

分析化学におけるデータ科学的解析

片 山 建 二

1 はじめに

分析化学におけるパターン認識といえば、多くの研究者はまずケモメトリクスを思い浮かべるだろう。赤外分光法や近赤外分光法で得られる混合物のスペクトルから主要な成分を見つけ出し、何が含まれているか（定性）、どのくらい入っているか（定量）を調べる手法は、分析化学の重要な技術として長い間発展してきた。これらのデータ処理技術は、主成分分析（PCA）や部分最小二乗回帰（PLS）といった統計的手法を使って、複雑な分析データから価値のある情報を取り出す方法として確立されている。

最近では、データサイエンス分野の目覚ましい発展によって、従来の分析化学データ処理に新しい風が吹いている。機械学習技術の発達が一番わかりやすい例であるが、他にも様々な手法が登場している。ケモメトリクスの基となっている行列分解手法、必要な特徴やパラメータを自動で見つけてくれる最適化手法、さらには理論計算で作ったシミュレーションデータと実際の測定データを組み合わせるデータ同化手法など、分析化学と相性の良い様々なデータ科学の手法が提案されている。

こうした新しい手法は、従来のケモメトリクスでは難しかった課題を解決してくれる可能性がある。例えば、大量のデータを効率よく処理したり、複雑な非線形の関係を見つけ出したり、リアルタイムで異常を検出したり、複数の測定方法で得られた異なる種類のデータを統合して解析したりすることができる。特に、人工知能（AI）技術を使うことで、パターンを見つける精度が向上し、より正確な予測ができるようになったり、品質管理を自動化できるようになったりすることが期待されている。

本レビューでは、分析化学分野で頻繁に扱われる各種データ（スペクトルデータ、時系列測定データなど）に対して、どのようなデータ科学的手法が適用可能であるかを俯瞰的に述べる。分析化学では、パターン・特徴・異常を検出することが重要で、これらの手法が分析化学

の実践にもたらす利益と課題について説明する。

このようなデータ科学を利用した計算手法は、日進月歩であるが、基本的な事項は共通するものが多い。いくつかのテキストや参考書を読むよりも、実際に計算してみることで深く理解することができる。本稿で記述する内容についても、手法名をキーワードとして、AIやWikiで検索し、明治大学の金子先生のWeb解説を読み¹⁾、試しにgoogle colab（gmailのアカウントに付随するpythonプラットフォーム）で計算してみたい。本稿では、化学を専門とする読者にとって、とっつきにくい専門の参考文献は引用しなかった。データ科学を始めようとしている読者は恐らく、用語と分野さえわかれば、まずAIに聞いてからの方が、データ解析の扉を開いてくれるものと思う。そのための入り口となってくれることを願い、本稿にそった試用プログラムを作成したので、ダウンロードしてご自分のWeb browserで実行してほしい²⁾³⁾。

2 分析化学におけるデータの種類

データサイエンス的手法を分析化学に適用する際、まず重要なのは対象となるデータの性質を理解することである。分析化学において扱うデータの種類について、その次元性に注目して整理する（図1）。

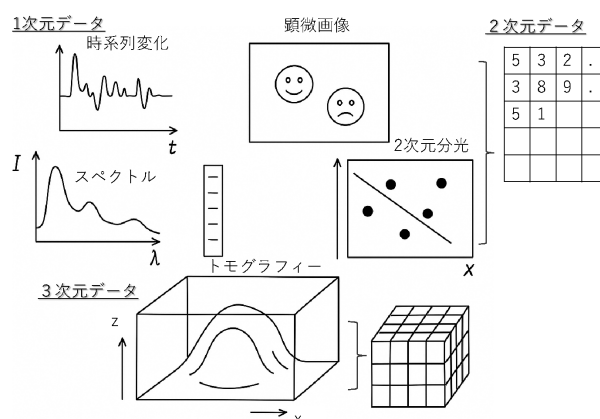


図1 分析化学で観測される様々なデータの形

2.1 1次元データ

分析化学で最も頻繁に遭遇するデータ形式として、波長やエネルギーに対するスペクトルデータが挙げられる。紫外可視吸収分光法 (UV/Vis)、赤外吸収分光法 (IR)、ラマン分光法、X線光電子分光法 (XPS)、核磁気共鳴分光法 (NMR) などから得られるスペクトルは、横軸に波長、波数、化学シフトなどの物理量、縦軸に吸光度や強度を取る。横軸は共通のパラメータであり、試料を表す分析データは縦軸に表される信号の部分であるため、これらのスペクトルデータは1次元の数値変化として表現される。

また、時間軸に対して変動する時系列データも重要なカテゴリである。クロマトグラフィー (HPLC, GC, IC など) から得られるクロマトグラム、各種スペクトル強度のリアルタイムモニタリングデータ、反応速度論的測定データなどがこれに該当する。これらのデータは、スペクトルデータと同様に時間を共通のパラメータとする1次元ベクトルとして取り扱うことができる。

2.2 2次元データ

より複雑なデータ構造として、2次元データが存在する。代表例の一つは、スペクトルデータの時間変化を捉えたものである。反応過程における分子種の変化を追跡する際のスペクトル測定や、クロマトグラフィーと分光法を組み合わせたハイフネーション技術 (LC-MS, GC-MS など) では、(波長, 時間) または (m/z , 時間) の2次元データマトリクスが得られる。

また、画像データも重要な2次元データである。電子顕微鏡 (SEM, TEM)、X線イメージング、光学顕微鏡などから得られる画像は、X, Y方向の2次元空間における数値配列 (行列データ) として表現される。分析化学においては、微小空間での高解像度観察に対する需要が高く、これらの画像解析技術の重要性は増している。

2.3 3次元データおよび高次元データ

画像データの時間変化を観察する動画撮影は、(X, Y, 時間) の3次元データを生成する。例えば、触媒反応における表面変化の観察、細胞内での分子の動態追跡、材料の劣化過程の可視化などがこれに該当する。

さらに、医療分析分野においては、コンピュータ断層撮影 (CT) や磁気共鳴イメージング (MRI) により、直接的に3次元空間データを取得することが可能である。これらの技術は体内の3次元構造情報を提供し、広義的分析化学の範疇に含まれる。

2.4 複数試料測定における次元の拡張

上記のような観点から分析化学で扱うデータ群を整理すると、多くの場合、以下のように分類できる：

- 1次元データ：スペクトル, 時系列変化
- 2次元データ：画像, スペクトルの時間変化
- 3次元データ：動画, 立体画像, 時空間変化データ

ただし、実際のデータサイエンス的解析では、単一の試料だけを扱うことは稀である。複数の試料に対して同じ測定を行い、それらのデータを比較・分析することで初めて有意義な情報が得られる。このため、複数試料の測定では以下のような次元の拡張が起こる：

- 1次元データ：スペクトル, 時系列変化→試料数分のデータが集まり2次元行列として解析
- 2次元データ：画像, スペクトルの時間変化→試料数分のデータが集まり3次元テンソルとして解析
- 3次元データ：動画, 立体画像, 時空間変化データ→試料数分のデータが集まり4次元テンソルとして解析

つまり、データサイエンス的解析においては、個々のデータの次元に「試料数」の次元が加わった高次元データを扱うことになる。この次元の拡張を理解することは、適切な解析手法を選択する上で重要である。

3 次元別データ解析

分析化学において、なぜデータを処理する必要があるのだろうか。生データをそのまま観察しても、ノイズが多い、複雑すぎる、などの要因により、データ中のどこに注目すべきかを判断することは困難である場合が多い。例えば、測定装置から得られる信号には様々なノイズ成分が重畳しており、目的とする化学情報を直接的に読み取ることは容易ではない。ケモメトリクスの例からも明らかなように、複数の成分が含まれた混合物のスペクトルでは、各成分由来のピークが重なり合い、スペクトル全体が複雑化する。この結果、何がどれだけ含まれているかを定量的に把握することが困難となる。また、スペクトルデータや時系列データにおいて本当に重要な情報は、特定のピークや特徴的な変化部分に集約されていることが多い。

2次元・3次元の画像データにおいても同様の問題が存在する。画像の大部分は類似の情報しか含んでおらず、問題となる箇所を特定する必要がある。このような課題は、材料の不良検出、医療における悪性部位の検出、環境汚染物質の解析など、分析化学の様々な場面で現れる。

以下では、基本的な解析手法を中心に、その目的と適用方法について説明する。

3.1 データをもとに何がしたいか

1次元のスペクトルデータ ($1 \times N$) が試料数 M 個測定される場合を考えよう。これは図2(a)に示すよう

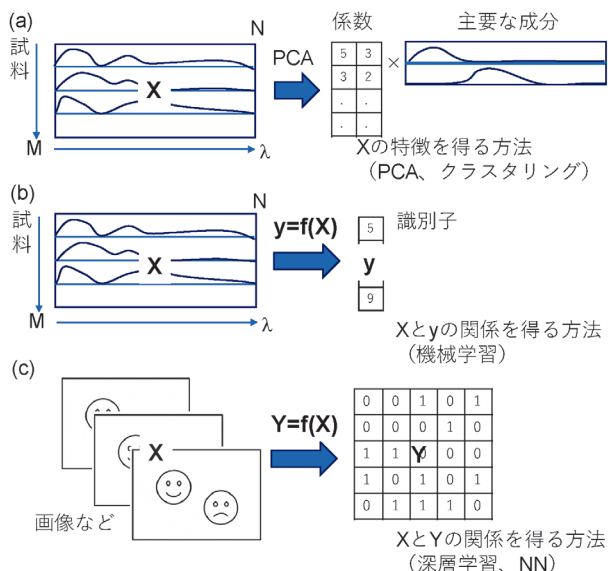


図2 データの種類と解析の目的

に $M \times N$ の行列データとして表示される。この入力データを X と表記する。

3・1・1 データの特徴抽出（教師なし学習）

与えられた X のデータの中の特徴や性質を取り出すことを教師なし学習と呼ぶ。最も頻繁に用いられる主成分分析（PCA）は、この X 内における特徴（例えば、どの波長にピークが集中しているかなど）を取り出すものである。（図2（a））PCAと類似の目的を持つ手法として、統計的に独立な成分を抽出する独立成分分析（ICA）や、観測変数の背後にある潜在因子を見つける因子分析がある。これらは特徴抽出・次元削減手法と呼ばれ、データの本質的な構造を理解するために用いられる。

一方、データを類似性に基づいてグループ分けするクラスタリング手法も重要である。代表的な手法として、データを類似する k 個ずつのクラスターに分割する k 近傍法、樹形図による階層的なグループ化を行う階層クラスタリング、密度ベースでクラスターを形成するDBSCANなどがある。これらは試料間の関係性を理解し、類似した特性を持つ試料を特定するために用いられる。

さらに、高次元データを人間が理解しやすい低次元空間に投影して可視化する手法も存在する。t-SNE（t-distributed Stochastic Neighbor Embedding）やその改良版であるUMAP（Uniform Manifold Approximation and Projection）は、複雑な高次元データの構造を2次元や3次元の散布図として表現し、データの全体的な分布や潜在的なパターンを視覚的に把握することを可能にする。

3・1・2 予測モデルの構築（教師あり学習）

スペクトルデータ X に対応する目的変数、例えばある物質の濃度 y ($M \times 1$, 各試料ごとの濃度データ) が与えられている場合、 X と y の関係式を求めて新しいスペクトルデータから濃度を予測する問題を教師あり学習と呼ぶ（図2（b））。目的変数 y の性質により、以下に分類される：

回帰問題： y が濃度のような連続値の場合

分類問題： y が化学物質の活性（あり/なし）のような離散値の場合

この関数形に、様々な方法があり、一般的なフィッティング問題でも用いられる線形回帰から、特定の関数形を決めないサポートベクターやガウス過程回帰/分類などの高度な計算手法があるが、いずれの場合も問題の目的設定は同じである。主成分を用いる点でPCAと方法が似ている部分最小二乗法（PLS）は、分析化学でもよく用いられる。ここでは X 中の主成分が y データを最もよく説明できるように主成分を決める方法で、 y を説明するための教師あり学習であり、PCAとは使用目的が異なる。

近年では、AIの進歩という文脈で用いられる機械学習は、多くの場合、深層学習（deep learning）に分類されるもので、画像に用いられる畳み込みニューラルネットワーク（CNN）や時間変化データによく用いられるリカレントニューラルネットワーク（RNN）や長短期記憶ネットワーク（LSTM）などがある。上記の機械学習との違いは、基本的に複数データ入力（ X ）で複数データ出力（ Y ）に対応するものであり（図2（c））、関数作成のための時間も長くなり、高性能な計算機が必要となる。

3・1・3 手法選択の指針

計算手法の選択は、まず、目的を明確にする必要がある。 X データの特徴を知りたいのか、 X と y の関係を知りたいのか、回帰問題なのか、分類問題なのか、複数データの出力が必要なのかによって、おおよそ計算手法が決まり、計算資源などを総合的に考慮して決定する必要がある。

一般的な戦略として、線形手法を最初に試行し、予測精度が不十分な場合に非線形手法を適用することが推奨される。これにより、計算コストを抑えつつ、必要に応じて複雑なデータ構造に対応できる。

分析化学の実践においては、これらの手法を単独で用いるのではなく、前処理としてのノイズ除去、特徴量抽出による情報の凝縮、次元圧縮による可視化という一連の流れの中で適切に組み合わせることが重要である。

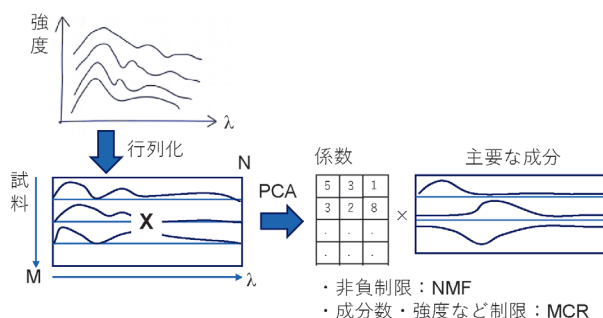


図3 1次元データのPCAなどの解析の模式図

3.2 1次元データの解析

1次元データの代表例として、スペクトルデータを考える。横軸に波長や化学シフト、縦軸に吸光度や強度を取ったベクトルデータ ($1 \times N$) が、試料数 M 個測定されると、全体は $M \times N$ の行列データとして表される (図3)。

このとき、まず関心があるのは「測定されたスペクトルがどのような成分 (特徴スペクトル) から構成されているか」である。そのために広く用いられるのが成分分析 (PCA) である。PCA では、複数のスペクトルに共通する特徴成分を抽出し、各試料のスペクトルをそれら成分の線形和として表すことができる。ただし、数学的には行列分解に基づくため、得られる成分には正負の寄与が混ざる場合がある。分析化学では物理的意味から「スペクトルは非負であること」が望ましい場合が多く、そのようなケースには非負値行列分解 (NMF, nonnegative matrix factorization) が有効である。さらに進んだ手法として、多変量曲線分解 (MCR, multivariate curve resolution) があり、成分数や非負性、濃度範囲といった制約条件を付与して解析することができる。MCR では交互最小二乗法 (ALS, alternating least squares) を用いて、スペクトル成分と各試料における寄与率を交互に最適化しながら解を求める。

次に、教師あり学習を用いるアプローチがある。これは、スペクトル ($X: M \times N$) と、その試料に対応する既知の情報 ($y: M \times 1$ あるいは $M \times L$) を対応づけ、未知の試料を予測・分類する手法である。例えば、UV/Vis や IR のスペクトルから化学物質の種類や構造を推定することは、従来はスペクトルデータベース検索で行われていたが、近年は SVM (サポートベクターマシン) やランダムフォレスト、ニューラルネットワークなどの機械学習モデルが応用されている。 y が濃度のような数値であれば回帰問題、物質の有無のようなラベルであれば分類問題となる。化学物質を計算に用いる場合には、化学物質を数値列に変換する。化学物質をどのように数値列として扱うかについて、疑問に思うかもしれない。どのような計算でも数値列にならないとコンピュータでは計算できない。機械学習でも例外ではなく、 y の値と

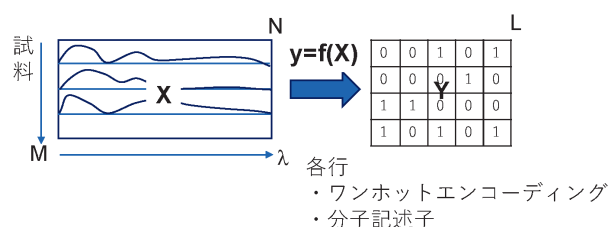


図4 1次元データから分子などに関係づける機械学習

して名前のような文字列で与えられる場合には数値列に変換する必要がある。分類したい数の分だけ、対応する数値列を作成することをワンホットエンコーディングという。より化学的な取り扱い方法としては、化学物質の共通の記述方法として SMILES という方法がある (図4)。

解析に先立って、前処理も重要である。バックグラウンド信号の除去や平滑化、ピーク検出による「位置と強度」の情報への変換といった基本処理を行うことで、データサイズを削減し、計算負荷を下げると同時に、解析精度を向上できる。単純な処理であっても、結果の解釈性と効率に大きく寄与する。

また、クロマトグラフィーにおけるクロマトグラムは、代表的な1次元データである。時間に対するピーク列として表され、従来は保持時間やピーク面積から物質同定・定量が行われてきた。近年は、ピーク検出やベースライン補正を自動化するアルゴリズムが一般化し、さらに複数成分が重なったピークを数値的に分離する手法も用いられている。これに加えて、データベース照合や機械学習を活用した物質同定の自動化が進んでおり、SVM やランダムフォレストによる分類報告されている。また、クロマトグラム全体を時系列データとして扱い、RNNなどを適用することで、保持時間の補正やピーク予測が行われている。このようにクロマトグラフィーデータ解析は、基礎的なピーク処理から機械学習を利用した高精度な予測・分類へと発展しており、今後さらに自動化と精度向上が期待される。

このように、1次元データの解析では、(1) 行列分解による成分抽出、(2) 教師あり学習による予測や分類、(3) 適切な前処理による精度向上、という三つの流れを組み合わせる利用することが効果的である。

3.3 2次元データの解析

2次元データとしては、分析化学の分野ではまず2次元分光 (2D-NMR, 2D-IR など) が思い浮かぶであろう。これらは1次元スペクトルを拡張したマトリクス形式のデータであり、1次元データの解析と同様に、化学物質の種類や濃度と結び付けて機械学習を適用することが可能である。一方、実際に最も頻繁に利用されるのは顕微鏡画像などの2次元画像データである。これは、機

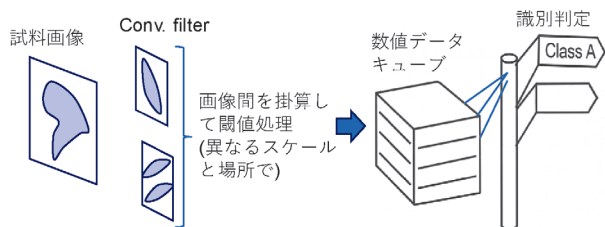


図5 CNNを用いた画像分類の模式図

機械学習の入門例としてよく挙げられる「犬と猫の写真分類」と本質的に同じ構造を持っており、画像から対象を識別する枠組みがそのまま分析化学データにも適用できる。

画像データは、図1に示した通り基本的に2次元の数値行列として表され、各画素の数値が明暗の強度を表す。カラー画像では、各ピクセルにおけるR・G・B強度を持つため、3層の2次元データとなる。このようなデータ解析には、畳み込みニューラルネットワーク（CNN）が一般的に用いられる（図5）。CNNは「畳み込みフィルタ」を用いて複数のパターンを画像から抽出し、それらの有無を数値化していく。さらにプーリングによって空間的な平均化や縮小を行い、異なるスケールでの特徴を取り込みながら処理を進める。これを繰り返すことで、画像を特徴量の数列に変換し、最終的に分類や回帰に利用する仕組みである。

CNNはこの10年で急速に発展し、ResNetやVGGなど多様な構造が提案され、画像認識精度は飛躍的に向上した。一般ユーザーが一から学習するのではなく、既存の汎用ネットワークを自分のデータに合わせて最終層を学習し直す「転移学習」が実用的である。転移学習では、最終層（fully connected layer）をリセットし、自分のデータ（入力 X と教師ラベル y ）を用いて短時間で再トレーニングすることで、高精度かつ効率的に解析する方法である。

ただし、CNNの判定はしばしば「どの部分を根拠に分類しているのか分からない」というブラックボックス性を持つ。これを補うため、Grad-CAMやLIMEといった「説明可能AI（Explainable AI）」の手法が開発され、判定に寄与した画像領域を可視化する試みが進んでいる。しかし、解釈性の完全な保証には至っておらず、今も研究課題である。

一方で、画像解析においては判定に必要な情報が全体の一部に限られる場合が多い。そのため、画像から特徴を抽出し、1次元ベクトルとして扱うアプローチも有効である。例えば粒子画像から輪郭を検出して、粒径・形状・分布といった特徴量を数値化すれば、1次元データ解析と同様の機械学習手法を適用でき、計算量削減と解釈性向上の両立が可能となる。

さらに、画像全体を数値行列として扱い、特異値分解

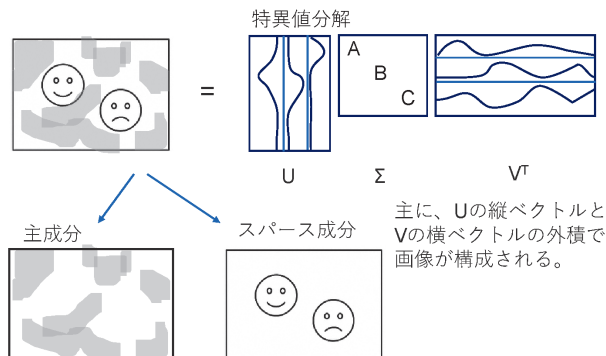


図6 画像の主成分分析による分解と背景・特徴画像の分離

（SVD）やPCAにより主要な強度パターンを抽出する方法もある（図6）。このとき、行列分解における固有値行列 Σ は縦横方向の強度パターンに対する係数行列と理解できる。いくつかの主要な画像要素でその画像が構成されていることを示す。主要な成分を除去して残差を取り出すと、画像の特徴的部分（スパース成分）が明らかになる。より発展的な手法としては、ロバスト主成分分析（Robust PCA）があり、背景成分と特徴成分を明確に分離できるため、欠陥検出や異常領域の抽出に有効である。

3・4 3次元以上のデータの解析

3次元以上のデータの代表例としては、CTやMRIの医用画像解析が挙げられる。MRI画像はNMR信号の3次元展開版と理解でき、分析化学にもなじみの深いデータである。また、共焦点レーザー顕微鏡による3次元画像、材料をスパッタリングしながら組成を観測するSIMS（secondary ion mass spectrometry）、皮膚断層の観察に用いられるOCT（optical coherence tomography）など、分析化学分野でも多様な3次元データが利用されている。さらに、複数の試料を比較したり時間変化を観察することで、空間3次元に時間軸を加えた4次元データも得られる。

このような3次元データ解析には、2次元CNNを拡張した3次元畳み込みニューラルネットワーク（3D-CNN）が用いられる。入力データをボクセル（体素）として扱い、3次元的なパターンを抽出することで、画像中の立体的特徴を自動的に学習できる。特に医療応用では、腫瘍や病変部位の位置を教師データとして学習し、診断支援に役立てる例が多い。こうした3次元モデルをゼロから構築するには大量のデータと計算資源が必要なため、U-Netなどの学習済みモデルを転移学習して再利用する手法が実用的に広く用いられている。

分析化学そのものからは少し離れるが、分子の3次元構造を入力として扱うグラフニューラルネットワーク（GNN）も近年活発に研究されている。分子を原子（ノード）と結合（エッジ）で表し、機械学習によって

構造と物性の関係を学習するもので、材料探索や創薬への応用が期待されている。

ただし、むやみにデータを高次元化すると、計算コスト増大や解釈困難といった問題を生じることが多い。実用的には、まず3次元データから特徴量を抽出して次元を圧縮し、それを機械学習に入力することが望ましい。2次元データにおけるPCAのような汎用的手法は3次元以上では限られており、代わりにテンソル分解が利用される。代表的なものにCP分解やTucker分解があるが、一意に分解できない場合やノイズ耐性の問題など課題も多く、分析化学での普及はまだ限定的である。現実的には、まず次元を削減してから機械学習に用いることが効率的かつ解釈しやすいアプローチとなる。

4 ま と め

分析化学におけるデータサイエンス手法の適用で最も重要なのは、扱うデータの次元性を正しく把握することである。本稿で述べたように、1次元のスペクトルや時系列データ、2次元の画像やマトリクスデータ、さらには3次元・4次元の時空間データに至るまで、それぞれに適した解析手法が存在する。

分析化学におけるデータ解析の目的は、化学的性質を人間が理解できるようにすることである。そのためにむやみにデータ量を増やすのではなく、高次元の複雑なデータから本質的な情報を抽出し、理解可能な形に変換するのがよい。主成分分析による2次元データの可視化、CNNによる画像特徴の抽出、テンソル分解による高次データの成分分解などを利用して可能な範囲でデー

タサイズを小さくし、教師あり学習（機械学習）を用いて予測を行うのがよいだろう。

実際の分析化学の実践においては、これらの手法を単独で用いるのではなく、データの性質、解析目的、利用可能な計算資源、ドメイン知識などを総合的に考慮した上で、複数の手法を組み合わせることが重要である。例えば、前処理として教師なし手法によるノイズ除去や次元圧縮を行い、その後に教師あり学習による予測モデルを構築するといったハイブリッドアプローチが有効である。分析化学者は自らが扱うデータの特性を深く理解し、適切な次元での解析手法を選択・組み合わせることで、データに内在する化学的な知見を効果的に抽出できるものと考えられる。

文 献

- 1) 明治大学 データ化学工学研究室 Web, <<https://datachemeng.com/summarydataanalysis/>>.
- 2) Google colab の 使 い 方, <<https://www.kikagaku.co.jp/kikagaku-blog/google-colab-howto/>>.
- 3) <<https://colab.research.google.com/drive/1jKAutKeNXcvVwCD9OCQ4RIba9dFXdT7W?usp=sharing>>.



片山 建二 (KATAYAMA Kenji)
中央大学理工学部 (〒112-0003 東京都文京区春日 1-13-27). 東京大学大学院工学系研究科応用化学専攻中退. 博士 (工学).
《現在の研究テーマ》データ科学・時間分解顕微鏡・光触媒・太陽電池・液晶.
kkata.33g@g.chuo-u.ac.jp

原 稿 募 集

「技術紹介」の原稿を募集しています

対象：以下のような分析機器、分析手法に関する紹介・解説記事

- 1) 分析機器の特徴や性能および機器開発に関わる技術、2) 分析手法の特徴および手法開発に関わる技術、3) 分析機器および分析手法の応用例、4) 分析に必要な試薬や水および雰囲気などに関する情報・解説、5) 前処理や試料の取扱い等に関する情報・解説・注意事項、6) その他、分析機器の性能を十分に引き出すために有用な情報など

報など

新規性：本記事の内容に関しては、新規性は一切問いません。新規の装置や技術である必要はなく、既存の装置や技術に関わるもので構いません。また、社会的要求が高いテーマや関連技術については、データや知見の追加などにより繰り返し紹介していただいても構いません。

お問い合わせ先：

日本分析化学会『ぶんせき』編集委員会

[E-mail : bunseki@jsac.or.jp]