物質科学における新しい帰納的アプローチ

Keyword: マテリアルズ・インフォマティクス

1. はじめに

古典力学，電磁気学，量子力学，統計力学などの物理の法則や周期表の確立などの原理や法則は，観測データからの推論によって得られたものであり，帰納的な研究の成果であった（帰納（induction）というよりは，仮説的推論（abduction）がより適切）．その後の物質科学の研究は，これらの原理や法則による因果関係を使って演繹的に結果を得ることが主流となってきた．一方，大方のこのような演繹的研究から離れているのは物質開発である．物質科学の発展を支える要因は多岐に亘る新物質であり，それらが発現する物性や現象に触発された新しい原理の構築である．

物質科学で注目されてきた大半の新奇物質は，経験に基づく「洞察」，「ひらめき」，「勘」，それに「偶然」などの産物として見出されてきた．物質開発の別の動きは，技術の進歩からの要請に因るものである．この方向の物質開発では目標とする機能が明確であり，「物質設計」というキャッチフレーズが盛んに行われるようになった．

物質設計が困難な理由は，それが演繹的研究による順問題ではなく，帰納的な逆問題の性格を持っており，一般に逆問題が順問題よりはるかに難しい，ということにある．順問題では物質を与えてその物性や機能を調べるが，物質設計では望ましい物性や機能を与えて，それを満たす物質を予測しなければならない．物質設計は主として演繹的方法である計算科学が主導してきたが，非常に労力と時間を要する．抜本的に物質設計を効率化し一般化するには，帰納的なアプローチを導入することが必要である．

2. マテリアルズ・インフォマティクスのスタート

世の中のデータ量の増加は指数関数的であり，データの活用が経済活動や医療活動において重要な役割を演じている．世の中の強い関心を引いたアルファ碁やその発展形の“Master”は人工知能の高い能力を見せつけた．バイオの分野では1990年にスタートしたヒトゲノム計画の推進とともに，大量のデータが出力されるようになり，バイオ・インフォマティクスが発展した．また，ケモインフォマティクスがそれに続いた．データ量の増加は物質科学の世界でも見られる．計算機の進歩は出力量の増大をもたらし，計測における精緻化，大型化，自動化により膨大なデータが出力されるようになり，それらの活用は重要な課題である．

マテリアルズ・インフォマティクス（MI）の本格的な始まりは2011年にオバマにより提唱された米国でのMGI（Materials Genome Initiative）を待つことになる．[1] 我が国での本格的なプロジェクトは2015年からスタートした「情報統合型物質・材料開発イニシアティブ」が最初である．[2] データが「経験」と「膨大な試み」に取って替わり，機械学習などのデータを活用した帰納的な逆問題解析が「ひらめき」とか「勘」に替わる「処方箋」を与えるような物質設計を可能とすることが，MIの主要な目標である．複雑な現象を支配する要因を帰納的に解析し理解する，という重要な役割も期待される．また，計測データの自動解析や計測の効率化が重要課題となっており，関連研究活動は「計測インフォマティクス」とも呼ばれる．

3. 機械学習

機械学習とは，訓練データからそれが内包するルールを計算機が学習し，そのルールを使って探索したい対象の性質を予測することである．[3]「機械学習」に似た内容を意味する言葉として「データマイニング」がある．それらは区別されないで用いられることも多々あるが，「機械学習」は予測に重きが置かれており，「データマイニング」はデータからの意味の抽出と知識発見に重きが置かれている．

機械学習の概要を図1として示す．[4] 実際の先端的研究では分類問題から相関分析・次元削減等の4つの扱いを使っての物質・材料研究が展開されている．

4. 記述子

物質に関するデータは，一般的にはその組成と構造の情報（Z）と，その物質の物性の情報（Y）の組み合わせで与えられる．そのデータを機械学習に活用するための第一歩は，扱いたい物質とその物性を定量的に特徴づける記述子（X）

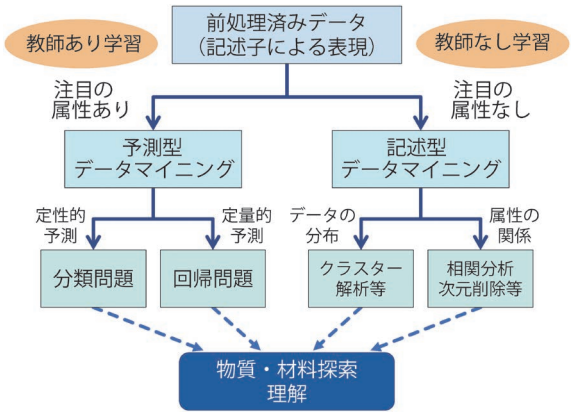

図1　機械学習の概略の体系．文献4のFig. 4を日本語にして改変したもの．

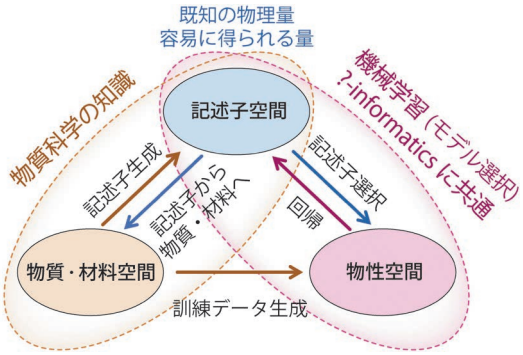

図2　マテリアルズ・インフォマティクスにおける機械学習の基盤となる記述子の位置づけを示す．

を与えることである．Xは一般にベクトルである．記述子については代数演算が行える必要があり，構造情報を記述子として表す種々の方法が提案されている．図2は機械学習における記述子の位置づけを示す．記述子が与えられると，記述子（X）と物性（Y）との関係づけをする機械学習は，一般的な「？インフォマティクス」の問題となる．ここで「？」は「バイオ」でも，「ケモ」でも，「マテリアルズ」でもよい．ただし，可能であれば，注目する物性に応じたモデルの導入が有用であり，そこに対象領域の知識が反映される．また，左の物質・材料空間と記述子空間の関係はまさにMIの特徴を表す．物質・材料の記述子の選択と活用はMIの根幹に関わる重要課題である．物質・材料の記述子（X）については，(1) 問題にする物性（Y）を制御する物理量であり，(2) その物性に関して，対象間の類似性を測るのに適している，の2つの要請がある．MIの記述子に関しては後述するように更に2つの重要な問題点がある．

5.　逆問題への対応

機械学習の逆問題への対応の分かりやすいアプローチは「仮想スクリーニング」である．例として回帰の場合を考える．機械学習によって訓練データから問題とする物性情報（Y）と記述子（X）の間の相関関係$Y=f(X)$を得る．この相関関係を用いて，探索したい物質群の個々の物質の物性（Y）を予測し，望みの物性を持つ物質を選択する．因果律を用いる演繹的アプローチに比べ，相関関係を用いる予測は圧倒的に高効率である．ただし，後者では交差検定という過程を経て相関関係の信頼度を検証することが行われる．この仮想スクリーニングをより効率化し，予測値だけでなくて，その分散も与えることができるベイズ最適化は，開発（exploitation）と探索（exploration）を状況に応じて選択できる手法として適応型（adaptive）法とも呼ばれる．

逆問題の観点における一つの重要な問題は，相関関係$Y=f(X)$から問題の物性を最適化する最適記述子X_{op}が得られたとして，それに対応する物質を決めることが一般的にはできない，という点にある．この問題は，有限の分子ではなくて無限の凝縮系を扱う場合には特に困難である．記述子が絡むもう一つの重要な問題は，機械学習の手法の選択とも関連するが，予測性能の向上と，理解の向上とが必ずしも一致しないことである．これらの2点は，MIの重要な今後の課題である．

6.　ベイズ推定と逆問題

ベイズ推定は，逆問題への対応の一般的指針を与えてくれるし，物質科学におけるデータ解析において物質科学の知識に基づくモデルを下記の尤度に対して導入しやすく，予測の信頼性の情報も与えてくれるという利点がある．[5] 基本の式は，2つの事象AとBの同時確率に対するベイズの関係式から得られる$P(\mathrm{A}|\mathrm{B})=P(\mathrm{B}|\mathrm{A})P(\mathrm{A})/P(\mathrm{B})$である．ここで，$P(\mathrm{A})$はAの実現する確率，$P(\mathrm{B}|\mathrm{A})$はAが実現しているという条件下でのBの実現する条件付確率である．ここで，Aが原因（物質），Bが結果（物性）を表す場合を考えると，この式の右辺は原因（物質）から結果（物性）という因果関係を表しており，左辺はそれを反転した，結果（物性）から原因（物質）という逆の関係を表す．右辺の尤度$P(\mathrm{B}|\mathrm{A})$はケモインフォマティクスでは定量的構造物性相関（qspr）と呼ばれており，訓練データからの学習で求められる．事前分布$P(\mathrm{A})$は，物質の実現性についての事前知識である．左辺の事後分布$P(\mathrm{A}|\mathrm{B})$は逆定量的構造物性相関（iqspr）と呼ばれる．望ましい物性（B）を持つ物質を探索することは，左辺の$P(\mathrm{A}|\mathrm{B})$が大きくなる物質（A）を探すことになる．この探索は有機分子についてモンテカルロ法によって具体的に行われている．

7.　その他の話題

本稿で殆どあるいは全く触れられなかった重要な話題には，深層学習，強化学習，スパースモデリング，モンテカルロ木探索，計測インフォマティクスなどがある．計測インフォマティクスでは，薄膜における組成の空間分布の自動解析などの重要な研究が進められている．

参考文献

1）https://www.mgi.gov/
2）http://www.nims.go.jp/MII-I/
3）標準的な教科書の一つとして，C. M. ビショップ著，元田　浩，栗田多喜夫，樋口知之，松本裕治，村田　昇監訳，『パターン認識と機械学習（上，下）』（丸善出版，2012）．
4）H. C. Dam，寺倉清之，表面科学，**36**, 507 (2015)．
5）U. von Toussaint, Rev. Mod. Phys. **83**, 943 (2011)．

寺倉清之〈物質・材料研究機構　TERAKURA.Kiyoyuki@nims.go.jp〉

（2017年10月17日原稿受付）