AI, IoT 時代の問題解決法と人材育成-有効なデータ解析の背後にある基本原理-

独立行政法人統計センター理事長・椿　広計

１．はじめに

Tukey (1962)が予言した「データ解析の未来」を実現するために．Tukey(1977)自身が「探索的データ解析（EDA, Exploratory Data Analysis)」という運動を興した．これが，データ解析という考え方の嚆矢である．Tukey は，ベル研究所にデータ解析のためのプログラミング言語 S を 1970 年代に開発させるなど，データ解析に必要なソフトウェアを開発することも指導した．探索的データ解析は，それまでの統計的方法が，検定など仮説検証のための方法論であったことに対して，データから仮説を探索するためのデータ解析を志向していた．探索的データ解析は，1980 年代には，データマイニングと呼ばれる第 2 世代人工知能活動とその代表的ソフトウェアとしての CART（Classification and Regression Tree，Breiman et al.，1984)と，それを Random Forest などに発展させた Brieman(2001)らの独創性によって，統計的機械学習全盛の時代を迎えた．実際，統計的機械学習は，第 3 世代人工知能の中心的方法論となっている．

一方，日本科学技術連盟は，奥野他(1971)が当時先端的な多変量データ解析の産業界での利活用促進のために「多変量解析研究会(通称，MA 研)」を発足させた．日本は統計的実験計画法の産業利用では世界をリードしたが，多変量データ解析でも MA 研や応用統計学会を中心に探索的データ解析や，上述した第 2 世代人工知能，グラフィカルモデリング等をいち早く産業界に普及する活動を展開した．特に芳賀(1984)の対話的データ解析とそれをサポートするソフトウェア開発が，日本科学技術研修所によって行われた．

MA 研はその後 20 世紀末まで活動を継続し，わが国産業界にデータ解析プロフェッショナル人材を多く育成する役割を果たした．MA 研ないしは日本規格協会の「データ解析研究会，通称 DAC」は，データ解析の数理を研究したのではない．産業界から提供された難易度の高い事例を共有し，それを研究会の中で解決することで，データ解析に基づくソリューション提供のプロセスを検討したのである．その種の活動の成果が，奥野他(1986)，吉澤，芳賀(1992, 1997)である．これらの活動に類似したものは海外産業界では見られない．

筆者は，データ駆動型時代(Society 5.0)の基幹ツールとなった統計的機械学習が，先人のデータ解析における戦略や戦術とは無縁ではないと考えている．この報告では，統計的機械学習でも活用されていると考えられる多変量データ解析の基本原理を振り返ってみたい．それと共に，品質工学（田口メソッド）や実験計画法で培われてきた主要な考え方のデータ解析への導入についても論じたい． なお講演では，独立行政法人統計センターが，2018 年 6 月に統計分析教育のために公開した 1741 自治体×111 変数からなる「教育用標準データセット(SSDSE, Standardized Statistical Data Set for Education)」（統計センター，2018）．を用いて本論文で記載したデータ解析の原理を紹介する．

２．データ解析に何を期待するのか－目的の分類

はじめに，データ解析でどのような知識価値を取得するかについて振り返りたい．伝統的統計科学が目指したのは，一般化可能な知識の獲得である．経営や技術開発に資する汎用性の高い知識をデータ解析が提供するのである．一方，統計的機械学習では，高精度予測が出来ること自体に価値を求めることが多くなっている．前者は，科学的データ解析，後者は技術的データ解析と呼ぶべきものである．また，データに基づく方法が，産業界の問題解決に果たして来た役割は，椿(2018a)が総括したように，「発見機能: 統計的問題の発見と絞り込み」,「分析機能: 統計的問題に影響を与えている原因の同定」,「最適化機能: 対策の最適化」の 3 つである．データ解析教育の大半は分析機能，すなわち結果と原因との関係性を定量的に示し，予測や最適制御にも活用

可能な統計モデルの当てはめに充てられている．そしてモデル当てはめには様々なノウハウが蓄積している．

一方，データ解析の基本機能である発見・分析・最適化は，椿(2018b)が，初中等教育のために示した図 1 のように，マネジメントの科学的サイクルである Deming と石川らの PDCA サイクルの Check，Action, Plan にそれぞれ対応している．

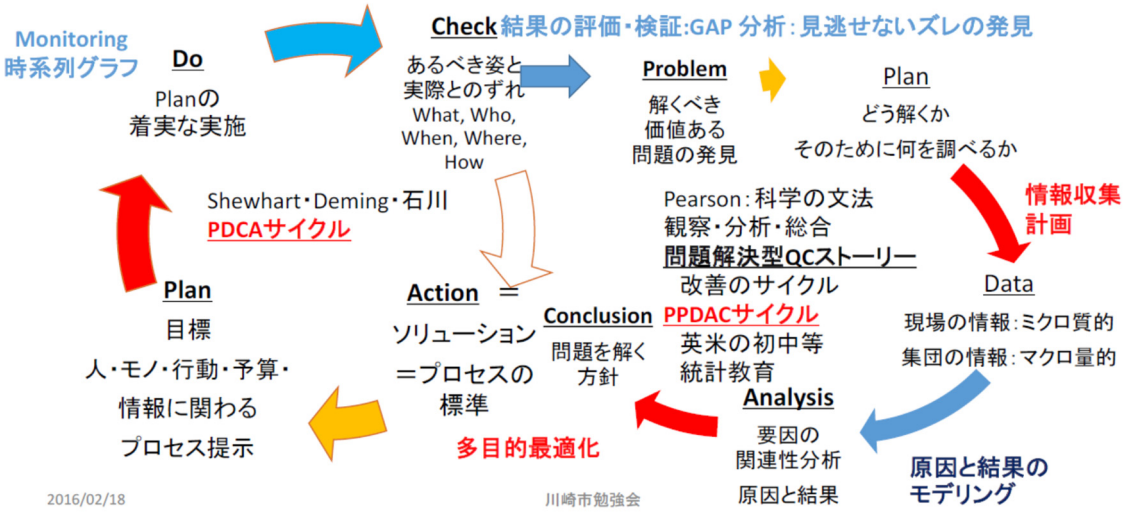

図 1　Check に基づく問題発見をトリガーとする問題解決のサイクル（椿，2018b より）

現時点で，企業が有する最適の統計的予測モデルから期待される値と有意に異なるデータが生じることが統計的問題である．つまり，想定外の外れ値や異常値の検知が統計的問題発見に資する方法となる．統計的ないしはデータサイエンス的マネジメントにおける Check とは，企業パフォーマンスに影響を与える可能性を秘めた，常とは異なる現象が無いかどうかを判断する行為である．筆者は，図 1 のサイクルを回すことが，AI, IoT 時代でも基本理念であると考えている．

この種の統計的問題が提起された上で，これまで想定していない異常の原因系を探索するのが分析機能である．Cox and Donnelly(2011)は，この分析プロセスを次のようにまとめており，下記のプロセスを“Ideal Sequence”と呼んでいる．

研究すべき問題，仮説の定式化
関連するデータの探索と適切なデータを採取する研究の計画と実施
データ解析
適切な意思決定に繋がる結果の解釈

これは，実証的科学の文法(Pearson, 1892)の確立を目指し，記述統計的方法を整備した近代統計科学の設計意図から視れば，極めて正当なプロセスである．しかし，統計的品質管理のパイオニアである Shewhart(1939)の示唆により，最善の統計モデル当てはめに基づく異常検知こそ研究すべき統計的問題を示唆することが，図 1 でも示唆したように日本の産業界では常道として確立していたはずである．

統計的機械学習発展前は，異常値ないしは外れ値が生じるのは，当てはめた統計モデルの不完全性，例えば非線形性・交互作用・同定可能な潜在構造・潜在クラス（変化点）の無視に起因して生じることが多かった．これをデータ解析の過誤ないしは未熟から生まれた「見かけ上の外れ値」と呼びたい．しかし，今日の統計的機械学習技法は，これらの構造はその同定に足る教師データさえ系統的に入手できれば，自動的に考慮できる

ようになった．ただ，アルゴリズムが学習した入出力構造の可視化ないしは解釈が難しい，いわゆるブラックボックス型モデル当てはめになっただけである．

逆に，未知の原因系による「本質的な外れ値」の原因を分析するには，既存データではなく，その原因に関する様々な仮説に基づく追加変数の新たな採取は不可欠の筈である．ただし，これまで管理してきた入力変数が，出力の期待値だけではなく，極値(Extreme Value)としての外れ値発生に影響を与えるといったことは否定できない（椿, 2015）．異常値自体をある確率で検知する統計モデルを考えることも可能ということである．

最適化機能は，実験計画法や品質工学の狙いそのものであるが，これまでの多変量解析ではその前工程としての分析機能が，後工程の便を意識しないことが多かった．統計的機械学習を含めた予測・制御を目的とするデータ解析でも実験計画法や品質工学で用いられてきた基本原理は，十分意識すべきである．

３．科学的データ解析に潜む基本原理

ここでは，メカニズムが支配的なモノつくりに関わるデータ解析原理を再考察する．

1）目的変数の選択原理

1-1)因果仮説の配慮：目的変数は結果変数

産業界で統計モデルの当てはめを行おうとする素朴な目的は，現象の解釈よりも重要な特性を安価に予測したいということが多い．統計的機械学習を漫然と使いたいといっても何を予測して，価値を得たいのかを自覚してなければ話が始まらない．

一方，その変動のメカニズムを考えた上で予測するのが解釈可能な統計モデルを当てはめるのが，これまで前提であった．純粋に予測を最適化問題に帰着させる統計的機械学習との最大の違いは，統計モデルとは，原因が結果に与えるメカニズムを記述したものということである．メカニズムの解釈ができれば，データから得られた知識は，学習に用いたデータを超えた範囲での利用可能性，すなわち外挿可能性が生じる．

実際，結果から原因を予測する逆問題において両者は根本的に異なる予測方式を与える．トレーニングデータが予測すべきデータの標本空間からの代表性のあるサンプルでない限り，統計的機械学習から導かれる予測は，逆問題では，予測バイアスの原因となる情報までもデータから学習するために失敗する．田口の T 法は，単一の特性を逆問題において予測しようとするときのある意味で最適な方法を示唆している（寺本,椿, 2018）．

このようにデータ解析では，目的変数は結果変数あるいは被説明変数と呼ぶべきである．逆に説明変数は結果変数の変動の原因候補となる要因でなければならない．相関分析は，この種の因果関係に踏み込まないが，回帰分析は分析者が仮説として想定している因果関係を前提にしている．

1-2）潜在的結果変数の探索～２つの異なる相関発生の原因：共線構造と因子構造

解析目的が事前に定まっていないのは，基本的に好ましくない．しかし，多変量データの相関分析によってデータに隠れた予測式構造等を探索することが可能になる場合がある．これを通じて何を制御するためには何を使うことができるのかといった知識を獲得できる可能性もある．

先ず，ある観測変数群が他の観測変数の変動の原因となっている場合，当然両者には強い相関が発生する，これは直接的な関係性から発生する相関で多重共線性と呼ばれることが多い．決定係数が高い回帰関係が多変量データに潜むということは，多変量データの上手い一次式（目的変数-説明変数による回帰関係）をつくると，その分散が著しく小さくなることである．従って，主成分分析を行えば，最小固有値に対応する主成分負荷量を観察すれば，その種の共線関係の発生が示唆される場合が多い．通常は，負荷量の絶対値が最も大きい変数が，目的変数候補となる．Mardia et al.(1979)は，この種の手続きを行い，共線性で説明される結果変数を

あらかじめ除外することが，潜在因子で相関を説明する因子分析の前処理として必須と考えた．なお，共線性や因子の存在の有無の推測には，固有値の対数 Scree plot が有効である(椿，2011)．

もちろん，共線構造のみが顕著な相関構造の場合にはグラフィカルモデリング(日本品質管理学会テクノメトリックス研究会編, 1999)を用いることで線形共線構造の全貌が理解できる．しかし，逆に潜在因子からの影響を多くの変数が影響を受ける構造が存在する場合は，潜在因子を明示的に統計モデルに取り入れないとグラフィカルモデリングでは簡便な現象解釈は困難になる．

2) 主要な説明変数の選択－信号因子の選択と効率性分析

製造業の問題解決では，目的変数と単純な関係性があるべきと考えられる原因系変数が存在することがある．特に比例関係が成立する主要な原因系変数は，品質工学では信号因子と呼ばれている．計量経済学における生産関数同定でも，この種の変数の導入は常套手段である．電流を倍にすれば電力は倍になる，企業が従業員数や資産を全て定数倍すれば，規模の効果により，売上高も定数倍されるといった概念である．統計的機械学習全盛期に何故このような概念が必要かというと，２つの意味がある．

一つは，データ解析で得たい知識の中でも，信号因子の目的変数に対する影響を他の説明変数がどのように修飾するかを知ることの意義が大きいことである．統計的機械学習でこの種のことを考慮しようとすると，目的変数をそれと比例関係にある説明変数で割った効率変数（生産性）を目的変数とする工夫がいる．

第２の意味は，目的変数の「分散関数(Variance Function)」，すなわち平均値と分散との関係性が分かることである．予測最適化のための統計的機械学習では，単純に予測最小二乗誤差を最適化がなされる．しかし，平均と分散との関係を推論して，適切な情報量規準を最適化しなければ，情報量的観点からは最適予測をしたことにはならない．目的変数比例関係にある主要な説明変数が定数倍されれば，目的変数の平均も定数倍されるが，このとき目的変数の分散が変わらないのか，あるいはやはり定数倍になるのかといったことを知り，分散関数に整合した尤度(正確には擬似尤度)に基づく最適化をすることが，統計モデル当てはめでのみならず，統計的機械学習でも重要である．歴史的には，比例関係にある目的変数と説明変数の両者に対数変換や平方根変換などを施し，分散が一定となる変数変換を探すことも行われた．この変数変換が分かれば，最小二乗型統計的機械学習でも，目的変数と主要な説明変数についてはその種の変換（分散安定化変換）を施すのが良い．

3) 説明変数（原因系変数）の分類―外生性・制御性・因果性

主要な制御変数としての信号因子が少数個選択出来たら，次に結果系変数と信号因子との関係性に影響を与える可能性のある要因と呼ばれるデータを採取するのが，古典的データ解析の常道であった．わが国の統計的実験計画法の業績の一つとして，田口(1957)による「因子の分類」がある． 信号因子のように技術的に変数の値が制御可能な制御因子と，計測はできるが制御不可能な標示因子は特に重要である．探索的データ解析でもこの種の観点での説明変数の分類は必要である．更に，説明変数間の相関構造を共線性として，共線性の原因となる説明変数を解析から除外するのではなく，どの説明変数群がどの説明変数群に影響を与えているかといった事前知識の整理が重要である．

多変量データ解析では，事前に存在し制御性の一切ない標示因子を外生変数と呼ぶことがある．一方，ある種の外生変数は，制御因子変数の値の設定に影響を与える場合もある．気温や湿度の高い日には装置の設定を変えるといった具合である．この多変量データの因果関係上の最上流に位置づけられる外生変数群を第 0 次説明変数群と呼ぶことにしよう．なお，第 0 次説明変数群は直接目的変数に影響を与えるかもしれない．

因果階層上．その次に位置するのが信号因子を含む完全制御可能な説明変数群で，これを第 1 次説明変数群と呼びたい．

更に完全に制御可能ではないが, 第一説明変数群の影響を受け, かつ目的変数に影響を与える変数群がある. これは, 第 2 次説明変数群と呼びたい. McCullagh が指摘したように喫煙と死産との関係性を分析しようというときに, 喫煙の影響を受ける早産という中間変数をデータ解析に導入すると喫煙の直接効果がマスキングされる問題が生じる. ブラックボックス型統計的機械学習で制御因子の影響を分析したい場合に, あまり知恵を使わない方法が必要ならば第 2 次説明変数群を全て解析から除外して, 重要度指標などを算出することが, 統計家としては情けないのだが勧められる.

これら, 0 次から 2 次までの変数群の影響を結果系変数群は影響を受けると考えるのである. もし, 制御変数自体も時系列的に順次段階的に設定されるのであれば, その段階間にある標示因子的説明変数を一括りに分類する必要がある.

4) 説明変数の階層化 – 説明変数の分解

日本の実験計画法では, 原料条件と加工条件との最適化のために分割法実験が用いられてきた. MA 研では, 芳賀, 奥野(1984)が, データ解析にこの考え方を導入し, 分割型回帰分析と呼んだ. 今日, マルチレベルモデルないしは階層ベイズモデルと呼ばれる考え方である. 大きさ n のデータセットにおいて説明変数 x_i, i=1,…,n は, 本来それを中心化した説明変数 $(x_i-\bar{x})$に変換しても, 説明変数に切片項が含まれている限り, 本質的に同等の分析結果が得られる. 一方, 説明変数が事前に層 j, j=1,…J に層化されている場合, 説明変数を一元配置的に $x_{ji} = \bar{\bar{x}} + \left(\bar{x_j} - \bar{\bar{x}}\right) + \left(x_{ji} - \bar{x_j}\right)$と分解する. このとき, 説明変数を分解して右辺第 2 項の群平均変数と全体平均との差と第 3 項のデータの群内平均からの偏差の 2 変数にする. 当該説明変数自体のミクロな影響を示す$\left(x_{ji} - \bar{x_j}\right)$と, 群平均のマクロな影響を示す$\left(\bar{x_j} - \bar{\bar{x}}\right)$との両者が予測に有用な可能性があると考えるのである.

5) 第 1 次・第 2 次説明変数群の回帰残差への変換

説明変数のもう一つの分解は, 直交性に第 0 次説明変数群から第 1 次説明変数群への影響, 第 0 次, 第 1 次説明変数群から第 2 次説明変数群への影響を考える場合にも拡張できる. 第 1 次説明変数群に属する変数 x_{1i} に対して, 第 0 次変数群を説明変数とした回帰予測値を$\widehat{x_{1i}}$, 回帰残差を e_{1i} と表現し, x_{1i} の代わりに e_{1i} を新たな説明変数とするのである. 回帰残差は説明変数とは無相関になる. 一方, 第 1 次説明変数群の残差間相関は, 偏相関と呼ばれる. 同様に第 2 次説明変数群については, 第 0 次, 第 1 次説明変数群から説明されない回帰残差に変換するのである.

6) 第 0 次変数群ならびに第 1 次変数群, 第 2 次変数群の回帰残差の尺度化

深層学習と呼ばれるニューラルネットワーク系の統計的機械学習の本質は, ネットワーク構造の利用と, 変化構造をも含む非線形性の自動検出である(Asano, Tsubaki, Yosizawa, 2002). このネットワーク構造とは, 説明変数の尺度化（線形結合）の利用である. 多変量データ解析として利用されてきた尺度化は, 主成分分析や因子分析と呼ばれる方法論である. 予測のためのデータ解析でも, 主成分回帰分析や PLS のような方法論が活用されてきた. 第 0 次から第 2 次の説明変数群にも相関構造や偏相関構造がある以上, 主成分分析で直交尺度化を説明変数群に行って回帰分析を行うことが有用である.

一般に, 主成分分析ではどの主成分に意味があるかということを示す外的基準はないが, 主成分回帰分析では, 目的変数との相関係数の絶対値順に, 説明変数の重要性が位置づけられる. 相関係数の絶対値に一定の閾値を与えて, 必要な主成分までを用いて予測を行うことができる. MA 研では, モデル選択基準による説明変数選択が推奨され, あまり主成分回帰分析の有用性については検討されなかった. 何故ならば, 結果の解釈可能性に難があったからである. これは, 今日の深層学習がブラックボックスアプローチであり, 解釈困難とされていることと類似である. Miyamoto and Tsubaki(2001)は, 予測に有効と判断された主成分空間に対して因

子分析で良く用いられてきた直交回転(Varimax Rotation)を用いて回転後の主成分を説明変数とすることでにより，主成分回帰分析の解釈可能性を改善する方法を提案した．直交説明変数を絞り込むモデル選択には統計的検定，赤池情報量規準，最近では Lasso も使えるが，説明変数が直交実験計画同様，無相関になっているため，解釈可能性が本質的に高くなると共に，尺度変数については信頼性の高い重要度指標が得られる．

なお，0 次から 2 次までの説明変数群．更に目的変数を第 3 次変数と呼ぶことにすれば，より高次の変数群に対しては全て要因効果があるというモデルから出発して順次変数減少法によって必要な変数間だけに関係性が存在するというモデルへの絞り込みをモデル選択基準で行うこととなる．

グラフィカルモデリングでいう階層的因果構造と有効な尺度化戦術がデータに当てはまるのならば，ニューラルネットのような統計的機械学習でも，階層構造（深層化）を導入する方が効率的予測に繋がることが予期される．

7) 説明変数の非線形効果と説明変数間の交互作用効果への配慮

古典的重回帰分析では，説明変数 $x_1,\ldots,x_p$ と目的変数 y との関係を正規線形モデル(Normal Linear Model)で表現するため，２つの戦術が用いられてきた．一つは目的変数と説明変数とを変数変換し，$f_0(y)=\beta_0+\beta_1 f_1(x_1)+\ldots+\beta_p f_p(x_p)+\varepsilon$ といった統計モデルを当てはめることである．目的変数と説明変数の関係性の固有技術的考察，あるいは散布図を眺めたり，回帰分析の決定係数をチェックしたりといったデータ解析的検討を通じて，線形性や等分散性の成立しやすい変換を探る手続きである．

もう一つの戦術は，説明変数の二乗項 x_1^2 や 2 つの説明変数の積の項(交互作用効果項)x_1x_2 を説明変数群に追加して，回帰曲面の近似度を上げるというものである．基本的には多項式回帰，あるいは実験計画法でいう応答曲面法(Response Surface Design)を予測分析に適用する方法である．二次式項を追加することで，最適な制御条件なども議論することができる場合もある．

一方，1970 年代に勃興した Nelder and Wedderburn(1972)の一般化線形モデル(GLIM, Generalized Linear Model)では，目的変数の期待値パラメータ μ ではなく，それをリンク関数と呼ばれる非線形関数に変換した $\eta(\mu)$ を推論の対象とする接近が提唱された．更に GLIM とセミパラメトリック回帰と結合して，GAM(Generalized Additive Model) が誕生した(Hastie and Tibshirani, 1990)．GAM では，リンク関数は与えたうえで，$\eta(\mu)=\beta_0+\beta_1 f_1(x_1)+\ldots+\beta_p f_p(x_p)$ というモデルが当てはめられるが，変換の非線形関数形 $f_j, j=1,\ldots,p$ は，スプライン関数などを用いて自動的にデータから生成される．

3 層ニューラルネットワーク回帰は，基本的に応答曲面を $\beta_0+\beta_1 f_1(z_1)+\ldots+\beta_q f_q(z_q)$ としたもので，GAM に極めて近い．重要な差は，説明変数の尺度化 $z_j=\sum_{k=1}^{p} w_{jk}x_p$ を自動的に探索しているということ，f_j はノンパラメトリックにデータから決めるのではなく，ロジスティック関数など既知の非線形関数形を用いて，そのパラメータを推定することである．

ニューラルネット系の予測モデルの特長は，２つある．

第 1 の特長は，尺度化が，ネットワークのウェイトとして自動的に学習されることである．実際，ニューラルネット系の統計的機械学習では，尺度化と非線形性を上手く利用することで，説明変数の交互作用の表現も達成している．例えば，最も単純な説明変数間の交互作用を表す説明変数の積の項は，$4x_1x_2=(x_1+x_2)^2-(x_1-x_2)^2$ 和の尺度と差の尺度をと二乗項とを用いると表現できる．第 2 の特長は，典型的なニューラルネットワークであるボルツマンマシンでは，基底関数に，ロジスティック関数という，温度パラメータ（尺度母数）が低い時には一次関数といった滑らかな関数，温度パラメータが高いときには，ステップ関数という不連続関数を近似できる関数形が用いられていることである．

実は，3 層ニューラルネットは，一般化加法モデルと尺度の最適化とを結合した Friedman and Stuetzle(1981)の射影追跡回帰(Projection Pursuit Regression)の特殊な場合に過ぎない．ただし射影追跡回帰では，一般化加法モデルを継承しているために，基底関数にスプライン関数を用いるために，基底関数に滑らかさの制約が導入されている（実際にはある程度の急変化は表現される）．このため，潜在クラス（変化点構造)，すなわち予測式自体の教師無し分類が必要な場合には，ニューラルネット系機械学習より若干予測精度が落ちる場合がある．しかし，尺度や基底関数は明示されるので解釈可能性は高い．

なお，交互作用や非線形性を可視化する，より初等的接近は，探索的層別を基本原理とする第 2 世代人工知能の CART（Breiman et al. 1984）を使うことである．メカニズムに支配されるモノつくりでは，主要なメカニズムの同定する汎用知識の獲得は重要だが，メカニズムよりはサービスプログラムの前後関係などの有効性を調べたいのならば，今でも適切な説明変数を用意した CART を用いることが勧められる．

４．統計的機械学習への途とこれからの人材教育

3 章あるいは椿(2018c)で概説したように，今日の人工知能，統計的機械学習技術が全くこれまでの統計解析から独立に生まれてきたわけではない．統計的機械学習が導入している統計学やデータ解析の基本原理の中で，これまで紹介しなかったものには，正則化，アンサンブル学習，カーネルトリックがある．正則化はモデル選択（説明変数選択）に代わる概念で，統計的にはパラメータに事前分布を想定するベイズ統計利用のことである．アンサンブル学習は，複数の予測値を適当なウェイトを付けて，束ねより良い予測値を構成する方法である．基本的は品質工学における T 法は，その代表的な技法である．カーネルトリックは，超高次の多項式回帰のための計算を効率的に内部処理する方法で，数理的には計量的多次元尺度構成法と類似である．この種の数理は Support Vector Machine などで利用される．

実務的に重要なのは，統計的機械学習，すなわち人工知能のユーザー基礎教育は，デヴェロッパー教育としての，Python や R での統計的機械学習プログラミングといった情報教育ではない．筆者は，２つの側面を重視すべきと考える．一つは，一般論として何で機械はデータ解析が賢くできるのかという基本原理の習得である．もう一つは，適切なデータセットとソフトウェアに基づく統計的機械学習プロセスの実践活用である．その上で，統計的機械学習から導かれた結果を解釈すること，つまり自身のデータセットについて，人工知能はどのような賢いことを行ったかを理解し，背景にあるメカニズムという汎用知識に迫ることである．椿(1999)は，第 2 世代人工知能が普及した際に，社会人に対する予測モデル考案のためのデータサイエンス教育を下記の「仮説成長型データ解析」に一新することを提唱し，筑波大学大学院ビジネス科学研究科の多変量解析第 1 の講義を下記の順番で実施した．

STEP 1 第 2 世代人工知能による論理型（層別型）知識（仮説）の獲得

STEP 2　散布図平滑化技法（GAM）による論理型知識をパターン型知識に変換

STEP 3　GAM で発見されたパターン型知識をメカニズム型（数式型）知識に変換

第 2 世代人工知能や GAM といった柔軟な予測モデルに人間が創る回帰モデルが既に敵わないことを前提にした教育で，それらを目標にして解釈可能なメカニズム型統計モデルを構築することが，必要という立場であった．しかし，今日の統計的機械学習は，当時の水準をはるかに超えている，科学的データ解析の様々な原理を駆使して，なぜその種の予測精度が生じるかを探り，外挿可能性の高い予測モデルを提示することが統計家の使命となりつつある．そのような時に，ブラックボックス型統計的機械学習に学ぶための統計教育はやはり仮説成長型に沿ったものにすべきという考えは強くなりつつある．20 年前と異なるのは，ベンチマーク用

の方法論が，CART や GAM ではなくなり，深層学習，Random Forest, PPR といった強力な予測ツール群に変化しただけである．

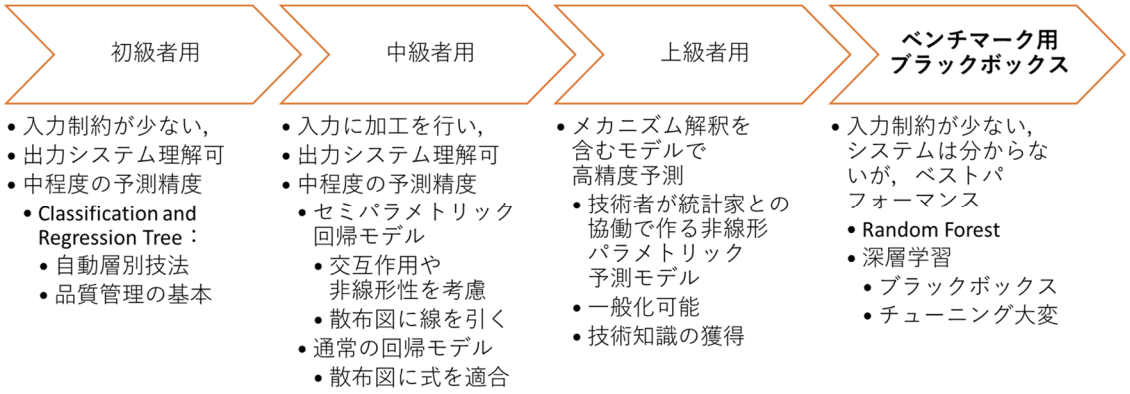

図 2 次世代データ解析教育の階層性

その意味では， 図 2 に示したように，QC 七つ道具の層別原理さえ理解できれば，簡単に理屈が分かる第 2 世代人工知能を先ず初級者は大規模データで実践的に学習し，それをアンサンブル学習によって予測最適化した Random Forest をブラックボックス型接近として教え，これをベンチマークとして，メカニズムに迫るデータ解析の知，すなわち統計モデルアプローチを徐々に深化させる教育に進めるのが良いと考えている..

総務省，（独）統計センター，（一社）日本統計協会は，SSDSE を用いた第 1 回統計データ分析コンペティションを，研究開発法人科学技術振興機構，（一社）日本統計学会の後援で実施した．このコンペティションで高校 1 年生の伊藤(2018)が Random Forest を使った分析を行い，審査員特別賞を受賞した．コンペティションの目的は社会問題のデータに基づく解決教育の促進が趣旨だったので，総務大臣賞には選出されなかったが，高校生が，独習で Random Forest に関する一定水準の統計技術報告をで行えたことは，次世代データサイエンス教育に様々な可能性があることを示したものである．

SSDSE のようなデータサイエンス教育用の標準データセットとして様々なものが今後も開発され，ある程度複雑な構造を持つデータセットを多くの方が様々な分析目的を設定し，データ分析や統計的機械学習を行い，更にその実践の知を社会で共有することも，今後のデータサイエンス教育にとって重要と考える．MA 研全盛時代に，吉澤，芳賀(1992,1997)が，企業の実データファイルを公表できたことは，日本の産業競争力の源泉でだったと考えている．知の共有の文化をデータ駆動型時代を前に再興することを切望して本稿を閉じたい．

参考文献

Asano, M., Tsubaki, H. and Yosizawa, T. (2002) Effectiveness of neural networks to regression with structural changes, *Applied Stochastic Models in Business and Industry*, 18(3), 189-195.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification And Regression Trees*, Chapman and Hall.

Breiman, L. (2001) Random Forests, *Machine Learning*, 45 (1), 5–32.

Cox, D. R. and Donnelly, C. A. (2011) *Principles of Applied Statistics*, Cambridge University Press.

Friedman, J. H. and Stuetzle, W, (1981) Projection Pursuit Regression, *Journal of the American Statistical Association*, 76, 817-823.

Hastie, T. and Tibshirani, R, (1990) *Generalized Additive Models*, Chapman and Hall.
Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*, Academic Press.
Miyamoto, M. and Tsubaki, H. (2001) Measuring Technology and Pricing Differences in the Difital Still Camera Industry Using Improved Hedonic Price Estimation, *Behaviormetrika*,28(2), 111-152.
Nelder, A. and R. Wedderburn, W. M.(1972) Generalized Linear Models, *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384.
Pearson, K. (1891) *The Grammar of Science*, Walter Scott.
Shewhart, W.A. (1939) *Statistical Method from Viewpoint of Quality Control*, Graduate School of the Department of Agriculture.
Tukey, J. (1960) The Future of Data Analysis, *The Annals of Mathematical Statistics,* Vol. 33(1), 1-67.
Tukey, J. (1977) *Exploratory Data Analysis*, Pearson.
伊藤寛子(2018)機械学習による 15 歳未満人口の推定，https://www.nstac.go.jp/statcompe/doc/2018H-tokubetu.pdf
奥野忠一，久米均，芳賀敏郎，吉澤正(1971)多変量解析法，日科技連.
奥野忠一，上郡長昭，入倉則夫，片山善三郎，伊東哲二，藤原信夫(1986)工業における多変量データの解析，日科技連.
田口玄一(1957)実験計画法（上），丸善.
椿広計(1999）データサイエンスの社会人教育（特集 データサイエンス）--（第 1 部 データサイエンス登場), Keio SFC review, 3(1), 38-43.
椿広計(2011)多変量データとパネルデータの相関構造に関する注意と試み，Latent Dynamics 研究会第 2 回 Latent Dynamics Workshop （LD-2）予稿集．1-5．http://latent-dynamics.net/02/LD-2_proc.pdf
椿広計(2015）ビジネスは統計科学足りえるか？，応用統計学(応用統計学会誌)，Vol. 44(1), 17-30.
椿広計(2018a)データサイエンスと品質マネジメント－その方法と教育，品質(日本品質管理学会誌），Vol. 48(4), 27-32.
椿広計(2018b）小学校・中学校における算数・数学教育の中に如何にして統計的考え方を導入すべきか？,特集「統計教育の新展開」，統計数理，66(1), 3-14. http://www.ism.ac.jp/editsec/toukei/pdf/66-1-003.pdf
椿広計(2018c）統計を深く知る 古典統計学対話 : 統計学から観た統計的機械学習,統計(日本統計協会機関誌), 69(1), 35-41.
寺本顕武，椿広計(2018)計測のための統計，計測自動制御学会編，計測制御テクノロジーシリーズ第 4 巻，コロナ社.
統計センター(2018）教育用標準データセット（SSDSE）の解説, https://www.nstac.go.jp/SSDSE/SSDSE2018_kaisetsu.pdf
日本品質管理学会テクノメトリックス研究会編(1999)グラフィカルモデリングの実際，日科技連.
芳賀敏郎(1984)対話型データ解析システム，応用統計学(応用統計学会誌)，Vol. 13(3), 125-138.
芳賀敏郎，奥野忠一(1984)分割型データの回帰分析，日本品質管理学会第 14 回年次大会.
吉沢正，芳賀敏郎編(1992)多変量解析事例集第 1 集，日科技連.
吉沢正，芳賀敏郎編(1997)多変量解析事例集第 2 集，日科技連.

■公開資料ページ
弊社ウェブページで各種資料をご覧いただけます http://www.i-juse.co.jp/statistics/jirei/

■お問い合わせ先
(株)日科技研 数理事業部 パッケージサポート係 http:/www.i-juse.co.jp/statistics/support/contact.html