

「なぜ今データサイエンスなのか～AI技術とのシナジーが もたらす進化と展望～」に関する報告

植田 和憲

東北公益文科大学総合研究論集第49号 抜刷

2025年2月21日発行

シンポジウム等記録

「なぜ今データサイエンスなのか～AI技術とのシナジーがもたらす進化と展望～」に関する報告

植田 和憲

1. はじめに

2024年（令和6年）11月18日、および11月29日、東北公益文科大学鶴岡キャンパスにおいて、「なぜ今データサイエンスなのか～AI技術とのシナジーがもたらす進化と展望～」が開催された。本企画は、2024年度（令和6年度）鶴岡市課題解決事業情報科学連続講座による事業の一環である。この連続講座は全三回開催され、本稿ではその一回目と二回目について報告する。

一回目の講演は、11月18日、高知工科大学情報学群・データ&イノベーション学群の吉田真一教授より「統計モデル・深層学習の基礎とデータ活用の実際」（講演時間90分：午後6時30分～8時）と題して行われた。二回目の講演は、11月29日、高知工科大学情報学群・データ&イノベーション学群の吉田真一教授および東北公益文科大学公益学部公益学科メディア情報コースのノヴァコフスキ・カロール講師より「データ分析・AIモデル活用の実際」（講演時間90分：午後6時30分～8時）と題して行われた。

これらの講演は、IoTやAI関連技術の進化が目覚ましい今日、収集され処理される膨大なデータが存在することを背景として、インターネットに蓄えられたそのようなデータがどのように扱われ活用されているのかについての基礎知識及び研究等から得られた知見について解説し、将来の展望について考える機会を提供することを目的として企画された。講師には、高知工科大学でデータ&イノベーション学群の設置に関わり、現在もデータサイエンスに関わる教育や研究に携わる吉田真一教授と東北公益文科大学にて自然言語処理を専門とし資料となるデータが非常に限られる危機言語に関する研究を行っているノヴァコフスキ・カロール講師を招いた。

一回目の講演では、データサイエンスの基礎となる統計モデル、および機械

学習、深層学習についての導入となる解説、および、それらの手法の応用となるデータ活用事例についての説明が行われた。具体的な事例として、高知県でのユズの栽培現場における画像処理を応用した収穫量予測についての成果が紹介された。

二回目の講演では、データを活用した研究に焦点を当てた講演が行われた。吉田教授からは、一回目の講演で紹介されたユズの栽培現場での事例についてさらに詳細な紹介があり、その研究でデータをどのように扱うかについてさまざまなアプローチおよびその有効性について説明された。ノヴァコフスキ講師からは、危機言語の記録保存のための音声認識技術に関する研究の成果が紹介された。

以下に、「なぜ今データサイエンスなのか～AI技術とのシナジーがもたらす進化と展望～」の講演の内容について報告する。

2. 統計モデル・深層学習の基礎とデータ活用の実際

はじめに、データサイエンス、統計モデル、機械学習の概要を示す。データサイエンスは、統計学、計算機科学（ここでの計算機とはコンピュータのことを指す）、場合により数学（統計学以外のもの）を要素とするものと考えることができる。統計モデルとは、実世界のデータに基づき、データがどのように生成されるかを模式的に表現したものあるいはメカニズムを表したものである。機械学習とは、コンピュータで多くのデータを処理し、それら同士の関係性を導き出すデータサイエンスの一領域である。関係性の導出によって、対象物の認識、現象等の理解、未来や過去に加え現在に関する予測を行うことが期待できる。深層学習（ディープラーニング）やニューラルネットワーク（人工神経回路網）も機械学習に含まれる。

データサイエンスは、データから知識を得る、価値を創出する科学と考えることができる。データサイエンス関連の要素技術は、従来活用されてきたものであり新規の学問領域ではないと考えられるが、大規模なデータ収集や高速な処理技術を前提とすることができるようになったことに伴い、これまで対象とすることが難しかった膨大なデータや処理することが難しかった複雑なモデル

を取り扱うことができるようになったことで注目された。それらの対象として、前述の深層学習やニューラルネットワークなどが含まれており、この領域におけるブレイクスルーにつながっている。

統計モデルは、前述したようにモデルを構築することによって実世界でデータが得られるメカニズムを利用して、対象や現象を理解することを可能とするものである。モデルの構築手法として数理モデルを採用することにより、コンピュータ上で現実世界において発生する現象の再現や予測がある程度可能となる。統計モデルでは、現実世界での事象の発生を確率分布に基づく数式で記述する。このモデルは不確実性を含んでおり、具体的な例として、確率で移動する方向を決定するランダムウォークや経済指標の動きや価格変動などの経済・金融モデルなどがある。

回帰分析とは、統計モデルのひとつである回帰モデルを用いて、特定の要素が他の要素に与える影響を説明、予測する手法である。特定の要素は複数の場合も考えることができ、一つの場合は単回帰、複数の場合は重回帰と呼ばれる。関係を一次式で表す場合のモデルは線形単回帰モデルと呼ぶ。線形単回帰モデルの式を以下に示す。

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

ここで、 x を説明変数、 y を被説明変数、 β を回帰係数と呼ぶ。この式は、 x が y に与える影響の度合いを表しており、 y が x によって説明される、と表現する。回帰分析では、この式における回帰係数 β を求めることを目的としており、データから変数同士の関係性を見出す方法と考えることができる。機械学習も目的や手法は同様であり、統計学では β を推定 (estimation) し、機械学習では単回帰モデルのあてはめ (fitting) を行う。

機械学習は、コンピュータによる学習ととらえられる。機械学習では、入力としてデータを、出力としてデータの内容や予測をそれぞれ与え、入力が出力に変わるまでの過程である処理を規則や関数として学習する。機械学習の応用例として、人の手によって描かれた数字の画像と対応する数値を与え、数字が描かれていると思われる未知の画像に対してその数値を判定する、というもの

がある。この例では、学習のために用いられるデータは数千から数万あり、十分なテストを経て実用化される。

機械学習には、解答が与えられる教師付き学習とそうでない教師なし学習がある。教師として与えられるものが質的なデータである場合は分類、量的なデータである場合は回帰と呼ばれる。質的な教師データの例としては、グループ分けにおけるグループが挙げられる。教師なし学習では大まかな分類や似た要素の統合などが行われ、大まかに分類する手法を特にクラスタリング、似た要素の統合は次元削減とそれぞれ呼ぶことがある。

機械学習では、複雑なデータから複雑な数式モデルを導出できる場合があり、直線のような単純な数式に比べ性能が向上する期待がある。しかし一方で、汎化能力という一般的、普遍的に予測や判定を行う能力という概念があり、本当の適切な式に比べてモデルや数式が複雑すぎると過去のデータに偏りすぎてしまい、実用的でなくなるという側面もある。この複雑すぎるモデルを構築してしまうことを過学習と呼ぶ。

ここで実際の例として、吉田教授の行っている熟練農業者の知識をAI化するという試みについての解説も行われた。高知県で盛んに栽培されているユズを対象として、ユズの収穫量を予測する、あるいは収穫量に影響する枝の剪定に活用する、というものである。枝の剪定はいわゆる熟練の技、すなわち新規参入者や経験の浅い従事者には難しい作業である、として熟練者の目を対象としてAI技術を適用することが主な内容である。成果の説明において、機械学習を行った結果についての解説や過学習の例などが示された。この研究内容については二回目の講演にて詳細な解説が行われるため、ここではこの程度にとどめる。

最先端のAI関連技術としてニューラルネットワークおよび深層学習についても解説が行われた。ニューラルネットワークは、人の脳の神経細胞を模したニューロンによって構成されるものであり、ニューロン同士が接続され他のニューロンに電気信号が送られるモデルである。ニューラルネットワークの考え方に基づく計算モデルや学習モデルが1940年代以降提案されてきた。これまで、ニューラルネットワークとしてさまざまなモデル等が提案されており、特徴的な理論や手法が登場すると一時的に活発に研究や開発が行われるという

歴史を繰り返してきたが、現在の深層学習が広まることによってふたたびブームが起きている状況となり現在でもそれが続いている。深層学習では、ニューラルネットの階層を大きく（3層以上）するが、階層が増えると処理が膨大になるため、最前層から事前学習を行わせ段階的に学習を行わせることによって効率的な学習を実現する。

ニューラルネットワークおよび深層学習については、階層数の増加といった処理時間の増大につながる傾向に基づいて状況が変化しており、こういった処理を行うためのハードウェアとしてGPUを活用することが一般化しているが、それでも処理リソースが不足する場合があります、クラウドのリソース、AI用のスーパーコンピュータや共同研究機関の利用が必要な状況になっている。

最後に、吉田教授の行っている研究から、同じくユズの栽培における着果状況を画像認識によって推定するという内容について紹介があった。この研究では、ユズの着果数を推定することによって生育状況の把握を自動化し、収量予測や出荷予測を行うことによってユズが最も売れる年末の時期における予約の受付を最適化することを目的としている。高知県では、IoP（Internet of Plants）の取り組みとして農業を取り巻く少子化や高齢化への対応をAIやIoT関連技術を活用したスマート農業によって行おうとしていることも紹介された。

3. データ分析・AIモデル活用の実際

この回は、高知工科大学の吉田教授より「高知でのデータ活用・AIの活用」と題して、東北公益文科大学のノヴァコフスキ講師より「データ分析・AIモデル活用の実際～危機言語の言語処理技術に着目して～」と題してそれぞれ講演があった。以下にて、それぞれの講演の概要を説明する。

3.1 高知でのデータ活用・AIの活用

データサイエンスという言葉が広く用いられ、注目を集めている昨今であるが、実際の課題に対するアプローチを検討し、解決策を提示するということを考えたとき、得られたデータから知識や価値を創出する方法等に目が行きがち

である。しかしながら実際には、データを収集する方法やそれを転送する方法、モデル化や学習に関する手法が基盤技術である。高知工科大学データ&イノベーション学群では、都市交通、金融経済、ヘルスケア、農業といった分野で実際に使用されているシステムを踏まえ、総合的なデータサイエンス技術について推進していくための活動を行っている。

吉田教授は、一回目の講演でも紹介した高知県におけるデータサイエンスと農業とを組み合わせたIoT事業の一環としても研究を行っている。収集されたデータに基づいた分析やそれによって得られた情報や知識に基づいた栽培・経営を行うことにより、作業の効率化や収量の増加といった効果が期待できることに加え、農業従事者に時間的な余裕についても提供できる可能性がある。なお後者に関しては、農業においては長期間の放置あるいは作業等からの離脱が難しいという側面があることからきている。

ユズの栽培に関する問題として、熟練した農業従事者による適切な剪定が必要であることと予約相対取引という前もって取引を行う方式であることが挙げられる。そのため、剪定の参考になる知見を得ることやある程度の精度で事前に収量予測ができることが求められる。前者に対する解決方法として、葉面積指数という土地面積あたりの葉の片面の総面積を算出する方法を採用した。後者に対する解決方法として、スマートホンやドローンなどを使って撮影された画像を解析して果実の収量を予測する方法を採用した。画像の解析に当たっては、人間から見て不自然なものと判断される程度の合成画像を使い学習に用いることで画像データの不足を補う方法を採用し、結果として性能が向上したと結論付けた。この成果は、基礎となるデータの不足が問題となる場合にこの手法が適用できる可能性を示唆したといえる。

また、脳ドックによって撮影されたMRIのデータと高齢者の運転能力との相関についての研究や胸部X線画像の診断支援に関する研究についても紹介された。後者はAIによる判定結果に対する理由説明の可能性を追求したもので、そのために着目した部分や複数の画像の差分を示す。この研究は、AIとは何であるのかや知能の仕組みについての命題にもつながるものである。

結びとして、さまざまな分野で人手不足が深刻な状況であり、その一助となる成果が求められている、人のリソースをAIやデータサイエンスの範疇では

ない領域に集中すべきである、との考えが示された。また、データモデルを適切に構築し状況等が判断できるようになってきており、データサイエンスがそのことに寄与できる、とした。

3.2 データ分析・AIモデル活用の実際～危機言語の言語処理技術に着目して～

自然言語処理とは、人間が話すことばをコンピュータによって処理することに基づいた技術である。自然言語処理の応用技術としては、スペルチェック、機械翻訳、文章校正、音声認識・合成、対話型AIなどが挙げられる。ただしこれらの応用例は、膨大なテキストあるいは音声データを前提として発展しており、アイヌ語のような基礎をなすデータが少ない「低資源言語」に対して適用するには多くの課題がある。

世界には絶滅の危機に瀕している言語が多いが、言語は貴重な無形文化財であり記録して保存することは後世のために重要であり急務である。そのために自然言語処理技術が果たせる役割に対して大きな期待がある。しかし、前述のように危機言語が低資源言語でもある場合、資源のデータ化について課題がある。その一つとして、文字起こし、すなわち音声データのテキストデータ化があり、人間が作業を行う場合に非常に時間がかかることが、データ化の促進を阻害する結果につながる。

そのため、音声認識技術を適用した文字起こし作業の効率化、あるいは言語研究における音声資料解析の効率化を目的として研究を進めている。対象となる言語は樺太アイヌ語であり、研究プロジェクト「言語の記録保存のための音声認識技術の研究：樺太アイヌ語に着目して」を推進している。対象となるデータは樺太アイヌ語の専門家によって1960年から1970年代に渡って録音された20時間以上の音声データである。

次に、言語処理技術の最新動向について解説する。前述したように、今日の自然言語処理技術を応用した各種のシステムは大量のデータを背景として機械学習を行った成果が活用されている。機械学習における自然言語処理では、音声認識に対して実用的な精度を達成するためには1,000時間以上の教師データが必要であるが、低資源言語・危機言語ではこの点が非常に大きな問題である。

機械学習やディープラーニングに関しては、よりよい結果を得るための手法を採用するよりも多くのデータを使用するのが効果的であるという報告がある。

自己教師学習という手法では、段階が二つに分かれ、まず膨大なデータを用いて事前学習済みモデルを構築し、続いて、少量の教師データを使ったファインチューニングと呼ばれる微調整を行う。また、マルチリンガル事前学習という手法があり、これは共通する特徴や語彙を持つ言語のデータを用いることで、それぞれの言語を単独で学習するよりも良い結果が得られるというものである。研究プロジェクトでは、この手法をアイヌ語に適用させる方法について検討を進めている。

研究方法として、53か国語の音声データを使って事前学習されたモデル、新規言語であるアイヌ語の比較的小規模の200時間の音声データを用いた自己教師学習を継続する実験、および樺太アイヌ語に加え、同じ語群に属する北海道アイヌ語、音韻関係の類似度が比較的高い日本語、関係のない英語を用いたファインチューニング、とがある。結果として、前者の結果より、新規言語（アイヌ語）の音声データを用いた継続事前学習が効果的であったこと、系統的に近い北海道アイヌ語をファインチューニングの際に追加することでより良い結果が得られたことがそれぞれ示された。また考察として、教師データにおいて語と語とを離して表記する分かち書きにおける同一の語に対する表記の違いである揺れが影響したこと、専門家の表記と比較して機械による結果が正確な場合があった、ことが説明された。

4. おわりに

いずれの回においても、一般的で背景や将来展望にとって重要なものから、専門的で知識のある聴衆にとって興味深いものまで、幅広く質問やコメントが出され、充実した質疑となった。一部を紹介すると、データ利用にあたっての当事者の許諾はどのようになっているか、データを持っていて活用したい人がそれを扱える専門家をどのように探せばよいか、AI関連で使用されるGPU (Graphical processing unit) と比較してNPU (Neural processing unit) にはどのような特徴があるか、などの質問が出され、それらに対して各講師からの

回答が得られた。

また、連続講座の一回目および二回目のテーマは「なぜ今データサイエンスなのか」とした。これは、データサイエンスの基礎をなす知識や解析手法は従来使用されてきたものであり、また、データから意味のある情報を抽出して活用することは以前よりデータマイニング技術として知られており、新しく発見された、あるいは提案されたものではないためである。この問いに対して、一回目の講座では吉田教授から、二回目の講座ではノヴァコフスキ講師から、それぞれ回答を得たので紹介する。

吉田教授からは、人手不足が深刻な状況であり、そのためにデータサイエンスやAIに関する技術を活用することが求められる、との回答を得た。紹介された研究結果や応用例は、専門的に培われた技術や知識を人間の代わりに再現可能とするものや視覚化・言語化されていない知見を具現化するものを含んでおり、述べられている目的に合致する。世間にはAI技術の発展により仕事が奪われるという懸念があるものと思われるが、それはAI技術によって人の代わりとなるものを生み出せると思われているということである。もし、そうであるとするならば、人手不足という現状に対する解決方法のひとつとして考えることができる。

ノヴァコフスキ講師からは、データは石油みたいなのである、との回答を得た。石油にはさまざまな成分が含まれており、精製過程を経てガソリン、灯油、軽油といった燃料や石油を原料とする製品などが製造される。データにもそのような側面があり、収集されたままのデータではなく、それを適切に選別、分析、解析等を行うことによって得られたデータを各種の目的に沿って活用する。電子デバイスに関する技術、ネットワーク技術などの発展により、今日では日々膨大で多様なデータが計測、収集されており、それらを有効に活用するために必要なデータサイエンスの重要性が高まっているといえる。

前述の電子デバイス技術やネットワーク技術に限らず、コンピュータ関連技術は発展を続けており、膨大なデータを集めることが可能になったのと同時に、膨大なデータを処理することも可能となった。従来は処理能力の限界から断念した解析手法が、今日では現実的な選択肢のひとつとなることもあり、データ処理、あるいはデータサイエンスを取り巻く状況も変化している。近年（執筆

時、2024年までの数年）ではデータサイエンスは情報分野の一種の流行となっており、新しい学部を設置や教育プログラムの認定などが盛んに行われている。このような動きがどれほど継続するかは不透明であるが、従来からデータ解析手法が基礎的なものとして活用されてきたように、今後もデータサイエンスは重要な技術分野として扱われるものと考ええる。

本講座は全三回の開催であり、第三回は公開され広く利用できるオープンデータの活用に関する総説および演習、具体的な生成AI技術でありテキスト生成AIとして広く知られるChatGPTに関する基本的な解説、を予定している。その報告は別途行う予定である。本講座の開講が、本学におけるデータサイエンスを活用した研究や教育の一助になれば幸いである。



11月29日の様子（吉田教授とノヴァコフスキ講師）

この事業は、「令和6年度鶴岡市地域課題解決事業」として実施されました。