データサイエンスにおけるモデリングのアプローチ

田村光太郎

CONTENTS

Ⅰ　データサイエンスプロジェクトでのモデル構築のアプローチ
Ⅱ　データサイエンスにおける２つのモデリング
Ⅲ　モデルの活用事例と課題
Ⅳ　最後に（モデリングで求められること）

要　約

1　コロナ禍において、蓄積してきた過去の大量のデータと直近のデータで性質が大きく異なり、データサイエンスの分析が難しくなっている。昨今は、データの不足やデータが存在しない中でのモデルの構築方式に注目が集まる。

2　モデルの構築は、解釈性と精度という２軸がトレードオフの関係となっている。この２軸は、解釈性が高いがドメイン知識に基づいて立式する難しさがある理解先行型モデルと、精度が高まるが複雑なモデルになる傾向があるタスク解決型モデルに大別される。

3　理解先行型モデルは、データが不足あるいは存在しない中でのモデリングを可能として、シミュレーションから得られる知見を通し、タスク解決型モデルのモデリングに反映していくという方式がとられることが多く、その使い分けが重要となる。

4　精度を高めることを中心に構築されるモデルは、複雑なモデルになる傾向があるが、これらモデルに解釈性を付与する技術が開発されている。特に、LIMEやSHAPなどが注目されている。

5　モデリングは技術だけで完結することは難しい。モデルに解釈性を付与する技術が開発されてはいるが、複雑な現象を端的に表現するための高いドメイン知識は依然必要となる。

Ⅰ データサイエンスプロジェクトでのモデル構築のアプローチ

最近は、コロナ禍で生活行動が極端に変わり、直近１、２年のデータと長年蓄積してきたデータの性質が大きく異なるようになった。たとえば、コロナ禍以前は季節周期などで変動していた売上や客数が、報道される感染者数の変動に連動したり、緊急事態宣言の発出の有無に左右されたりする傾向が見られるようになっている。通常は、データが時間軸方向で一貫した性質であれば、データ量が多いほど、より良いモデルが構築できると期待されている。しかし、過去の学習データとモデル利用時のデータの性質があまりにも異なる場合や、そもそもデータが極端に少ない場合は、これまでの統計や機械学習で対処するには難しい状況となってしまう。そのため、コロナ禍では少量のデータを使ったモデリングや、データを使わないモデリングの方式が再考され、データサイエンスにおける新しいモデリングの選択肢として注目が集まっている。

特に、新型コロナの感染者数予測は、８割の人流制限などモデルのシミュレーション結果によって、早くから意思決定が行われたことから、学習データが存在しない中でのモデリングが注目された。このようなモデリングは、複雑系科学やシステム科学の分野で長く扱われてきたアプローチであり、既にさまざまな分野で適用され、多くの知見や事例が存在する。本論考では、この機械学習やAIのモデリングアプローチと、複雑系科学のアプローチについて簡単に触れ、それぞれの役割や目的の違いについてまとめ、相互に必要となる技術であることを述べたい。

本章では、モデリングにおける昨今の課題と背景を述べた。第Ⅱ章では、モデルとはどのようなものかを述べ、第Ⅲ章では、具体的に実課題に応用されているモデルについて述べる。最後の第Ⅳ章で、今後のモデリング技術の発展について述べる。

Ⅱ データサイエンスにおける２つのモデリング

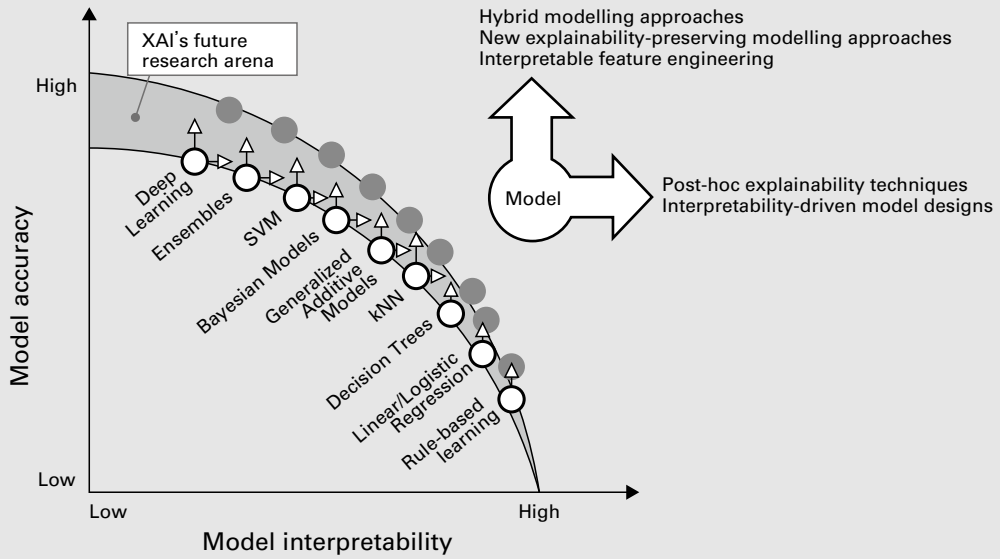
１ モデルに必要な「解釈性」と「精度」

モデルという用語は、ある物事を代表的なパーツだけで表した、現実を簡易的に表現したもののことである。模型や現実の近似と表現されることもある。モデルは、ある目的をもって組み上げられることがほとんどであるが、物事の理解や予測精度を高めるなどの目的ごとに、モデルの作り方はさまざまである。

このモデリングで重要となる観点として、「解釈性」と「精度」の２つが挙げられる。解釈性とは、モデルによる判断結果を、人間が解釈できる度合いのことであり、精度とは、あるタスクの解決具合のことをいう。

図１は、さまざまなアルゴリズムを使った場合に、モデルが持つ「解釈性」と「精度」がどのように達成されるかを表している。現在のモデリングの技術では、この２つを同時に追い求めることは非常に難しい。たとえば、図１中では、統計における主要な回帰（Linear／Logistic Regression）であれば、モデルがシンプルであるため解釈性は高いが、精度は低くなる。一方、深層学習（Deep Learning、AIモデルを構築するときの中心

図1　モデルの解釈可能性とパフォーマンスのトレードオフ



出所）A. B. Arrieta et al, "Explainable Artificial Intelligence（XAI）: Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", Information Fusion, Vol. 58, p.p 82-115,（2020）Figure12より作成

となるアルゴリズム）では、複雑なモデルが組み上げられるため、解釈性は低くなるが、精度が高くなることが表されている。

また、今後研究が進む中で、各アルゴリズムが獲得していくと思われる範囲（図１の影部）としても、大きく傾向が変わることがないと考えられている。

ここでは、トレードオフとなる「解釈性」と「精度」に注目して、モデルを大きく「理解先行型モデル」と「タスク解決型モデル」の２つに分類した上で、その特徴を紹介したい。

2　現実の本質を捉える「理解先行型モデル」

理解先行型モデルは、現実のある側面の数理的本質を捉えることを目的としたモデルである。さらに細かくいうと、マクロ変数の関係を記述する現象論モデルと、現象の構成要素から組み上げる還元モデルと呼ばれるものに分類される。前者は、統計モデルのように簡単な回帰分析などを利用して、データに内在する関係を発見するために構築され、データサイエンスとしても探索的分析の際に利用される。また、後者は物理モデル、確率モデルなどのように注目する現象を引き起こすのに最低限必要な動的効果と要素を使って、現象を説明しようと構築するものである。

両者は、ともにモデルは簡易なものほどよいと考えられ、結果への寄与が低い要素や効果は削ぎ落としていくことが好まれる。そのため、データや現象の本質のみが残り、結果を理解することが容易になっていく。しかし、このモデルは単純であるため高い精度を上げることはあまり期待できない。このモデルをベースとして、複雑なモデルを構築する際のヒントとし、第一段階のモデリングとして扱われることが多い。

一般に、データをどう分析するかという前提から考えるデータサイエンスに対しては、還元型モデルは事実をモデルに組み込んでいく方法でモデリングが行われるため、なじみの薄いアプローチである。しかし、次章で紹介する例（ウェブサーファの動きや感染者数予測）のように、分析時でデータが存在しないという状況下でのモデリングには、このアプローチが取られることもあり、社会シミュレーションなどの領域では重宝されている。

現実の現象や業務を大まかに再現する数式を作ることに難しさがあるが、箱庭的なシミュレーションが可能となるため、さまざまな仮想的な環境を設定して、多くの知見が得られることが利点である。

3　出力値の精度を重視する「タスク解決型モデル」

タスク解決型モデルは、数値の予測、文章の分類、画像や映像の認識など、ある程度の精度が求められているタスクを解決するときに組まれるモデルである。このモデルは、精度を向上させるための、多くの「テクニック」が存在するとともに、モデル自体も複雑になる傾向がある。複雑とは、入力された変数が何らかの形に変換されたり、2つ以上のモデルの結果が相互に組み合わされて、統合された結果が出力されたり、と意思決定するには困難な表現に代わってしまうことによる。つまり、なぜその結果となるのか理解できないモデル、いわゆるブラックボックスなモデルとして組み上がってしまう。そのため、実際の利用では出力値の精度をもってモデルの結果を信じることが要求される。

しかしながら、それが運用上難しいことも多く、解釈性に問題があり、運用が失敗してしまった事例はいくつか報告されている。たとえば、アメリカのミシガン州フリントにおいて、水道管劣化の予測のために運用していたAIモデルが、隣家同士で水道管工事の要否が異なることを自治体が住民に説明できず、批判を浴び運用が停止されてしまうことがあった[文献1]。

精度の良いAIモデルというだけでは、なかなか受け入れてもらうことは難しく、運用が見送られたり停止されたりしてしまうこともある。複雑なモデルの運用における解釈性の問題は、モデルにおいて非常に大きなテーマとなっている。このため、昨今は精度の良いモデルを構築する技術とともに、構築したモデルに解釈性を与える技術も多数開発されている。

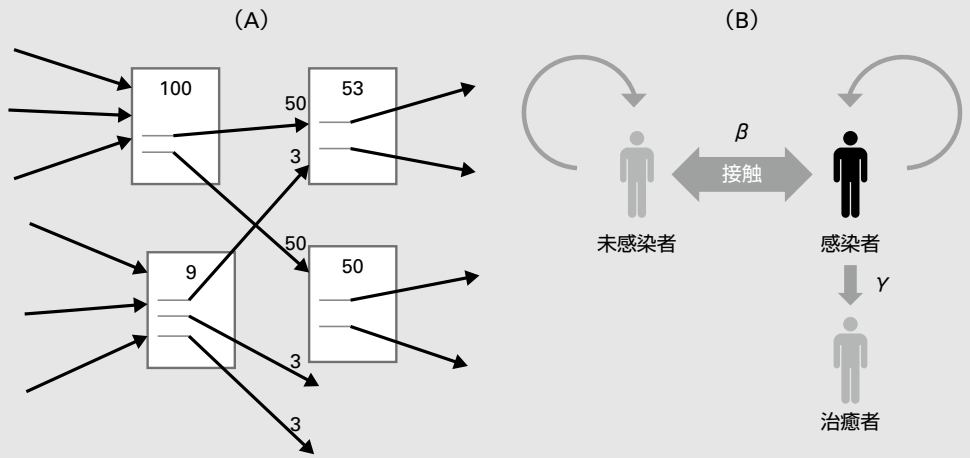
Ⅲ　モデルの活用事例と課題

2つのモデリングアプローチの実課題への応用にはどのようなものがあるか、代表的な利用について挙げてみたい。

1　ネットワークを使った理解先行型モデル「PageRank」

PageRankは、GoogleがWebサイトの重要度を評価するために使っている指標である（ここで紹介するものはGoogle創業当時に提案されたものである）。この指標は、Webサイトのハイパーリンク関係を大規模ネットワークとして捉えて、ネットワーク構造の重要な部分に重要度スコアを付ける手法である。Googleの創業者のセルゲイ・ブリンとラリー・ペイジが1998年に、ウェブサーファがラ

図2　PageRank（A）とSIRモデル（B）のシミュレーションプロセス



出所）（A）L.Page et al." The PageRank Citation Ranking: Bringing Order to the Web.", Technical Report. Stanford InfoLab.（1998）Figure2より作成

ンダムにWebサイトを巡る中で、Webサイトの行き着きやすさ（滞在確率が高い）が、Webサイトの重要度であると考えて発表した[文献2]。

スコアをつけるために、ウェブサーファのWebサイト間の移動を、「リンクをクリックすることによるWebサイト間の遷移」と「自然検索でのWebサイトへの遷移」の2つと考え、モデリングを行っている。図2（A）のように、個々のWebサイト（四角）のウェブサーファがほかのWebサイトへのリンク（矢印）を通して、遷移していく過程でモデル化される。具体的には、あるWebサイトに訪れるウェブサーファが100人いたとしたとき、このウェブサーファが、ハイパーリンク関係でつながる別の2つのWebサイトに50人、50人と当配分で移っていくという仮定である。このようにウェブサーファがWebサイトのネットワーク上を巡るシミュレーションを通して、たどり着きやすさを指標化するのである。

このプロセスを基礎方程式で表すことができて、PageRankはその解として表される。ここでは、式の紹介にとどめるが、ウェブサーファの総数N人に対して、時刻tのWebサイトiのサーファの人数$x_i(t)$は、同Webサイトのハイパーリンクでの接続先数k_iと、Webサイトi,j間の$i \to j$のハイパーリンク関係を表す行列A_{ij}と離脱率rによって次のように再帰的に表される。

$$x_i(t+1) = r\sum_{j=1}^{N}\frac{A_{ji}x_j(t)}{k_j} + (1-r)N$$

この基礎方程式により、「重要なWebサイトからリンクが張られているWebサイトほど重要なWebサイトである」という観点で数値化される。それに加えて、このような基礎方程式を理論的に解析することで、モデルの特徴を詳細に調べることができる。たとえば、多くの被リンクを得ることでPageRankを高めることができるといったSEOでよく聞くテクニックも、この式から導くことがで

きて、現象を表すように作ったモデルを再解析することで、新たな知見を得ることもできる。

PageRankがこのようにウェブサーファの動きの基礎方程式が立てられて構築された背景には、Webサイト間遷移のミクロデータが当時存在しなかったことが大きな理由となる。しかし、技術的進歩によって、Webサイト間遷移をデータとして取得できるようになったことや、Webサイト内のテキストデータが蓄積されていることから、現在のモデルは、このモデルから大きく進展していると考えられる。

PageRankによるビジネスの成功は、説明するまでもないが、このアプローチの功績として、ビッグデータ分析の流行のきっかけとなったことと、これを基点に大規模ネットワーク構造のさまざまなスコア手法や社会シミュレーションのモデルが組み上げられたことにある。たとえば、PageRankを一般化したHyperlink-Induced Topic Search（HITS）[文献3]のように、グラフ構造でのノードのランキングを測る指標はさまざまに展開され、引用関係のネットワーク構造から特許や論文などの書誌を評価する方法として提案されている。また、企業間の売上を企業間取引ネットワークの構造から説明するモデル[文献4]などに応用したり、SNSやブログの口コミ時系列の持つ性質の解析[文献5]にも参照されたりしていて、社会の中で広がっていく現象をモデル化するときに有用とされている。

2 コロナの感染者数予測で脚光を浴びた「SIRモデル」

PageRankと同じ複雑系のモデルとして、昨今の感染者数予測に利用されるW・O・カーマックとA・G・マッケンドリックが1927年に提唱したSIRモデルというモデルがある[文献6]。

総数N人の集団を、未感染者S（Susceptible）、感染者I（Infected）、治癒者R（Recovered、または隔離者や死亡者Removedとも解釈される）のグループに分けて、未感染者Sと感染者Iの接触によって感染が広がり、感染者は一定時間の後に治癒者（死亡者）Rになるという仮定の下、シミュレーションを行うものである。このモデルやその派生モデルは、感染症ダイナミクスを理解するために、新型コロナウイルス感染症パンデミックだけでなく、1905年のボンベイにおけるペスト流行から、最近では2014年の西アフリカ・エボラ出血熱流行などにおいて解析に利用された。

実際に今回の感染者数予測が行われたモデルは、ここで紹介する初期に提案されたモデルとは異なることを注意しておく。

具体的には、各グループの人数は次の方程式に従うと仮定される。

$$\frac{dS(t)}{dt} = -\beta S(t) I(t)$$

$$\frac{dI(t)}{dt} = \beta S(t) I(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

ここで、「接触時に感染する確率β」や「回復率γ」がパラメータとなる。未感染者と感染者が接触したときに感染が起き、感染者数が増えていくことが未感染者数と感染者数の積$S(t)I(t)$で表されている。この積の項が感染者数の増加に寄与する。

モデル中の感染確率βや回復率γのパラメータに、仮想的な値やデータから得られた値を設定して推移をシミュレーションすることができる。また、実際の感染者数のデータが順次手に入ることで、モデルのシミュレーション結果と実データを比較することができるので、実データに合うように、これらの感染確率や回復率のパラメータ値を推定するということもできる。

より現実的なモデルとしていくためには、感染から発症までの潜伏期間を設定したり、人の状態や属性を細かく分けたり、人流の効果などを反映させたりして改良していく。このようにして、感染動向を理解し、感染をコントロールする技術に応用しようと研究が重ねられている。

また、このモデルは人と人との接触によって広がっていくという感染に関する現象に対し、広く扱うことができる汎用的なモデルである。そこで、噂の伝播などの社会活動による情報の拡散現象も同系統の現象として考えられており、SIRモデルの感染確率や回復率といったパラメータなどの各部分を読み替えて利用される。具体的には、災害時のデマの広がりにおいて、「デマ情報に触れ、デマを信じてしまう確率」や「時間とともにデマから覚める確率」として解析[文献7]されている。ほかにも、商品のバイラルマーケティングにおけるインフルエンサーの効果やSNS上の炎上においても、長くこのモデルを使った研究がなされている。特に、人同士の接触などを介して広がっていく現象のモデル化に応用されるため、社会での消費者間のコミュニケーションをモデル化するときに有用と考えられている。

3　理解先行型モデルの課題と有用性

PageRankもSIRモデルも、データの存在しない状況や不足する状況で行われたモデリングとして、注目する対象固有の基礎方程式を作るという過程を経る。このようなアプローチで、データ分析や予測を行う領域として気象、都市工学、環境、回路など、科学的な分野において多岐にわたる。これらは、対象に応じて、個別性の高いモデリングを行っていくことになり、数式を自在に操れるスキルが必要という点で難しさはあるが、シミュレーションや理論解析を通じて得られる知見は非常に多い。一方、マーケティングや流通を中心に扱うデータサイエンスでは、これまで統計学と機械学習を使うことが多かった。

2つのモデリングアプローチは、どちらもデータとデータの背後にある事実を理解する方法であるが、前者のモデリングから得られる知見を通し、データを蓄積し、後者のモデリングに反映していくという方式が採られることが多い。また、機械学習や最適化の手法を使って、前者のモデルのパラメータ推定を精緻に行っていく方法などもあり、使い分けは重要となる。

4　タスク解決型モデルでは「解釈性の付与」がポイント

タスク解決型モデルは先に述べた通り、データが一定程度存在し、分析手法が先にあるときに選択されるアプローチである。個別性の高い立式を行っていく理解先行型モデルとは異なり、アルゴリズムにデータを当てはめていく（学習させる）ことで精度を高めていく。一方で、先に述べたように構築したモデルは複雑になる傾向があり、解釈性の低い結

果をどう提示していくかは、運用の上ではよく課題となる。そのため、モデルの解釈性を高める技術が、昨今、多数提案されるようになってきている。この複雑なモデルの解釈性を高めていく技術をどのようにして使いこなしていくかについて、仮想的なケースを例にして説明したい。

まず、ある保険契約者が満期を迎えるときに、契約更新を行ってくれるかどうかを判定するモデルを作るとしよう。ここでは話を簡単にするために、契約の更新を月額の保険料と営業員とのコンタクト回数のような変数で説明することを考える。

一般的な分析方法として、回帰分析の標準的な手法であるロジスティック回帰のような線形分類を使ったとき、契約者の「契約更新●」と「契約非更新○」を判定するモデルは、図3（A）のような直線の境界線をもって判別することになる。また、図3の影部は、モデルによってその領域内にあるデータが契約更新と判定される領域となる。

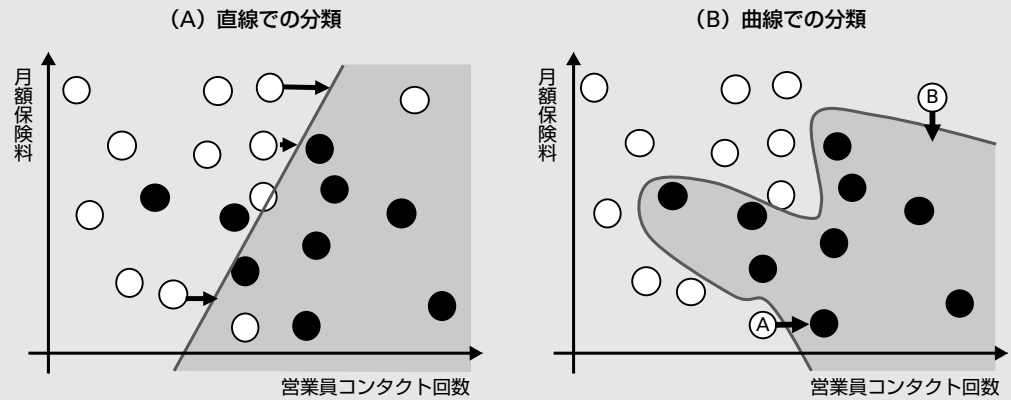
線形分類の結果だと、結果に寄与する各説明変数の重み（回帰係数）を知ることができて、解釈は非常にシンプルになる。図3（A）では、契約更新と判定される領域に契約非更新の契約者を近づけるためには、契約者全体に対して営業員コンタクト回数を増やしたり月額保険料を減額したりする方策が有効であることと、またそのどちらがより有効であるかを知ることができる。

このようなアプローチが好まれるのは、各変数の業務改善につながる方向が明瞭で、どのようにアプローチをとればよいか、ネクストアクションにつなげやすいことにある。ただし、この線形分類のモデルは、顧客全体への改善施策として総論となる結論しか得られないことと、単純が故に精度が期待できず、誤判定が多く出る傾向があることに不足を感じることも多い。

一方で、曲線や直線の組み合わせでこの分類タスクを行う高度な方法（非線形分類）では、分類が図3（B）のように（少し極端だが）精度よくできることが多い。しかし、先ほどの直線のモデルと比較して違うのは、分類の結果として、次にどうすればよいのかという方向性が読み取りにくく不明瞭であることである。

このようなモデリングのときに、モデルを読み解くために解釈性を与える必要がある。その方法としてLocal Interpretable Model-ag-

図3　仮想的な事例に対する分類モデル適用の様子



nostic Explainations（LIME）[文献8]、Shapley Additive Explanations（SHAP）[文献9]などの手法が存在し、ライブラリがオープンソースとして公開されている。SHAPはLIMEを一般化したもので、どちらもモデルの出した個々の結果に対して、各説明変数がどの程度の寄与しているかを明らかにすることができる。

たとえば、作成した複雑なモデルに、これらの解釈性を付与する技術を適用することで、より細かく改善施策を示すことができるようになる。

図３（B）からは、契約者Aを表す点Aでは営業員コンタクト回数を増やす方向が、B点では月額支払料を下げる方向が、契約更新確率を高めるための最も良い方法であることが分かる。前述の単純なモデルでは、総論としての改善の方向感のみが示されたため、契約者Bに対する営業員コンタクト回数の増加は、改善につながらないことが予測されるが、今回の使い方で個々の契約者に対して、どのような方策が必要であるかの方向を、より細かく示すことができるようになる。

このように、もともとのモデルは複雑性の高いものとして構築したが、解釈性を付与する技術を適切に利用することで、予測精度をある程度高めた上で、なおかつ個々のデータ点（＝契約者）に対して、個別に改善する方向を提示することができるようになる。

図解をすると非常に簡単な説明になるが、多数の説明変数が存在すると、このような発見を行うことは一般に難しくなる。複雑なモデルをそのまま利用することは意思決定が難しくなるため、実際のプロジェクトでは避けられることが多いが、これらの技術をうまく使いこなすことで、改善施策の提案に個別性を持たせることができ、詳細に方策の方針を決めることができる。より実務的にも処方的分析に役立つものと期待される。

Ⅳ 最後に（モデリングで求められること）

本論考で、モデルという言葉をその役割と具体的な例を用いて説明した。

モデルのタイプとして、理解先行型モデルとタスク解決型モデルのそれぞれのアプローチがある。どちらも、構築されるモデルは現象の一部を端的に表現するという点は共通である。

理解先行型モデルは、ドメイン知識を持って固有の立式が必要であるが、仮想的な設定でのシミュレーションができ、データ外の知見を得られること、タスク解決型モデルは精度を高められるが内部のアルゴリズムによって、複雑な変換がなされてしまうことで、解釈性が課題となってくることを述べた。

2つのモデルは、データの質や量に応じてモデリング技法を変え、モデルが求められている環境に応じてその特徴を活かす方法を選択する必要がある。また、適用業務や適用業種は異なっても、モデリングの上では同系統の問題を扱うことがある。本論考で紹介したPageRankとSIRモデルのように、本来の設計された領域での利用を超えて、社会調査や経済モデリングのほか、マーケティングや流通など他分野に応用されたものもある。モデルを応用・調整する技術とドメイン知識があれば、広く他業種への展開も可能となる。

このようなモデリング技術への注目は、技術的なトレンドに加えて、昨今の社会変化に

よってもたらされたところが大きく、今後も、コロナ禍からの回復に伴う社会変化の中でニーズが高まっていくと考えられる。実際にデータサイエンスの大学教育の中で、社会シミュレーションの技術が見直され、講義も増えてきていて、今後のデータサイエンティストが身に付けるスキルの一つとなっていくことが予想される。それによって、われわれデータサイエンティストが、データの少ない場でも活躍できるようになるかもしれないし、モデリングを通して、知見や必要なデータを明らかにするという進め方も増えていくと思われる。

昨今行われたモデリング技術を振り返ることで、データサイエンスによる意思決定の幅は大きく広がることが期待される。

参考文献

1 https://www.theatlantic.com/technology/archive/2019/01/how-machine-learning-found-flints-lead-pipes/578692/

2 L.Page et al. “The PageRank Citation Ranking: Bringing Order to the Web.”, Technical Report. Stanford InfoLab.（1998）

3 J.M.Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, Journal of the ACM. 46 (5), (1999)

4 K. Tamura et al., “Diffusion-localization transition caused by nonlinear transport on complex networks”, Scientific reports, Vol 8 (1), (2018)

5 Y. Sano et al, “ソーシャルメディアの書き込み数における揺らぎ”, 人工知能学会全国大会（第29回）, 2015

6 W. O. Kermack and A. G. McKendrick, “A Contribution to the Mathematical Theory of Epidemics”. Proc. Roy. Soc. of London. Series A 115 (772), (1927)

7 M.Takayasu et al, “Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study”, PLoS ONE 10 (4): e0121443, (2015)

8 M. T. Ribeiro et al, “Why Should I Trust You?” Explaining the Predictions of Any Classifier”, . In Proc. of the 22nd ACM SIGKDD. ACM, 1135–1144, (2016)

9 S. M. Lundberg and Su-In Lee, “A Unified Approach to Interpreting Model Predictions”, NIPS 2017

著者

田村光太郎（たむらこうたろう）
野村総合研究所（NRI）データサイエンスラボ主任データサイエンティスト
NRI認定データサイエンティスト、博士（理学）
専門は複雑系物理学、データサイエンスなど