

きらり。中のヒト

第18回



グラフに始まり グラフに終わる？ とある生命科学系 データサイエンティスト の日々 山口敦子[†]

データサイエンスとは？

みなさん、はじめまして。東京都市大学でデータサイエンスを教えている山口と申します。「データサイエンス」という言葉が世の中に出てきてずいぶんたちます。データサイエンスを専門とする学部や学科もどんどん出てきています。でも、データサイエンスっていったい何でしょう。

データサイエンスを専門としている方に話を聞いても、バラバラな答えが返ってきます。どこで何をどうするものか、結構ぼんやりしているように思えます。データサイエンス系の学部のカリキュラムを見ても、大学によってバラバラで、データ解析に力を入れている大学がある一方、データを使った問いや仮説形成に力を入れている大学などさまざまです。

ちなみに、私が講義などでデータサイエンスを説明する時には「データに基づいて、科学のおよび社会的に有益な知見を引き出そうとするアプローチ・技術」という言い方をしています。この言い方も我ながら、具体性に欠けイメージしにくいですよ。そこで、実際私がどんな研究をどのようにやっているか、そして、どうしてこのような研究をするようになったのか、経歴を含め、紹介していきたいと思います。

「グラフ」との出会い

私は元々数学科の出身です。しかも最初は純粋数学を志していました。高校時代から「大学への数学」という雑誌を購読し、美しい数学の理論を構築することを夢見て大学に入りました。ところが、大学で学ぶ数学は確かに美しく楽しいのですが、すでに完成されていて、私如きが付け入る隙がありません。卒論の時期になっても、純粋数学の分厚い本を読んで理解するのがいっぱい。修士課程に行っても新しいことをできる気がしませんでした。そのころ、一人の先輩が悩む私を情報科学の研究室に連れて行ってくれました。そこでは、みんな新しいことを見つけようとする情熱でいきいきしていました。そして、その光景に感動した私は純粋数学の道を諦め、情報科学の道を歩き始めたのでした。

修士課程では、人工知能とアルゴリズムの研究室に入ったのですが、なかでも、グラフ理論とグラフアルゴリズムに興味を持ちました。修士課程を修了する頃にはグラフ理論とアルゴリズムの研究をもっとやりたいと思っていました。幸い、ある企業の研究所でグラフアルゴリズムの研究をつづけることができることになりました。ただし、企業なので何か製品に結びつく必要があります。そこで、LSIの最適な配置を計算する問題をグラフに変換して解くアルゴリズムを考えたり、要求仕様書をグラフで表現して手戻りを減らしたり、グラフを見やすく表示するアル

ゴリズムを研究していました。こうして私は、企業の研究所におけるグラフアルゴリズムの研究からキャリアを始めました。

生命科学とデータサイエンス

そのころ、私が修士の時にいた情報科学の研究室が生命科学に注力し始めました。バイオインフォマティクスという言葉が始めた頃の話です。そして、私がいた会社もバイオインフォマティクスを始めようとしていました。修士の時にいた研究室が生命科学ですっかり有名になってしまったため、「君、〇〇研究室の出身だね」と言われ、私はバイオインフォマティクスの勉強も兼ねて、生命科学の研究所へ出向することになりました。受験は物理と化学を選択しましたし、大学に入ってから生物は避けていたので、最初はついていくのが本当に大変でした。

さて、生命科学では一部の例外を除き、あまり数式などの理論に頼りません。実験などで取ったデータがすべてという世界です。しかも生命科学データは計測機器が発展するにつれ、爆発的に大きくなっていきます。これらのデータを使って新しい知識を引き出そうとするアプローチが「バイオインフォマティクス」と呼ばれる、当時としては新しい学問だったわけです。今考えると、当時からまさに「データサイエンス」をやってたんだなあと思います。

[†] 東京都市大学
"From Graphs to Graphs? The Journey of a Certain Life Science Data Scientist" by Atsuko Yamaguchi (Tokyo City University, Tokyo)

ウェットとドライ

人によっては、他人が取ったデータをあまり信用せず、自分で同じようなデータを取り直す人もいます。お金の無駄遣いじゃないのか、とその当時は思っていました。実は意味があることを後に知りました。この論文¹⁾によると、1,576人の研究者を対象として調査をした結果、70%が他の科学者の実験を再現しようとして失敗した経験があり、半数以上が「自分の」実験の再現に失敗したことがあるそうです。特にがん研究の再現率²⁾はかなり低く、他の科学者の実験は11%しか再現できなかったそうです。これでは、他の人のデータを鵜呑みにできない気持ちもわかります。再度データを取り直したくなるでしょう。

そして、そうして実験室で細胞や生き物を用い、生命科学の実験をしてデータを生産する研究者の方々(バイオインフォマティクス業界では、このような研究者をウェット研究者と呼びます。生物の実験では水を使うからでしょうか?)から提供されたデータを用いて、私は計算機上でいろんな解析をしてきました。私のように生命科学実験を直接行うことなく、他の人が生産したデータを用いて知識を引き出す研究者はドライ研究者とよべれます。ちなみに、私は典型的なドライ研究者ですが、データを理解するためにはウェット研究者の実験の現場を知らなくては、と思い、ウェット研究者から手順を教えてもらい、大腸菌を使った実験を行ったことがあります。しかし、手順通りにやったのに大腸菌は全滅し(一緒に同じ実験をしたウェット研究者の大腸菌は元気そうでした、ちょっと悔しかったです)、自分にウェットの才能がないことをまざまざと思い知らされました。そして、ドライ研究者として、現在の言い方では、データサイエンティストとして、生きていくことを強く心に決めたのでした。そして、今、生命科学の研究をしながら、データサイエンスを教える教員として大学

にいます。

映像画像データの重要性と将来性

さて、これは「映像情報メディア学会」の原稿なので、少し画像や映像に関する話題も出しておきたいと思います。小児科の病気、特に遺伝性の病気の子供の顔には、病気ごとに独特な共通する特徴があります。例えば、有名なところでは、ダウン症(21番の染色体が通常2本のところ、3本で生まれてくる病気)の子供の顔立ちには、目がつり上がり気味という特徴があります。病気ごとに顔の特徴が違うので、小児科の医師は、子供の顔を見て病気を推測していたそうです。ある有名な小児科の医師は、たくさんの子供の病気の顔の特徴を覚えていて、他の病院では診断がつかなかった子供の病気を、その子供の顔から診断していたという話を聞いたことがあります。ちなみに、今では子供の顔写真を入力すると病気の候補が出てくるソフトウェアがあります³⁾。

また、CTやMRIをはじめとした画像によって診断がつく場合も少なくありません。最近は機械学習によって、医療画像を読む研究が盛んに行われており、がん細胞の発見など、いくつかのタスクではすでに人の能力を超えていると言われます。近年は動画も、運動障害を記述する場合などに使われることが普通になってきたように思います。文章で「右足に歩行障害」と書くより、実際の歩行を動画で撮る方が、より正確かつ詳細に歩行の状態を記録できます。実際、医療系の雑誌には、動画を投稿できるJournalが出てきています。

最近の私の興味：希少疾患

さて、現在、私は「希少疾患」に興味を持って研究をすすめています。希少疾患というのは、患者数が少ない病気のことです。患者数が少ないので、その病気を知らない医師が多く、診断までに何年もかかってしまい、その結果、

治療開始が遅れたり、補助をもらうことが遅れたりします。また、患者数が少ないので、治療方法が確立されておらず、薬の開発も売れる数が少ないので進まないことが多いです。希少疾患の診断には、全部の遺伝子を調べる解析や、遺伝子ではない領域も含めたゲノム全体の解析が主流です。ただ、全部の遺伝子を調べて病気が特定できる割合は30%~40%、ゲノム全体を調べても50%程度に留まっています。私は遺伝情報に加え、何か他のデータを加えなければ、これ以上数字は上がらない可能性があると思っています。

そういう意味では、近年、生成AIが話題ですが、マルチモーダルな生成AIは医療データに親和性が高そうだと感じています。医療データは、医師が書く電子カルテの文章を中心として、画像や動画も含め、検査結果のさまざまな形のデータが紐づいています。そして、マルチモーダルな生成AIを用いて、電子カルテと関連情報を読み込んで、患者を分類し、病気の候補を挙げて診断支援を行うなど、希少疾患に対してさまざまな可能性があると思っています。しかし、生成AIは基本的にブラックボックスのため、なぜそうなのか説明できないという問題があります。医師から「なぜかわからないですが、あなたはこの病気だと生成AIが言っています」と言われても困りますよね。

そして「知識グラフ」へ

私は今、説明可能な人工知能ということで、知識グラフの利用を考えています。知識グラフとは、データをグラフの形でかいたものです。グラフ構造なので、表形式より、柔軟にスキーマを定義できます。また、多様性が高いデータの場合、表形式だと疎なデータとなってしまう。一方、グラフ構造であれば、データがある部分だけを記述すれば良いので、利用しやすいです。また、グラフ構造なので結合もやりやすいです。同じものを指す頂点を重ねることで複数のグラフを結合し、一つのグラフにすることが容易です。そして、さまざまなデータから一

つの知識グラフを作り、それと機械学習を組み合わせてみます。具体的には、機械学習の結果を知識グラフ上の構造と照らし合わせて、結果を説明できないかと考えています。一方で、グラフの問題は計算時間と計算領域を多く取る問題が少なくありません。少しでも速いアルゴリズムを今後も考えていきたいです。

また、さまざまなデータを知識グラフにしていくことで、データの再利用

性があがります。知識グラフでデータを提供することで、データをきれいにしつつ一箇所で管理できます。そこで、データを知識グラフにして提供するサービスの開発も行っています。

そういうわけで、私の研究キャリアは、グラフで始まりましたが、一周回って、グラフで終わりそうです。

(2024年4月4日受付)

【文 献】

- 1) M. Baker: "1,500 scientists lift the lid on reproducibility", Nature, 533, pp.452-454 (2016)
- 2) C.G. Begley and L.M. Ellis: "Raise standards for preclinical cancer research", Nature, 483, pp.531-533 (2012)
- 3) Face2 Gene, <https://www.face2gene.com/>



未来へ向かって

宣言

私たち映像情報メディア学会は、将来の映像情報メディア社会の未来を描き、その社会実現に向けてなができるかを議論して、実行していきます。

夢のある未来の実現に向けて――

未来ビジョン

- 映像情報メディア社会の未来（10年後、30年後の社会）
 - ・ディスプレイはペーパー？カメラはどうなる？
 - ・テレビ番組は？
- 学会の未来ビジョン
 - ・夢の未来社会を現実のものにするための研究・開発とはなにか
 - ・それを実現していくための学会施策とは

ランドデザインに向けて

- 社会貢献活動の推進
- 学術新領域の創出
- 会員へのサービス向上
- 活動のグローバル化
- 国内外標準規格開発への参画

マスタープラン

- 社会貢献制度の拡充と学会内外へのアピール
- Webサービスの充実
- 長期的な財政計画、学会財政安定化策
- 他学会との連携・協調

映像情報に特化した通信・放送技術、映像メディアに関する技術について、会員間の学術・技術交流を図るとともに、この分野の学術・技術を広く社会に発信し、社会情勢の変化に対応して国内外の学術・技術発展に主導的役割を果たしていきます。