計測インフォマティクスと機械学習
Measurement Informatics and Machine Learning

石井真史
Masashi ISHII
国立研究開発法人物質・材料研究機構　統合型材料開発・情報基盤部門
MaDIS, National Institute for Materials Science
〒305-0044 つくば市並木 1-1
1-1 Namiki, Tsukuba 305-0044
e-mail: ISHII.Masashi@nims.go.jp
分類番号：12.3

計測に機械学習など、統計的な手法を導入するときの共通課題を、関連技術の歴史を織り交ぜつつ解説します。材料が持つ物理・化学的コンテキストの多様性と、本来計測が持つべき汎化性を評価軸にして、代表的な手法の位置づけと応用における制約を考え、計測インフォマティクスを材料開発などに展開する上での、知識基盤やデータ連携の重要性に言及します

1. まえがき

計測分野で大量に得られるデータを統計数理的な手法で処理し、新しい知見を見出そうという「計測インフォマティクス」が盛んになっています。本稿では、その背景にある技術を発展の歴史にも触れつつ俯瞰して、課題やスコープを述べます。本稿で頻出する言葉として「コンテキスト」があります。背景、状況といった意味ですが、一般的に材料の背後にある情報を限定することで、計測に関する特定の解決策が見つけられる可能性が高まりますが、その一方で本来計測が持つべき汎化性は低下してしまいます。計測インフォマティクスの基幹をなすものとして機械学習がありますが、歴史的に見てもその他の技術も取り入れた、コンテキストの問題の解決が必要になります。図 1 に本稿で俯瞰する機械学習・デジタル画像処理・電子素子の歴史の概略をまとめておきます。

図 1　機械学習・デジタル画像処理・電子素子の歴史の概略

インフォマティクスの中で頻繁に表れる「機械学習」という言葉は、決して新しくはなく、計算機の生みの親と言われる A. Turing は 1950 年の論文[1]の中で、「学習する機械」という言葉を使っています。1960 年頃には、学習する機械が日本でも試作されており、真空管の一種であり放電位置を変えることで荷重を離散的に変化させるデカトロン[2]や、電気化学的な金属析出を使ったメミスタ[3]と言った「学習素子」が開発されました。手元の古いテキスト[4]では、パーセプトロン第一号はサーボモータ付き可変抵抗器によって作成されたという記載がみられます。この史実は、電子技術の発展と、機械学習の取り組みが今に至るまで同時に進んでいたことを示しています。

最近は人が思いつかないことを機械が気づくことに期待が寄せられています。そこには、一つ一つは人が思いつけるものではありながら、データ量が膨大であるために整理できないものを、客観的にまとめ上げる能力も含まれています。Turing は、「機械が考える」ことを、機械が人を模倣できること、と定めており、数値計算能力は人に勝ることは知りつつ、人知を超えた機械が現れることははっきりとは予想していなかったようです。デカトロンが離散値でしか調整できず、メミスタにしても荷重を変える応答時間（50%応答）が 15 秒であった[4]ことを考えると、機械学習の意味や意義は当時のスコープを越えているようです。

こうした想定以上の展開は、計算機の高速化、大容量化に負うところが大きく、先日発表されたオークリッジ国立研究所の FRONTIER は、エクサ FLOPS の計算処理能力を持っていて[5]、図 2 に示す通り世界最初の大型コンピュータ ENIAC から 10^{15} 倍もの計算処理能力を持っています。また計測分野における検出器の革新は、例えば画像素子に関しては 2000 年に現れていることは以前に本誌で紹介しました[6]。これらのことから、機械学習と電子素子の進展の相乗効果から計測インフォマティクスが生まれたと考えられます。しかし固体

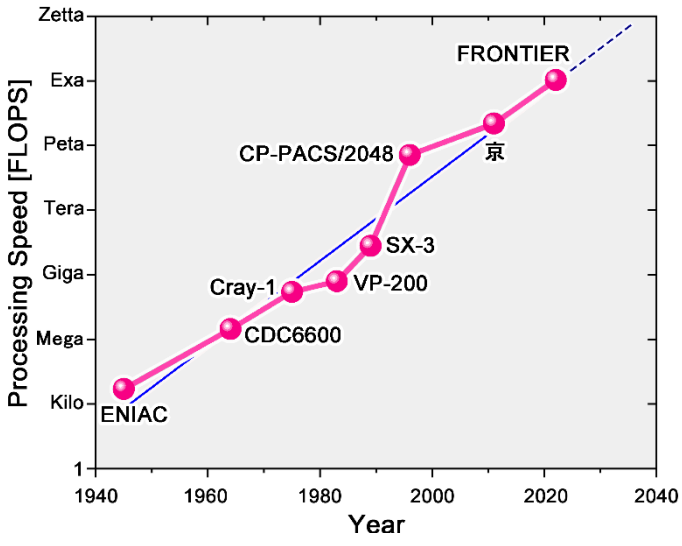

図 2　大型コンピュータの処理速度の変化

素子の高い信頼性に対して、機械学習は、人工知能の著しい発展を見た現在でも、制約は各所で議論されています[7-9]。このことは、計測インフォマティクスにおいても考慮する必要があると思われます。

2. 計測インフォマティクスの素過程

計測インフォマティクスのフローを、素過程として「(1)オブジェクト認識」、「(2)特徴量抽出」と「(3)識別」に大別してみます。(1)と(2)に関しては、デジタル画像処理と重なる部分が多くあります。デジタル画像処理は 1960 年代のアポロ計画などで急速に進展し、データの前処理技術もこの時代以来、混然一体となって発展してきたといえます。ここではデジタル処理技術の詳細は逐一触れませんが、計測インフォマティクスの背景技術として位置づけられることは間違いありません。以下では、計測インフォマティクスの素過程として(1)-(3)を順に概観し、コンテキストと汎化の相互関係を考えてみます。

2.1 オブジェクト認識

例えば、画像における輪郭抽出がこれに相当します。こうした処理は、現在ではオープンライブラリを使うことで容易に可能になっています。コントラストが明確であるオブジェクトに対しては、輝度の一回微分の最大値や二回微分のゼロクロッシング点を輪郭と定義することができます。更にカラーヒストグラム（三原色の画素分布）等を特徴量として、探索領域の時間的な変化を、類似度から決定して追跡する動体認識のアルゴリズムは、良く知られています。更に、オブジェクトが存在する確率（尤度）を考慮して離散的な点の集合（Particle）を選別して効率化を図る Particle filter（逐次モンテカルロ法）[10]などの手法は追跡時間を高速にしました。強調したいことは、こうしたオブジェクトに一般的に内在する特徴量（多くのものは透明ではない）を計測インフォマティクスに適用する重要性になります。すなわち、オブジェクトに付随する背景（コンテキスト）を求めない限り、共通性が高くなり、多くの物体に適用できるようになります。更に、混雑した中での同一人物の追跡など機械にしかできない能力が実装されている現状は、コンテキストがインフォマティクスに必須ではないことを示しています。

2.2 特徴量抽出

オブジェクトが決まった後の特徴量抽出においては、コンテキストを考慮しない方法として、主成分分析(Principle Component Analysis, PCA)などのクラスタリングがあります。PCA では、多くのデータから分散が大きな変換軸を順に決めてゆくことで、次元圧縮して特徴量を抽出することができます。この分析対象のコンテキストを考慮しない成分分解は、手法の適用性を高め、汎化性を生むことになります。一方で、コンテキストに踏み込まない以上、抽出される特徴量に正確な物理的な意味を求めることはできません。勿論、結果的に各主成分に対応する物理・化学的な特徴が定性的に現れることはあります。多少 2.3 の識別に踏み込むことになりますが、例えば、適当なフーリエ変換赤外線分光（FTIR）スペクト

ルのセットから PCA によって特徴量を抽出すると、図 3 のようになります。図 3(a)の第一主成分は、FTIR の特徴を一般的に表しているだけで、この形状に物理・化学的な意味はありません。図 3(b)の第二、第三主成分は、それぞれ 1,750cm^{-1}、3,000cm^{-1} 付近で対照的な振る舞いしています、FTIR の経験がある方でしたら、これらの波数が C=O とアルキル基に相当することがすぐに想像できるかもしれません。ただし、その形状も、どちらが第二・第三になるかも、母集合となったスペクトル次第で変化します。つまり、主成分には正確な帰属や定量性はありません。一方で定性的とはいえ、分散最大という条件から一般的な化学情報（コンテキスト）が定性的でも導けたことが重要で、PCA が汎化に成功していると言えます。

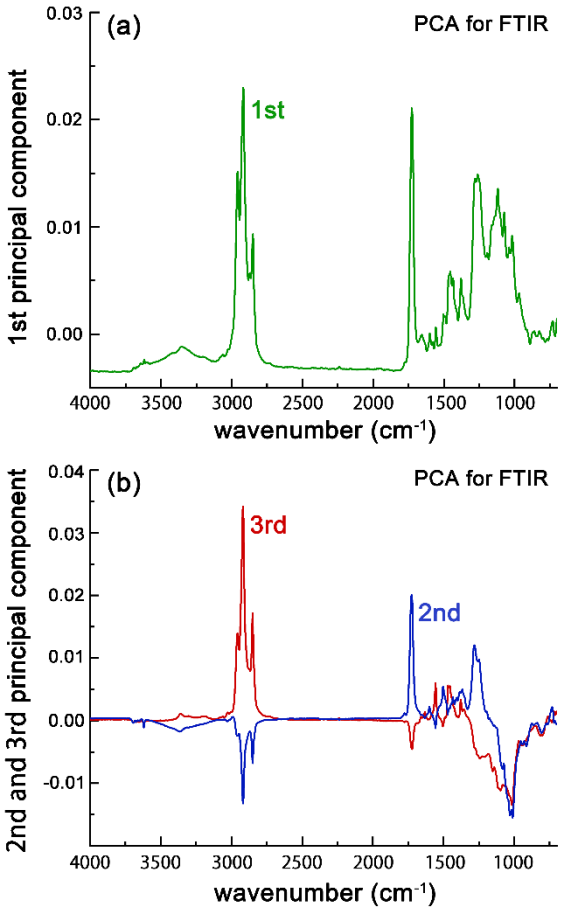

図 3　FTIR スペクトルのセットから PCA によって抽出した特徴量

特徴量抽出のアルゴリズムは多く考案されています。応用物理分野で良く知られているＸ線回折像は、幾何構造の特徴量を巧みに抽出しています。周期関数を使った展開によって、対象の平行移動に無頓着な点対称な特徴量が得られます。同様な幾何構造の特徴量の抽出にパーシステントホモロジーが最近良く取り上げられています [11)]。いずれにしても抽出された特徴量を 2.3 で述べる識別過程の入力とすれば、揺らぎ成分を除去した分析が可能になります。このような利点の一方で、Ｘ線や電子線を一次ビームに使った像の場合、散乱や吸

収などにより、成分元素に固有のコンテキストが像に強く含まれます。こうした場合、幾何構造だけではなく材料の多様性が取り込まれ、汎化が急に難しくなります。

2.3 識別

ここでは、三つの識別を取り上げてみます。

①クラスタリング

計測インフォマティクスでは、特徴量空間で興味ある対象の集合（クラスタ）が可視化されれば十分な例は少なくありません。この場合は、コンテキストは必要なく、数学的な指標を使った次元削減、例えば先に述べたPCAの他、独立成分分析(ICA)[12]、特異値分解(SVD)[13]などで目的を達成できます。図 3 で述べた通り、正確な物理・化学的意味は少ない代わりに、未知の対象に対しても同じモデルを適用できる可能性は高まります。

②分類

教師データを準備して、対象を限られた数のクラスに、ヒューリスティックに、あるいは機械学習で分類するタスクは、状態や成分などの分析として計測インフォマティクスにおける一つのテーマになります。この場合、クラスに応じて分散が大きい特徴量を定め、それに照らして分類が実施されます。特徴量抽出と識別を一括してブラックボックス的に深層学習で行うことは最近の主流ですが、例えば1963年に考案されたサポートベクトルマシン（SVM）[14]のように、特徴量空間の中で、幾何学的にクラス分けをする方法は、決して新鮮ではありませんが簡易にして強力です。いずれにしても、このタスクがもつ問題は、教師データを作成する際にコンテキストが狭められる事が多い点です。計測では化学成分など広い対象を扱う必要がありますが、現状では膨大な教師データを体系的に準備できないため、特定の対象に限って検証されます。この方法は、手法を検討する上で有効ですが、汎化性を落とすことになります。つまり、学習対象の化学成分は判別できるようになりますが、それ以外の化学成分を識別することは一般には困難になります。

③照合

①と②は考え方は異なりますが、インスタンスをどのクラスに分けるか、という意味では見かけ上アウトプットには違いはありません。ただし、そのクラスがどのような物理・化学的意味を持つかに踏み込むと、得られる情報量は限られています。実際に精密な計測や材料開発のツールとして計測インフォマティクスを使うことを想定すると、ある特徴量を介して詳細なコンテキストがオブジェクトと識別先の間で関係づけられることが理想的と思われます。この場合の識別は、分類上は顔認証や指紋認証のような「照合」の技術と位置づけられます。照合先に知識基盤であるデータベースを設定し、特徴量の類似性から、オブジェクトとデータベース内の可能性があるレコードを結び付けることで、コンテキストに間接的に到達する考え方です。このポイントは、類似性を使った照合自体にコンテキストが殆ど含まれないことです。指紋認証でいえばマニューシャ照合[15]のように、特徴量の抽出の仕方に固有の特徴はありますが、そこに個人情報は入っていません。しかし、一旦認証が成立すればその人の詳細情報（コンテキスト）を利用できる、良好な汎化性を保証しています。

計測インフォマティクスに話を戻せば、X線回折パターンを使って、データベースのレコードとの類似性や同一性を判断する「物質認証」は可能でしょう[16]。最初から深いコンテキストを正解ラベルとするのではなく、知識のネットワークの中のある接点（ノード）間の一つの枝（エッジ）の有無を機械学習で確率的に、あるいは別の方法で決定論的に決め、各ノードに紐づく情報は別のタスクとして得ることは、計測インフォマティクスを使った材料開発の現実的な解になると思われます。

3.まとめ

機械学習をはじめとする統計的手法を取り入れた計測インフォマティクスが盛んになる中で、基盤をなす技術を歴史的な背景を含めて紹介しました。材料の物理・化学的な特性（コンテキスト）の多様性は、本来計測が目指していた広い測定対象を網羅して理解する考え方を変え、コンテキストを限ることで、特定の高度な課題を機械学習で解決する展開を生んでいます。一方で、コンピュータを中心とする技術の発展は、人では事実上実行しきれない計測をこなすような、高速・大量処理をベースとする展開をもたらしました。このとき、計測対象から取得できるコンテキストは決して多くはありません。材料開発に必要な広く深いコンテキストは、照合を使って測定結果とデータベースを結び付けた後で、連鎖的に獲得することが有効と考えらえます。

1) A. M. Turing: Journal of the Mind Association **LIX**, 433-60 (1950).
2) 西野治, 梶浦正孝, 江川英晴, 倉持義徳: 応用物理 **24**, 276 (1955).
3) B. Widrow, W. H. Pierce, J.B. Angell: Technical Report No. 1552-2/1851-1 (1961)
4) 志村正道: パターン認識と学習機械, 昭晃堂, 1970.
5) https://www.top500.org/news/ornls-frontier-first-to-break-the-exaflop-ceiling/
6) 石井真史: 応用物理 **85**, 223 (2016).
7) 松原仁: 人工知能学会誌 **18**, 564 (2003).
8) https://dl.sony.com/ja/deeplearning/about/disadvantage.html
9) ヤン・ルカン: ディープラーニング　学習する機械（講談社, 2021).
10) 樋口知之: 電子情報通信学会誌 **88**, 989 (2005).
11) 平岡 裕章, 西浦 廉政: 日本物理学会誌 **72**, 632 (2017).
12) 池田 思朗: 日本神経回路学会誌 **9**, 181 (2002).
13) Steven L. Brunton and J. Nathan Kutz: "Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control 1st Edition", Cambridge University Press (2019). http://www.cs.columbia.edu/~djhsu/AML/lectures/notes-pca.pdf
14) V. Vapnik and A. Lerner: Automation and Remote Control **24**, 774 (1963).
15) 例えば Jianjiang Feng: Pattern Recognition **41**, 342 (2008).
16) 石井真史, 上杉文彦, 小澤哲也: 日本結晶学会誌 **62**, 35 (2020).

著者紹介
1995 年　大阪大学基礎工学研究科　博士課程修了　博士（工学）。計測・データ駆動研究に幅広く取り組みつつ、高分子データベース PoLyInfo、X 線吸収データベース MDR XAFS DB を中心に、データベースの構築と、データの概念化と統合を主専門分野とする。

図 1　機械学習・デジタル画像処理・電子素子の歴史の概略
図 2　大型コンピュータの処理速度の変化
図 3　FTIR スペクトルのセットから PCA によって抽出した特徴量

Common issues when introducing statistical methods such as machine learning to measurements are explained, referring to the history of related technologies. The diversity of the physical and chemical context of materials and the universal applicability of measurements are taken up as assessment criteria, and the position of typical methods and their restrictions in applications are discussed. The importance of knowledge base and data linkage in the application of measurement informatics to materials development will then be addressed.