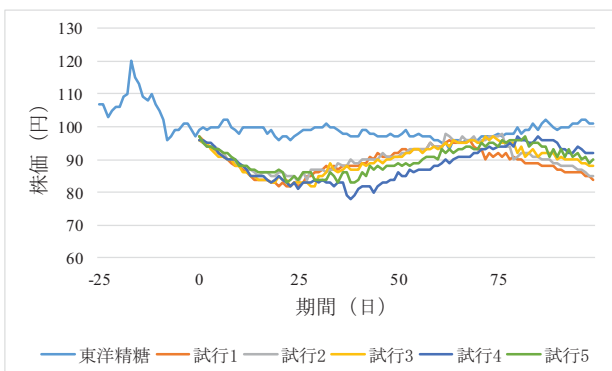


図 13 東洋精糖の株価変動

図 11 において、計算結果の株価変動は導入期間の上昇トレンドを継承し、各試行ともに株価の上昇を確認することができた。図 12 において、計算結果の株価変動は導入期間の均衡気味の株価変動を継承することができず、各試行ともに株価が下落していることを確認した。図 13 において、計算結果の株価変動は導入期間の下落トレンドを継承し、各試行ともに株価の下落を確認することができた。

4. 考察

順張り 3 割、逆張り 7 割時の株価変動は順張り 4 割、逆張り 6 割時の株価変動に比べて、株価の下落幅が小さかった。これは、下落トレンドが発生している際に、順張り投資家及びランダム投資家の売り注文を買い支える逆張り投資家が多かったためであると考えられる。この逆張り投資家の投資行動により、株価の下落幅が小さくなったと考えられる。

そして、導入期間の株価トレンドによっても株価変動に大きな差が出ることを確認した。導入期間が上昇トレンドであるブルボンでは、計算結果においても上昇トレンドが継承された。だが、導入期間の株価が均衡気味の森永乳業では株価の大きな下落がみられた。本モデルの順張り投資家は、過去の株価

をもとに株価トレンドを判断し、購入または売却を決定する。その際、順張り投資家は過去の株価に対してあてた回帰直線の傾きを基にトレンド判断を行うが、均衡気味の株価変動では正確に均衡気味と判断することができない。このような順張り投資家の行動により、森永乳業のように株価が均衡しているような銘柄では、均衡気味の株価変動を継承することができなくなったと考えられる。

そして、本研究で着目した心理的要因が株価に与える影響が小さいことが分かった。これは、本モデルの投資家がシステムティックに取引したことが影響していると考えられる。1 例として、実際の投資家は下落トレンド時に含み損が発生していると、損失を確定したくないがために、損切りをすることが出来ずに大損をしてしまうことがある。本モデルの投資家はこうした一時的な心理状態の変化により投資判断基準が大きく変わってしまうことがない。このような理由により、心理的要因が株価に与える影響が小さかったと考えられる。

5. 今後の課題

今後の課題として、順張り投資家のトレンド考慮方法の再検討が挙げられる。本モデルでは、トレンド考慮期間を一様乱数で決定している。だが、実際の投資家のトレンド考慮期間が一様分布に従っているとは考えにくい。そのため、投資家のトレンド考慮期間について再検討したい。

最後に、本モデルにて投資家の心理的要因が株価に与える影響を十分に確認することができなかった。だが、実際の市場において、損失を確定させることを嫌がるという心理が働き、損切りができない投資家が存在する。そのため、投資家の心理状態によって変化する投資行動、そして心理的要因の取り入れ方を再検討していきたい。

参考文献

- [1] 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)
- [2] 塩沢 由典, 中島 義裕, 松井 啓之, 小山 祐介, 谷口 和久, 橋本 文彦：知的エージェントで見る社会 3 人工市場で学ぶマーケットメカニズム-U-Mart 経済学編, 立出版(2006)
- [3] 喜多 一, 森 直樹, 小野 功, 佐藤 浩, 小山 祐介, 秋元 圭人：知的エージェントで見る社会 4 人工市場で学ぶマーケットメカニズム-U-Mart 工学編, 共立出版(2009)
- [4] 大垣 昌夫, 田中 沙織：行動経済学 伝統的

- 経済学との統合による新しい経済学を目指して, 有斐閣, (2001)
- [5] 角田 康夫: 行動ファイナンス 金融市場と投資家心理のパズル, 金融財政事情研究会, (2001)
- [6] 角田 康夫: 行動ファイナンスII 例題と用語集で読み解く非合理の謎, 金融財政事情研究会, (2004)
- [7] 並河 悠介, フェイ ザイ, カン シェン, 北栄輔: 行動ファイナンス理論に従うエージェントの市場取引の影響について, 情報処理学会論文誌 数理モデル化と応用 vol.48 No. SIG(TOM 17), p51-64, (2007)
- [8] Kahneman,D. and Tversky,A. :A prospect theory of decisions under risk, *Econometrica*, vol.47, p.263-297, (1979)
- [9] Tversky,A. and Kahneman,D. :Advances in prospect theory:Cumulative Representation of Uncertainty, *Journal of Risk and Uncertainty*, vol. 5, p297-323, (1992)
- [10] 日本証券所グループ : <http://www.jpx.co.jp/>.

Analysis of group behavior bias in financial markets using artificial market

Wang Yating¹, Toriumi Fujio¹

¹School of Engineering, The University of Tokyo

Abstract: In the past few decades, researchers have realized that human psychology can affect traders in decision-making process and finally affect the financial market based on behavior finance theory and cognitive psychology. Group behavior bias is one of them. Group. Some studies have been done on group behavior bias from behavior finance viewpoint to tell the differences of group behavior bias. However they only qualitatively analyzed macro phenomenon without quantitatively measure the detail micro thinking process in individuals. In this paper, we proposed and validated three types of group behavior bias models including majority following, winner following and hub following models based on different nationalities. We also compared these models and figured out that the majority following bias is the easiest to form in the market, however the market impact is the least. On the other hand, hub following bias is the hardest to emerge but the market impact is the most. Besides, we introduced short selling regulation as well as multi rate regulation and find both of them can lead heavier market impact in the market with group behavior biases.

1, Introduction

1.1 Research Background

In the past 20 years, researchers have realized that human psychology can affect traders in decision-making process and finally affect the financial market based on behavior finance theory and cognitive psychology. Group behavior bias is a kind of human psychologies. It refers to the psychology phenomenons that humans have a strong tendency to belong to a group and groups have influence on decision-making processes of individuals in a variety of ways, such as groupthink, deindividuation etc.

Recently, some researchers start to pay attention to group behavior bias because they think traders in the same group or the same market will share the same chart and the same channel even in the same dealer room should have more communication and affect each other in their decision-making process. They may be more likely to follow the groupthinking instead of their own thinking[1]. It is believed that group behavior bias varies in different groups, basing on different culture and different thinking mode in various countries.

However all previous analysis on group behavior biases are from behavior finance point of view. They are conducted from empirical literature and qualitative

aspect, which can't measure the detail micro thinking process in individuals and reveal the mechanism about how individual thinking accumulate to affect the market. Besides there is no detail model for group behavior bias.

1.2 Research Objectives

Based on the problems on group behavior bias related researches, we conduct this research with three objectives:

- 1) To propose the models of group behavior bias which is never discussed in previous studies.
- 2) To figure out the differences in these group behavior biases.
- 3) To reveal the mechanism about how group behavior biases influent the market and how the asset price changes when the group behavior bias is engaged by using artificial market.

2, Artificial Market Model

To quantitatively measure the group behavior bias and the market impact of financial market based on each individual and their interaction, we use artificial market, which is one kind of multi-agent based models to simulate financial market. We also apply the CNN

network model to describe the underlying structure of communication and relationship among traders in our artificial market model and then combined it with the mechanics behind the spread of interaction.

The whole artificial market model includes base model, which is created on the basic artificial market model, as well as group behavior bias models, which is based on agents' different decision-making processes, including the group behavior of majority following, winner following and hub following models.

2.1 Base model

The base model for usual agent of our simple artificial market is built on the basis of the model from Mizuta[2], in which agents follow the combination of fundamental strategy and technical strategy to decide whether to buy or sell.

$$r_{e,j}^t = \frac{1}{\sum_i w_{i,j}} (w_{1,j} \log \frac{p_f}{p^t} + w_{2,j} r_{k,j}^t + w_{3,j} \epsilon_j^t)$$

The first term stands for fundamental strategy, the second term stands for technical strategy and the last term is the noise item.

Learning process is also engaged in this model by comparing the evaluation term with each strategy term separately. If the fundamental term or the technical term is the same signs with evaluation term, the weight of this term will increase, otherwise will decrease.

2.2 Group behavior bias models

For group behavior bias models, we propose three different types on different decision-making processes in different groups. These three kinds of models are: majority following, winner following and hub following policy based group behavior bias. Majority following bias represents the bias to follow the decision made by majority people in the group. This stands for thinking process in Japan. Winner following bias represents the bias to follow the decision made by the most profitable trader in the group. This kind of thinking mode matches the way of some result-driving countries such as America. Hub following bias represents the bias to follow the decision made by the authority trader in the group which is an important trading rule for traders in China.

Agents which employ majority following policy is to follow the most adopted decision by surroundings, like herding behavior. Traditional herding is defined as a switch in traders' decision into the direction of the crowd,

however in our majority following model, agents make the decision based on both self-signal and others' signal. Agent will set a threshold and compares it to the result percentage of majority, then decide whether to follow the majority or insist his own decision. The agent observes the trading decision of neighbors within one step in the network, then calculate the buy ratio and sell ratio (Details in [4]).

For winner following model, we consider the agent with the most net assets within one step distance as "winner". In winner following model, the agents will copy the decision made by winner agent. The agent observes the net assets of neighbors (including himself) within one step in the network and find the most profit agent, which we call it "winner agent" here. If he earns the most, he will follow previous step to make the decision. The equation to calculate net asset is

$$V_i^t = \sum_{j=1}^k P_j^t \times S_{i,j}^t + C_i^t$$

where P_j^t is the market price of the asset. $S_{i,j}^t$ is the number of share for each asset, C_i^t means the amount of cash.

Hub following policy is a policy to follow the decision made by the agent who has the most resources or the most ways to gather information. We call this kind of agent the "hub" agent. In the network, such kind of agent can be marked as the node with most links. The decision made by hub agent will be imitated by other agents.

3, Market Settings

In this part, we divide market settings into three parts: regulation settings, evaluation values and parameter settings. We also extend single risky asset market to multi-assets markets.

3.1 Regulation settings

We import short selling regulation and multi rate regulation to see how these regulation work when group behavior bias agents exist.

Short selling regulation stipulates the holding number of the asset that should not be less than one share when making a sell order, and the rest cash should not be less than the price of the asset when making a buy order.

The multi rate regulation means that each asset has a maximum threshold in the net asset of the agent who owns it. The equation to calculate threshold is

$$\frac{|P_j^t \times S_{i,j}^t|}{V_i^t} \leq \theta$$

P_j^t is market price at time t of asset j , $S_{i,j}^t$ is the number of share hold by each agent. V_i^t means the total net asset.

Agent will check whether the holding asset ratio exceeds the maximum ratio before making order. When the rate of assets larger than maximum ratio, the order will be cancelled. Otherwise the order will be proceeded. If one asset ratio already exceeds the threshold θ due to abrupt price change, the following order will try to adjust the ratio to a reasonable range by selling the exceeding asset or buying the other asset until the ratio rebalances.

3.2 Evaluation Values

We introduce three features to explain how the markets are effected by group behavior biases. These features are group behavior bias coefficient, market impact and critical point.

3.2.1 Group Behavior Bias Coefficient

To measure group behavior bias phenomenon level in the market, we introduce group behavior bias coefficient. The Group behavior coefficient is an average measure of group behavior in the market during the whole simulation period T , it represents the synchronized level of decisions made by group agents in the artificial market. We calculate the group behavior bias phenomenon on the base of the coefficient described in [3]

$$\sigma = \frac{2}{N} \sqrt{\left(\frac{1}{T} \sum_{t=1}^T (N_{bt} - \frac{N}{2})^2\right)}, \sigma \in (0, 1)$$

Here N means the total number of agents, and N_{bt} is the number of agents who decide to buy one unit at time t . T is the total simulation period. The bigger the value of σ is, the larger group behavior bias. If it becomes close to 1, it means that the action of the agents are synchronized to one direction. If it becomes close to 0, it means there is no group behavior bias phenomenon and the market is in a balance status.

3.2.2 Market Impact

we follow [2] and use the value of market impact to measure market efficiency. Market impact is defined to analyze how much the current price deviates the fundamental price of the risk asset. By calculating the value of it, we can check the efficiency of the market.

$$MI = \frac{1}{n_a} \sum_{j=1}^{n_a} \frac{P_a^j - P_f}{P_f}$$

Here n_a means the number of group behavior bias agents in a simulation period. P_a^j is traded price for each group behavior bias agent. The larger value of MI is, the heavier the market impact caused by group behavior bias agents.

3.2.3 Critical Point

Critical point is the maximum ratio of group behavior bias agents exist in the market to ensure the market efficiency. We only compare previous two evaluation values with the ratio of group behavior bias agents ranges from 5% to the critical point in the following experiments.

To figure out the limit action to keep efficient market, we carry on experiments and see how market impact changes. We define the maximum value of group behavior bias ratio which keeps the market efficient as the critical point r_{max} . When group behavior bias ratio is lower than it, the market is still efficient. We follow below steps to calculate the critical point:

- 1) To find the convergent time of each simulation run by locating the time where the price variations afterward are all lower than a threshold according to empirically set.
- 2) Mark the maximum value of group behavior bias ratio where 70% of simulation runs can finally get convergent as critical point.
- 3) For multi-assets market, the smallest critical point value among all the assets is considered as the critical point.

3.3 Parameter Settings

	Values
Agent Number	$N = 1000$
Maximum Size of Time	$T_{max} = 2000$
Fundamental Price of Single Market	$P_f = 102$
Fundamental Price of Multi Market	
Asset 1	$P_1 = 102$
Asset 2	$P_2 = 102$
Fundamentall Price after Jump	$P_j = 80$
Jump Time	$T_j = 600$
Tick Size	$d_p = 0.1$
Threshold of Most Trading Policy	$\Theta = 0.7$
Maximum Ratio of Ownership Rate Regulation	$\theta = 0.5$
Initial Share	100
Initial Cash	4000
Cancel Time	$T_c = 10000$
Probability in CNN model	$P = 0.75$

4, Simulation Results

4.1 Verification of the artificial market

In this part, we conduct experiment to confirm whether the models we built are valid or not. We calculate kurtosis, volatility and autocorrelation coefficients of squared return of the artificial market. As a result, the values of kurtosis and autocorrelation coefficients of squared return in all the markets are positive and are similar with actual markets. Meanwhile, autocorrelation coefficients of squared return decays as the lag interval goes by which is also like the actual market. Hence we reproduce the stylized facts including fat-tail and volatility clustering with different models. Therefore, the artificial markets we built with group behavior bias models are valid and the models replicate the characteristics of short term micro structure in real financial markets.

4.2 Experiment Results

We took 30 times simulation runs for each situation. The following result is the average of 30 simulation which runs for different markets on different random seeds.

4.2.1 Group Behavior Bias and Market Efficiency

We observe the price flow changes of the market, when with and without group behavior biases. We assume there is an abruptly jump on fundamental price of risky asset in the market. This experiment helps us to figure out whether group behavior biases can make the market inefficient or not.

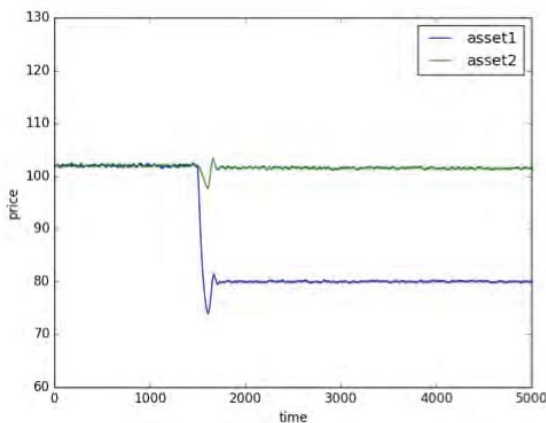


Fig.1(a) Price change figure of multi-assets market without group behavior bias agents

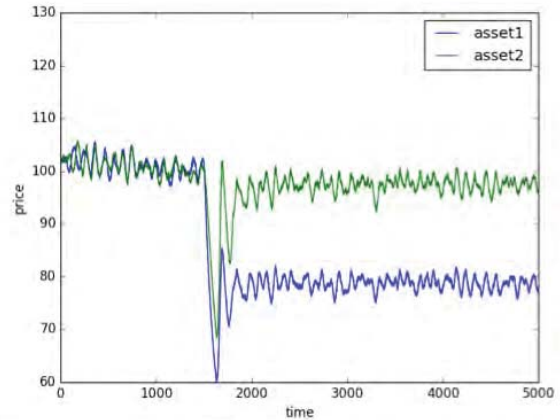


Fig.1(b) Price change figure of multi-assets market with group behavior bias agents

Fig.1 (a) and (b) show the price changes of assets without and with group behavior biases during a simulation run. X-axis is simulation time period. Y-axis is the market price of the asset. According to figures, we can see the price changes of asset will change more when group behavior bias exists.

4.2.2 Differences of group behavior bias models

We adjust the group behavior bias ratio in the market and compare the group behavior bias coefficient, market impact of different group behavior bias models to figure out the how each group behavior bias influences the market.

● Results of Group Behavior Bias Coefficient

We compare the group behavior bias coefficient of different group behavior bias models in the same market to figure out which kind of group behavior bias is more easy to form in the market.

Fig.2 shows group behavior bias coefficient of three kinds of group behavior bias models in the market basing on the CNN network. X-axis is bias agent ratio and y-axis is group behavior bias coefficient

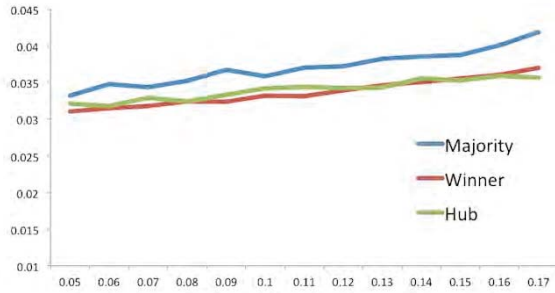


Fig.2 Group behavior bias coefficient of three kinds of group behavior bias model in single asset market without regulation.

● Group Behavior Bias and Market Efficiency

Fig.3 expresses the market impact with different group behavior bias models in the market. X-axis is the bias agent ratio in the market. Y-axis is the value of market impact. We only take the results within critical point to the validation of the results.

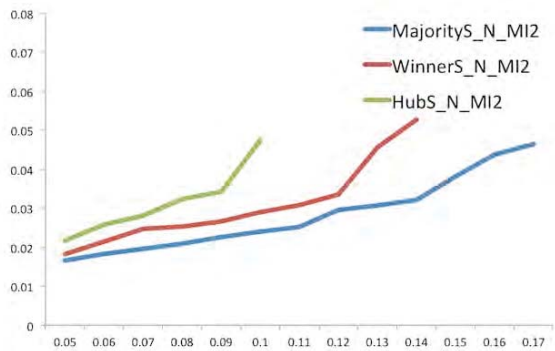


Fig.3 Market Impact by three group behavior bias models in single market without regulation

we can see that the group behavior bias coefficient with majority following model is the highest and lowest with hub following model. On the other hand, the market impact of hub following model is largest and majority is the smallest. We conclude that the majority following bias is the easiest to form in the market, however the market impact is the least. On the other hand, hub following bias is the hardest to emerge but the market impact is the most.

4.2.3 Regulation results

We compare the market impact of group behavior bias models when the market with or without different regulations respectively (Table I and Table II). Fundamental price jumps abruptly during the simulation run. In this way, we make clear how different regulations

work with group behavior bias models.

TABLE I

Market impact in different single asset markets with 10% hub following bias agent

	With short selling regulation	Without any regulations
Single Market	0.066	0.043

TABLE II

Market impact in different multi-assets markets with 10% hub following bias agent

	With regulations	With short selling regulation	With multi rate regulation	Without any regulations
Asset 1	0.049	0.065	0.039	0.034
Asset 2	0.068	0.024	0.063	0.023

The group behavior biases show heavier market impact with short selling regulation exists and the market in which short selling is allowed is more stable. Hence we suggest that it will be better not to introduce short selling limitation in the market when group behavior biases exist to ensure the stability of financial market.

The multi rate regulation is the main factor to increase impact for non-risk asset. This is because agents will not take any action on non-risk asset to make sure the limitation number of risk asset within the threshold. In all the situations, the market impact will increase when multi rate regulation is engaged. Therefore, it is better not to employ multi-rate regulation to keep market efficiency. At last, no matter when the regulations exist or not, the market impact for hub following model is the largest and least for majority following model.

4.2 Conclusion

In this paper, we propose three different types of group behavior bias models based on different decision-making processes in various groups. We integrate these group behavior bias models to artificial markets and conduct experiments to figure out the differences among them. Besides, we reveal how the decision made by individuals get accumulated with group behavior biases and get reflect on the different markets.

We use KPI of group behavior bias coefficient, market impact to compare the differences of these group behavior bias models. We notice that it is the majority following bias is the easiest to form in the market, however the market impact is the least. On the other hand, hub following bias is the hardest to emerge but the

market impact is the most. This matches the results of empirical analysis about financial markets which state financial market in China is more dramatic[5] and Japanese financial market is comparatively stable[6]. Therefore our models can reproduce the characteristics of group behavior biases in different countries.

we also figure out that group behavior biases show heavier market impact with short selling regulation exists. Hence we suggest that it will be better not to introduce short selling limitation in the market when group behavior biases exist to ensure the stability of financial market. The multi rate regulation is the main factor to increase impact for non-risk asset. This is because agents will not take any action on non-risk asset to make sure the limitation number of risk asset within the threshold. In all the situations, the market impact will increase when multi rate regulation is engaged. Therefore, we do not suggest market with multi rate regulation for market efficiency. At last, no matter when the regulations exist or not, the market impact for hub following model is the largest and least for majority following

References

- [1] Masayuki Susai and Hiroshi Moriyasu. Consistency of risk attitude and other investment behavior of japanese professional fund managers: Questionnaire- based analysis. Waseda University Institute of Finance, Working Paper Series, pages 1–13, 2007.
- [2] Takanobu Mizuta, Kiyoshi Izumi, and Shinobu Yoshimura. Price variation limits and financial market bubbles: Artificial market simulations with agents' learning process. In 2013 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), pages 1–7. IEEE, 2013.
- [3] Wang Yating, Toriumi Fujio. Market impact analysis on group behavior by artificial market. SIG-FIN-016-04, 2016
- [4] Abhijit V Banerjee. A simple model of herd behavior. The Quarterly Journal of Economics, pages 797–817, 1992.
- [5] Shujie Yao and Dan Luo. The economic psychology of stock market bubbles in china. The World Economy, 32(5):667–691, 2009.
- [6] 高野陽太郎 and 櫻坂英子. 日本人の集団主義 ”と “ アメリカ人の個人主義. 心理学研究, 68(4):312–327, 1997.

深層学習と拡張手がかり表現による業績要因文への極性付与 Assigning Polarity to Causal Information by Deep Learning and Extended Clue Expressions

酒井浩之^{1*} 坂地泰紀¹ 山内浩嗣^{2†} 町田亮介^{2†} 阿部一也^{2†}
Hiroyuki Sakai¹ Hiroki Sakaji¹ Hiroaki Yamauchi² Ryosuke Machida² Kazuya Abe²

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 三菱UFJトラスト投資工学研究所

² Mitsubishi UFJ Trust Investment Technology Institute

Abstract: In this paper, we propose a method of assigning polarity to causal information extracted from summary of financial statements of companies. Our method assigns polarity (positive or negative) to causal information in accordance with business performance, e.g. “Orders of semiconductor manufacturing equipments were strong”. First, we assign polarity to extended clue expressions to be used to extract causal information. Using them, our method automatically generates training data and assigns polarity to causal information by deep learning. We evaluated our method and confirmed that it attained 86.7% precision and 95.4% recall of assigning polarity positive, and 90.0% precision and 73.9% recall of assigning polarity negative, respectively.

1 はじめに

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

投資家にとって、企業の業績に関する情報を収集することは重要であるが、実際の業績に関する情報だけでなく、その業績要因も重要である [4][5]。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。しかしながら、証券市場の上場企業数は東京証券取引所の上場企業だ

けでも 2017 年 2 月 17 日現在、3,537 社と多いうえに¹、近年では年に 4 回、決算発表がある。さらに、大幅な業績の修正を行う場合にも業績修正発表を行う必要があるため、人手によって多くの企業の業績要因を取得するには多大な労力を要する。そのために酒井らは、企業が業績発表を行った直後に企業の Web サイト等で公開される決算短信 PDF に着目し、その中から業績要因を含む文（例えば「半導体製造装置の受注が好調でした。」）を抽出する手法を提案した [6]。

決算短信より抽出された業績要因を含む文は、例えば、証券アナリストへの支援材料として利用できる。（以降、業績要因を含む文を「業績要因文」と定義する。）業績要因文の応用例として、抽出された業績要因文を検索対象とし、例えば「太陽電池」で検索することで、「太陽電池」を含む業績要因文（例：「太陽電池は、国内の住宅用及びメガソーラーなど産業用の販売が伸長したことから、売上高は前年同期比 85.8% 増の 2,768 億円となりました。」）を検索することができ、そのような業績要因文が頻繁に出現する企業を検索することができる。本例の「太陽電池」では、「東洋炭素」、「フジプレアム」、「シャープ」、「カネカ」といった企業が検索される。しかし、より有効な情報として利用するためには、抽出した業績要因文に対して

*連絡先：成蹊大学 理工学部 情報科学科
〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp

†本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。

¹<http://www.jpx.co.jp/listing/stocks/co/index.html>

業績に関する極性（「ポジティブ」、「ネガティブ」）を付与する必要がある。例えば、業績要因文「半導体製造装置の受注が好調でした。」に対しては「ポジティブ」、「世界的な太陽電池市況の低迷により太陽電池製造装置の販売が減少しました。」に対しては「ネガティブ」の極性を付与する。業績要因文に対して極性を付与することで、業績要因を使用した景気動向予測、および、業績要因に基づいて株取り引きを行うコンピュータトレーディングにも応用することが期待できる。例えば、「太陽電池」で検索された業績要因文に極性が付与されていることで、太陽電池の市況を業績要因文に付与された極性の数で得ることができ、また、その決算短信PDFの発表日と組み合わせることで、太陽電池の市況の時系列推移を得ることができる。そのため、本稿では、決算短信PDFより抽出された業績要因文に対して極性（「ポジティブ」、「ネガティブ」）を自動的に付与する手法を提案する。具体的には、業績要因文を抽出するための手がかりとなる表現（以降、手がかり表現と定義）にいくつかの文節を追加することで、“拡張手がかり表現”と定義する文節列（例えば「受注が好調でした」）を獲得する。そして、“拡張手がかり表現”を使用することで極性付与のための学習データを自動的に生成し、生成された学習データを使用して深層学習を行い、業績要因文に極性を付与する。

2 決算短信から抽出した業績要因文への極性付与

まずは抽出した手がかり表現²に極性（ポジティブ、ネガティブ）を手で付与し、その手がかり表現の極性を使用して業績要因文への極性付与を行うことを試みる。しかしながら、例えば「推移しました」のような手がかり表現には極性を付与できない。なぜなら「堅調に推移しました」はポジティブとなるが、「厳しい状況で推移しました」はネガティブとなるからである。このように、手がかり表現に係る文節列（上記の例では「堅調に」「厳しい状況で」）によって極性が変わる手がかり表現が存在する。また、「増加」「増え」「減少」を含む手がかり表現は、その手がかり表現に係っている文節列によって極性が反転する場合がある。例えば、手がかり表現の「増加しました」はポジティブの極性を付与するが、「費用が」や「コストが」が係っている場合はネガティブの極性を付与する必要がある。逆に、手がかり表現の「減少しました」はネガティブの極性を付与するが、「費用が」や「コストが」が係っている場合はポジティブの極性を付与する必要がある。このように、極性を付与できない手がかり表現を含む業績

²172,951個の決算短信PDFから酒井らの手法[6]により163種類の手がかり表現を抽出した。

表 1: 極性付与された手がかり表現の例

手がかり表現	極性
堅調でした	positive
伸び悩んだ	negative
寄与し	positive
低調でした	negative
低迷し	negative
改善し	positive

要因文には極性を付与できず、さらに、極性を付与した手がかり表現のみでは業績要因文の極性を決定することができない場合がある。そのため、あらかじめ極性を付与した業績要因文を大量に作成し、それを学習データとした深層学習により、業績要因文に極性を付与する。しかしながら、あらかじめ極性を付与した業績要因文を手で大量に用意することは、多大な労力を要する。そこで本研究では、業績要因文への極性を付与するための学習データを自動的に生成する。

まず、抽出した手がかり表現のなかから、極性（ポジティブ、ネガティブ）を付与できる手がかり表現に対して手で極性を付与する。その結果、163種類の手がかり表現のうち、108種類の手がかり表現に極性を付与した。表1に、手で極性付与した手がかり表現の一部を示す。しかし、極性付与した手がかり表現を含む業績要因文を学習データとして全て使うことは適切ではない。なぜなら、業績要因文の抽出精度が83%程度であり、誤った業績要因文も多く含まれているからである。そこで、極性が付与できる手がかり表現にいくつかの文節を追加することで、“拡張手がかり表現”と定義する文節列（例えば「受注が好調でした」）を獲得する。これにより、例えば手がかり表現「好調でした」が「受注が好調でした」や「極めて好調でした」のような、より精度の高い手がかり表現へ拡張される。そして、例えば拡張手がかり表現「極めて好調でした」を含む業績要因文をポジティブ、拡張手がかり表現「極度の不振で」を含む業績要因文をネガティブとすることで学習データを自動的に生成する。

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から抽出した業績要因文から、極性を付与できる手がかり表現の“拡張手がかり表現”を獲得する

Step 2: ポジティブの拡張手がかり表現を含む業績要因文をポジティブ、ネガティブの拡張手がかり表現を含む業績要因文をネガティブとして、学習データを自動生成する。

Step 3: 拡張手がかり表現を含まない業績要因文に対して、Step 2で生成された学習データを使用し、

深層学習にて極性を付与する。

2.1 拡張手がかり表現の獲得

極性を付与した手がかり表現に文節列を追加して、拡張手がかり表現を獲得する手法について述べる。具体的には、手がかり表現 c に係る文節列 p に対して以下の式 1 でスコアを求め、このスコアがある閾値を上回る文節列を抽出する。

$$\begin{aligned} \text{Score}(p, c) &= -f(p, c)\sqrt{fp(p)}\log_2 P(p, c)H(p) \\ P(p, c) &= \frac{f(p, c)}{N(c)} \end{aligned} \quad (2)$$

ただし、決算短信 PDF から取得した業績要因文の集合において、

$P(p, c)$: 手がかり表現 c から取得される文節列 p の出現確率

$f(p, c)$: 手がかり表現 c から取得される文節列 p の取得回数

$N(c)$: 手がかり表現 c から取得される文節列の総数

$fp(p)$: 文節列 p に含まれる文節の数,

$H(p)$: 文節列 p がある企業の業績要因文に出現する確率に基づくエントロピー (後述)。

例えば、決算短信 A に「収益性が大幅に改善しました」という業績要因文が存在していたとすれば、手がかり表現「改善し」に係る文節列として「大幅に」、「収益性が大幅に」という 2 つの文節列を取得する。また、決算短信 B に「業績が大幅に改善しました」という文が存在していたとすれば、この文から「大幅に」、「業績が大幅に」という 2 つの文節列を取得する。そして、決算短信 A と決算短信 B からは「大幅に」が 2 回、「収益性が大幅に」、「業績が大幅に」をそれぞれ 1 回、取得したことになる。そのため、例えば「大幅に」の $f(p, c)$ の値は 2 であり、 c から取得する文節列の総数 $N(c)$ は 4 であるため、 $P(p, c)$ の値は $2/4$ となる。

$H(p)$ は文節列 p がある企業の業績要因文に出現する確率に基づくエントロピーであり、 $H(p)$ が高い文節列は多くの企業の業績要因文に出現している文節列であることが分かる。また、1 つの企業にのみ多く出現する文節列の $H(p)$ は 0 になるため、そのような文節列を除去できる。 $H(p)$ は以下の式 3 で求める。

$$H(p) = - \sum_{s \in S(p)} P(p, s) \log_2 P(p, s) \quad (3)$$

ここで、 $S(p)$ は文節列 p を含む業績要因文をもつ企業の集合、 $P(p, s)$ は文節列 p が企業 s の業績要因文に出現する確率を表す。

手がかり表現 c における $\text{Score}(p, c)$ の平均を求め、平均値より大きいスコアが付与された文節列 p を手がかり表現 c に追加し、それを拡張手がかり表現として獲得する。表 2 に 172,950 の決算短信 PDF から抽出された業績要因文を対象に、手がかり表現「改善し」から取得された文節列の中で $\text{Score}(p, c)$ が高い文節列をいくつか示す。表 2 に示すように、手がかり表現に追

表 2: 「改善し」から取得された文節列

文節列	$\text{Score}(p, c)$	拡張手がかり表現
大幅に	18,543	大幅に改善し
大きく	10,157	大きく改善し
利益率が	5,438	利益率が改善し
収益が	5,278	収益が改善し
採算が	3,066	採算が改善し

加する文節列は 1 文節であることが多いが、拡張手がかり表現が「企業の景況感が改善する」「当初の計画を上回って」「売上高の増加に貢献した」のように、複数の文節で構成される文節列が追加される場合もある。

2.2 拡張手がかり表現の極性

拡張手がかり表現の極性は、拡張前の手がかり表現の極性に準じる。例えば拡張手がかり表現「大幅に改善し」の場合、拡張前の手がかり表現「改善し」に付与された極性がポジティブであるため、ポジティブとなる。ただし、「増加」、「増え」、「減少」を含む手がかり表現は、拡張手がかり表現によっては極性が反転する場合がある。例えば手がかり表現「増加する」はポジティブの極性が付与されており、その拡張手がかり表現「売上高が増加する」はポジティブであるが、「人件費が増加する」はネガティブの極性を付与する必要がある。そのため、「増加」、「増え」、「減少」を含む手がかり表現に以下に示す語を含む文節列が追加されている場合、手がかり表現に付与されている極性の逆の極性を、拡張手がかり表現に付与する。

費, 支出, 借入金, 債権, 未払金, 債務, 負債, 損失, コスト, 貸出金

極性を付与した 108 種類の手がかり表現から、4,318 個のポジティブの拡張手がかり表現、1,898 個のネガティブの拡張手がかり表現を獲得した。

2.3 学習データの自動生成

拡張手がかり表現とその極性を使用して、学習データを自動生成する。ここで、業績要因文には例として以下に示すように複数の単文で構成される複文が含まれているとする。

粉ミルクが震災直後のまとめ買いの反動などによる需要減少により前年同期を下回りましたが、流動食は生産・供給能力回復によりほぼ前年同期並みとなりました

このような業績要因文の場合、接続助詞「が」で分割し、「粉ミルクが震災直後のまとめ買いの反動などによる需要減少により前年同期を下回りました」と「流動食は生産・供給能力回復によりほぼ前年同期並みとなりました」のそれぞれに対して極性を付与する必要があります。また、学習データもそれに準拠して生成する必要があります。そのため、業績要因文の文中に接続助詞「が」、および、接続助詞「ものの」が出現したら、そこで業績要因文を区切る。以降、接続助詞によって区切られた文も含めて業績要因文とする。

業績要因文に拡張手がかり表現が含まれていれば、その拡張手がかり表現の極性を業績要因文に付与して学習データとする。以下に自動生成された学習データの一部を示す。+1が正例（ポジティブ）、-1が負例（ネガティブ）であることを示す。また、下線は含まれていた拡張手がかり表現を示す。

+1 血液凝固阻止剤などの販売量が大幅に増加しました
+1 イーファイナンス事業の売上高の増加も大きく貢献しました
-1 健康機能性飲料が前年同期の売上げを大きく下回ったこと等により減収となりました

後述の評価では、340社を無作為に選択し、その業績要因文から47,953文の学習データを自動生成した。

3 深層学習による業績要因文への極性付与

3.1 素性選択

自動生成された学習データに基づき、深層学習により業績要因文へ極性を付与する。まず、入力層の要素となる語（素性）を選択する。具体的には、自動生成された学習データにおいて正例の業績要因文に含まれる内容語（名詞、動詞、形容詞）に対して、以下の式4にて重みを計算する。

$$W_p(t, S_p) = TF(t, S_p)H(t, S_p) \quad (4)$$

ただし、

S_p : 学習データにおいて正例に属する業績要因文の集合

$TF(t, S_p)$: 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 における各業績要因文に含まれる語 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語であることが分かる。 $H(t, S_p)$ は次の式5で求める。

$$H(t, S_p) = - \sum_{s \in S_p} P(t, s) \log_2 P(t, s) \quad (5)$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \quad (6)$$

ここで、 $P(t, s)$ は業績要因文 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、負例の業績要因文に含まれる内容語（名詞、動詞、形容詞）に対しても、同様に重みを計算する。

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

ただし、 S_n は学習データにおいて負例に属する業績要因文の集合である。

ここで、ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ の2倍より大きければ、その語 t を素性として選択する。もしくは、語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の2倍より大きければ、その語 t を素性として選択する。すなわち、以下の条件のどちらかが成り立つ語 t を素性として選択する。

$$W_p(t, S_p) > 2W_n(t, S_n) \quad (8)$$

$$W_n(t, S_n) > 2W_p(t, S_p) \quad (9)$$

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。以下に選択された素性の一部を例示する。

増加, 減少, 好調, 推移, 拡大, 利益, 影響, 大幅, 増収, 伸長, 順調, 減収

後述の評価における47,953文の学習データでは、5,920語が素性として選択された。

3.2 モデル

入力は、47,953文の学習データから抽出された5,920語を要素、語 t における $\log(W_p(t, S_p))$ 、もしくは、 $\log(W_n(t, S_n))$ 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5,920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。

表 3: 業績要因文への極性付与の評価結果 (本手法)

	positive	negative
精度 (%)	86.7 (210/242)	90.0 (91/101)
再現率 (%)	95.4 (210/220)	73.9 (91/123)

4 評価

本手法の評価を行うため、本手法を実装した。実装にあたり、形態素解析器として MeCab³、係り受け解析器として CaboCha[3] を使用した。ここで、340 社を無作為に選択し、その業績要因文から 47,953 文の学習データを自動生成した。テストデータは、学習データには含まれない (すなわち拡張手がかり表現を含まない) 業績要因文から無作為に 343 文の業績要因文を抽出して作成した。テストデータにおいて本手法によりポジティブの極性が付与された業績要因文を以下にいくつか示す。

- ・豪州の鉄鉱石事業が堅調に推移し、トレード関連では原油・石油製品取引及び鉄鋼製品取引も堅調に推移しました
- ・検索連動型広告においては、スマートフォン経由の売上が大きく伸び、売上が前年同四半期比で増加しました

また、テストデータにおいて本手法によりネガティブの極性が付与された業績要因文を以下にいくつか示す。

- ・天候不順の影響などにより、衣料品分野の売上は厳しいまま推移いたしました
- ・売上原価の増加積極的な研究開発活動を行い、研究費が増加したことなどにより、営業利益 583 億 47 百万円 (前年同期比 26.3% 減)、経常利益 609 億 22 百万円 (前年同期比 25.8% 減) となりました

テストデータの 343 文の業績要因文に対して人手にて極性を付与して正解データを作成し、本手法の精度、再現率を求めた。結果を表 3 に示す。比較手法として、人手によって極性付与された手がかり表現のみを使用し、業績要因文に含まれている手がかり表現の極性に従って、業績要因文の極性を付与する手法の精度、再現率を表 4 に示す。なお、もし「推移する」のように極性を付与できない手がかり表現のみが含まれていた場合は、極性を付与しない。

³<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

表 4: 業績要因文への極性付与の評価結果 (比較手法)

	positive	negative
精度 (%)	93.0 (120/129)	89.8 (62/69)
再現率 (%)	54.5 (120/220)	50.4 (62/123)

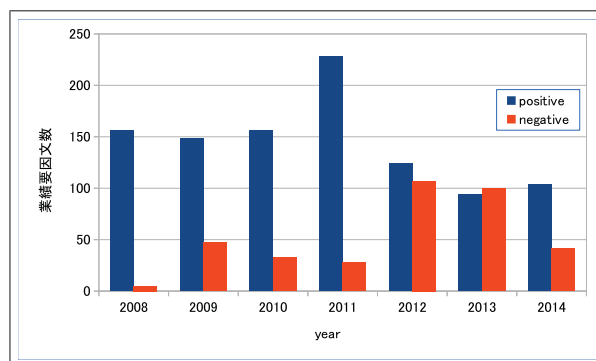


図 1: 「太陽電池」を含む業績要因数の推移

5 考察

本手法による業績要因文への極性付与手法の精度はポジティブで 86.7%、ネガティブで 90.0% であり、高い精度を達成した。また、再現率はポジティブで 95.4%、ネガティブで 73.9% であり、比較的良好な結果を得ることができた。一方で、人手によって極性付与された手がかり表現のみを使用する比較手法では再現率がポジティブで 54.5%、ネガティブで 50.4% であることから、本手法は比較手法と比べて、それほど精度を落とすことなく再現率が大きく向上していることが分かる。例えば、本手法では以下の業績要因文に対してポジティブの極性を付与することができたが、比較手法では極性を付与することができなかった。

セラミックス事業は、アジアを中心としたガラスエンジニアリング分野と国内の環境エネルギー分野で需要が旺盛であったため、概ね堅調に推移しました

本手法は、上記のような「推移しました」のような極性を付与できない手がかり表現を含む業績要因文に対しても極性を付与することができるため、比較手法に比べて再現率を向上することができた。

決算短信には発表年月日が記載されているため、あるキーワードを含む業績要因文が記述されている決算短信の発表日を得ることで、そのキーワードを含むポジティブな業績要因文数、ネガティブな業績要因文数の時系列推移を得ることができる。例として、図 1 に「太陽電池」を含む業績要因文数の時系列推移を示す。図 1 より、2011 年において「太陽電池」を含むポジティブ

な業績要因文数は最も多いが、2012年になると大幅に減少し、逆にネガティブな業績要因文数が大幅に増えていることが分かる。2012年の「太陽電池」を含むネガティブな業績要因文には、以下のような記述がある。

多結晶シリコンは、太陽電池パネルの供給過剰とパソコンの販売不振等を背景にした半導体ウエハーの在庫調整に伴う、販売数量の減少及び販売価格の下落により大幅な減収減益となりました

以上のように、極性付与した業績要因文は、企業における製品や事業の現状の把握から、製品や事業の動向の時系列推移とその要因など、様々な適用が可能である。

6 関連研究

酒井らは、経済新聞記事の企業の業績発表記事から抽出した業績要因文に対して極性を自動的に付与する手法を提案している [5]。酒井らの手法は、業績発表記事に対してSVMで高精度で極性付与できることが前提となっており、酒井らの手法をそのまま決算短信に適用するのであれば、決算短信に対して高精度で極性を付与するための、大量の学習用データを人手で用意する必要がある。高村らは2つの単語から成る表現に対して隠れ変数モデルを用い、機械学習を用いて構成語の属性をクラスタという形で抽出して確率モデルを構築し、複数語表現の感情極性を分類する手法を提案している [7]。それに対して、本手法では業績要因文を抽出するために有効な手がかり表現（「が好調でした」等）に極性を付与し、その手がかり表現を拡張した“拡張手がかり表現”を使用することで学習データを自動生成し、深層学習により業績要因文への極性付与を行った。

Kajiらは、評価文に含まれる評価表現（名詞＋格助詞＋形容詞）に対して極性を決定する際に、人手で作成した手がかり表現リストやパターン、規則を使用して評価文を抽出して極性を付与しておき、評価表現が好評文に出現する頻度、不評文に出現する頻度を使用することで評価表現の極性を決定する手法を提案している [2]。この手法では、同一の評価表現が不評文、好評文に3回以上、出現する必要があるが、業績要因文は多くの名詞や動詞で構成されるため、同一の業績要因文が出現することはまずない。そのため、本手法では業績要因文に含まれる手がかり表現のみに着目して、業績要因文への極性付与を行った。

7 まとめ

本稿では、決算短信より抽出された業績要因文に対して深層学習を用いて極性（「ポジティブ」、「ネガティブ」）を自動的に付与する手法を提案した。具体的には、業績要因文を抽出するための手がかりとなる表現にくつつかの文節を追加することで、“拡張手がかり表現”と定義する文節列を獲得した。そして、“拡張手がかり表現”を使用することで極性付与のための学習データを自動的に生成し、生成された学習データを使用して深層学習を行い、“拡張手がかり表現”を含まない業績要因文に極性を付与した。評価の結果、業績要因への極性付与手法の精度はポジティブで86.7%、ネガティブで90.0%、再現率はポジティブで95.4%、ネガティブで73.9%であり、良好な精度、再現率を達成した。

ブ)を自動的に付与する手法を提案した。具体的には、業績要因文を抽出するための手がかりとなる表現にくつつかの文節を追加することで、“拡張手がかり表現”と定義する文節列を獲得した。そして、“拡張手がかり表現”を使用することで極性付与のための学習データを自動的に生成し、生成された学習データを使用して深層学習を行い、“拡張手がかり表現”を含まない業績要因文に極性を付与した。評価の結果、業績要因への極性付与手法の精度はポジティブで86.7%、ネガティブで90.0%、再現率はポジティブで95.4%、ネガティブで73.9%であり、良好な精度、再現率を達成した。

参考文献

- [1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).
- [2] Kaji, N. and Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, *Proceedings of the EMNLP-CoNLL 2007*, pp. 1075–1083 (2007).
- [3] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).
- [4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).
- [5] Sakai, H. and Masuyama, S.: Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies, *IEICE Trans. Information and Systems*, Vol. E92-D, No. 12, pp. 2341–2350 (2009).
- [6] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信PDFからの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).
- [7] 高村大也, 乾孝司, 奥村学: 隠れ変数モデルによる複数語表現の感情極性分類, 情報処理学会論文誌, Vol. 47, No. 11, pp. 3021–3031 (2006).

経済テキストデータを用いた極性概念辞書構築とその応用

伊藤 友貴^{*1} 坪内 孝太^{*2} 山下 達雄^{*2} 和泉 潔^{*1}
Tomoki Ito Tatsuo Yamashita Tatsuo Yamashita Kiyoshi Izumi

^{*1}東京大学大学院工学系研究科 ^{*2}ヤフー株式会社
School of Engineering, The University of Tokyo Yahoo! JAPAN Research

現状、単語単位での極性辞書自動構築に関する既存研究はあるものの、似たような意味の単語がひとまとまりになった形での極性辞書、極性概念辞書の構築に関する研究はほとんどされていない。本研究では、極性概念辞書の構築手法を提案すると共にその有用性を検証した。まず、ロイターニュース記事・ヤフーファイナンス掲示板のテキスト情報、専門家によって与えられた約 100 単語の極性値、及び構造化されたニューラルネットワークモデルを用いて意味的にクラス分けされた極性概念辞書を作成した。その後、極性概念辞書を用いて市場動向の分析を試みた。

1. はじめに

情報通信技術の発達に伴い、金融テキストマイニングの技術に対する個人投資家及び機関投資家からの関心が高まってきている [1]。金融テキストマイニングとは、投資に有用な情報を SNS や記事のような大規模のテキストデータから抽出する技術である。この分野においては、極性辞書を使うことが有用であることがわかっている [2]。極性辞書とは単語の極性情報に関する辞書である。特定の文脈においてポジティブな意味を持つ単語にはプラスの極性値を、ネガティブな意味を持つ単語にはマイナスの極性値を与える形で各単語にその極性値を割り振っている辞書である。

通常、極性辞書は人手によって作成されるが、経済用語・ネットスラングについては十分な極性についての情報を極性辞書からは手に入れない。経済用語のポジネガ極性付与についての研究もいくつかされているが [3, 4]、今のところ確立された手法はない。また、これらの方法により作られるポジネガ辞書や現在世間に出回っている極性辞書からは、単語単位でのポジネガ情報は抽出できるものの、単語間の類似度は取ることができないため、異なる語群からなるテキストの情報を同じ指標による分析が難しい。このような分析には、似た意味をもつ極性単語がひとまとまりの概念になっている形での極性辞書が必要だが、そのような辞書の自動獲得手法は確立されていない。

自動獲得に関わる機械学習の手法として、近年深層学習の技術発展が著しいが、ニューラルネットワークモデルはその変数の多さ故、中身の分からないブラック関数に近い扱いをされており、モデルの解釈が難しい。

本研究の目的は市場動向分析に有用な単語の極性概念辞書を半自動的に構築するモデルの開発、モデルの解釈が可能でかつ予測精度が高いニューラルネットワークモデルの開発、語彙の異なる複数種類のテキストを同一の指標で分析することが可能な極性概念辞書の作成である。

まず、word2vec [5] を用いてニュース記事に出てくる単語にベクトルを与え、それをもとに各記事に特徴量ベクトルを与えた。これらの単語のうちの一部には経済専門家の手によってつけられた極性スコアが付与されている。その後、極性が不明な単語の極性を求めるために、各ニュース記事の特徴量と株価動向、及び Yahoo!Finance 掲示板上の投稿とその極性タグの対応の関係を分析した。機械学習の学習の過程で極性辞書に含まれる単語の極性スコアが極性辞書外の単語にも伝播することが期待できる。本手法によって得られる特徴量を用いた株価動

向・極性スコアの予測結果と既存手法によって得られる特徴量を用いた予測結果を比較することで、提案手法が与える極性値の妥当性を検証した。さらに、ロイターニュースとヤフーファイナンス掲示板から得られた二つの極性辞書をマージし、両種類のテキストを同時に分析できる極性概念辞書を作成した。

2. 手法

本節では単語への極性値付与に関して本研究で新たに提案する手法 II (importance infiltration) algorithm を紹介する。

2.1 Word classification and document representation 法 (CDR 法)

まず、word2vec によって単語に与えた分散表現を利用した文書の特徴量の生成手法、CDR 法 [6] の紹介をする。CDR 法は提案手法 II algorithm のベースとなった手法である。CDR 法では、意味の近い単語が同じクラスになるようにクラスタリングし、文書内に出現する各クラスの出現回数によって文書の特徴量を生成する。まず、word2vec [5] を使い、ニュース記事に出現する各単語にベクトル表現を与える。その後、クラス数 K を決めた後、K-means 法 [7] により単語のクラスタリングを行う。 K 個のクラスターを得た後、各文書の特徴量 $V_{\text{document}} (\in \mathbb{R}^K)$ を文書中に出現する各クラスの単語の回数をを用いて求める手法である。

2.2 Importance Infiltration propagation algorithm

word2vec で単語をクラスタリングすると、対義語同士が同じクラスに入ってしまうことがある。これは word2vec では各単語がどの言葉と組み合わせで使われるかによって単語の分散表現を獲得するためである。しかし、これは市場動向の分析をする上では望ましくない。そこで、これを改善するために、本論文ではニューラルネットワークモデルを利用したモデル (II (importance infiltration) algorithm) [8] を用いて極性辞書外単語に極性を与え、文書のベクトルを生成した。文書のベクトルを生成するにあたり、金融機関に所属する機関投資家が人手で作成した経済用語極性辞書の情報を用いた。II algorithm では図 1、図 2 のように表現されるニューラルネットワークモデルを用いた。入力層の次元は単語の数で、各単語に対応するノード (1 層目の各ノード) は、それぞれ属するクラスを表すノード (2 層目のノード) のみに結合するという構造のニューラルネットワークモデルである。ここで、 $W_{\text{polarity}} \in \mathbb{R}^{K \times m}$ (式 (5) にて定義、 K はクラス数)、 $W_3 \in \mathbb{R}^{2 \times K}$ 、 $W_3^* \in \mathbb{R}^{K^2 \times K}$ 、 $W_4^* \in \mathbb{R}^{2 \times K^2}$ は重み行列、 K^2

は層の数が4の場合における第3層目のノード数, $b_0 \in \mathbb{R}^2$ はバイアスペクトル, $y_{cls} \in \{0 \text{ (ネガティブ)}, 1 \text{ (ポジティブ)}\}$ は出力層の値である. y_{cls} は文書につけられるタグに該当する値である. 文書番号 j の文書内に出現する単語の頻度からなるベクトル V_{BOW_j} を式 (1) のように定義する.

$$V_{BOW_j} := Z_j^{(1)} = [Z_{j,1}^{(1)T}, Z_{j,2}^{(1)T}, \dots, Z_{j,K}^{(1)T}]^T \quad (1)$$

$Z^{(l)}$ は l 層目の出力を意味する. $Z_{j,k}^{(1)} \in \mathbb{R}^{n(k)}$, $Z_{j,k}^{(1)}[i]$ ($k = 1, 2, \dots, K$) は単語 $w_{k,i}$ ($w_{k,i}$ は単語のクラスが k でクラス内単語 ID が i の単語とする.) の文書番号 j の文書における頻度, $n(k)$ はクラス k である単語の数である. このとき, II algorithm で用いるニューラルネットワークモデル (図 1, 図 2) はそれぞれ3層時には式 (2), 4層時には式 (3) で y_j を表し, (4) のように y_{cls} を表すことで表現できる.

$$y_j = f_3(W_3(\tanh(W_{\text{polarity}} V_{BOW_j}) + b_0)) \quad (2)$$

$$y_j = f_4(W_4^*(\tanh(W_3^*(\tanh(W_{\text{polarity}} V_{BOW_j}) + b_0)))) \quad (3)$$

$$y_{cls_j} = \operatorname{argmax} y_j \quad (4)$$

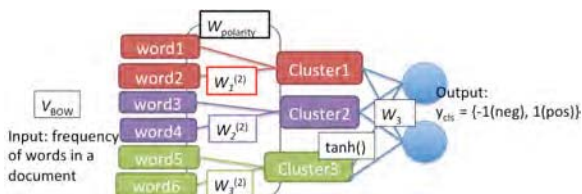


図 1: 本研究で用いたニューラルネットワークモデル (3層)

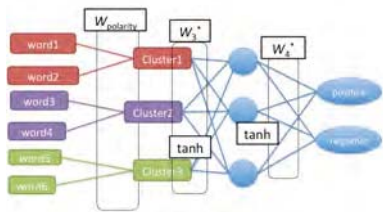


図 2: 本研究で用いたニューラルネットワークモデル (4層)

活性化関数 f_3, f_4 には Softmax 関数を用いた. また, 学習時の損失関数には Softmax cross entropy 関数を用いた. 過学習を防ぐために学習時には, Dropout 法 [9] を用いた. $W_{\text{polarity}} (\in \mathbb{R}^{K \times m})$ を以下のように定義する.

$$W_{\text{polarity}} := \operatorname{diag}(W_1^{(2)T}, W_2^{(2)T}, \dots, W_K^{(2)T}) \quad (5)$$

$W_k^{(2)} \in \mathbb{R}^{n(k)}$ であり, $W_k^{(2)}[i]$ は $w_{k,i}$ の極性値に対応する.

次に $W^{(2)}$ の初期値の与え方について説明する. 単語 $w_{k,i}$ が既存の極性辞書内の単語であり, 極性値が事前に専門家の手によって与えられている場合, 単語 $w_{k,i}$ の極性辞書値 $PS(w_{k,i})$ を用いて $W_k^{(2)}[i]$ の初期値を以下のように与える.

$$W_k^{(2)}[i] = \begin{cases} PS(w_{k,i}) & \text{(単語 } w_{k,i} \text{ が極性辞書単語のとき)} \\ 0 & \text{それ以外} \end{cases}$$

とする. 教師あり学習の過程で, W_{polarity} の値が更新されていく. 学習後に W_{polarity} の値を取り出すことで既存極性辞

書外単語の極性値を得ることができる. $W_k^{(2)}[i]$ の値が伝搬後の単語 $w_{k,i}$ の極性値に対応する.

以下に理論解析から判明した II algorithm (3層) の性質を述べる.

2.3 定義

学習時のミニバッチサイズ N を用いて入力値 $X (\in \mathbb{R}^{m \times N})$, j^+, j^-, p^+, p^-, U, Z を以下に定義する.

$$X := [V_{BOW_1}, V_{BOW_2}, \dots, V_{BOW_N}]$$

$$p^-(w_{k,i}) := p(w_{k,i} \text{ 出現文書がネガティブ} | w_{k,i} \text{ が文書に出現})$$

$$p^+(w_{k,i}) := p(w_{k,i} \text{ 出現文書がポジティブ} | w_{k,i} \text{ が文書に出現})$$

$$U^{(2)} := W_{\text{polarity}} X, Z^{(2)} := \tanh(U^{(2)})$$

$$U^{(3)} := W_3(Z^{(2)} + b_0)$$

便宜的に j を文書番号 j の文書がポジティブである場合に j^+ , 文書番号 j の文書がネガティブである j 場合に j^- と表記する. また, $w_{k,i^+}, w_{k,i^-}, w_{k,i^{neu}}$ を

$$w_{k,i} = \begin{cases} w_{k,i^+} & p_{w_{k,i}^+}^+ > p^{t^+} \\ w_{k,i^-} & p_{w_{k,i}^-}^- > p^{t^-} \\ w_{k,i^{neu}} & \text{otherwise} \end{cases}$$

と定め,

- $n^+(k)$: クラス k 内の単語で $p_{w_{k,i}^+}^+ > p^{t^+}$ である単語の数,
- $n^-(k)$: クラス k 内の単語で $p_{w_{k,i}^-}^- > p^{t^-}$ である単語の数,
- $n^{neu}(k)$: 上記二つの条件を満たさない単語の数

と定める.

2.4 極性伝搬条件 (II algorithm (3層))

II algorithm (3層) について以下に述べる単語への極性付与に関する性質が成り立つ.

- (i) ポジティブ単語 w_{k,i^+} に十分大きい正の極性値, ネガティブ単語 w_{k,i^-} に十分小さい負の極性値が付与
- (ii) $p^-(w_{k,i^-}), p^+(w_{k,i^+}), n^-(k), n^+(k)$ が十分に大きい
- (iii) $W_3[0][k] < 0$ 十分に小さく, $W_3[1][k]$ 十分に大きい
- (iv) ミニバッチサイズ N が十分に大きい

このとき, 任意の i^-, i^{neu}, i^+ について条件 (i)–(iv) を満たすように初期値が与えられるという理想的な場合には

$$E[\partial W_k^{(2)}[i^+] < 0, E[\partial W_k^{(2)}[i^-] > 0$$

となり, ポジティブ単語 w_{k,i^+} にはプラスの極性値が与えられ, ネガティブ単語 w_{k,i^-} にはマイナスの極性値が与えられる.

この主張は条件 (i)–(iv) が成り立つという条件下での誤差逆伝搬の様子を計算して求めることで示すことができる. 具体的な説明については, 付録 A にて紹介する.

2.5 極性伝搬条件 (II algorithm (4層))

II algorithm (4層) について以下に述べる単語への極性付与に関する性質が成り立つ. $U^{(3)}$ についての

$$U^{(3)} := W_3^*(Z^{(2)} + b_0)$$

と定義し直すと, 以下が成立する.

$$\begin{aligned} ((\operatorname{diag}(1 - (\tanh(U_j^{(3)}))^2) W_3^*)^T W_4^{*T})[0] < 0 \\ ((\operatorname{diag}(1 - (\tanh(U_j^{(3)}))^2) W_3^*)^T W_4^{*T})[1] > 0 \end{aligned} \quad (6)$$

が満たされ、かつ 2.4 節の条件 (i)–(iii) が任意の i^- , i^{neu} , i^+ について成り立つとき、単語の極性は

$$E[\partial W_k^{(2)}[i^+]] < 0, E[\partial W_k^{(2)}[i^-]] > 0$$

を満たすように伝搬する。また、条件 (i)–(iii)、及び「 $W_3^* > 0$ かつ $W_3^* > 0$ が十分に大きく、 $W_4^*[0] < 0$ が十分に小さく、 $W_4^*[1] > 0$ が十分に大きい (このとき式 (6) は成立)」を満たすように初期値が与えられるという理想的な場合にはポジティブ単語 w_{k,i^+} にはプラスの極性値が与えられ、ネガティブ単語 w_{k,i^-} にはマイナスの極性値が与えられる。

この主張も上記の条件が成り立つという条件下での誤差逆伝搬の様子を計算して求めることで示すことができる。具体的な証明については、長くなるので本稿では省略する。

3. 人工データを用いた極性伝搬テスト

提案手法の妥当性を示すために、人工データを用いた単語へのポジネガ付与に関する実験をおこなった。本実験は 2.4, 2.5 の確認を目的として行う実験である。

3.1 実験設定

まず、人工データを以下のように生成した。

- クラスタ数 $K = 200$, クラスタ内単語数 $n(k) = 30$ ($k = 1, 2, \dots, 200$)
- $k \leq 100$ の場合: $p^+(w_{k,i}) = \eta$ ($1 \leq i \leq 10$), $p^-(w_{k,i}) = \eta$ ($11 \leq i \leq 20$)
- $k > 100$ または ($k \leq 100 \wedge 20 < i \leq 30$) の場合: $p^+(w_{k,i}) \sim \text{Norm}(0.5, 0.05)$
- 単語 $w_{k,i}$ の文書への出現頻度 $\text{freq}(w_{k,i}) = 0.2$
- データ数: ポジティブ文書 800, ネガティブ文書 800
- ミニバッチサイズ $N = 10$
- $W_k^{(2)}$ の初期値は

$$\begin{cases} W_k^{(2)}[1] \sim U(0, 1) & (k \leq 60) \\ W_k^{(2)}[20] \sim U(-1, 0) & (21 \leq k \leq 80) \\ W_k^{(2)}[i] = 0 & ((k \geq 81) \vee (k \leq 80 \wedge i \neq 1 \wedge i \neq 20)) \end{cases}$$

で生成。

$\text{Norm}(u, \sigma)$ は平均値 u , 標準偏差 σ の正規分布, $U(a, b)$ は区間 $[a, b]$ 上の一様分布を指す。

人工文書データセットを η の値を

$\eta \in \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5\}$ について変えながら生成した。その後 それぞれのデータセットについて II algorithm (epoch 数 10) を用いて各単語にポジネガスコアを与えた後、各 $k \leq 100$ について

- $(W_k^{(2)}[i] > 0) \wedge (1 \leq i \leq 10)$ を満たす $w_{k,i}$ の数 (= $n_{\text{pred}}^+(k)$ とする。),
- $(W_k^{(2)}[i] < 0) \wedge (11 \leq i \leq 20)$ を満たす $w_{k,i}$ の数 (= $n_{\text{pred}}^-(k)$)

を求めた。 $n_{\text{pred}}^+(k)$, $n_{\text{pred}}^-(k)$ によって、以下の妥当性の指標

$$AC = \frac{\sum_{k=1}^{100} (n_{\text{pred}}^+(k) + n_{\text{pred}}^-(k))}{200},$$

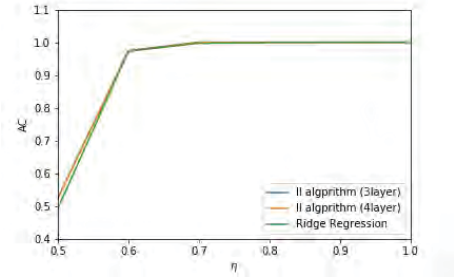


図 3: 極性伝搬の検証 (AC の平均値)

の試行回数 5 回での平均値を求め、妥当性を検証した。2.4 節, 2.5 節における理論解析の結果から、AC の値は η が大きくなれば 1.0 に近づくことが期待できる。

ここで、ニューラルネットワークモデルのパラメータ値 W_3 に初期値を II algorithm (3 層) を使う場合には

$$W_3[0][k] \sim U(-0.01, 0), W_3[1][k] \sim U(0, 0.01)$$

に従って与えた。また、II algorithm (4 層) を使う場合にはパラメータ値 W_3^* , W_4^* に初期値を

$$W_3^* \sim U(0, 0.01), W_4^*[0][k] \sim U(-0.01, 0),$$

$$W_4^*[1][k] \sim U(0, 0.01)$$

に従って与えた。

これは 2.4 節, 2.5 節を踏まえてのものである。II algorithm を用いた極性値付与手法の妥当性を確かめるため、リッジ回帰モデル [3] を用いた極性値付与手法による結果と比較する。ここで、リッジ回帰の L2 正則化のパラメータ値はグリッドサーチによって $\{0.1, 1.0, 10.0\}$ の範囲でチューニングして決めた。

3.2 実験結果

図 3 が実験結果である。II algorithm (3layer) が II algorithm (3 層), II algorithm (4layer) が II algorithm (4 層), Ridge Regression がリッジ回帰モデルをそれぞれ表す。 $\eta = 0.5, 0.6, 0.7, 0.8, 0.9$, の範囲で動かしたとき、期待通り、 η の値が大きいくほど AC の値は大きくなることを確認できた。この結果より、II algorithm による単語への極性値の付与が「 $p^+(w_{k,i})$ が大きい単語はポジティブ単語であり、 $p^-(w_{k,i})$ が大きい単語はネガティブ単語である」という観点のもとでは妥当であることを検証できた。

4. 実データを用いた極性伝搬の検証

実データを用いた検証によって、「II algorithm が付与する単語の極性値の有用性」と「II algorithm の予測モデルとしての有用性」を検証した。

4.1 データセット

有用性を調査するにあたり以下の実験をした。

4.1.1 株価動向分析

本実験においては 2013 年 1 月から 2015 年 12 月までの間に配信されたトムソンロイターの経済ニュース記事のうち、銘柄コードの入りの記事を用いた。これらの記事が個別銘柄の株価動向に与える影響を予測した。まず、ニュース記事の配信日を d , 記事に最初に銘柄コードが出てくる個別銘柄の時刻 t_p

の株価を p_{t_p} 、及び時刻 $t_a (> t_p)$ の株価 p_{t_a} (表 1 のパターン参照) から算出される株価リターンを用いて以下の y_{cls} で定めるタグを各記事につけた。

$$\Delta p(d) = \frac{p_{t_a} - p_{t_p}}{p_{t_p}}, y_{cls} = \begin{cases} 1 & (\Delta p(d) > 0.01) \\ -1 & (\Delta p(d) < -0.01) \end{cases}$$

その後、記事のタグの予測を行い、F 値の 10 交差検定平均スコアを算出し評価した。

表 1: 株価リターンの設定 (ロイターニュース)

パターン名	p_{t_p}	p_{t_a}
open_nowday_to_close_nextday	$d-1$ の始値	$d+1$ の始値
close_nowday_to_close_nextday	d の終値	$d+1$ の終値
close_nowday_to_open_nextday	d の終値	$d+1$ の始値
open_nextday_to_close_nextday	$d+1$ の始値	$d+1$ の終値
close_previousday_to_open_nextday	$d-1$ の終値	$d+1$ の始値
close_previousday_to_close_nextday	$d-1$ の終値	$d+1$ の終値

4.1.2 ヤフーファイナンス掲示板ボジネガタグ予測

Yahoo!Finance 掲示板^{*1} の投稿に付与されるタグの予測をおこなった。タグの種類は $\{1, 2, 3, 4, 5\}$ である。1 は「強く買いたい」、5 は「強く売りたい」を意味する。

以下の二つのケースについて実験を行った。

- 短期間における多種多様な銘柄の投稿に関する予測: 2014/9/1 ~ 2014/9/30 までの全銘柄の投稿について投稿日の古いタグ 1 とタグ 5 の投稿 10000 件ずつを抽出し、タグを予測し、その F 値の 5 交差検定平均スコアをもとに評価した。
- 長期間における少数銘柄の投稿に関する予測: 2014/11/18 ~ 2016/6/15 まで銘柄番号 998407, 9501, 4777, 8462, 4564, 6871 のタグが 1 または 5 の投稿を抽出し、2015/6/1 ~ 2016/5/31 について各月の投稿のタグを検証データ、各月の前月以前のタグが 1 と 5 の投稿を各 20000 件ずつを訓練用データとして予測精度を検証した。月毎に F 値を算出し、12 回の平均スコアをもとに評価した。

4.2 前処理

本実験において、形態素解析には MeCab [11] を用いた。

特徴量を生成するにあたり、実験 4.1.1 項の「株価動向分析」では動詞、形容詞、名詞を抽出 (全部で 21286 語) し、また、実験 4.1.2 項の [短期間における多種多様な銘柄の投稿に関する予測] においては動詞、形容詞、名詞を中心に単語 (全部で 25218 語) を抽出し、各単語の頻度から文書の特徴量を生成した。実験 4.1.2 項の [長期間における少数銘柄の投稿に関する予測] においては形態素を抽出した上で、極大部分文字列 [12] を抽出する手法により、文書全体にて 100 回以上出現する極大部分形態素列 16627 表現を抽出し、それらの頻度をもとに文書の特徴量を生成した。

4.3 ベースライン

4.3.1 II algorithm が付与する単語極性値の有用性

Label [13], SSL [4], Ridge Regression [3] の各既存手法及び、II algorithm (3 層) (II algorithm (layer3)), II algorithm (4 層) (II algorithm (layer4)) によって得られた各単語の極性

値 ($PS^*(word_l)$ を単語 $word_l$ とする) と CDR 法 (2.1 節) を用いて以下のように $W_{polaritybase}$ を生成し、 $V_{BOW} W_{polaritybase}$ の値を文書の特徴量とした。

$$W_{polaritybase} = (\delta_{ij}^*)^T$$

$$\delta_{ij}^* = \begin{cases} PS^*(word_l) & (word_l \in class_j) \\ 0 & \text{else} \end{cases}$$

これらの特徴量による予測結果に加え、CDR 法 (任意の l について $PS^*(word_l) = 1$ である場合に相当)、AntSyn [14] を用いた CDR 法 (AntSyn)、 V_{BOW} (BOW) により生成された特徴量による予測結果を比較した。予測モデルには線形 SVM を用いた。II algorithm の初期値付与に利用する単語極性値には金融機関に所属する機関投資家が人手で作成した経済用語極性辞書の値を利用した (約 200 単語)。II algorithm の精度を求めるにあたっては、5 回の試行平均値をもとに算出した。II algorithm (3 層・4 層) の 2 層目の次元を $K = 500$, II algorithm (4 層) の 3 層目の次元数、4.1.2 項における epoch 数 (学習回数) は訓練データ内におけるチューニングにより決定し、4.1.1 項では epoch 数 (学習回数) を 50 とした。

4.3.2 II algorithm の予測モデルとしての有用性

II algorithm を用いた場合の予測精度と多層パーセプトロンモデル [15] の 3 層モデル (MLP3)、及び 4 層モデル (MLP4) を用いた場合の予測精度の比較をおこなった。

4.4 実験結果

図 4 の上図が 4.3.1 項における「株価動向分析 (4.1.1)」の結果、下図が 4.3.2 項における「株価動向分析 (4.1.1 項)」の結果である。この結果からロイターニュースはパターン close_previousday_to_open_nextdayII.algorithm の株価リターンに最も影響を与えたとわかる。また、パターン close_previousday_to_open_nextdayII.algorithm の株価リターンの場合のより詳細な結果は表 2 の通りである。表 3 が実験 4.3.1 項における「ヤフーファイナンス掲示板ボジネガタグ予測 (4.1.2 項)」の結果、表 4 が 4.3.2 項における「ヤフーファイナンス掲示板ボジネガタグ予測 (4.1.2 項)」の結果である。既存手法との結果の比較により、II algorithm が与える単語の極性値から生成される特徴量から作られる予測モデルの方が他の手法から生成される特徴量により作られる予測モデルに比べて F 値を指標とした場合に高い予測力を持つことがわかった。また、II algorithm (3 層) が MLP3 に比べ、また II algorithm (3 層) が MLP 4 に比べそれぞれ F 値を指標とした場合に高い予測力を持つことがわかった。

本節では II algorithm を用いたマルチタイプ極性概念辞書の作成手法の説明、及びマルチタイプ極性概念辞書の出力結果の紹介を行う。

4.5 極性辞書マージ手法紹介

4.5.1 手法概要

ヤフー辞書内の単語クラスとロイター辞書内の単語クラスについて、クラス単位でのマージを行い、マージ後のクラスが似た意味の単語群で構成されることを目指す。

4.5.2 マージ手順

具体的なマージ手法の手順を以下に記す。

- ロイター辞書、ヤフー辞書のそれぞれの各クラスターのセントロイドを求め、 \cos 類似度をもとに類似度行列 $W_{ruiter} \in \mathbb{R}^{K_{ruiter} \times K_{ruiter}}$ と $W_{yahoo} \in \mathbb{R}^{K_{yahoo} \times K_{yahoo}}$ をそれぞれ作成 ($K_{ruiter} = 500, K_{yahoo} = 1000$) する。ここで、

*1 <http://textream.yahoo.co.jp/category/1834773>

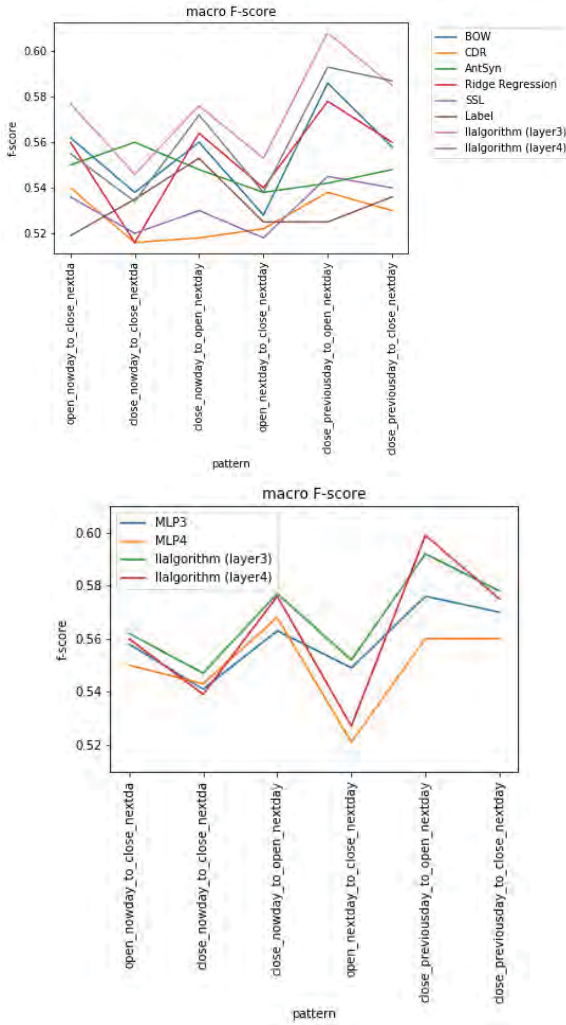


図 4: 株価動向分析結果 (F 値)

表 2: パターン close_previousday_to_open_nextday の結果

Methods	F 値 ($\pm\sigma$)
BOW	0.586
CDR	0.538
AntSyn	0.542
Ridge Regression	0.578
SSL	0.545
Label	0.525
II algorithm (layer3)	0.608 (± 0.005)
II algorithm (layer4)	0.593 (± 0.004)

K_{ruiter} , K_{yahoo} はそれぞれロイター辞書, ヤフー辞書におけるクラスタ数である.

- 類似度行列 2 つを対角に並べて W_{merge} を作成

$$W_{merge} = \begin{pmatrix} W_{yahoo} & 0 \\ 0 & W_{ruiter} \end{pmatrix}$$

表 3: ヤフーデータ掲示板ボジネガタグ予測

Methods	F 値 ($\pm\sigma$) (短期)	F 値 ($\pm\sigma$) (長期)
BOW	0.735	0.790
CDR	0.684	0.724
Ridge Regression	0.754	0.786
AntSyn	0.684	0.700
SSL	0.570	0.660
Label	0.633	0.651
II algorithm (layer3)	0.763 (± 0.001)	0.795 (± 0.001)
II algorithm (layer4)	0.760 (± 0.001)	0.791 (± 0.001)

表 4: ヤフーデータ掲示板ボジネガタグ予測

Methods	F 値 ($\pm\sigma$) (短期)	F 値 ($\pm\sigma$) (長期)
MLP3	0.742 (± 0.004)	0.774 (± 0.002)
MLP4	0.743 (± 0.004)	0.772 (± 0.002)
II algorithm (layer3)	0.775 (± 0.001)	0.800 (± 0.001)
II algorithm (layer4)	0.771 (± 0.001)	0.792 (± 0.001)

- ロイター辞書内の単語とヤフー辞書内の単語と人手極性辞書の共通単語群 $\{w_{com}[0], w_{com}[1], \dots, w_{com}[n-1]\}$ を利用して Algorithm 1 のようにロイター単語クラスとヤフー単語クラスをつなげる.

Algorithm 1 エッジ挿入

Input: W_{merge}

- for $i \leftarrow 0$ to $n-1$ do
- $W_{merge}[yclass(w_{com}), rclass(w_{com}) + K_{yahoo}] \leftarrow 1;$
- $W_{merge}[rclass(w_{com}) + K_{yahoo}, yclass(w_{com})] \leftarrow 1;$
- end for
- W_{merge} を返す

$yclass(w_{com})$ は単語 w_{com} のヤフー辞書におけるクラス番号であり, $rclass(w_{com})$ は単語 w_{com} のロイター辞書におけるクラス番号である.

- SVD [16] により欠損値の補完を行う.
- W_{merge} の成分のうち, 閾値 $T = 0.9$ より小さい値は 0 にし, それ以外は 1 にする.
- W_{merge} について SpectralClustering [17] を用いてマージを行う.

4.6 数値実験

4.6.1 実験設定

まず, II algorithm (3 層) によりヤフー辞書, ロイター辞書の作成をおこなった. その後, 4.5 節の方法でヤフーファイナンス掲示板から作成した極性辞書とロイターニュースから作った極性辞書のマージを行った. マージ後のクラスタ数は 100 とした. ロイター辞書にはパターン close_previousday_to_open_nextday の株価リターンとニュース記事の対応から IIalgorithm (layer3) を用いて作成した極性辞書を用いた. ヤフー辞書には「ヤ

フーデータ掲示板ポジネガタグ予測 (短期) から II algorithm (layer3) を用いて作成した極性辞書を用いた .

4.6.2 極性概念辞書出力結果

次に、実験 4.6.1 によって作成された極性概念辞書の一部を紹介する .

表 5-6 は辞書のマージ後に「上昇」と同じクラスの単語とその極性値を、マージ前のロイター辞書の単語クラス、ヤフー辞書の単語クラス別に表示した結果の一部である . 例えば、表 5 から「上昇」のプラスの極性値が II algorithm によって同じクラスの「急騰」や「急上昇」に伝搬し、逆に「下落」や「下降」のマイナスの極性値が II algorithm によって同じクラスの「急降下」や「暴落」に伝搬している様子がわかる .

5. まとめ

本論文において、提案手法の性質を理論的に解析すると共に、提案手法が市場動向分析において他の既存手法に比べ有用であることを実験的に示した . また、ロイターニュースとヤフーファイナンス掲示板の両方を同じ指標で分析可能な極性概念辞書を作成した . 今後の課題として、II algorithm (4 層) を利用したフレーズ単位での極性情報抽出手法の考案、極性辞書の情報が少ない場合や少数のラベルつきデータしか手に入れない状況でも安定的で高い予測精度を出すための改良法の開発、マルチタイプ極性概念辞書を用いた複数種類のテキストデータの分析方法の開発などが考えられる .

参考文献

- [1] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, D. C. L. Ngo, "Text mining for market prediction: A systematic review", Journal of Expert Systems with Applications, Vol. 41, Issue 16, pp. 7653-7670, 2014
- [2] W. Ye and F. Ren, "Learning sentimental influence in twitter", ICFCSA, 2011
- [3] K. Tsubouchi and T. Yamashita, "Positive / Negative Detection for Finance Contents via Stock Bulletin Boards Data", JSAI 2014, 2014
- [4] H. Yanagimoto, "Improvement of Sentiment Dictionary Using Neural Network Language Model", JSAI 2014, 2014
- [5] 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
- [6] Y. Yuan, L. He, L. Peng, Z. Huang, "A New Study Based on Word2vec and Cluster for Document Categorization", Journal of Computational Information Systems, Vol. 10, Issue 21, pp. 9301-9308, 2014
- [7] K. Hornik, I. Feinerer, M. Kober, C. Buchta, "Spherical k-Means Clustering", Journal of Statistical Software, Vol. 50, Issue 10, pp. 1-22, 2012
- [8] T. Ito, K. Izumi, K. Tsubouchi, T. Yamashita, "Polarity propagation of financial terms for market trend analyses using news articles", CEC 2016, 2016

表 5: クラス 551 (ヤフー辞書)

単語	極性値 (伝搬後)	(極性辞書値)
上昇	0.328	(0.5)
反発	0.232	(0.5)
動意	0.297	(0.0)
動き	0.276	(0.0)
急落	-0.525	(0.0)
反転	0.498	(0.0)
下落	-0.655	(-0.333)
下降	-0.818	(-0.333)
急騰	0.184	
下げ	-0.307	
短期間	0.007	
急上昇	0.278	
反騰	-0.015	
上下	0.137	
値動き	0.095	
急降下	-0.411	
上げ	0.017	
調整	0.366	
盛り上がり	0.096	
乱高下	0.081	
上下動	0.333	
上がり	0.208	
リバウンド	-0.240	
上げ過ぎる	-0.160	
暴落	-0.791	
暴騰	0.245	
連騰	0.240	
相場	0.083	
上げ下げ	0.245	

表 6: クラス 304 (ロイター辞書)

単語	極性値 (伝搬後)	(極性辞書値)
続伸	1.357	(1.118)
急伸	0.678	(0.5)
反発	0.594	(0.5)
上昇	0.443	(0.5)
急落	-0.070	(0.0)
推移	0.167	(0.0)
下落	-0.453	(-0.333)
低下	-0.384	(-0.375)
続落	-1.229	(-0.929)
反落	-1.250	(-0.938)
急騰	-0.088	
下げ	-0.120	
下押し	-0.096	
下押す	0.118	
鬼門	0.072	
下がる	-0.052	
伸び悩む	-0.034	
軟化	-0.171	
値上がり	0.025	
値下がり	0.060	
安	0.069	
買われる	0.060	
乱高下	0.261	
安い	0.155	
下げる	-0.178	
急上昇	0.089	
上げ	-0.160	
下げ止まる	0.200	
ドル安	-0.033	
伸し	0.082	

- [9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, Vol. 15, pp. 1929–1958, 2014
- [10] D. P. Kingma, J. L. Ba, "Adam: A Method for Stochastic Optimization", arXiv:1412.6980, 2014
- [11] T. Kudo, K. Yamamoto, Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis", EMNLP 2004, pp. 230–237, 2004
- [12] D. Okanohara, J. Tsujii, "The Categorization with all Substring Features", SDM 2009, pp. 838–846, 2010
- [13] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Scholkopf, "Learning with Local and Global Consistency", NIPS 2003, pp. 321–328, 2003
- [14] K. A. Nguyen, S. S. im Walde, N. T. Vu, Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction, ACL 2016, 2016
- [15] 岡谷貴之, "機械学習プロフェッショナルシリーズ 深層学習", pp12–14, 講談社, 2015
- [16] K. Yu, S. Zhu, J. Lafferty, Y. Gong, "Fast Nonparametric Matrix Factorization for Large-scale Collaborative Filtering", SIGIR 2009, 2009
- [17] U. V. Luxburg, "A tutorial on spectral clustering", Journal of Statistics and Computing, Vol. 17, Issue 4, pp. 395–416, 2007

付録 A II algorithm (3層) 理論解析

本節では、提案手法である II algorithm (3層) によってどのようにして単語に極性値が付与されるのかについて理論的な側面から説明し、II algorithm によって妥当に各単語に極性値が付与されることを説明する。

A1 定義

まず、出力値 y , ミニバッチ内文書番号 j の文書の目的値 d_j について以下のように定義する。

$\text{freq}(w_{k,i})$: 単語 $w_{k,i}$ が文書に出てくる頻度確率

$$\begin{aligned} \mathbf{y} &:= (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N) \\ &= \text{softmax}(\mathbf{W}_3(\tanh(\mathbf{W}_{\text{polarity}} \mathbf{X}) + \mathbf{b}_0)) \in \mathbb{R}^{2 \times N} \end{aligned}$$

$$\begin{cases} d_j = (0, 1)^T & (\text{文書番号}j\text{の文書がポジティブの場合}) \\ d_j = (1, 0)^T & (\text{文書番号}j\text{の文書がネガティブの場合}) \end{cases}$$

また、簡単のため、 $\mathbf{Z}_{j,k}^{(1)}[i] \in \{0, 1\}$ であると仮定する。

A2 パラメータの更新

$\mathbf{W}^{(3)}, \mathbf{W}_k^{(2)}$ ($k = 1, 2, \dots, K$) は次のように更新される。

$$\Delta_j^{(3)} := \mathbf{y}_j - d_j \quad (7)$$

$$= \begin{cases} (|\Delta_j^{(3)}[0]|, -|\Delta_j^{(3)}[1]|)^T & (d_j = (0, 1)^T) \\ (-|\Delta_j^{(3)}[0]|, |\Delta_j^{(3)}[1]|)^T & (d_j = (1, 0)^T) \end{cases} \quad (8)$$

$$\Delta^{(2)} := (1 - (\tanh(\mathbf{U}^{(2)}))^2) \odot (\mathbf{W}_3[1]^T \Delta^{(3)}[1] + \mathbf{W}_3[0]^T \Delta^{(3)}[0]) \quad (9)$$

$$\partial \mathbf{W}^{(3)} = \frac{1}{N} \Delta^{(3)} \mathbf{Z}^{(2)T}$$

$$\partial \mathbf{W}_k^{(2)} = \frac{1}{N} \sum_{j=1}^N \Delta_j^{(2)} [k] \mathbf{Z}_{j,k}^{(1)}$$

$$\mathbf{W}^{(3)} \leftarrow \mathbf{W}^{(3)} - \epsilon \partial \mathbf{W}^{(3)}, \mathbf{W}_k^{(2)} \leftarrow \mathbf{W}_k^{(2)} - \epsilon \partial \mathbf{W}_k^{(2)} \quad (10)$$

\odot はアダマール積(要素積)を指し、 $\epsilon (> 0)$ は更新時に Adam [10] によって計算される。

また、 $E_{j+} [|\Delta_{j+}^{(2)}[k]| \mathbf{Z}_{j+,k}^{(1)}[i] = 1]$ を $E_{j+} [|\delta_{j+}^{(2)}[k]|]$, $E_{j-} [|\Delta_{k,j-}^{(2)} [|\mathbf{Z}_{k,j-}^{(1)}[i] = 1]|]$ を $E_{j-} [|\delta_{j-}^{(2)}[k]|]$ と表記することにする。

A3 極性伝搬条件

このとき、2.4 節で述べた主張が成り立つ。これは、2.3 節における p^{t+}, p^{t-} がそれぞれ十分に大きい場合に、ポジティブ単語 $w_{k,i+}$ にはプラスの極性値が与えられ、ネガティブ単語 $w_{k,i-}$ にはマイナスの極性値が与えられることを保証するものである。

この主張が成り立つ理由について以下の A4, A5 にて説明する。まず、 $\mathbf{W}_3[0][k] < 0, \mathbf{W}_3[1][k] > 0$ の場合に十分大きい p^{t+}, p^{t-} が与えられたときに

$$E[\partial \mathbf{W}_k^{(2)}[i^+]] < 0, E[\partial \mathbf{W}_k^{(2)}[i^-]] > 0$$

が成り立つことを示す。次に条件 (i)–(iv) を満たすように初期値が与えられた場合に更新後も $\mathbf{W}_3[0][k] < 0, \mathbf{W}_3[1][k] > 0$ であることが維持されることを示す。これら二つを示すことによって、条件 (i)–(iv) を満たすように初期値が与えられるという理想的な場合にはポジティブ単語 $w_{k,i+}$ にはプラスの極性値が与えられ、ネガティブ単語 $w_{k,i-}$ にはマイナスの極性値が与えられることを説明できる。

A4 $\mathbf{W}_3[0][k] < 0, \mathbf{W}_3[1][k] > 0$ の場合の解析

$$\Delta^{(2)} = (1 - (\tanh(\mathbf{U}^{(2)}))^2) \odot (\mathbf{W}_3[1]^T \Delta^{(3)}[1] + \mathbf{W}_3[0]^T \Delta^{(3)}[0])$$

について

$$\Delta_{j+}^{(2)}[k] < 0, \Delta_{j-}^{(2)}[k] > 0 \quad (11)$$

であることを示す。(8) より

$$\begin{aligned} &(\mathbf{W}_3[0][k]^T \Delta^{(3)}[0])_j \\ &= \begin{cases} \mathbf{W}_3[0][k]^T |\Delta_j^{(3)}[0]| < 0 & (d_j = (0, 1)^T) \\ -\mathbf{W}_3[0][k]^T |\Delta_j^{(3)}[0]| > 0 & (d_j = (1, 0)^T) \end{cases} \quad (12) \end{aligned}$$

である。同様に、

$$\begin{cases} (\mathbf{W}_3[1][k]^T \Delta^{(3)}[1])_j < 0 & (d_j = (0, 1)^T) \\ (\mathbf{W}_3[1][k]^T \Delta^{(3)}[1])_j > 0 & (d_j = (1, 0)^T) \end{cases} \quad (13)$$

である。ここで、 $(1 - (\tanh(\mathbf{U}^{(2)}))^2) > 0$ なので、(9)(12)(13) より、(11) が言える。このとき、

$$E[\partial \mathbf{W}_k^{(2)}[i]] = E \left[\frac{1}{N} \sum_{j=1}^N \Delta_j^{(2)} [k] \mathbf{Z}_{j,k}^{(1)}[i] \right] = E_j [\Delta_j^{(2)} [k] \mathbf{Z}_{j,k}^{(1)}[i]]$$

$$= \text{freq}(w_{k,i})(-p^+(w_{k,i})E_{j+}[\delta_{j+}^{(2)}[k]] + p^-(w_{k,i})E_{j-}[\delta_{j-}^{(2)}[k]])$$

$$\therefore \begin{cases} E[\partial \mathbf{W}_k^{(2)}[i]] < 0 & \left(\frac{p^+(w_{k,i})}{p^-(w_{k,i})} > \frac{E_{j-}[\delta_{j-}^{(2)}[k]]}{E_{j+}[\delta_{j+}^{(2)}[k]]} \right) \\ E[\partial \mathbf{W}_k^{(2)}[i]] > 0 & \left(\frac{p^+(w_{k,i})}{p^-(w_{k,i})} < \frac{E_{j-}[\delta_{j-}^{(2)}[k]]}{E_{j+}[\delta_{j+}^{(2)}[k]]} \right) \end{cases} \quad (14)$$

となる．すなわち， p^{++}, p^{+-} がそれぞれ十分に大きい場合，ポジティブ単語 $w_{k,i+}$ には正の極性値，ネガティブ単語 $w_{k,i-}$ には負の極性値が与えられるようパラメータは更新される．

次に，任意の単語の頻度 $\text{freq}(w_{k,i})$ が一定値であり， p^{++}, p^{+-} が十分に大きいという場合にポジティブ単語 $w_{k,i+}$ ，ネガティブ単語 $w_{k,i-}$ に与えられる極性値の絶対値の大きさがニュートラル単語 $w_{k,i^{neu}}$ に比べて大きくなることを示す．

$$E[\partial \mathbf{W}_k^{(2)}[i]] = E \left[\left| \frac{1}{N} \sum_{j=1}^N \Delta_j[k] \mathbf{Z}_{j,k}^{(1)}[i] \right| \right]$$

$$= E[\Delta_j^{(2)}[k] \mathbf{Z}_{j,k}^{(1)}[i]]$$

$$= \text{freq}(w_{k,i}) - p^+(w_{k,i})E_{j+}[\delta_{j+}^{(2)}[k]] + p^-(w_{k,i})E_{j-}[\delta_{j-}^{(2)}[k]]$$

$$= \text{freq}(w_{k,i})p^-(w_{k,i})2E_{j-}[\delta_{j-}^{(2)}[k]] - E_{j+}[\delta_{j+}^{(2)}[k]]$$

故に，任意の単語の頻度 $\text{freq}(w_{k,i})$ が一定値の場合，

$$T_{pt} := \frac{E_{j+}[\Delta_{j+}^{(2)}[k] | \mathbf{Z}_{k,j+}^{(1)}[i] = 1]}{2E_{j-}[\Delta_{j-}^{(2)}[k] | \mathbf{Z}_{k,j-}^{(1)}[i] = 1]}$$

とすると

$$\begin{cases} E[\partial \mathbf{W}_k^{(2)}[i^+]] > E[\partial \mathbf{W}_k^{(2)}[i^{neu}]] & (pt - < T_{pt}) \\ E[\partial \mathbf{W}_k^{(2)}[i^-]] > E[\partial \mathbf{W}_k^{(2)}[i^{neu}]] & (pt + > T_{pt}) \end{cases} \quad (15)$$

が成り立つ．

A5 中間層と出力層の間のエッジの解析

$$E_{j+}[\Delta_{j+}^{(3)}[0] \mathbf{Z}_{j+}^{(2)}[k]]$$

$$= E_{j+}[\Delta_{j+}^{(3)}[0] \tanh(\sum_{i=1}^{n(k)} \mathbf{Z}_{j+,k}^{(1)}[i] \mathbf{W}_k^{(2)}[i])]$$

$$\sum_{i=1}^{n(k)} \mathbf{Z}_{j+,k}^{(1)}[i] \mathbf{W}_k^{(2)}[i] = \sum_{i^+=1}^{n^+(k)} \mathbf{Z}_{j+,k}^{(1)}[i^+] \mathbf{W}_k^{(2)}[i^+] + \sum_{i^{neu}=1}^{n^{neu}(k)} \mathbf{Z}_{j+,k}^{(1)}[i^{neu}] \mathbf{W}_k^{(2)}[i^{neu}] + \sum_{i^-=1}^{n^-(k)} \mathbf{Z}_{j+,k}^{(1)}[i^-] \mathbf{W}_k^{(2)}[i^-]$$

である．ここで，

$$E[\mathbf{Z}_{j+,k}^{(1)}[i]] = \text{freq}(w_{k,i})p^+(w_{k,i}) = \text{freq}(w_{k,i})(1 - p^-(w_{k,i}))$$

なので， $\mathbf{W}_k^{(2)}[i^+] > 0$ ， $\mathbf{W}_k^{(2)}[i^+] \gg \mathbf{W}_k^{(2)}[i^{neu}]$ であり，かつ $p^-(w_{k,i-})$ ， $p^+(w_{k,i+})$ ， $n^+(k)$ の値が十分に大きいとき，

$$E_{j+}[\Delta_{j+}^{(3)}[0] \mathbf{Z}_{j+}^{(2)}[k]] > 0$$

となる．同様に， $\mathbf{W}_k^{(2)}[i^-] < 0$ ， $\mathbf{W}_k^{(2)}[i^-] \ll \mathbf{W}_k^{(2)}[i^{neu}]$ であり， $p^-(w_{k,i-})$ ， $p^+(w_{k,i+})$ ， $n^-(k)$ の値が十分に大きいとき，

$$E_{j-}[\Delta_{j-}^{(3)}[0] \mathbf{Z}_{j-}^{(2)}[k]] > 0$$

よって，(i)(ii) が任意の i^+ ， i^{neu} ， i^- について成り立つならば，式 (12) より，

$$E[\partial \mathbf{W}_3[0][k]] = E \left[\frac{1}{N} \cdot \sum_{j=1}^N \Delta_j^{(3)}[0] \mathbf{Z}_j^{(2)}[k] \right] > 0$$

となる．同様に，(i)(ii) が任意の i^+ ， i^{neu} ， i^- について成り立つならば，

$$E[\partial \mathbf{W}_3[1][k]] < 0$$

となる．よって， $\mathbf{W}_3[1][k] > 0$ が十分に大きく， $\mathbf{W}_3[0][k] < 0$ が十分に小さいならば，

$$E[\mathbf{W}_3[1][k] - \epsilon \partial \mathbf{W}_3[1][k]] > 0, E[\mathbf{W}_3[0][k] - \epsilon \partial \mathbf{W}_3[0][k]] < 0.$$

が成り立つ．ここで， u ， σ^2 を

$$u := E[\Delta_j^{(3)}[0] \mathbf{Z}_j^{(2)}[k]] (> 0), \sigma^2 := \text{Var}[\Delta_j^{(3)}[0] \mathbf{Z}_j^{(2)}[k]].$$

と定義すると，大数の法則より，

$$P(\epsilon \partial \mathbf{W}_3[0][k] \leq -\mathbf{W}_3[0][k]) \leq \frac{\epsilon^2 \sigma^2}{2N \cdot (eu + \mathbf{W}_3[0][k])^2} \left(\xrightarrow{N \rightarrow \infty} 0 \right)$$

となる．よって，学習時のミニバッチサイズ N が十分に大きく， $\mathbf{W}_3[0][k] (< 0)$ が十分に小さいならば，

$$\mathbf{W}_3[0][k] < 0 \quad (16)$$

は更新後も維持される．

同様に，学習時のミニバッチサイズ N が十分に大きく， $\mathbf{W}_3[1][k] (> 0)$ が十分に大きいならば，

$$\mathbf{W}_3[1][k] > 0 \quad (17)$$

は更新後も維持される．故に，(i)-(iv) が任意の i^+ ， i^{neu} ， i^- について満たされるならば，

$$\mathbf{W}_3[0][k] < 0, \mathbf{W}_3[1][k] > 0$$

は更新後も維持される．

A4 における式 (14) 及び A5 における解析結果から，2.4 節で主張したことが説明でき，II algorithm がボジネガ付与の手法として妥当であることを保証できた．

Cross-lingual News Article Comparison Using Bi-graph Clustering and Siamese-LSTM

Enda Liu¹ Kiyoshi Izumi¹ * Kota Tsubouchi² Tatsuo Yamashita²

¹ School of Engineering, The University of Tokyo

² Yahoo Japan Corporation

Abstract: Calculating similarity score for monolingual text is a popular task since it could be used for various text mining system. However seldom research is focusing on multilingual text resources. On the other hand, machine learning based algorithms such as CBOW word embedding and clustering are widely used in extracting features of text. In this research, we develop and train a model that could calculate the similarity of the two finance news reports, by utilizing CBOW, spherical clustering, bi-graph extraction as well as the Siamese-LSTM deep learning model. In the end, we train the model by feeding news data that is closely related in the financial domain to help us to analyze the relationship among news reports written in different languages.

1 Introduction

Financial text mining has been widely used for financial analysis and one popular example is building financial prediction model with public database by means of machine learning algorithm and natural language processing [9].

However, there is only few studies about cross-lingual text mining and sometimes foreign news reports will also give impacts on local markets. Although there are some popular solutions for multilingual translation, they are neither designed nor specialized for financial domain. Hence, it becomes a needful work to construct a system that can establish relationships among multilingual financial text so that both developers and users are able to handle the foreign information promptly and accurately.

On the other hand, Vector representation of words such as skip-gram and CBOW[8] is efficient methods used in text mining. In this paper we present the cross-lingual study on English and Japanese text data respectively, investigating possible relationships between the two language models.

Finally, we also develop an piratical application basing on our research, which could calculate the similarity of the two cross-lingual finance news reports. Although calculating similarity score for monolingual

text is a popular task since it could be used for various text mining system, seldom research is focusing on multilingual text. Our application is basing on the results of the extraction of bi-graph structures as well as the Siamese-LSTM deep learning model[5]. It performs well, proving the efficiency even though the training news data are closely related in the financial domain and also helps us to analyze the relationship among news reports written in different languages.

2 Framework of bi-graph structure extraction

The overview framework of the bi-graph structure extraction is shown in Figure 1. Generally, our framework includes five main blocks: data retrieving and preprocessing, word2vec modeling, dictionary extraction(vector assignment), and then clustering with spherical k-means and finally mapping for cross-lingual clusters in order to extract the bi-graph structures.

2.1 Data preparation and preprocessing

There are mainly three steps for data preparation and preprocessing:

1. We first retrieve both the Japanese and English news reports related to finance and economics

*連絡先: Dep. of Systems Innovation, School of Engineering, the University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan
E-mail: izumi@sys.t.u-tokyo.ac.jp

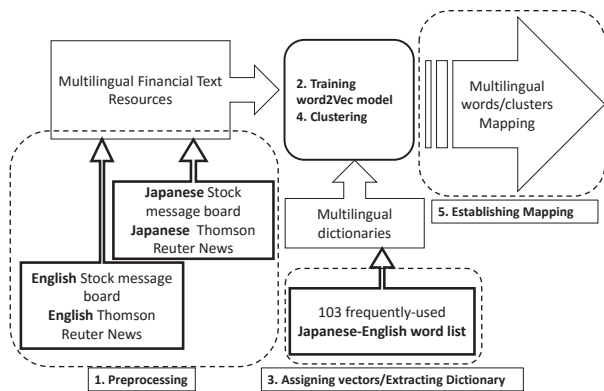


図 1: The data flow and framework for bi-graph clustering and structure extraction

from the database of Thomson Reuters ¹ of year 2010.

2. Tokenization, Tagging and lemmatization are then conducted on both text data in order to acquire more precise word2vec model by only using small scale training data[8]. For English text, we implement StanfordNLP[6] as tagger as well as NLTK[7] as lemmatizer and for Japanese, we employ the MeCab[3] with the neologism dictionary².
3. Finally, we remove unnecessary and meaningless semantic elements including determiners, such as the word *the*, punctuations, conjunctions, and foreign words. Further more, we also remove unsolvable elements such as special characters, http and email address, typo and facial expression in order for better word2vec models through regular expression.

2.2 Word embedding

Google's word2vec is a prevalent tool basing on skip-gram or continuous bag-of-words architecture which provides an efficient way for word embedding[8]. In this research we perform two training for the Japanese news and the English news respectively in 200 dimension.

¹Official websites of Thomson Reuter:
<http://www.reuters.com/>

²Neologism dictionary implementation on Mecab-ipadic:
<https://github.com/neologd/mecab-ipadic-neologd>

2.3 Dictionary Extraction

In this research, we manually choose 103 typical Japanese words frequently appeared in the finance domain. We first translate those words into English, then assign vector representation basing on the word2vec model trained for both English and Japanese words.

In case of phrases that are not detected by the phrase detector, a trick applied here that we first retrieve the vector representation for each word appeared in the phrase and then conduct a summation on these vectors, considering the result as the vector representation for the whole phrase. This is proposed basing on the recent report[4] that word2vec could represent many linguistic regularities, suggesting that the vector addition and subtraction could still be able to represent the relative meaning of the phrase.

2.4 Spherical k-means clustering

We conduct spherical k-means clustering for both Japanese and English financial word list with centroids k to be 10. Concerning word and document vector clustering, it is reported the spherical k-means, which combine the k-means algorithm with cosine similarity as the distance function, could produce a satisfied clustering results[2]. We initialize with k-means++ seeding and overwrite the algorithm[1] with the help of Scikit learn platform³.

A sample of clustering results for cross-lingual financial news reports of Thomson Reuters in both English and Japanese is shown in the Table 1 with the number of cluster centroid $k=10$. Here, we define the notation EN_i^m and JP_j^n , $i, j, m, n \in (1, 2, 3, \dots, k)$ referring to English cluster names and Japanese cluster names respectively where m and n indicate number of centroids we choose during clustering while i and j shows the cluster name in terms of number. Normally, as the the default size of cluster centroids of 10, the superscript $m=10$ and $n=10$ would be omitted.

This results indicate that for each cluster, the spherical kmeans approach indeed clusters words with similar meanings into the same cluster.

2.5 Mapping for cross-lingual clusters

The clustering results in Table 1 also shows that some Japanese words whose English translation are

³Scikit learn, official website:
<http://scikit-learn.org/stable/index.html>

表 1: Example of English words clustering basing on Thomson Reuters news

EN Cluster	EN Word	JP Word	JP Cluster
EN ₅	Improvement	改善	JP ₅
EN ₅	Jump	高騰	JP ₁
EN ₅	Fall	低下	JP ₁
EN ₅	Decline	減少	JP ₉
EN ₉	Decrease of profit	減益	JP ₂
EN ₉	Increase of income	増収	JP ₂
EN ₇	Concern	懸念	JP ₁₀
EN ₇	Risk	リスク	JP ₁₀
EN ₇	Aggravation	悪化	JP ₁₀

clustered into the same cluster, for instance, EN₇, are also categorized into the cluster JP₁₀, indicating that there might be potential relationships between English clusters and Japanese clusters.

Here, we define the concept *common words* for cluster EN_i and JP_j as any translation pair where its English translation belongs to EN_i while its Japanese translation belongs to JP_j, denoted as C_(i,j). According to the Table 1, for example, the Japanese-English pair *Decline* and *Genshou* is *common words* for cluster EN₅ and JP₉.

In order to identify the *cross-lingual cluster similarity* as the similarity among multilingual clusters quantitatively, we here define the *similarity between any two clusters* as:

$$sim(EN_i, JP_j) = \frac{2 \times size[C_{(i,j)}]}{size[EN_i] + size[JP_j]} \quad (1)$$

where C_(i,j) as mentioned previously denotes the the common words of EN_i and JP_j, while the notation size[A] refers to the size of a set A. In this equation, the similarity among clusters could reach 1 when size[C_(i,j)] = size[EN_i] = size[JP_j] whereas it becomes 0 when there is no common words, that is size[C_(i,j)] = 0.

In reality, there might not be a one-to-one mapping for cross-lingual clusters, instead it is also possible to 1-to-N mapping among them. Hence, we adjust the definition of the cluster similarity as:

$$sim(EN_i, JPF_p) = \frac{2 \times size[C_{(i,p)}]}{size[EN_i] + size[JPF_p]} \quad (2)$$

In this equation, considering full combinations of clusters from set JP_i j where j ∈ (1, 2, .., k) and k=10,

表 2: Extended Japanese clusters with maximum cross-lingual similarity respecting to English clusters

English clusters	Most similar cluster	Similarity scores
EN ₁	JP ₆	0.4
EN ₂	(JP ₁ +JP ₂)	0.5
EN ₃	JP ₅	0.4
EN ₄	JP ₁₀	0.32
EN ₅	JP ₁	0.64
EN ₆	(JP ₉ +JP ₇ +JP ₂)	0.49
EN ₇	(JP ₁₀ +JP ₃)	0.30
EN ₈	JP ₄	0.38
EN ₉	JP ₃	0.4
EN ₁₀	JP ₄	0.43

we then could obtain a set, denoted as JPF_p where p ∈ (1, 2, ..., Q). Here, Q is the number of the full combinations, normally calculated as:

$$Q = \sum_{n=1}^k \frac{k!}{n!(k-n)!} \quad (3)$$

To be more specific, the new sets after full combination for Japanese clusters, the JPF_p typically include:

- JP₁, JP₂, ... , JP₁₀
- (JP₁+JP₂) , (JP₁+JP₂) , ... , (JP₉+ JP₁₀)
- (JP₁+JP₂+JP₃) , (JP₁+JP₂+JP₄) , ... , (JP₈+JP₉+ JP₁₀)
- (JP₁+JP₂+...+JP₁₀)

Basing on the equation 2, we could derive the most similar combination of Japanese clusters in the regarding to each English cluster, as shown in the table 2.

3 An application: Cross-lingual Article Comparison

By utilizing the cross-lingual mapping results, we also develop an piratical application that could, when given a English news article, return the most similar Japanese news articles inside the database.

3.1 Framework of cross-lingual article comparison

There are mainly two part of this framework as shown in the Figure 2, three input layers and a Siamese LSTM module.

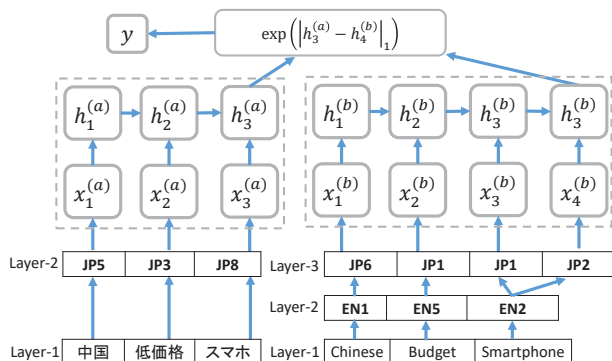


図 2: Framework of cross-lingual article comparison using two-LSTM modules with different embedding layer for the Japanese input and English input respectively.

3.1.1 Siamese LSTM

Siamese Long short-term Memory is a deep learning based machine learning model that could not only memorize the order information of input but also calculate the similarity between the given two text input [5]. This structure is normally applied for the monolingual text similarity task and has been proved efficiency, whereas it has not been applied in the cross-lingual situation so far. In our application, we adopts the same configuration which are claimed to be optimal.

3.1.2 Input layers

Differing from the monolingual case used in the Siamese LSTM module, there are either two or three layers appended for the text input in order to adapt for cross-lingual case, depending on whether it is Japanese text or English text.

1. Layer-1. The tokenized and normalized words are fed in term of fixed-length word2vec representation in the layer-1.

2. From layer-1 to layer-2, the cosine distance between each of the fed word and the centroid of each Japanese clusters will be calculated, and the the cluster with the closest distance is considered to be the representation of this word in the layer-2.
3. Layer-3 is only deployed in case of English input. It will map the cluster number in layer-2 from English domain to Japanese domain basing on the mapping results obtained in the bi-graph extraction section, so that the LSTM module could compare them equally.

3.2 The Data sets

Similar to the process conducted during the extraction of bi-graph structure extraction, we excavate data from Thomson Reuter News. With the help of existed news tags in the database, we extract the English and Japanese news pairs which are both relating to the same news events, which could be considered as similarity of 1. In contrast, any two cross-lingual news reports are defined to be dissimilar with a similar score of 0. In this experiment, we prepare 1000 cross-lingual pairs with similarity of 1 and 1000 pairs with similarity of 0 as training data.

3.3 Model training and evaluations

We perform the training 100 epochs times. In addition, instead of using mean-square-error as evaluation benchmark, we define a practical criteria, the @TOP-N indicator. Given a Japanese article, the comparison model would offer a list that lists all articles ordered from most the similar article to the least similar article. However there is only 1 correct answer with similarity score of 1, and therefore the @TOP-N indicator indicates whether the real similar article is identified among the top N similar articles suggested by the comparison system.

The results are shown in the table 3. As expected, the @TOP-N indicator for the training data is always slightly higher than testing. If we only focus on the testing data, there are about 33 articles are correctly paired within the TOP 10 range, which is not as good as our expectation. We consider this could be due to the high correlation among our test and training financial text.

表 3: The TOP-N score of the cross-lingual LSTM model

	Training Data	Testing Data
@TOP-10	58/1000)	33/1000
@TOP-5	32/1000	20/1000
@TOP-1	4/1000	6/1000

In the previous step for calculating the top-N score, we have obtained the ranking positions of the correct counterpart within the whole database for each of documents. Figure 3 shows this ranking results for training documents and test documents in red and in blue respectively. The more dots close to the top line are (rank = 1), the better the training results are. Though there are large amounts documents still outside the top-10 range, most of them are approaching to the top-1 line compared with the case of the random distribution indicating the effectiveness of our model.

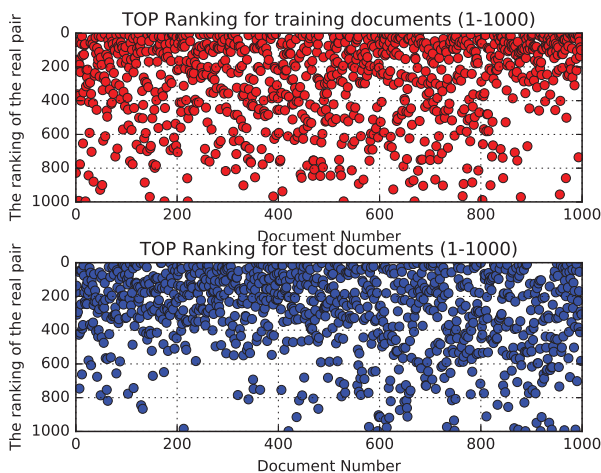


图 3: TOP ranking for training documents (red, above) and test documents (blue, below) respectively

Furthermore, when investigating the learning process of our model, we obtain the learning curve as shown in Figure 4, where the TOP-N (N=1,5,10) accuracy are increasing along with training epoches and become largest TOP accuracy at the 100 times training.

4 Conclusion

This paper give a cross-lingual solution for articles similarity comparison by using extraction of bi-graph structures basing on word2vec, spherical clustering

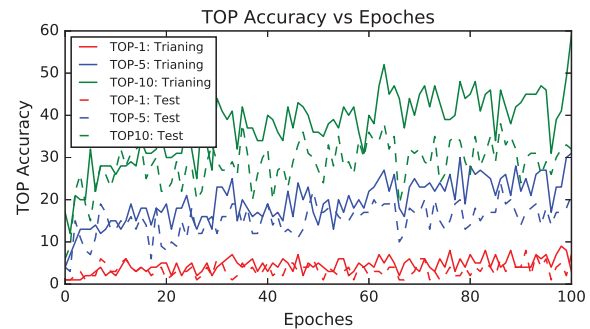


图 4: Learning curve in terms of TOP-1,5,10 accuracy regarding to learning epoches

and Siamese LSTM with multilingual news reports from Thomson Reuters. We evaluate our model with finance news data and prove the effectiveness of our model, though the accuracy still needs to be improved in the future work. As the final goal, this kind of model is expected to assist the financial analyst in their prediction model.

参考文献

- [1] BANERJEE, Arindam, et al. Clustering on the unit hypersphere using von Mises-Fisher distributions. In: *Journal of Machine Learning Research*. 2005. pp. 1345-1382.
- [2] DHILLON, Inderjit S.; GUAN, Yuqiang; KOGAN, J. Refining clusters in high-dimensional text data. In: *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at the Second SIAM International Conference on Data Mining*. 2002. pp. 71-82.
- [3] KUDO, Taku; YAMAMOTO, Kaoru; MATSUMOTO, Yuji. Applying Conditional Random Fields to Japanese Morphological Analysis. In: *EMNLP*. 2004. pp. 230-237.
- [4] MIKOLOV, Tomas; YIH, Wen-tau; ZWEIG, Geoffrey. Linguistic Regularities in Continuous Space Word Representations. In: *Hlt-naacl*. 2013. pp. 746-751.
- [5] MUELLER, Jonas; THYAGARAJAN, Aditya. Siamese Recurrent Architectures for Learning Sentence Similarity. In: *AAAI*. 2016. pp. 2786-2792.

- [6] TOUTANOVA, Kristina, et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003. pp. 173-180.
- [7] TAGHVA, Kazem; ELKHOORY, Rania; COOMBS, Jeffrey. Arabic stemming without a root dictionary. In: *Information Technology: Coding and Computing, 2005. ITCC 2005*. International Conference on. IEEE, 2005. pp. 152-157.
- [8] MIKOLOV, Tomas, et al. *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013.
- [9] VU, Tien-Thanh, et al. An experiment in integrating sentiment features for tech stock prediction in Twitter. In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*. The COLING 2012 Organizing Committee, Mumbai, India, 2012. pp. 2338.

銘柄固有の金融極性辞書の構築

Constructing Financial Sentiment Lexicons for Individual Stocks

関 和広^{1, 2*} 柴本昌彦²
Kazuhiro Seki^{1,2} Masahiko Shibamoto²

¹ 甲南大学 知能情報学部

Faculty of Intelligence and Informatics, Konan University

² 神戸大学 経済経営研究所

Research Institute for Economics & Business Administration, Kobe University

Abstract: This paper reports on our ongoing work to construct sentiment lexicons in the financial domain. Our approach takes advantages of news headlines and a given financial variable, such as stock prices, so as to generate initial sentiment lexicons. The initial lexicons are then filtered based on their co-occurrences with financial seed words and are subsequently expanded by analogical reasoning by using distributed representation of words. Evaluative experiments on around 12 years' worth of news data show that the resulting lexicons are mostly reasonable. As a possible application of the lexicons, trading simulation is also carried out, showing promising results.

1 まえがき

近年、マイクロブログやユーザレビュー、質問応答型の様々なフォーラムなど、誰もが即座に利用可能な Consumer Generated Media (CGM) のプラットフォームが急速に普及し、世界中で利用されている。そして、これらのプラットフォームに投稿される情報は、商品の購入や訪れる店舗の選択といったユーザの多様な意思決定に日常的に影響を及ぼしている。このような背景から、WWW 上に溢れるテキスト情報を有効に活用するためのテキストマイニングの研究が盛んに行われている [Imran 15].

テキストマイニングの技術の一つとして、分析対象のテキストに表現される書き手の感情を抽出、あるいはスコア化する「感情分析」がある [Liu 15]. そして、感情分析を行う際に有用な言語資源として、極性辞書や感情表現辞書がある。極性辞書は肯定・否定という二極分化的な観点から構築され、感情表現辞書は、喜びや悲しみや怒りといった複数の感情の種類を区別して構築される。これらの辞書では、肯定や否定、あるいは喜びや悲しみといった感情の強弱を考慮して、辞書に収録される各単語に感情を点数化して付与していることが多い [Takamura 05].

一方、金融・経済の分野でもテキストマイニングの研

究が長年行われてきている。たとえば、ニュース記事を分析して株価などの金融指標を予測する研究がある [Schumaker 09]. これらの研究の手法は2つのアプローチに大別することができる。1つ目のアプローチでは、ニュース記事等のテキストと目的変数（金融指標）の関係を過去のデータから Support Vector Regression などの何らかの回帰モデルで学習し、予測を行う。もう一方のアプローチでは、予測対象とする金融指標に関してテキストが肯定的な内容を含んでいるか、否定的な内容を含んでいるかを分析し、前者であれば金融指標が上昇、後者であれば下降するものと予測する。このアプローチは前述の極性辞書を用いた感情分析と類似している。

本研究では、金融テキストマイニングにおける第一歩として、金融指標の短期的な変動に関わる語を収録した金融極性辞書の構築を目指す。なお、極性辞書に含むべき語は、個別銘柄の株価など、対象とする金融指標によって異なる可能性があるため、対象とする個々の指標に応じた個別の辞書を構築する。辞書の構築は、過去のニュース記事と金融指標の変動との関係を利用して、自動的に行う。さらに、構築した辞書を用いて金融指標の短期的な変動を予測し、その結果に基づいて行った投資シミュレーションの結果についても報告する。

本論文の構成は以下の通りである。まず、2節で感情分析や株価予測の関連研究を紹介する。続いて、3節

*連絡先：甲南大学知能情報学部知能情報学科
〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1
E-mail: seki@konan-u.ac.jp

で提案手法の詳細を述べ、4節で実データを用いて行った評価実験について報告する。最後に、5節でまとめと今後の課題について述べる。

2 関連研究

2.1 極性辞書の自動構築

自動的に極性辞書を構築する方法は、辞書を用いる方法とコーパスを用いた方法に大別できる。辞書を用いた方法では、WordNet [Fellbaum 98] のようなシソーラスを用いて、極性が既知の語の関連語に極性を伝播させることによって極性辞書を構築する。なお、以降では極性が既知の語を「種表現 (seed word)」と呼ぶ。例えば、「important」が肯定的な種表現であるとき、シソーラスを参照することで、その同義語である「essential」「momentous」等は肯定的、反意語である「immaterial」「fiddling」「lilliputian」等は否定的な表現であるとみなす。実際には、シソーラスを語のネットワークとして見たときの語間の距離や、語の定義文 (gloss) の類似度を考慮して極性を伝播させることが多い [Baccianella 10, Kamps 04]。辞書を用いた方法の欠点としては、辞書あるいはシソーラスに含まれる語しか扱うことができないこと、ドメインを区別しない一般的な極性辞書しか構築できないことが挙げられる。

コーパスを用いた手法でも、辞書による方法と同様に、種表現を用いて極性が未知の語の極性を推定する。推定の手がかりとしては、大規模なコーパスに現れる語の言語的特徴や統計的特徴を利用する。前者の例としては、接続詞を用いる方法が提案されている [Hatzivassiloglou 97]。この方法は、「but」のような逆接の接続詞の前後の語は逆の極性を持つことが多く、「and」のような順接の接続詞の場合は同じ極性を持つことが多いことに着目する。たとえば、effective が肯定的な種表現であったとき、「effective but costly」という句からは costly が否定的な表現であること、「efficient and effective」からは efficient が肯定的な表現であることが推測できる。

コーパスを用いた統計的な手法としては、共起関係を用いた方法 [Turney 02] がよく利用される。この方法では、少数の種表現を用いて、極性が不明な語 w とそれら種表現との共起頻度を計数する。そして、共起頻度を基に肯定的な種表現と w の自己相互情報量、および否定的な種表現と w の自己相互情報量を算出し、その差を極性スコアと定義する。そして、このスコアの符号と絶対値の大きさをそれぞれ極性の方向 (肯定対否定) と強さとする。

また最近では、語の分散表現 [Mikolov 13a] を用いた手法がいくつか提案されている。分散表現とは、ニュー

ラルネットワーク言語モデルを構築する際に副次的に獲得される語のベクトル表現であり、単語類推問題¹などに応用できる。このような分散表現を利用し、たとえば佐藤ら [佐藤 16] は、分散表現を素性として語の極性を出力する分類器を学習することを提案している。ただしこの手法では、極性が既知な語が学習データとして相当数必要となる。

2.2 金融ドメイン極性辞書

極性辞書に収録されるべき単語やその極性は、対象ドメインや記述の対象によって異なる [Bross 13]。たとえば、ホテルのレビューを対象に感情分析を行うのであれば、(ベッドや風呂などが)「小さい」は否定表現であるのに対して、ノートパソコンなどの電化製品に関するレビューでは同じ語が肯定表現になりうる。本研究では、金融ドメインを対象とするため、一般ドメイン向けの極性辞書では適切な感情分析は難しい [Loughran 11]。

金融ドメインについては、Loughran and McDonald [Loughran 11] によって人手で構築された英語の極性辞書がいくつかの研究で利用されている [Jegadeesh 13, Lee 14, Wang 13]。極性辞書の自動構築については、日本語の文章を対象とし、Hatzivassiloglou ら [Hatzivassiloglou 97] と類似の言語的特徴を利用した手法が試みられている [前川 13]。また、前述のような分散表現を用いた手法もいくつか試みられている [片倉 15, 伊藤 16]。

以上で述べた関連研究と比較し、本研究で提案する手法は、種表現がほとんど必要なく人手で収集する必要もない点、テキストデータに加えて金融指標の時系列データも用いている点、個別銘柄の株価など金融指標固有の極性辞書を構築できる点に特長がある。

3 金融極性辞書の構築

3.1 概要

本研究では、以下の手順で、対象とする金融指標固有の極性辞書を構築する。

1. 金融指標の変動を利用した極性表現候補の抽出
2. 極性表現候補のフィルタリング
3. 反意語の獲得

以下、個々の処理について詳述する。

¹ 「フランスに対するパリ」の関係から、「日本に対する x 」の x を推定するような問題 (正解は東京)。

3.2 金融指標の変動を利用した極性表現候補の抽出

ある時間 t にニュース記事 d が配信され、その後 $t+\Delta t$ に金融変数 y が $\theta_y\%$ 以上変動したとき、ニュース記事 d はこの変動に影響を与えたと仮定する。もちろん、ニュースでは報道されない他の様々な要因も考えられるため、この仮定は必ずしも成り立たない。しかし、金融指標の変動に影響を与えた記事を正確に同定することは困難であるため、本研究ではこの単純な仮定を用いて議論を進める。

上述の仮定において、時間窓の幅 Δt と金融指標の変動の幅 θ_y を適当な値に設定すれば、目的変数 y の正方向への変動の要因となった記事集合 D^+ 、および負方向への変動の要因となった記事集合 D^- を自動的に同定することができる。これらの記事集合 $D = D^+ \cup D^-$ を用いて、 y の変動の要因となる記事に特徴的な語を見つける。

具体的には、記事に現れる各語 w について式 (1) のカイ二乗値を算出し、この値が高いほど特徴的な語であると判断する。

$$\chi^2(w) = \frac{|D| \times (n_w^+ n_w^- - n_w^+ n_w^-)^2}{(n_w^+ + n_w^-)(n_w^+ + n_w^-)(n_w^+ + n_w^-)(n_w^+ + n_w^-)} \quad (1)$$

ここで、 $|D|$ は総記事数を表し、 n_w^+ は n^+ のうち w を含む記事数、 n_w^- は n^- のうち w を含む記事数、 n_w^+ は n^+ のうち w を含まない記事数、 n_w^- は n^- で w を含まない記事数とする。

なお、この指標は D^+ と D^- のいずれかに特徴的な語であれば大きい値を取る。よって、カイ二乗値だけではどちらに特徴的な語か判断できないため、 D^+ と D^- における語 w の相対頻度を比較して、どちらに特徴的な語であるか判定する。そして、 D^- に特徴的な語である場合は、便宜上、カイ二乗値にマイナスの符号を付して区別する。このようにして算出されたカイ二乗値の絶対値が閾値 θ_χ より大きい語を「極性表現候補」と呼ぶ。

3.3 極性表現候補のフィルタリング

前節で得られた極性表現候補には、目的変数 y 、あるいは金融指標一般の変動とは関わりがないと思われるような語も含まれてしまう。このような不適切な語を除去するため、コーパスを利用した極性辞書の構築でよく用いられる指標 [Turney 02] を用いる。具体的には、式 (2) に示す自己相互情報量 I の差 δ を算出する。

$$\begin{aligned} \delta(w) &= I(w, seed^+) - I(w, seed^-) \\ &= \log_2 \frac{h(w, seed^+)h(seed^-)}{h(w, seed^-)h(seed^+)} \quad (2) \end{aligned}$$

ここで $h(\cdot)$ は、引数 (語) の出現文書数を表す。 $\delta(w)$ の絶対値が大きいほど、種表現 $seed^+$ あるいは $seed^-$ と語 w の共起がいずれかに偏っていることを意味し、従来研究ではこの値が極性値として使われている。本研究では、 $\delta(w)$ の絶対値がある閾値 θ_δ 以上の場合のみ、語 w は金融指標の変動と関連した文脈で使われることが多いと判断し、極性表現と見なす。

3.4 反意語の獲得

前節で得られた極性表現は、抽出・フィルタリングの過程を経て獲得されるため、あまり多くの表現は得られない。そこで、得られた表現を基にこれを拡張する。表現の拡張には、語の統語的・意味的特徴を表現できる分散表現 [Mikolov 13b] を用いる。具体的には、単語類推問題と同様の方法で、前節で獲得した極性表現 w の反意語 w' の候補を類推する。すなわち、「 $seed^+$ に対する $seed^-$ 」あるいは「 $seed^-$ に対する $seed^+$ 」は「 w に対する w' 」であると考え、それぞれの語に対応する分散表現の加減算によって得られたベクトルと類似した分散表現を持つ語 (本研究では上位 20 語) を w' の候補と考える。ベクトル間の類似度としては、コサイン類似度を用いる。

さらに、この過程で同意語が反意語として獲得されてしまうことを避けるため、3.3 節で利用した自己相互情報量の差 (δ) の符号の変化に着目する。具体的には、語 w についての $\delta(w)$ と、類推に基づいて得られた語 w' についての $\delta(w')$ が異符号のとき反意語、同符号のときには同意語と判断する。なお、3.3 節と同様に $|\delta(w')| \geq \theta_\delta$ が成り立つ w' のみを獲得の対象とする。

4 評価実験

3 節で述べた提案手法の有効性を検証するため、実データを用いて評価実験を行った。評価実験は、得られた極性辞書の定性的評価、および極性辞書を用いた株価動向予測による投資シミュレーションからなる。以下、4.1 節で実験設定について述べ、4.2 節で極性辞書の評価、4.3 節で投資シミュレーションの結果について報告する。

4.1 実験設定

4.1.1 データ

評価実験に用いたデータは、Thomson Reuters News (TRN) アーカイブの 2003 年 1 月 1 日～2014 年 6 月 30 日分 (11 年 6 ヶ月分) の日本語ニュース記事と、その期間に対応する株の売買を記録した歩み値 (ティッ

配信日時	2013-01-04 00:28:22
見出し	ホットストック：自動車株が軒並み高、円安加速で収益改善期待強まる
本文	[東京 4日 ロイター] 自動車株が軒並み高となっている。ドル／円が87円台後半まで上昇するなど...
関連銘柄	7203.T 7261.T 7267.T 7270.T
...	...

図 1: Thomson Reuters News アーカイブ収録の記事の一部 (記事 ID: nTK0584247)。

ク) データである。TRN のニュース記事には、記事配信の日時、記事見出し、記事本文、記事に関連する銘柄のコードなどが記載されている (図 1)。評価実験では、銘柄コードの情報に基づいて特定の企業に関するニュース記事を抽出し、その企業の株価の動きと対応を取ることで株価の変動に影響を与えるニュース記事 ($D^+ \cup D^-$) を同定した。

本節の実験では、特に他の記載がない場合、TRN の 2003~2012 年 (10 年分) の記事の見出しを極性表現候補の抽出に利用し、残り 1.5 年分のデータを投資シミュレーションに用いた。TRN の記事見出しは形態素解析器 MeCab²で解析し、(複合) 名詞と動詞だけを極性表現候補として抽出した。なお、株価増加に影響を与えた記事集合 D^+ 、減少に影響を与えた記事集合 D^- を決定するための時間窓の幅 Δt は 30 分、株価変動の幅 θ_y は $\pm 1\%$ とした。また、カイ二乗値による極性表現候補の抽出時の閾値 θ_x は 2.706 (有意水準 10%)、自己相互情報量の差に関する閾値 θ_δ は 3.0 とした。seed⁺ には「売り」、seed⁻ には「買い」をそれぞれ用いた。なお、ニューラルネットワーク言語モデル CBOW [Mikolov 13a] の構築には gensim³ を利用し、TRN の 2003~2012 年の 10 年分の記事の見出しと記事本文を入力とした。パラメタ値については、語の最低出現頻度を 10 とした他はデフォルトの設定を用いた。

4.1.2 投資戦略

自動的に構築された金融極性辞書の応用例の一つとして、単純な投資シミュレーションを行った。ニュース記事 d が配信された際、その記事の関連銘柄 $code$ についての投資の判断は、次式の極性平均値に基づいて行った。

$$s_{avg}(code, d) = \frac{1}{|d|} \sum_{w \in d} s(code, w) \quad (3)$$

ここで $s(code, w)$ は $code$ 固有の辞書を基に w の極性値を返す関数であり、肯定表現であれば +1、否定表現で

²<http://taku910.github.io/mecab/>

³<https://radimrehurek.com/gensim/>

あれば -1 とした。code の辞書に w が含まれていない場合は、全銘柄データから作成した辞書を基に極性値を返すものとした。| d | は記事 d の長さ (語数) である。

そして、 $s_{avg}(code, d)$ の絶対値が閾値 θ_s (本実験では 0.1) を超えたとき、この値が正であれば買い、負であれば売りと判断した。売買単位はいずれも 1 株とした。続いて、株価があらかじめ設定した比率 (本実験では 1%) だけ変動した際に反対売買を行い、利益確定あるいは損切りした。

4.2 極性辞書の構築

3 節で述べた方法で、全銘柄のニュース記事を対象に構築した銘柄共通の極性辞書を表 1 に示す。なお、時間の都合上、本節の実験では、TRN の 2013 年の記事だけを用いた。結果を見ると、「下方修正」「売り」「公募増資」「赤字」など、株価の低下につながるような語が負のカイ二乗値を持ち、否定的な語として獲得されている。一方、「上方修正」「好感」「自社株買い」など、株価の上昇につながるような語は正の値を示しており、肯定的な語として獲得されている。この結果から、ニュース記事の変動とその後の実際の株価との間には、直感に合致するような関連があることが見て取れる。

さらに、分散表現を用いた関連語の同定と相互情報量の差の符号変化に基づく極性判定により、「下方修正」に対する「上方修正」、「売り」に対する「買い」、好感に対する「嫌気」などの語が獲得されていることが確認できる。また、同意語としても「売り」に対する「売り越し」、「買い」に対する「買われる」などが獲得されている。また、同意語ではないものの、「株式分割」から「自社株買い」や「消却」などのいずれも肯定的な語が獲得できていることが分かる。一方で、「売り」の同意語として「買い越し」など、本来は極性が反対である語を獲得してしまっている場合もあり、さらなる改善が必要である。

表 1 の辞書は、東証一部上場のすべての銘柄に関するニュース記事と株価を利用し、銘柄を区別せずに構築している。しかしながら、同じ表現であっても、業種や個別の企業によってそれぞれの株価への影響は異なると考えられる。たとえば、為替相場が業績にどのように影響するかは、輸出産業か輸入産業かなどで異なる。前者の場合、円高が進めば海外での売り上げが日本円では目減りするため、減益につながる。すなわち、同じ「円高」という語でも、株価の上昇につながる肯定的な語と捉えられるか、否定的な語として捉えられるかは注目する銘柄によって異なる。

本研究の手法では、入力に用いる金融指標のデータを変更すれば、その金融変数に応じた極性辞書が自動的に構築される。そこで、提案手法によって銘柄ごと

表 1: 共通金融極性辞書構築の結果. 否定的な表現のカイ二乗値には便宜的に負号を付している (3.3 節参照).

極性表現 w	χ^2	δ	極性表現 w'	δ
下方修正	-89.55	-3.76	上方修正	3.81
売り	-39.16	-4.38	買い	4.21
			買い越し	-3.22
			換金	-3.04
			売り越し	-5.07
			売りで	-3.09
			押され	-4.20
公募増資	-12.20	-3.48	自社株買い	3.33
			株式分割	4.81
			消却	3.14
赤字	-9.20	-4.08	黒字	3.14
自社株	16.18	4.29	公募増資	-3.48
			自社株買い	3.33
			株式分割	4.81
			消却	3.14
上方修正	30.33	3.81	下方修正	-3.76
好感	75.64	5.16	嫌気	-3.56
自社株買い	148.77	3.33	公募増資	-3.48
			株式分割	4.81
株式分割	148.86	4.81	公募増資	-3.48
			消却	3.14
			自社株買い	3.33
材料視	223.20	3.56	嫌気	-3.56
買い	800.59	4.21	売り	-4.38
			売りで	-3.09
			売られ	-3.35
			買われる	3.86
			買われ	3.86

に極性辞書を構築し、実際に「円安」あるいは「円高」が逆の極性で獲得された銘柄を抽出して比較した. 表 2 に結果を示す.

表 2: 銘柄による円安・円高の極性の違い.

円安が肯定的	円高が肯定的
キャノン ジェイエフイーホールディングス 安川電機 新日鐵住金 川崎重工業 シャープ 富士通	本田技研工業 (ホンダ) パナソニック

いずれの企業も、一般的に円安メリットを享受すると思われる企業である. しかしながら、ホンダは他の自動車メーカーと比較して海外生産の比率が非常に高く、円安効果がほとんどないと言われている. また、パナソニックは国内での売上比率が高く、大手電機メーカーの中では円高の影響を受けにくい. これらの点から、ホンダ、パナソニックでは円高のニュースの配信後に株価が相対的に上昇した可能性がある.

なお、銘柄固有の極性表現として、トヨタでは「米自動車販売」、ソフトバンクでは「アジア株」、シャープでは「新興国」、東京電力では「見送り」「不透明感」「処分」などの語が獲得されていた.

4.3 投資シミュレーション

4.1.2 節で述べた戦略に従い、投資シミュレーションを行った. 売買が行われた銘柄の異なり数は 1,066 であり、取引総数は $2,367 \times 2$ 回であった. 利益の増減の結果を図 2 の「ind. & common」のグラフに示す. 投資期間の前半で利益が大きく伸び、最終的には約 1.7 百万円の利益が得られた. なお、「individual」で示したもう一方のグラフは、(全銘柄共通の辞書を用いず) 銘柄固有の辞書のみを利用した場合の結果である. 取引回数が減少することもあり、前者ほどの利益の伸びは見られないものの、最終的には約 10 万円程度の利益が得られた. 投資シミュレーションの期間が 18ヶ月と比較的短期であり、この実験の結果を一般化することはできないものの、本研究で自動構築された金融極性辞書が有用である可能性が示された.

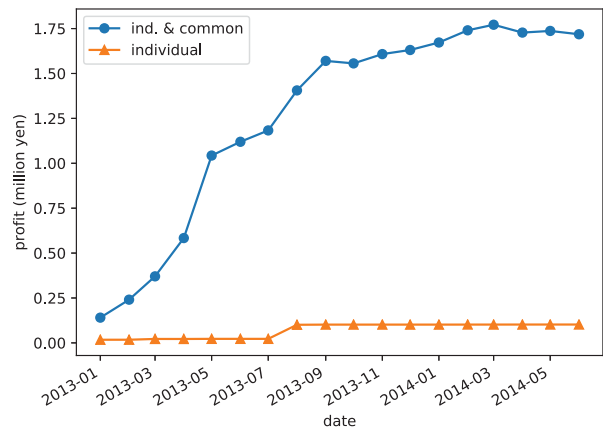


図 2: 投資シミュレーションの結果.

参考まで、投資シミュレーションで特に利益幅が大きかった 5 銘柄と小さかった 5 銘柄の個々の結果を表 3 にまとめる.

表 3: 利益幅上位・下位 5 銘柄の投資シミュレーションの結果.

銘柄	取引回数	最終利益 (千円)
プロロジ	8	142
SBI ライフ	2	141
産業ファンド	6	132
ワコム	4	105
日本高純度化	4	84
星野	2	-68
アコモ F	4	-68
ナノキャリア	6	-71
NVC	2	-74
ヴィレッジ V	2	-80

5 むすび

本研究では、ニュース速報コーパスと任意の金融指標を用いて金融極性辞書を構築する枠組みについて報告した。この枠組は、既存のシソーラスや辞書を必要とせず、与えた金融指標に応じた辞書を自動的に構築できる点に特長がある。実際に構築された辞書を確認したところ、「下方修正」「赤字」「嫌気」「公募増資」などが否定的、「上方修正」「黒字」「好感」「自社株買い」などの語が肯定的な語として獲得されており、おおむね直感に合う結果が得られていた。また、企業ごとの辞書を比較したところ、たとえば「円安」という語は企業によって極性が異なっていることが確認できた。また、極性辞書の応用例の一つとして、2013年から1.5年のデータを用いて投資シミュレーションを行ったところ、単純な方法ながらある程度の収益が得られることが示された。

一方、実際には単語（あるいは複合語）単位で極性を判断することはできない場合も多いため、節や述語項関係などより大きい単位での極性辞書を構築する必要がある。また、同一の表現であってもその極性は時間とともに変化すると考えられるため、固定的な極性ではなく、今後、時系列的な極性の変化についても検討していく。

謝辞

本研究の一部は、私立大学等経常費補助金特別補助「大学間連携等による共同研究」、およびJSPS科研費15H05729の助成によるものである。

参考文献

- [Baccianella 10] Baccianella, S., Esuli, A., and Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, in *Proceedings of the International Conference on Language Resources and Evaluation 2010*, pp. 2200–2204 (2010)
- [Bross 13] Bross, J. and Ehrig, H.: Automatic Construction of Domain and Aspect Specific Sentiment Lexicons for Customer Review Mining, in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1077–1086 (2013)
- [Fellbaum 98] Fellbaum, C. D.: *WordNet: an electronic lexical database*, MIT Press (1998)
- [Hatzivassiloglou 97] Hatzivassiloglou, V. and McKeown, K. R.: Predicting the Semantic Orientation of Adjectives, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 174–181 (1997)
- [Imran 15] Imran, M., Castillo, C., Diaz, F., and Vieweg, S.: Processing Social Media Messages in Mass Emergency: A Survey, *ACM Computing Surveys*, Vol. 47, No. 4, pp. 67:1–67:38 (2015)
- [Jegadeesh 13] Jegadeesh, N. and Wu, D.: Word power: A new approach for content analysis, *Journal of Financial Economics*, Vol. 110, No. 3, pp. 712–729 (2013)
- [Kamps 04] Kamps, J., Marx, M., Mokken, R. J., and Rijke, de M.: Using WordNet to Measure Semantic Orientations of Adjectives, in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1115–1118 (2004)
- [Lee 14] Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D.: On the Importance of Text Analysis for Stock Price Prediction, in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1170–1175 (2014)
- [Liu 15] Liu, B.: *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, Cambridge, United Kingdom (2015)
- [Loughran 11] Loughran, T. and McDonald, B.: When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance*, Vol. 66, No. 1, pp. 35–65 (2011)
- [Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)
- [Mikolov 13b] Mikolov, T., Yih, W., and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, pp. 746–751 (2013)

- [Schumaker 09] Schumaker, R. P. and Chen, H.: Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System, *ACM Transactions on Information Systems*, Vol. 27, No. 2, pp. 12:1–12:19 (2009)
- [Takamura 05] Takamura, H., Inui, T., and Okumura, M.: Extracting semantic orientations of words using spin model, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pp. 133–140 (2005)
- [Turney 02] Turney, P. D.: Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424 (2002)
- [Wang 13] Wang, C.-J., Tsai, M.-F., Liu, T., and Chang, C.-T.: Financial Sentiment Analysis for Risk Prediction, in *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 802–808 (2013)
- [伊藤 16] 伊藤 友貴, 坪内 孝太, 山下 達雄, 和泉 潔 : ニュース記事を用いた経済専門用語のクラスタリングと極性付与, 第 30 回人工知能学会全国大会論文集 (2016)
- [佐藤 16] 佐藤 貴俊, 高村 大也, 奥村 学 : 分散表現を用いた単語の感情極性抽出, 情報処理学会研究報告自然言語処理 (NL) , Vol. 2016-NL-228, No. 12, pp. 1–6 (2016)
- [前川 13] 前川 浩基, 中原 孝信, 岡田 克彦, 羽室 行信 : 大規模ニュース記事からの極性付き評価表現の抽出と株価収益率の予測, オペレーションズ・リサーチ: 経営の科学, Vol. 58, No. 5, pp. 281–288 (2013)
- [片倉 15] 片倉 賢治, 高橋 大志 : 金融市場ニュースの分散表現学習による辞書作成と金融市場分析, 第 29 回人工知能学会全国大会論文集 (2015)

株主招集通知における議案別の開始ページの推定

高野海斗^{1*} 酒井浩之¹ 坂地泰紀¹ 和泉 潔²
岡田奈奈³ 水内利和³

Kaito Takano¹ Hiroyuki Sakai¹ Hiroki Sakaji¹ Kiyoshi Izumi²
Nana Okada³ Toshikazu Mizuuchi³

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 東京大学大学院 工学系研究科

² School of Engineering, The University of Tokyo

³ 株式会社日経リサーチ

³ NIKKEI RESEARCH INC.

Abstract: In this research, we aim to predict start pages of proposals stated in notice of the meeting of shareholders and classify which proposal the page is. We propose two methods that classification method of proposals. The first method heuristically predicts the page on which the proposal is described. Moreover our method extracts specialized terms of each proposal and assigns weights to them. After that, our method classifies proposals by specialized terms. The second method classifies proposals using deep learning. Each methods were evaluated, and the effectiveness of each methods was verified.

1 はじめに

人工知能分野の手法や技術を、金融市場における様々な場面に応用することが期待されており、例えば、膨大な金融情報を分析して投資判断の支援を行う技術が注目されている。その一例として、日本銀行が毎月発行している「金融経済月報」や企業の決算短信、経済新聞記事をテキストマイニングの技術を用いて、経済市場を分析する研究などが盛んに行われている [1][2][3][4][6]。

日経リサーチでは、有価証券報告書などの公開資料を収集し、必要な箇所のデータを抽出する作業を行っている。データ作成にあたっては、例えばXBRL形式のように値に付与されたタグ様の付加情報を利用し、作成しているものもあるが、データ分類用付加情報が付与されているデータはまだ少数で、人手による作業が大半を占めている。

手作業で必要な情報を抽出するには、専門的知識や経験が必要となる。人工知能分野の手法や技術を用いることで、正確性を担保しつつ、データを作成時間の短縮化をめざしている。そのための一環として、まずは株主招集通知に掲載されている議案の開始ページの

推定を行う。従来は抽出したい議案（「取締役選任」「剰余金処分」などの十数区分の項目）が報告書のどのページに記載されているか人手により確認し、データを作成していたが、各社で報告書のページ数や議案数が異なるため、確認に時間を要していた。抽出したい議案がその報告書にあるのか、どのページに記載されているのかが自動で推定できれば、業務の効率化につながる。具体的には、株主招集通知の各ページが議案の開始ページであるかそうでないかを判別し、さらに、開始ページであると判断されたページに記述されている議案が、どのような内容の議案であるかを分類する。

関連研究としては白田らが、日本銀行政策委員会金融政策決定会合議事要旨のテキストデータから、トピック抽出の研究を行っている [7]。また酒井らは、決算短信PDFから業績要因の抽出の研究を行っている [5]。それらの研究を踏まえ本研究では、株主招集通知のデータを扱う点や、トピックや業績要因の抽出ではなく、議案の開始ページの推定とその議案分類をする点が異なっている。

*連絡先：成蹊大学 理工学部 情報科学科
〒180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: dm166208@cc.seikei.ac.jp

2 提案手法 1 特徴語による議案分類

本章では、特徴語による議案分類について説明する。この提案手法は以下の3つのステップで議案の分類を行う。

- Step 1: 議案ごとの特徴語の獲得
- Step 2: 議案がある開始ページの推定
- Step 3: 議案の分類

2.1 議案ごとの特徴語の獲得

2.1.1 特徴語候補の抽出

株主招集通知に出現する議案を分類するために、議案ごとの特徴語の抽出をする。例えば、「取締役選任」の特徴語として、「現任取締役」のような語を抽出する。議案ごとの特徴語の獲得をするために、2016年4月から8月までの株主招集通知における議案別の開始ページとその議案の分類が記述されたデータ(4,729件)を学習データとして使用する。特徴語の抽出は上記の学習データを形態素解析し、それから以下の条件のもと、各分類議案の開始ページに2回以上出現する語を特徴語の候補とする。

- 条件1 名詞を対象
- 条件2 分割は N-gram 単位
- 条件3 25文字以上の長すぎる複合名詞は除外

2.1.2 特徴語候補への重み付け

特徴語の候補 n_i に対して分類ごとに重み付けを行い、特徴語を選択する。重み付けの式には式1を用いる。

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

ここで、学習データにおいて、

$C(t)$: 議案分類 t の開始ページの文書集合。

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

$H(n, C(t))$: $C(t)$ の各文書である 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。以下の式2によって求める。

$$H(n, C(t)) = - \sum_{d \in C(t)} P(n, d) \log_2 P(n, d) \quad (2)$$

ここで、 $P(n, d)$ は d に名詞 n が出現する確率である。

$df(n)$: 名詞 n を含む文書の数。

N : 学習データにおける文書の総数。

エントロピーを用いた理由は、各議案分類の文書集合中で多くの文書に分散して出現している語の方が、少数の文書に集中して出現している語と比較して、よりその議案分類の特徴を表し、特徴語としても有効であるという仮定に基づく。

2.1.3 特徴語の選択

各議案分類ごとの特徴語候補の重み付けの平均値を算出し、平均値よりも重みの高いものの特徴語として選択する。すなわち、以下の条件が成り立つ語 n_i を特徴語として選択する。

$$W(n_i, C(t)) > \frac{1}{m} \sum_{j=1}^m W(n_j, C(t)) \quad (3)$$

m : 議案分類 t の特徴語候補の総数。

例えば、取締役選任の特徴語の一部を表1に示す。

表1: 取締役選任の特徴語

特徴語	重み
リーダーシップ	18.13
当社代表取締役社長就任	18.00
現任取締役	17.85
在任取締役	17.49

2.2 議案がある開始ページの推定

議案がある開始ページは、以下の条件のもと推定した。

- 条件1 議案がある開始ページには、「議案」または「第 X 号議案」が先頭に含まれるページを対象とする
- 条件2 目次ページが存在するため、目次ページは議案がある開始ページから除外する
- 条件3 参考ページ¹以降の開始ページ推定の対象とする

¹株主招集通知において、議題について書かれている最初のページを参考ページと呼ぶことにする。多くの場合、第1号議案の開始ページが参考ページとなる。

「決議事項」または「目的事項」という表現が含まれているページを 目次ページ し、「参考書類」、「議題及び参考事項」または「議題および参考事項」という表現が含まれているページを 参考ページ とする。

「第 X 号議案」という語が含まれている場合、その株主招集通知に含まれる議案数は「 X 」の最大値である。また「第 X 号議案」という語が含まれていない場合、議案という語が含まれていれば、議案数は 1 であると推定できる。これらの情報を基にページの推定を以下のように行う。

Step 1: 保存されている参考ページの中で最も後ろのページから最後のページまでに、「第 1 号議案」から「第 N 号議案」の順に出現するかどうかを調べる。

Step 2: 出現しない場合は、保存されている参考ページの一つ若いページを用いて、同様の探索を行う。

Step 3: 参考ページが目次ページ以下になった場合、目次ページの次のページから同様の探索を行い、それでも見つからない場合は議案はないものとする。

2.3 議案の分類

2.1 節で得られた重み付けと 2.2 節で推定した議案の開始ページを用いて、開始ページごとの議案の分類を行う。議案分類 t の開始ページ j に対するスコア付与は式 4 を用いる。

$$score(j, t) = \frac{\mathbf{V}(t) \cdot \mathbf{V}(j)}{|\mathbf{V}(t)| |\mathbf{V}(j)|} \quad (4)$$

ここで、

$\mathbf{V}(t)$: 議案分類 t の特徴語を要素、特徴語の重みを要素値とするベクトル

$\mathbf{V}(j)$: 開始ページ j の名詞 N-gram を要素、出現数を要素値とするベクトル

複数の議案が同ページに存在する場合、スコアが上位のものから順に選ばれるものとする。

3 提案手法 2

深層学習による議案分類

提案手法 1 では、学習データから各議案の特徴語を抽出し、それに基づいて議案を分類している。この学習データを使用すれば、機械学習手法に基づく手法でも議案分類が可能である。そこで、本研究では深層学習を用いた議案分類も試みた。

3.1 素性選択

株主招集通知に記載されている議案の開始ページの議案分類を、深層学習により行う。すなわち、議案の開始ページが、ある議案分類であるかそうでないかを判別する分類器を議案分類の数だけ生成し、テストデータとなる議案の開始ページがどの議案分類に属するかを判定する。従って、例えば議案分類「取締役選任」を判別するための学習データは、「取締役選任」の開始ページが正例、それ以外の議案分類の開始ページが負例となる。また、テストデータは、学習データとして使用した株主招集通知を除き、株主招集通知を 1 ページごとに分割したうえで、「号議案」の文字列が含まれているページとした。

まず、入力層の要素となる語（素性）を選択する。具体的には、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、以下の式 5 にて重みを計算する。

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

ただし、

S_p : 学習データにおいて正例に属する文の集合

$TF(t, S_p)$: 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語であることが分かる。 $H(t, S_p)$ は次の式 6 で求める。

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

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \quad (7)$$

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、負例に含まれる内容語（名詞、動詞、形容詞）に対しても、同様に重みを計算する。

$$W_n(t, S_n) = TF(t, S_n)H(t, S_n) \quad (8)$$

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。

ここで、ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ 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が成り立つ語 t を素性として選択する。

$$W_p(t, S_p) > W_n(t, S_n) \quad (9)$$

$$W_n(t, S_n) > 2W_p(t, S_p) \quad (10)$$

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、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。以下に議案分類「取締役選任」を判別するための学習データから選択された素性の一部を例示する。

取締役, 監査, 議案, 配当, 株主, 社外, 変更, 事業, 代表, 現任, 責任, 部長, 社長

上記の学習データでは、2,845 語が素性として選択された。

3.2 モデル

入力は、学習データから抽出された語（素性）を要素、語 t における $\log(W_p(t, S_p))$ 、もしくは、 $\log(W_n(t, S_n))$ 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（すなわち素性の数）と同じとし、隠れ層は、ノード数 1000 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。

4 評価

本手法を実装した。学習データとして、2016 年 4 月から 8 月までの株主招集通知から、人手にて議案開始ページとその分類を作成し使用した（学習データ数は 4,729 件）。実装にあたり、形態素解析器として MeCab² を使用した。

評価において、正解データとして、2016 年 9 月から 10 月までの株主招集通知から、人手にて議案開始ページとその分類を作成した（正解データ数は 345 件）。次に、正解データと同じ 9 月から 10 月までの株主招集通知から、各手法を用いて議案開始ページとその議案分類を推定した。表 2 はある企業の株主招集通知における正解データと提案手法 1 による議案開始ページと議案分類の推定結果を示す。そして、各手法の推定結果と正解データが一致すれば正解とし、議案ごとの適合率、再現率、F 値を算出した。評価結果を表 3 に示す。表 3 の手法 1 は提案手法 1 の特徴語による議案分類による手法であり、手法 2 は提案手法 2 の深層学習による手法であることを示す。

²<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

表 2: 提案手法 1 による議案開始ページと議案分類の推定結果とその正解データ

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
34	剰余金処分	34	剰余金処分
35	定款変更	35	定款変更
37	取締役選任	37	取締役選任
38	監査役選任	38	監査役選任
39	退職慰労金	39	退職慰労金

表 3: 評価結果

議案分類	手法	適合率 (%)	再現率 (%)	F 値
会計監査人選任	手法 1	100.00	100.00	1.000
	手法 2	100.00	100.00	1.000
監査役選任	手法 1	72.73	96.00	0.828
	手法 2	85.19	90.20	0.876
企業再編	手法 1	100.00	50.00	0.667
	手法 2	50.00	25.00	0.333
ストックオプション	手法 1	81.25	100.00	0.897
	手法 2	61.11	84.62	0.710
退職慰労金	手法 1	52.63	100.00	0.69
	手法 2	90.91	100.00	0.952
定款変更	手法 1	94.74	94.74	0.947
	手法 2	94.44	89.47	0.919
取締役選任	手法 1	95.79	87.27	0.910
	手法 2	95.79	82.73	0.888
役員賞与	手法 1	83.33	83.33	0.833
	手法 2	100.00	100.00	1.000
役員報酬	手法 1	84.21	94.12	0.889
	手法 2	100.00	66.67	0.800
剰余金処分	手法 1	93.06	93.06	0.931
	手法 2	96.97	88.89	0.928

5 考察

提案手法と深層学習を比較すると、会計監査人選任の項目の分類推定は、両手法ともに良好な結果を示している。ストックオプション、定款変更、取締役選任、役員報酬、剰余金処分の項目は提案手法 1 が良好な結果が得られ、監査役選任、退職慰労金、役員賞与の項目は提案手法 2 によって良好な結果が得られた。提案手法 2 の適合率は全体的に高い傾向にあるが、これは「号議案」の文字列が含まれているページに限定して分類を行っているため、議案数が 1 つしかない株主招集通知は分類ができていないためである。よって、全体的に再現率が低くなってしまっている。提案手法の分類推定が誤っていたものを確認したところ、誤分類は 46 件だった。その誤分類の詳細を確認したところ、取締役選任の項目が、監査人選任の項目に誤分類されている件数が 15 件あった。これはどちらの項目も選任の件であり、分類が難しいことと、議案としての出現確率が高いことに起因している。また、分類にはその他といった項目が存在するが、今回はその他への分類をしていないため、15 件が誤分類となった。そして、同一ページに対し同じ議案が出てこないことが前提で議案分類を行っているため、同じ議案が複数出てきた開始ページの推定に影響を与え、6 件の誤分類となった。

また、その場合の推定は退職慰労金に分類される傾向にあり、それに起因して退職慰労金の分類適合率が低くなってしまった。その際の正解データと出力結果を表4に示す。

表 4: 同じ議案が複数出てきた開始ページ

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
40	剰余金処分	40	剰余金処分
40	定款変更	40	定款変更
48	取締役選任	48	取締役選任
50	取締役選任	50	取締役選任
52	役員賞与	52	役員賞与
52	退職慰労金	52	退職慰労金
53	役員報酬	53	役員報酬
53	役員報酬	53	退職慰労金

提案手法1の分類推定を向上させるためには、取締役選任と監査人選任の項目の特徴語の選択をヒューリスティックに調整することが考えられる。また、その他の項目も、同様の手法で分類できるようにすることも考えられる。同じ議案が存在するページに関しては、退職慰労金の項目に分類されることが多いため、退職慰労金への分類に制約を与えることで、解消されることが考えられる。

6 まとめ

本研究では、株主招集通知における議案の開始ページを推定し、その議案を分類する手法を提案した。議案の開始ページ推定は、議案がある開始ページには「議案」または「第 X 号議案」が先頭に含まれるといった規則に基づく。議案分類の推定は、議案ごとの特徴語を抽出し、その特徴語のスコアに基づき分類した。また、深層学習を用いた議案分類も行った。評価の結果、特徴語による議案分類、深層学習による議案分類ともに良好な適合率、再現率を達成した。

今後の展望としては、各手法には得手不得手が存在するため、本研究での評価に用いた結果を開発データとし、ハイブリッド手法の実現を予定している。

参考文献

- [1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).
- [2] 北森詩織, 酒井浩之, 坂地泰紀: 決算短信 PDF からの業績予測文の抽出, 電子情報通信学会論文誌 D, Vol. J100-D, No. 2, p. 150161 (2017).

- [3] Peramunetilleke, D. and Wong, R. K.: Currency exchange rate forecasting from news headlines, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pp. 131–139 (2002).
- [4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).
- [5] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信 PDF からの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).
- [6] 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).
- [7] 白田由香利, 橋本隆子, 佐倉環: LDA 方式による金融政策トピック抽出, 第159回DBS・第115回IFAT合同研究発表会 (2014).

ネットワークの表現学習を用いた金融専門極性辞書の構築

Lexicon Creation for Financial Sentiment Analysis using Network Embedding

伊藤 諒^{1*} 和泉 潔¹ 須田 真太郎^{2†}
Ryo Ito¹ 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: It is necessary to build a comprehensive polarity dictionary specialized for financial policy to improve the accuracy of lexicon-based sentiment analysis in evaluating texts written on financial policy. In this research, we acquire distributed representation of words using feature learning of dependency network of words and create the polarity dictionary by bootstrap method using the distributed representation of words.

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、Kumar Ravi et al.(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。

このような中、Jegadeesh and Wu(2015) の研究を代表として、近年テキストマイニングを応用して金融政策の効果を分析する新たな研究が登場している。Jegadeesh and Wu(2015) は、米国の金融政策を策定する委員会である Federal Open Market Committee

*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。

(FOMC) の議事録からトピックを LDA により抽出し、Loughran and McDonald(2011) により導入された辞書を用いて、トピック別にセンチメントを付与し、各トピックのセンチメントがマクロ変数や資産価格に対して与える影響を分析している。ここにおいて、Loughran and McDonald(2011) の辞書は同じファイナンスの分野であれど、企業の財務報告書を基に作成された辞書であり、金融政策の分析に特化して作成された辞書ではない。さらに、伊藤他(2017) は Jegadeesh and Wu(2015) の手法を拡張したトピック別センチメント付与方法を用いて、得られたトピック別センチメントと市場参加者の期待形成に関するイベントスタディ分析を行っているが、ここで用いられている辞書は、金融政策の分析に特化した人手で作成された辞書であり、テキストに出現する単語の種類が多さから、全ての極性語を網羅した辞書を構築出来ていないと言えない。

そこで本研究では、金融政策の分析に特化した極性辞書の自動構築を行うことを目的とし、単語の特徴を定量的に扱うために、とりわけ近年研究の進んでいるネットワークの表現学習を用いた辞書自動構築方法を提案する。

2 関連研究

本章では、極性語を対象にした辞書構築 (Lexicon Creation) に関する先行研究について述べる。

極性辞書構築に関して数多くの研究が行われているが、大きく分類して、辞書ベースのアプローチと、コーパスベースのアプローチに分けられる。

辞書ベースのアプローチとしては、辞書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。

また、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、片倉・高橋(2015) は、極性語同士は類似した分散表現を持つという考えの下、CBOW モデル (Continuous bag-of-words model) によって単語の分散表現を得た後に、シードとなる極性語と類似した分散表現を持つ単語を取得する事で、Loughran and McDonald(2011) の辞書の拡張を行なっている。さらに、コーパスベ

ースのアプローチにおいて、テキストと極性ラベルの関係を学習させた後に、学習済みモデルの各単語に対する係数を得る事で、各単語の極性スコアを得るといふ、教師あり学習を用いた研究も存在する。坪内・山下(2014) は、リッジ回帰を用いてセンテンスとポジティブ・ネガティブフラグの関係を学習し、モデルの各フレーズに対する係数をフレーズの極性値として付与している。

コーパスベースのアプローチでは人手で構造化されていない情報を入力とするため、一般に辞書ベースのアプローチよりも辞書構築の精度は劣るが、入力として専門領域について書かれたテキストを入力にする事で、専門領域に特化した極性辞書を構築できるという利点がある。

以上の背景の下、本研究ではコーパスベースのアプローチを用いて、金融政策のセンチメント分析に特化した極性辞書を作成する事を目的とする。また、本研究ではネットワークの表現学習によって得られた分散表現を用いた、新たな極性辞書構築の為の手法を提案する。

3 ネットワークの表現学習を用いた極性辞書の構築

本章では、ネットワークの表現学習を用いた極性辞書の構築方法について提案するが、はじめに本研究の着想について述べ、次に具体的なフレームワークについて述べる。

3.1 本研究の着想

極性を持つセンテンス中において出現する極性語が、他の単語とどのような関係を有しているかを考える。図 1 は極性を持つセンテンスについて、極性語の係り先を図示したものである。

この例で特徴的な点は、これらの極性語が同じ単語である rate に掛かっている点である。increased は decreased と対義語の関係にあるが、両者とも数量の大きさを評価するという概念を持ち、評価対象の単語は類似した単語になることが考えられる。また、increased や decreased に対する類義語も、同様に数量の大きさを評価するという概念を持つことから、評価対象の単語は類似した単語になることが考えられる。以上の観察から、極性語同士は係り先となる単語が類似しており、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ると考えられる。

さて、単語間の係り受け関係を表現するものとして、ネットワークによる表現がある。ここで、先の観察を

¹<https://wordnet.princeton.edu/wordnet/>

ネットワークの観点で表現すると、ネットワーク上において極性語同士はある別の単語を介して繋がっており、二次近接の関係にあると言える。よって、単語間の係り受けネットワークにおける二次近接の関係性を以って単語を定量化する事で、単語の持つ極性語らしさを定量的に扱えると考えられる。そこで、本研究では初めに金融政策テキストから単語間の係り受けネットワークを構築する。次に、ネットワークの各ノード（単語）に対して、二次近接の関係性を、J Tang et al.(2015)によってネットワークの表現学習手法として提案されたLINE (Large-scale Information Network Embedding)を用いて定量化する。そして、定量化された単語の分散表現と種表現の極性語を元に、ブートストラップ法を用いて極性語辞書構築を試みる。

分散表現を用いて極性辞書構築を行う研究として、前章で述べた片倉・高橋 (2015) の研究が存在するが、この研究において単語の分散表現の獲得にはCBOWモデル (Mikolov et al. 2013) が用いられている。CBOWモデルではある単語 x が現れている周辺の単語を用いて x を推定する中で、単語の分散表現を獲得するが、周辺単語の組み合わせの多さ故、CBOWモデルで獲得される分散表現はLINEよりも意味の広い分散表現となることが考えられる。一方、LINEではCBOWモデルよりも焦点を絞った近接関係を元にして分散表現を得る為、極性語の獲得において、より良い分散表現であると期待される。

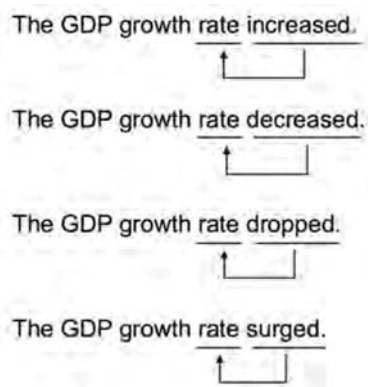


図 1: 極性語の係り先の類似性

3.2 係り受けネットワークの構築

はじめに、解析対象となるテキストから単語間の係り受けネットワークを構築する。本研究では金融政策の評価に特化した辞書を作成するという目的から、解析対象となるテキストとして、米国の金融政策を策定する委員会の議事録であるFOMC議事録を用いる。

まずFOMC議事録のHTMLを、FRBのホームペー

ジ²よりウェブクローラーによって収集する。次に収集した議事録のHTMLをパースし、テキスト部分のみを抽出する。そして、取得したテキストに対して係り受け解析を行うことで、単語の係り受けペアを抽出する。この際、各単語に対してレンマ化を行う。さらに、単語の係り受けペアから重み付き有向グラフを作成する。ここで、ネットワークの各ノードは各単語に相当し、ネットワーク上のあるノード v_i から v_j に貼られるエッジの重み w_{ij} は、その単語の係り受けペアの出現回数に相当する。また、エッジの方向性は係り元から係り先としてネットワークを構成する。

3.3 LINEによる単語の分散表現獲得

前節で得られた単語間の係り受けネットワークを用いて、単語の分散表現をネットワークの表現学習手法であるLINEによって得る。LINEには一次のLINEと二次のLINEが存在するが、ここでは二次近接の関係にある単語に対して、類似した分散表現を獲得させたいため、二次のLINEを用いる。

二次のLINEではノード間で共通して隣接しているノードが多ければ、それら2つのノードが類似した分散表現を持つように学習をする。具体的には、あるノード v_i から他のノードに対してエッジが貼られる確率分布が、分散表現から予測される分布と観測される分布とで類似した分布になるように学習を行う。ここであるノード v_i に対してノード自身のベクトル \vec{u}_i とノードの文脈ベクトル \vec{u}_i' を導入する。分散表現から予測されるあるノード v_i から他のノード v_j に対してエッジが貼られる確率は以下の式で表される。

$$p_2(v_j|v_i) = \frac{\exp(\vec{u}_j'^T \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_k'^T \cdot \vec{u}_i)}$$

また、ノードの集合として V 、エッジの集合として E を持つ、ネットワーク G から観測可能な、あるノード v_i から他のノード v_j に対してエッジが貼られる確率は以下となる。

$$\hat{p}_2(v_j|v_i) = \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}}$$

$N(i)$ はノード i からエッジを貼られているノードの集合を表す。

ここで、上記の2つの分布間の距離の重み付き総和を目的関数とし、これが最小になるように学習を行う。目的関数は以下の式で表される。

$$O_2 = \sum_{i \in V} \lambda_i d(\hat{p}_2(\cdot|v_i), p_2(\cdot|v_i))$$

²<https://www.federalreserve.gov/>

λ_i はノード i のネットワークにおける重要度を表し、 $\sum_{k \in N(i)} w_{ik}$ の値が用いられる。二つの分布間の距離 $d(\cdot, \cdot)$ を KL 情報量として、定数項を除くと以下の形を得る。

$$O_2 = - \sum_{(i,j) \in E} w_{ij} \log p_2(v_j | v_i)$$

そして、この目的関数を負例サンプリングを用いて最適化する事で、各ノードに対する分散表現を得る。

3.4 ブートストラップ法による極性語の獲得

次に LINE によって得られた単語の分散表現を用いて、ブートストラップ法により、極性語候補の獲得を行う。

まず、ブートストラップを行う際にシードとなる極性語を人手により与え、これらのシードとなる単語を極性語候補リスト C へ登録する。また、ある単語 d がどれほど極性語らしいかを表す値である信頼度 P_d を導入するが、ここでシードとなる全ての極性語の信頼度を 1 として与える。

次に極性語候補リストとなる単語以外の各単語に対し、信頼度 P_d を、単語 d と極性語候補リスト中の各単語 c との類似度と、単語 c が持つ信頼度 P_c との積の平均により算出する。

$$P_d = \frac{1}{|C|} \sum_{c \in C} \text{sim}(\vec{u}_c, \vec{u}_d) P_c$$

$\text{sim}(\vec{u}_c, \vec{u}_d)$ は 2 単語間の分散表現の類似度を表す関数であり、ここではコサイン類似度を用いる。また \vec{u}_d は単語 d の持つ、LINE によって得られたノード自身のベクトルに対応する。

$$\text{sim}(\vec{u}_c, \vec{u}_d) = \frac{\vec{u}_c \cdot \vec{u}_d}{|\vec{u}_c| |\vec{u}_d|}$$

さらに、極性語候補リストとなる単語以外の各単語に対し信頼度 P_d を計算した後、信頼度 P_d のスコアが上位 L 件以内に含まれる単語を、次の反復において使用するために極性語候補リスト C に追加する。

以上のステップを繰り返し行うことによって、極性語候補リスト C を拡張し、停止条件の回数 M に達した場合、信頼度のスコアが上位 N 件以内に含まれる単語を、システムの最終的なアウトプットとして出力する。

4 実験

本章では、各種実験設定と評価方法について述べる。

4.1 分散表現のマッピング

はじめに、LINE によって得られた分散表現がどのような特性を有しているかを、分散表現を可視化する事で検証を行った。また本検証を行うに当たり、極性語の周辺にどのような単語が存在するかを中心に考察した。

単語間の係り受けネットワークの構築において、1993 年 1 月から 2016 年 11 月の間に公表された FOMC 議事録 190 件を用いた。また分散表現の次元数は 50 次元として学習を行った。分散表現の可視化にあたり t-SNE 法 (Maaten and Hinton 2008) を用いて次元削減を行い、各単語を各座標に配置することで可視化を行った。

また、比較実験として、CBOW モデルを用いて分散表現を獲得した場合の可視化も同様に t-sne 法により行った。CBOW モデルを用いた場合の分散表現の次元数は LINE と同じく 50 次元とした。また、CBOW モデルにおけるウィンドウサイズは 4 として学習を行った。

4.2 極性辞書構築実験

ここでは、LINE によって得られた分散表現を用いて、ブートストラップ法を行う事で極性語を取得する実験を行った。また比較実験として、CBOW モデルによって得られた分散表現を入力とした場合と比べて、どのような語が取得されるかを検証した。ブートストラップにおける各パラメーター L , M , N はそれぞれ 1, 30, 30 とした。さらにシードとなる極性語として、increase · high · improvement の 3 単語を用い、LINE と CBOW モデルにおける分散表現は、両者とも分散表現のマッピング実験にて用いた分散表現を用いた。

5 結果と考察

本章では、実験結果と考察について述べる。とりわけ、LINE によって得た単語の分散表現と CBOW モデルによって得た単語の分散表現の相違と極性語の取得可能性の観点から論じる。

5.1 分散表現のマッピング

図 2 は、LINE によって得られた単語の分散表現を 2 次元平面上にマッピングをしたものである。図から見て取れるように、マッピングされた分散表現は上部と下部に分断されている。上部に出現する単語は下部に出現する単語と比較して、数字や副詞などの係り先になりやすい単語が存在していた。このように LINE によって得られた分散表現は、当然のことではあるが、

単語の係り受け関係を反映した分散表現となっていることが確認された。

図3は図2の一部分を拡大して表示したものである。increase, decrease, rise, dropを始めとして、動詞を中心とした極性語が多く分布していた。また、他の部分においても形容詞の極性語が多く分布する部分や、名詞の極性語が多く分布する部分が観察された。これらの結果は、係り受けネットワークにおける二次近接の関係性を元に、シードとなる単語を与える事で極性語を自動獲得出来る可能性を示している。



図 2: LINE による単語分散表現の可視化

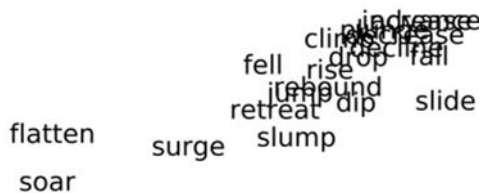


図 3: 動詞極性語集中部位 (LINE)

一方、図4はCBOWモデルによって得られた単語の分散表現を、2次元平面上にマッピングをしたものである。図から見て取れるように、LINEを用いた場合と比較して、明確に二つの部位に分かれるということはない。図5は図4の一部分を拡大して表示したものである。図から分かるように、increase, decrease, growth, expansionを始めとして多くの極性語が集中して存在していた。このことは、CBOWモデルを用いた場合においても、シードとなる単語を与える事で極性語を自動獲得出来る可能性を示している。

LINEとの比較についてであるが、LINEでは極性語は比較的品詞ごとに分かれて存在していた。一方、CBOWモデルではLINEと比較して極性語がある程度品詞に依らず纏まって分布する傾向がある事が分かった。これは、LINEでは係り受け関係を元に分散表現

を獲得するため、文法的構造を含んだ分散表現になるが、CBOWモデルでは自身の周辺単語から自身の単語自体を予測する中で分散表現を得るので、LINEよりもより文脈の意味情報を含んだ分散表現が獲得されるためと考えられる。



図 4: CBOW モデルによる単語分散表現の可視化



図 5: 極性語集中部位 (CBOW モデル)

5.2 極性辞書構築実験

表1は、各モデルによる分散表現を入力として、ブートストラップ法を用いる事で獲得された単語の内、信頼度上位の20単語を示したものである。

LINEによる分散表現を用いた場合において、decrease, drop, advance, fallなどをはじめとする極性語が取得できている事が分かる。とりわけ、climbなどの単語は日常的に用いられる文脈では極性を持たないが、物価等が上昇するという意味合いで、金融政策テキスト中には多く出現する表現であり、このような単語を極性語として取得する事が出来ている。しかしながら、thereなどの極性語ではない単語も含まれる結果となった。

一方、CBOW モデルによる分散表現を用いた場合においても、decline, rise, drop などの極性語を取得する事が出来ており、さらに climb などの、金融政策の文脈において極性を持つ単語も取得する事が出来ている。

LINE による分散表現と CBOW モデルによる分散表現を、それぞれ用いた場合の差異について述べる。ブートストラップ法において、反復処理を行う中で目的となるインスタンス以外のインスタンスが抽出されてしまう問題は、意味ドリフト (Curran et al. 2007) と呼ばれているが、LINE による分散表現を用いた場合、この意味ドリフトが比較的早く起きている事が観察された。具体的には極性語ではない differently が獲得された後に、differing が獲得されるというように、意味がドリフトしていく様子が観察された。意味ドリフトが起こる原因の解明や抑制は今後の課題である。

表 1: ブートストラップ法で取得された単語の内、信頼度上位 20 件以内の単語

LINE	CBOW モデル
rise, decline, decrease, drop, advance, climb, showing, fall, jump, there, faster-than-anticipated, quicken, constraining, weaker-than-anticipated, slower-than-expected, differently, better-than-anticipated, differing, higher-than-expected, approximating	decline, rise, drop, decrease, advance, fall, fell, surge, jump, decelerate, climb, rebound, contract, accelerate, deceleration, run-up, gain, step-up, weaken, recover

6 まとめ

専門領域について書かれたテキストに対して語彙ベースのセンチメント分析を行う上で、専門領域に特化した網羅的な辞書がセンチメント分析の精度向上のために必要である。本研究では、金融政策の分析に特化した極性辞書の自動構築を行うことを目的とし、極性語同士は係り受けネットワークにおいて二次近接に位置する度合いが高いという考えの下、二次近接の関係を定量化するネットワークの表現学習手法である、二次の LINE を用いた辞書自動構築方法を提案した。

実験により、LINE による分散表現を用いた辞書構築の可能性が示唆された。また、CBOW モデルによる分散表現とは異なる性質を持つ分散表現が獲得されて

いることを確認し、LINE を用いた分散表現と CBOW モデルを用いた分散表現の相違について考察をした。

今後の課題として多くの課題が挙げられるが、主に次の四点が重要な課題として挙げられる。

まず第一点目として、提案手法による辞書構築の精度評価を定量的に行うことである。またその中において、CBOW モデルを用いた場合と提案手法を用いた場合の精度を比較する必要がある。さらに、シードとして与える極性語に対する辞書構築精度のロバストさを両手法とも検証する必要がある。

また第二点目として、提案手法を用いた場合、単語のポジティブ・ネガティブに関係なく極性語候補に追加されるが、取得された単語に対してポジティブ・ネガティブに分類した辞書を構築できるよう、提案手法を拡張することが課題である。これを実現する上で、単語の類義語・対義語関係を考慮した分散表現を得る必要があるが、K. A. Nguyen et al.(2016) は、Skip-gram モデルの目的関数に類義語・対義語関係を考慮した項を追加した上で目的関数を最適化することで、単語の類義語・対義語関係を考慮した分散表現を得ることが出来、取得された極性語をポジティブ・ネガティブに振り分けた辞書を構築することが可能であると考えられる。

さらに第三点目として、ブートストラップ法を用いた場合に、反復を進めるにつれて極性語以外の単語が取得される意味ドリフトに対処する必要がある。意味ドリフトに対処する方法としては、各単語に対する信頼度を算出する関数の再設計や、ブートストラップ法と学習器を組み合わせる方法が考えられる。また、Curran et al. (2007) の研究に見られるように、意味ドリフトを起こす事が分かっている単語をストップクラスとして用意し、単語の抽出に制限をかける方法を用いる事が考えられる。

最後に第四点目として、本提案手法によって拡張された辞書を用いた場合と、拡張前の辞書を用いた場合とに分けた上で、金融政策テキストに対して伊藤他 (2017) の手法によってセンチメント分析を行い、拡張された辞書を用いた場合に、センチメント分析の精度が向上するかを検証する必要がある。拡張された辞書が正しい極性値のついた網羅的な辞書であるならば、その辞書を用いる事でセンチメント分析の精度が向上する事が期待される。さらに、センチメント分析の精度が向上しているのであれば、得られたセンチメントの値と市場期待やエコノミスト予想との関連性について検証を行いたい。

参考文献

- [1] Bollen, J., and Huina, M. (2011) Twitter mood as a stock market predictor, *Computer* **44**: 91-94.
- [2] Curran, J. R., Murphy, T., and Scholz, B. (2007) Minimising semantic drift with mutual exclusion bootstrapping, *In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*: 172-180.
- [3] Jegadeesh, N., and Wu, D. (2015) Deciphering FedSpeak: The Information Content of FOMC Meetings, *2016 AFA Annual Meeting Working Paper* (<https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=1136>).
- [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] 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.
- [6] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.
- [7] Maaten, L. V. D., and Hinton, G. (2008) Visualizing data using t-SNE, *Journal of Machine Learning Research* **9(Nov)**: 2579-2605.
- [8] 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*.
- [9] 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.
- [10] 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.
- [11] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015) Line: Large-scale information network embedding, *In Proceedings of the 24th International Conference on World Wide Web*: 1067-1077.
- [12] 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.
- [13] 片倉賢治, 高橋大志 (2015) 金融市場ニュースの分散表現学習による辞書作成と金融市場分析 2015年度人工知能学会全国大会 (第29回) .
- [14] 坪内孝太, 山下達雄 (2014) 株価掲示板データを用いたファイナンス用ポジネガ辞書の生成 2014年度人工知能学会全国大会 (第28回) .

人工知能を用いた金融政策予想と市場予測分布に基づく 為替の投資戦略

Monetary policy analysis and investment strategy using artificial intelligence and economists' forecast distribution

上田 翼^{1*} 東出 卓朗^{1,2}
Tsubasa Ueda¹ Takuo Higashide^{1,2}

¹ 三井住友アセットマネジメント株式会社

² 一橋大学大学院国際企業戦略研究科

Abstract: Central bank's monetary policy is one of the major interests for market participants. In this paper, we clarified Reserve Bank of Australia's monetary policy reaction function, predicted its policy change, and applied them to investment strategy. First of all, assuming perfect foresight by the central bank, we estimated an extended Taylor rule using bidirectional Recurrent Neural Network. Next, we combined it with distributed representation of Monetary Policy Committee minutes to develop a classifier of interest rate decision. Both the extended Taylor rule and the classifier showed improved performance. Finally, we formulated profitable Foreign Exchange strategy based on the classifier's prediction and market economists' forecasts.

1 はじめに

中央銀行の金融政策のあり方には、様々な議論が存在する。最も有名な政策ルールは、[2]Taylor(1993)が示したテイラー・ルールであり、需給ギャップとインフレギャップ（現実のインフレ率の目標に対する乖離幅）に対応して政策金利が調節されるというものである。

テイラー・ルールは、もともと現実の政策金利を単純な式でよく近似できることで知られていたが、金融危機以降は効果的に説明できない事例も増えてきた。こうした環境もあり、中銀の政策反応関数の捉え方も変化している。[1]Curdia, et al.(2010)は金融機関の貸出スプレッドをテイラー・ルールに加えた。また、先行きの経済見通しを考慮するのも一般的であり、FRBの[3]Yellen(2012)は最適管理（optimal control）を提唱した。いずれにせよ、市場参加者にとっては、中銀の政策反応関数を明らかにできれば金融政策の予想に役立つ。

勿論、中央銀行は常に一定のルールに従うわけではなく、その時々状況を踏まえたうえで意思決定していると考えるのが自然だろう。そうした議論の痕跡は、金融政策決定会合（MPC）の議事録などに残されている可能性が高い。

これまで、深層学習手法で中央銀行の政策を予想する

試みはあった。[5]塩野(2016)は日銀の公表した文書と経済統計データから、政策変更の有無を予想した。しかし、直接的に政策金利を予想したケースは少ない。中銀の政策反応関数を特定することで、金利の変化幅、緩和・引き締め局面変化、経済シナリオを前提にした金利パス予想などに役立てられる可能性がある。また、RNNを用いた時系列予測の事例も数多く存在する。しかし、双方向RNNを用いた研究は自然言語解析や音声解析分野以外では稀である。

そこで本稿では、深層学習手法によりテイラー・ルールを拡張し、MPC議事録テキストの情報と組み合わせることで、政策変更を予想するシステムを提案する。まず、中央銀行の完全予見を仮定し双方向RNNを用いて拡張されたテイラー・ルールを推定する。次に、ルールの示唆と政策金利の乖離を導出し、MPC議事録の分散表現を加えることで、政策変更を予測する分類器を構築した。この分類器を豪州中銀（RBA）の金融政策に適用したところ、拡張されたテイラー・ルールは、単純なスプレッド調整型テイラー・ルールと比べて実際の政策金利をより正確に捉えた。また、構築した予測器は、議事録テキストのみを用いるケースより予測性能が高かった。

また上述の手法で得られる金融政策の予測情報に加えて、エコノミストの金融政策に対する見通しの情報も加味し、オーストラリアドルの対米ドルの為替（AUD/USD）を対象とする投資戦略を構築した。重要指標発表前の相

*三井住友アセットマネジメント株式会社
東京都港区愛宕2丁目5番1号
愛宕グリーンヒルズ MORI タワー 28 階
Email:tsubasa_ueda@smam-jp.com

場は膠着する傾向にあることと、発表後は相場が一定時間ボラタイルに動くことが経験的に知られていることから、発表 30 分前にポジション保有し、発表 30 分後にポジション手仕舞いするといった投資戦略の提案を行う。実証分析の結果、当該投資戦略のシャープレシオは 9 を上回る良好なパフォーマンスを獲得した。

2 分析手法

2.1 テイラー・ルール

Taylor は米国の FF 金利が⁵ (1) で近似できることを示した。

$$i_t = \pi_t^* + r_t^* + \alpha(\pi_t - \pi_t^*) + \beta(y_t - y_t^*). \quad (1)$$

すなわち、政策金利は ①均衡名目金利 ($\pi^* + r^*$)、②現実のインフレ率 (π) とインフレ目標 (π^*) の乖離、③現実の GDP (y) と潜在 GDP (y^*) の乖離に応じて調節される。[1]Curdia, et al.(2010) は、金融市場へのショックを考慮するため、テイラー・ルールに (2) のように貸出スプレッド項を加えた。

$$i_t = \pi_t^* + r_t^* + \alpha(\pi_t - \pi_t^*) + \beta(y_t - y_t^*) + \gamma\omega_t. \quad (2)$$

現実には、金融機関は調達金利（政策金利）にスプレッド (ω) を加えて貸出金利を決定する。それゆえ、金融ショック等で貸出スプレッドが変化した場合、適切な政策金利も上下すると考えられる。本稿では、同モデルをベースにテイラー・ルールの拡張を試みた。

2.2 双方向 RNN

RNN は内部に閉路を持つニューラルネットであり、時刻 t における中間層の出力が次の時刻 $t+1$ の中間層の入力となる。この構造により、系列データの中の“文脈”を捉えるのに適しているといわれる。系列データの全体がある場合、系列データを逆向き ($t = T, T-1, \dots, 1$) に RNN に入力することもできる。順向きの RNN と逆向きの RNN を統合したものを双方向 RNN と呼ぶ。

$$\hat{y}_t = o(\mathbf{W}^o[\vec{\mathbf{h}}_t^L; \overleftarrow{\mathbf{h}}_t^L]),$$

$$\vec{\mathbf{h}}_t^l = f^l(\overrightarrow{\mathbf{W}}^l[\vec{\mathbf{h}}_t^{l-1}; \overleftarrow{\mathbf{h}}_t^{l-1}; \vec{\mathbf{h}}_{t-1}^l]),$$

$$\overleftarrow{\mathbf{h}}_t^l = f^l(\overleftarrow{\mathbf{W}}^l[\overleftarrow{\mathbf{h}}_t^{l-1}; \overleftarrow{\mathbf{h}}_t^{l-1}; \overleftarrow{\mathbf{h}}_{t+1}^l]).$$

ただし、 $\{l|1 \leq l \leq L\}$ は隠れ変数ベクトル \mathbf{h} の階層を示し、 $\mathbf{h}_t^0 = \mathbf{x}_t$ とする。 \mathbf{W}^o は予測関数 o のパラメータ、 f はシグモイド関数などの非線形関数である。 $\overrightarrow{\mathbf{h}}$ 、 $\overleftarrow{\mathbf{h}}$ は前向き、 $\overleftarrow{\mathbf{h}}$ 、 $\overrightarrow{\mathbf{h}}$ は後ろ向きの隠れ変数ベクトルとパラメータを示す。金融政策は、過去から足元までの経済状況の変遷から判断するのが基本だが、将来の経済予測に基づいて判断するフォワードルッキングな要素も持つ。中

央銀行が将来の経済変数を完全予見すると仮定した場合、中間層が過去と未来の状態を記憶する双方向 RNN によりテイラー・ルールを近似できると考えた。

具体的には、経済時系列データを双方向 RNN に入力し、その出力と貸出スプレッドを回帰層に入力、出力結果に名目均衡金利を加算して政策金利を導出する「双方向テイラー・ルール・モデル」を考えた。

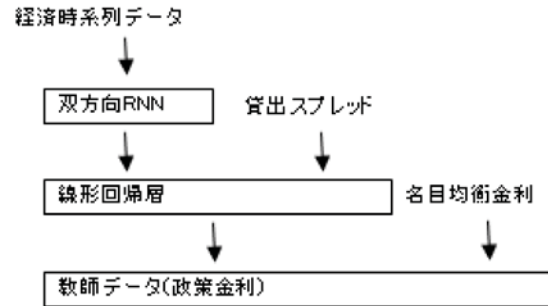


図 1: 双方向テイラー・ルール・モデルの概略

2.3 Doc2Vec による文章分散表現

現実的には、中銀は常に一定のルールに従うわけではない。テイラー・ルールが考慮しない情報を踏まえて、意思決定していると考えるのが自然である。ルールが示唆する政策金利水準から逸脱することを、いわゆるタカ派・ハト派スタンスと解釈することもできよう。そのようなニュアンスは MPC 議事録に表れている可能性がある。そこで、[5] 塩野 (2016) の手法にならい、Doc2Vec を用いて各議事録の分散表現を獲得し、先の双方向テイラー・ルールが示唆する水準と実際の政策金利を考慮して、政策変更を予想する分類器を構築した。

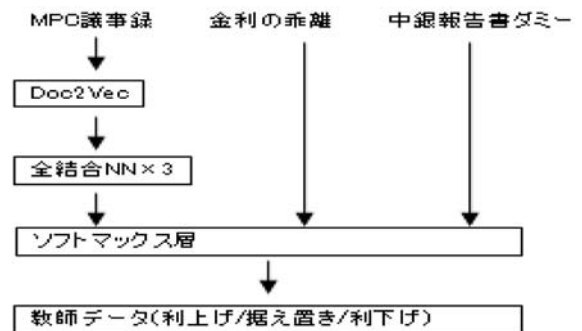


図 2: 政策変更予測器の概略

3 実証分析

3.1 双方向テイラー・ルール

双方向テイラー・ルールのモデルに使用するデータは次の通りである。

双方向 RNN に入力する経済時系列データは、①CPI 前年比、②基調（コア）CPI 前年比、③失業率ギャップとした。失業率ギャップは、OECD の均衡失業率推計を HP フィルターで平滑化してから実際の失業率との差を取った。回帰層に入力する貸出スプレッドは、RBA が公表する住宅ローン変動金利と現実の政策金利の差をとった。最後に加算する名目均衡金利は、潜在成長率＋インフレ目標とした。潜在成長率は OECD の潜在 GDP 推計を HP フィルターで平滑化して導出し、インフレ目標は RBA のターゲット・レンジ中央の 2.5 % とした。

教師データはその月に開かれた MPC で決定された政策金利とし、対応する入力 MPC 開催時点で公表され入手可能なデータ（過去 6 カ月分）とした。

以上のデータセットを 2000 年 2 月～2014 年 12 月を訓練期間、2015 年 2 月～2016 年 12 月テスト期間として分割して利用した。RNN には正則化を行った。さらに結果を一般的なテイラー・ルールと比較した。

テスト期間における平均二乗誤差は、双方向テイラー・ルールが 0.111、単純なスプレッド調整型テイラー・ルールが 0.151 となり、拡張したモデルの優位性が確認された。

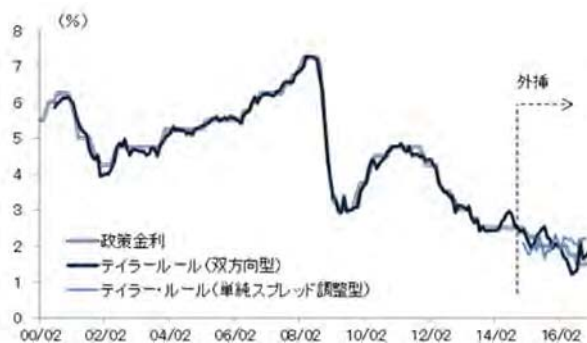


図 3: テイラー・ルールの性能比較

3.2 政策変更予測器

双方向テイラー・ルールが実際の政策予測に有効かどうか予測器を構築して検証した。

入力データは、①前回の MPC 議事録テキスト（概ね MPC の 2 週間後に公開）、②その月の中銀報告書ダミー、③MPC 前の政策金利と双方向テイラー・ルールの乖離を用いた。教師データは、MPC の政策金利決定を利上げ、据え置き、利下げの 3 通りに分類したものとした。

以上のデータセットを 2006 年 10 月～2014 年 12 月を訓練期間、2015 年 2 月～2016 年 12 月をテスト期間として分割して利用した。これをモデル 1 とし、議事録のみから作成した予測器（モデル 2）と比較した。

テスト期間における正解率は、モデル 1 が 0.909、モデル 2 が 0.818 となり、双方向テイラー・ルールを加えることで予測性能が改善することが示された。そもそも、モデル 2 では各時点で利上げ／利下げがほぼ当確率で予測されており、政策変更は察知できるが政策金利の上下を判別できていない。

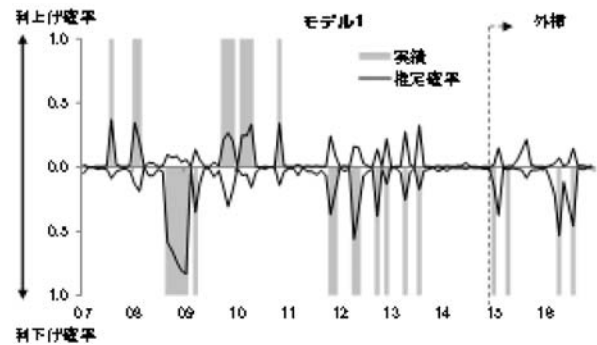


図 4: 政策変更予測器の性能比較 1

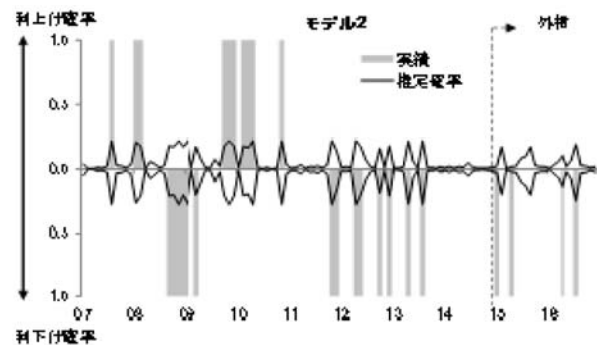


図 5: 政策変更予測器の性能比較 2

3.3 投資戦略

ここでは、投資戦略の応用について述べる。本稿でのシミュレーションはすべて統計ソフト R 上で行った。まず、”政策変更予測器”を投資戦略に応用する方法について説明する。モデルから算出される利上げ、据え置き、利下げ各々の確率から最も大きい確率のシナリオが起こると判断する。例えば、{利上げ, 据え置き, 利下げ} = {0.2, 0.6, 0.2} であれば据え置きと想定する。ポジション保有のタイミングは金融政策発表 30 分前、ポジション手仕舞いのタイミングは金融政策発表 30 分後としてシミュレーションを行った。価格データは Bloomberg より各売買方向の気配値を使用した。ポジション保有量

は、利上げあるいは据え置きであった場合ロング1単位保有、利下げであった場合はショート1単位保有とする。据え置きの場合でもロングポジションを保有する背景には、AUD/USDはキャリー等の妙味から選好して買われる傾向にあるためである。

次に Bloomberg から提供されている主要エコノミストによる見通しをポジション保有量に加味する。具体的には Bloomberg 独自の算出方法に基づき、予測正答率に裏付けされた上位エコノミストランキングの見通し情報を利用する。また上位エコノミスト4人の情報に限定する。この情報を次のようにポジションに反映させる。まず、毎回主要エコノミストが政策金利の見通しの回答をしているとは限らない。背景には様々な理由はあるが、この場合エコノミストは自信をもって予測することが困難な状況であると判断する。この場合、エコノミストの予測情報は投資戦略に利用しないものとする。一方で、主要エコノミストが全員回答している場合は有益な情報として利用する。次にこれらエコノミストの予測最頻値を取る。利上げ、据え置き、利下げの多数決によりポジションを決める。“政策変更予測器”から得られる情報と同様に、多数決の結果が、利上げあるいは据え置きであった場合ロング1単位保有、利下げであった場合はショート1単位保有するものとする。もし予測が据え置き2人、利下げ2人のように見方が分かれた場合には、予測力が過去最も高かったエコノミストの意見を採用するものとした。加えて、もし“政策変更予測器”と“エコノミスト情報(多数決)”の結果が異なっている場合は、ポジション保有しないものとする。

“政策変更予測器”から得られる情報のみを用いた投資戦略では、2015年2月～2016年12月までの間でシャープレシオは4を上回る結果となった。また“エコノミスト情報”もポジションに反映させた結果、シャープレシオは9を上回る良好な結果となった。また売買コストは保守的に見積もり1回の取引あたり5pipsとした。

表 1: シミュレーション結果

統計量	政策変更予測器	政策変更予測器 エコノミスト情報
最小値	-2.05 %	-0.42 %
第一四分位	-0.20 %	-0.10 %
中央値	0.19 %	0.10 %
平均	0.14 %	0.21 %
第三四分位	0.44 %	0.41 %
最大値	1.75 %	1.75 %
合計	3.03 %	4.65 %
SharpRatio	4.46	9.54

4 結論と今後の課題

本稿では、時系列データから中銀の従う政策反応関数を導出し、議事録テキストと合わせて政策変更予測に用い、さらに投資戦略に応用することを試みた。RBAの金融政策に適用したところ、一定の予測性能と収益性を得ることに成功した。また投資戦略においてシャープレシオも9を上回る良好な結果を得た。今後は、その他諸外国においても同様の手法で解析を行い、国ごとに議事録の持つ情報とエコノミストの見通し情報がどの程度有益なのかを見ていく。もっとも、多くの先進國中銀の多くはゼロ金利制約に直面し非伝統的な金融政策に移行したため、本稿で用いたテイラー・ルールがどの程度他国に適用できるかは未知数である。また今回エコノミストの見通し情報は、主モデルを捕捉する形で基本的な統計量を用い保守的に利用したが、今後はエコノミストの行動パターンを検知したうえで情報の精緻化を行うことも課題の1つである。

参考文献

- [1] Curdia, Vasco, and Michael Woodford.(2010). "Credit spreads and monetary policy". *Journal of Money, Credit and Banking* 42.s1: 3-35.
- [2] Taylor,John B.(1993)."Discretion versus policy rules in practice". *Carnegie-Rochester conference series on public policy, Vol. 39, North-Holland.*
- [3] Yellen,Janet L.(2012)."Revolution and Evolution in Central Bank Communications". *speech delivered at the Haas School of Business, University of California at Berkeley.*
- [4] 開発壮平, 黒住卓司, 寺西勇生.(2010)."今次金融危機の経験を踏まえた金融政策ルールの拡張について". *日銀レビュー 2010-J-6.*
- [5] 塩野 剛志.(2016)."文書の分散表現と深層学習を用いた日銀政策変更の予想". *人工知能学会研究会資料,SIG-FIN-016-11.*
- [6] 坪井 祐太.(2015)."自然言語処理におけるディープラーニングの発展". *オペレーションズ・リサーチ, Vol.60, No.4, pp. 205-211.*

金融レポート、およびマクロ経済指数によるリアルタイム 日銀センチメントの予測

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

余野 京登¹ 和泉 潔¹

Kyoto Yono¹, Kiyoshi Izumi¹

¹ 東京大学大学院

¹Graduate School of Engineering, The University of Tokyo

1 はじめに

金融市場を取り巻く情報は日々増加している。物価指数や産業活動指数など数値データもあれば、企業の有価証券報告書、ロイターなどの配信している経済ニュースなどのテキストデータも含まれている。この膨大な量の情報から、投資家やトレーダーは、それぞれの投資戦略に従い、必要な情報を選び出し、投資判断を下している。

ニューラルネットワークや遺伝的アルゴリズム等を数値データに用いて市場分析を行うものがあり、一定の成果を上げている[1]。一方で、近年、データマイニング技術を用いて、市場動向を分析する研究が多く行われている。ニュースのヘッドラインを分析し、為替市場の短期予測を行った研究[2]や、中央銀行の発行するテキストを用いた研究などがある。日本銀行のテキストを対象にしたものでは、ニューラルネットワークを用いることでレポートを指数化する研究[3]、政策変更確率を予測する研究[4]、米国のFOMCのテキストを対象にしたものでは、トピックモデルによる分類と極性語による指数化に関する研究[5]がある。

本研究では、日本銀行の景気に対するセンチメントを日本銀行の発行するテキストに対して、トピックモデルおよびニューラルネットワークを用いて、トピック事に分解した指数化を行った。

今後の展望として、よりリアルタイム性を出すためにテキストの発行以前において、マクロ経済指数を外挿することによりリアルタイムな指数の予測を目指す。

2 手法

2.1 データ

日本銀行の景気に対するセンチメントを知るために分析の対象となるテキストは主に以下の通りであるが、今回の分析では金融経済月報を対象とする。

分析の期間は1998年1月から2015年12月である。なお、2016年以降、金融経済月報は「経済・物価情勢の展望」(展望レポート)に集約された。

表 1 日本銀行の発行するテキスト

テキストの種類	タイミング	頻度	分量	フォーマット
金融政策決定会合議事要旨	1~2ヶ月後	月次	20-30	定形
主な意見	1週間後	月次	5	定形
総裁定例記者会見要旨	翌営業日	月次	20	不定形
経済・物価情勢の展望	翌営業日	四半期毎	10/100	定形
金融経済月報	翌営業日	月次	20-30	定形
講演	不定	月1回程度	10~20	不定形
記者会見	不定	月3回前後	1~2	不定形

2.2 LDA によるトピック分類

金融経済月報は、金融政策決定会合における政策判断の背景となる金融経済情勢を説明する資料として公開されており、その内容は、実体経済、物価情勢や海外経済など多岐にわたる。Latent Dirichlet Allocation (LDA) [6] を用いてこれらのトピックの抽出を行う。

LDA とは、一つの文章に複数のトピックが存在すると仮定したトピックモデルである。文章 d ごとにトピック分布 θ_d があり、このトピック分布 θ_d はパラメータ α のディリクレ分布に従う。トピック z ごとに単語分布 Φ_z があり、これらによって単語生成 w がさせる。なお、単語分布 Φ_z はパラメータ β のディリクレ分布に従う。全文章 D が生成される確率は以下の式のように表される。

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d.$$

この生成確率が最大化されるようにパラメータ推定を行う。

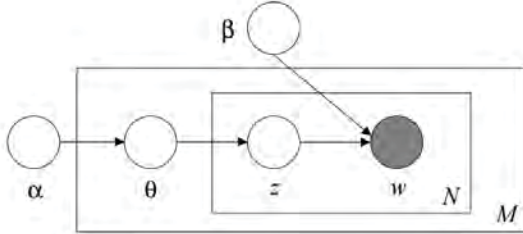


図 1 LDA のグラフィカルモデル

LDA を用いることにより、トピックの単語の生成確率が算出される。このモデルを各月の金融経済月報に適用し、トピックの比率を計算する。

2.3 文章のセンチメント付与

文章のセンチメント指数を計算する際に、極性語を用いたアプローチ[5]もあるが、今回の分析においては、先行研究[3]で用いられた景気ウォッチャー調査を学習データに LSTM RNN で構築したモデルでのセンチメント付与を行う。

内閣府の景気ウォッチャー調査は業種ごとにおける景気判断（良い、やや良い、変わらない、やや悪い、悪いの 5 段階評価）とその理由となるテキストがセットとなっており、かつデータ量が十分にあり、センチメント付与の学習データとしては良質なものである。

このデータセットに対して、LSTM RNN により、分類器を構築する。

RNN は音声や言語、動画などの系列データを扱うのに適したニューラルネットであり、その中でも LSTM は中間層にメモリーユニットを有しているため、長い文脈をとらえてモデル構築が可能である。

LSTM ブロックは記憶セル $c(t)$ 、入力ゲート $i(t)$ 、出力ゲート $o(t)$ 、忘却ゲート $f(t)$ の 4 つの要素からなり、それぞれ以下の式で記述される。

$$a(t) = f(W_c \mathbf{x}(t) + U_c \mathbf{h}(t-1) + b_c)$$

$$i(t) = \sigma(W_i \mathbf{x}(t) + U_i \mathbf{h}(t-1) + b_i)$$

$$f(t) = \sigma(W_f \mathbf{x}(t) + U_f \mathbf{h}(t-1) + b_f)$$

$$o(t) = \sigma(W_o \mathbf{x}(t) + U_o \mathbf{h}(t-1) + b_o)$$

$$c(t) = i(t) \odot a(t) + f(t) \odot c(t-1)$$

$$h(t) = o(t) \odot g(c(t))$$

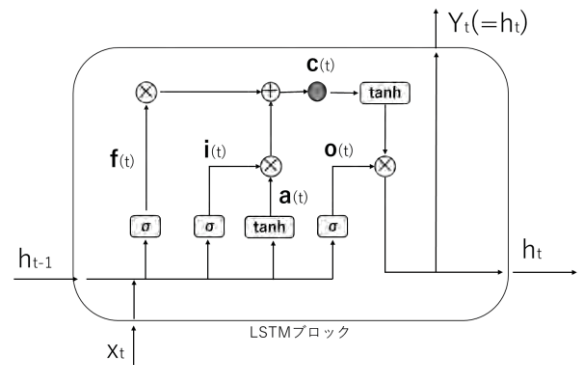


図 2 LSTM のグラフィカルモデル図

学習した LSTM RNN モデルに金融経済月報を一文ずつ判断させ、その平均値をその金融経済月報のセンチメントとする。

3 結果

3.1 トピック分類の結果

LDA の結果は以下の通りとなった。

表 2 各トピックの上位単語

個人消費	公共投資	在庫	設備投資
個人消費	公共投資	資本財	機械投資
住宅投資	公共工事請負金額	部品	実質
海外経済	実質輸出	財別	船舶
所得環境	所得環境	情報関連	輸送機械
設備投資	公共工事出来高	実質輸入	設備投資
景気	有効求人倍率	自動車関連	電力
公共投資	完全失業率	電子部品	機械受注
鉱工業生産	労働需給	中間財	全国百貨店売上高

物価	マネーサプライ	鉱工業生産	企業金融
消費者物価	マネーサプライ	在庫調整	金融機関
物価	電子部品	鉱工業生産	貸出態度
国内企業物価	C D	実質輸出	企業金融
国際商品市況	デバイス	最終需要	設備投資
ターム物金利	輸送機械	素材	資金供給面
生鮮食品	マネタリーベース	情報関連財	資金調達コスト
短期金融市場	業種別	鉄鋼	信用力
輸入物価	在庫バランス	海外経済	製造業

マーケット	貿易	資金供給	企業収益
株価	実質貿易収支	C P	製造業
国債	実質ベース	社債	企業収益
流通利回リスプレッド	サービス収支	民間銀行貸出	設備投資
為替相場	世帯	発行環境	雇用者所得
社債	海外経済	発行残高	中小企業
対米ドル相場	黒字幅	資金調達	賃金
流通利回り	二人以上	資金需要面	常用労働者数
日経平均株価	消費財	資金需要	所定外給与

金融経済月報を対象に LDA を用いて、トピック分類すると、個人消費、公共投資、在庫、設備投資、物価、マネーサプライ、鉱工業生産、企業金融、マーケット、貿易、資金供給、企業収益の 12 トピックに分類することができた。各月のトピック割合の推

移を図示すると以下になる。

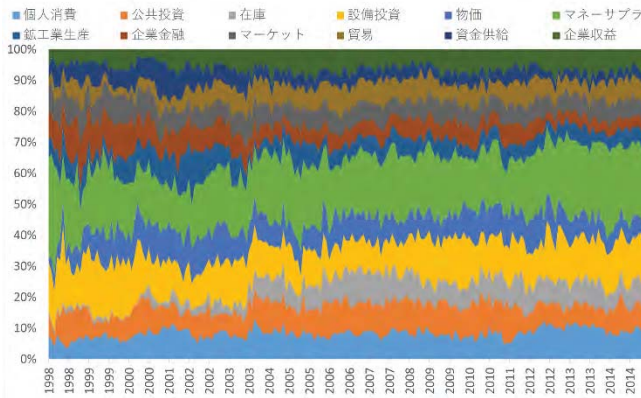


図 3 トピック推移

3.2 センチメント付与

まず、金融経済月報をトピックに分類せずに全文に対して LSTM RNN モデルでセンチメント指数を算出すると以下のようなになる。先行研究[3]とほぼ同じ結果となった。

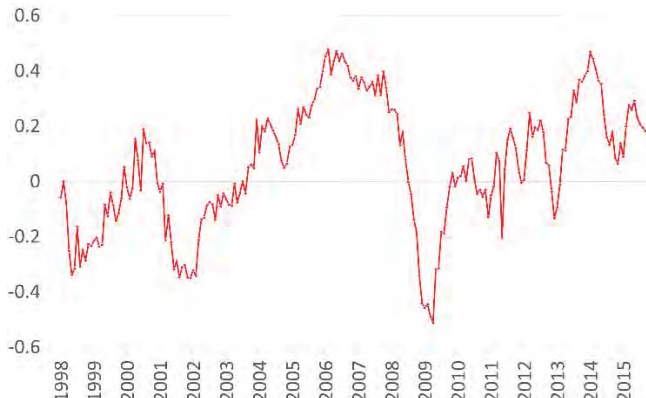


図 4 金融経済月報のセンチメント指数

一方、LDA でトピック分類を行い、トピック毎に対してセンチメントを付与し、全体に対する寄与度をトピック別に算出すると以下のようなになる。

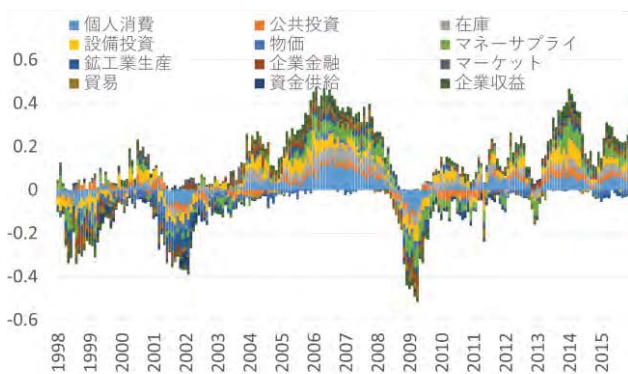


図 5 トピック別の寄与度

トピック別のセンチメントを見ると、ほとんどのトピックは同様な正負の推移を見せている。2015 年のセンチメントの寄与度に注目すると。物価トピックは全体に対してマイナスの寄与を与えている。これは、アベノミクスにより景気は回復し、物価以外のマクロ経済は良い状態に対して、物価は政策目標である 2%に届いていないため、金融経済月報にもそれが表れているのではないかと考える。

4 まとめ

本研究では、先行研究[3]で行われた、金融経済月報を対象にした日本銀行の景気に対するセンチメントの指数化に対して、LDA によるトピック分類を加えることで、センチメントをトピック別に算出し、寄与度に分解することで、どのマクロファクターがセンチメントに影響を及ぼしているかについて、考察を行った。

今後の展望としては、よりリアルタイム性を出すためにテキストの発行以前において、マクロ経済指数を外挿することによりリアルタイムな指数の予測を目指す。

参考文献

- [1] 山口 和孝: ニューラルネットと遺伝的アルゴリズムを用いた株式売買支援システム, (2002)
- [2] D. Peramunetilleke, R. K. Wong. : Currency exchange rate forecasting from news headlines, Australian Computer Science Communications Volume 24 Issue 2, (2002)
- [3] 山本 裕樹, 松尾 豊: 景気ウォッチャー調査の深層学習を用いた金融レポートの指数化, 第 30 回人工知能学会全国大会,(2016)
- [4] 塩野 剛志: 文書の分散表現と深層学習を用いた日銀政策変更の予想, 第 16 回 SIG-FIN,(2016)
- [5] 伊藤諒, 須田真太郎, 和泉潔: トピック別極性値付与方法による FOMC 議事録の評価, 第 17 回 SIG-FIN,(2016)
- [6] Blei, D. M., Ng, A. Y., Jordan, M. I. : Latent Dirichlet Allocation, Journal of Machine Learning Research, pages 993-1022, (2003)

深層学習と高頻度データを用いた株式注文状況の推定

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: For algorithmic trading, it is important to reduce market impact and opportunity costs that closely related to market liquidity. In this work, we propose a tick-based approach to prediction of the liquidity. Our method utilizes order data encoded according to its flexibility and a Long Short-Term Memory(LSTM) that predict a next order. Accuracy of the model outperforms by a large margin maximum occurrence ratio of order labels. Furthermore, we examine the embedding layer of the trained model and find out that it obtains difference and similarity between each order.

1 はじめに

近年の金融市場では、HFT (High Frequency Trading) やアルゴリズム・トレードといった機械的取引が台頭している。いずれもコンピュータを利用して自動的に発注と売買を行う点で共通するが、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、アルゴリズム・トレードは比較的低頻度にかつより高度な戦略を用いて執行する。マーケット・インパクト、機会コストといった取引コストを考慮しつつ、イントラデイでの大口取引を小口に分割し執行することが求められる。

アルゴリズム・トレードの主な執行戦略としてはVWAP (volume-weighted average price) 戦略 [1] が挙げられる。VWAP 戦略とは、取引執行価格を出来高の加重平均をとった価格に近づくように、大口注文を小口に分割し発注する戦略である。さらに、深層学習による価格予測をVWAPと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る。取引コストをいかに縮小するかは、アルゴリズム・トレード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][3]。

さらにミクロな間隔では、執行自身が与える需給の影響により取引が不利になるコストであるマーケット・インパクトや、板の希薄さによって予定した価格での執行ができないリスクである機会コストなどに関する研究が行われている。これらのコストには、市場の需給を表す板のスプレッドや厚さ、回復速度といった流動性が大きく影響する。

一方、金融市場の電子化と高速化による注文量の増加にともない、蓄積される注文情報も膨大化しており、大量の学習データが必要なディープラーニングへの応用の土台は整っている。J. Sirignano は、板情報に対するデータドリブンなアプローチとして、オンラインで学習できる点、大量のデータを扱える点、確率を出力できる点をメリットとして指摘した上で、ディープラーニングは最良の選択肢であると述べている [4]。

2 関連研究と本研究の目的

深層学習を用いて市場予測を行った関連研究を紹介する。

N. Du らはマーク付き点過程の強度関数をRNNを用いた非線形関数とすることで、約定の方向とその時刻の予測を行い、通常のRNNより良い成果を得た [5]。J. Sirignano は、板に対してファットテールに強いニューラルネットワークを提案し、ベストアスクとベストビッドの分布を予測し一定の成果を得た [4]。やはり前述の通り、多次元の情報を持ち非線形で複雑に作用する情報をモデリングする場合にディープラーニングは有効であると言える。

取引主体は何らかの情報を得た上で発注する。市場に集まる注文のパターンにはその背後にある情報が反映されると考えられ、深層学習を用いることでそのような注文パターンを獲得できる可能性がある。

そこで本研究では、板の変化量や価格の変化の予測をティックベースで行うことによって、マーケット・インパクトコストや機会コストに影響を与える将来の板の厚さや板の回復度を予測できるモデルの構築と分析を行う。応用として、あらかじめ流動性の減少を予測し、早期に執行を行うといったことなどが考えられる。

*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研究室, 〒 113-8654 東京都文京区本郷 7-3-1, E-mail: m2016dtashiro@socsim.org

3 LSTMを用いた注文状況予測

ここでは、モデルへの入力の前処理であるオーダーの符号化について説明する。次に、LSTMを用いて直後のオーダーを予測するモデルの構築を行う。

3.1 オーダーの符号化

まず前処理の方法について述べる。オーダーの自由度には指値注文/成行注文/キャンセル注文といった注文のタイプの他、価格、注文量などがある。ここではタイプと価格のみに限定する。まず各指値注文に対して、注文の入った時刻の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、指値価格と仲値との差の絶対値を保持する。

次に Figure 1 に示すフローに沿って、各注文に識別符号を与える。まず、注文のタイプとして売成行/買成行/売指値/買指値/売キャンセル/買キャンセルに分ける。成行注文二種に関しては、そのまま識別符号を付与し、指値注文とキャンセル注文に関しては、仲値からの価格差によって分類後、符号を付与する。なお、オーダーと識別符号間の変換は一意であり、符号からオーダー、つまりタイプと価格差へは復元可能となっている。

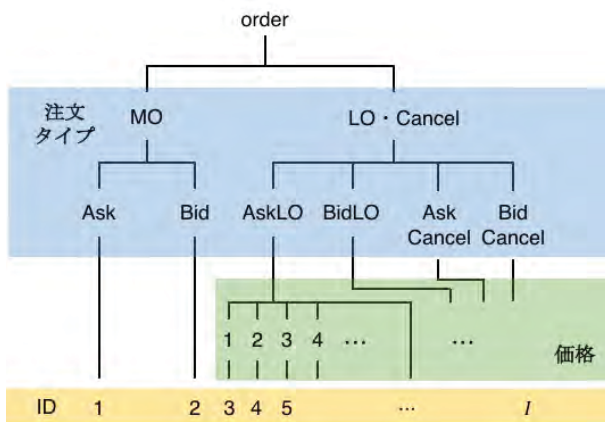


Figure 1: オーダーの符号化フロー。注文タイプ MO/LO/Cancel はそれぞれ/Market Order/Limit Order/Cancel 注文を、Ask/Bid はそれぞれ売り/買いを指す。

3.2 問題設定

本節では、オーダー予測の問題設定を行う。モデルの入力となる系列の集合を $C = \{S^1, S^2, \dots\}$ と表す。各系列 S^i は日毎の寄り付き後から引け前までの前場と後場のいずれかのザラバに対応する。ザラバと板寄せでの注文の入り方では意味が異なるため、ザラバのティックデータだけを扱い、同日の前場と後場は異なる系列

として扱う。ここで、 t 番目のオーダーを x_t とおくと、各系列 S^i は $S^i = (x_1^i, x_2^i, \dots)$ で表される。各オーダー x_t^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、過去のオーダー系列 $(x_1, x_2, \dots, x_{t-1})$ が与えられたときの次のオーダー x_t 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 $p(x_t|x_1, \dots, x_{t-1}; \theta)$ 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。

本研究のモデルは RNN Encoder-Decoder[6][7] を模して設計した。学習後、次のオーダーの予測を繰り返してシミュレートすることで t から T 先までのデコードする場合の条件付き確率分布は以下ようになる。

$$p(x_t, \dots, x_{t+T}|x_1, \dots, x_{t-1}; \theta) = \prod_{t'=t}^T p(x_{t'}|x_1, \dots, x_{t'-1}; \theta) \quad (1)$$

複数オーダーの予測の場合は、ビームサーチを用いればよい。同時確率を最大とするオーダーの系列を探索し、それを最適なオーダー系列とし、さらに板に足し合わせるにより板の変化量や価格を予測することができる。

3.3 モデル

モデルの RNN には、LSTM (Long Short-Term Memory) [8] を用いる。LSTM は、再起的な入出力を行うノードの活性化関数として恒等関数を導入することで、通常の RNN で長い系列データを学習する際に発生する勾配消失問題を回避する。さらに、メモリノードと 3 つの判断ノードから構成される集合体を一つのメモリセルとして扱うことにより表現力の低下を防ぐ。

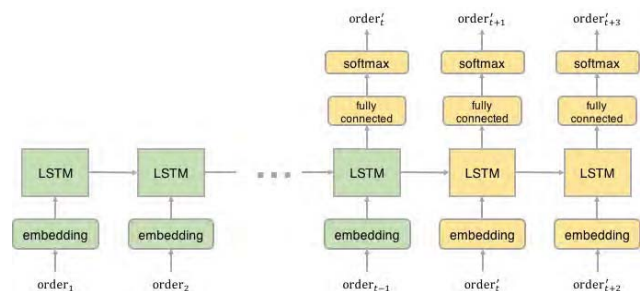


Figure 2: モデルの時系列展開図。

本研究で用いるモデルを時系列で展開した様子を Figure 2 に示す。onehot ベクトルで表される t 時点でのオーダーと embedding 行列 W_{embed} の積をとりオーダーの埋め込みベクトル e_t を得る。 e_t と $t-1$ 時点で

の LSTM セルの出力 h_{t-1} が t 時点での LSTM セルの入力となる。ここで t 時点でのメモリセルのベクトルを c_t , 重みを W_{LSTM} とおくと, LSTM の出力は次のように更新する。

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ g_t \end{bmatrix} = \begin{bmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \text{tanh} \end{bmatrix} W_{LSTM} \begin{bmatrix} h_{t-1} \\ e_t \end{bmatrix} \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (3)$$

$$h_t = o_t \odot \tanh(c_t) \quad (4)$$

さらに, LSTM の出力は全結合層 (W_{full}), softmax 層へと経て出力の条件付き確率を得る。

3.4 学習

(5) 式の負の対数尤度の最小にするようパラメータ $\theta = (W_{embed}, W_{LSTM}, W_{full})$ を探索し, 最適パラメータ $\theta = \theta^{opt}$ を得る。

$$\mathcal{L} = - \sum_{i=1}^C \sum_{t=1}^{S^i} \log p(x_t^i | x_1^i, \dots, x_{t-1}^i; \theta) \quad (5)$$

ただし, $C = |C|$ かつ $S^i = |S^i|$ である。

4 実験

4.1 データセット

データセットには 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。使用する銘柄と期間は大和ハウス工業 (銘柄コード 1925) の 2013 年 7 月 1 日から 2014 年 6 月 30 日までと, キリンホールディングス (銘柄コード 2503) の 2013 年 7 月 1 日から 2014 年 3 月 31 日までとした。各銘柄のデータの日数を 8:2 に分割し, 日時の古い期間のデータを学習データ, 新しい期間のデータをテストデータとする。

4.2 前処理

3.1 で述べた手法で, オーダーの前処理を行う。本実験では, オーダーの自由度を注文タイプと価格のみに限定する。価格の深さの自由度は 20 として, それ以上の価格に入ったオーダーは 20 として分類した。前処理後, オーダーの自由度を表す識別符号のサイズ I は $I = 82$ となる。分類後の主なオーダーの出現回数比を Table 1 に示す。

Table 1: 銘柄 1925 における分類比率 (比率による降順)。

分類	注文タイプ	価格差	比率
	BidLO	1	0.076
	AskLO	1	0.074
	BidCancel	1	0.058
	AskCancel	1	0.057
82 分類 (注文タイプ+価格)	BidLO	2	0.056
	AskLO	2	0.056
	⋮		
	AskMO	—	0.019
	BidLO	5	0.018
	BidMO	—	0.018
	⋮		
	BidLO	—	0.259
	AskLO	—	0.251
6 分類 (注文タイプ)	BidCancel	—	0.231
	AskCancel	—	0.223
	AskMO	—	0.019
	BidMO	—	0.017

4.3 実装

(5) 式が最小となるように, バッチサイズを 5 としたミニバッチを用いた確率的勾配降下法で最適化を行う。同系列内では LSTM のメモリセルの状態は保持しつつ, 新しい系列セットを用いる際に状態を初期化する。学習時には, 100 オーダー毎に勾配情報を破棄した。ハイパーパラメータは epoch 数 50, 隠れ層のユニット数は全て 40, ドロップアウトの比率 0.5, LSTM の層 1 とした。深層学習フレームワークとして chainer を使用し実装した。

5 結果と考察

5.1 モデルの評価

テストデータを用いて, 学習済みモデルによる予測と評価を行う。まず, LSTM のメモリセルの状態を得るため, 予測を行わず先頭から 100 オーダーを入力する。次に, 残りの系列を用いて 1 オーダーごとの予測と評価を逐次的に行う。モデルは直後の 1 オーダーのみを未知のものとし, Encoder-Decoder モデルのごとく出力を自らの入力として与えることはここでは行わない。

結果の評価は, 前処理での分類による 82 分類そのままの評価と, 6 分類での評価を行う。82 分類では, 出

力で得られる条件付き確率の最大となる符号を採用し、正解ラベルと比較する。6分類評価では、その82分類で予測した注文のタイプと価格のうち注文のタイプだけを対象とした6分類での評価を行う。例えば、得られた予測がタイプ AskLO で価格差2で正解ラベルがタイプ AskLO で価格差3の場合、82分類では誤と判別するのにに対し、6分類評価では正と評価する。6分類評価では、粒度を粗くすることによって、モデルが注文の方向性に対してどの程度説明できるかを評価する。

結果を Table 2 に示す。82分類と6分類いずれも、Table 1 に示す出現回数最大の比率を上回る。このことから、モデルは十分に学習しており、オーダーの予測は可能であると言える。

Table 2: 評価方法と精度。

評価方法	銘柄	精度
82 分類	1925	0.201
	2503	0.240
6 分類	1925	0.460
	2503	0.439

5.2 隠れ層のクラスタリング

次に embedding 層の重みを分析する。embedding 層の各オーダーに対応する重みは、embedding 層での各オーダーを表す固有のベクトルに対応する。この40次元82個のベクトルに主成分分析を施す。得られた第一主成分から第十主成分までを用いて k-means クラスタリングを行う。結果を第一主成分と第二主成分上にプロットしたものを Figure 3 に示す。

まず注文のタイプの全体的な分布だけを見ると、 $x = -0.3$ 付近を境界として売り側/買い側で大まか分類できていると言える。特にキャンセルは、指値/キャンセルと売り/買いでいずれもはっきりと分かれる。さらに PC1, PC2 上では線形の関係にあるように見える。

クラスタリングに関しては、キャンセルは売り/買いだけでなく価格の大小によっても分かれている。一方、指値注文は売り/買いに分類できていない。ここから、モデルが指値に関する情報よりもキャンセルに関する情報を捉えていると解釈できる。ただ、このクラスタリングは embedding 層の重みの PC1 と PC2 のみによるものであり、複雑で非線形なモデルであるニューラルネットワークでは、さらに上の層で指値の情報を識別する機構がある可能性もあり、分析の余地が残る。

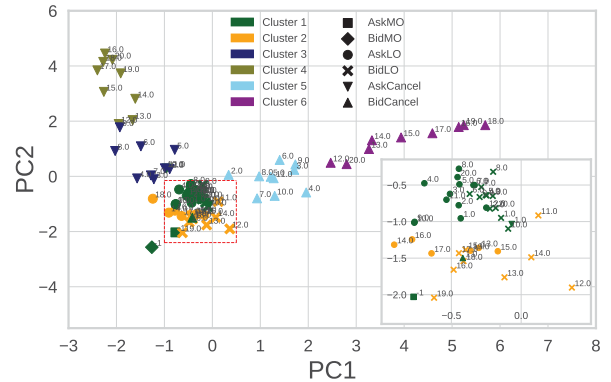


Figure 3: embedding 層の主成分分析とクラスタリングの結果。PC1: 第一主成分, PC2: 第二主成分。各プロットに対応する数字は各指値/キャンセル注文の仲値からの価格差である。成行注文は-1として表現する。プロットの形は注文のタイプを表し、色はクラスタリングの結果を表す。図中右下の図は、赤破線枠の指値の注文が多く集まる範囲の拡大図。

6 まとめ

本研究では、オーダーの予測を行い、さらにニューラルネットワークの重みからオーダーの類似性を見ることができた。具体的には、前処理としてオーダーの符号化を行い、LSTM を用いて次のオーダーを予測するモデルの学習・予測の実験を行った。さらに学習済み embedding 層の重みを分析することで、各重みに対応するオーダーのベクトルのクラスタリングを行った。今回のオーダーの予測以外の問題でも、オーダーベースでニューラルネットワークを学習することによって、オーダーの関係性や類似度が得られる可能性がある。問題設定によって、その関係性がどのように変わるかは興味深いテーマだと思われる。

本研究のモデルへの入力、板の状態に関する情報を含めない、オーダーベースでかつ価格情報は仲値からの価格差を用いた。しかし板の状態によってオーダーの意味は変わると考えられ、改善の余地がある。例えば、同じ価格差のオーダーであっても、スプレッドが大きい時ではスプレッドの中に入り得、逆ではスプレッドの外に入ることになる。また 100 円に対する 1 円と 1000 円に対する 1 円とでは意味が異なるように、価格の大きさに対する価格差が持つ意味も異なる。

これを踏まえて今後の展望としては、板の状態やオーダーの量、時間差といった入力情報の拡充、価格差の比率によって前処理の分類を行うなどの工夫が挙げられる。

参考文献

- [1] R. Almgren and J. Lorenz. Adaptive Arrival Price. *Algorithmic Trading III, Institutional In-*

vestor, 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] 杉原慶彦 「執行戦略と取引コストに関する研究の進展」『金融研究』第31巻第1号, 227–292頁, 2012年.
- [4] J. Sirignano . Deep learning for limit order books, *arXiv 1601.01987v7*, 2016.
- [5] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent Marked Temporal Point Processes Embedding Event History to Vector. *Knowledge Discovery and Data Mining*, 2015.
- [6] I. Sutskever, O. Vinyals, and O. Le. Sequence to sequence learning with neural networks. *Neural Information Processing Systems*, pp. 3104–3112. 2014.
- [7] K. Cho, B. Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the Empirical Methods in Natural Language Processing*, 2014
- [8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, pp. 1735–1780. 1997.

二重マッピングとSCWによる株価変動予測

Predicting stock fluctuations using Two-level Mapping and SCW

福田 ムフタル^{1*}
Muhtar Fukuda¹¹ 名古屋産業大学環境情報ビジネス学部¹ Faculty of Environmental and Information Studies, Nagoya Sangyo University

Abstract: Due to high uncertainty in the stock market, it is difficult to predict the future fluctuations of stock prices even if we use the state-of-the-art techniques of machine learning, such as Deep Learning. However, in some cases with choosing an appropriate machine learning algorithm, feature values and outputs for the prediction, we can have desirable predicted results, especially on short-term stock fluctuations about some market indices. Some initial reliable results have been achieved in our related work, by using Soft Confidence-Weighted (SCW) Learning, which is one of online learning. In this paper, we propose a predicting method using two-level mapping and SCW. We will focus on feature transformations using the two-level mapping. The first one is based on the mathematical concept of the Singular Value Decomposition (SVD), to get strong convergence and higher accuracy. The second one is to make the predicted Fluctuation Strength (FS) more precisely, in which we use pre-learned outputs and do relearning.

1 はじめに

株式市場には、常に需要と供給があり、そのバランスによって市場価格が形成され、またはその価格で取引されると今までは考えられてきた。そして投資家たちはそれぞれの判断基準に基づいて株式投資を行ってきた。しかし、ICT技術を基盤とした株式投資環境の発展に伴い、個人投資家を含めた市場参加者の拡大、マーケットのグローバル化、SNS等を手段としたソーシャルメディアネットワークの普及により、株価形成に影響を与える要素が複雑化し、不確実性も増大した。そういったことから市場動向を伝統的な経済学や金融工学の理論だけでは説明できないケースが増え、一定以上の投資パフォーマンスの確保が一層困難になった。株価変動はランダム・ウォークであると言われる場合がある。そうであれば一定の確率分布に従うはずであり、何らかの理論的な説明ができる。しかし、最近の変動はランダムどころか、何重にも攪乱に遭い、本来の姿が消えてさえているとも思える。

上で述べたように、説明しにくくなった金融市場において、経験の浅い個人投資家だけではなく、プロの機関投資家にとっても如何にして利益の最大化とリスクの最小化を実現できるかがこれまで以上に重要になってきた。一方、人工知能の分野では前述の状況と並行し

て、本来の姿が表面的には見えないが、大量に蓄積された過去のデータから潜在的な相関関係を見出し、それらに対して機械学習を行い、様々な予測や判断に利用できる機械学習手法が考案され、著しい成果が上がっている。その中で特に注目を浴びているのがディープラーニング(DL)[1]であろう。

ディープラーニングは人間の脳の働きをモデル化したものの一つであり、ベテランの金融アナリストや投資家たちが過去の市場動向や長年の投資経験から多くを学んだのと同様に学習させ、投資判断に応用できると期待されている。その応用例として Deep Belief Network(DBN)[2]を用いた為替レート変動の予測[3]や上場企業倒産の予測[4]等の研究が報告され、ある程度の有効性が示されている。しかし、多階層のネットワーク構造であるDLを用いて金融時系列データの学習を行う場合、依存性の高いパラメータが多数あり、その値を微調整するだけで収束性と予測結果が激変するし、かなりの学習時間も要する。金融時系列データには時間的従属関係があり、学習期間と学習のための入力データである属性値(feature)の構成要素を常に最新のものに更新して再学習を繰り返す必要があり、更新のたびに収束性と予測結果に大きなばらつきが生じると、予測結果の信頼性が揺らぐことになる。

学習の安定的な収束とよい予測結果が期待できるもう一つの機械学習モデルとして Support Vector Machine (SVM)[5]があり、SVMによる高精度のトレンド予測

*連絡先：名古屋産業大学環境情報ビジネス学部
〒488-8711 愛知県尾張旭市新居町山の田 3255-5)
E-mail: fukuda@nagoya-su.ac.jp

を実現した研究 [6, 7] もある。しかし、SVM を用いて金融時系列データの学習と予測を行う場合、その正解率が学習データのサイズと属性値の選択に依存するところが大きい [7]。また、金融時系列データは非線形であり、SVM モデルの予測正解率を上げるためには、非線形カーネルを使用する必要がある。非線形カーネルにはやはり複数のパラメータがあり、SVM の振る舞いにかなり影響を及ぼすため、その調整を慎重にかつ時間をかけて行わなければならない、学習期間と属性値を常に更新する機械学習では問題となる。

また、以上の各モデルを含め、従来の予測モデルの殆どは翌日、あるいは次の一定期間の株価変動の方向性一致率に重点をおいており、ある銘柄の次の日の株価を「上がる」と予測し、実際の株価が例えば 0.01% でも上がった場合、それが正解とされる。それは理論的には正解となっても、実際の投資目的には不十分である。株価予測の場合、株価が上がるか下がるかだけではなく、その騰落の程度を定量的に示すことがより重要である。買って売ってもコストに見合ったリターンが望めない場合は何もせず、一定以上のリターンが望めるタイミングで売買を行うのが実用的である。

本研究では、より最近の過去データが次の株価変動の予測により大きな影響を与え、かつ近隣の学習データができるだけ独立であるべきだと考えており、ディープラーニングも SVM も基本的にはバッチ学習であるため、本研究の考え方にはそぐわない。

著者は先行研究として、以上のような諸問題を避けつつ、株価が上がるか下がるかの方向性だけではなく、株式の売買タイミングにも重点をおいた株価変動予測の一手法 [9] を提案し、一定の有効性を示した。その中では、機械学習アルゴリズムにオンライン学習 (逐次学習) の一つである Soft Confidence-Weighted Learning (SCW)[8] を利用した。

本研究では、更に進んで、二重マッピングと SCW[8] による機械学習を組み合わせることで、特に売買タイミングを正確に知るための予測騰落強度の精度向上をはかる手法を提案する。第一段のマッピング (一次マッピング) では、収束性と安定性を備えた機械学習を実現するため、一次入力データに対して線形代数の概念である特異値分解に基づいたマッピングを行い、その結果に対して機械学習を行う。第二段のマッピング (二次マッピング) では、同じ教師データと一回目の学習過程で得られる方向性判断値、及び重みを利用し、株価騰落率をより正確に反映できるような騰落強度を得るためのマッピングを行い、その結果に対して再学習を行う。次の第 2 節では本研究の基本的な考え方について述べる。第 3 節では本研究独自の学習用データセットについて述べる。第 4 節では本研究で提案した手法について述べる。第 5 節では日経平均株価指数を対象にした予測実験について述べる。

2 基本的な考え方

本研究では、文献 [9] でも述べた通り、以下の 6 つの考え方に基づいた株価変動の予測を行う。

1) 予測しやすい、かつリスクが低い銘柄・指数だけを予測の対象とする。これまで、日経平均株価指数 [10] を最も安定的に学習・予測できることを実験で確認しており、今回もこの指数を主に扱う。日経平均株価指数に一对一に連動する ETF が幾つもあり、流動性が非常に高く、いつでも売買でき、よく知られている指数でもあるため、実用的な株式投資対象としても好都合である。それに比べて個別銘柄の株価はいろいろな要素の影響を受ける。次の日に上がると予測して株を購入し、次の市場が始まる前に悪材料が出た場合、予測が大きく外れることになる。格付け機関によってレーティングが変更されたり、目標株価が僅かに変更されただけで株価が激変し、関連銘柄にも影響が出る場合がよくある。特に個別銘柄に関しては市場ではなく、格付け機関によって価格が形成される場合も有りである。

2) 一つの銘柄の株価は独立に動くことは殆どなく、複数の銘柄と直接または間接的に影響しあって相関的に変動する。日経平均株価指数に関して言えば、その構成銘柄の株価変動の影響を受けやすい。本研究では、一つの銘柄の値動きを予測するためには関連性の高い複数の銘柄の時系列データの組み合わせを入力データとし、その関連銘柄群 (ベース銘柄集) を一定の手順 (後述) に基づいて常に機械的に選択し直す。株価は、需給、株価材料、投資家心理、為替、経済政策、経済状況や地政学、要人の発言、マスメディアや SNS からのニュース (例えば偽ニュースであっても) など様々な要因の影響を受けることは周知の通りであるが、第 1 節で述べたように、現在のマーケットを取り巻く環境は劇的に変化しており、株価変動要因になり得る何らかの出来事が発生した場合、瞬時に株価に反映されるため、基本的にベース銘柄集の株価時系列データだけを学習の入力データとする。

3) 学習期間 (図 1 の (c)) として予測ポイント (図 1 の (b)) から数日間離れた最終学習ポイント (図 1 の (a)) より過去 3600 日分 (約 15 年間) の株価時系列データを使用するが、ベース銘柄集と最終学習ポイントを常に更新し続けることにより、最新の株価変動要因が株価時系列データに反映されている状態を可能な限り維持し続ける。

4) 本研究では、一日の株価変動ではなく、数日間のウィンドウ内の株価変動を学習、又は予測のターゲットにする。このウィンドウをターゲットウィンドウと呼び、学習ターゲットウィンドウ (図 1 の (d)) と予測ターゲットウィンドウ (図 1 の (e)) の二つに分けて考える。日数がターゲットウィンドウのサイズになる。ここでもやはり予測しやすいターゲットウィンドウサイズを

考える。

5) 学習ポイント (図1の(a)) から見た過去一定期間内の株価のトリプルボトムとトリプルトップの学習ポイントから見た騰落率を入力データの一次属性値 (feature) とし、その期間を属性ウィンドウ (図1の(f)) と呼ぶ。そして、ベース銘柄集の各銘柄の属性値の組み合わせを一次入力データとする。

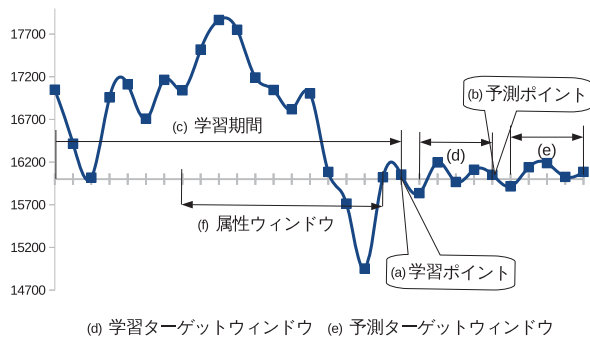


図1: 学習ポイント・予測ポイント・ウィンドウ

6) 本研究では、株価変動の方向性、即ち次の期間内に株価が上るか下がるかだけでなく、売買タイミングの割り出しにも重点をおいた予測を行う。騰落の程度を表すことのできる定量的な指標として「騰落強度」(式(2))を使用し、騰落強度が大きければ大きいほど、売買タイミングに近付いたと判断する。

SCWを含めたオンライン学習では、入力 $x_t \in \mathcal{R}^d$ と機械学習によって更新される重み $w \in \mathcal{R}^d$ に対して、

$$\hat{y}_t = \text{sgn}(x_t \cdot w), \hat{y}_t \in \{+1, -1\} \quad (1)$$

でもって入力 x_t が属するクラスを予測する。

本研究では、全ての入力 x_t が d 次元空間の点であり、それらの点が法線ベクトル w を持つ超平面によって二つに分離され、点と分離超平面との距離が大きければ大きいほどそのクラスへの属性が強いと見做し、その距離(符号付き)を「騰落強度」と定義する。即ち、騰落強度 FS (Fluctuation Strength) を次のように定義する。

$$FS = \frac{x_t \cdot w}{|w|} \quad (2)$$

ここで、 $|FS|$ は騰落の強さ、 $\text{sgn}(FS)$ は上がるか下がるかの方向性になる。しかし、騰落強度が最も大きい場合、その騰落率も一番大きくなるという訳ではなく、その傾向が強いことを表す量になる。騰落強度の妥当性に関しては、先行研究 [9] での一連の実験によって検証済みであるが、本研究の主な目的の一つは株価騰落率をより忠実に反映できるような騰落強度を獲得することである。DL、あるいは最小二乗法を用いた回帰計算等によって株価騰落率そのものを推定することも考えられるが、本研究の実験範囲内では期待できる結果が得られなかった。

3 機械学習用データセットの準備

機械学習を行う場合、学習アルゴリズムと並んで学習のためのデータセットがその学習と予測結果に大きな影響を与えることがよくある。さまざまな機械学習の典型的なアルゴリズムの例題プログラムとそれ用に提供されるデータセットを Web などから見つけては実行すると、見事な結果が出たりする。さすがに今どきの機械学習であると感激したりもする。しかし、いざその学習アルゴリズムのサンプルプログラムを自分の研究中のデータセットに合わせてカスタマイズし実行してみると、期待はずれの結果が出てきたりする。従って、実用レベルの機械学習や予測結果を出すためには独自の学習データセットを用意することが重要になってくる。

本研究では次のようにして独自の学習データセットを用意する。まず第一に、ベース銘柄集を自動的に選択する。第二に、先行研究 [9] と同様、ベース銘柄集の各銘柄の属性ウィンドウ (図1の(f)) におけるトリプルボトムとトリプルトップの学習ポイント (図1の(a)) から見た騰落率の組み合わせを入力データの一次属性値とする。第三に、従来の研究では株価の騰落率に関して 0 を基準にし、学習ターゲットウィンドウ (図1の(d)) の学習ポイントから見た騰落率を計算し、騰落率 ≥ 0 は「上がり」(+1)、そうではない場合は「下がり」(-1) とラベル付けをして学習と予測を行うことが多いが、ここでは「上がり」と「下がり」の基準値を別々に決め、学習と予測も別々に行う。第四に、入力データの一次属性値に対して特異値分解の概念に基づいたデータ変換を行い、新たな入力データを生成する。これは所謂二重マッピングの第一段に相当する。

以上の第二の方法で決める一次属性値の妥当性であるが、本研究の可能な範囲内でいくつかの種類値を属性値として学習と予測実験を行い、トリプルボトムとトリプルトップベースの属性値が最もよいことが分かればよしとする。学習データセットの準備の詳細に関しては以下の各サブ節で述べることにする。

3.1 ベース銘柄集の自動選択

日経平均株価指数、あるいは特定の業種に連動するように設定されたETFなどは独立に動くことは殆どなく、複数の銘柄と直接または間接的に影響しあって相関的に変動すると思われる。日経平均株価指数を考えた場合、構成銘柄リストに含まれていない場合でさえその可能性がある。従って、一つの銘柄の値動きの予測を行う場合、連動性の高い銘柄群の時系列データの組み合わせを学習の入力データとするのが自然であろう。ここでは、そのような銘柄郡をベース銘柄集と呼ぶ。

ベース銘柄集をどの株式市場の上場銘柄から選ぶのか、国内外から同時に選ぶのかなどを検討する必要がある。

るが、今回は日経平均株価指数の変動を学習と予測の対象とし、東京証券取引所で取引されているものの中で一定の条件を満たす銘柄をベース銘柄集の選択対象とする。

ベース銘柄集を次のような手順で選択する。便宜上、選択対象の全銘柄の集合を S 、学習期間を T 、銘柄 $s \in S$ の学習期間中の売買成立日数を T_s 、 s の最大連続売買不成立日数を N_s ($T \geq T_s + N_s$)、基準最低取引成立日数を T_0 、基準最大連続不成立日数を N_0 とする。

Step0 予測対象銘柄 $s_0 \in S$ を一つ決定する。話を進めやすくするため、全期間中において売買不成立日のない銘柄を選ぶ。即ち、 $N_{s_0} = 0$ になるような銘柄を選ぶ。日経平均株価指数は正にこの条件を満たす。

Step1 $S_1 = \{s \in S \mid T_s \geq T_0, N_s \leq N_0\}$ とする。即ち、できるだけトータルで売買不成立日が少くかつ連続の売買不成立日の少ない銘柄を選ぶ。 $s \in S_1$ に売買不成立がある場合、補間を行う。

今回の研究では $T = 3600$ 、 $T_0 = 0.99T = 3564$ 、 $N_0 = 1$ としており、そうした場合、 $|S_1|$ は大抵 1100 前後におさまる。

Step2 S_1 から一様平均流動性の高い上位 $P_1\%$ の銘柄を選び、 S_2 とする。ある銘柄 $s \in S_1$ の予測ポイントから過去 T 日間の株価終値、及び出来高の時系列をそれぞれ $p_t, v_t, t = 1, 2, \dots, T$ とした場合の s の一様平均流動性 $AvgLiq(s)$ を次のように定義する。

$$AvgLiq(s) = \left(\prod_{k=0}^K \left(\frac{1}{R} \sum_{t=1}^R p_t \cdot v_t \right) \right)^{\frac{1}{K+1}}$$

ここで、 $K \geq 0, 2^{K+1} \leq T, R = \lfloor \frac{T}{2^k} \rfloor$ 。

Step3 S_2 から一様平均騰落率の高い上位 $P_2\%$ の銘柄を選び、 S_3 とする。ある銘柄 $s \in S_2$ の予測ポイントから過去 T 日間の株価終値時系列を $p_t, t = 1, 2, \dots, T, T+1$ とした場合の s の一様平均騰落率 $AvgPC(s)$ を次のように定義する。

$$AvgPC(s) = \left(\prod_{k=0}^K \left(\frac{1}{R} \sum_{t=1}^R \left| \frac{p_t - p_{t+1}}{p_{t+1}} \right| \right) \right)^{\frac{1}{K+1}}$$

ここで、 $K \geq 0, 2^{K+1} \leq T, R = \lfloor \frac{T}{2^k} \rfloor$ 。

Step4 $S_3 \cup \{s_0\}$ から s_0 の s に対する一様平均絶対ベータ値の高い上位 $P_3\%$ の銘柄を選び、それらを s_0 のベース銘柄集 $S(s_0)$ とする。前の各 **Step** で s_0 が落とされる可能性があるため、ここで s_0 がベース銘柄集に入るようにしておく。本研究では、 $|S(s_0)| = N$ とし、 $P_1 = P_2 = P_3$ 、

$P_1\% \cdot P_2\% \cdot P_3\% \cdot |S_1| = N$ になるように P_1, P_2, P_3 の値を決めている。例えば $N = 320$ の場合、 $P_1 = P_2 = P_3 \simeq 66$ 。日経平均株価指数の場合、ベース銘柄集のサイズを 320 にしているが、対象銘柄によって増減する必要がある。

s_0 とある銘柄 $s \in S(s_0)$ の予測ポイントから過去 $T + L$ 日間の株価終値時系列をそれぞれ $p_t^0, p_t^s, t = 1, 2, \dots, T, T+1, \dots, T+L$ とし、 s_0 の s に対する一様平均絶対ベータ値 $AvgBeta(s_0 | s)$ を次のように定義する。

$$AvgBeta(s_0 | s) = \left(\prod_{k=0}^K \left| \frac{\alpha_k \cdot \beta_k}{\beta_k} \right| \right)^{\frac{1}{K+1}}$$

$$\alpha_k = \left\{ \frac{p_1^0 - a_1^0}{a_1^0}, \frac{p_2^0 - a_2^0}{a_2^0}, \dots, \frac{p_R^0 - a_R^0}{a_R^0} \right\}$$

$$\beta_k = \left\{ \frac{p_1^s - a_1^s}{a_1^s}, \frac{p_2^s - a_2^s}{a_2^s}, \dots, \frac{p_R^s - a_R^s}{a_R^s} \right\}$$

$$a_t^0 = \frac{1}{L} \sum_{i=1}^L p_{t+i}^0, \quad a_t^s = \frac{1}{L} \sum_{i=1}^L p_{t+i}^s$$

ここで、 $K \geq 0, 2^{K+1} \leq T, R = \lfloor \frac{T}{2^k} \rfloor$ 、 L は属性ウィンドウのサイズ、 a_t^0 と $a_t^s, t = 1, 2, \dots, T$ はそれぞれ L 日間の移動平均値である。 $K (= 7)$ を折り畳み回数と呼ぶことにする。一般的に個別銘柄の日経平均に対するベータ値を計算するが普通であるが、ここではその逆である。

以上の手順で日経平均株価指数のベース銘柄集 $S(s_0)$ を自動的に選択した場合、 $S(s_0)$ には日経平均構成銘柄の約 30% 前後が含まれておらず、 S_1 に含まれる日経平均構成銘柄を全部入れた上で残りの銘柄を上記の手順で選択してできるベース銘柄集 $S'(s_0)$ と $S(s_0)$ とで別々に学習・予測実験を行ったところ、 $S(s_0)$ のほうが良い結果になっており、この選択手順が妥当であると言える。また、日経平均株価指数だけに限って言えば、その構成銘柄の一部が場合によって必ずしも指数の変動により大きな影響を与えるとは言えないことも意味する。

3.2 一次属性値の決定

機械学習の場合、何を入力データの構成要素にするのが重要なファクターの一つであり、その構成の仕方によっては不十分な学習になったり、過学習に陥ったり、識別率に大きな影響を及ぼしたりする。入力データの構成要素として移動平均値と RSI (Relative Strength Index) の組み合わせを用いたり [6]、 n 日間の株価時系列データをそれぞれウィンドウ内と全学習期間内で正規化したデータの組み合わせを用いた [11] 研究がある。

本研究では、ベース銘柄集の各銘柄の属性ウィンドウ (図1の(f))内のトリプルボトムとトリプルトップの騰落率を一次属性値の構成要素とし、それらの組み合わせを一次属性値として使用する。

ベース銘柄集の一銘柄のある時点(学習ポイントか予測ポイント) t_0 における株価終値を p_0 , その属性ウィンドウのサイズを $L = 3 \times 2 \times stride + slack$, 属性ウィンドウ内の株価終値時系列を $p_t, t = 1, 2, \dots, L$ とした場合のトリプルボトムの時点 t_0 に対する騰落率を次のように求める。

Step1 $c_t = \frac{(p_t - p_0)}{p_0} 100, t = 1, 2, \dots, L$ とする。

Step2 $t_1, t_2, t_3 \in \{1, 2, \dots, L\}, t', t'' \in \{t_1, t_2, t_3\}, |t' - t''| \geq stride$ を満す下位の $\{c_{t_1}, c_{t_2}, c_{t_3}\}$ を選択する。即ち、時系列的に隣ではなく、互いに $stride$ 日以上離れた下位三つの値を選択する。

Step3 $\{c_{t_1}, c_{t_2}, c_{t_3}\}$ を元の時系列順に戻し、トリプルボトムの騰落率 (l_1, l_2, l_3) を得る。

上記の手順と同様、上位3つの値を選択することでトリプルトップの騰落率 (h_1, h_2, h_3) を求めることができる。今回は $stride = 2, slack = 2 (L = 14)$ としているが、ターゲットウィンドウ (図1の(d), (e))のサイズに合わせて $stride$ と $slack$ の値を適切に調整する。

次に、予測対象銘柄 s_0 のベース銘柄集

$$\mathcal{S}(s_0) = \{s_1, s_2, \dots, s_N\}, N = |\mathcal{S}(s_0)|$$

において、 $s \in \mathcal{S}(s_0)$ の時点 t のトリプルボトムとトリプルトップの騰落率を $\alpha_t^s = (l_{t_1}^s, l_{t_2}^s, l_{t_3}^s, h_{t_1}^s, h_{t_2}^s, h_{t_3}^s)$ とし、 $x_t^0 = (\alpha_t^{s_1}, \alpha_t^{s_2}, \dots, \alpha_t^{s_N})$, $t = 1, 2, \dots, T$ を入力データの一次属性値とする。一次属性値 $x_t^0, t = 1, 2, \dots, T$ を行に持つ行列を X^0 とする。便宜上、この行列を一次属性値行列、 x_t^0 を一次属性値ベクトルと呼ぶことにし、以降、同様の名称で参照する。

3.3 ラベル付け基準値の決定

入力データの一次属性値が決まったところで、次に問題になるのは教師データであるラベルを何を基準に決めるかということである。第2節でも述べたように、一日の株価変動ではなく、数日間のターゲットウィンドウ内の株価変動の学習と予測を行う。また、「上がるかどうか」と「下がるかどうか」の基準値を別々に決め、学習と予測も別々に行う。

学習・予測対象銘柄 s_0 の「上がり」、「下がり」それぞれの基準値は次のように決める。

まず、学習期間内の全学習ポイント (図1の(a)) から見たターゲットウィンドウ (図1の(d)) の平均騰落率 $PC(t)$ (Percentage change) を計算する。即ち、 p_t を学

習ポイント t の株価終値、 M をターゲットウィンドウのサイズ、 $p_{t-1}, p_{t-2}, \dots, p_{t-M}$ をターゲットウィンドウ内の株価終値の時系列とし、 $t = 1, 2, \dots, T$ に関して

$$PC(t) = \frac{(a_t - p_t)}{p_t} 100, a_t = \frac{1}{M} \sum_{i=1}^M p_{(t-i)}$$

を計算し、それらを昇順に並べ替えた

$$r_1, r_2, \dots, r_{\lfloor \frac{T}{3} \rfloor}, \dots, r_{\lfloor \frac{2T}{3} \rfloor}, \dots, r_T$$

の中の $LC = r_{\lfloor \frac{T}{3} \rfloor}$ を「下がり」の基準値 (Lower Criterion), $UC = r_{\lfloor \frac{2T}{3} \rfloor}$ を「上がり」の基準値 (Upper Criterion) とする。ラベル付けの基準値 UC と LC が決定されたところで、

$$y_t^U = \begin{cases} +1 & \text{if } PC(t) \geq UC \\ -1 & \text{otherwise} \end{cases}$$

$$y_t^D = \begin{cases} +1 & \text{if } PC(t) \leq LC \\ -1 & \text{otherwise} \end{cases}$$

$$t = 1, 2, \dots, T$$

によってそれぞれ「上がるかどうか」と「下がるかどうか」のラベル付けを行う。そして、

$$Y^U = (y_1^U, y_2^U, \dots, y_T^U)^\top, Y^D = (y_1^D, y_2^D, \dots, y_T^D)^\top$$

$X = X^0$ (第3.2節) とし、 $\{X, Y^U\}$ と $\{X, Y^D\}$ でそれぞれ「上がるかどうか」、「下がるかどうか」の学習ができるようになる。ここでも便宜上、 Y^U と Y^D をラベルベクトルと呼ぶことにする。

4 二重マッピングと二重学習

第3.2節の方法で作成される一次属性値行列 X^0 と第3.3節の方法で作成されるラベルベクトル $Y = Y^U$ (あるいは Y^D) でも学習を行えるが、期待できるほどの結果が得られない。そこで、先行研究 [9] では、

$$X^0 = U \Sigma V^\top \quad (3)$$

のように特異値分解し、 $X = X^0 V$ として属性値行列を作成し、この X とラベルベクトル Y に対して SCW アルゴリズムを利用した機械学習を行った結果、僅かではあるが実用レベルの成果が出た。即ち、一次属性値ベクトル x_t^0 を直接学習の入力とせず、 $x_t = x_t^0 V$ による変換を行い、ベクトル x_t を学習の入力とした。これは一種のマッピングであり、一次マッピングと呼ぶことにする。行列 V が一次マッパーになる。

SCW の機械学習アルゴリズムでは、入力ベクトル $x_t \in \mathcal{R}^d$ と学習によって更新される重み $w \in \mathcal{R}^d$ 対

して、 $\hat{y}_t = \text{sgn}(x_t \cdot w)$, $\hat{y}_t \in \{+1, -1\}$ を見て入力 x_t が属するクラスを予測する。本研究では $x_t \cdot w$ を方向性判断値と呼ぶことにする。この方向性判断値が限り無く 0 に近い場合でも判断を行い、方向が一致すれば、重み w の更新を行わず、次の入力を待つ。しかし、実際の株式投資では想定 of 基準値以上に上がる、あるいは基準値以下に下がるという予測ができなければ、中々決断できない¹。それが故に本研究での「上がるかどうか」、「下がるかどうか」の判断基準値を 0 ではなく、第 3.3 節で述べた方法で決めている。

しかし、それでもなお上がるならよく上るかどうか、下がるならよく下がるかどうかを定量的に計ることができれば、決断の更なる手助けになる。著者は、そのために式 (2) のように騰落強度 FS を定義し、その有用性を実験によって示した [9]。

本研究では一歩進んで、学習過程においては騰落率が大きければ大きいほど騰落強度が大きくなり、予測においては騰落強度が大きければ大きいほど騰落率も大きくなるという傾向を更に強くすることを考える。そこで、まず一次マッピングの結果に対して予備の学習 (一次学習) を行い、得られた方向性判断値、重み、及び教師データの元である騰落率を利用して新たなマッパーを作成し、その二次マッピングの結果に対して同じ SCW による再学習を行うという手法を導入する。本手法の詳細は次の通りである。

一次学習を行った結果、得られた重みベクトルを $w = (w_1, w_2, \dots, w_n)$ 、方向性判断値ベクトルを $d = (d_1, d_2, \dots, d_T)$ とすると、次の式が成り立つ。

$$X^0 V w^T = d^T \quad (4)$$

しかし、 d の各成分の絶対値の大きさと騰落ベクトル $f = (f_1, f_2, \dots, f_T)$ の各成分の絶対値の大きさの方向性が必ずしも一致しない。そこで、 f のある成分の絶対値が大きければ、それに対応する d の成分の絶対値も大きくなるような成分の数ができるだけ多くなるようにしたい。そうなれば、式 (2) の騰落強度 FS の識別能力が高まると期待できる。

上記の発想から、次の式 (5) を満たすようなもう一つのマッパー行列 M を作ればよいことになる。

$$X^0 V M w^T = \frac{|d|}{|f|} f^T \quad (5)$$

ここで、 $W = w^T w$ とし、 $W = U_1 \Sigma_1 V_1^T$ のように特異値分解し、更に式 (3) から、

$$M = \Sigma^{-1} U^T \frac{|d|}{|f|} f^T w V_1 \Sigma_1^{-1} U_1^T \quad (6)$$

になり、この M を二次マッパーとする。二つのマッパーが揃ったところで、一次属性値ベクトル x_t^0 , $t =$

¹少なくとも著者はそうである

$1, 2, \dots, T$ に対して一次マッパー V と二次マッパー M によって次の式のような二重マッピングを行い、その結果であるベクトル x_t を最終入力ベクトルとして再学習を行う。

$$x_t = x_t^0 V M, \quad t = 1, 2, \dots, T \quad (7)$$

5 予測実験

第 2 節で述べたように、予測しやすい銘柄・指数だけを予測の対象とする。これまでは、日経平均株価指数 [10] とそのベース銘柄集の各銘柄、更に日経平均株価に連動する複数の ETF に関して、学習と予測実験を行った結果、日経平均株価指数が最もよかった。その善し悪しを次の三つの観点から観察した。

一つは学習のエポック回数である。学習が 100% 終るまでにかかったエポック回数が日経平均株価指数よりも少ない回数で済んだ銘柄が銘柄数全体の 6% 前後であった。日経平均株価指数のベース銘柄集を他の銘柄のベース銘柄集として学習を行っても大きな違いがなかった。

二つ目はベース銘柄集に含める銘柄の数である。日経平均も個別の銘柄もその株価は関連性の高い銘柄と影響し合って変動することは同じであるが、日経平均に影響を与える銘柄が沢山あるのに対し、個別銘柄は必ずしもそうではない。関連性の薄い銘柄をできるだけベース銘柄集に含めないようにすると、学習用入力属性値ベクトルの次元が不十分になり、安定した学習ができない。属性値ベクトルの次元を適切なサイズにしようとすると、関連性の薄い、或いは関連性のない銘柄まで含まれてしまい、学習は安定的になるが予測は安定的に行えない。

三つ目は予測の正解率である。日経平均株価の予測の正解率が実験を実施した他のどの銘柄よりも高く安定的であった。日経平均株価に連動する複数の ETF と比べても平均的に 3% 前後の差があった。第 3.1 節の手順でそれらの ETF のベース銘柄を選択すると、日経平均のベース銘柄集と 1, 2 銘柄しか差がないにも関わらずである。

しかし、本研究の主な目的は二重マッピングと二重学習の有効性を示すことであり、ここでは今回提案した手法に基づいたシステムの予測正解率と騰落強度の売買タイミングにおける識別能力に関する実験結果だけ報告する。

5.1 実験に使用したデータ

今回の実験には、東京証券取引所に上場している銘柄のうち、東証 1 部、東証 2 部、JASDAQ を問わず、入手

できる株価データ全てを使用した。それには 3800 もの銘柄が含まれているが、第 3.1 節で述べたベース銘柄選択手順の **Step1** を実行するだけでその数が 1100 前後に減り、残りの各 **Step** の実行後に 320 銘柄になるように関連パラメータの値を決めており、一定の期間内の予測を行う場合、実際に使用されるのが 370 前後の銘柄である。実験には Yahoo!ファイナンス [12] のサイトから 2001 年 1 月から 2016 年 12 月までの 16 年分の時系列データをダウンロードして使用した。その時系列データには「日付」、「始値」、「高値」、「安値」、「終値」、「出来高」の他に「調整後終値」も含まれており、そのまま実験に使用した²。

5.2 予測正解率に関する実験

システムの予測正解率が何%までなら識別能力があると判断されるかは困難であるが、この実験では、一定の予測期間内の各タイミングで「上がるかどうか」(「下がるかどうか」)に関して予測を行い、「上がる(Yes)」(「下がる(Yes)」)と予測されたタイミング(その回数を A とする)の内、本当に上がった(下がった)回数(その回数を B とする)を勘定し、 $\frac{B}{A}100$ を正解率(%)とする。本研究では、「上がるかどうか」と「下がるかどうか」に関して別々に予測を行う。つまり、二つの予測器で別々に予測を行う。

今回は、ターゲットウィンドウ(図 1 の (d, e)) のサイズを 3 とし、2016 年 7 月 1 日から 2016 年 12 月 30 日までの 123 日間(6ヵ月間)を予測期間、それ以前の 3600 日間を学習期間として「上がるかどうか」と「下がるかどうか」別、及び一次マッピングと二重マッピング別に実験を行った。その結果を表 1 に示す。

表 1 の結果から、一次マッピングの場合の正解率は「上がるかどうか」と「下がるかどうか」の両方とも二重マッピングの場合より 3% 前後よいことが分かり、後者が悪化したのは二回目のマッピングで属性値ベクトルにノイズが加わったと考えられる。従って、「上がるかどうか」(「下がるかどうか」)だけを予測したい場合、二重マッピングを行わないほうがよさそうである。

マッピング	「上がる?」		「下がる?」	
	一次	二重	一次	二重
予測 Yes 回数	44	41	43	44
真の Yes 回数	36	32	34	33
正解率 (%)	81.81	78.04	79.06	75.0

表 1: 予測正解率に関する実験結果

また、二重マッピングの場合の再学習にかかったエポック回数が一回目の学習に比べて平均的に倍もかかっており、この状況からも最終入力属性値ベクトルが悪化したことが分かる。

5.3 騰落強度の識別能力に関する実験

これまで、「上がるかどうか」(「下がるかどうか」)に対して「上がる」(「下がる」)と予測された場合、その「上がる」(「下がる」)程度がどれぐらいかを定量的に測るために「騰落強度」(式 (2) を参照)を使用し、騰落強度が大きければ大きいほど、それに伴うターゲットウィンドウの騰落率も大きくなる傾向があり、それが売買タイミングの見極めに有効であることを示した [9]。従って、予測騰落強度が充分大きくなった場合、その時点で売買を決断するわけである。

本研究では、一歩進んで、騰落強度の大きさに伴って騰落率も大きくなる相関性を更に強くさせるようなマッパーを導入した。そして二重マッピングを行った場合、予測騰落強度と実際の騰落率の相関性がそうでない場合に比べて強くなったのかどうかについてやはり「上がるかどうか」と「下がるかどうか」別、及び一次マッピングと二重マッピング別に実験を行った。その結果を表 2 に示す。ここでもターゲットウィンドウのサイズや予測期間等の条件は第 5.2 節と同様である。

マッピング	「上がる?」		「下がる?」	
	一次	二重	一次	二重
相関度	0.68	0.79	0.61	0.73

表 2: 騰落強度の識別能力に関する実験結果

表 2 から、二重マッピングの場合の予測騰落強度と実際の騰落率の相関度が「上がるかどうか」と「下がるかどうか」の両方とも一次マッピングの場合に比べてかなり改善されたことが分かる。

6 結論

二重マッピングと SCW [8] による機械学習を組み合わせることで、特に売買タイミングを正確に知るための予測騰落強度の精度向上を実現する手法を提案し、その有効性を示した。また、「上がるかどうか」を主眼にした学習と「下がるどうか」を主眼にした学習に分けて学習を行った。これには二つ目的があった。

一つは「上がるかどうか」(「下がるどうか」)のラベル付け基準値の近辺での予測が僅かな「上がり」(「下がり」)であっても、できるだけ実用レベルの「上がり」(「下がり」)を得るためである。学習期間内のターゲッ

²株価時系列データに「調整後終値」もあることで大掛かりな加工を施さなくても使用でき、一貫性も保たれるため、大変助かる。この場を借りて Yahoo!ファイナンスに感謝の意を表したい。

トウィンドウの学習ポイントから見た全騰落率のソート後の $\frac{1}{3}$ 範囲内が「上がり」(「下がり」)になるようにラベル付け基準値を決め、その範囲外にも敢えて一部の「上がり」(「下がり」)を残すことでこの目的を実現した。2クラス分類問題では、教師データの正例と負例のバランスが取れていない場合、予測結果が数の多いほうに傾きやすいことが一般的に知られているが、SCW の場合はそうならない [8] になっているため、本研究でも問題にならない。今回の実験結果からもその点が確認できる。

もう一つは、株式市場が「上昇気流」に乗っている場合は緩やかに上昇するが、「調整気流」になった場合は短期間で降下する傾向にあり、騰落率が 0 以上になるターゲットウィンドウの数が多いため、「上がり」(「下がり」)のラベル付け基準値を 0 にすると、「下がり」の予測がしにくくなり、投資機会が明らかに少なくなってしまう。従って、本研究のアプローチが効果的である。

更に、株式投資は、毎日その変動を予測して毎日売買を行わなくてはならないことではないため、「上がる」、「下がる」可能性が充分高くなった時点で売買を行えばよい。確信を持った時だけ売買を行い、一定以上の利益を確保する、あるいは損だけは出さないことは投資家にとって重要である。本研究では、その観点から、これまでの研究 [9] で導入した予測騰落強度の認識力の向上をはかった手法を提案し、その有効性を示した。

今回は日経平均株価指数に関して評価実験を行ったが、他に検討したい指数としてダウ平均株価指数とナスダック総合指数がある。ダウ平均株価指数は構成銘柄数が少なく、30 銘柄数であり、本研究のアプローチの有効性が高くない可能性がある。また、ナスダック総合は加重平均指数であり、構成銘柄数が逆に大過ぎて高い正解率が期待できない可能性がある。日経平均株価指数のベース銘柄数が多すぎても少なすぎても正解率が悪くなることを関連実験で確認しており、ベース銘柄数を 320 にしたのもそのためである。従って、今後の課題として、上記の 2 つの指数に関しても同様の実験を実施し、今回提案した手法の有効性を調べたい。

謝辞

研究を進めるにあたり、その方法・考え方等多方面に渡ってご指導を頂いた恩師であり、名古屋大学名誉教授の稲垣康善先生に厚く御礼を申し上げます。

参考文献

- [1] Hinton, G. , Osindero, S. and Teh, Y. W. : A fast learning algorithm for deep belief nets, *Neural Computation*, Vol. 18, No. 7, pp. 1527–1554 (2006)
- [2] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H.: Greedy Layer-Wise Training of Deep Networks, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, accessed February 1, 2017, <http://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf> (2007)
- [3] Chao, J., Shen, F. and Zhao, J.: Forecasting exchange rate with deep belief networks, *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1259–1266(2011)
- [4] Yeh, S. , Wang, C. and Tsai, M.: Corporate Default Prediction via Deep Learning, In *The 34th International Symposium on Forecasting (ISF '14)*, (2014)
- [5] Boser, B. E., Guyon, I. M., and Vapnik, V. N.: A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory (COLT '92)*, pp. 144–152 (1992)
- [6] 淵井 亮, 鍾 寧: サポートベクターマシンを利用した株価の値動き予測, 電子情報通信学会技術研究報告, *KBSE2010-42*, pp. 43–48 (2011)
- [7] Shen, S., Jiang, H., and Zhang, T.: Stock market forecasting using machine learning algorithms, *CS229 (Machine Learning) at Stanford University*, accessed February 1, 2017, <http://cs229.stanford.edu/proj2012/ShenJiangZhang-StockMarketForecastingusingMachineLearningAlgorithms.pdf> (2012)
- [8] Wang, J., Zhao, P., and Hoi, S. C. H.: Exact Soft Confidence-Weighted Learning, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 121–128 (2012)
- [9] 福田ムフタル: SCW を用いた株価変動予測, 人工知能学会研究会資料, SIG-FIN-016-02, (2016)
- [10] 日経平均プロフィール, <http://indexes.nikkei.co.jp/nkave/>
- [11] 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, 人工知能学会研究会資料, SIG-FIN-012-08, (2014)
- [12] Yahoo!ファイナンス, <http://finance.yahoo.co.jp/>

SVM の Limit order Book への応用と精度向上

Improvement of Support Vector Machine Applied to Limit Order Book

木島隼人¹ 高田英行²

Hayato Kijima¹ and Hideyuki Takada²

¹ 東邦大学大学院理学研究科 情報科学専攻 ² 東邦大学理学部情報科学科

¹Graduate School of Information Science, Toho University

²Department of Information Science, Toho University

Abstract: Market participants place their limit/market orders by taking into account both the trajectory and current status of the limit order book. This behavior is based on the policy that the shape of the limit order book is quite informative for predicting future direction of a traded asset. In this paper, we employ Support Vector Machine combined with conformally transformed Gaussian RBF kernel to forecast the mid price dynamics. Our empirical studies show that the conformal transform methods improved the precision more than 3% in average compared to the standard Gaussian RBF kernel.

1 はじめに

高頻度データの蓄積に伴い、リミット・オーダー・ブックのダイナミクスに関心が集まっており、その確率的挙動のモデル化、マーケットインパクト関数の推定、最適執行戦略など比較的新しいテーマが議論されるようになった。

機械学習を応用する先行研究には、例えば Patel, Shah, Thakkar and Kotecha [5] がニューラルネットやランダム・フォレストなどの手法を時系列データに適用し、ミッドプライスの予測を試みている。Zheng, Moulines and Abergel [9] はリミット・オーダー・ブックに対してロジスティック回帰を用いてミッド・プライスの予測を提案している。Fletcher and Shawe-Taylor [3] や Kercheval and Zhang [4] 等ではリミット・オーダー・ブックの状態とその単位時間の変化量など様々な指標を学習データとする3クラス分類サポート・ベクター・マシンを提案している。

本研究では、[3][4]と同様に、サポート・ベクター・マシンを利用し、リミット・オーダー・ブックの状態からミッドプライスの超短期の予測を試みる。とくに、リーマン幾何学の観点から Amari and Wu [1] によって提案されたカーネル関数の共形変換により、標準的なガウスカーネルと比べて平均的に3%以上の精度向上が見込まれる点について報告する。

2 データ

我々は、大阪取引所に上場・取引されている日経

225先物のヒストリカルデータ（2012年4月~6月）のうち、レギュラー・セッション（午前9時~午後3時10分）を対象に実証分析を行った。オリジナルのデータレコードは約20マイクロ秒毎であるが、分析の効率化を優先し本研究では1秒毎のデータに加工して用いた。日経225先物の取引には、オークション方式が採用されており、リミット・オーダー・ブックの上で主に指値注文（リミット・オーダー）と成行注文（マーケット・オーダー）とそれらのキャンセルによって取引が展開されている。ここでは、呼び値の刻み幅を δ (=10円)と表記する。時刻 t において、そのときの最良買値から $k \cdot \delta$ 円だけ離れた価格に待機中の指値売注文の量を

$$Q_{sell}^{k\delta}(t) \in Z_+$$

と定め、最良売値から $k \cdot \delta$ 円だけ離れた価格に待機中の指値買注文の量を

$$Q_{buy}^{k\delta}(t) \in Z_+$$

と定める。時刻 T_i ($0 \leq i \leq M$)においてリミット・オーダー・ブックに待機している指値売買注文の状態を表す $2n$ 次元ベクトルを次のように表す。

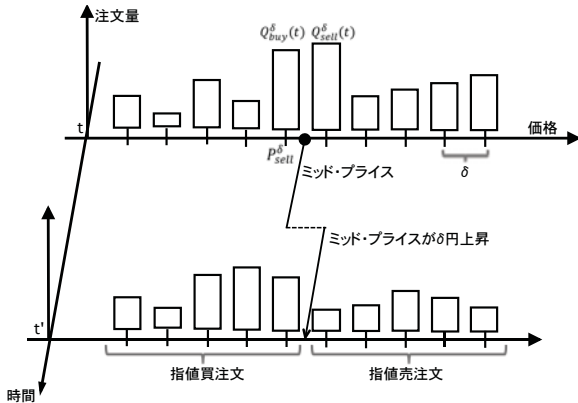
$$\mathbf{x}_i = \left(Q_{buy}^{n\delta}(T_i), Q_{buy}^{(n-1)\delta}(T_i), \dots, Q_{buy}^{\delta}(T_i), \right.$$

$$\left. Q_{sell}^{\delta}(T_i), \dots, Q_{sell}^{(n-1)\delta}(T_i), Q_{sell}^{n\delta}(T_i) \right).$$

最良買値 $P_{buy}^{\delta}(T_i)$ と最良売値 $P_{sell}^{\delta}(T_i)$ の中間値

$$\pi(T_i) := (P_{sell}^{\delta}(T_i) + P_{buy}^{\delta}(T_i))/2$$

をミッド・プライスと呼ぶ（下図参照）。



時刻 T_k 以降で初めてミッド・プライスが動いた時刻を表すインデックスを $I(T_k) \in \mathbb{Z}_+$ とする。

$$I(T_k) = \min\{j | \pi(T_{j+1}) \neq \pi(T_k), j \geq k\}.$$

3 サポート・ベクター・マシン

本研究で利用するサポート・ベクター・マシンの学習データはベクトルの列 $\{\mathbf{x}_i \in \mathbb{R}^{2n}\}_i$ であり、そのクラスを次のように付与する（教師あり学習）。

$$y_i = \begin{cases} +1 & \pi(T_{I(T_k)}) > \pi(T_{I(T_k)-1}) \\ -1 & \pi(T_{I(T_k)}) < \pi(T_{I(T_k)-1}) \end{cases}$$

すなわち、時刻 T_i 以降で初めてミッド・プライスが動いたとき、上昇していた場合は $y_i = +1$ 、下降していた場合は $y_i = -1$ と付与する。

一般に、カーネル関数 K を利用したサポート・ベクター・マシンは、入力空間 $I = \mathbb{R}^{2n}$ から特徴空間 $F = \mathbb{R}^N$ ($2n \leq N$)への写像 $\Phi: I \rightarrow F$ をうまく選ぶことによって分類能力を向上させる。学習を終えたサポート・ベクター・マシンは分離境界

$$f(\mathbf{x}) = \sum_{i \in SV} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$$

を得ており (SV はサポート・ベクター全体の集合)、新たに観測した未分類のデータ \mathbf{x} に対して

$$\mathbf{y} = h(\mathbf{x}) := \text{sign}(f(\mathbf{x})) \in \{+1, -1\}$$

により予測値を返す。

4 カーネル関数の幾何学的解釈

一般に、写像 Φ によって写されたデータ $\mathbf{z} = \Phi(\mathbf{x})$ は、特徴空間 F 内の $2n$ 次元超曲面 S 上の点とみなすことができる。 Φ が全ての方向に連続微分可能であれば、 S を $F = \mathbb{R}^N$ に埋め込まれた部分多様体と見るこ

とができ、 F から自然に誘導されるリーマン距離を持つ。リーマン距離によって、 S 上の異なる2点間の距離を測地線に沿った線積分として表すことができるが、その線素 ds は

$$(ds)^2 = \sum_{1 \leq i, j \leq 2n} g_{ij} dx^i dx^j$$

と表される。ただし g_{ij} はリーマン距離である。

$2n$ 次元の無限小ベクトル $d\mathbf{x} = (dx^1, \dots, dx^{2n})$ は、写像 Φ によって $d\mathbf{z} = \Phi(\mathbf{x} + d\mathbf{x}) - \Phi(\mathbf{x})$ に写されるので、線素は F 内で次のように表される。

$$(ds)^2 = \|d\mathbf{z}\|^2 = \|\Phi(\mathbf{x} + d\mathbf{x}) - \Phi(\mathbf{x})\|^2$$

テイラーの定理により、線素はカーネル関数 K を用いて次のように表すことができる。

$$\begin{aligned} (ds)^2 &= \sum_{k=1}^N \left\{ \sum_{j=1}^n \frac{\partial \Phi_k(\mathbf{x})}{\partial x^j} \right\}^2 \\ &= \sum_{k=1}^N \sum_{1 \leq i, j \leq 2n} \frac{\partial \Phi_k(\mathbf{x})}{\partial x^i} \frac{\partial \Phi_k(\mathbf{x})}{\partial x^j} dx^i dx^j \\ &= \sum_{1 \leq i, j \leq 2n} \left(\frac{\partial^2 K(\mathbf{x}, \mathbf{x}')}{\partial x^i \partial x'^j} \right)_{\mathbf{x}=\mathbf{x}'} dx^i dx^j \end{aligned}$$

従って、リーマン距離はカーネル関数を使って

$$g_{ij}(\mathbf{x}) = \left(\frac{\partial^2 K(\mathbf{x}, \mathbf{x}')}{\partial x^i \partial x'^j} \right)_{\mathbf{x}=\mathbf{x}'}$$

と書けることがわかる。直感的には、 $I = \mathbb{R}^{2n}$ 内の点 \mathbf{x} のまわりの局所領域は写像 $\Phi(\mathbf{x})$ によって $g_{ij}(\mathbf{x})$ 倍に伸縮されることになる。(議論の詳細は参考文献[1][6][7][8]を参照頂きたい)

5 カーネル関数の共形変換

サポート・ベクター・マシンの精度を向上させるため、Amari and Wu [1] と Wu and Amari [8] はカーネル関数を共形変換することを提案した。共形変換とは、二つのベクトルの成す角を保存するかわり、長さを変える変換のことである。Amari and Wu [1] が最初に意図した考えは、カーネル関数を共形変換することで分離境界の近くで倍率 $g_{ij}(\mathbf{x})$ を大きくし、かわりに分離境界から離れたところで倍率 $g_{ij}(\mathbf{x})$ を小さくし、全体を保ちつつ判別力を上げるというものであった。こうすることで、分離境界の近くで誤判別を減らせると期待できるからである。

具体的に、カーネルの共形変換は次で定義される。

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = D(\mathbf{x})D(\mathbf{x}')K(\mathbf{x}, \mathbf{x}')$$

ただし D は I 上の正値関数で、これを適切に設定する

ことがサポート・ベクター・マシンの精度向上のために重要である。共形変換後のリーマン距離 \tilde{g}_{ij} は

$$\tilde{g}_{ij}(\mathbf{x}) = D(\mathbf{x})^2 g_{ij}(\mathbf{x}) + D'_i(\mathbf{x}) D'_j(\mathbf{x}') K(\mathbf{x}, \mathbf{x}') + 2D'_i(\mathbf{x}) D(\mathbf{x}) K'_i(\mathbf{x}, \mathbf{x}')$$

ただし、

$$K'_i(\mathbf{x}, \mathbf{x}') = \left. \frac{\partial K(\mathbf{x}, \mathbf{x}')}{\partial x_i} \right|_{\mathbf{x}=\mathbf{x}'}; \quad D'_i(\mathbf{x}) = \frac{\partial D(\mathbf{x})}{\partial x_i}$$

である。我々は、多くの研究が選択しているようにガウシアン RBF カーネルとその共形変換を考える。

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right).$$

更に我々は、Williams, Li, Feng and Wu [6] に従い、共形変換を定義する関数 D として

$$D(\mathbf{x}) = e^{-\kappa f(\mathbf{x})^2}$$

の形に表されるものを採用する。ここで $f(\mathbf{x})$ は既に求めてある分離境界、 κ は正の定数である。Williams, Li, Feng and Wu [6] が示しているように、このように定めた共形変換によって、分離境界の近くで $\tilde{g}_{ij}(\mathbf{x})$ を大きくし、分離境界から離れたところで小さくできるのは次のような理由である。

まず、 $g_{ij}(\mathbf{x})$ を並べた行列を $g(\mathbf{x})$ と書くとき、このリーマン距離に関する（無限小）体積要素は

$$\sqrt{\det(g(\mathbf{x}))} dx^1 \dots dx^{2n}$$

であることに注意する。従って、共形変換によって行列 $g(\mathbf{x})$ の行列式がどのように変化するかを見ればよい。ガウシアン RBF カーネルの場合は

$$g_{ij}(\mathbf{x}) = \frac{1}{\sigma^2} \delta_{ij}$$

であり、

$$\sqrt{\det(g(\mathbf{x}))} = \frac{1}{\sigma^n}$$

となる。一方、共形変換されたリーマン距離は

$$\tilde{g}_{ij}(\mathbf{x}) = \frac{D(\mathbf{x})^2}{\sigma^2} \delta_{ij} + D'_i(\mathbf{x}) D'_j(\mathbf{x}')$$

であるから、

$$\sqrt{\det(\tilde{g}(\mathbf{x}))} = \frac{D(\mathbf{x})^{2n}}{\sigma^{2n}} + \frac{D(\mathbf{x})^{2n-2}}{\sigma^{2n-2}} \sum_{i=1}^{2n} D'_i(\mathbf{x})^2$$

と書ける。定義よりベクトル $\nabla D(\mathbf{x}) = D(\mathbf{x}) \nabla \log D(\mathbf{x})$ の各成分を $D'_i(\mathbf{x})$ と書いていたので、伸縮の比率は

$$\frac{\sqrt{\det(\tilde{g}(\mathbf{x}))}}{\sqrt{\det(g(\mathbf{x}))}} = D(\mathbf{x})^{2n} \sqrt{1 + \sigma^2 \|\nabla \log D(\mathbf{x})\|^2}$$

と書き直され、 $D(\mathbf{x})$ の具体的な形から更に

$$\sqrt{\frac{\det(\tilde{g}(\mathbf{x}))}{\det(g(\mathbf{x}))}}$$

$$= \exp(-n\kappa f(\mathbf{x})^2) \sqrt{1 + 4\kappa^2 \sigma^2 f(\mathbf{x})^2 \|\nabla f(\mathbf{x})\|^2}$$

と表すことができる。方程式 $f(\mathbf{x}) = 0$ が分離境界を表すが、右辺の支配項である $\exp(-n\kappa f(\mathbf{x})^2)$ が分離境界上で最大値をとり、離れるにつれ急激に減少していくことから、当初意図した通りの性質を備えていることが確認できる (Williams, Li, Feng and Wu [6] を参照)。

6 数値計算アルゴリズム

共形変換後のカーネルを用いたサポート・ベクター・マシンの学習のためには、分離境界 $f(\mathbf{x}) = 0$ を事前に求めておく必要がある（これをプライマリー・カーネルと呼ぶ）。従って、アルゴリズムは合計2回の学習ステップを経ることになる。

高精度の予測を行うためには学習データとして十分な量を確保する必要があるが、巨大な最適化問題を数値的に解くためには多大な計算時間も必要となる。我々は、学習データとして予測時刻から遡ること過去1時間のリミット・オーダー・ブックの履歴を用いることにした。即ち、レギュラーセッションのオープニング時刻を $T_0 = 09:00:00$ 、その1時間後を $T_k = 10:00:00$ とし、区間 $[T_m, T_{m+k}]$ に記録された履歴を学習データとして使い、 $T_M = 15:10:00$ （クロージング）まで $m = 0, 1, 2, \dots$ と動かしながら予測を行う¹。ただし、実際にはミッド・プライスに変化していないと予測時点において既知の情報としてクラスを付与できないため、過去1時間の履歴が全て学習データとして使える訳ではないことに注意を要する。数値計算アルゴリズムは次のようになる。我々のデータセットでは $T_{m+1} - T_m = 1$ 秒である。

アルゴリズム

- Step 0. パラメータ C, σ, κ をセットし、学習データを区間 $[T_0, T_k]$ に得られるデータとする。
 $m = 0$ とする。
- Step 1. プライマリー・カーネル K を用いたサポート・ベクター・マシンにより、分離境界 $f(\mathbf{x})$ を求める。
- Step 2. 共形変換したカーネル \tilde{K} を用いたサポート・ベクター・マシンにより、分離境界 $\tilde{f}(\mathbf{x})$ を求める。

¹ ここで言う予測とは、「次に初めてミッド・プライスが動いたとき、上昇か下降かを予測する」ことである。

Step 3. 新しく観測されたデータ x_{m+k+1} に対して、
 $y_{m+k+1} = \tilde{h}(x_{m+k+1}) = \text{sign}(\tilde{f}(x_{m+k+1}))$ を
予測値として返す。

Step 4. $m \leftarrow m + 1$ とし時間を 1 秒進め、学習デ
ータを区間 $[T_m, T_{m+k}]$ に得られるデータと
し、 $T_{m+k} < 15:10:00$ まで Step 1 に戻る。

このアルゴリズムが実際に使えるためには、Step1
から Step 4 までの一連の計算時間が 1 秒より短く
なくてはならないが、4GHz Intel Core i7, 32GB 1867
MHz DDR3 の環境で Matlab を用いた計算では約 0.17
秒と十分な早さを保っている。

7 結果

共形変換したカーネル関数を用いたサポート・ベ
クター・マシンのパフォーマンスを評価するため、
以下で定義される指標を用いる。はじめてミッドプ
ライスが変化するとき、上昇すると予想して実際に
上昇した件数を $\sigma_{1,1}$ とする (true positive)。また、上
昇すると予想したにもかかわらず実際には下降した
件数を $\sigma_{1,-1}$ とする (false positive)。 $\sigma_{-1,1}$ と $\sigma_{-1,-1}$ も
同様に定める。このとき、指標 $\text{Precision}_{(+1)}$ は、+1
(ミッドプライスが上昇) と予想したうち、実際に
+1 であった (上昇した) 割合を表し、 $\text{Precision}_{(-1)}$ は、
-1 (下降) と予想したうち実際に -1 であった (下降
した) 割合を表す。

$$\text{Precision}_{(+1)} = \frac{\sigma_{1,1}}{\sigma_{1,1} + \sigma_{1,-1}},$$

$$\text{Precision}_{(-1)} = \frac{\sigma_{-1,-1}}{\sigma_{-1,1} + \sigma_{-1,-1}}$$

比較対象として (共形変換を施さない) ガウシアン
RBF カーネルを考える。共形変換のパラメータ
 $\kappa = 0$ がガウシアン RBF カーネルに対応している
ことを思い出しておく。 κ を大きくすればそれだけガ
ウシアン RBF カーネルからずれることになる。こ
こでは $\kappa = 1.5$ とおいたが、より適した値を探し出す
方法は今後の課題となる。

また、我々は共形変換の効果に関心があるので、
 C, σ のチューニングは行わないこととする。そこで、
 C, σ の値として $(C, \sigma) \in \{(1,1), (10,1), (100,1)\}$ の組
み合わせで数値実験を行ったあと、これら 3 つの場
合の指標 Precision の平均で評価を行う。

表 1 から表 3 は $\text{Precision}_{(+1)}$ の値を日時でプロッ
トしたものである。折れ線は (共形変換を施さない)
ガウシアン RBF カーネルを用いたサポート・ベク
ター・マシンによる結果を表し、実線は共形変換され
たガウシアン RBF カーネルを用いたサポート・ベク
ター・マシンによる結果を表す。

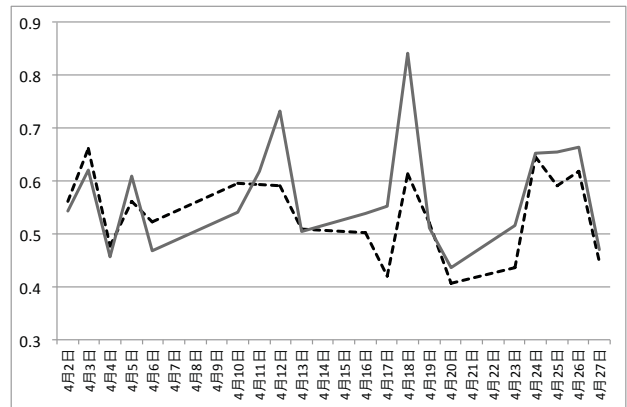


表 1 Precision₍₊₁₎ (2012年4月)

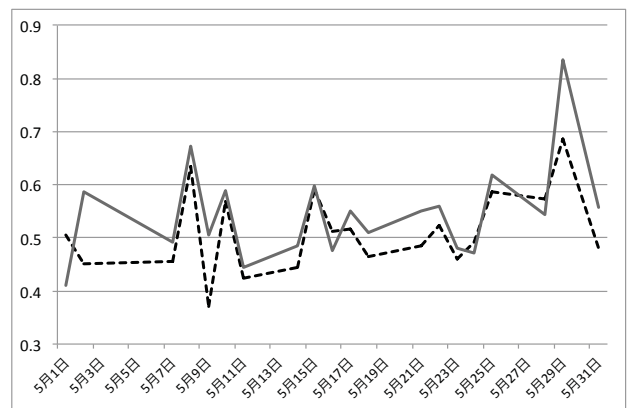


表 2 Precision₍₊₁₎ (2012年5月)

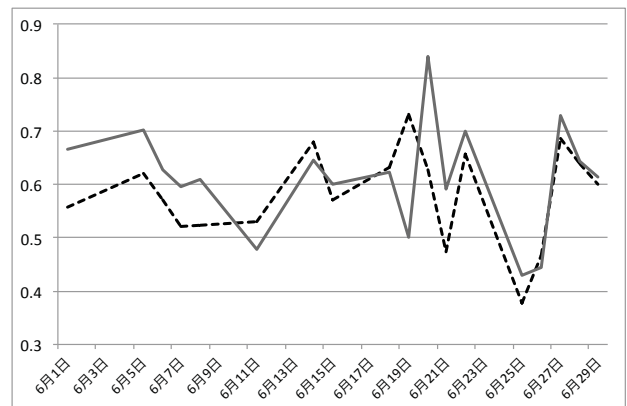


表 3 Precision₍₊₁₎ (2012年6月)

日時で両者を比較すると、常に共形変換を組み合わ
せた方法の精度が優れているわけではないが、2012
年 4 月～6 月の $\text{Precision}_{(+1)}$ を平均すると、

ガウシアン RBF カーネル … 54.340%
共形変換後 RBF カーネル … 57.757%

となり、カーネルの共形変換により Precision が 3% 以上改善されたことがわかる。

以下の表 4 から表 6 は Precision₍₋₁₎ の値を日時でプロットしたものである。

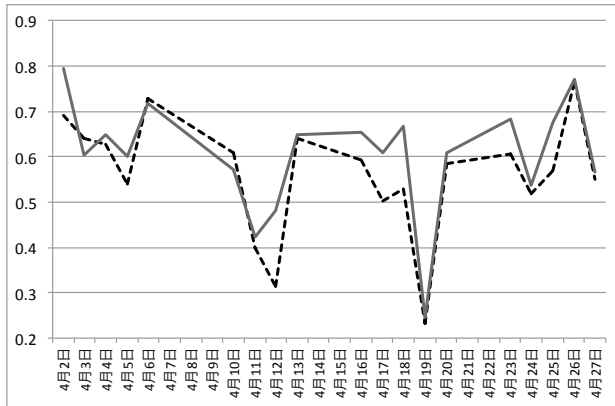


表 4 Precision₍₋₁₎ (2012 年 4 月)

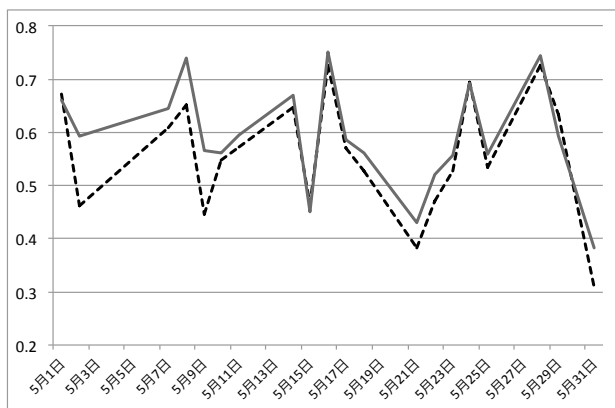


表 5 Precision₍₋₁₎ (2012 年 5 月)

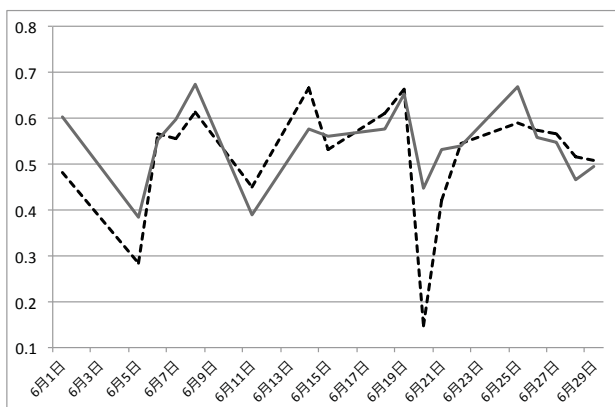


表 6 Precision₍₋₁₎ (2012 年 6 月)

2012 年 4 月～6 月の Precision₍₋₁₎ を平均すると、

ガウシアン RBF カーネル … 54.566%
共形変換後 RBF カーネル … 58.220%

となり、こちらもカーネルの共形変換により Precision が 3%以上改善されたことがわかる。

今後は、共形変換を決定するパラメータ κ の調整により、より精度を向上させることができないか調査することになる。

参考文献

- [1] Shun-ichi Amari and Si Wu: *Improving support vector machine classifiers by modifying kernel functions*, Neural Networks, Vol. 12, 783-789, (1999)
- [2] Shangkun Deng, Akito Sakurai and Kei Shioda: *Prediction of Foreign Exchange Market States with Support Vector Machine*, 10th International Conference on Machine Learning and Applications and Workshops, Vol. 1, 327-332, (2011)
- [3] Tristan Fletcher and John Shawe-Taylor: *Multiple Kernel Learning with Fisher Kernels for High Frequency Currency Prediction*, Computational Economics, Vol. 42, Issue 2, 217-240, (2013)
- [4] Alec N. Kercheval and Yuan Zhang: *Modelling high-frequency limit order book dynamics with support vector machines*, Quantitative Finance, Vol. 15, Issue 8, 1315-1329, (2015)
- [5] Jigar Patel, Sahil Shah, Priyank Thakkar and K Kotecha: *Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques*, Expert Systems with Applications, Vol 42, 259-268, (2015)
- [6] Peter Williams, Sheng Li, Jianfeng Feng and Si Wu: *Scaling the kernel function to improve performance of the Support Vector Machine*, Advances in Neural Networks, Volume 3496 of the series Lecture Notes in Computer Science, 831-836, (2005)
- [7] Gang Wu and Edward Y. Chang: *Adaptive feature-space conformal transformation for imbalanced-data learning*, Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), Washington DC.
- [8] Si Wu and Shun-ichi Amari: *Conformal transformation of kernel functions: A data-dependent way to improve the performance of support vector machine classifiers*, Neural Processing Letter, Vol. 15, 59-67, (2002)
- [9] Ban Zheng, Eric Moulines and Frederic Abergel: *Price Jump Prediction in a Limit Order Book*, Journal of Mathematical Finance, Vol. 3, 242-255, (2013)

マルチタスク最大マージントピックモデルによる収益予測

Revenue Prediction based on Multi-task Max-margin Topic Models

中川 雄太^{1*} 上野 良輔¹ 江口 浩二^{1†}¹ 神戸大学大学院システム情報学研究科

Abstract: Researchers and practitioners in the economic and financial field recently have a keen interest in discovering new ideas by making full use of large-scale data, such as in the form of document data of company valuation in online news and the form of numerical data of company financial indices. One promising approach to analyzing such large-scale data is topic modeling, typically by Maximum Entropy Discrimination LDA (MedLDA). MedLDA is a supervised topic model that can improve accuracy of latent topic estimation by making use of the side information associated with each document. In this paper, we generalize Multi-task MedLDA (MultiMedLDA) that simultaneously addresses classification and regression tasks in an extension of MedLDA. In this paper, we evaluate the effectiveness of MultiMedLDA through experiments with enterprise evaluation documents associated with continuous labels of change rate of operating incomes and discrete labels of categories of business, and discuss it compared with single-task MedLDA.

1 はじめに

近年、情報技術の発達によって、情報の発信形態は多様化しており、世の中に存在するテキストデータの量は増加し続けている。これに伴って大規模データを解析する重要性が増しており、代表的なモデルの1つとして潜在的ディリクレ配分法 (Latent Dirichlet Allocation: LDA)[1] を代表とするトピックモデルが挙げられる。トピックモデルでは、文書に存在するテキストデータを bag-of-words 形式で離散表現している。bag-of-words とは文書中に現れる単語の語順を無視して単語の出現頻度に着目した表現形式のことである。また、文書中には複数の潜在的な潜在トピックが存在していると仮定しており、この潜在トピックは文書集合中に出現する語彙からなる分布で表現される。更に、LDA では文書ごとに潜在トピック分布が存在していると仮定している。これら2つの分布によって文書集合に含まれる単語を表現している。

また、企業や金融機関を中心に経済・金融分野の情報解析への関心が高まっている。経済データは社会のあらゆる事象を反映し、刻一刻と変化する情勢の中で生み出されており、このデータから新たな知見を得る取り組みがなされ始めている。経済・金融分野での文書データに存在する特徴として、企業データや数値を

含むデータのような多種多様な付加情報が存在することが挙げられる。この付加情報を考慮した、より有望な解析手法が望まれている。

このように、実世界に存在する文書データは付加情報として離散値 (たとえば、ジャンルや著者の性別) や連続値 (たとえば、著者の年齢や評価点) を持つことが一般的である。付加情報を学習に活かすことで潜在トピック推定の精度を向上させているモデルとして、教師有りトピックモデル (Supervised topic models)[2] や最大マージントピックモデル (Maximum Entropy Discrimination LDA: MedLDA) [3] が存在する。MedLDA には、離散値を伴う文書データに対応した MedLDA regression と、連続値を伴う文書データに対応した MedLDA classification が存在する。文書の付加情報を潜在トピックの推定に活かすことによって、予測精度の向上を図っている。また、サポートベクターマシン (Support Vector Machine: SVM) [4, 5] におけるマージン最大化法を取り入れ、推定した潜在トピックを特徴量として利用している。ただし、MedLDA は単一の付加情報を持つ文書データを対象としており、離散値ラベルと連続値ラベルの両方を持つデータに対して適用できない。

本論文では、金融・経済テキストデータから企業の収益性に関する指標を予測するモデルを実現することを目的とする。この目的のもと、MedLDA を拡張して分類タスクと回帰タスクを同時に解決するマルチタスク最大マージントピックモデル (MultiMedLDA) に一般化する。MultiMedLDA では複数種類のラベルが付与された文書データを対象としており、複数種類の付加情

*連絡先: 神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: nakagawa@cs25.scitec.kobe-u.ac.jp

†連絡先: 神戸大学大学院システム情報学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: eguchi@port.kobe-u.ac.jp

報を同時に考慮しながら潜在トピックの推定を可能にしている。また、複雑化した最適化問題を解く為に双対分解 [7] と呼ばれる手法を導入している。双対分解は効率的に解くことができない目的関数に対するアプローチの1つであり、目的関数がいくつかの関数に分解でき、それぞれの関数の最適解が効率的に解くことができる場合に適用することが可能である。MultiMedLDAではより多くの付加情報から潜在トピックを推定しており、予測精度の改善が期待される。本論文では、業種の離散ラベル、営業利益変化率の連続ラベルを伴う企業評価テキストを用いて MultiMedLDA の有効性を評価し、MedLDA の分類タスクおよび回帰タスクと比較して議論する。

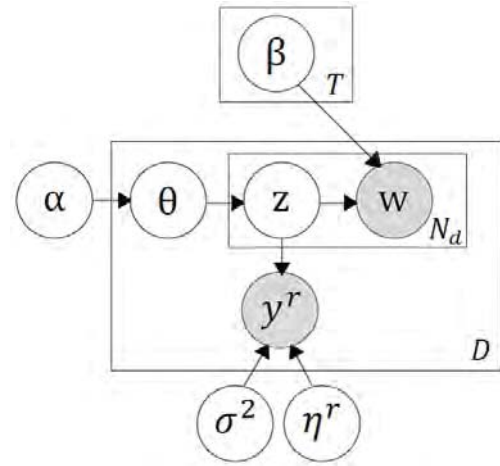


図 1: MedLDA のグラフィカルモデル

2 関連研究

提案手法の基礎研究として、教師ラベルを考慮して潜在トピックの推定を行う MedLDA, および双対分解について説明する

2.1 最大マージントピックモデル

Maximum Entropy Discriminated LDA (MedLDA) [3] は、最大エントロピー識別 [6] と呼ばれる教師付き学習の枠組みにおいて、潜在トピックを特徴として用いつつ、その潜在トピックを推定する手法である。MedLDA のグラフィカルモデルを図 1 に示す。図 1 中の y は各文書に付与された教師ラベルである。また、 η はラベル評価時の各トピックに対する重み係数であり、トピックに対する関係度を表す。MedLDA の生成過程を以下に示す。

1. 文書 $d (d \in 1, \dots, D)$ に対して、 $\theta \sim \text{Dirichlet}(\alpha)$ を選択。
2. 文書 d 内の N_d 個の単語 $w_{d,n} (n \in 1, \dots, N_d)$ に対して、
 - (a) トピック $z_{d,n} \sim \text{Multinomial}(\theta_d)$ を選択。
 - (b) 単語 $w_{d,n} \sim \text{Multinomial}(\beta_t)$ を選択。
3. D 個の文書に対してラベル $y_d \sim F(\eta, z_d) (y_d \in 1 \dots R)$ を選択する。

なお、教師ラベル y が連続量であるときの MedLDA は回帰モデル、離散量であるときは分類モデルに位置づけられる。それぞれについて 2.2.1 節と 2.2.2 節にて後述する。

2.1.1 回帰問題を想定した最大マージントピックモデル

連続ラベルを持つ文書データを扱う MedLDA Regression (MedLDA-Reg) について説明する。MedLDA-Reg ではマージン最大化を考慮することにより、以下のような最適化問題が定義される。

P1 (MedLDA-Reg) :

$$\min_{q(\mathbf{Z}, \Theta, \eta), \alpha, \beta, \delta^2, \xi, \xi^*} \mathcal{L}(q(\mathbf{Z}, \Theta, \eta)) + C \sum_{d=1}^D (\xi_d + \xi_d^*)$$

$$\text{s.t. } \forall d : \begin{cases} y_d - E[\eta^\top \bar{z}_d] \leq \epsilon + \xi_d & [\mu_d] \\ -y_d + E[\eta^\top \bar{z}_d] \leq \epsilon + \xi_d^* & [\mu_d^*] \\ \xi_d \geq 0 & [v_d] \\ \xi_d^* \geq 0 & [v_d^*] \end{cases}$$

$\mu_d, \mu_d^*, v_d, v_d^*$ はラグランジュ乗数である。上式中の $[\]$ はラグランジュ関数を求める際の制約式とラグランジュ乗数の対応を表している。また、各変数は $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_D\}, \mathbf{z}_d = \{z_{d,1}, \dots, z_{d,N_d}\}, \Theta = \{\theta_1, \dots, \theta_D\}, \mathbf{Y} = \{y_1, \dots, y_D\}, \mathbf{W} = \{w_1, \dots, w_D\}, \mathbf{w}_d = \{w_{d,1}, \dots, w_{d,N_d}\}, \beta = \{\beta_1, \dots, \beta_K\}$ を表す。制約式中の ξ, ξ^* は訓練データの誤差を吸収する程度を示すスラック変数であり、 ϵ は許容誤差である。ここからは MedLDA-Reg の更新式の導出を行う。目的関数の \mathcal{L} は

$$\mathcal{L}(q(\mathbf{Z}, \Theta, \eta)) = -E[\log p(\theta, \mathbf{z}, \eta, \mathbf{Y}, \mathbf{W} | \alpha, \beta, \delta^2)] - \mathcal{H}(q(z), \theta, \eta) \quad (1)$$

である。 \mathcal{H} は変分事後分布 $q(\mathbf{Z}, \theta, \eta)$ のエントロピーであり、 $\mathcal{H}(q) = -\sum q \log(q)$ である。最適化問題 P1 は一般的に解くことが困難であるため、変分近似を行い

q についての独立性を与える.

$$q(Z, \Theta, \eta | \gamma, \phi) = q(\eta) \prod_{d=1}^D q(\theta_d | \gamma_d) \prod_{n=1}^N q(z_{d,n} | \phi_{d,n}) \quad (2)$$

ここで, γ_d および $\phi_{d,n}$ は変分パラメータであり, γ_d はディリクレ分布パラメータの K 次元のベクトル, $\phi_{d,n}$ は K トピックの多項分布である. そして変分 EM アルゴリズムを行い, 各パラメータを最適化する. 変分 EM アルゴリズムでは次の 2 ステップを繰り返し行う.

1. **E-Step**: 潜在変数の事後分布を推定

2. **M-Step**: 未知変数を推定

更新式の導出では変分下限を最大化する各パラメータを求める. また, 最適化問題 P1 の制約式を目的関数に組み込み, ラグランジュ関数 L^r を定義する.

$$\begin{aligned} L^r = & \mathcal{L}(q) + C \sum_{d=1}^D (\xi_d, \xi_d^*) - \sum_{d=1}^D \mu_d (\epsilon + \xi_d - y_d + E[\eta^\top \bar{Z}_d]) \\ & - \sum_{d=1}^D (\mu_d^* (\epsilon + \xi_d^* + y_d - E[\eta^\top \bar{Z}_d]) + v_d \xi_d + v_d^* \xi_d^*) \\ & - \sum_{d=1}^D \sum_{n=1}^N c_{d,n} \left(\sum_{t=1}^K \phi_{d,n,t} - 1 \right) \end{aligned} \quad (3)$$

この L^r を各パラメータに関して最適化することにより更新式を得る.

E-Step:

- γ に関して L^r を最適化: γ は α と ϕ から決定する.

$$\gamma_d \leftarrow \alpha + \sum_{n=1}^N \phi_{d,n} \quad (4)$$

- ϕ に関して L^r を最適化: $\partial L^r / \partial \phi_{d,n} = 0$ とし, 次式が得られる.

$$\begin{aligned} \phi_{d,i} \propto & \exp(E[\log \theta | \gamma] + E[\log p(w_{d,n} | \beta)]) \\ & + \frac{y_d}{N_d \delta^2} E[\eta] \\ & - \frac{2E[\eta^\top \phi_{d,-i} \eta] + E[\eta \circ \eta]}{2N^2 \delta^2} \\ & + \frac{E[\eta]}{N} (\mu_d - \mu_d^*) \end{aligned} \quad (5)$$

なお, $\phi_{d,-i} = \sum_{n \neq i} \phi_{d,n}$ であり, 単語 $\phi_{d,i}$ 以外の ϕ の総和を表す. $\eta \circ \eta$ はアダマール積であり, η 同士の各要素の積からなるベクトルである.

- $q(\eta)$ に関して L^r を最適化: A を, 各行がベクトル \bar{Z}_d^\top からなる $D \times K$ 行列と定義する. $\partial L^r / \partial q(\eta) =$

0 として, 次式を得る

$$q(\eta) = \frac{p_0(\eta)}{S} \exp\left(\eta^\top \sum_{d=1}^D (\mu_d - \mu_d^* + \frac{y_d}{\delta^2}) E[\bar{Z}_d] - \eta^\top \frac{E[A^\top A]}{2\delta^2}\right) \quad (6)$$

また, $E[A^\top A] = \sum_{d=1}^D E[\bar{Z}_d \bar{Z}_d^\top]$, $E[\bar{Z}_d \bar{Z}_d^\top] = 1/N^2 (\sum_{n=1}^N \sum_{m \neq n} \phi_{d,n} \phi_{d,m}^\top + \sum_{n=1}^N \text{diag}\{\phi_{d,n}\})$, S は定数である. 得られた $q(\eta)$ を L^r に代入することによって, 以下の双対問題が得られる.

$$\max_{\mu, \mu^*, \epsilon} -\frac{1}{2} a^\top \Sigma a - \epsilon \sum_{d=1}^D (\mu_d + \mu_d^*) + \sum_{d=1}^D y_d (\mu_d - \mu_d^*) \quad (7)$$

なお, $a = \sum_{d=1}^D (\mu_d - \mu_d^* + y_d / \delta^2) E[\bar{Z}_d]$ であり, この双対問題を SVM-light¹ などのソルバーによって解くことで $q(\eta)$, μ , μ^* を得る.

M-Step: β と δ^2 の更新式は以下の通りである.

- β に関して L^r を最適化:

$$\beta_{k,w} \propto \sum_{d=1}^D \sum_{n=1}^{N_d} I(w_{d,n} = w) \phi_{d,n,t} \quad (8)$$

$I(w_{d,n} = w)$ は, 単語 n の語彙が w である場合にのみ $\beta_{k,w}$ に加算することを意味する.

- δ^2 に関して L^r を最適化:

$$\delta^2 \leftarrow \frac{1}{D} (y^\top y - 2y^\top E[A] E[\eta] + E[\eta^\top E[A^\top A] \eta]) \quad (9)$$

なお, $E[\eta^\top E[A^\top A] \eta] = \text{tr}(E[A^\top A] E[\eta \eta^\top])$ であり, tr は行列の対角成分の和を表す.

2.1.2 分類タスクを想定した最大マージントピックモデル

離散ラベルを持つ文書データを扱う MedLDA Classification (MedLDA-Cla) について説明する. MedLDA-Cla では以下のような最適化問題が定義される.

P2(MedLDA-Cla):

$$\min_{q(Z, \Theta), \alpha, \beta, \xi} \mathcal{L}(q(Z, \Theta))(q) + KL(q(\eta) \| p_0(\eta)) + C \sum_{d=1}^D (\xi_d)$$

$$\text{s.t. } \forall d: \begin{cases} y \neq y_d \\ E[\eta^\top \Delta f_d(y)] \geq 1 - \xi_d \\ \xi_d \geq 0 \end{cases}$$

¹ <http://svmlight.joachims.org/>

各変数は $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_D\}$, $\mathbf{z}_d = \{z_{d,1}, \dots, z_{d,N_d}\}$, $\Theta = \{\theta_1, \dots, \theta_D\}$, $\mathbf{Y} = \{y_1, \dots, y_D\}$, $\mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_D\}$, $\mathbf{w}_d = \{w_{d,1}, \dots, w_{d,N_d}\}$, $\beta = \{\beta_1, \dots, \beta_K\}$ を表す. 制約式中の ξ は訓練データの誤差を吸収する程度を示すスラック変数であり, 文書ごとに設定する. KL は2つの確率分布の差異を表すカルバック・ライブラー情報量であり, 次式で表される.

$$KL(q(\eta) \parallel p_0(\eta)) = \int q(\eta) \log \frac{q(\eta)}{p_0(\eta)} \quad (10)$$

ここからは MedLDA-Cla の更新式の導出を行う. 目的関数において,

$$\mathcal{L}(q(Z, \Theta)) = -E[\log p(\theta, \mathbf{z}, \mathbf{Y}, \mathbf{W} \mid \alpha, \beta, \delta^2)] - \mathcal{H}(q(Z, \Theta)) \quad (11)$$

$$\Delta \mathbf{f}_d(y) = \mathbf{f}(y_d, \bar{Z}_d) - \mathbf{f}(y, \bar{Z}_d) \quad (12)$$

である. MedLDA-Reg と同様に最適化問題 P2 に関しても変分近似を行い, q についての条件付独立性を与える.

$$q(\Theta, Z \mid \gamma, \phi) = \prod_{d=1}^D q(\theta_d \mid \gamma) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}) \quad (13)$$

そして目的関数に制約式を含めたラグランジュ関数 L^c を次のように定義し, L^c を各パラメータに関して最適化することで更新式を得る. なお, γ, β に関しては MedLDA-Reg と更新式が同じであるため省略する.

$$\begin{aligned} L^c &= \mathcal{L}(q(Z, \Theta)) + KL(q(\eta) \parallel p_0(\eta)) \\ &+ C \sum_{d=1}^D \xi_d - \sum_{d=1}^D v_d \xi_d \\ &- \sum_{d=1}^D \sum_{y \neq y_d} \mu_d(y) (E[\eta^\top \Delta \mathbf{f}_d(y)] + \xi_d - 1) \\ &- \sum_{d=1}^D \sum_{n=1}^{N_d} c_{d,n} \left(\sum_{t=1}^K \phi_{d,n,t} - 1 \right) \end{aligned} \quad (14)$$

E-Step :

- ϕ に関して L^c を最適化 : $\partial L^c / \partial \phi_{d,n}$ とし, 次式が得られる.

$$\begin{aligned} \phi_{d,n} &\propto \exp(E[\log \theta \mid \gamma] + E[\log p(w_{d,n} \mid \beta)]) \\ &+ \frac{1}{N} \sum_{y \neq y_d} \mu_d(y) E[\eta_{y_d} - \eta_y] \end{aligned} \quad (15)$$

最初の2項は MedLDA-Reg と同様である.

- $q(\eta)$ に関して L^c を最適化 :

$$q(\eta) = \frac{1}{Z} p_0(\eta) \exp(\eta^\top \mu_\eta) \quad (16)$$

ただし, $\mu_\eta = \sum_{d=1}^D \sum_{y \neq y_d} \mu_d(y) E[\Delta \mathbf{f}_d(y)]$.

2.2 双対分解

双対分解 (dual decomposition) は, 複雑な目的関数を効率的に求める手法である. 直接的に求めることが困難な目的関数をいくつかの関数に分割でき, それぞれの関数の最適解が効率的に求まる場合に適用可能である. 効率的に解くことができない次の関数を対象として, 双対分解の例を示す.

$$\arg \max_y f(y) + h(y) \quad (17)$$

$\arg \max_y f(y)$, $\arg \max_y h(y)$ は効率的に求まると仮定する. このとき, 次の問題は上の問題と同じ意味を持つ.

$$\arg \max_{y,z} f(z) + h(y) \quad (18)$$

$$\text{s.t. } y = z \quad (19)$$

この問題の解を L^* とする. そして, この問題に対してラグランジュ緩和を適用する.

$$L(u, y, z) = f(z) + h(y) + u(y - z) \quad (20)$$

u はラグランジュ乗数である. 次に $L(u, y, z)$ に関して最大値をとるものを考える.

$$\begin{aligned} L(u) &= \max_{y,z} L(u, y, z) \\ &= \max_z (f(z) - uz) + \max_y (h(y) + uy) \end{aligned} \quad (21)$$

この関数は $y = z$ の制約を持たないため, 最初の問題より広い解空間を持ち, $L^* \leq L(u)$ が成り立つ. これにより, 元の最適化問題の上限を与えている. よって, 双対定理により以下が成り立つ.

$$L^* = \min_u L(u) \quad (22)$$

$\min_u L(u)$ は凸関数であるので, u に関する勾配を求めることができれば, 勾配降下法により最適化できる. よって, 劣微分の1つである d_u は次のように求めることができる.

$$d_u = y^* - z^* \quad (23)$$

$$z^* = \arg \max_z f(z) - uz \quad (24)$$

$$y^* = \arg \max_y f(y) - uy \quad (25)$$

そして, 勾配法に基づき以下のように u を更新する.

$$u \leftarrow u - \mu(y^* - z^*) \quad (26)$$

μ はステップ数である. この更新を繰り返して $L(u)$ を小さくし, $y^* = z^*$ となる時が主問題と双対問題の値が一致したときなので, 最適解を求めることができる.

3 双対分解を利用したマルチタスク最大マージントピックモデル

3.1 モデルの定義

2.2節で述べたように、連続値または離散値の付加情報を持つ文書データの解析を行うためには MedLDA を利用すればよい。しかし、MedLDA では連続値と離散値の両方の付加情報を持つ文書データの解析を行うことができない。この問題を解決する為に、我々は双対分解を利用したマルチタスク最大マージントピックモデル (Multi-task MedLDA: MultiMedLDA) を提案する。MultiMedLDA では複数種類のラベルが付与された文書に対して適用可能なモデルであり、MedLDA を双対分解を利用して拡張している。以下に MultiMedLDA の生成過程を示す。

1. 文書 $d (d \in 1, \dots, D)$ に対して、 $\theta \sim \text{Dirichlet}(\alpha)$ を選択。
2. 文書 d 内の N_d 個の単語 $w_{d,n} (n \in 1, \dots, N_d)$ に対して、
 - (a) トピック $z_{d,n} \sim \text{Multinomial}(\theta_d)$ を選択。
 - (b) 単語 $w_{d,n} \sim \text{Multinomial}(\beta_k)$ を選択。
3. D 個の文書に対して、連続ラベル $y_d^r \sim F(\eta^r, z_d)$ 、離散ラベル $y_d^c \sim F(\eta^c, z_d)$ を選択。

なお、 η^r, η^c は重み係数である。MultiMedLDA のグラフィカルモデルを図 2 に示す。

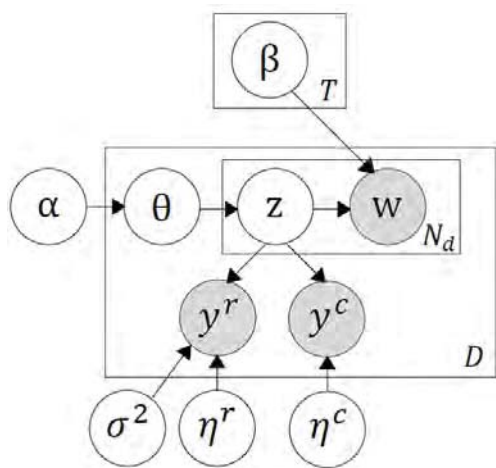


図 2: MultiMedLDA のグラフィカルモデル

3.2 モデルの推定

連続ラベル $y^r \in R$ と離散ラベル $y^c \in \{1, \dots, M\}$ が各文書に付加されている、データセットについて考える。このとき、MedLDA-Reg の最適化問題と MedLDA-Cla の最適化問題を統合することによって、以下のような最適化問題を定義することができる。なお、目的関数第 1, 2 項が回帰タスクの目的関数、目的関数第 3, 4 項が分類タスクの目的関数である。同様に制約式第 1, 2, 3 行が回帰タスクの制約式、制約式第 4, 5 行が分類タスクの制約式である。第 6 行は双対分解のための制約である。

$$\begin{aligned}
 \text{P3(MultiMedLDA)} : \quad & \min_{q(Z^r, \Theta^r, \eta^r), q(Z^c, \Theta^c, \eta^c), \alpha, \beta, \delta^2, \xi^r, \xi^{r*}, \xi^c} \\
 & \mathcal{L}^r(q(Z^r, \Theta^r, \eta^r)) + C^r \sum_{d=1}^D (\xi_d^r + \xi_d^{r*}) \\
 & + \mathcal{L}^c(q(Z^c, \Theta^c, \eta^c)) + C^c \sum_{d=1}^D \xi_d^c \\
 \text{subject to } \forall d \quad & \begin{cases} y_d^r - E[\eta^{r\top} \bar{z}_d] \leq \epsilon + \xi_d^r \\ -y_d^r + E[\eta^{r\top} \bar{z}_d] \leq \epsilon + \xi_d^{r*} \\ \xi_d \geq 0, \xi_d^* \geq 0 \\ y^c \neq y_d^c : E[\eta^{c\top} \Delta \mathbf{f}_d(\mathbf{y}^c)] \geq 1 - \xi_d^c \\ \xi_d^c \geq 0 \\ \phi^r = \phi^c \end{cases}
 \end{aligned}$$

各変数は $Z^r = \{\mathbf{z}_1^r, \dots, \mathbf{z}_D^r\}$, $\mathbf{z}^r = \{z_{d,1}^r, \dots, z_{d,N_d}^r\}$, $Z^c = \{\mathbf{z}_1^c, \dots, \mathbf{z}_D^c\}$, $\mathbf{z}^c = \{z_{d,1}^c, \dots, z_{d,N_d}^c\}$, $E[z_{d,n}^r] = \phi_{d,n}^r$, $E[z_{d,n}^c] = \phi_{d,n}^c$, $\Theta = \{\theta_1, \dots, \theta_D\}$ である。そして、 ξ^r, ξ^{r*}, ξ^c はそれぞれ訓練データの誤差を吸収する程度を示すスラック変数、 ϵ は許容誤差である。また MedLDA-Cla と同様に $\Delta \mathbf{f}_d(\mathbf{y}^c) = \mathbf{f}(y_d^c, \bar{z}_d) - \mathbf{f}(y, \bar{z}_d)$ である。 $\mathbf{f}(y^c, \bar{z}_d)$ は $(y^c - 1)K + 1$ から $y^c K$ の要素がベクトル \bar{z} であり、他の要素が 0 であるような特徴ベクトルであり、 $\bar{z}_d \leftarrow (1/N_d) \sum_{n=1}^N z_{d,n}$ である。

以下では回帰タスクに関する目的関数第 1, 2 項を $\mathcal{L}(R)$ 、分類タスクに関する目的関数第 3, 4 項を $\mathcal{L}(C)$ とする。この最適化問題に対してラグランジュ緩和を行い、次の最適化問題を得る。なお、簡単のため制約式は省略している。

$$L(U, \phi^r, \phi^c) = \mathcal{L}(R) + \mathcal{L}(C) + U(\phi^c - \phi^r)$$

U はラグランジュ乗数である。次に $L(U, \phi^r, \phi^c)$ を最小化する ϕ^r, ϕ^c を考える。

$$\begin{aligned}
 L(U) &= \min_{\phi^r, \phi^c} L(U, \phi^r, \phi^c) \\
 &= \min_{\phi^r} (\mathcal{L}(R) - U\phi^r) + \min_{\phi^c} (\mathcal{L}(C) + U\phi^c) \quad (27)
 \end{aligned}$$

この関数には最適化問題 P3 の制約式第 6 行にある $\phi^r = \phi^c$ が考慮されていないので、より広い解空間を持つ。これにより、 $L^* \geq L(U)$ が成り立つので、最適化問題 P3 の下限を与えている。また、双対定理より $L^* = \max L(U)$ が成り立つ。よって、 $L(U)$ の劣微分の 1 つである d_U 、および ϕ^{r*} 、 ϕ^{c*} 、ラグランジュ乗数 U は以下ようになる。

$$d_U = \phi^{c*} - \phi^{r*} \quad (28)$$

$$\phi^{r*} = \arg \min_{\phi^r} \mathcal{L}(R) - U\phi^r \quad (29)$$

$$\phi^{c*} = \arg \min_{\phi^c} \mathcal{L}(C) - U\phi^c \quad (30)$$

$$U \leftarrow U - \mu(\phi^{c*} - \phi^{r*}) \quad (31)$$

なお、 μ はステップ幅であり、本研究では反復回 S の逆数を用いている。回帰タスクと分類タスクで潜在トピック ϕ^r 、 ϕ^c の推定を行った後、この更新を繰り返すことで下限 $L(U)$ の最大化を行う。そして $\phi^{r*} = \phi^{c*}$ になった時が主問題と双対問題の値が一致したときなので最適解に到達したことが保証される。

これにより得られた ϕ^{r*} 、 ϕ^{c*} をそれぞれの最適化問題に与えなおすことによって、もう一方の影響を考慮した潜在トピックの推定が可能となる。なお、Multi-MedLDA の最適化問題は、MedLDA とは異なり制約式に $\phi^r = \phi^c$ が追加される。これにより、 $-U\phi^r$ 、 $U\phi^c$ の項が偏微分をした後にも残る。よって ϕ^r 、 ϕ^c の更新式は以下ようになる。

$$\begin{aligned} \phi_{d,n}^r &\propto \exp\left(E[\log \theta | \gamma] + E[\log p(w_{d,n} | \beta)]\right) \\ &\quad - \frac{2E[\eta^\top \phi_{d,-i} \eta] + E[\eta \circ \eta]}{2N^2 \delta^2} \\ &\quad + \frac{E[\eta]}{N}(\mu_d - \mu_d^*) + \frac{y_d}{N_d \delta^2} E[\eta] - U \end{aligned} \quad (32)$$

$$\begin{aligned} \phi_{d,n}^c &\propto \exp\left(E[\log \theta | \gamma] + E[\log p(w_{d,n} | \beta)]\right) \\ &\quad + \frac{1}{N} \sum_{y \neq y_d} \mu_d(y) E[\eta_{y_d} - \eta_y] + U \end{aligned} \quad (33)$$

4 実験

4.1 データセット

本研究では、データセットとして東洋経済新報社が発行する会社四季報を使用した。これは、四半期ごとに発表される上場企業 3675 社 (2017 年度新春版) の、企業名をはじめとした上場コード、業種、営業利益、株価、短評などを載せている記事である。2014 年度新春版から 2017 年度新春版までの 13 四半期分のデータを使用した。文書データには短評、離散値として業種、連続値として前年度からの営業利益の変化率を使用した。

各企業には上場する際に登録された 32 種類の業種のうち 1 つが選ばれており、その中で登録企業数の多い上位 10 種類 (サービス業、情報・通信業、小売業、卸売業、電気機器、機械、化学、建設業、食料品、輸送用機器) の業種の企業データを使用した。各業種の企業数には偏りが存在するため、1 業種につき 88 企業を無作為に選択している。また、全文書中で 3 文書未満しか出現しない低頻度語を除外している。なお、文書データは MeCab¹ を用いて形態素解析を行い、助詞や接続詞といった機能語を除外している。以上の処理を行ったデータセットの情報を表 1 に示す。

表 1: 四季報データセット

	Shikiho
Number of documents	2657
Number of words	437539
Size of vocabulary	6945
Number of labels	10

また、次節以降の実験を行う為に、データセットの分割を行った。まず各年度 (2014 年, 2015 年, 2016 年) で文書を 3 分割し、2015 年と 2016 年のデータに関しては企業単位で 20% をテストセットとして確保した。残りは最適パラメータを導出する実験を行う為に、60% の学習セットと 20% の検証セットに分けた。

4.2 実験設定

提案手法である MultiMedLDA、データセットの時系列性を考慮した MultiMedLDA-Seq、既存手法である MedLDA-Reg の 3 つのモデルで最適パラメータを導出する実験を行った。実験で使用するデータセットは、全ての企業に対して 3 年分の時系列データとなっており、MultiMedLDA-Seq では前年の学習で得られた β と η を初期状態として与えている。ただし、初年度である 2014 年のデータは初期状態を乱数によって定めている。ハイパーパラメータは $\alpha = 0.1$ 、損失パラメータは $l = 1$ に設定した。また、提案した 2 手法の学習において、MedLDA-Reg と MedLDA-Cla の特徴を活かすため、双方の学習結果に影響させない burn-in period を 5 回目の反復までに設定している。これにより、回帰タスクと分類タスクの特徴を活かした状態で潜在トピックの統合が図られている。学習を停止する条件として、テストセット対数尤度 [8] の変化率が負であることを設けている。

正則化パラメータ C^r 、 C^c を以下の手順で決定する。 $C^r \in \{0.25, 1, 4, 16, 64\}$ 、 $C^c \in \{0.25, 1, 4, 16, 64\}$ の範

¹<http://taku910.github.io/mecab/>

囲で値を変化させ、2015年と2016年の検証セットからRMSEを計算する。そして2015年と2016年のRMSEの平均が最も優れているものを最適な正則化パラメータ C^r , C^c とする。RMSEは4.3.2節で説明する。

次に回帰問題における許容誤差 ϵ の決定について述べる。2014年、2015年、2016年の各年度に関して $\epsilon \in \{0.01, 0.1, 1, 2, 4\}$ の範囲で値を変化させ、2015年と2016年の検証セットからRMSEを計算する。そして2015年と2016年のRMSEの平均が最も優れているものを最適な許容誤差パラメータ ϵ として決定した。

最後に得られた最適パラメータ C^r , C^c , ϵ を設定して、2015年と2016年のテストセットを使用した連続値ラベルの予測を行う。以上の実験をトピック数 $T \in \{15, 20, 25\}$ で行い、提案手法である MultiMedLDA-Seq, MultiMedLDA と既存手法である MedLDA-Reg を比較することによって、提案手法の評価を行う。

4.3 評価尺度

4.3.1 Root Mean Squared Error : RMSE

Root Mean Squared Error(以下 RMSE) はモデルの予測能力を表す指標のひとつである。モデルの予測値と真値から算出される相対的な評価指標である。RMSE は以下の式で表される。

$$RMSE = \sqrt{\frac{1}{D} \sum_{d=1}^D (y_d - \hat{y}_d)^2} \quad (34)$$

y_d はモデルの予測値であり、 \hat{y}_d は真値である。予測値が真値から離れているほど大きい値をとるため、0に近いほど優れている。

4.4 予備実験

本実験の前に、提案手法である MultiMedLDA-Seq と MultiMedLDA, 既存手法である MedLDA-Reg において正則化パラメータ C^r および C^c , 許容誤差 ϵ の最適値を導出する予備実験を行った。本実験用データを除いたデータセットを4つのデータセットに分割し、3セットを学習用、1セットを検証用とした。そして上述のRMSEで評価することによって、最適なパラメータを決定した。各手法のトピック数 $T \in \{15, 20, 25\}$ での最適値を表2, 表3, 表4に示す

4.5 連続値ラベルの予測実験および考察

4.4節で導出した C^r , C^c , ϵ を用いて各モデルの連続ラベル予測性能比較実験を行った。評価指標はRMSEを

表 2: トピック数 $T=15$ での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	64	4	1
MultiMedLDA	0.25	16	1
MedLDA-Reg	-	0.25	0.1

表 3: トピック数 $T=20$ での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	16	4	0.01
MultiMedLDA	64	1	0.01
MedLDA-Reg	-	4	0.01

表 4: トピック数 $T=25$ での最適パラメータ

	C^c	C^r	ϵ
MultiMedLDA-Seq	16	64	1
MultiMedLDA	4	0.25	0.01
MedLDA-Reg	-	0.25	0.01

用いている。結果は図3のようになった。図3から、全てのトピック数に関して提案手法である MultiMedLDA-Seq および MultiMedLDA が MedLDA-Reg よりも RMSE で改善された結果を示していることがわかる。これは、提案手法は連続値ラベルと離散値ラベルの両方を考慮しているので、より精度の高い連続値ラベルの予測が行えていると考えられる。また、本研究における実験では MedLDA の原著論文で使用されていたデータセットよりも小規模なデータセットを使用している。このため、単一のラベルのみを考慮した MedLDA-Reg では文書の潜在トピック構造をうまく解析できていないと考えられる。それに対して MultiMedLDA では複数種類のラベルを付加情報として用いているので、文書データによる潜在トピック構造の推定を、付加情報がうまく補完していると推測する。更に、MultiMedLDA-Seq は時系列性を考慮しない MultiMedLDA よりも RMSE で改善された結果を示している。これは、MultiMedLDA-Seq がデータセットの時系列性を捉えることによって、連続値ラベルの推定に適したトピックの構築が可能になっていると考えられる。最後に、RMSE が最も低くなるトピック数は、既存手法では $T = 25$ であるのに対して、提案手法では $T = 15$ である。このことから、既存の単一の回帰問題を想定した MedLDA では回帰問題を一部のトピックでのみ対応し、残りのトピックは単語の事後確率最大化にのみ寄与するような分担が起こっていると考えられる。それに対して、提案する MultiMedLDA では双対分解によって回帰と分類と事後確率最大化を共通のトピックで実現していると考え

られる。以上より、提案手法は、複数種類のラベルを持つ文書データに対して、既存手法である MedLDA-Reg よりも優れた手法であるといえる。

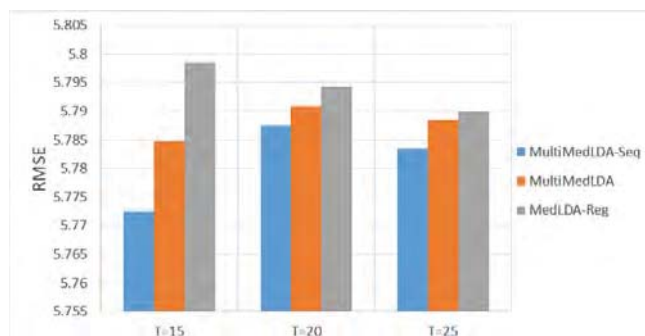


図 3: 四季報データセットでの各手法の RMSE

5 おわりに

本論文では、複数種類のラベルが付加されている文書に対して適用することができるモデルである MultiMedLDA を提案した。MultiMedLDA は既存手法である MedLDA とは異なり、複数の分類問題や回帰問題を伴うマルチタスク問題を最適化することが可能である。そこで、多くの付加情報を持つ会社四季報の企業データにたいしてこのモデルを適用した。実験では連続値ラベルの推定を行い、各年度の RMSE の平均値において MedLDA よりも優れた結果を示したことから MultiMedLDA の優位性を示すことができた。また、データセットの時系列性を考慮することによって更なるモデルの改善が可能であることも示せた。

モデルの改善点としては、分類問題と回帰問題でそれぞれの性能が発揮されるトピック数が異なる場合に、どちらかのトピックを選択しなければならない問題が挙げられる。この問題を解決し、それぞれに適切なトピック数を設定することによって、MultiMedLDA の予測性能を更に高めることが期待できる。

今後の展望として、より多くの付加情報を考慮することが挙げられる。MultiMedLDA は連続値と離散値の複数種類の付加情報を持つ文書データに対応したモデルであり、本論文ではそれぞれ 1 つずつ持つ四季報データセットを使用した。より多くの連続値ラベルと離散値ラベルを考慮することで、更なる予測精度の向上が期待される。また、単語が属す話題を表す潜在トピックの推定に加え、文章の修辭的機能を解析する研究 [9] も存在する。このような異なるモデルを組み合わせることで、モデルの精度を向上させることも期待できる。最後に、経済・金融業界において大規模データの解析に対する関心は高まっており、今回扱った会社四

季報の企業活動データだけでなく、株価、為替、債権などより複雑な要因を持ち、多くの付加情報を持ちうるデータセットに対して MultiMedLDA を適用することで、モデルの性能を示すことができるかもしれない。

謝辞 本研究を行うにあたり、有益な助言を頂いた神戸大学大学院経済学研究科の羽森茂之教授と金京拓司教授に感謝する。本研究の一部は科学研究費補助金基盤研究 (B)(15H02703) の援助による。

参考文献

- [1] David M Blei, Andrew Y Ng, and Michael I Jordan.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, (2003)
- [2] David M Blei and Jon D. McAuliffe.: Supervised topic models, *Advances in neural information processing systems*, pp. 121–128, (2008)
- [3] Jun Zhu, Amr Ahmed, and Eric P Xing.: MedLDA: Maximum Margin Supervised Topic Models, *Journal of Machine Learning Research*, Vol. 13, pp. 2237–2278, (2012)
- [4] Vladimir Vapnik.: *Statistical Learning Theory*, John Wiley and Sons, New York, (1998)
- [5] Alex J Smola and Bernhard Schölkopf.: A tutorial on support vector regression, *Statistics and Computing*, Vol. 14, pp. 199–222, (2004)
- [6] Tommi Jaakkola, Marina Meila, and Tony Jebara.: Maximum entropy discrimination, *Neural Information Processing Systems*, Vol. 12, pp. 470–476, (1999)
- [7] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright.: Optimization for Machine Learning, *Neural Information Processing Systems*, Mit Press, (2012)
- [8] Yee Whye Teh, David Newman, and Max Welling.: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *Neural Information Processing Systems*, Vol. 6, pp. 1378–1385, (2006)
- [9] Bei Chen, Jun Zhum Nan Yang, Tian Tian, Ming Zhou, and Bo Zhang.: Jointly Modeling Topics and Intents with Global Order Structure, *arXiv preprint arXiv*, 1512, 02009, (2015).

グローバルなジオ・ポリティカル（地政学的）リスクの 投資運用における計測方法についての検討

A discussion on how to measure and manage global geopolitical risk
for the purpose of investment risk management

西山 昇

Noboru Nishiyama

Dragons' Desk Ltd. / 千葉商科大学

Abstract: The purpose of this discussion is to illustrate in practice the investment risk analysis of geopolitical events such as Brexit, China shock, and similar events, which traditionally have been forecasted using qualitative approaches. I discuss what quantitative methods are available for modeling and forecasting geopolitical risk from the perspective of investment practitioners, and provide usage case examples.

Key words: Statistical Multi-factor model, GARCH, Variance-Covariance matrix estimation, VaR, Geopolitical Risk, Asset Allocation, Downside Risk.

1. はじめに

フォワードルッキングなリスク計測には、統計的に、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。

昨年（2016年）には、Brexit（英国のEUからの離脱）の国民投票での勝利、米国大統領選挙など（世論調査の結果から予想される結果なら）資産価格に波及する程度が小さいと見られていたイベントが、（世論調査の結果と逆になり）資産価格の変動に大きく影響する事例が発生した。

これらイベントは、事後的に分析することもでき、事前に何が発生するかの要因（ファクター）を特定することができあ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は可能なのだろうか？

本稿では、説明変数を最初から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してモデルの適用可能性を探る。

検証作業の詳細は後述するが、イベント発生日の前後に日次でモデル推定を実施して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ではあるものの、モデルによる推定を実施する前日までのデータを利用してその先を予測するアウトサンプルの方法をとっている。

2. 地政学的リスクとは

2.1 地政学的リスクの事例

昨年（2016年）1月、イアン・ブレマー氏（Eurasia Group）は2016年に予想される地政学的リスクとして次5点をあげていた。[1]

- (1) 同盟関係の分断（THE HOLLOW ALLIANCE）— 米国が世界の指導的立場ではなくなり、欧州における影響力が低下することなど。
- (2) 閉じられた欧州（A CLOSED EUROPE）— 開かれた欧州と閉じられた欧州との間での軋轢が顕在化すること、EU設立の目的である国境をなくすとの方向性が揺らぐこと、英国が欧州から離脱する可能性が高いことなど。
- (3) 中国の足跡（THE CHINA FOOTPRINT）— 経済大国としてグローバル経済に影響を与えていくなかで、国内政治と国外経済活動とのアンバランスが発生することなど。
- (4) テロ組織 ISIS と他の組織との連携（ISIS AND "FRIENDS"）— テロ活動が ISIS を中心に世界各

地に拡大することなど。

- (5) サウジアラビアの亀裂(SAUDI ARABIA'S RIFTS)
—中東地域でのサウジアラビアの政治的な孤立と近隣のイランとの関係の緊張化など。

これらのことから、地政学的リスクの重要な一面には、国同士の政治的な緊張関係の大きな変化があることがわかる。英国の EU 離脱、中国の金融政策等、2006 年を振り返ると地政学的リスクとして捉えられる複数の事象が頭に浮かぶ。

ここで考えることは、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のかとの点である。

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) にとってエマージングリスクの一つとして取り扱われている。[2]

エマージングリスクとは、それまで無関係とされていたファクター (要因) 同士に相関が突然発生することによ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を、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。

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するのではなく、ファクターと感応度を同時に推定できる APT 型の統計的ファクターモデルを適用する。[3]

R : return, F : risk factors, β : sensitivity

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

(式 2-1) 統計的マルチファクターモデルのリターン表現

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 2-2) 統計的マルチファクターモデルのリスク表現

(式 2-2) において、 Σ はトータルリスクとしての分散共分散、右辺第二項 D_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計測する。

実務的なリスク属性分析としては、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寄与の計測、トラッキングエラー、ベータ、リスクファクター、投資スタイルのトレンド分析などの推定がある。

3.2 過去データのシミュレーション

外国株式の仮想ポートフォリオを作成していくつかのイベントが発生したポイントで統計的マルチファクターモデルによるリスク分析をおこなう。

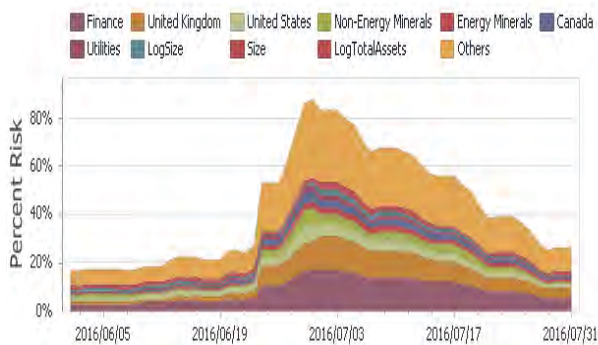
仮想ポートフォリオは、世界の株式の時価総額上位から 200 銘柄を抽出する。また 200 銘柄は等加重で投資を想定している。よって本ポートフォリオは特別な運用スタイルをとっていない。世界の各国の代表的な大型株に分散投資するスタイルをとる。

一つ目のイベントは、2016 年 6 月 24 日に発生した Brexit(英国の EU 離脱を決めた国民投票)である。

(図 3-1) は、リスクファクターとしてのアクティブエクスポージャ (スタイル) の種類と前後約 1 カ月 (2016 年 6 月 1 日~2016 年 7 月 31 日) 間の変動の推移である。

Brexit (2016 年 6 月 24 日) 直前まで漸増していた Finance(金融) (グラフ最下層) と United Kingdom(英国) (グラフ最下層から 2 番目) のリスクファクターが、2016 年 6 月 24 日以降急増している。

また 2016 年 7 月 31 日 (最右端) では、Finance(金融) と United Kingdom(英国)のアクティブエクスポージャが Brexit 直前の水準まで低下していることが観察される。



(図 3-1) Brexit 前後のアクティブエクスポージャの推移
(Analytics by EM Applications)

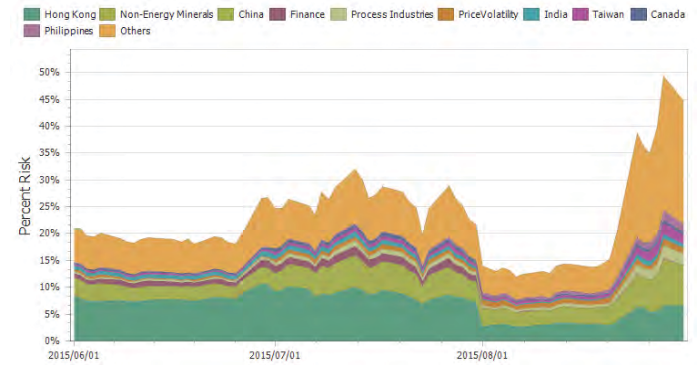
二つ目のイベントは、2015 年 7 月に発生した中国ショックと呼ばれる株安である。

中国人民銀行は、8 月 11 日に人民元の切り下げに踏み切った。そこから世界的な株安がスタートした事例である。

(図 3-2) では、アクティブエクスポージャ (スタイル) の種類と前後 (2016 年 6 月 1 日~同年 8 月 31 日) 間の変動の推移である。

グラフからは、Hong Kong(香港) (グラフ最下層) と China(中国)(グラフ最下層から 2 番目に大きな層)のアクティブエクスポージャが高いことがわかる。また中国のアクティブエクスポージャは、この期間

を通じて不安定な動きをみせていることが観察される。



(図 3-2) 中国ショック前後のアクティブエクスポージャの推移
(Analytics by EM Applications)

4. 結果の考察

4.1 異なるリスクファクターの意味

本稿では、二つのイベントが発生した期間に、まったく同じ等加重のポートフォリオ (200 銘柄) を統計的マルチファクターモデルの同じグローバル株式モデルによりリスク分析している。

Brexit の時期には、Finance(金融) と United Kingdom(英国)のリスクファクターが主にポートフォリオのリスクに大きな影響を与えており、中国ショックの時期には、Hong Kong(香港)と China(中国)がリスクファクターとしての影響が大きいことが計測により示されている。

フォワードルッキングなリスク計測のためには、トータルリスクの短期変動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枠組みとして統計的マルチファクターモデルの適用できることを確認した。

グラフで示したのは、(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。

このあとの課題は、これらのリスクファクターを実際の投資運用においてフォワードルッキングなリスク計測指標としてどう活用するかになる。

4.2 実際の運用のために必要なこと

各ファクターに対するエクスポージャの大きさがわかれば、全体のリスクを低下させるために必要な行動が明確になる。

単純な方法としては、リスクエクスポージャの大

きな銘柄から保有量を減らすことによりリスクを低下させる調整ができる。また最適化を使ってリスクを最小化する目的関数によりポートフォリオを再構築する方法がある

また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。

グローバルな投資運用会社では、このプロセスを通じて、ファンドマネージャと相談しながら運用リスク管理の側面から投資の意思決定を実行している。

モデルは複雑なものより説明しやすいシンプルなものが望ましい。それは地政学的リスクにおいては、リスクファクターが複雑にからみあう状況で計測され、あとで振り返ったときに結果の解釈が難しくなるからである。

5. おわりに

本稿では、フォワードルッキングなリスク計測のために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条件だった。

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ー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は有望な検証対象である。

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 Andorsoni 氏に感謝したい。

参考文献

- [1] The Top 5 Geopolitical Risks for 2016, Ian Bremmer, January 7, 2016, (<http://time.com/4170785/the-top-5-geopolitical-risks-for-2016/>)
- [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://emapapplications.com/index.php?q=research/statistical-factor-model/stat-factor-model>)

AI トレーダー コンピュータビジョン編

AI trader computer vision edition

河合 継¹ 小澤 昂^{1,2}

Kawai Kei, Ozawa Takashi^{1,2}

¹クリスタルメソッド株式会社

¹Crystal method co.ltd

²東京工業大学/数学科 4 年/数理ファイナンス専攻

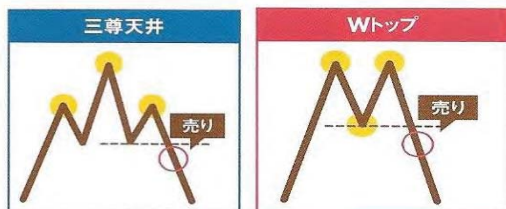
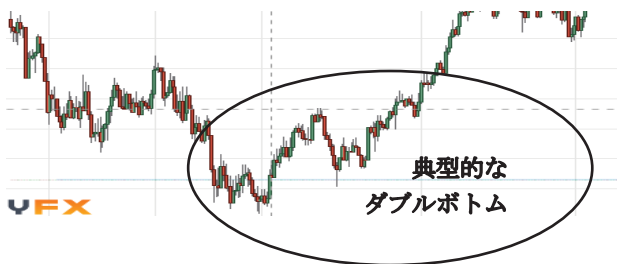
²Tokyo Institute of Technology Department of Mathematics

1. はじめに

1-1. トレーダー

2010 年末からトレーダーになる為、半年ほど FX スクールに通って取引の勉強をしていた。当時やっていたのは、全くの裁量取引で、人工知能とはかけ離れたもの。毎日 FX のチャートを 5 通貨ペアほどプリントアウトして、日足に決められた線を引き、傾向分析を行い、15 分足でその日のレンジなどを確認したり、支持線・抵抗線などを定めたりしていた。

そんな中で勝率が高いケースがいくつか存在していた。“三尊”・“ダブルトップ”・“フラッグ”などチャートのロウソク足の並びに名前を付けたものを目印にトレードしていた。



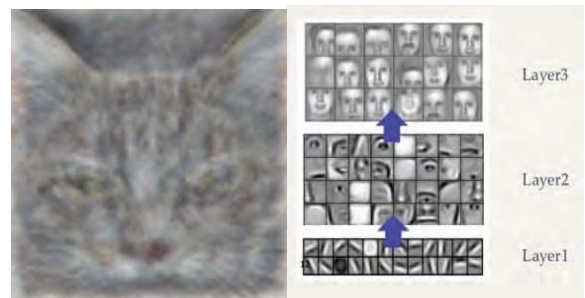
詳しいトレード手法は別の機会としたいが、ルールが守られている限りは 20 勝 1 敗というような成績を残すことが可能で、すぐに資産は 2 倍ほどに成長した。

ただし、そのトレードは非常に重労働であったし、見ているのは単純な値動きだった。私は自分がこの

世界に役に立っていない人間だと思えて来て、トレードのモチベーションを失った。自動で同じようなトレードが出来ないかという欲求には駆られていた。

1-2 AI や Deep Learning について

2012 年に Google が猫を判定したというのはあまりにも有名なニュースだった。ニューラルネットの中間層に猫の特徴をあらわした画像が出ており、人間の顔の特徴が出てきていたのがとても印象的だった。Deep Learning の利点は人間が気付かないような特徴も分析結果として現れるという事だった。ここから世間で語られていない全く新しい取引方法が見いだせる可能性があると考えた。



私もこの後、機械学習に触れ始めるようになっていたが、チャートの特徴分析を行う事によって、一度諦めたトレードが再度できるようになるのではないかと考え始めた。目で見て値を判定するという事にこだわったので、一番初めに行った検証は、画面上の画像としてのレートと文字情報としてトレード判定条件に利用できると考えて MNist 数字を判定する事で、機械学習の道に入った。

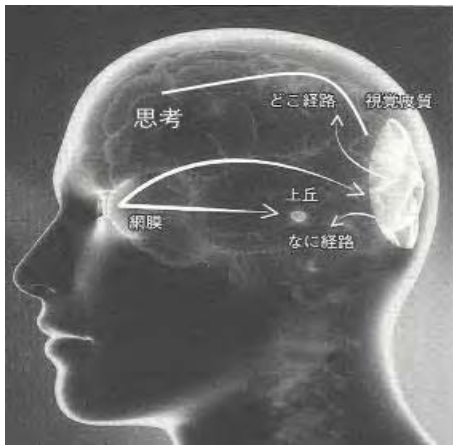
1-3 AI トレーダー

次は勝率の高いトレードを再現するために、レートの数字のみでなくチャートを目で見て判定してい

るように振舞うアプリケーションを開発できないか検討を始めた。

人間の目の働きは網膜に映った情報が神経を経て脳の後ろの視覚野に到達する。その後それは何か？などの判定に入るのだから、認識判断速度としては少なくとも300ミリ秒以上通常400ミリ秒ほどかかる。ただ、それよりもショートカットされた中脳の上丘（じょうきゅう）でも情報は認識される。視覚認識に15ミリ秒ほどと非常に早く、100ミリ秒で反応を行える。その経路は熟練したスキルを持つものが無意識に利用する経路である。

AIトレーダーの実現方法は人間の認知の仕組みと機械学習の利点を組み合わせようと考えた。上丘の動きを模した部分、特徴を認識し学習するする前頭葉にあたる部分、前意識と呼ばれる、勘などで反応ができる熟練性である。前頭葉では情報を本などから入力することにより学習し、学習したことを利用することによって熟練する。



2. 準備

2-1 準備 1

・チャート画像取得プログラム MQL4 C++
メタトレーダーのチャートをコマ送りするプログラムを作成することにより、2005年1月から2016年末までのUSD/JPY・15分足チャート画像30万枚をスクリーンショットとして保存した。同時にチャートの値動きと時刻情報をラベリングした。+-5 PIP以上の値動きをそれぞれラベリングし、5 PIP以内は同じラベルとした。

例 12005#01#17#23#00.png -5 以下-
22016#07#27#10#00.png 5 以上+
32009#05#05#14#30.png 5PIP 以内
ラベル年#月#日#時#分

・画像サンプル



・チャート画像加工プログラム

Png形式で保存されたデータに対し、前処理として、ビット深度やサイズ変更をするようなプログラムを作成した

・Chainer CNN 学習プログラム Python

Pythonで畳み込みニューラルネットワーク(3章で後述)を利用し、上記の画像を学習するプログラムを作成した。

・Chainer 学習結果返却プログラム Python

上記で学習した学習済みデータセットに2016年中の画像を入力として与えることで、売りが買いか何もしないかを返却するプログラムを作成した。

・チャート価格等取得プログラム MQL4

対象の画像時刻(チャート左端)チャートの現在時刻(チャートの右端)次の足時刻(チャートの右端+1 右端の終値、右端+1の終値、両者の価格差をすべての15分足で取得した。

・結果検証プログラム Python Postgres Excel

学習済みデータセットが返却した売買情報と上記のデータベースに対してSQLを発行し、Excelにわかりやすいようにまとめるプログラム

・ハードウェア

自作PC i7920 メモリ16G Ubuntu 16.04 GTX1060 6G

2-2 準備 2

ショートカット経路のあるあまり思考することなく反応する経路について、高速に作動するVisualC++を利用し、与えたテンプレートに対して、雰囲気似ている事を検知できるプログラムを作成した。

2-3 準備 3

通常人間は、本やその他の情報を入力として勉強し、スキルを身につけていく。今回はMT4のハーモニックパターン検出システムを調査した

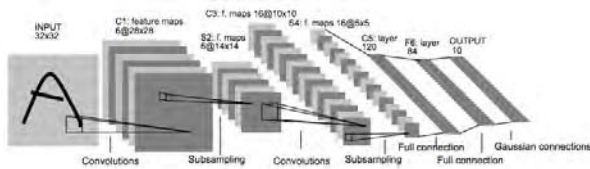
3. 実証分析

Deep Learning の利点は人間が気付かないような特徴も分析することが出来て結果にあらわれるということ。思考部分であり、今回の研究メインである。

3-1 畳み込みニューラルネットワークでのチャート
チャートを見て、視覚情報をもとに売買するトレードを模するために、畳み込みニューラルネットワークモデル(CNN)を採用した。

CNN は視覚野の構造を取り入れた多層パーセプトロンモデルであり、今回の分析目的に最も適していると考えられる。

局所領域（受容野）の畳み込みとプーリングを繰り返す多層ネットワークであり、段階的にスケールを変えながら、局所的な相関パターンを抽出することが可能である。



3-1-1 各種ネットワークでの実行結果

Chainer-imagenet というカラー画像を分類するプログラムがあるが、優秀な各種ニューラルネットワークが定義されており入れ替えて利用できる。

画像入力数も何種類か実施したが、10 年分 30 万画像に関してはボリュームがありすぎて、現状の構成ではメモリエラーがでたり、loss の値が収束しなかったりなどで、あまり多すぎる画像は控えるようにした。

・実施結果表

	Net	入力数	Epoch	収益 (PIP)	取引回数	勝率
1	NIN	32000	10	112	52	52
2	GoogLeNet	300000	10	15	2	100
3	Alex	300000	10	1300.50	5764	50
4	Alex	2015 年データ 24194	70	1350	5098	52.5

4 に関しては accuracy（学習率）97%に到達

いずれのケースも負け越すことは無いので多少の特徴分析は行えているようだが、明確に表れてこないためこの後のフェーズでは全体の検証と工夫を行った。

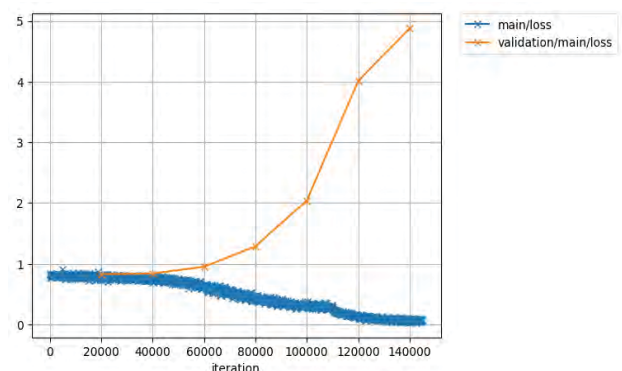
3-1-2 各年度での検証結果

2006 年中の画像を入力し、2007 年のデータをテストサンプルとしてモデルを学習した。2008 年から 2016 年の期間のデータでバックテストを実行した。

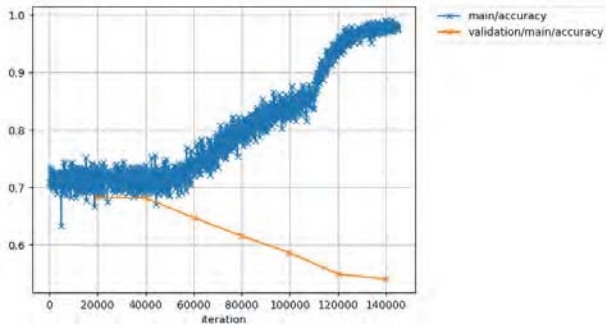
検証年度	取引数	勝率	損益 (PIPS)
2008 年	7835	52.1%	-2.4
2009 年	7726	52.4%	-4.84
2010 年	7690	53.7%	3.659
2011 年	7581	51.0%	-0.836
2012 年	7817	50.1%	-3.194
2013 年	7675	49.7%	-9.862
2014 年	7555	49.6%	-10.802
2015 年	7544	50.7%	5.107
2016 年	6710	49.1%	-14.271
合計	68133	51.0%	-37.439

勝率 50%ほどの結果になっている。また負け越している年度もあり有効なトレードの結果ではないと思われる。勝率が 51%なので、誤差の範囲かもしれないが、少しは特徴の分析が出来たと思われる。

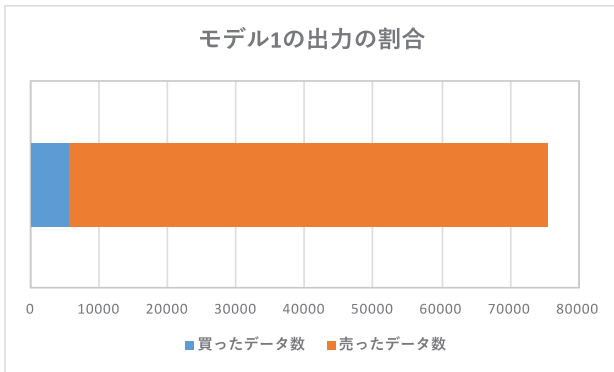
・損失関数の返却値推移



・ 正答率の推移



モデル1の出力の割合



・ 値上がりしているデータと値下りしているデータの割合はほぼ 1:1 にも関わらず、
モデルの出力に大きな偏りが見られる。

・ 買いデータの特徴が上手く抽出されていないと考えられる。

3-1-3 ラベリング変更と畳み込み対応

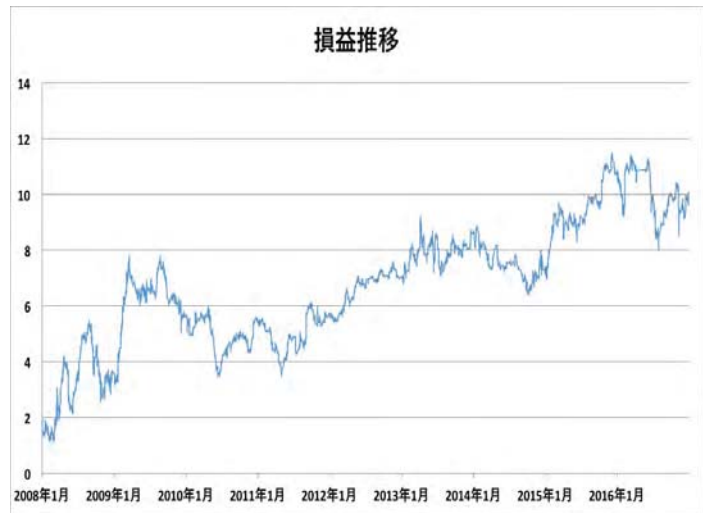
今の画像とラベリングでは違いをもたらす違いが無かったと仮定した場合、準備時に 5PIP 以上の差をラベリングしたが、それを倍の 10PIP に変更した場合どのようなことになるのか検証した。

結果としては、10PIPS の値幅画像での検証は勝率 50.1%で損益合計の推移も順調に増えている。もう少し勝率が上がればもしかしたら実用化も視野に入ると考えた。

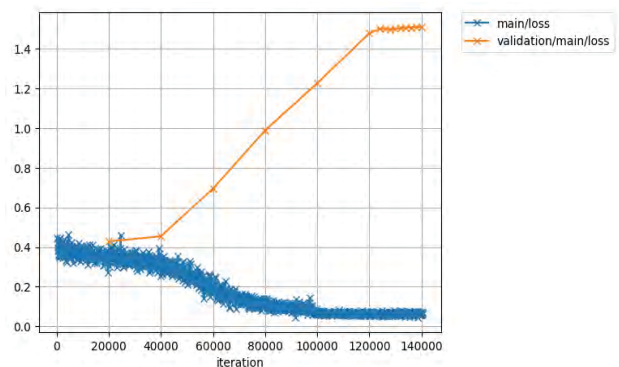
実施結果表

検証年度	取引数	勝率	損益
2008 年	680	51.5%	1.55
2009 年	646	49.5%	2.02
2010 年	593	50.1%	-0.20
2011 年	706	50.7%	0.28
2012 年	666	52.3%	1.31
2013 年	619	52.0%	1.70
2014 年	600	46.3%	-1.56
2015 年	613	48.5%	3.53
2016 年	611	49.3%	-1.05
合計	5734	50.1%	7.59

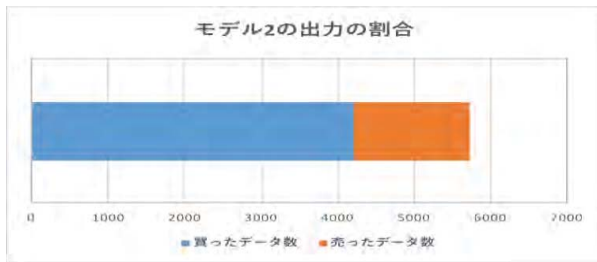
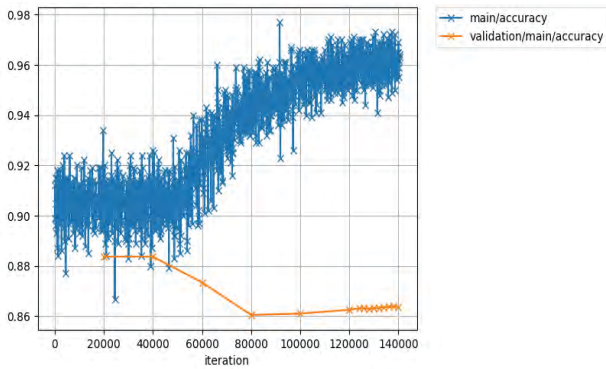
損益推移



・ 損失関数の返却値推移



・ 正答率の推移



出力の偏りはある程度改善されたが、理想の出力の割合は 1:1 のため、まだ改善の余地が見込まれる。より性能の高い GPU を用意し、学習データ数と中間層のフィルタ枚数を増やす必要がある。

3-2 ショートカット回路

ショートカット回路に入力した画像は以下のダブルトップ画像で、各種パラメータ設定を行う事である程度の精度で似ている画像の抽出が行えた。

テンプレート画像



選択後の様子



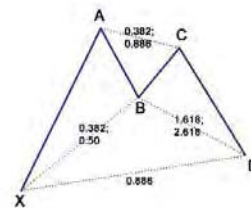
3-3. MT4 のハーモニックパターン検出 EA

ハーモニックパターンの種類 “三尊”・“ダブルトップ”・“フラッグ”などと近い、“ガートレー”、“バット”と呼ばれる各種パターンが成立したところからの反転をトレードする。

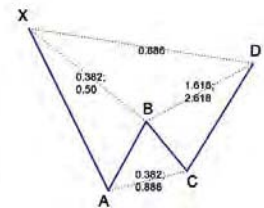
例 バットパターン

The Bat Pattern

Bullish



Bearish



メタトレーダーを利用した EA でハーモニックパターンが成立した時刻がわかるものがあったのだが、私の方でうまく成績がだせず、学習素材として利用できなかったのも、利用できるものを探して利用する。今回のメインテーマではないので、こちらはここまでとしたい。

4. 結論

人間の仕組みを模した人工知能を利用し、Deep Learningの特徴分析能力を利用することによりAIトレーダーは構築できるかどうかを検証した。

1. 前頭葉を模した特徴分析

勝率 53% 勝ち PIP は少ないがある程度、特徴は分析できたと考えている。

過去の研究会では学者をツールとして使ってくださいというお話もあり、お言葉に甘えて本研究会に出席されている、諸先生方のお知恵をお借り出来れば、もっと素晴らしいものになると確信している。

2. 上丘（前意識）を模した判定

VC++で開発することにより、高速動作し、ぱっと見似ているという判定までは行う事が出来た。

3. チャートの形による学習

ハーモニックパターンの各分類でのトレード実行結果が芳しくないので、もう少し相応しい学習素材をさがし学習させることでAIトレーダーに利益が出るトレードを学習させていきたい。

4. 熟練行程でのフィードバック

1で解析した中間層の画像を2の入力判定として利用することにより素早く判定できる仕組みとできることが分かった。

現在の技術でも実際に目で見て考えるような、AIトレーダーは構築出来るのではないかと考えております。また別の分野への転用の幅が広い技術だと思います。

謝辞

東工大の学生の皆様、色々教えていただきありがとうございました。またSIG-FINの先生方にあたたかいお言葉を頂き、私のような至らぬものがこのような研究を行えたこと、非常に誇りに思っています。

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
- [2] 新納 浩幸 Chainerによる実践深層学習
- [3] ラリーペサント フィボナッチ逆張り売買法
- [4] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-

Based Learning Applied to Document Recognition,
Proceedings of the IEEE, 86(11):2278-2324, 1998.

[5] 株お得意ベストセレクション 普遊舎

深層生成モデルによる時系列ネットワークの低次元埋め込み

Embedding time-dependent networks into low-dimensional space using deep generative models

円道 滉一郎^{1*} 江口 浩二¹ 羽森 茂之² 金京 拓司²
Koichiro Endo¹ Koji Eguchi¹ Shigeyuki Hamori² Takuji Kinkyō²

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

² 神戸大学大学院経済学研究科

² Graduate School of Economics, Kobe University

Abstract: Network embedding is one of the approaches to effectively analyzing the network data. Almost all the existing network embedding methods adopt *shallow* models without having deep architecture that is commonly used in deep learning studies. However, shallow models cannot capture highly non-linear network structures that are often observed in real-world, complex networks. To solve this problem, Structural Deep Network Embedding (SDNE) was proposed as a *deep* model for network embedding. In this paper, we focus on Generative Stochastic Network (GSN) for network embedding, in an extension of Autoencoder. GSN robustly captures latent features of data by adding random noises in the process of learning. The framework to capture the latent structure of network is similar to that of SDNE. As a target network in this study, we focus on the time-dependent networks. In order to address the dependency between time intervals and to capture the tendency of previous time interval, we propose *time-dependent pretraining* that uses the parameters learned from the previous time interval as initial states of the current time interval while in the learning process. In the experiments, we use time-dependent financial network data, where each node (or vertex) represents a bank and each link (or directed edge) represents a per-month transaction between a pair of banks, resulting in a series of per-month networks.

1 はじめに

現実世界には、人間関係、企業間関係などといったノードとリンクで構成されるネットワークで表現できる関係データが数多く存在し、ネットワーク表現を分析することで、コミュニティの検出や未観測の関係性の有無の予測などを行うことができる。その分析における有効的な方法の1つとして、ネットワークを低次元潜在空間に埋め込むアプローチ（以下、「ネットワーク埋め込み」と呼ぶ）がある。今日まで様々なネットワーク埋め込み手法が提案されてきたが、そのほとんどは浅い構造を持つモデルであり、複雑である潜在的なネットワーク構造を十分に捉えることが困難であった。そのような中、自己符号化器 [1] と呼ばれるニューラルネットワークを用いた次元圧縮を行うモデルを深い構造にした深層自己符号化器 [1] に基づく構造的深層ネットワーク埋め込

み (Structural Deep Network Embedding: SDNE)[2] と呼ばれるモデルが提案された。

本論文では、以上に述べた動向を踏まえた上で、時系列のネットワークに着目する。とりわけ、ネットワークの時間変化を分析する際に時間的依存性を反映させるようなネットワーク埋め込み手法について提案する。時系列のネットワークの中でも、銀行間の取引を表した金融データに着目し、月ごとの取引をネットワークで表現した時系列のネットワークデータを用いて分析を行った。そして、ネットワークの構造が保存されているかどうかの検証と過去の取引の動向を利用することによる未知の関係性の予測を行った。また、ネットワークのノード間関係性は数量的に表すこともできる。そこで、本論文では、金融データを月ごとの取引金額を重みとした重み付きネットワークでも表現した。そして、手法により重みの大小関係を捉えられているかの検証を行い、未観測リンクに関する重みの大小関係の予測を行った。

*連絡先：神戸大学大学院システム情報学研究科
〒657-0013 兵庫県神戸市灘区六甲台町 1-1
E-mail: endo@cs25.scitec.kobe-u.ac.jp

2 関連研究

ネットワーク埋め込みにおいて、ネットワークの複雑な構造をどのように捉えるかという方法を考えるのは重要な課題である。ネットワークの複雑な構造を捉えるモデルとして、構造的深層ネットワーク埋め込み (Structural Deep Network Embedding: SDNE) というモデルが提案され、良い性能であったことが文献 [2] に述べられている。本研究でネットワーク埋め込みを行う際に参考にした SDNE を紹介するにあたり、まずその基本となる自己符号化器について述べる。次にネットワーク埋め込みの際に考慮する問題点を述べてから SDNE の詳細について述べる。本研究におけるネットワーク埋め込みにおいても自己符号化器に着目しており、ネットワーク埋め込みに使用した自己符号化器の拡張モデルも加えて紹介する。

2.1 自己符号化器

自己符号化器 (Autoencoder: AE)[1] はニューラルネットワークを用いて次元圧縮を行い、データの特徴表現を得るためのモデルである。モデルの構成は入力層、中間層、出力層の 3 層構造のニューラルネットワークを成しており、入力層と出力層の次元数を同じにして、出力されたデータが入力されたデータを復元するように学習させるものである。自己符号化器の構造の例を図 1 に示す。

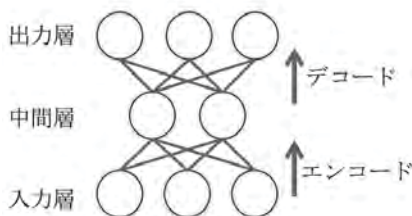


図 1: 自己符号化器の構造。

中間層の次元数は恒等写像の学習にならないように、データの次元数より小さくする必要がある。入力層 \mathbf{X} に対して、中間層 \mathbf{Y} は $\mathbf{Y} = f_{\theta}(\mathbf{X}) = \phi(\mathbf{W}\mathbf{X} + \mathbf{b})$ と定義され、出力層 \mathbf{X}' は $\mathbf{X}' = f'_{\theta'}(\mathbf{Y}) = \phi'(\mathbf{W}'\mathbf{Y} + \mathbf{b}')$ と定義される。ここで、パラメータ θ, θ' はそれぞれ $\theta = (\mathbf{W}, \mathbf{b}), \theta' = (\mathbf{W}', \mathbf{b}')$ であり、 \mathbf{W}, \mathbf{W}' は重み、 \mathbf{b}, \mathbf{b}' はバイアスである。また、 ϕ, ϕ' は活性化関数と呼ばれ、活性化関数にはシグモイド関数やランプ関数 (ReLU 関数) などが用いられる。活性化関数とパラメータを用いて入力層から中間層の表現を得ることをエンコード、中間層の表現から出力層の表現を得ることをデコードと呼ぶ。 \mathbf{W}' に \mathbf{W} の転置行列を使う考え方があり、そ

の考え方は "tied weight" と呼ばれるが、"tied weight" の場合は学習で考慮するパラメータが実質 $\mathbf{W}, \mathbf{b}, \mathbf{b}'$ となる。

このように定義された出力層 \mathbf{X}' を入力層 \mathbf{X} に近づけるようにパラメータを学習する。具体的には、損失関数を L とすると、 $L(\mathbf{X}, \mathbf{X}')$ の値の最小化することが学習となる。損失関数には 2 乗誤差や交差エントロピーなどがよく用いられ、損失関数の最小化には確率勾配降下法 [4] などが用いられる。

学習が終了し、得られたパラメータを用いて表される中間層の表現が次元圧縮された入力データの特徴表現である。

2.2 生成的確率ネットワーク

自己符号化器の一種に雑音除去自己符号化器 (Denoising Autoencoder: DAE)[5] というモデルがある。基本的な学習の考え方は自己符号化器と同様であるが、違う点は、自己符号化器が $L(\mathbf{X}, \mathbf{X}')$ の最小化を考えるに対して、DAE は入力データ \mathbf{X} にノイズを加えた表現 $\tilde{\mathbf{X}}$ にエンコードとデコードを行い得られた表現 $\tilde{\mathbf{X}}'$ を用いて、 $L(\mathbf{X}, \tilde{\mathbf{X}}')$ の値の最小化を考えることである。そして、DAE を一般的な確率モデルとして捉え直し、以下に示した \mathbf{X} と $\tilde{\mathbf{X}}$ を交互に生成するマルコフ連鎖を考える。図示すると図 2 のようになる。

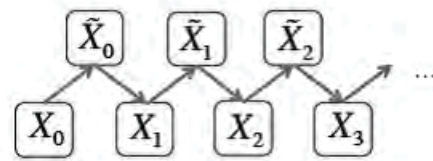


図 2: 雑音除去自己符号化器のマルコフ連鎖。

このマルコフ連鎖において、 \mathbf{X}_0 が入力データ \mathbf{X} に対応し、 \mathbf{X}_1 が入力データ \mathbf{X} を再構築したのに対応する。マルコフ連鎖において、サンプリングを繰り返していくと誤ったモードを推定してしまう可能性がある。それを回避するため、マルコフ連鎖のサンプリングの過程で得られた $\tilde{\mathbf{X}}_t$ も訓練データとして用いることによって誤ったモードを元のモードに引き戻すことを行う。これを Walkback アルゴリズムと呼ぶ。この Walkback アルゴリズムを適用した DAE[6] も提案されている。そして、Walkback アルゴリズムを適用した DAE を拡張したモデルに生成的確率ネットワーク (Generative Stochastic Network: GSN)[7] がある。より柔軟で、複雑な構造を持つことができるように DAE に潜在層を加えたものが GSN である。例として潜在層を 2 層持つ GSN のマルコフ連鎖を図 3 で示す。

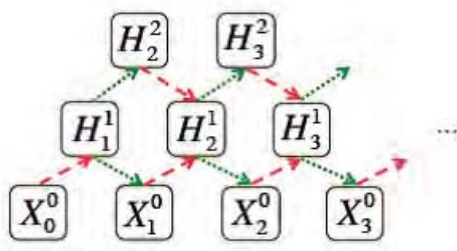


図 3: 生成的確率ネットワークのマルコフ連鎖の例.

GSN の入力層の表現は 1 層目の潜在層によって更新され、最上層以外の潜在層は上下の 2 層によって更新され、最上層は 1 つ下の潜在層によって更新される。上図の各層の表現は次のように示される。

$$\mathbf{X}_t^0 = \phi(\mathbf{W}_1^T \mathbf{H}_t^1 + \mathbf{b}_0) \quad (1)$$

$$\mathbf{H}_t^1 = \phi(\mathbf{W}_1 \mathbf{X}_{t-1}^0 + \mathbf{W}_2^T \mathbf{H}_t^2 + \mathbf{b}_1) \quad (2)$$

$$\mathbf{H}_t^2 = \phi(\mathbf{W}_2 \mathbf{H}_{t-1}^1 + \mathbf{b}_2) \quad (3)$$

ここで、 $\mathbf{W}_1, \mathbf{W}_2$ は重み、 $\mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2$ はバイアスである。GSN の学習では Walkback アルゴリズムが適用され、サンプリング回数が T 回の時のモデルの構造は図のマルコフ連鎖を T 回で打ち切ったものとなる。DAE 同様、データにノイズを加えた表現を GSN に入力し、再構築されたものが元のデータに近づくようにパラメータを学習する。GSN の学習では奇数目の層と偶数目の層を交互にサンプリングして更新する。そして、再構築された表現にまたノイズを加える。また、潜在層の表現は活性化関数に代入する前後にノイズを加える [8]。本研究では、GSN をネットワーク埋め込みに適用する。

2.3 ネットワーク埋め込みに関する問題定義

ネットワーク表現を学習するにおいて、3つの問題点がある。それは、高度非線形性、構造の保存、スパース性である。まず、高度非線形性についてであるが、高度非線形なネットワーク構造をよく捉えるために、Structural Deep Network Embedding(SDNE)[2]と呼ばれるディープラーニングの手法を用いた深層モデルが提案された。本研究では、ネットワーク埋め込みを行う際に SDNE の考え方を採用している。SDNE の詳細については、次節で述べる。そして、構造の保存やスパース性の問題に対処するため、1次近接性と2次近接性をネットワーク埋め込みを行う際に取り入れる必要がある。1次近接性は局所的な構造に着目したもので、ノード間にリンクを持つならば、それらのノードは似ているという考え方を仮定している。ある論文はその論文が引用している論文といくつかの共通した話題が含まれている可能性があるというようなことが例

として挙げられる。2次近接性は全域的な構造に着目したもので、同じような隣接構造をもつノード同士は似ているという考え方を仮定している。多くの共通の友達がいる人同士は友達である可能性があるといったようなことが例として挙げられる。現実世界に存在するネットワークはリンクが欠損している場合もあり、2次近接性はリンクが欠損していたとしてもノードの近接性を捉えることができるので、ノードの関係性をより豊富な情報を用いて捉えることができる。これらの近接性を考慮することにより、局所的にも全域的にもネットワークの構造を捉え、多い情報を用いてネットワークを特徴づけることができるので、ネットワークの構造の保存やスパース性の問題を緩和できる。1次近接性はノード間のリンクの有無のみに着目しているため、ネットワークが無向グラフの場合にのみ有効であることに注意が必要である。

2.4 構造的深層ネットワーク埋め込み

本節では、構造的深層ネットワーク埋め込み (Structural Deep Network Embedding: SDNE)[2] について述べる。SDNE は自己符号化器を深層にした深層自己符号化器に基づき、枠組みは図4のようになる。

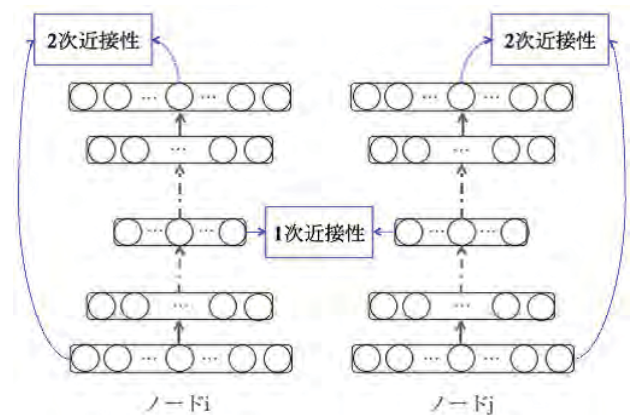


図 4: 構造的深層ネットワーク埋め込みの枠組み.

ネットワークデータは隣接行列で表現され、それを入力とする。隣接行列の i 行 j 列要素を $s_{i,j}$ 、ネットワークデータのノード数を N 、SDNE における入力層の表現を \mathbf{X} 、出力層の表現を $\tilde{\mathbf{X}}$ 、最も低次元な層の表現を \mathbf{Y} とする。まず、1次近接性からノード間にリンクを持つならば \mathbf{Y} も似たようなと考えられ、1次近接性の損失関数 L_{1st} は以下のように示される。

$$L_{1st} = \sum_{i,j=1}^N s_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \quad (4)$$

次に2次近接性から似た隣接構造を持つノードは似たような再構築表現になると考えられ、2次近接性の損失関数 L_{2nd} は以下のように示される。

$$L_{2nd} = \|(\tilde{\mathbf{X}} - \mathbf{X}) \odot \mathbf{B}\|_F^2 \quad (5)$$

ここで、 \odot はアダマール積であり、 $\|\cdot\|_F$ はフロベニウスノルムである。アダマール積とは同じサイズの行列に対して行列の要素ごとに積を取ることで定まる演算であり、フロベニウスノルムは行列 A の i 行 j 列要素を $a_{i,j}$ とすると、 $\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2}$ で定義される。そして、 \mathbf{B} はバイアスであり、 \mathbf{B} によって、ネットワークのスパース性に対処する。 \mathbf{B} の i 行 j 列要素を $b_{i,j}$ とすると、もし $s_{i,j} = 0$ ならば $b_{i,j} = 1$ 、そうでなければ $b_{i,j} = \beta > 1$ とする。ここで、 β はパラメータである。再構築誤差を最小化することで、再構築の手順で似た隣接構造を持つノードは似た潜在表現を持つようになる。

以上の2つの損失関数を組み合わせ、さらに過学習を抑制するための正則化項 L_{reg} を加えたものを、SDNEの目的関数とする。目的関数 L は次のように表現される。

$$L = \alpha L_{1st} + L_{2nd} + \gamma L_{reg} \quad (6)$$

ここで、 α, γ はパラメータである。目的関数を最小化することにより、1次近接性と2次近接性を考慮したネットワーク埋め込みを行うことができる。尚、1次近接性は無向グラフの場合のみ有効であるので、有向グラフの場合はパラメータ α を $\alpha = 0$ とする必要がある。

3 時系列ネットワーク埋め込み

本研究では、構造的深層ネットワーク埋め込み (Structural Deep Network Embedding: SDNE) が深層自己符号化器に基づくことに着目し、同じ自己符号化器の系統の生成的確率ネットワーク (Generative Stochastic Network: GSN) をネットワーク埋め込みに用いることを考える。尚、GSNでネットワーク埋め込みを行う際、ネットワークデータの構造を捉える枠組みはSDNEと同様のものを使用する。本章では、まずGSNを用いたネットワーク埋め込みに関する詳細について述べ、次に、GSNで時系列ネットワークデータの埋め込みをどのように行うかを提案する。

3.1 生成的確率ネットワークにおけるネットワークデータの適用

SDNEのネットワーク埋め込みの枠組みを参考にし、GSNでは最上層で1次近接性、最下層で2次近接性を考慮する。ネットワーク埋め込みにおけるGSNの1次

近接性と2次近接性を捉えた目的関数は次の式で表すことができる。尚、ネットワークデータのノード数を N 、隣接行列を $S = \{s_{i,j}\}_{i,j=1}^N$ 、GSNの潜在層の数を K 、入力層の表現を $\mathbf{X}_t^k = \{\mathbf{x}_{i,t}^k\}_{i=1}^N$ 、最上層の表現を $\mathbf{H}_t^K = \{\mathbf{h}_{i,t}^K\}_{i=1}^N$ 、Walkback回数を T 回とする。

$$\begin{aligned} L &= \frac{\alpha}{T-K+1} \sum_{t=K}^T (L_{1st})_t + \frac{1}{T} \sum_{t=1}^T (L_{2nd})_t + \gamma L_{reg} \\ &= \frac{\alpha}{T-K+1} \sum_{t=K}^T \left(\sum_{i,j=1}^N s_{i,j} \|\mathbf{h}_{i,t}^K - \mathbf{h}_{j,t}^K\|_2^2 \right) \\ &\quad + \frac{1}{T} \sum_{t=1}^T \|(\tilde{\mathbf{X}}_t^0 - \mathbf{X}_t^0) \odot \mathbf{B}\|_F^2 + \gamma L_{reg} \end{aligned} \quad (7)$$

SDNEと同様、 α, γ はパラメータであり、 $(L_{1st})_t, (L_{2nd})_t, L_{reg}$ は、それぞれWalkbackが t 回目の1次近接性の損失関数、Walkbackが t 回目の2次近接性の損失関数、重みを用いた正則化項である。そして、 \mathbf{B} はバイアスであり、 \mathbf{B} の i 行 j 列要素を $b_{i,j}$ とすると、もし $s_{i,j} = 0$ ならば $b_{i,j} = 1$ 、そうでなければ $b_{i,j} = \beta > 1$ とする。しかし、本研究において時系列ネットワークデータを扱う上での例外として、もしある時点において $\mathbf{s}_i = \{s_{i,j}\}_{j=1}^N = \mathbf{0}$ である。すなわち、どのノードともリンクを持たないならば、ノード i を欠損扱いとし、 $\mathbf{b}_i = \{b_{i,j}\}_{j=1}^N = \mathbf{0}$ とする。このように工夫することで学習の際に欠損ノードが直接影響を及ぼすことがなくなる。ネットワークが無向グラフの時はパラメータ α は $\alpha = 0$ となる。GSNにおけるネットワーク埋め込みは、ネットワークデータを隣接行列で表現したものを入力とし、上式の最小化を行うことによってネットワークの潜在構造を捉え、保存することを目的とする。

3.2 生成的確率ネットワークにおける時系列ネットワークデータの低次元埋め込み

本節では、前節で述べたネットワーク埋め込みを行うGSNを用いて、時系列のネットワーク埋め込みを考える。本研究では、ネットワークデータの傾向や特徴を捉えるためのパラメータに着目し、時系列を考慮する。尚、ここで述べているパラメータとは、層間の関係を表す重みと各層のバイアスのことを示している。そして、本研究では、事前に学習したパラメータを初期値とするという点に着目した。過去のデータの学習によって得られたパラメータを現在のデータの学習における初期値とする、ある種の事前学習 (プレトレーニング) を行うことで時区間の間の依存性を捉え、過去の動向を踏まえた学習ができるのではないかと考え、時系列ネットワーク埋め込み手法として提案する。

例として、本研究で着目する時系列ネットワークデータである金融データを用いる場合を考える。1 か月間の企業間のやり取りの関係を示す時系列ネットワークデータに時系列プレトレーニングを適用する時、当月のデータを用いて学習する際にパラメータの初期値を先月のデータを用いて学習して得た最適パラメータとする。そして、翌月のデータに対して学習を行う際に、先月の学習で得たパラメータを初期値とした当月の学習によって得られたパラメータを初期値とすれば、先月と当月の傾向をある程度考慮した状態で翌月のデータを学習することができる可能性が考えられる。

4 実験

時系列ネットワーク埋め込みを GSN で行い、性能評価実験を行う。本研究では、時系列ネットワークを重みなしと重み付きの2種類の隣接行列で表したデータとして使用する。この章において、時系列で過去のデータの傾向を考慮し学習する事を時系列プレトレーニングと呼ぶ。

4.1 データセット

本研究では時系列ネットワークの中でも金融データを使用する。実験で用いた金融データのデータセットはリーマンショックが起こった2008年の銀行間の取引のデータである。欧州15か国の銀行間の取引を1年分集めたデータであり、銀行は全部で177行、取引数は93067である。1つの取引において実験で着目する情報は、取引の契約日、取引を持ちかけられる側の銀行名(quoter)、持ちかける側の銀行名(aggessor)、最終的な取引金額、取引の状態である。最終的な取引金額は百万ユーロ単位で示されており、取引の状態は”Sell”か”Buy”の2つで示されている。そして、この1年分の銀行間の取引関係のネットワークデータを有向グラフと捉え、 177×177 の重みなし隣接行列と重み付き隣接行列として月ごとに表現する。ここで、隣接行列を S とする時、 S の i 行 j 列の要素を $s_{i,j}$ で表すとする。有向グラフの向きは取引の状態(”Sell” or ”Buy”) で決めており、取引を持ちかけたかどうかで向きを決めていないことに注意が必要である。 S をある月の重みなし隣接行列と考える場合、取引の有無で隣接行列を作成するため、 $s_{i,j}$ は0か1の値を持つ。そして、 S をある月の重み付き隣接行列と考える場合、最終的に成立した取引額がいくらであったかという数値を契約が成立した取引の規模と考え、その数値で隣接行列を作成するため、 $s_{i,j}$ は月ごとの取引の規模の合計の値となる。重みなしと重み付き隣接行列のリンク密度の月ごとの平均は約0.0585である。元の隣接行列のリンク

数の月ごとの平均は1833.25であり、欠損ノード数の月ごとの平均は25.25である。ここで、欠損ノード数は、どのノードともリンクを持たないノードの数であり、GSNの学習において、欠損ノードは目的関数に影響を与えないよう工夫している。

4.2 重みなし隣接行列を用いた実験

この節では、重みなし隣接行列を GSN の入力とした時の性能を隣接行列の再構築と未観測のリンク予測というタスクで評価する。実験を行う際、元の隣接行列のリンクをランダムに15%除いた隣接行列を使用する。Walkback 回数 T は $T = 5$ 、 $(\alpha, \beta, \gamma) = (0, 10, 10^{-6})$ とする。これらのパラメータの値は、本実験で使用するデータを扱う際に適切になるよう予備実験により定めた。そして、本実験で使用する GSN は入力層が177次元、1層目の潜在層が100次元、2層目の潜在層が50次元の3層構造である。

4.2.1 隣接行列の再構築に関する評価

再構築の評価はネットワーク埋め込みにおいて最も基本的な評価である。最後の Walkback で再構築された隣接行列の各要素の値に着目し、リンクを持つ可能性の高い順番でランキングを行い、実際のリンクをどの程度適合させることが平均適合率 (Average Precision) で評価する。評価には時系列プレトレーニングの有無による月ごと平均適合率の差を図5で示し、性能を比較する。1月より過去のデータが存在しないため、1月に関しては時系列プレトレーニングが行うことができず、時系列プレトレーニングの有無によって数値に変動がないことに注意が必要である。

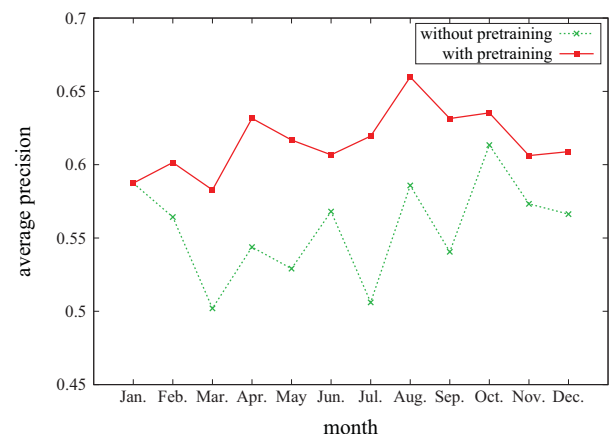


図5: 重みなし隣接行列再構築における平均適合率。

月ごとの平均適合率に着目すると、すべての月において時系列プレトレーニングを行った方が性能は良く、改善率(パーセント)の月ごとの平均が11.9%で、その標準標準偏差が5.8%であった(以下、月ごとの性能に着目した改善率を表記する時は「 $11.9 \pm 5.8\%$ 」のようにする)。また、月ごとの改善率 I は評価値 E から E' になった時、 $I = E'/E \times 100 - 100$ と定義される。

4.2.2 未観測リンクの予測に関する評価

リンク予測は過去の傾向から現在の観測されていないリンクをどの程度予測できるかを評価するものである。あらかじめ除いた15%のリンクを未観測データ、残りのリンクを観測データとする。そして、先月のパラメータを用いて当月のデータを再構築し、再構築された隣接行列において、観測されているリンクを持つノード対を除外して残りのノード対においてリンクを持つ可能性の高い順番でランキングを行い、未観測データのリンクをどれだけ捉えられるかを評価する。スパース性が大きく増すので、問題としての難易度は高くなる。リンク予測における時系列プレトレーニングは先々月のパラメータを初期値として学習した先月のパラメータを用いて当月を予測することである。時系列プレトレーニングの有無による月ごと平均適合率の差を図6で示し、性能を比較する。1月より過去のデータが存在しないため、2月に関しては時系列プレトレーニングが行うことができず、時系列プレトレーニングの有無によって数値に変動がないことに注意が必要である。

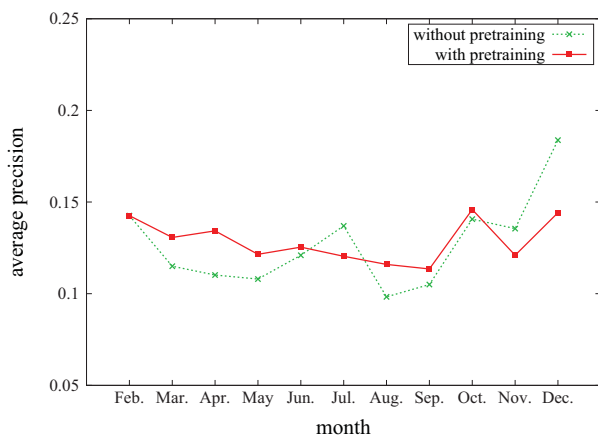


図 6: 重みなし隣接行列のリンク予測における平均適合率。

月ごとの平均適合率に着目すると、7月、11月、12月以外は時系列プレトレーニングを行った方が平均適合率の値は良く、時系列プレトレーニングの有効性が部分的に確認された。

4.3 重み付き隣接行列を用いた実験

この節では、重み付き隣接行列を GSN の入力とした時の性能を隣接行列の再構築と未観測の重みの予測というタスクで評価する。2次近接性を捉える際に関与するパラメータ β は $\beta = 40$ と設定している。そして、重みなし隣接行列を用いた時と同様に、入力層が177次元、1層目の潜在層が100次元、2層目の潜在層が50次元の3層構造の GSN を使用する。

4.3.1 隣接行列の再構築に関する評価

重み付き隣接行列を用いた時のネットワーク埋め込みの基本的な評価として、隣接行列の再構築を検討する。重み付き隣接行列の再構築では、リンクを持つノード対のみに着目する。そして、重みの重い順にノード対をランキングし、そのランキングが GSN の学習によって再構築されているかをケンドール順位相関係数を用いて評価する。ケンドール順位相関係数 τ は $|\tau| \leq 1$ であり、分布などを仮定せず、ランキング同士の相関度合いを測る評価指標である。このタスクにより取引の規模の大小関係を捉えているかを判断する。結果を月ごとに図7で示す。1月より過去のデータが存在しないため、1月に関しては時系列プレトレーニングが行うことができず、時系列プレトレーニングの有無によって数値に変動がないことに注意が必要である。

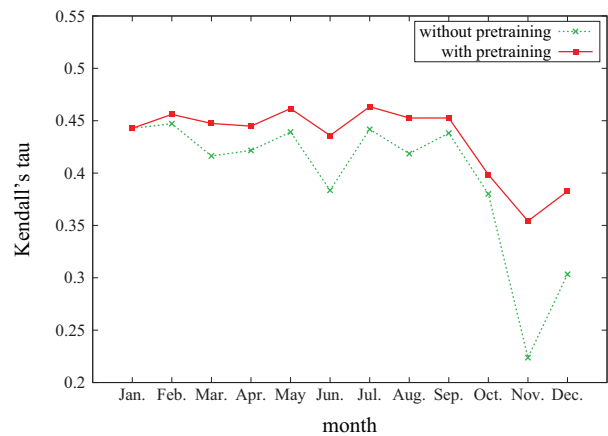


図 7: 重み付き隣接行列再構築におけるケンドール順位相関係数。

ここで、「ケンドール順位相関係数が0である(無相関である)」という仮説を帰無仮説とした有意確率を求めた時、いずれの月でも有意水準1%よりも下回り、帰無仮説が棄却されることが確認された。したがって、いずれの月についても正解となる未観測の重みのランキングと再構築された隣接行列に基づく未観測の重みのランキングに統計的に有意な相関があると言うことが

でき、提案手法の有効性を示すことができた。時系列プレトレーニングを行った時のケンドール順位相関係数の改善率は約 $12.7 \pm 15.8\%$ であり、時系列プレトレーニングの有効性が確認できた。ただし、いずれの月も $2.0 \sim 58.3\%$ の正の改善率を示しており、標本標準偏差が大きいのは改善の程度が月によって幅があることを意味することに注意が必要である。

4.3.2 未観測重みの予測に関する評価

未観測の重みの予測は過去の傾向から現在の未観測リンクの重みの大小関係をどの程度予測できるかを評価するものである。あらかじめ除いた 15% のリンクを未観測データ、残りのリンクを観測データとする。まず、過去のデータを用いて学習されたパラメータを用いて現在の観測データを再構築する。そして、未観測データでリンクを持つノード対のみに着目し、観測データを再構築して得られた隣接行列の各要素から重みのランキングを行う。そして、そのランキングが未観測データのリンクの重みのランキングとどの程度相関があるかをケンドール順位相関係数を用いて評価する。時系列プレトレーニングを行う時は、先々月までの傾向を考慮し学習されたパラメータを先月の傾向を学習する際の初期値とし、学習されたパラメータを用いて当月の観測データを再構築したものから当月の未観測データのランキングを予測する。時系列プレトレーニングの有無による月ごとケンドール順位相関係数の変化を図 8 で示す。隣接行列の再構築のタスクの時と同様に、1 月より過去のデータが存在しないため、2 月に関しては時系列プレトレーニングが行うことができず、時系列プレトレーニングの有無によって数値に変動がないことに注意が必要である。

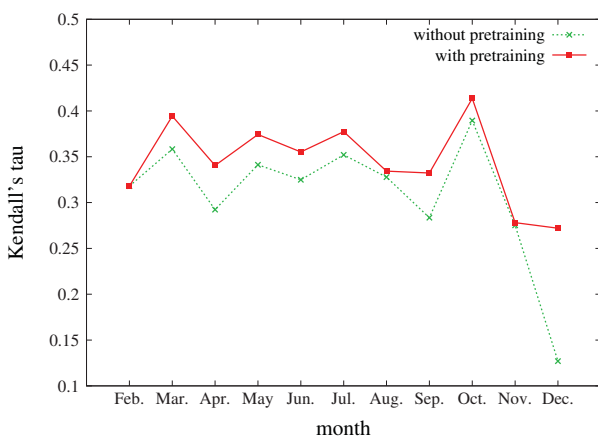


図 8: 重み付き隣接行列の重み予測におけるケンドール順位相関係数。

ここで、隣接行列の再構築の時と同様に「ケンドール順位相関係数が 0 である (無相関である)」という仮説を帰無仮説とした時、いずれの月でも帰無仮説が有意水準 1% の下で棄却されることが確認されたので統計的に有意な相関があるということができ、提案手法の有効性を示すことができた。全ての月において、時系列プレトレーニングを行うことによる性能改善が見られ、ケンドール順位相関係数の改善率は約 $19.3 \pm 32.0\%$ であった。以上の結果より、時系列プレトレーニングの有効性が確認できた。

4.4 考察

この節では、実験の結果について考察する。重みなし隣接行列の再構築では、時系列プレトレーニングを行うことで性能改善を確認でき、時系列プレトレーニングは有効であると言える。リンク予測では、時系列プレトレーニングにより、7 月、11 月、12 月以外は性能の改善が見られた。この 3 つの月で性能の改善が見られなかった原因として、ランダムに除かれた 15% のリンクの中に時系列の傾向を捉える際に重要なリンクが含まれていた可能性があることや取引があった銀行が月ごとに局所的であったことなどが挙げられる。平均適合率の値自体が高くない原因については、観測されたリンクを除くことでスパース性が増し、リンク予測がかなり困難なタスクになったためであると考えられる。重み付き隣接行列の再構築と未観測の重みの予測では、時系列プレトレーニングを行うことでケンドール順位相関係数の値は統計的に有意に大きな値を示していることから、時系列プレトレーニングは有効であると言える。

さらに、重み付き隣接行列は取引金額の値を利用していることから、取引金額と金利の結びつきがあると考え、時系列プレトレーニングを行った時の GSN の最上層の潜在表現に金利と相関があるかどうかを追加的に分析した。まず、銀行間の取引における金利の期待値を月ごとに求め、それを各要素に持つ金利期待値隣接行列を考える。そして、金利期待値隣接行列の列ごとに金利の期待値の平均を取り、それを銀行ごとの金利ラベルとする。この金利ラベルは、各銀行が他行から平均的にどの程度の金利で融資を受けたかを示し、金利ラベルの値が大きいとき、その銀行の経営が悪化していることの指標となり得る。GSN の最上層は 50 次元であり、それが銀行 177 行分だけ存在するので、 177×50 の行列を得ることができる。その行列の列に着目し、列ベクトル (177 次元) が金利ラベルベクトル (177 次元) とどの程度相関があるかを調べる。このようにすることで、50 次元の最上層の中で金利を捉えるようと機能している潜在表現を確認できる可能性がある。相関を

調べる際、ピアソンの相関係数を用いた。ピアソンの相関係数を r とすると、 $|r| \leq 1$ である。そして、金利との相関をピアソンの相関係数を用いて調べたところ、一部の潜在表現に金利との有意な相関が認められた。例として、4月では $r = -0.1959, 0.3298, -0.1840$ となる3つ、9月では $r = -0.2450$ となる1つ、10月では $r = -0.2175, -0.2821$ となる2つの潜在表現が確認された。ここで、いずれの月においても「ピアソンの相関係数が0である(無相関である)」という帰無仮説の下で、有意確率は有意水準5%よりも下回り、帰無仮説が棄却されることが確認されたので、統計的に有意な相関があると言えることができる。上記の通り、金利と相関のある潜在表現の数と基底が月により変動があった。9月にリーマンショックが起こったことから、金利が非定常な動きをしていたことが予想され、潜在表現との相関も他の月と比して有意なものが観察されにくい状況が認められる。潜在表現と金利の関係についてより詳細な分析を行うことは今後の課題の一つである。

5 おわりに

銀行間の取引のデータに対し、重みなし隣接行列における隣接行列の再構築とリンク予測、重み付き隣接行列における隣接行列の再構築と未観測の重みの予測の4つのタスクで評価することで、GSNでネットワーク埋め込みを行うことの有用性と時系列プレトレーニングの有効性について確認することができた。

本研究における今後の課題として、まず、本実験で用いた銀行の取引データに関して、取引の有無や金額だけでなく、取引にかかった期間や取引の頻度に着目してネットワーク表現を獲得し、解析することが挙げられる。さらに、取引における金利まで考慮し、金利の予測を行うことができれば、より解析の幅は広がる。また、ネットワーク埋め込みを行うモデルは様々なので、それらのモデルを用いた時系列ネットワーク埋め込みの性能と比較を行う必要があると考えられる。

謝辞

本研究を行うにあたり、有益な助言、協力を頂いた神戸大学大学院システム情報学研究科谷口隆晴准教授、同大学経済学研究科の田中克幸助教に感謝する。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] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *science* 313.5786 (2006): 504-507.
- [2] Daixin Wang, Peng Cui, Wenwu Zhu. "Structural Deep Network Embedding." *KDD Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(2016)*: 1225-1234.
- [3] Bengio, Yoshua. "Learning deep architectures for AI." *Foundations and trends in Machine Learning* 2.1 (2009): 1-127.
- [4] Ngiam, Jiquan, et al. "On optimization methods for deep learning." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.
- [5] Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.
- [6] Bengio, Yoshua, et al. "Generalized denoising auto-encoders as generative models." *Advances in Neural Information Processing Systems*. 2013.
- [7] Bengio, Yoshua, et al. "Generalized denoising auto-encoders as generative models." *Advances in Neural Information Processing Systems*. 2013.
- [8] Bengio, Yoshua, Nicholas Leonard, and Aaron Courville. "Estimating or propagating gradients through stochastic neurons for conditional computation." *arXiv preprint arXiv:1308.3432* (2013).