

AIとデータ解析をはじめの前に

宇野 毅明*

Introduction to Artificial Intelligence (AI) and Data-Analysis

Takeaki Uno*

本原稿は、2020年1月24日に大阪市東淀川区においてアドバンスソフト主催で実施したAI技術セミナー「CAE分野におけるAI技術適用可能性」において、宇野毅明教授にご講演いただいた内容を、アドバンスソフトが書き起こした内容です。また、章立てに関する情報もアドバンスソフトで付加したものです。

Keywords: 人工知能、データ解析、ビッグデータ、機械学習、ディープラーニング

1. はじめに

最近、一般新聞紙にAIという言葉が技術の意味で掲載されるようになってきました。この先どこまで投資や技術が伸びるのだろうかという話題が出てきています。

それでは、これまで人間が行ってきた活動が全てAIにとって変わられるかという、それはそこまではいかないだろうと考えられています。例えば、今のAIブームの火付け役の1つであるIBMのワトソンは、聞かれた質問の解答となる事実をデータから探し出すタスク（ファクトアンサー）に対する高性能なAIを作り、アメリカのクイズ番組でチャンピオンを破り優勝しました。このファクトアンサーAIを発展させることで、ある分野の数万本の論文を学習し、その分野の事実を解答できるAIを作れるのではと期待されました。創薬分野で例えるなら、新薬の自動開発ではなく、患者の症状に対して最適な薬剤や副作用の懸念を提示することになります。しかし、2018年にこのプロジェクトチームは解散同然となりました。これが意味することは、この方向性のAIはまだまだ実用レベルに達することは難しいだろう、ということです。人の知識を学習し、自動的に解答を導き出すようなAIを作ることはまだまだ難しいのです。

* 国立情報学研究所 情報学プリンシプル研究系 教授
Professor, National Institute of Informatics (NII)
<https://researchmap.jp/uno/>

ではここで、人工知能の知能とは何か、考えてみましょう。人間の知的活動にはさまざまなものがあります。記憶、計算、整理、、、推理、推論、関連づけ、判断、、、認知、認識、類別、判別、、、理解、読解、想像、経験、、、状況理解、説明、、、感情、創造性、愛、感覚、直感、、、少し考えただけでも多岐にわたります。よく『人工知能』と一括りにまとめられますが、そんな単純なものではないことがここからも分かるでしょう。では、この中でどの知的活動がコンピューターで可能かを考えてみます。最初の、記憶、計算、整理などはコンピューターが誕生したときにできるようになりました。推理、推論、関連づけなどは、1980年ごろに、論理学などから推論のルールを導入し、人間がルールをデータベース化することで実現しました。認知、認識、類別、判別などは、最近、深層学習と統計モデルによって格段に精度良くできるようになりました。これらは、人間がルールを教えることなくデータからルールを学ぶ帰納型のアプローチで、これにはビッグデータの存在が欠かせません。それ先の、理解や読解、感情や理解などはいまだ難しく、目処すら立っていません。ただ、状況理解や状況の説明は、全ての常識を詰め込んだ巨大常識データベースがあればできるのではないかと考える研究者もいます。人手でこの巨大データベースを作成するのは困難ですので、膨大なテキストデータや実世界の観測データからからなんらかの方法で学習するの

でしょうが、まだ見通しは立ちにくい状況です。物事の理解、文章の読解などはまだまだ実現困難で、創造性や感情はさらに困難です。最近、AIが作曲や作画をするといわれますが、これは人間の創造活動を模倣したにすぎませんので、芸術性を理解し創造しているのではないのです。確かに、人間の知的活動の一部はAIにとって代わられるでしょうが、世間でいわれているようなAIが人間の仕事を全て奪い、人間を支配するようなことははるか未来にならないと起きようもないでしょう。

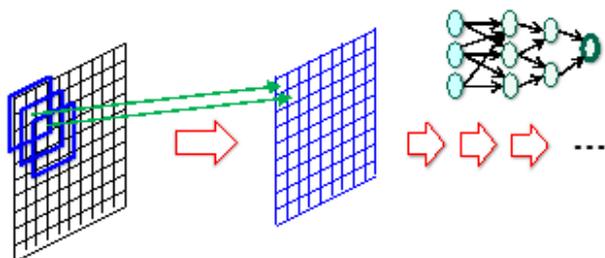


図1 ディープラーニングが学習する多層的なネットワークの模式図

一方で、最近ビジネスや産業でAIが注目されているのは、人間を抜きそうだからではなく、技術として新しく高度なものがあるからです。それはディープラーニングと統計モデルの爆発的な発展が大きな理由になっています。ディープラーニングは画像や音声の認識の精度を格段に向上させ、統計モデルの進化は自動翻訳を大いに発展させました。両者とも、画像や文章の意味を理解するのではなく、パターンとして認識し、データに付き合わせて最もそれらしいものを解答するだけで、何かを理解しているわけではないのですが、このアプローチの方がコンピューターにはしっくりくるため、ビッグデータの利用と合わせて従来よりも飛躍的な精度向上が起きました。

この文章を読んでいる皆さんの多くは、業務でのコンピューターによる計算といえば、シミュレーションが多かったと思います。シミュレーションとこれら最近のAI技術の違いを簡単にいうなら、シミュレーションは物理現象などメカニズムが分かっているときに、初期状態を与えるとそれがどのように変化していくかをメカニズムに従

って計算するものです。このようなアプローチをモデル演繹的といいます。一方で最近のAIは、データからメカニズムを推測する、つまり逆のことをします。メカニズムは不明だが、入力と結果（認識結果）だけは分かります。この状態でメカニズムを推定することで、入力から結果（認識結果）を計算できるようにするのが、機械学習、ディープラーニングの考え方です。

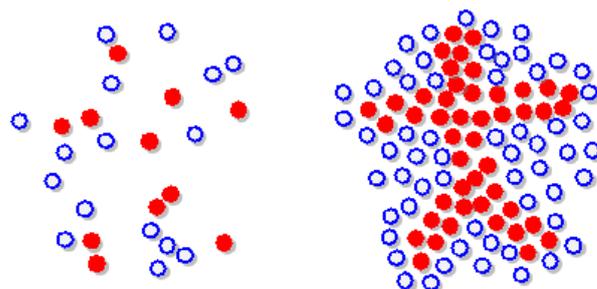


図2 データ量の増大によって、データが何を表しているかが分かりやすくなる例

ディープラーニングは、ニューラルネットという、人間の脳の仕組みを単純にモデル化したものを、複数階層積層したものを学習します。画像処理では、例えば図1のように、1ドットごとに分割された画像データを、例えば3×3の9ドットをまとめて少しだけパターンのある、パターンに抽象化されたデータとして次の層に送ります。これを順に繰り返すことでよりパター的な抽象度を増し、最後の層でYesかNoかの認識を行います。巨大な訓練用のデータベースを用意し、正解データではYes、それ以外ではNoが必ず出るようにこのネットワークを最適化することで、たとえデータベースに存在しないデータでも精度の高く判別ができるようになるのです。

ニューラルネットワークの考え方自体は1980年代から既にありましたが、それが今になってブレイクしているのには理由があります。ニューラルネットの最適化は比較的計算コストが高く、過去の計算機では2-3層が限界でした。近年スーパーコンピューターなどにより50層以上の最適化が可能になり、ようやく高い精度が得られるようになったのです。さらに、Web技術、センサー技術の発展によりビッグデータの利用が可能になったことも大きな要因です。

上でも述べたビッグデータですが、データが巨大であるだけで精度が増すことは非常に多くあります。図2を見てください。左図は、右側の図から多くのデータを除去したもののなのですが、これでは何が書いてあるのか分かりませんね。しかし、右側のように十分点があれば、赤い点がどのような形をしているか簡単に分かるのです。従来、統計学が盛んに研究されてきましたが、これはデータが少なくても何かを正確に理解、予測したいという欲求から出てきたものですが、巨大なデータがあれば、統計的手法が必要とする分布や確率の仮定、モデルがなくとも、データの分析が簡単にできるようになってきているのです。

ビジネス書などでは「AIはなんでも解決できる魔法の道具」「ビッグデータはなんでも情報が入っている宝の山」のように書かれていることが多いと思います。実際のところは、「近年発達したAIはものを認識する、いわばセンサー」「ビッグデータは(ざっくりとした、見れば分かるような)物事が分かりやすくなるデータ」として理解した方が良いでしょう。

この言葉にのせられてか、3~5年前に、多くの企業でビッグデータ・AI部門が立ち上がりました。ただ、目的意識、技術の理解なしに立ち上がったところも多く、迷走状態であったところも多かったようです。私のところにも、AI研究という看板を頼って、「こんなデータがありますが、どうしたらいいのでしょうか?」「AIを使ってどうやって儲ければいいのでしょうか?」「どういうデータを使えばいいのでしょうか?」のような質問が多くきました。とりあえずAIをやるとはいつてみたものの、何をやっていいのかわからない、いわゆる”迷子の状態”です。『写真や録音機が、画像や音声を認識するセンサーになります』『巨大なデータで顧客の隠れた行動の一部が見えるようになります』のように、AIを自身が利用すべき「技術」として捉えていれば迷子にはならなかったでしょう。目的意識を持ち、技術としてAIを見ていれば、自身の目的に合った技術を探し、技術のうまい使い方を考えるでしょう。総合的に企業の舵取りを考えるはずです。

これらの企業、問題意識がないわけではありません。自社の現状に対する危機意識や、さらなる発展の必要性を認識しているからこそ、AIに踏み込み、模索しているのです。上記のような迷子の方々も、丁寧にヒアリングをすると抱えている問題が言語化され、明確になります。AIに対する関心と危機意識は、そのトリガーになっているといえるでしょう。

実情としてAIに求められているものは、多くの企業で、顧客理解と業務改善のようです。これは、既存技術の導入でシンプルに効果を得ることが難しいため、長年積み重なった課題が複雑に絡み合い、大きな問題意識となっているからでしょう。商品開発や技術改善では初めから技術を求めているので迷子にはなりません。

迷子の企業をさらに深掘りすると、企業や業界の構造的な問題、組織や労使の人間関係的な問題などが本質的なものとして出てくることもあります。このような場合は、より初歩の情報技術、「情報の共有に対する価値観の醸成」「直感だけでなくデータを利用する」「分からないことにふたをせず、専門家にきちんと聞く」といった点が解決のポイントになることも多く、これを行うためにAIやIT技術をどのように利用するか、という議論になっていきます。

ある企業から営業日報のデータを見せてもらう機会がありました。「何が書いてあるかをよく知らないけど、営業日報から何か有用な情報が引き出せたら嬉しいです」という相談でした。中を見ても、きちんと書いてあるものとほぼ何も書かれていないものがあり、大事なことが書かれているものはすべからず備考欄の記述が長いという傾向がありました。つまり、長い備考欄だけ抽出するだけでも、なかなか有益で可視性の高いデータ処理ができることになります。深掘りすれば、さらに有益なことも出てくるでしょう。

自身の目的と、自身が持つデータの質が明確に理解できれば、AI技術、あるいは初歩の初歩のデータ解析技術だけでも、その利用法が分かり、効果的な結果を得ることができます。対して、経営判断をAIにやらせてみたいというような話も聞

きますが、このように、企業人として自身にとって最も大事な仕事を放棄せず、きちんと自身の業務、および AI と向き合っていたいただきたいと思います。AI 技術を持つ企業は、多分に、この向き合い、言語化し、解決策を模索する、コンサルタントの色が濃いのだと思います。

2. 東大ロボに見る AI の力

メディアでも盛んに取り上げられましたので、ご存じの方も多いたと思いますが、国立情報学研究所の新井先生のプロジェクト、「東大にロボットは入れるのか？」のご紹介をしたいと思います。これは、AI の考える力が、最新の研究ではどこまで進んでいるのか説明するのに良い例だからです。

このプロジェクトが AI の力試しの対象として入試問題を選んだのには理由があります。入試では、問題が明確に矛盾なく完備に記述されており、人間の常識に基づいた推定を行わなくても問題が解けるからです。人間にとって入試問題は難しいですが、コンピューターにとってはその明確さが非常に扱いやすいのです。約 5 年間の研究を経て、表 1 のような試験結果を得ることができました。旧帝大は少し難しいが、日東駒専、MARCH のレベルならギリギリ合格できそうな学力です。まずこの表からは、AI の強み・弱みが見えてきます。

AI にとって読解が難しいことは既に述べましたが、そのことは、国語の点数があまり伸びていないところに現れています。ただ、多くの場合読解問題の解答は問題文の中に埋め込まれており、それを検索する問題と考えれば、AI に手が出る問題になり、ここに関しては手が出せます。逆にとても強いのが社会科（日本史や世界史）です。社会の問題は、最初に述べたファクトアンサー型の質問が多く、つまりデータベースの構築で多くが解答可能なのです。何を聞かれているかだけ自然言語解析でおぼろげにでも理解できれば、あとはデータベース検索で最もそれらしい単語を提示します。例えば、『徳川幕府 7 代目の将軍の名前を答えよ』という問題では、「名前を答えよ」という問いは認識する必要がありますが、「徳川幕府 7 代目将軍」を検索すれば、自然と出てくる名

前は一通りに定まるでしょう。検索機能が気を利かせて答えを見つける必要はなく、「徳川幕府 7 代目将軍」を含む Wikipedia や辞書の文章から、人名辞典に載っているもの提示する程度で十分です。特に、選択問題では、全ての選択肢について、それが Wikipedia の該当ページでどれだけ頻度高く使われているか、ということ調べるだけでも十分です。人間にとっては Wikipedia の記憶こそが難しいのですが、コンピューターには優しいのです。

表 1 東大ロボの得点表

大学入試センター試験の模試 (マーク式)			
	得点	全国平均	偏差値
英語 (筆記)	95 (80)	92.9	50.5 (48.4)
英語 (リスニング)	14 (16)	26.3	36.2 (40.5)
国語 (現代文+古文)	96 (90)	96.8	49.7 (45.1)
数学 I A	70 (75)	54.4	57.8 (64.0)
数学 II B	59 (77)	46.5	55.5 (65.8)
世界史 B	77 (76)	44.8	66.3 (66.5)
日本史 B	52 (55)	47.3	52.9 (54.8)
物理	62 (42)	45.8	59.0 (46.5)
合計 (950 点満点)	525 (511)	437.8	57.1 (57.8)
2 次試験の模試 (記述式)			
地理歴史	16	14.5	51.8
数学 (文系)	46	19.9	68.1
数学 (理系)	80	30.8	76.2

一方で、社会科でも AI には難しい問題もあり、例えば写真や図から状況を読み解く問題はとても苦手です。人間はイラストをなんとなく理解できますが、AI では全ての対象を定義（例えば雪国の暮らしのイラストであれば、白地に黒い線で囲まれている部分は雪である、水色の点々は消雪パイプから出てきた水であるなど）して、訓練データとしないと問題を解くことができません。猫を

認識できる AI にキティーちゃんが認識できないのと同じように、イラストと写真はまったく別物で、人間はその意味を推察できるので理解できるのです。しかし、このような問題は少ないので、社会科は AI が苦手な問題が少ない、得意な科目なのです。

英語については、皆さんがご存じの通り、AI は大の得意です。Web 検索で充分解けるものも多く、「arrive () seminar room」のような穴埋め問題なら、at や to や on を入れて検索し、ヒット件数が一番多いものを選べば十分です。言葉はみんなが使うものが真実なので、検索してヒット数が多いものを選ぶのはある意味で真実です。英訳、和訳問題は上述の翻訳機能で十分でしょう。

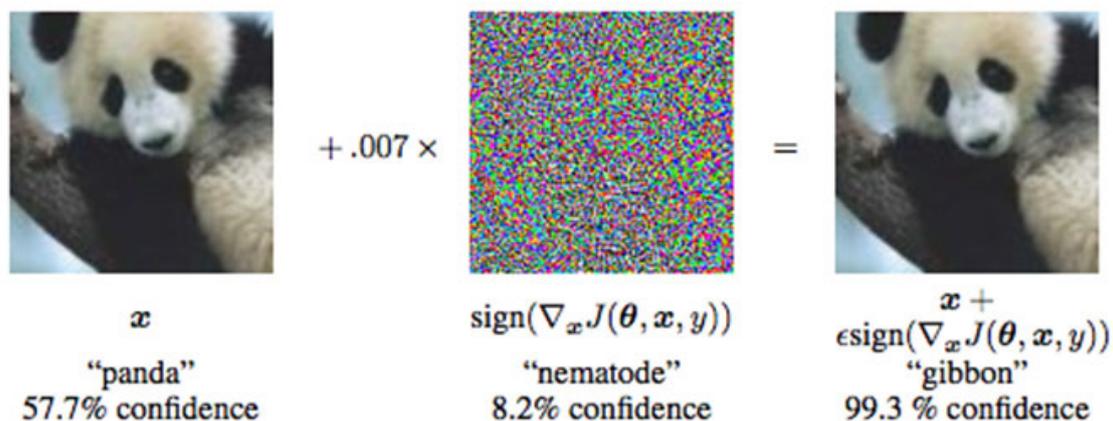
一方で苦手な問題もあります。社会科と同様に、イラストの中身を読み解く問題は苦手です。また、英語のリスニングの点数があまり高くないことを不思議に思われた方もいるかもしれません。最近では、音声認識が発達したのに、なぜリスニングの問題が解けないのか、それは音声認識の難しさではなく、会話文の難しさがネックだからです。会話には「雨が降れば体が濡れる」「雨が降りそうなら人は傘を携帯する」「雨が降りそうなのに傘を持っていない人は、傘を家に忘れた」「傘を忘れた人は雨が降るかどうかになり、不安になる」といった常識のデータが不可欠です。この常識がない AI にはできません。「なんでお花を買ったの」というセリフの次に「誕生日だからだよ」という解答を選ぶことは、常識を持たない AI には不可能です。

その点、世界の常識を必要としない数学は、AI の得意科目の 1 つです。例えば数学の証明は、与えられた条件から、証明の命題にあるゴールとなる条件を導けるかどうかの問題なので、自動定理証明システムでだいたい解けます。文章問題を読み解き数式に変換するところは大変ですが、残りほぼ自動でできます。苦手なものは、統計で表の中の穴埋め問題（表の中に抜けている部分があって、周りの数字から抜けている部分を想像する問題）のようなものですが、これも問題数が少ないので、落としても大きなダメージにはなりません。

AI はこれだけ問題が解けますが、東大の筆記試験には不十分です。選択でなく筆記式、意味を深く問われる問題が多いなど、AI には簡単に解けない問題ばかりです。一方で、センター試験の点数で合否判定が出される大学には、AI は多くの大学に合格できることとなります。日本では、子供達が何年もの時間を費やして受験勉強をするのですが、その多くは AI が簡単に解いてしまうのです。これが、AI が（給料の良い）仕事を奪ってしまう、という言説のもととなっているのでしょう。

では、AI にどのような仕事が奪われてしまうのか、実際に考えてみましょう。まず、人にいわれたことを模倣するタイプの仕事、知識を詰め込んでそれを出すだけの仕事は AI に取って代わられる可能性が高いです。コールセンター、居酒屋の店員などの業務の一部は既に IT 化され、既に仕事が一部奪われているといいでしょう。知識を蓄えることが必要な職種に、国家試験の合格が必要な士業（弁護士、税理士など）があります。薬剤師が、処方箋どおりの薬を出して、薬の効用と注意を説明するだけなら、薬剤師は AI にとって代られてしまうでしょう。しかし、患者とのコミュニケーションにこそ仕事の本質があるのであれば、AI にすぐにとって代られることはないでしょう。税理士や弁理士も、書類の書き方や規則を教えるだけなら、AI で十分でしょう。しかし、特許の新しさや重要性を認識し、それを主張する方法を考える、収入や支出の意味を解釈し税法の適用がどのようになるかを考える、という理解と創造性を伴う作業はまだ AI には難しいです。AI が仕事の手伝いをすることはできますが、完全に奪うことはないでしょう。これら士業の方々は今までの業務を効率化する一方で、より新しい価値を提供するように仕事が変わっていくのだと思います。会計士に至っては、その仕事は相手の嘘を見抜くこととなります。コンピューターは基本的に、いわれたこと（データ）を信じて、その通りに動くものなので、嘘を見抜くには、嘘の見抜き方を教えないとできません。

知り合いの税理士さんの話では、「スナックでの売り上げを推定するとき、ピーナッツの仕入



© <http://karpathy.github.io/2015/03/30/breaking-convnets/>

図3 人間とコンピューターの理解方法の差から生じる認識の違い

れ量で推測する」のだそうです。おつまみにピーナッツを出すことが多いので、仕入れたピーナッツの購入量と客数が比例すると考えるのでしょう。このような、アイデアを出すこと、はAIにはとてもできません。

3. AIの理解の仕方

AIにとって「理解」はとても難しいことです。人間の意味するところの「理解」は難しいとしても、猫や犬の画像を認識できるようになるからには、データから何かを獲得しているはずですが、これはいったいなんなのでしょう。ディープラーニングを例に説明しましょう。

ディープラーニング図1のような階層化したニューラルネットの形、重み付けを最適化して認識精度を上げます。つまり、この最適化したパラメータなり重みなりネットワークの形なりが、コンピューターが画像などを認識するために獲得したものとなります。実用上、十分な精度が出ますので、これは画像を表す知識になっており、ある種の理解といってもいいのかもしれませんが、人間のいうところの理解の形とは大きく異なるのは分かると思います。理解の仕方が全く違うので、人間には猫に思えるイラストでも、AIには猫と認識できず、猫のイラストを認識するためには猫のイラストを学習する必要があります。

AIの理解の仕方を見る、興味深い例があります。図3を見てください。引用元のサイトではパンダの画像の認識に関する研究の紹介をしています。

図の左と右に同じパンダの画像があるように見えますが、右の画像は左の画像に、中央のカラフルなノイズを、色の強さを7/100にして足し合わせてできたものです。われわれには両方とも全く同じパンダに見えますが、AIは、左の画像は57%の確信度でパンダと認識するのですが、右の画像は99.3%の確信度でテナガザルと認識します。驚きの結果です。

これはまさに、人間とAIの理解の仕方の違いを表しています。ディープラーニングは、ニューラルネットの画像の細かい特徴をまとめ上げるようにして全体的な特徴を捉えます。図3のノイズの情報は、パンダの画像の細かい部分がディープラーニングにとって「テナガザルっぽく」なるように設定したものです。人間にはパンダに見えますが、ディープラーニングにとっては、「画像の至るところがテナガザルっぽい」画像なのです。

このような技術で仮に自動運転の車を作ったと想像してください。何かフィルターを車のカメラにつけたら、人が通ったときに、ゴミ袋が舞っているように見せられるかもしれません。自動運転の車を作る側は、テストをたくさん行うでしょうが、このような「セキュリティ面での攻撃」を全て防ぐことは難しいでしょう。カメラにフィルターを貼り付けるほかにも、チカチカした光を当てるとか、人がなんかの服を着ていた場合とか、そういうことを全て試すことは不可能でしょう。しかも、1回これで車を暴走させることができるとしたら、同じ仕組みで動いている車が全て暴走

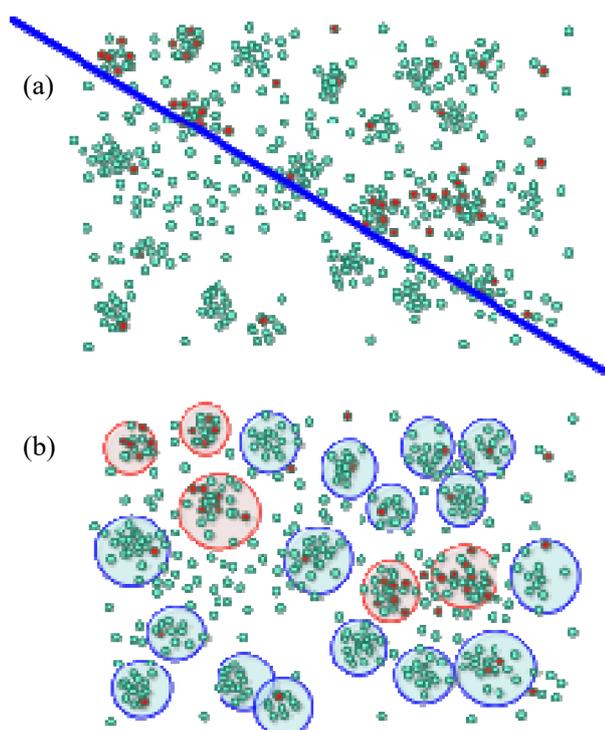


図 4 コンピュータによるデータ分類(a)と人によるデータ理解(b)

することになります。

逆に、何の妨害をしていなくても、目の前に出てきた人が、着ぐるみを着ていたら、はたして人と判断できるでしょうか？急ブレーキが危険な状況なら、はねる判断をしてしまうかもしれません。こういう問題があるので、自動運転を本当の意味で使うようになるには難しいかと思われま。理解の仕方が異なると、想定外のケースを大量に考えなければなりません。実証実験をたくさん行っても、想定内のから外に出た事象に対する不安はなかなかぬぐえないでしょう。

このような、AIの理解の仕方が人間にとって理解不能であることを、AIのブラックボックス化といいます。このように極端な危険を考えることだけではなく、一般にはAIの判断基準が分からないことを意味します。判断基準が分からなくても、常に正しい答えを出してもらえらるのなら問題ないのですが、正解が明確ではないもの、想定外の事象がありえるものなどは、その場合にどのような判断が行われるか予測しづらいため、AIの利用に懸念が生じます。そのため、理解しやすい解を出すAIの技術開発に注目が集まっています。

もし、ディープラーニングの判断基準を示せと

いわれたら、われわれにできることはディープラーニングが獲得した50階層以上のニューラルネットワークの形状を示すことしかできません。しかしこれでは、相手も納得しないでしょう。たとえ全体像が把握できたとしても、それがどのようなメカニズムになっているのか、概略をつかむことはほぼ無理です。巨大ニューラルネットを直接的に理解しようとするのは、あまり筋の良くないアプローチでしょう。

ディープラーニングとは逆に、単純なルールでルールを学習するという方向もあります。しかし、それはそれで理解しにくい場合もあります。例えば、図4のようになんらかの属性値が縦軸、横軸に対応する平面に分布している点のデータがあるとします。赤のマークが目撃したいデータ（事故、大切なお客さまなど）だとします。どのように赤い印が分布しているのかを、サポートベクターマシンという単純なルールを出してくれる機械学習に入力すると、図4(a)の青線を出してくれます。この線の右上が赤い点がたくさんある領域です、ということですね。このルール、単純ですし、精度も高い、モデルとしては完璧です。しかし、人間にとっては全く完璧ではありません。なぜこの場所に線が引かれるのか、なぜこの傾きなのか、という疑問に対する説明が全くできないからです。

人間は、何か大きなこと、複雑なこと、メカニズムの分からないことを理解しようとするとき、まずはその物事を観察します。そして、こんな性質があるのかな、こういう仮説が成り立つのかな、こういうルールなのかな、などと考えます。データがたくさんあるなら、抽象化して、一般化して、特殊なものに注目して、得られた観察から、状況を鑑みて最もらしい事柄を、仮説なり法則なりとしてたて、総合的に全体を俯瞰していきます。このような方法で、図4のデータを理解しようとするならば、点に群れがあることに注目し、群れごとに分けて考えてみるのが自然でしょう。分けてみると、図4(b)のように赤い点が多く入ったグループがいくつかあることが分かります。このあたりにある点、赤くなる可能性が高いのだろう、

と考えます。理由が知りたければ、グループの中を見ます。グループ内の点は、図の上で近くにあるということは、何らかの共通点があるということになるので、その共通点を観察することで、そこに赤い点が多いことの原因を探ることができるのです。

人間がデータを見るときは、その着目点は予測の精度ではなく、何か特徴を理解しようとしています。理解しやすいAIのためには、AIがそもそも人間のこの思考法、抽象化や共通性の抽出、といったことをサポートすべきでしょう。

4. 精度の落とし穴

近年のコンピューターシステムによるものごとの自動化では、多くの場合、システムの評価の指針は精度になります。認識にしても、推薦にしてもそうです。では、理解したいという課題に対する精度は何でしょうか。現在の機械学習では精度が価値の中心であり、精度が評価尺度にあるタスクにしか適用されていない、というのが実情でしょう。しかし、精度を追い求めてばかりいると、意外なところが抜け落ちてしまうことがあります。その例をご紹介します。

クラスタリングは、データ分析が昔から行われてきた分野、例えば小売業などで使われてきた技術です。さまざまな顧客のニーズに応えるため、顧客を複数のグループに分割し、それぞれのグループのニーズに対応する店作り／商品設計を行うときに用います。グループ分けと聞くと、男女、年齢といった顧客の属性値に基づいた分割を行えば簡単に思われますが、若い女性にもいろいろな好みの人がいて、そのニーズを的確にくみ取れるわけではありません。クラスタリングは、顧客の買った品物の特徴に基づいてグループ分けをすることで、「同じような買い方をしている人」「買う品物に共通性がある人」というようなグループ分けを行うことで、顧客の好みが捉えやすいようなグループを得ます。

クラスタリングの狙いは大きく分けて2つあります。1つは全項目、例えば顧客、を2~5程度のグループに分けたい、というもの、もう1つは特

徴的な共通性を持つ人のグループを見つけ出したい、というものです。後者は、新聞記事などのデータではトピックマイニング、ソーシャルネットワークの分野ではコミュニティマイニングと呼ばれます。共通性を持つ記事を集めれば、それはなんらかのトピックに対応するだろう、という仮説から来ています。

この2つ狙いの両方とも、それほど上手に的確に解ける問題ではありません。グループ分けをする基準はいくらでも考えられますし、共通性もいくらでも考えられます。スーパーの顧客であれば、和食・洋食、ファミリー、独身、高級志向、健康志向、エスニック、共通性の候補はたくさんありますが、唯一の正解があるわけではなく、それぞれの分け方はそれぞれいい面を持っています。このように判断に揺らぎのある問題を解くことは、コンピューターにも人間にとっても難しい問題です。クラスタリングでは、昔からさまざまなモデルが考えられ、いろいろな方法でクラスタリングができます。しかし、このモデルが人間の持つこのモヤモヤとした揺らぎを正確に表現できないので、当てがはずれた結果が出ることも多いです。典型的なものが新聞データで、過去に使われてきた方法を使うと、いろいろな記事に使われる言葉、「今日」「東京」「起きた」などの単語でつながった超巨大なグループと、誰ともつながっていない小さなグループが大量にできます。もちろん、超巨大グループは特徴が見えませんが、孤立グループは何が一般性であるのか分かりません。顧客分類でも、普通に買う物を普通に買う「普通の人グループ」ができてしまい、その方々が持つちょっとした興味や指向は全部隠されてしまいます。これらの結果は、今まで研究されてきた評価値に基づいて、精度高く解を求めた結果得られたものです。しかし、人間の評価基準がよく分からないときには、いくら精度を追求してもいい結果は得られません。しかし、現実には、解の評価をするものが精度しかないのです。きちんとした質を担保したいと考えれば考えるほど、莫大な計算コストをかけて、この「精度」を追求することになるのです。これが精度の落とし穴です。精

度を良くしていい知見を得たいと思っているのですが、精度を良くしてもたいした結果は得られないことが多くあるのです。

さらにいえば、精度とは数値なので、解の質や形については何もいっていません。クラスタリングや学習の問題では、評価値は同じだが様相の違う解が出てきます。例えば画像認識であれば、毎回全く違う巨大分類ルールを出してくるのですが、精度はほぼ同じです。クラスタリングも、分け方が毎回バラバラですが、評価値はほぼ同じです。図5の例は、冗談や誇張はなく、実際にこれくらい異なる分け方が毎回出てくるのです。精度がほぼ同じなら、データ解析を依頼してきたクライアントに「なぜこのグループ分けを選んだか」を説明するときに根拠が持てません。実際には、熟考してクラスタの意味付けを考えて解釈を与え、その解釈の最もらしさを根拠としています。

結局のところ、AIのデータ解析現場において難しいのは、高い精度の求解ではなく、結果の解釈です。図5に戻ると、各グループが持つ共通性や特徴を理解した上で、グループの顧客の嗜好やライフスタイルなどを想像し、クライアントのビジネスの方向に合う分け方を提示するのでしょうか。深く考えれば、その提示にどのグループ分けを使うかによって、クライアントのビジネスの方向は変わります。開設する店舗、閉める店舗も変わってくるかもしれません。自分の選択で多くの人の運命が変わってしまうかと考えると、選択肢があること自体が大きなプレッシャーにもなります。このような経営判断を伴わなくても、去年と今年はほとんど変化がないのに結果が大きく違う、ということにもなり、信頼性の大きく損ねることになります。精度が高いことよりも、解の選択肢がない、つまり安定していることの方が高い価値を持つことが往々にしてあるのです。

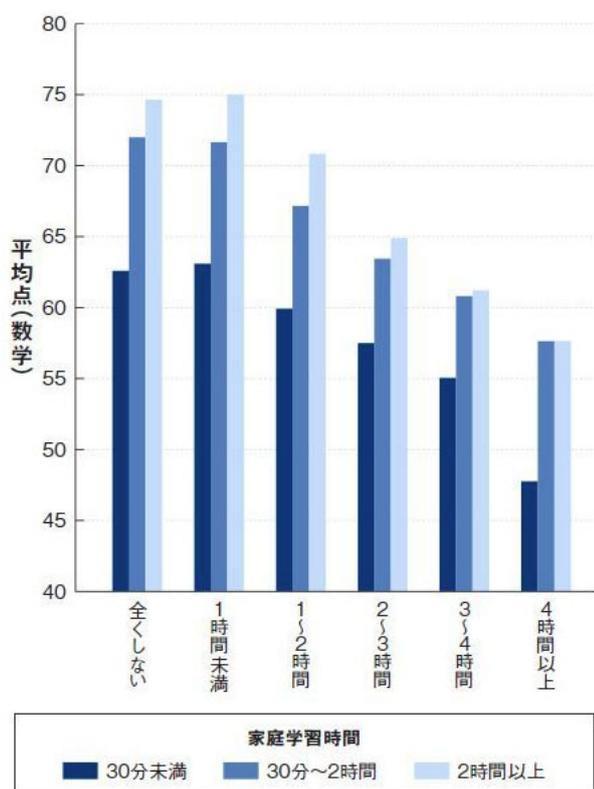
データサイエンティストの間には、『バラバラの解が出てくるけれども、データの中には必ず真実の分け方が存在していて、究極的にモデルをフィットさせたなら、必ず真実の解が見つかるはずだ』といった都市伝説のようなものがあります。われわれはこの都市伝説の検証をしてみ

ました。同じクラスタリングアルゴリズムを1,000回実行して1,000個のグループ分けを得て、アルゴリズムで使われている評価値が高いTop10を出して比べてみました。もし、精度が高くなると真実の解に近づくのであれば、精度の低い解はバラバラと異なるものが多く、精度の高い解はあまり異なりがないはずですが、しかし結果は、どのような精度でも同じようにバラバラの違いがあり、都市伝説は棄却されてしまいました。これ以上精度を突き詰めても、バラバラなのはバラバラなのでしょう。科学技術的に基づいたもの、という目線でAIやデータ解析技術を見ると、どうしても精度が神様のように見えますが、そんなことはなく、ある種の思い込みに近いと考えています。

5. データの解釈

上記の例でも出しましたが、計算結果の解釈は非常に難しく、手間がかかります。最近ではデータを使って議論することは広く市民権を得ていますが、解釈に主観が入り、作業としても一番大変だという認識はされていません。解釈の難しさはデータの分かりにくさや大きさに決まるものではありません。例えば、図6を見てください。3つ値を同時に表示しているのだから分かりにくいですが、子供のスマホの利用時間と勉強時間と学校の成績の関係です。横軸は1日にスマホを使う時間、3本ある色の違う棒は1日の家庭での勉強時間に対応し、勉強時間とスマホの時間で子供をグループ分けして、それぞれのグループの数学の平均点を縦軸で表しています。この図から、「せっかく勉強してもスマホを使うと勉強したことを忘れてしまう」という論説が出ているようですが、これはグラフから読み取れるものとしては間違っています。全くスマホを使わない人にスマホを使わせてこのような成績になったのであれば、多少説得力があるのですが、現実にはそうではないからです。これは、明らかに解釈違いですが、こう解釈してしまう人もいます。

この間違いの原因の1つは、因果と相関の違いを理解していないことにあります。スマホをするから学力が落ちるのか、学力が落ちたから勉強が



©プレジデント <https://president.jp/articles/-/24764>

図 5 平日の携帯・スマホの利用時間と家庭学習時間・数学の平均点

いやになってスマホをしているのか、はたまた勉強をしない人はスマホを使う時間がたっぷりあるだけなのか。スマホと学力に相関がある、つまり数値の上下に関係があることは見て取れますが、何が原因で何が結果なのかは分かりません。勉強が嫌いで遊んでばかりの子供が、成績が振るわないのは過去も現在も同じですので、スマホに限らず、勉強以外のことばかりしていれば学力は下がるのでしょうか。因果関係としては、これくらいしか導けません。

この解釈違いにはさらなる原因もあります。図からは、たとえスマホを1日4時間以上使っても、勉強すれば成績が上がる、という観察もできます。これは先ほどの解釈とは正反対の結果です。スマホをすれば成績は下がるが、スマホをしていても勉強すれば成績は上がる。これはスマホの時間が成績を決める根本的な原因ではないということを示唆しています。

われわれはデータを読み解くときに、因果関係が知りたいのです。図6のようなグラフには相関

しかありません。どの相関のところに、どっち向きに因果が付いているかを現実の状況などに当てはめて、仮説を立てて考えるのが解釈です。ですから、たとえ小さなデータでも、仮説の立て方には広い幅があるのです。ちゃんと納得がいて、真実に近いと確信が持てる結果を得ることが大事でそのためには、因果と相関の違いや、仮説を支持する観察と支持しない観察がどれくらい得られるかとか、そういうことを総合的に考える必要があります。

数字の理解は実は簡単です。グラフにすれば、数字の特徴、大きい小さいか、大小関係や順位順番は、簡単に理解できます。難しいのは、そのグラフの意味解釈をして、因果関係を見つけることです。間違えると、本当にとんでもないことになります。間違えやすいのは、自分が強く信じる仮説を持っているとき、特に思い込みから出てくる仮説は大敵です。例えば、「最近、物騒になったよね」という話はよく聞きますが、年間の殺人件数や強姦、窃盗の数はここ50年ほど減り続けています。交通事故の死者も昔に比べればずっと減少しています。データでは、世の中は平和になってきていますが、ニュースなどを見て物騒になったと勘違いしているだけなのです。

こういう思い込みには、メディアの功罪もあります。ニュースで児童相談所への通報件数がうなぎ登りであることを示せば、虐待が増えていると視聴者は思うでしょう。しかしこれは、児童虐待と思われるときに第三者が通報するという、慣習の変化しか表していません。虐待の実態を表す、多分正しいと思われるデータは、幼児の他殺件数だと考えられます。10歳未満の子供が殺された場合、ほとんどの犯人は親であり、つまり虐待をしたこととなります。つまり、虐待数と子供の他殺件数はほぼ比例すると考えられます。幼児の他殺件数の推移を見ると、これも50年くらい下がり続けています。恐ろしいのは、現在の親御さんではなく、昭和時代の親御さんだということになります。このような間違った認識をしないようにするのはメディアの責任ですが、同時にデータ解析者の責任でもあるわけです。

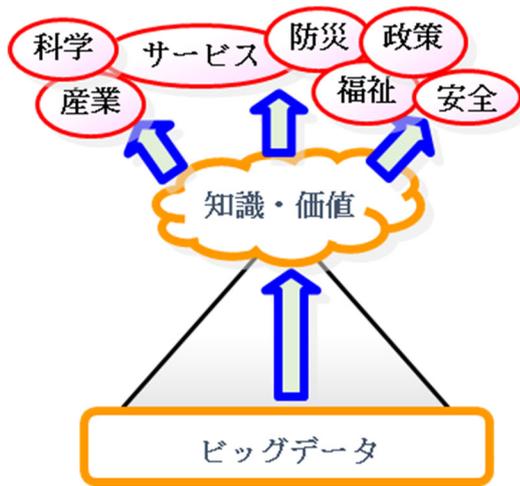


図 6 ビッグデータをビジネスに結びつける模式図

人間にとって、物事に安易に結論をつけることは、ある種の快樂です。あれを考え、これを考え、それでも結論が出ないことはとても大きなストレスで、「全ては神様がお決めになるのです」のような簡単な価値観、簡単な判断基準があった方が悩まなくて済みます。企業や役所で前例踏襲や横並びで物事を決めるのは、この決めごとをするストレスから逃れるためでしょう。AIに判断をゆだねる、ゆだねたい、というのにも、多分にこのような心理が働いているのでしょう。しかし逆にいうと、その苦勞を乗り越えて考え続けることが、進歩につながっていくと考えられます。AIを出す解をやみくもに信じるのではなく、しっかりと解釈をしてその意味を捉えることこそが、AI時代のビジネスにとって最も大事なことなのかもしれません。

6. まとめにかえて

AIに関わる記事や文献には、頻繁に未来の予測が出てきます。ここでも、AIの未来について考えてみましょう。まずはじめに断っておきますが、われわれ人類は未だに未来予測がとても苦手です。例えば、スマホや携帯電話の爆発的な普及を予言した人は皆無でした。電話を持ち運べれば大変便利になる、それで情報検索をすれば生活の質が上がる、こういう発想はありませんでした。証拠に、ほとんどのSFには、腕時計型の通信機器は出てきますが、携帯電話や情報端末は出てきません。さらにいうならば、インターネットと情報

検索のシステム、ブログやSNSも出てきません。昔の未来予測を見ると、町中をロボットが歩いたり、空飛ぶ車が飛んでいたり、宇宙旅行が実現されていたり、どれも実現されてないものばかりです。

では、実際にわれわれの生活で変わったことはなんでしょう。最も大きな者は、LINEやYouTubeなどインターネットサービスでしょう。さらにいえば、サービスの登場ではなく、その使い方が文化や社会を変えてきていることが、われわれ人間社会にとって最も大きな変化でしょう。このような使い方の変化、それに伴う新しい価値の登場まで考えると、未来予測が難しいことをうなずけます。このように正確な予測が難しいときには、それぞれの人がそれぞれの立場から自分の考えを持ち、それを議論していくうちに、だんだんと実像が固まっていくのだらうと考えます。そこで、ここでは多少主観が入りますが、ある未来のシナリオを考えたいと思います。

本稿で紹介したようなデータの解析、今後技術が進めばデータの理解が今よりもはるかに容易になるでしょう。人々はデータから物事や社会の仕組みや現状を、今よりも簡単に理解できるようになります。そのような世界では、物事に興味を持った人が手軽にデータを解析し、その結果に自身の解釈をつけて発信していく、そういうことがごく普通に行われる社会になるのではないかと考えます。人々が、社会や経済について自分の意見をいいたい、ということは、Twitterやニュースのコメント機能の盛況ぶりを見れば明らかでしょう。さらに近年の行政や企業のデータのオープンデータ化は、この動きに拍車をかけると考えられます。

このようにデータに基づいた意見がなされるようになると、それに基づいた議論が盛んに行われるようになるでしょう。今は政治や経済の議論という、感情的だったり、自身の友達や過去の経験のみに基づいていたり、当てにならないものが多いのですが、落ち着いた議論をする社会に成熟するのではないかと考えます。例えば、スポーツの分野は、既にそのようなデータに基づく議

論を盛んに行っています。サッカーでは、勝率やボール支配率や選手の運動量など、野球では打率やOPSなどを駆使して、データに立脚した議論をしています。データで議論をされると、感情や経験のみに基づく意見は、説得力を急に失うでしょう。このようにデータで話をすると、あまり喧嘩が起きません。意味のない中小のような批判も起きません。じゃれあいのような悪口だけになると考えられます。

逆に、現在の議論の中で、最もデータが入っていないのは政治でしょう。各党の説明にはなんの根拠も数字もありません。時々、2,000万円とかGDP-0.5%などの具体的な数字が出てくると、その解釈の異なりから、大きな論争が起きることもあります。データに基づいて議論する世界ならば、不毛な戦いはせず、しっかりと政策を立案できるかもしれません。



図 7 宇野教授ご講演風景

次に、データサイエンスの話をしていきます。データサイエンスを業務とする人をデータサイエンティストと呼びますが、彼らの仕事はデータからなんらかの知見を見だし、それを産業やビジネスに応用することになります。しかし、上述したように、データ解析の技術自体はそれなりに上がっているのですが、出てきた解の解釈や意味付けが難しく、それとビジネスアイデアを結びつけることが困難であるため、なかなかスムーズな普及には至っていません。ビッグデータが花開き始めたころ、1つのデータから100の価値が出るといわれていましたが、1の価値を出している事例も少ないのが実情です。ビッグデータでニュースリ

リースを検索したところ、自社でビッグデータはじめたというリリースはたくさんありましたが、知見が出てビジネスに活かせましたというリリースは、大阪ガスさんただ1社だけでした。

このような状況ですが、シリコンバレーをはじめとする欧米では、データサイエンティストは成功しているように思えます。これはひとえに、マーケットの大きさが大きな要因と考えられます。日本で500万の投資が付くビジネス案件は、マーケットが10倍大きなアメリカに行けば5,000万の資金調達ができることとなります。これなら、はるかに仕事はしやすく、小さな可能性でも芽が出てくるでしょう。では、日本は勝ち目がないのかといえば、それはそうでもないと考えています。一般に、日本の企業の強みは、現場力といわれています。現場の状況を現場の方々が理解し、それを改善する思考力を持っています。現場力の強さをデータ解析に活用する潮流さえ生まれれば、日本はデータサイエンス業界に強い国になるはずですよ。

以上、ざっくりでしたが、AIのビジネスにおける現状を、技術と、使い方と、難しさと、未来像の点からご紹介させていただきました。もう少しAIとビッグデータの、ざっくりした話をお読みになりたい方は、ぜひ下記の解説をお手にとりいただければと思います。

参考文献

[1] 宇野 毅明、池田 亜希子「情報研シリーズ しっかり知りたいビッグデータとAI」、丸善出版

※ 当原稿は、宇野毅明教授のご講演をもとに、アドバンスソフトが構成したものです。写真は、ご講演当日の写真です。

※ スライド資料をアドバンスソフトのHP シミュレーション図書館に公開しております。ダウンロードの上、本稿をみていただければより詳細な情報となります。

http://www.advancesoft.jp/support/download/37simlib_seminar_20200124_all.html