配信先：文部科学記者会、科学記者会

Press Release

2024 年 8 月 7 日
東京工業大学
横浜市立大学

機械学習により有望物質群とその設計指針を抽出

－所望の特性を持つ無機材料のパターンを自動検出する手法を開発－

【要点】

○無機材料データから所望の光学特性を持つ物質群に共通な特徴を検知
○機械学習予測モデルに基づいたクラスタリングにより物性を考慮した物質分類を実現
○マテリアルズインフォマティクスにより物質・材料科学的な知識を獲得

【概要】

東京工業大学 科学技術創成研究院 フロンティア材料研究所の佐藤暢哉研究員（研究当時）、高橋亮助教、清原慎 JSPS 特別研究員（研究当時。現：東北大学 助教）、大場史康教授は、横浜市立大学大学院生命医科学研究科の寺山慧准教授、物質・材料研究機構 マテリアル基盤研究センターの田村亮チームリーダーと共同で、無機材料の分類および設計指針抽出のための新たな機械学習手法を開発した。

所望の材料機能の発現の鍵となる構成元素や原子配列の特徴を見出すことは、材料設計指針の構築や機能発現機構の解明において重要である。本手法は、機械学習の物性予測モデルに基づいて物質の分類を行うことにより、物質群・物性の種類を問わず、任意の無機材料データから所望の物性に応じて有望な物質のパターンを抽出することを可能にした。これにより、1,000 種類以上の物質を含む無機材料データから各エネルギー領域のバンドギャップを持つ物質や、広いバンドギャップと大きい屈折率を両立する物質に共通な特徴を事前知識無しに自動的に検知することに成功した。本手法により**マテリアルズインフォマティクス**（用語 1）を用いた物質・材料科学的な知識の獲得が明確・容易になり、さまざまな無機材料の研究・開発や学理構築が加速されることが期待される。

本研究成果は 8 月 5 日付（現地時間）で「*Advanced Intelligent Systems*」誌に掲載された。

●背景

材料科学分野では、物質の構成元素や原子配列の特徴に着目し、一定の基準において物質をさまざまな物質群に分類することが頻繁に行われてきた。例えば金属と酸素の化合物であるシリカ（SiO_2）やアルミナ（Al_2O_3）は酸化物という物質群にまとめられ、窒化物や硫化物などと区別して取り扱われる。また価電子構造に基づいた II-VI 族半導体（CdTe、ZnSe、ZnO など）、III-V 族半導体（GaAs、GaP、GaN、AlN など）といった分類や、結晶構造の観点で岩塩型構造やペロブスカイト型構造といった分類を用いながら、材料設計指針や機能発現機構についての議論が行われてきた。一般に材料設計指針や機能発現機構について考える際に、あらかじめ有望な物性を持つ物質群やその構成元素・原子配列の特徴を知ることができれば有益である。

ある機能を発現する鍵となる特徴や望ましい機能を持つ有望物質群は、想定している物性（例えば電気特性、光学特性、磁気特性、力学特性）や、材料の用途（電子材料、光学材料、磁性材料、構造材料など）に応じてさまざまである。したがって構成元素・結晶構造のどういった特徴・基準を用いるかについて無数の分類法が考案されており、所望の機能発現の鍵となる物質の構成元素・原子配列の特徴と、有望な物質群をその都度見出すことが必要となる。

一方で、近年は機械学習がさまざまな分野で爆発的に流行しており、材料科学分野も例外ではない。最も典型的な応用の一つは、物質の化学式や結晶構造から物性を高速に予測することであり、ここ 20 年ほどで非常に多くの研究例がある。このような研究では多数の物質について物性値を算出した結果をまとめた**第一原理計算**（用語 2）データベースがよく使われる。さらに最近では、機械学習によりデータを解釈・説明する手法も流行しており、大規模データから構成元素や結晶構造の特徴と物性の関連性を人間が理解できる知識として抽出するための手法も提案されている。例えば、**クラスタリング**（用語 3）と呼ばれるデータ分類法を用いると、あらかじめ人間が選択した**特徴量**（用語 4）について類似した物質群の分類が可能となる。しかし、通常のクラスタリング手法を適用した場合、あらかじめ構成元素・原子配列の特徴量を選択する必要があるため、上述したような用途に応じた物質分類を事前知識無しに行うことができないという問題がある。そこで本研究では、機械学習による物性・機能予測とクラスタリング手法を融合させることで、専門的な事前知識を必要とせずに、想定している物性と物質の構成元素・原子配列に基づいて合理的かつ自動的に物質群の分類を定義する手法を開発することとした。

●研究成果

本研究で用いたクラスタリング手法は、Breiman らの提案したランダムフォレスト分類器に基づいたクラスタリング手法を改変し、回帰モデルに適用できるようにしたものである。

通常、ランダムフォレストの予測モデルは多数の**決定木**（用語 5）から構成されているが、まずは本手法の概要を説明するために、1 本の決定木で分類が行われる様子を図 1 に示す。決定木は特徴量を用いた不等式から構成されており、例えば、図 1 では原子番号や原子間距離に基づいて各物質に物性値パターンを割り当てている。ある物性に関するデー

タを学習した決定木における不等式で使われる特徴・基準は、その物性を予測する上で適切なものが自動的に選択される。したがって、興味ある物性データを学習した決定木上で「同じ経路を辿った物質は類似度が高い」、「そうでないものは類似度が低い」と定義して物質の分類を行えば、対象とする物性に対して適切な基準で構成元素・原子配列の類似度を定義し、また類似度の高い物質をまとめて物質群を定義できるというのが本手法の骨子となるアイデアである。

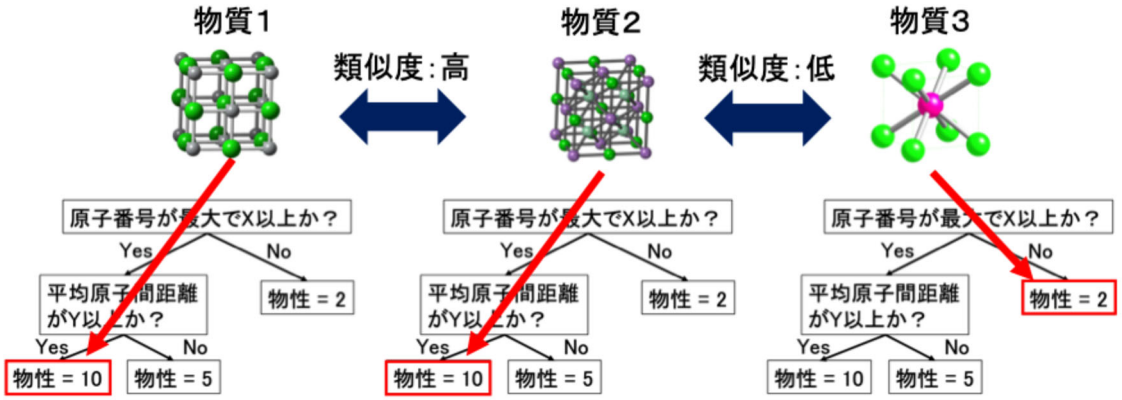

図 1　決定木による物性予測モデルの概略図

これを踏まえて具体的な手法の概略図を図 2 に示す。図 2(a)では物質の特徴量空間において、決定木のパターンがどのように表されるかを示している。実用的にはランダムフォレストモデルは予測精度の向上のため、多数（数百～数千程度）の決定木を用いる（図 2(b)）。したがって、その全ての決定木によるパターン分類を考慮した上で物質群の分類を行う必要がある。しかしながら、標準的なクラスタリング手法はこうした多数のパターンにより扱われる情報を直接適用できないため、各データ点（この場合は各物質に割り当てられたパターン）を表形式の数値データで表す必要がある。そのため、本手法では物質がそれぞれの決定木でたどり着いたパターンを **one-hot encoding**（用語 6）で表現して、特徴量空間（x 空間）から物性予測モデルに基づいた新しい空間（z 空間）に変数変換する。すなわち、図 2(c)のように、x 空間上で長方形（実際は多次元上の超直方体）の重なりとして表現されていた多数のパターンを、z 空間上で通常のベクトル値として表現することが可能となる。したがって、決定木の各パターンを表形式の数値データとして扱うことが可能となり、標準的なクラスタリング手法が適用可能となる。

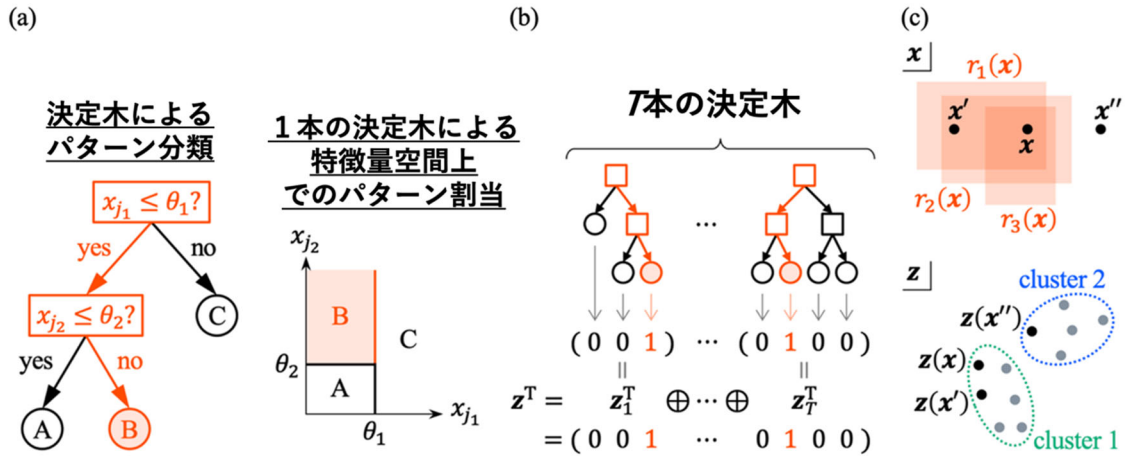

図 2　本手法による物質分類手法の概略図。(a)決定木によるパターン分類の特徴空間での振る舞い、(b)ランダムフォレストモデルの one-hot encoding による z 空間への変数変換、(c)特徴空間（x 空間）上で $\boldsymbol{x}$, $\boldsymbol{x}'$, $\boldsymbol{x}''$の特徴量で表される物質に対して T 本の決定木によるパターン割当を行い、z空間上で物質の分類を行う様子。

上記の手法の応用例を示すため、**Materials Project データベース**（用語 7）から約 1,000 種類以上の酸化物のデータを取得し、さらに **matminer コード**（用語 8）を用いて化学式や結晶構造に基づいた特徴量を約 700 個生成した。そのデータについて生成エネルギー・バンドギャップ・電子系誘電率の機械学習予測モデルを構築して、その予測モデルに基づいたクラスタリング、すなわち物質の分類を行った。ここでは代表的な結果として電子系誘電率の観点での分類について解説する。電子系誘電率はその平方根を取ると光の屈折率となり、光学用途において重要な物性である。特にバンドギャップが広く電子系誘電率の大きい物質は光学コーティングなどの用途で重要であるが、バンドギャップと電子系誘電率は一般的にトレードオフの関係にあり、広いバンドギャップと大きな電子系誘電率を両立するような物質の設計は難しい。

本研究では電子系誘電率に基づいてクラスタリングを行い、取得した酸化物データを 20 種類に分割することで図 3(a)のような結果が得られた。それぞれのクラスターで確かに電子系誘電率の値が類似した物質がまとまっている傾向が確認でき、そのうちの一つの物質群が比較的広いバンドギャップを持ちながら大きな電子系誘電率を持つことが分かる。さらにその物質群の特徴量の分布を全データと比較することで、有望物質群を特徴づける因子を特定した。例えば図 3(b)に示すように、八面体型配位構造が含まれるかどうかの指標に着目すると、全データの分布と比べて有望物質群が明らかに高い値を持つ傾向があることが分かる。このような解析から、この物質群の分類基準は解釈しやすいように簡略化して言えば「八面体配位した遷移金属元素が結晶構造に含まれること」であることが分かった。実際、図 3(c)で示すように、この物質群はペロブスカイト型構造やその類似構造を多く含んでおり、確かに「八面体配位した遷移金属元素」を有していることが分かる。さらにこうした物質の電子状態密度の第一原理計算データについて詳細な解析を行うことで、八面体配位したカチオンがバンドギャップの上端（伝導帯の下端）近傍の

電子状態の起源となっており、広いバンドギャップと高い誘電率を両立するための鍵となる因子であることが裏付けられた。

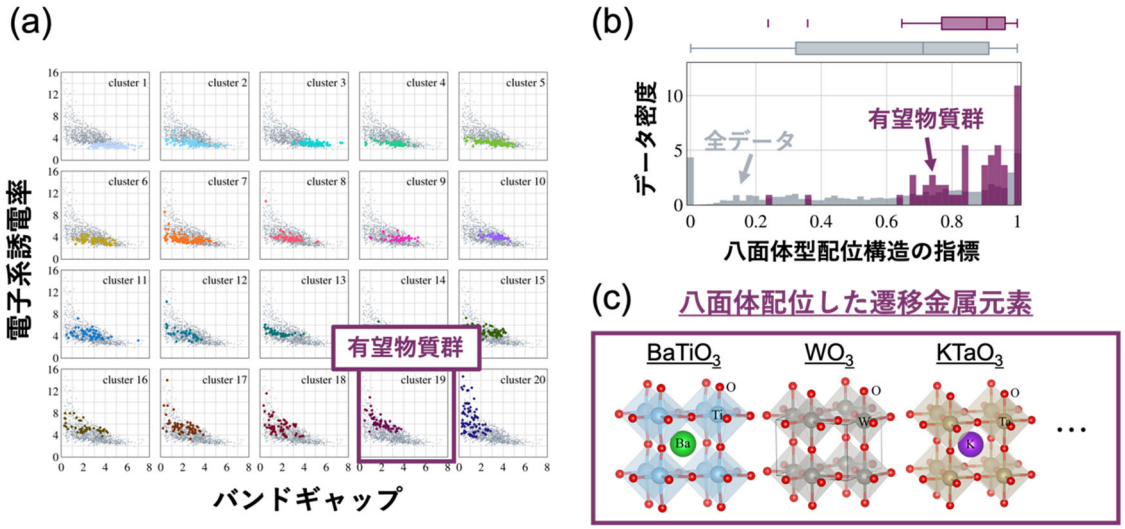

図 3 (a)電子系誘電率を基準として酸化物を物質群に分類した結果、(b)八面体型配位構造の指標の全データおよび有望物質群における分布、(c)有望物質群に分類された物質の例。

●社会的インパクト

近年、材料科学のさまざまな分野でロボットやシミュレーション・プログラミング技術を用いることにより実験・計算データの生成効率が飛躍的に向上しており、大規模データベースが続々と構築されている。本手法を用いて、その都度着目すべき物質群・分類基準が示されれば、人間が理解可能な形で材料設計指針や材料科学の学理構築に寄与することが期待される。さらに材料科学以外の分野でも「ある特性・機能を考慮した上で、データの入力情報・制御因子に着目してデータを分類したい、データ群を見出したい」というニーズも多分にあると見込まれるため、本手法はより広範な科学分野に応用できる可能性を秘めている。

●今後の展開

本研究グループは、高精度な第一原理計算により約 10 万物質の電子・光学的な特性を評価した大規模な計算材料データベースを保有しており、このデータベースと本手法を用いて半導体材料や光学材料の設計のための指針を提案する。また本研究では単一の物性のみを考慮したが、今後は本手法を拡張して複数の物性を勘案してクラスタリングを行うことで、より実用的な知識抽出を行う。

●付記

本研究は科学技術振興機構 戦略的創造研究推進事業 CREST（JPMJCR17J2)、日本学術振興会 科学研究費助成事業（JP20H00302、JP 21K14401)、文部科学省 データ創出・活用型マテリアル研究開発プロジェクト事業（JPMXP1122683430)、国際・産学連携イン

ヴァースイノベーション材料創出プロジェクト拠点 DEJI²MA プロジェクト、KISTEC 脱炭素化対策事業の助成を受けて行われた。

【用語説明】

(1) **マテリアルズインフォマティクス**：材料科学における実験および理論計算の結果に対して機械学習などのデータ科学手法を適用することで、膨大な種類の材料やその性質を扱うアプローチ。

(2) **第一原理計算**：量子力学の基本原理に基づいた理論計算。物質の電子構造やエネルギーを計算することにより、電子・光学・磁気特性や安定性、力学特性など非常に多様な物性や分子・結晶などの構造を予測することができる。

(3) **クラスタリング**：機械学習の手法の一つで、類似した特徴を持つデータ点をグループ（クラスター）にまとめる方法。

(4) **特徴量**：機械学習モデルに入力するデータの特徴を表す属性。例えば、材料科学における物性予測では、電気陰性度や近接原子間距離などの原子および原子配列の基礎的な特徴を用いることが多い。決定木予測モデル構築のアルゴリズムでは、学習データから物性予測のために適切な特徴量が自動的に選択される。

(5) **決定木**：データの特徴に基づいて不等式を繰り返しながらそれぞれのデータに特性のパターンを割り当て、特性の予測を行う機械学習の手法。

(6) **one-hot encoding**：カテゴリー型データを数値データに変換する手法。各カテゴリー（本研究では決定木により割り当てられたパターン）を 0 と 1 の組み合わせで表現する。

(7) **Materials Project データベース**：材料科学分野の大規模オープンデータベース。第一原理計算により得られた物質の構造や特性に関するさまざまな情報を提供しており、2024 年 7 月時点で 15 万種類以上の無機物質データを掲載している。(https://next-gen.materialsproject.org)

(8) **matminer コード**：材料データの取得、処理、機械学習用の特徴量抽出を行うための Python ライブラリ。材料科学とデータ科学を橋渡しし、マテリアルズインフォマティクスの研究を支援するツール。(https://hackingmaterials.lbl.gov/matminer/)

【論文情報】

掲載誌：*Advanced Intelligent Systems*
論文タイトル：Target Material Property-Dependent Cluster Analysis of Inorganic Compounds（対象物性に依存した無機化合物のクラスター分析）
著者：Nobuya Sato, Akira Takahashi, Shin Kiyohara, Kei Terayama, Ryo Tamura, Fumiyasu Oba（佐藤暢哉、高橋亮、清原慎、寺山慧、田村亮、大場史康）
DOI：10.1002/aisy.202400253