公開

研 究 終 了 報 告 書

「イベント情報を活用する高精度時系列モデリング技術の構築」

研究期間: 2019年 10 月 ～ 2023年 3 月
研 究 者: 小林 亮太

1. 研究のねらい

Web・センサー・社会・経済・生物など多様なシステムから時系列 (時間情報を含むデータ) が得られるようになり、データ活用の重要性は日々高まりつつある。ビッグデータのデータサイズとその活用分野は日々拡大を続けている。ビッグデータの活用事例では、**時間情報を無視して深層学習 (機械学習) を適用するデータ駆動型アプローチが主流**であり、時系列自体がモデル化されることは少ない。その一方で物理学や工学分野などでは、時系列の数理モデル (微分方程式等) を構築し、実験や観測によって数理モデルを検証し改善を行う、という方法論により発展してきた。時系列を数理モデル化できれば、① 対象システムの理解が深まる、② シミュレーションにより将来を予測できる、③ 対象システムを理論的に制御できる (制御理論)、など様々な利点がある。こうした利点があるにも関わらず、ビッグデータの活用において時系列がモデル化されないのはなぜだろうか？ この原因として、計測されたビッグデータの多くは複雑システム (複雑系) から得られたものであり、以下の **3 つの困難**があるためであると考えられる。

困難 A システムが非定常性・非線形性を持つ。
困難 B. システムの支配方程式が不明。
困難 C. 全ての変数を計測できない。

本研究では、時系列の中でも「イベント時系列」(点過程) に着目する。イベント時系列とは、あるイベントが起きた時刻のデータであり、Twitter (Web)、会話履歴 (社会)、購買履歴 (マーケティング)、神経スパイク (脳) など分野横断的に見られる。

本研究では、上述した 3 つの困難を解決し、複雑システム (脳やソーシャルネットワークなど) から得られたイベント時系列から数理モデルを構築する技術を開発することを目的とした。また、開発技術は Web アプリや Github などにより公開する。さらに、開発技術を実問題に適用することにより、その有用性を実証するとともに、脳科学・Web 情報学などの分野へ貢献することも目指した。

2. 研究成果

(1) 概要

時系列データから数理モデルを構築する上での 3 つの困難を解決するため、以下の研究テーマを設定して研究を進めた。

テーマ (1) 多次元イベント時系列データから相互作用を推定する技術の開発
テーマ (2) ソーシャルメディアにおける情報拡散の数理モデリング
テーマ (3) イベント情報を活用する時系列分析技術の開発

テーマ (1) では、点過程の数理モデルに基づく統計的手法 (Kobayashi et al. 2019) やニューラルネットワーク (代表的な論文 [1]: Endo, Kobayashi et al. 2021) を提案することにより、神経スパイクデータから神経細胞間の相互作用の強さを高精度に推定する技術を開発した。さらに、これらの開発技術を手軽に試すことができる Web アプリを作成して公開した。(https://s-shinomoto.com/CONNECT/) また、Python コードは GitHub で公開している。(https://github.com/NII-Kobayashi/GLMCC)

テーマ (2) では、フェークニュースが Twitter 上で拡散するパターンを高精度に予測することができる数理モデルを開発した (代表的な論文[2]: Murayama et al. 2021)。特に本研究の興味深い点は、「フェークニュースは初めには珍しいニュースとして認知されるが、その後偽情報に人々の認知が変わると情報拡散の速度も変わる」という仮説を提案した点にある。

テーマ (3) では、イベント情報を活用する時系列解析技術を開発した。ただし、新型コロナウィルスの感染拡大の影響を受け、テーマ(3)の一部を 「新型コロナワクチンをめぐる人々の話題・関心の変化の分析」 に変更した。本研究では Wikipedia 記事へのアクセス数の時系列に焦点を当て、記事の内容 (イベント情報) を考慮に入れた時系列分析技術を開発した (Kobayashi et al. ICWSM 2021)。そして、イベント情報の活用によって予測精度が大きく向上することを示した。次に、大規模ツイートデータ (1.1 億ツイート) を分析することにより、新型コロナワクチン接種に対して日本の人々がどのように反応し、関心が変化したかを調べた (代表的な論文[3]: Kobayashi et al., 2022)。

(2)詳細

本項目では、テーマ (1)、(2)、(3)に関連する研究成果を述べる。

テーマ (1) 多次元イベント時系列データから相互作用を推定する技術開発

本課題の研究成果として、我々が開発した技術 GLMCC (Kobayashi et al., Nat Commun, 2019) を説明する。この手法は、神経データの解析技術である相互相関解析 (Cross-Correlation) とイベント時系列解析を融合させたものである (図 1)。ここでは、単純のため 2 つのイベント時系列データ (神経細胞が 2 つ) である場合を説明する。より多数の神経細胞の場合には、各ペアについて以下に説明する分析を繰り返せばよい。

イベント時系列データを $\{t_1^A, t_2^A, \cdots, t_{n_A}^A\}$, $\{t_1^B, t_2^B, \cdots, t_{n_B}^B\}$ 、ただし、$t_k^{A(B)}$は細胞 A (B) の k 番目のスパイク時刻である。まず、2 つのイベント時系列データのイベント時刻の差を計算することで、1 つのイベント時系列データ (Cross-Correlation データ) $\{s_{i,j}\}$ (ただし、$s_{i,j} = t_j^B - t_i^A,\ 1 \le i \le n_A,\ 1 \le j \le n_B$) に変換する。

次に、時間差データ $\{s_{i,j}\}$ を非一様 Poisson 過程を使ってモデル化する:

$$\lambda(s) = \exp[a(s) + J_{\mathrm{BA}}\kappa(s) + J_{\mathrm{AB}}\kappa(-s)], \qquad (1)$$

ただし、$a(s)$は共通入力、J_{BA} は A から B への相互作用の強さ、$\kappa(s) = \exp\left[-\frac{t-d}{\tau}\right]$ $(d \le t),\ \kappa(s) = 0\ (t < d)$ は相互作用関数である。さらに、共通入力 $\{a(s)\}$ の事前分布として、滑らかな関数を仮定する:$p(\{a(s)\}) \propto \exp\left[-\frac{1}{\gamma}\int_{-W}^{W}\left(\frac{da}{ds}\right)^2 ds\right]$.

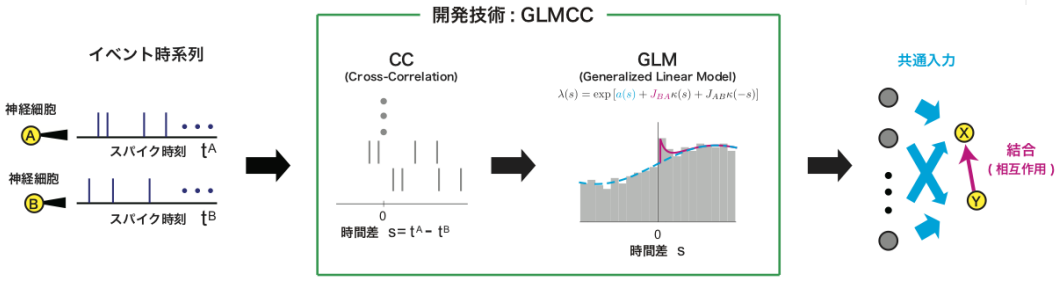

図 1: 開発技術 GLMCC の模式図。

あとは、ベイズ推定の枠組みを用いることで変数 J_{BA}, J_{AB} を推定するアルゴリズムを導出できる。また、相互作用がない $(J_{\mathrm{BA}} = 0)$ という仮説検定を考えることで相互作用 (因果関係) がないことの有意性を議論できる。開発手法の妥当性を確認するため、脳の局所回路に近い 1,000 個のモデルニューロンからなる大規模神経回路のシミュレーションを行い、この人工データを分析した。この結果、**開発技術 (精度 97%, MCC 0.7) は既存技術の推定精度 (精度 77%, MCC 0.4) を大幅に向上することが示された。**さらに、提案手法をラット脳海馬から計測された実験データに適用して神経細胞間の結合を推定した結果、神経活動の様子などから専門家が判定した結果と一致することを確認した。

しかし、我々が開発した技術 GLMCC には、ユーザ (データ分析者) がいくつかのパラメータをデータに合わせて調整しなければならないという問題点があった。そこで、この問題点を解決する機械学習技術 CONNECT (図 2) を開発した (代表論文[1]: +Endo, +Kobayashi et al., Sci Rep 2021)。開発手法 CONNECT は、シナプス結合の有無を判定する問題を Cross Correlation ヒストグラムのパターン認識の問題と定式化し、畳み込みニューラルネット (CNN) により推定を行う。次に、シミュレーションデータに適用して開発手法の精度を評価した。この結果、**開発手法 CONNECT はパラメータを調整することなく、GLMCC と 同程度の推定精度を実現する**ことを確認した。

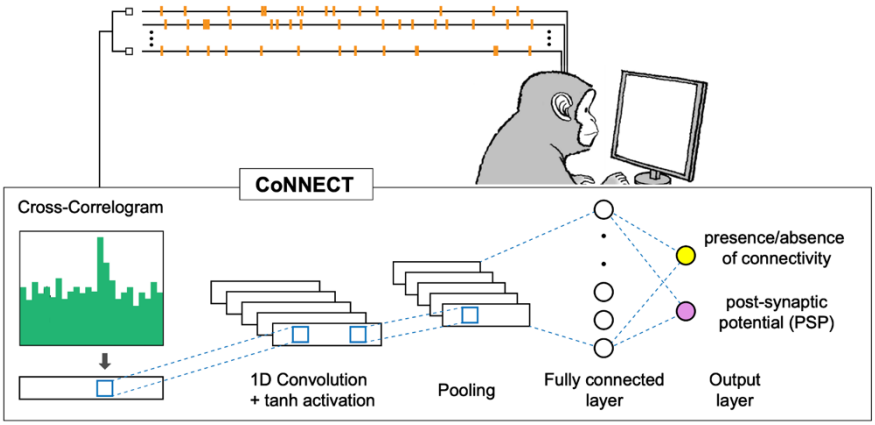

図 2: 提案技術 CONNECT: 本技術では、ペアの神経細胞ごとに Cross-Correlation データを計算し、入力を Cross-Correlation、出力を神経細胞間に結合があるかないか (結合があるとしたらその強さ)、とするニューラルネットを開発した。

テーマ (2) ソーシャルメディアにおける情報拡散の数理モデリング

ソーシャルメディア（Twitter, Facebook）は社会に悪影響を及ぼす可能性のあるフェイクニュースの温床となりつつある。そこで、Twitter 上におけるフェイクニュース拡散の様子を記述する数理モデル構築を行った（代表論文[2]：Murayama et al., PLoS ONE, 2021)。提案モデルでは、① 通常のニュースとしての拡散、② ニュースが偽であるという事実の拡散、の 2 段階でフェイクニュースは拡散する。我々は、2 種類のデータセット（2019 年の米国データ、2011 年東日本大震災直後の日本語データ）を用いて、提案モデルの予測精度を調べた。どちらのデータセットに対しても提案モデルは既存技術に比べて高い予測精度を達成した。

テーマ (3) イベント情報を活用する時系列分析技術の開発

ソーシャルメディアデータでは、イベントの発生した時刻だけでなく、イベント自体の情報 (投稿メッセージや動画の内容) が重要であるにもかかわらず、時系列解析では無視されることが多かった。本課題では、イベント情報を活用する時系列解析の技術開発を行なった。本研究では Wikipedia 記事へのアクセス数時系列に焦点を当て、記事の内容(イベント情報)考慮に入れた時系列モデルの開発を行った (Kobayashi et al., ICWSM 2021)。 その結果、

人々の興味の変化についての時間スケールは記事の内容 (選挙、サッカーの試合など) に大きく依存することが明らかになった。また、イベント情報を活用することによって、アクセス数時系列の予測精度も大きく向上した (図 3)。

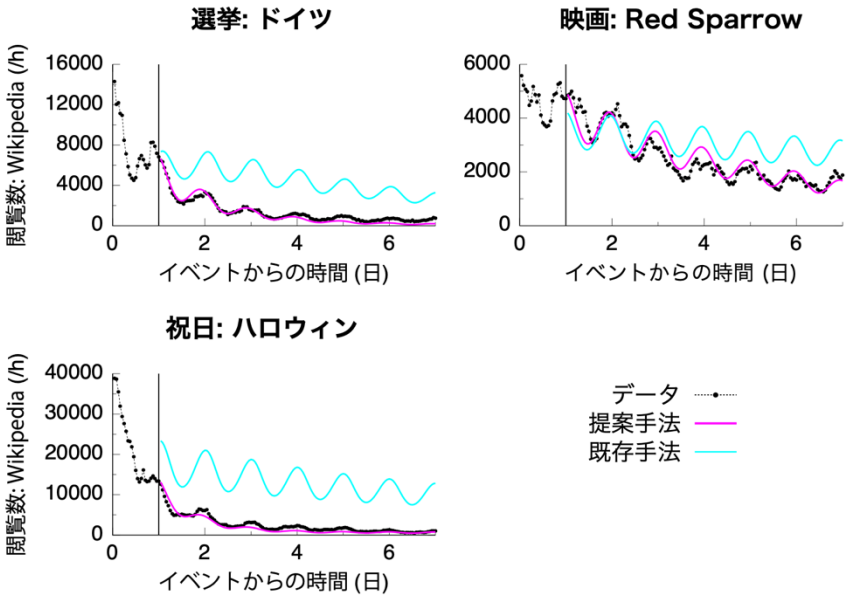

図 3: イベント情報を活用する時系列予測: 本研究では、イベント情報 (Wikipedia 記事の内容) を活用して時系列を予測する方法を開発した。提案手法 (シアン) は、既存技術 (マゼンタ) に比べて予測精度を大きく改善させた。

日本での新型コロナワクチンの接種は、ワクチンの安全性・有効性に対する不安やワクチン接種に関する政策への不満があったにもかかわらず、欧米諸国と比べると迅速に進んだ。短期間で高い接種率に達する過程で、人々が何を考え、何に関心を抱いたかを知ることは公衆衛生上の重要な問題である。そこで我々は、イベント情報を活用する Twitter データの分析技術を開発し、2021 年 1～10 月 (ワクチン接種期間) に投稿された「ワクチン」を含む全ての日本語ツイート (1.1 億ツイート) を分析した。 この結果、2021 年 6 月の職域接種の開始を境に、ワクチン政策、ワクチンの有効性、ワクチン関連ニュースなど社会的トピックに関するツイートの割合が減り、接種を受ける予定、接種後の体調の変化、副反応の報告など個人的事柄に関するツイートの割合が増えたことを発見した。また、陰謀論関連のキーワードが含まれるツイートの割合は 6%であり、その中には冗談をつぶやいているツイートも多かったため、日本語 Twitter はワクチンの陰謀論の温床にはなってないことが示唆される。この成果は、医療情報学の主要雑誌 Journal of Medical Internet Research に掲載された (代表的な論文 [3] : Kobayashi et al, JMIR 2022)。また、この成果は日経産業新聞で紹介されるなど社会的注目も集めつつある。

3. 今後の展開

本研究では、神経データの分析技術: テーマ (1)、Web・ソーシャルメディアデータの分析

技術: テーマ(2), (3) について研究を進めた。以下では、神経データ分析と Web・ソーシャルメディア分析における今後の展開を説明する。

神経データ分析としては、神経スパイクデータから高精度に神経結合を推定する技術 (GLMCC, CONNECT) を開発できた。今後は、学会に積極的に参加したり、シンポジウムを企画したりして開発技術の宣伝を行い、より多くの研究者に使ってもらえるよう広報したい。直近では、2023 年 8 月の神経科学学会においてシンポジウム「ビッグデータは神経科学を変えるのか」を企画予定である。

Web・ソーシャルメディア分析については、イベント情報を活用することによって時系列解析の精度が向上するという結果が得られ、本研究アプローチの有用性を支持する結果を得た (Kobayashi et al., ICWSM, 2021)。しかし、開発技術では、イベント情報が「スポーツ」、「選挙」のように人手で意味付けされていることが仮定されているため、現状では Wikipedia など限られたデータにのみ有効である。今後は、自然言語処理技術を用いて意味構造を抽出する技術開発を進め、Web データ全般に適用可能なイベント情報の活用技術を開発したい。このような技術は、大規模 Twitter データ分析を効率化させるためにも重要である。

4. 自己評価

研究目的の達成状況

研究開始時には、以下の 3 つの課題を進めることを目的に研究を進めた。

テーマ (1) 多次元イベント時系列データから相互作用を推定する技術開発

テーマ (2) ソーシャルメディアにおける情報拡散の数理モデリング

テーマ (3) イベント情報を活用する時系列分析技術の開発

テーマ (1), (2), (3) のそれぞれで学術論文、国際会議論文を出版できたので、研究目的を十分に達成できたと評価できる。研究構想時に達成できると予想していたテーマ (1), (2) については当初想定通りに目標を達成できた。一方で、挑戦的な課題と予想していたテーマ (3) の達成度は 60%程度である。テーマ (3) に関連してベイズ統計に基づくイベント情報活用技術 (Kobayashi et al., ICWSM 2021) を提案したが、この技術を Web データ一般に適用するのは難しいからである。新型コロナウィルス感染拡大の影響を受け、テーマ (3)の一部を「新型コロナワクチンをめぐる人々の話題・関心の分析」という課題に変更して研究を進めた結果、 医療情報学の主要雑誌 Journal of Medical Internet Research に掲載される想定外の成果を出すことができた (代表的な論文[3] : Kobayashi et al, 2022)。この成果は日経産業新聞で紹介されるなど社会的注目も集めつつある。

以上をまとめると、新型コロナウィルスの感染拡大なども影響もあり、当初の研究構想とは一部違う方向に研究は進んだが、最終的には満足のいく研究成果が得られたと言える。

研究の進め方(研究実施体制及び研究費執行状況)

研究構想時には研究費の大部分は、国内外の数学系研究者と新しい共同研究を行うための旅費として計画していた。残念ながら、新型コロナウィルス感染拡大の影響を受け、さきがけ

研究期間のうち 2 年半ほど自由に出張できない状況が続いた。そのため、研究費を計画通りに執行できなかった。新型コロナウィルス勃発当初は、予定通り研究を進捗させるため、対面打ち合わせをZoomなどのオンラインミーティングに代用することを試みた。しかし、全く新しい研究を進めるためのブレインストーミングにはオンラインミーティングは非効率的であり、予定通りに研究を進捗させることは不可能であった。限られた研究期間で成果を最大化するため、当初想定していた挑戦的課題 (テーマ (3)の一部) を変更し、社会的に重要な問題であった新型コロナワクチンに関する研究を進めた。

研究成果の科学技術及び社会・経済への波及効果

テーマ (1) の成果である神経シナプス結合の推定技術 (Kobayashi et al, Nat Commun, 2019; 代表的な論文 [1]: +Endo, +Kobayashi et al, 2021) を使うことで、多数の神経細胞から計測されたスパイクデータから脳における情報の流れや情報処理様式を分析できる。本技術は神経科学におけるデータ分析ツールとして、広く使われることが期待される。また、新型コロナワクチンについての研究成果 (代表的な論文[3]) によって、ワクチン接種に対して、日本の人々がどのように反応し、気持ちや関心が変化したかを調べることができた。今後は、インターネット上にあふれる人々の「生の声」を効率的にデータ分析するための数理・情報基盤技術の開発を進めたい。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数: 11 件

以下、＊はCorresponding Author、+はEqual Contribution を示す。

1. +Endo D, +Kobayashi R, Bartolo R, Averbeck BB, Sugase-Miyamoto Y, Hayashi K, Kawano K, Richmond BJ, and ＊Shinomoto S. “A convolutional neural network for estimating synaptic connectivity from spike trains”, Scientific Reports 11: 12087, (2021).
概要: 本論文では、神経細胞のスパイクデータから神経細胞間のシナプス結合 (脳の回路図) を推定するニューラルネットワーク (CONNECT) を開発した。シミュレーションデータに適用した結果、開発手法は従来手法に比べてはるかに高い推定精度を有することを確認した。計測技術の驚異的進展により、様々な脳領野から大規模スパイクデータが計測され始めている。開発手法によって、脳における情報の流れや情報処理様式が明らかにされることが期待される。
2. Murayama T, Wakamiya S, Aramaki E, and ＊Kobayashi R. “Modeling the Spread of Fake News on Twitter”, PLOS ONE 16(4): e0250419, (2021).
概要: 本論文では、フェークニュースが Twitter 上で拡散する様子を再現するモデルを構築した。さらに提案モデルに基づく予測技術を開発した。その結果、英語と日本語の 2 種類の

データセットに対して、提案モデルはフェークニュースの拡散の規模を高精度に予測できることを示した。
3. ＊+Kobayashi R, +Takedomi Y, +Nakayama Y, +Suda T, Uno T, Hashimoto T, Toyoda M, Yoshinaga N, Kitsuregawa M, and Rocha LEC. “Evolution of the public opinion on COVID-19 vaccination in Japan: Large-Scale Twitter Data Analysis”, *Journal of Medical Internet Research*, 24, 12, e41928, (2022).
概要：本論文では、日本における COVID-19 ワクチン接種期間中の 2021 年 1〜10 月に Twitter で投稿された「ワクチン」を含む1億件以上の日本語の全ツイートデータを時系列的に分析した。6 月から開始された職域接種を境に、人々の関心がニュース、政治などの社会的トピックから接種後の感想、副反応などの個人的事柄へ推移したことを発見した。本研究は、Twitter による個人的体験の共有がワクチン接種への安心感を醸成した可能性を示している。

(2)特許出願

研究期間全出願件数：０件（特許公開前のものも含む）

(3)その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

① Web アプリ：スパイクデータから神経結合を推定する。
https://s-shinomoto.com/CONNECT/

② メディア報道：研究成果 (代表的な論文 3) が日経産業新聞 (2023 年 1 月 23 日) で紹介された。
https://www.nikkei.com/article/DGXZQOUC138ZS0T10C23A1000000/

③ メディア報道：これまでの研究成果 (代表的な論文 1, 2 など) がインターネットメディア「一歩先への道しるべ」(日経 BP 総合研究所) に掲載された。
https://project.nikkeibp.co.jp/onestep/feature/00021/111600001/

④ プレスリリース：新型コロナワクチンをめぐる人々の話題・関心の変化を分析
―1億超の大規模 Twitter データを読み解く―
https://www.jst.go.jp/pr/announce/20221223/index.html

⑤ 国際会議発表 (招待講演): “Estimating synaptic connectivity from parallel spike train,” Ryota Kobayashi, 2021 International Conference on Mathematical Neuroscience