《第66回》とても不完全なデータを分析可能とすることを目指して

幸　島　匡　宏

1.　はじめに

筆者は現在，NTT 人間情報研究所にて人の状態や行動を推定・予測・制御する技術に関する研究開発を行っています．専門は機械学習で，これまでに個人や集団のさまざまな活動，たとえば商品の購買[1),2)]や都市における移動[3),4)]を記録したデータを分析する確率モデルの研究を進めてきました．本稿では，筆者の研究経歴と，近年取り組んでいる不完全なデータの分析手法に関する研究を紹介します．

2.　筆者の背景

筆者は修士課程を修了した 2012 年に NTT に入社しました．“ビッグデータ”が注目ワードとして取り上げられたり，「データサイエンティストが 21 世紀で最もセクシーな職業である」という論考が発表されるなど，データ活用の機運が世の中に高まっている時期でした．入社したばかりで右も左もわからない中，データ活用に関する社内外の打合わせ・議論に参加したことを覚えています．

現場の方々との議論を通して得た 1 番の収穫は，「手元にあるデータからできそうなこと」を考えてしまう自分の思考のバイアスに気づけたことです．たとえば，ある小売店や EC サイトの会員であるユーザの購買履歴データが利用できる場合，当時の私はデータ中の会員の購買傾向を分析するという方向で検討しがちでした．そんな時に，現場の方は「会員ではない潜在/見込みユーザや他社のユーザなど市場に存在する人々全体のことを考えて戦略を考えている」ことを議論を通して知れたことで，「会員だけを分析すること」を考えていた自身のバイアスに気づきました．この経験から，分析の目的や活用の仕方を考えて研究に取り組みたいという思うようになりました．

本稿で紹介する不完全なデータを分析する手法は上記の思いから作られたものです．不完全なデータというと欠損値や生存分析における打ち切り[5)]のあるデータを思い浮かべる方も多いと思います．しかし，すべてを網羅したデータを手に入れることは不可能ですので，欠損値の有無にかかわらず，分析の目的によってはどのようなデータも不完全なデータになりえます．たとえば，会員データは，会員でない多くの人の情報が含まれていないため，市場の人々全体を知るという目的のためにはとても不完全なデータだと言えます．なお，筆者は市場全体に関する情報に相当するユーザ集団（男性 30 代全体など）のデータが入手可能であるという条件のもと，会員と市場全体のユーザの両方を統合して分析する手法を提案しています．この手法については解説記事[1)]および論文[2)]をご参照ください．

3.　現在の研究内容

ここでは近年筆者が提案した不完全なデータを分析する手法を 2 つ紹介します．両手法ともデータが生成される過程を確率モデルとして数学的に表現することで，目的とする分析を可能としています．

不完全な移動データの分析：都市計画や店舗出店などのマーケティング活用のため，スマートフォンや IC カードで収集された人の位置情報や移動を記録したデータの分析が盛んです．ただし位置情報データは，たとえばプライバシー保護のために数十分間隔で大きく間引くよう加工されたり，ほかにも地下施設などの測位不可能なエリアが存在することにより，分析に利用できるデータは**図 1** に示す経由地点を訪問した情報が欠落した不完全な移動データとなります．筆者はこのような不完全なデータから任意の地点間の遷移確率を推定することのできる手法を提案しました．具体的には，データ中の人の

JL 0007/24/6307–0385

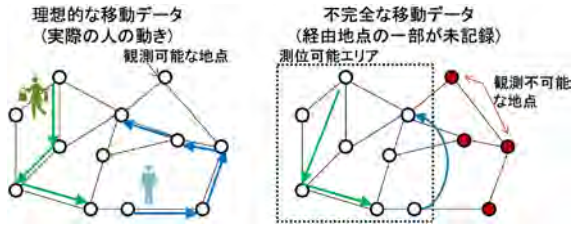

図 1　不完全な移動データの例

グラフは都市の空間構造を表わす．空間構造に従う実際の人の動き（左）から経由地点の情報が欠落している．

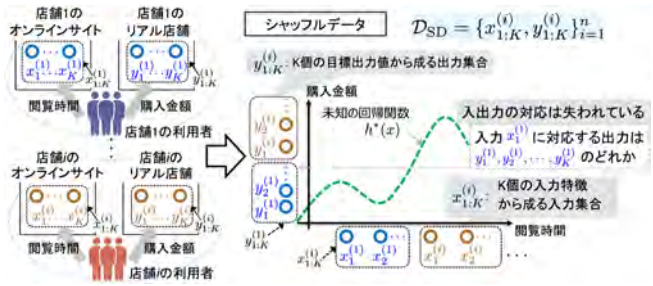

図 2　不完全な入出力データの例

閲覧時間と購入金額の情報が別々に収集されたことで，入出力の対応関係が失われている．

移動は複数回の遷移の結果であることを確率モデルにより表現することで，“遷移行列のべき乗の混合モデル”[3]と“センサーマルコフ連鎖の理論に基づくモデル”[4]を導き，推定アルゴリズムを構築しました．これにより都市の空間構造に基づいた人の移動傾向を把握できるようになり，店舗出店場所の検討などに活用できます．

不完全な入出力データの分析：企業が自社商品への顧客のエンゲージメントを高める要因を知りたい，学校が生徒の学業成績向上に寄与する要因を特定したいなど，入力変数（例：年齢や性別，オンラインサイト閲覧時間，勉強時間）と出力変数（例：購入金額，試験のスコア）の定量的関係を把握するために回帰分析が広く利用されます．回帰分析を行うためには，入力と出力の対応のあるデータが必要です．しかし，入力と出力に対応する情報が，オンラインサイトと実店舗のように異なる部門や組織で収集されていたり，個人が識別できないように集団単位で収集される場合などは，入出力の対応は失われます．そのため，利用できるデータはシャッフルデータと呼ばれる不完全なデータとなります（**図 2**）．筆者は，この不完全なデータが生成される過程を深層学習に基づく確率モデルにより表現し，シャッフルデータから回帰関数を推定する手法を提案しました[6]．これまでも真の回帰関数が線形であるなど限定的な状況で利用できる手法は存在していましたが，深層学習モデルの高い表現力を活用した提案手法を用いることで，非線形な回帰関数を対象とする回帰分析が可能となります．

4.　おわりに

実世界のさまざまな不完全なデータの分析を実現することで，データ分析の適用範囲および推定・予測の対象を広げることができると考えています．紙面の都合上触れられませんでしたが，観測が制限された状況での制御・意思決定に関する研究[7]や大規模イベント実施時の混雑解消[8]，生活習慣の改善[9]を目指した研究も並行して進めてきました．今後も推定・予測・制御を通じて，より良い社会の実現を目指していきます．

（2024 年 3 月 29 日受付）

参　考　文　献

1) 幸島，松林，澤田：複合データ分析技術と NTF (1) 複合データ分析技術とその発展，電子情報通信学会誌，**99**–6，543/550 (2016)
2) M. Kohjima, T. Matsubayashi, and H. Sawada: Learning of Nonnegative Matrix Factorization Models for Inconsistent Resolution Dataset Analysis, *IEICE Transactions on Information and Systems*, **102**–4, 715/723 (2019)
3) M. Kohjima, T. Kurashima, and H. Toda: Learning with Labeled and Unlabeled Multi-Step Transition Data for Recovering Markov Chain from Incomplete Transition Data, *International Joint Conferences on Artificial Intelligence*, 2412/2419 (2020)
4) M. Kohjima, T. Kurashima, and H. Toda: Inverse Problem of Censored Markov Chain: Estimating Markov Chain Parameters from Censored Transition Data, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 297/308 (2023)
5) M. Kohjima, T. Matsubayashi, and H. Toda: Variational Bayes for Mixture Models with Censored Data, *Machine Learning and Knowledge Discovery in Databases*, 605/620 (2018)
6) M. Kohjima: Shuffled Deep Regression, *AAAI Conference on Artificial Intelligence*, 13238/13245 (2024)
7) M. Kohjima, M. Takahashi, and H. Toda: Censored Markov Decision Processes: A Framework for Safe Reinforcement Learning in Collaboration with External Systems, *IEEE Conference on Decision and Control*, 3623/3630 (2020)
8) H. Kiyotake, M. Kohjima, T. Matsubayashi, and H. Toda: Multi Agent Flow Estimation Based on Bayesian Optimization with Time Delay and Low Dimensional Parameter Conversion, *Principles and Practice of Multi-Agent Systems*, 53/69 (2018)
9) M. Takahashi, M. Kohjima, T. Kurashima, and H. Toda: Can Reinforcement Learning Lead to Healthy Life?: Simulation Study Based on User Activity Logs, *International Conference on Pattern Recognition*, 4865/4872 (2021)

［著　者　紹　介］

幸島 匡宏（こうじま まさひろ）君

2009 年東京工業大学工学部情報工学科卒業．2012 年同大学院総合理工学研究科知能システム科学専攻修士課程修了．同年，NTT 入社．2019 年東京工業大学情報理工学院数理・計算科学系博士課程修了．博士（理学）．現在，NTT 人間情報研究所主任研究員．機械学習，特に確率モデル・不完全データ分析の研究に従事．

E-mail: masahiro.kohjima@ntt.com

所属：日本電信電話株式会社　NTT 人間情報研究所　横須賀市光の丘 1–1