分析データの MI 活用

鈴　木　啓　幸

1　は じ め に

本解説では，様々な分析装置から得られるデータを機械学習（machine learning）を用いて解析する方法について解説する．具体的な例としてプラスチックの特性予測について紹介する．機械学習に馴染みのない材料研究者が材料開発や機能開発の際に機械学習を利用するきっかけになれば幸いである．

分析データは材料組織構造と材料特性を関係づける上で重要であることは論をまたない．実験系の材料研究者は試作・分析・特性評価・考察のサイクルを繰り返すことで，高性能な材料を開発したり新奇な機能の発見に至る．ここに理論や計算科学の知見も加わると考察が更に深まりこのサイクルが深謀遠慮する．考察は，試作条件，分析データ，特性データとの関係性を知識も交え捉えて生じている現象を推察/理解し，次のサイクルに何をすべきかを見いだすことが要で，属人化していることが多い．傑出した材料研究者はこの属人的な要素が秀でている．一方，人工知能（artificial intelligence, AI）の力を用いてサイクル数を減らす試みは各材料分野を横断して近年非常に盛んであり，マテリアルズインフォマティクス（materials informatics, MI）と呼ばれている（材料研究の行動変容を促す 10 年プロジェクトが国の研究機関を主体に進行している[1]）．上記の属人的なプロトコルに対して標準化する取組みと見て取れなくもない．AI は膨大なデータの中から人が気付かなかったデータ間の関係性を見出して材料研究者に提起し，材料研究者はその関係性を起点に新たな気付きを得て次にすべきことを賢く実行するのである．限定されたパラメータの中で特性を適正化するようなのサイクルの場合には次にすべき試作条件をベイズ最適化（Bayesian optimization）などを使用した逆解析により提起することもできる．この MI を遂行する上で重要なのがデータである．試作・分析・特性評価の各工程で生成されるデータ（以下，実験データと総称）の処理方法について，一例として紹介していきたい．

Analytical-data Utilization for MI.

2　機械学習とは

本節では，材料情報科学の分野に頻繁に登場する用語を導入しつつ機械学習について概説する．MI に供するデータは具体には実験データの各パラメータであり記述子（Descriptor）と称する．記述子の中で予測したい記述子を目的変数（objective variable），その予測に用いる記述子を説名変数（explainable variable）と称する．前節になぞると特性評価で得られる性能指標が目的変数で，試作や分析の各パラメータが説明変数に該当する．ただし，ユースケースによって当然変わる．目的変数ベクトルの集合 $\mathbf{Y}$（部分集合を $\mathbf{y}_i$），説明変数ベクトルの集合 $\mathbf{X}$（部分集合を $\mathbf{x}_i$）とすると $\mathbf{Y}=\boldsymbol{F}(\mathbf{X})$ と表記され，機械学習は関数（/写像）$\boldsymbol{F}$ を明らかにする．この集合間の関係性を示す関数 $\boldsymbol{F}$ は物理化学の法則・理論として知られているものも含み，機械学習は解析関数に限定されずに決定木関数やネットワーク関数といった非常に柔軟な関数で関数 $\boldsymbol{F}$ を記述できることに利がある．

機械学習の手法には，教師あり学習（supervised learning），教師なし学習（unsupervised learning），強化学習（reinforcement learning）といった分類の仕方がある．教師あり学習は学習用データを用いて上記の関数 $\boldsymbol{F}$ を明らかにし，未知のデータに対して既知となった関数 $\boldsymbol{F}$ を適用して Y を予測する手法である．関数 $\boldsymbol{F}$ が未知データに対しても当てはまる場合には予測精度は高いが，生じている現象が異なるなど関数 $\boldsymbol{F}$ が担保されない場合には予測精度は低い．教師なし学習は $\mathbf{X}$ だけから中に潜む関係性 $\boldsymbol{F}$ を学習する手法であり，$\mathbf{Y}$ を必要としないのが最大の長所であるが教師あり学習より予測精度が劣る．強化学習は教師ありと教師なし学習の間に位置する．要素ベクトル間の $\mathbf{y}_i=\boldsymbol{F}_i(\mathbf{x}_i)$ という関係性を学習（教師あり学習）して次の xj を算出して行動（実験）して yj を取得して $\mathbf{y}_j=\boldsymbol{F}_j(\mathbf{x}_j)$ の関係性を学習する，というステップを行動報酬に基づいて繰り返すことで関数 $\boldsymbol{F}$ を自律的に明らかにしていく手法である．状況に応じて関数 $\boldsymbol{F}$ が変化するような場合に有効である．表 1 には機械学習の代表的な各手法を記載した（亜種も沢山ある）．ユーザはユースケースに応じて適切な手

表 1　機械学習の手法例

機械学習の分類	使用される手法の例
教師あり学習	ガウス過程（Gauss process），リッジ（Ridge），ラッソ（Lasso），サポートベクターマシン（Support-vector machine），ナイーブベイズ（Naive Bayes），ランダムフォレスト（Random forest），勾配ブースティング決定木（Gradient boosting decision tree），ニューラルネット（Neural network, NN）など
教師なし学習	次元削減・可視化：主成分分析（Principal component analysis, PCA），変分オートエンコーダ（Variational autoencoder, VAE），非負値行列因子分解（Non-negative matrix factorization, NMF），t 分布型確率的近傍埋め込み（t-distributed stochastic neighbor embedding, t-SNE），均一多様体近似と射影（Uniform manifold approximation and projection, UMAP）など クラスタリング：K−平均法（K-means），凝集型クラスタリング（Agglomerative clustering），階層密度に基づくノイズあり空間クラスタリング（Hierarchical density-based spatial clustering, HDBSCAN）など
強化学習	モンテカルロ（Montecarlo, MC），時間差（Temporal difference, TD），Q 学習（Q-learning），深層 Q ネットワーク（Deep Q network, DQN），サルサ（State-action-reward-state-action, SARSA），Actor-critic など

法を選択するために各手法の特徴を把握していることは望ましい．ただし，必ずしもベストな手法を選ぶ必要はなくて（何がベストプラクティスか分からないことがほとんど）ユースケースの目的が達成できる精度が出れば十分である．

3　実験データの処理

本節では，一例として「ポリプロピレン（PP）の機械特性を分析データから予測する」というタスクにおける実験データの処理フローを紹介する．

まず目的変数を設定する．今回の場合は引張機械特性になる．続いて引張機械特性に影響を与える材料因子について，材料知識を基に列挙して更にそれらを評価可能な計測手法も列挙する．ここに生成 AI を使うこともできる（検索拡張生成（retrieval augmented generation, RAG）技術[2)3)]によって専門化していないと材料研究者が満足する回答は得られないであろう）．データ取得コストも鑑みて説明変数に選ぶべき計測手法を選定する．ここでタスクの成否がまず裁定される．本タスクの場合の関係性の一部を切り出すと図 1 のようになる．分子量分布，エチレン分率，混合物（フィラー，エラストマー）のサイズ・量・種類など，酸化生成物，結晶比率，水酸基量，などが少なくとも引張機械特性に影響を与え，熱重量（thermal gravity, TG）計測，示差熱分析（differential thermal analysis, DTA），フーリエ変換赤外線分光（fourier transform infrared spectroscopy, FT-IR）計測，広角 X 線回折（Wide angle X-ray scattering, WAXS）計測などで直接/間接的に評価することができる．これら

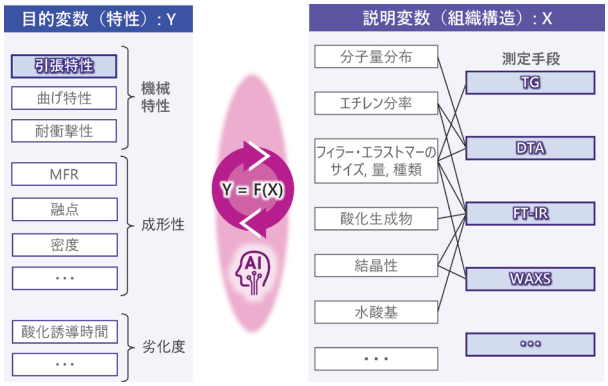

図 1　PP における特性と組織構造の関係性

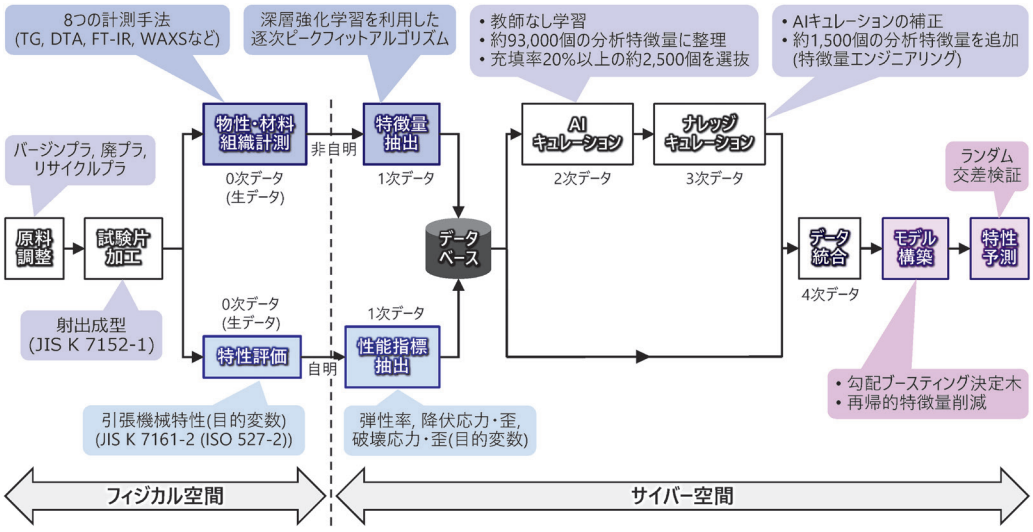

図 2　PP における実験データの処理フロー

計測手法から得られる情報が説名変数になる．目的変数の値は多くの現象が重畳された結果であるので，多方面から材料科学の関係性 $\boldsymbol{F}$ を探ると取りこぼしが抑えられる．熱，光，音といった具合に異なるプローブでの計測（マルチモーダル計測）となるように各計測手法を選定すると良い．

ここまでがタスクの「設計」で次に「データ処理」に移る（図 2）．フィジカル空間での 0 次データ（生データ）の取得が完了してもサイバー空間において機械学習にデータを供するまでには，計測データにおいて特徴量抽出とキュレーション，全データの統合と多くの前処理（データ加工）を踏まなければならない．最大で 4 次まで加工する計測データがある．依然として多くの労力を費やすことになるが，この前処理はタスクの成否が裁定される 2 番目のポイントになる．分析データの処理方法は材料種や現象に依存して必ずしも自明ではないので，これら処理を自動化するための解析アルゴリズム，キュレーション方法，データ統合方法の開発に著者は労力を費やした．データのレコード数（表で言うと行数）と分析データ数（表で言うと列数）が増えるとそもそも人力で処理することはほぼ不可能であるので，データ規模が大きくなると避けて通ることはできない．

では具体的に見ていこう．処理が必要な計測データはスペクトルや画像で得られることが多く，ここからエッセンスの情報を抽出しなければならず特徴量と呼ばれる．要は解析である．スペクトルデータに対して深層強化学習を用いた逐次ピークフィットアルゴリズムを考案してピーク特徴量を網羅的に抽出した．ピーク追加・削除と適正化を繰り返すサイクルにおいて，アルゴリズム内の動作パラメータを残差状況に合わせて動的に調整しながら逐次的にピークを追加・削除して全パターンフィッティングを行う，というものである．ピーク特徴量（位置，幅，面積など）という形で特徴量を抽出した．ピークフィット以外にも，バックグラウンド除去や平滑化などの処理を施したスペクトルを，ニューラルネットで解析して潜在空間の特徴量を抽出する方法もある（例えば，50 レコードの WAXS データ（9500 組の xy データ）を変分オートエンコード（variational autoencoder, VAE）で適正化処理すると潜在空間の特徴量はたったの 15 個になる）．ただし，潜在空間の特徴量は物理化学的な意味が失われているので，材料科学の知識に基づいたキュレーション（AI キュレーションの補正と特徴量エンジニアリング）を行うことが困難である．なお，説明可能 AI（explainable AI, XAI）技術[4)~6)]を使うとスペクトルのどのあたりが予測に貢献したかということは分かる．次に，抽出した分析特徴量と性能指標を紐づけてデータベース（Database）に格納して管理する．データベースでのデータ管理は，データの検索や共有など MI 以外の用途でも利便性が向上するが本タスクにおいては必須ではない．続いて，分析特徴量を表の形にするためにレコードごとに分析特徴量を整形・整理するキュレーションを行う．各レコード間でピークを比較し，由来が同じピークを同じ列に配置する．ピークに帰属ラベルを付けるなど，データにタグやメタデータを付けることをアノテーション（Annotation）と呼ぶ．ここに記載の AI キュレーションとは教師なし学習により説明変数となる分析特徴量をクラスタリングすることを指す．クラスタリング手法は各種ありそれらを組合せることもできる．データセットの規模を考慮しつつ試行錯誤で手法を選択する．多くの手法にハイパーパラメータ（Hyperparameter）と呼ばれる人為的に設定するパラメータがあり，これも通常は試行錯誤で適切な値を設定する．このようにキュレーションされたデータは材料科学の知識と照らし合わせると誤配置したものがあるので，ナレッジにより補正すると予測精度が向上する．さらにその際に既に知られている材料科学の法則・理論，解析方法を適用して新たに分析特徴量を算出すること（特徴量エンジニアリング）は，予測精度を向上させる上で非常に有効である．例えば今回の場合，ブロック共重合体のポリエチレン（polyethylene, PE）の含有比率，結晶相の比率，無機添加物の比率などである．前処理の最後の工程は，キュレーションした 8 種類の分析データ（説明変数）と機械特性データ（目的変数）を一つの表にデータ統合する．こうして機械学習に供する表形式のデータが整うことになる．後は機械学習を行うのみである．

4　機械学習による特性予測

本節の機械学習は教師あり学習で行う．教師あり学習は回帰，ランキング，クラス分類に大別される．回帰は値そのもの，ランキングは順位，クラス分類はクラスのラベルを予測するタスクである．本タスクは特性予測なので回帰である．モデル構築工程で学習用データを用いて関数 $\boldsymbol{F}$ を明らかにする．次の特性予測工程で未知のテスト用データについて明らかになった関数 $\boldsymbol{F}$ を適用して予測特性値を出力して実測値と比較して関数 $\boldsymbol{F}$ の精度を評価する（図 2 参照）．予想したいテストデータがある場合を除き，手持ちのデータセットを学習用とテスト用に分けてモデルの精度を評価する簡便な方法があり，交差検証（cross-validation）と呼ぶ．交差検証にもランダム分割，シャッフル分割や説明変数の特定のラベルを使用する層化 k 分割，グループ付き，など多くのデータ分割方法がある．この際に決してテスト用データの目的変数の値を使用してはならない．テスト用に分割したデータは最後の関数 $\boldsymbol{F}$ の精度評価にしか使用してはならず，ホールドアウト（holdout）しておかなければならない．予測結果を次の試行に使用してもならない．テスト用データの目的変数の値が何らかのデータ操

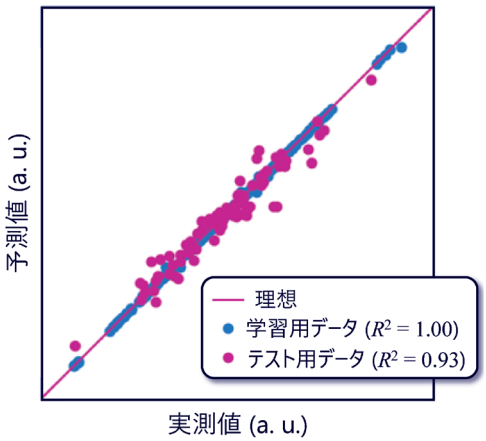

図 3　引張弾性率の yy プロット

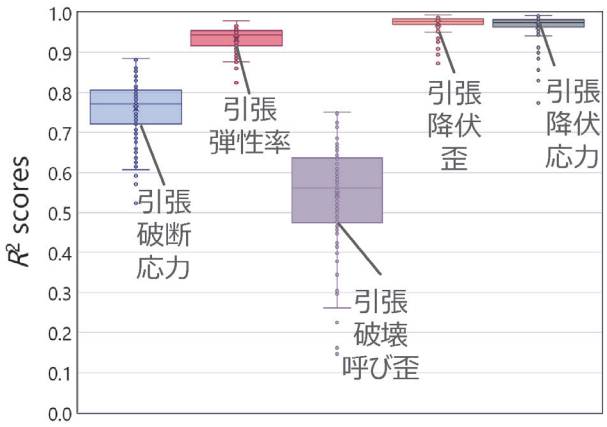

図 4　特性毎の予測精度

作によって学習用データに漏れることをデータリーケージ（data leakage）と呼ぶ．意図せずに漏れることが多分にあり，実態に沿ったデータ分割形式になっているかを考えるとミスを防ぐことができる．予測モデルの精度を評価するには交差検証を複数回実施してその統計量で予測モデルの精度を議論すべきである．特定のデータ分割で高精度な場合には分割形式を調査することでモデルの適用範囲を把握すべきである．モデルの評価指標には，回帰の場合には決定係数（R^2），平均絶対誤差（mean absolute error , MAE），平均二乗誤差（root mean squared error, RMSE）など，クラスタリングの場合には精度，適合率（precision），再現率（recall），f−値など，ランキングの場合には PR 曲線（precision-recall curve），平均相互順位（mean reciprocal rank, MRR），平均適合率の平均（mean average precision, MAP），減損累積利得（discounted cumulative gain, DCG）など，がある．これらは機械学習の手法とユースケースから設定する．複数の評価指標を出力してモデルを改善するために役立てることもできる．分割方法と評価指標の設計は極めて重要で，タスクの成否を裁定する最後のポイントである．

それでは具体的に見ていこう．機械学習の手法は勾配ブースティング決定木を使用した．説明変数がスパースで大量にある場合に非常に強力な非線形手法である．学習用データ 80 % とテスト用データ 20 % になるようにランダム交差検証を 50 回実施した．評価指標は決定係数に設定した．テストデータはホールドアウトし，学習用データについては更に 3 回繰返し 6 分割交差検証を実施して平均決定係数から予測モデルを構築した．学習用データの中をモデル構築用データとモデル検証用データに交差検証を用いて分割していることになる．この際に再帰的特徴量削減（recursive feature elimination, RFE）を併用すると効果的である．モデル構築とはパラメータの値を決めることである．ハイパーパラメータの適正化はベイズ最適化を使用した．予測精度を視覚的に

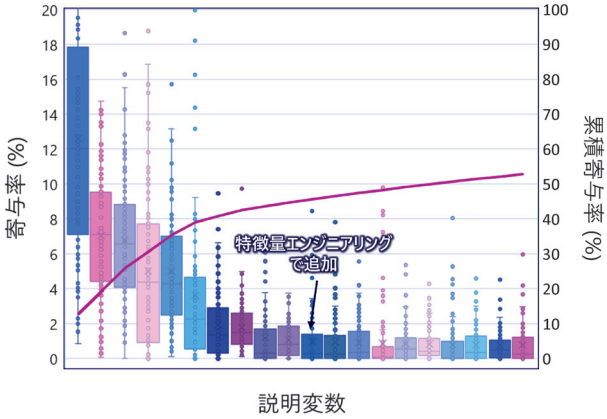

図 5　予測に有効な説明変数

表現する方法として横軸に実測値，縦軸に予測値を取った yy プロットと呼ばれるグラフが良く用いられる（図 3 参照）．学習用データの予測精度がテスト用データの予測精度より著しく高い場合は過学習（overtraining）と呼ばれる．学習用レコードに対してモデルの自由度を大きくし過ぎると生じる．学習用レコードを増やす，正則化などを使用してモデルの自由度を下げる（RFE 含む），モデル構築時の交差検証方法を見直す，モデル検証用データを用いた早期打ち切り（early stopping）などを行うことで回避できる場合が多い．予測精度は学習用データの方がテスト用データより若干高くなるのが常である．ランダム交差検証を 50 回実施した予測精度を箱ひげ図で示す（図 4 参照）．総じて良好に予測できていることが分かる．さらに XAI 技術を適用すると有効な説明変数の上位 20 個で寄与率 50 % 程度を越すことが分かる（図 5 参照）．予測に大きく寄与する説明変数はそれほど多くないことを示している．

5　最　後　に

紙面の都合で非常に駆け足で分析データを用いた MI についての一例を紹介した．材料研究の目的からとかく最後の機械学習を用いた特性予測に注目が集まりがちだ

が，データの前処理は極めて重要でそこでも教師なし学習や強化学習といった機械学習を利用できる．機械学習は作法を心得ていればライブラリで容易に実行できる．ただモデル構築方法にはノウハウがある．機械学習の専門的な良書[7]は背景技術を理解する上で有効である．一方で実用性を重視するならまずは「scikit-learn」という機械学習のライブラリを使用することをお勧めしたい．様々な機械学習の手法を簡便に試すことができる．開発者の著書[8]では本解説で紹介した用語も満遍なく紹介されている．

文　献

1） 文部科学省，〈https://dxmt.mext.go.jp/〉.
2） P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela: arXiv : 2005.11401.
3） Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang : arXiv:2312.10997.
4） 亀谷由隆：*IEICE Fundamental Review*, **16**, 83（3033）.
5） 恵木正史：日本セキュリティ・マネジメント学会誌 34（1）20（2020）.
6） S. Lundberg, S.-I. Lee : arXiv:1705.07874.
7） C.M. ビショップ：“パターン認識と機械学習” 上・下,（2012），（丸善）.
8） A. C. Muller, S. Guido :“Python ではじめる機械学習”,（2017），（オライリージャパン）.

後　送

鈴木 啓幸（Suzuki Hiroyuki）
株式会社日立製作所研究開発グループ計測インテグレーションイノベーションセンタナノプロセス研究部.（〒350−0395 埼玉県比企郡鳩山町赤沼 2520 番地）. 京都大学工学研究科材料工学専攻. 博士（工学）.
《現在の研究テーマ》マテリアルリサイクルを高度化する分析データの利活用技術.
《趣味》3 歳の娘と一緒に出掛けること.