

令和 5 年 6 月 22 日現在

機関番号：14301

研究種目：基盤研究(B)（一般）

研究期間：2019～2022

課題番号：19H04116

研究課題名（和文）エビデンスベースの投資支援に向けたエンティティ指向投資ビッグデータ分析基盤の構築

研究課題名（英文）Entity-oriented Investment Big Data Analysis Foundation for Evidence-Based Investment Support

研究代表者

馬 強 (Ma, Qiang)

京都大学・情報学研究科・准教授

研究者番号：30415856

交付決定額（研究期間全体）：（直接経費） 13,300,000円

研究成果の概要（和文）：本研究では、投資活動を支援する基盤技術について研究開発を行った。時系列データ分析の共通基盤では、時系列データを効率よく管理・処理する表現モデルと分析手法を開発した。投資商品の要因分析では、基準価額のトレンドの変化点を干渉要因として導入して動的状態空間モデルベースの要因分析手法を開発した。また、投資の重要な情報源であるニュース記事の経済的な影響力の推定手法も開発した。投資者の特徴分析では、ポートフォリオ理論を導入してソーシャルトレーディングサービスにおける投資者であるトレーダの特徴分析技術の開発を行った。
英文論文誌論文5本、査読付国際会議論文6本と国内研究会発表11本の成果発表を行った。

研究成果の学術的意義や社会的意義

本研究では、ニュース・報告書、マーケット情報やユーザの投資履歴など多様な投資ビッグデータを横断して分析し、「金融・投資商品に影響を及ぼす要因(事象)」や「投資のエキスパートがいつ・何を取引したか」を明らかにする。これにより投資に必要な知識やエビデンスを発見して意思決定のプロセスと結果の可読性を向上させることで、従来サービスにおける一般投資者の不安を緩和し、安心して投資できる仕組みの構築に貢献する。

研究成果の概要（英文）：In this study, we conducted research and development of the basic technology to support investment activities. As the analysis foundation, we developed representation models and analysis methods that efficiently manage and process time series data. For factors analysis, we introduced the trend change point of the benchmark price as an interference factor and developed the factor analysis method based on the dynamic state space model. We also developed a method to estimate the economic impact of news articles, which are an important source of information for investment. For investors analysis, we introduced portfolio theory and developed the feature analysis technology of traders, who are investors in social trading services.
We published four journal articles in English, six reviewed international conference papers, and 11 domestic conference papers.

研究分野：データ工学

キーワード：投資情報学 意思決定支援 要因分析 時系列データ 深層学習 金融工学 ポートフォリオ

様式 C - 19、F - 19 - 1、Z - 19 (共通)

1. 研究開始当初の背景

人工知能(AI)を活用して投資またはその支援を行うシステムは多くあるが、投資行動の理由やエビデンスを示さず、プロセスと結果の可読性が低く、ブラックボックス化されているため、一般利用者の拡大に繋がっていない。

2. 研究の目的

本研究の目的は、エビデンスに基づいて投資の意思決定を支援するための投資ビッグデータ分析基盤の確立である。これにより投資に必要な知識やエビデンスを発見して意思決定のプロセスと結果の可読性を向上させることで、従来サービスにおける一般投資者の不安を緩和し、安心して投資できる仕組みの構築に貢献する。

3. 研究の方法

本研究では、FinTechをはじめ金融・投資サービスによって生み出される多様な投資ビッグデータを対象に、時系列データを効率よく管理・分析する基盤技術、投資商品の基準価額に影響を及ぼす要因分析技術及び投資者の特徴分析技術について研究開発を行う。

4. 研究成果

本研究では、ニュース・報告書、マーケット情報やユーザの投資履歴など多様な投資ビッグデータを横断して分析し、投資活動を支援する基盤技術について研究開発を行い、以下のような成果を得られた。

(1) 時系列データの分析基盤

投資ビッグデータの多くが時系列データであり、スパース性が顕著である。それを克服するためのデータ補完・増殖手法を開発した。また、時系列データを効率よく管理・処理する表現モデルとその効率的に学習できる手法を開発した。

データ補完・増殖手法[1]

本研究では系列の数と長さの観点から時系列データを増殖する手法を開発した。

系列数を増やす増殖手法としてLiuらが提唱しているCrop, Mask, Reorder, Substituteを利用する。Crop, Mask, Reorderはランダムな増殖手法であり、対象となる要素(データアイテム)の性質を考慮しない。一方、Substituteは対象となる要素と別の要素との類似度を用いた増殖手法である。要素の類似度としてはコサイン類似度を用いた。

系列長を増やす増殖手法としてInsert, TenureWeightedInsertの2つのデータ増殖手法を開発した。系列の要素の特徴を考慮しながら系列長を増し、パディング処理の役割も担う手法である。Insert法では、要素間の類似度を用いて系列に要素を挿入する。TenureWeightedInsert法では、個々の要素の時間範囲などから推定される要素の重要度に即して要素を挿入する。

実験結果では増殖手法の有用性を確認した。特に、Insert法とReorder法がよい予測精度を得られた。系列長を増やすInsert法とTenureWeightedInsert法は短い系列の予測に有効であることを確認できた。

時系列データの表現モデルとその学習手法[2,3,4]

表現学習は、時系列データの分析において重要な役割を果たし、データ分析、機械学習モデル、意思決定プロセスの有効性と精度に大きな影響を及ぼす。時系列データを処理するには、時刻 (Timestamp) レベル、セグメントレベルなど粒度を考慮する必要がある。しかしながら、これまでの研究では、様々な粒度レベルのばらつきを考慮していなかったため、異なる情報源からの情報の統合活用などには不十分である。そこで、本研究では、時系列データの多粒度性を考慮した表現モデルとそれを学習するための教師なしフレームワーク(図1)を開発した。具体的には、粒度の細かい表現と粗い表現の間の関連付けを行うために、粒度横断するトランスフォーマーを提案した。また、時系列の多粒度表現を学習するためタスクとして、検索タスクを導入して、教師なしの学習を実現できた。さらに、教師なし学習によって時系列の包括的な多階調表現を得るために、新しい損失関数を設計した。実験の結果、提案されたフレームワークは、他の表現学習モデルに対する優位性が明らかになった。

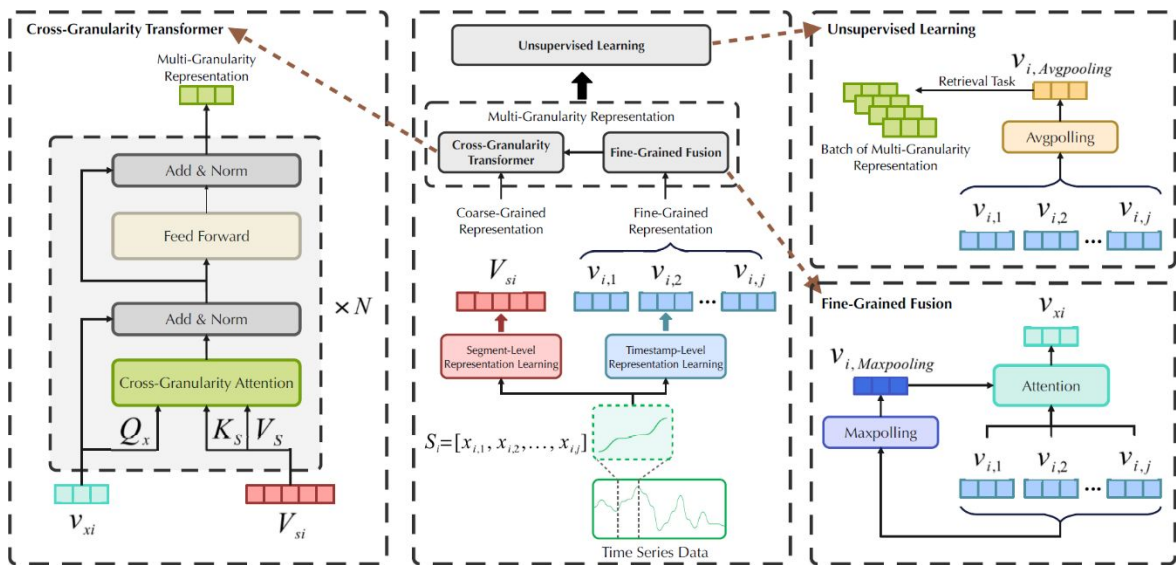


図1 多粒度の時系列データ表現モデルとその学習フレームワーク[4]

(2) 投資商品分析

投資商品に影響を及ぼす要因分析について研究開発を行った。動的状態空間モデルを用いて投資商品に影響を及ぼす要因の分析手法について研究開発を行い、トレンドの変化点を干渉要因として導入したトレンドシフトモデルの開発に成功した。また、投資の重要な情報源であるニュース記事の経済的な影響力の推定手法も開発した。

トレンドシフトモデル[5]

投資を支援するため、市場の変動要因に関する様々な研究が行われている。我々は先行研究で、投資に関するテキスト(運用報告書やニュース)と数量データ(投資商品の基準価額など)を併用して、投資信託商品の要因について定量的に分析手法を提案した。これらの研究では、上昇トレンドや下降トレンドなどの局面で分析を行うことが、投資信託商品の局所的分析の精度向上に有用であることを明らかにしているが、将来の予測などの大局的な分析について十分ではない。

そこで、本研究では基準価額を対象に、トレンドの変化点を動的に検出し、その変化点を考慮したトレンドシフトモデルを用いて各要因の影響度を分析する手法を開発した。トレンドを検知するため、従来のTBSM(Trend Based Segmantation Method)手法を拡張し、パラメータへの

依存を緩和した TP-TBSM(Two Phase TBSM) 法の開発に成功した。TP-TBSM 法は、トレンドにおける誤差を評価して分割するトレンドを選択する評価フェーズと、トレンドをさらに分割するセグメントフェーズを交互に繰り返していくことで、トレンドの変化点を検出するのに必要なパラメータ数を軽減した。

TBSM 法では、あらかじめ設定する 3 つの閾値が検知されるトレンドを大きく左右するため、時系列データに合わせて適切なパラメータを決めることが困難である。そこで、我々はトレンド誤差という概念を導入し、トレンド誤差が大きいトレンドをさらに分割することで、パラメータへの依存を緩和する。トレンド誤差とは、トレンドの始点と終点を結んだトレンド直線(線形関数で表現出来る)と各データ点の距離について足し合わせ、そのトレンドの長さで割ったものであり、トレンド直線から離れる程度を測る尺度である。この誤差が小さいほどトレンドがよい。

さらに、本研究では、要因分析するため、検出されたトレンド変化点を傾きシフト干渉変数として動的状態空間モデルに導入したトレンドシフトモデルを開発した。トレンドシフトモデルは、従来の状態空間モデルにおけるトレンド項よりも急激な変化を表現できる。また、トレンドシフトモデルでは、特定の局面における要因の動きだけではなく、トレンドを横断して要因を大局的に分析しているため、基準価額の予測にも応用可能である。評価実験では、トレンドシフトモデルは従来手法より予測誤差と AIC でよい性能を達成したことを確認した。

ニュース記事の経済的な影響力の推定[6]

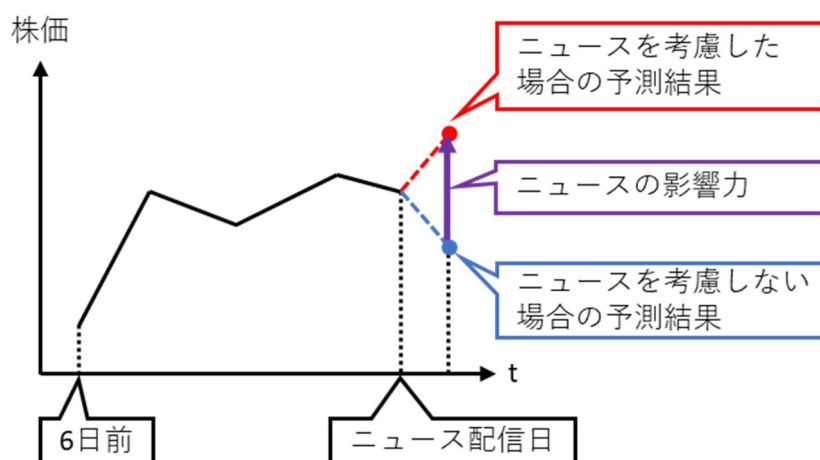


図2 ニュース記事の経済的影響力の推定[6]

ニュース記事の経済的な影響力の推定手法として、ニュース記事を考慮した株価指数を予測するモデルとニュース記事を考慮しない予測モデルの予測結果の差を用いた手法を開発した。具体的には、株価指数のテクニカル指標とニュース記事に基づく素性から株価を予測するモデルを構築し、テクニカル指標とニュース記事を用いた場合とテクニカル指標のみを用いた場合の差に着目し、差が大きいほどニュース記事で報じられている内容の影響力が大きいとする。予測モデルでは、テクニカル指標の系列を LSTM に入力して得られたベクトルとニュース記事のベクトルを連結したものを全結合層に入力し、翌日の株価の変化量を予測する。テクニカル指標としては、日ごとに Stochastic %K, Stochastic %D, Momentum, Rate of Change, William's %R, A/D Oscillator, Disparity 5 の 7 つの指標を用いてベクトルを作成し、ニュースの配信日を含む過去 7 日分を LSTM に入力する。また、ニュース記事の部分の処理には、以下の 2 つの手法を検討した。

- 朝日新聞の新聞記事を用いて GloVe で学習した単語分散表現である朝日単語ベクトル

でニュース記事の各単語のベクトルを表現し、その平均を用いる。この手法では新聞記事がない場合は0ベクトルが入力されたとする。

- ニュース記事をBERTに入力し、各トークンの隠れ状態の平均を用いる。この手法では新聞記事がない場合はニュース記事がない場合は空のニュース記事が入力されたのみならず。

株価の予測精度によって2つの手法を比較したところ、BERTを用いた手法の方が良好な結果となったため、提案手法としてはBERTを用いた手法を採用した。評価としては、10種類の業界（エネルギー資源、医薬品、機械、建設・資材、自動車・輸送機、情報通信・サービスその他、食品）に対して、業界ごとに予測モデルを構築してニュース記事の影響力を推定して有効性を確認した。さらに、モデル内で使用しているAttentionを用いて、ニュース記事において影響力の推定の根拠となった部分を可視化する方法を開発した。

(3) 投資分析

ソーシャルトレーディングとは、ソーシャルネットワークサービス(SNS)の取引プラットフォームで、ユーザが自分の投資経験や知識を共有することができる。ソーシャルトレーディングの特徴は、他のトレーダをフォローすることで、そのトレーダの取引を真似ることができる点である。したがって、信頼できるエキスパートのトレーダを見つけてフォローすることがソーシャルトレーディングの重要な課題である。そのため、トレーダをランク付けするシステムや手法が開発されているが、それらをサポートする金融理論が存在しない。

そこで、本研究では金融工学におけるポートフォリオ理論を導入して、ソーシャルトレーディングサービスにおけるトレーダの特徴を分析する技術を開発した[7]。リスク管理やポートフォリオ理論を活用して、全投資者からなるポートフォリオの効率的フロンティア曲線を構築し、効率的フロンティア曲線までの距離に基づいて投資者をランキングする手法を開発した。さらに、ポートフォリオ理論を拡張してトレーダポートフォリオという新しい概念を提案し、それに基づいてトレーダの特徴を分析してランキングを行う手法を開発した。実際のデータセットを用いた実験結果から、本手法が従来の手法より優れていることが示された。さらに、提案手法で見られたエキスパートトレーダーは、任意のリスクレベルに対してより良いポートフォリオを構築することが可能であることを実験で示した。

[1] 福知侑也, 馬強: データのスパース性を考慮した企業推薦手法の提案, DEIM2022

[2] Chengyang Ye, Qiang Ma: Representation Learning of Time Series Data with High-Level Semantic Features, DEIM2022

[3] Chengyang Ye, Qiang Ma: TS2V: A Transformer-Based Siamese Network for Representation Learning of Univariate Time-Series Data. CSCWD 2022: 1245-1250

[4] Chengyang Ye, Qiang Ma: GP-HLS: Gaussian Process-Based Unsupervised High-Level Semantics Representation Learning of Multivariate Time Series. DASFAA (1) 2023: 221-23

[5] Makoto Kirihata, Qiang Ma: A Trend-Shift Model for Global Factor Analysis of Investment Products. IEICE Trans. Inf. Syst. 102-D(11): 2205-2213 (2019)

[6] 井口勝太, 湯本高行: BERTを用いた株価への影響力のあるニュース記事の推定, DEIM2022

[7] Zimao Liu, Qiang Ma: Portfolio-Based Ranking of Traders for Social Trading. IEEE Access 8: 145363-145371 (2020)

5. 主な発表論文等

〔雑誌論文〕 計5件（うち査読付論文 5件/うち国際共著 5件/うちオープンアクセス 5件）

1. 著者名 Zhang Yang, Ma Qiang	4. 巻 -
2. 論文標題 Dual Attention Model for Citation Recommendation with Analyses on Explainability of Attention Mechanisms and Qualitative Experiments	5. 発行年 2022年
3. 雑誌名 Computational Linguistics	6. 最初と最後の頁 1~59
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/coli_a_00438	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する
1. 著者名 Liu Zimao, Ma Qiang	4. 巻 8
2. 論文標題 Portfolio-Based Ranking of Traders for Social Trading	5. 発行年 2020年
3. 雑誌名 IEEE Access	6. 最初と最後の頁 145363~145371
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2020.3011713	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する
1. 著者名 Zhang Yang, Ma Qiang	4. 巻 8
2. 論文標題 DocCit2Vec: Citation Recommendation via Embedding of Content and Structural Contexts	5. 発行年 2020年
3. 雑誌名 IEEE Access	6. 最初と最後の頁 115865~115875
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2020.3004599	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する
1. 著者名 KIRIHATA Makoto, MA Qiang	4. 巻 E102.D
2. 論文標題 A Trend-Shift Model for Global Factor Analysis of Investment Products	5. 発行年 2019年
3. 雑誌名 IEICE Transactions on Information and Systems	6. 最初と最後の頁 2205~2213
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2018EDP7420	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1. 著者名 ZHANG Yang, MA Qiang	4. 巻 E105.D
2. 論文標題 MP-BERT4REC: Recommending Multiple Positive Citations for Academic Manuscripts via Content-Dependent BERT and Multi-Positive Triplet	5. 発行年 2022年
3. 雑誌名 IEICE Transactions on Information and Systems	6. 最初と最後の頁 1957 ~ 1968
掲載論文のDOI (デジタルオブジェクト識別子) 10.1587/transinf.2022EDP7034	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 該当する

[学会発表] 計17件 (うち招待講演 0件 / うち国際学会 6件)

1. 発表者名 Chengyang Ye
2. 発表標題 GP-HLS: Gaussian Process-Based Unsupervised High-Level Semantics Representation Learning of Multivariate Time Series
3. 学会等名 DASFAA2023 (国際学会)
4. 発表年 2023年

1. 発表者名 Yang Zhang
2. 発表標題 Recommending Multiple & Positive Citations for Manuscript via Content Dependent Modeling and Multi Positive Triplet
3. 学会等名 WI-IAT 2021 (国際学会)
4. 発表年 2021年

1. 発表者名 Chengyang Ye
2. 発表標題 A Transformer-Based Siamese Network for Representation Learning of Univariate Time-Series Data
3. 学会等名 CSCWD2022 (国際学会)
4. 発表年 2022年

1. 発表者名 福知侑也
2. 発表標題 データのスパース性を考慮した企業推薦手法の提案
3. 学会等名 DEIM2022
4. 発表年 2022年

1. 発表者名 平野瑠登
2. 発表標題 複数ニュース記事の株に及ぼす影響の分析手法
3. 学会等名 DEIM2022
4. 発表年 2022年

1. 発表者名 Chengyang Ye
2. 発表標題 Representation Learning of Time Series Data with High-level Semantic Features
3. 学会等名 DEIM2022
4. 発表年 2022年

1. 発表者名 井口勝太
2. 発表標題 BERTを用いた株価への影響力のあるニュース記事の推定
3. 学会等名 DEIM2022
4. 発表年 2022年

1. 発表者名 Yang Zhang
2. 発表標題 Dual Attention Model for Citation Recommendation
3. 学会等名 COLING 2020 (国際学会)
4. 発表年 2020年

1. 発表者名 Satoshi Yoshida
2. 発表標題 Generating Dialogue Sentences to Promote Critical Thinking
3. 学会等名 DEXA 2020 (国際学会)
4. 発表年 2020年

1. 発表者名 前川丈幸
2. 発表標題 ニュース記事の読み方の判断支援に関する研究
3. 学会等名 DEIM2021
4. 発表年 2021年

1. 発表者名 福知侑也
2. 発表標題 キャリアパスを考慮したネクスト企業推薦手法
3. 学会等名 DEIM 2021
4. 発表年 2021年

1. 発表者名 米田宏生
2. 発表標題 ニュース記事の考慮の有無による 株価指数の予測結果の差に基づく経済的影響力の推定
3. 学会等名 情報処理学会第83回全国 大会
4. 発表年 2021年

1. 発表者名 Yang Zhang
2. 発表標題 Citation Recommendations Considering Content and Structural Context Embedding
3. 学会等名 BigComp 2020 (国際学会)
4. 発表年 2020年

1. 発表者名 桐畑 誠
2. 発表標題 金融市場における局面検出
3. 学会等名 ソーシャルビッグデータ研究会
4. 発表年 2019年

1. 発表者名 吉田聖
2. 発表標題 批判的思考を促進する対話文生成に関する研究
3. 学会等名 DEIM2020
4. 発表年 2020年

1. 発表者名 米田宏生
2. 発表標題 有価証券報告書の分析に基づく重要な新着ニュースの発見.
3. 学会等名 第169回データベースシステム・第136回情報基礎とアクセス技術合同研究発表会
4. 発表年 2019年

1. 発表者名 伊藤優希
2. 発表標題 受容性を考慮した多様な意見の検索手法
3. 学会等名 DEIM2023
4. 発表年 2023年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

DEIM2022 スポンサー賞（ウォンテッドリー賞）データのスパース性を考慮した企業推薦手法の提案福知侑也（京都大学），馬強（京都大学） WebDB2019 企業賞（Gunosy賞），有価証券報告書の分析に基づく重要な新着ニュースの発見 米田 宏生（兵庫県立大学）

6. 研究組織			
	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	湯本 高行 (Yumoto Takayuki) (20453152)	兵庫県立大学・社会情報科学部・准教授 (24506)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8 . 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------