様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ４ 日現在

機関番号：８２１０８
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ０４７１６

研究課題名（和文）材料インフォマティクスのスモールデータ問題に対する階層的機械学習モデリング

研究課題名（英文）Hierarchical machine learning for small data problems in materials informatics

研究代表者
小山　幸典（KOYAMA, Yukinori）

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・主幹研究員

研究者番号：２０４３７２４７

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：物性を予測する機械学習モデルの構築においてデータ数が少ない「スモールデータ問題」に対応するために、材料データや物性の関係を活用できる機械学習モデルの構築を検討した。複数の目的変数の間に階層的な関係があり、欠測パターンが単調なデータに対しては、コクリギング法を用いた機械学習モデルが有効に機能することを確認した。一方、現実のデータセットの欠測パターンは必ずしも単調ではない。そこで、単調でない欠測パターンに対してもデータ間の相関から欠測値を推定する多重代入法を用いることで、目的変数間に明確な階層的な関係がなくても機械学習を用いた予測が機能することを明らかにした。

研究成果の学術的意義や社会的意義
機械学習を用いた物質の物性値推定はマテリアルズ・インフォマティクスの主要な課題のひとつである。しかし、個々の研究者が注目する特定の材料群や物性に限るとデータ数は少なく、これが材料研究における機械学習を困難にする「スモールデータ問題」が存在している。本研究では、データの階層構造や物性の相関関係を考慮し、複数の物性を同時に取り扱う機械学習が「スモールデータ問題」に対して有効な対応策となることを示すことができた。

研究成果の概要（英文）：Amount of data in materials science is small for specific material classes and properties, and this causes difficulty in machine learning. This is the small data problems. In this study, several machine learning techniques that can take account of relationship among multiple properties have been examined to overcome the small data problems. Cokriging is confirmed as an effective technique for multiple properties having hierarchical relationship and datasets of monotonic missing patterns. Missing patterns of real data are, however, not monotonic in general, and cokriging is not effective for such datasets. Multiple imputation techniques have been examined for datasets whose missing patterns are not monotonic and target properties have no explicit hierarchical relationship. The results suggest that the multiple imputation techniques are useful even for properties having no explicit hierarchical relationship.

研究分野：無機材料および物性

キーワード：マテリアルズ・インフォマティクス　機械学習　多重信頼度モデル　多重代入法

## １．研究開始当初の背景

機械学習とは、データ解析の技法を大量のデータに適用することで、そのデータから有用な知識を抽出する、さらに、その抽出した知識を用いて、新たなデータについて予測を行なう技法である。この機械学習は、近年注目されているマテリアルズ・インフォマティクス（材料インフォマティクス）の中核技術のひとつである。例えば、化学組成と生成熱などの物性値の相関を既知のデータから発見し、その相関を用いてデータに含まれていない物質の性質を予測することを目指している。化学組成と物性値のような複雑な関係を見出すためには多数のデータが必要である。無機化合物の種類は、化学組成だけを考える場合でも 3 元系で約 800 万との見積りがある。一方、既知の無機化合物の数はと言うと、国立研究開発法人物質・材料研究機構が公開している無機材料データベース AtomWork Adv.に収録されている物質の数は 13 万件に限られる。さらに、物性値の収録数は物性により大きく異なるが、収録数が数千件という物性は希であり、数百件から数十件のデータしか収録されていない物性が大半である。機械学習を用いた物性値の推定はマテリアルズ・インフォマティクスの主要な課題のひとつであるが、多くの報告では、対象物性を体積、生成熱、バンドギャップなど、第一原理計算で容易に計算可能で多数のデータが利用可能なものとする（例えば、Materials Project には約 7 万件の計算データが収録されている）、あるいは、比較的少数のデータに適用するために物質の範囲を限定しているというのが実情である。マテリアルズ・インフォマティクスを発展させていくためにはデータベースの拡充が重要であるが、大量のデータの収集は一朝一夕にできるものではない。いわゆる「スモールデータ問題」はマテリアルズ・インフォマティクスで共通の課題である。一方、材料科学では物理モデルや経験式など様々な知識が蓄積されている。この「事前知識」を活用することで「スモールデータ問題」を克服できるのかという問いが本研究の主題であり、この解決がマテリアルズ・インフォマティクスにおける「スモールデータ問題」のブレイクスルーとなることは間違いない。本研究では物性値の推定に焦点を絞るが、合成や分析、ミクロ・マクロな構造においても同様の「スモールデータ問題」が存在していることは言うまでもない。

## ２．研究の目的

本研究計画では、材料科学の知識（物理モデルや経験式）である物性の関係性や階層構造を機械学習モデルに組み込み、材料データベースを用いて関連するデータ科学手法の有用性を具体的に実証し、より先進的なモデリング方法へと発展させる。ここでは、物質 X の物性 Y を推定する場合について基本的なアイデアを説明する。図 1 の「従来型」に示すように、機械学習では $Y = f(X)$ となる関数 f（の近似式）を求めるという問題となる。Y のデータが少ない場合、f として複雑なモデルを用いることはできず、高い推定精度を得ることは困難である。マテリアルズ・インフォマティクスではこのような状況がよく発生している。一方、別の物性 Z は目的物性 Y と相関が強いことが知られており、Z は Y よりもより多くのデータが得られているとする。この場合、 $Y = f(X)$ の関係を直接学習するのではなく、 $Y = g(Z, X)$ と $Z = h(X)$ の関係を学習するという方策（図 1 の「階層型」）が考えられる。Y と Z の相関が強ければ、その関係 g は簡単なモデルで表すことができ、また、Z のデータは Y よりも多いため、X との関係 h をより高い精度で推定できると期待できる。したがって、少ない Y のデータでも十分な精度での推定が可能になると期待できる。なお、この説明では単一の関連物性 Z を用いた 2 段階の推定を例に挙げたが、複数の関連物性や多段階の推定に拡張可能であることは容易に想像できる。

このように、機械学習のモデリングに材料科学の事前知識を反映させるアイデアは、関係性をデータから発見することを目指すデータ駆動型研究の弱点を補強するものである。また、上記の 2 段階推定は、単に物性 Z を記述子として用いるということではない。Z を記述子として用いると、Z が未知の物質では Y を推定することはできず、材料探索の範囲が限定される。これも、従来研究でしばしば見られる状況である。そうではなく、Y の推定では Z を、その不確かさも含めてベイズ推論的に考慮することが、本アイデアの重要な点である。

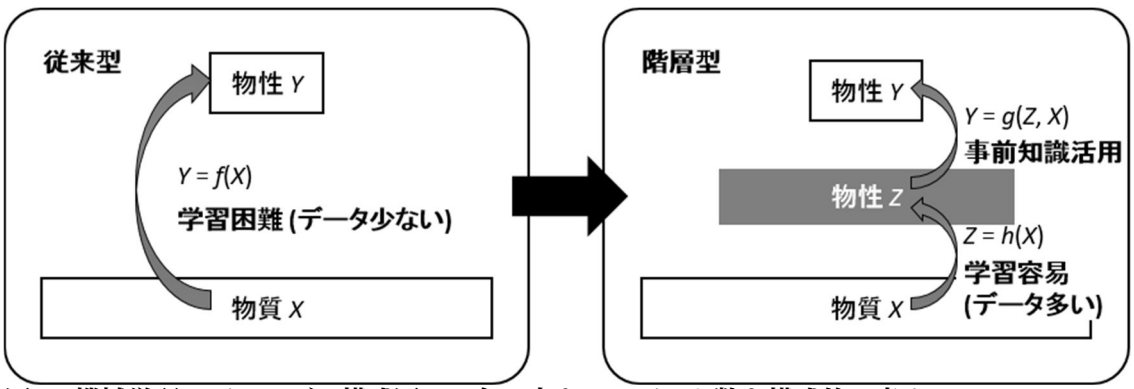


図 1. 機械学習モデリングの模式図。四角の大きさはデータ数を模式的に表す。

## ３．研究の方法

目的物性 Y と関連物性 Z の関係として以下の 3 通りを考えている。ケース①からケース③になるほど Y と Z の関係が弱く、問題は難しくなる。したがって、ケース①からケース③の順に、前段階で得られた知見を生かしながら、問題を解決していく計画である。

### ケース① Y と Z は本質的に同等だが、評価精度が異なる

本ケースは multi-fidelity model（多重信頼度モデル）と呼ばれ、具体的な使い方は、Y が実験値、Z が計算値、あるいは、Y に対し Z は簡便な方法での実験・計算結果という場合への適用である。Z は Y に対して系統的な誤差を含むと考えられる。多重信頼度モデルは流体力学シミュレーションの分野での適用例が多いが、物質への適用は限られている。

### ケース② Y と $Z_1$, $Z_2$, …の関係（物理モデル、経験則）が知られている

例えば、融点はデバイ温度、弾性率、凝集熱などに比例するという関係が知られている。後者の物性は、計算値も含めれば多数のデータが利用可能である。一方、融点は計算が困難な物性であり、利用可能なデータ数は少ない。そこで、デバイ温度、弾性率、凝集熱など関連物性の推定と、関連物性から融点の推定を組み合わせ、融点を一気通貫で推定する機械学習モデルを構築する。

### ケース③ Z は物質 X の重要な特徴であるが、全ての物質について既知ではない

結晶構造や電子状態などは物質の根源的な特徴であり、これらの特徴を取り入れることで物性推定の精度が向上するとの報告がある。結晶構造や電子状態は多数のデータが利用可能であるが、新物質の探索という状況ではこれらの特徴が使用できると限らない。そこで、結晶構造や電子状態を利用した物性推定の機械学習モデルと、それらの特徴を推定する機械学習モデルを構築し、一気通貫の物性推定を目指す。

## ４．研究成果

「研究の方法」で挙げたケース①の代表例として、第一原理計算で求めたバンドギャップを化合物記述子から推定するという問題を設定し、多重信頼度モデルにおいて一般的に用いられているコクリギング法を適用した。下位モデルは、計算コストが低い GGA レベルで求めたバンドギャップを化合物記述子からガウス過程回帰を用いて推定するものとした。一方、上位のモデルでは、計算コストがより高いハイブリッド法で求めたバンドギャップを、化合物記述子と GGA バンドギャップから単純な線形モデルを用いて推定した。このような階層モデルを用いることで、ハイブリッド法バンドギャップのデータだけを用いて線形モデルやガウス過程回帰で予測した場合よりも、高い予測精度が得られることを確認した。また、GGA バンドギャップが得られていない場合でも、GGA バンドギャップの確率分布を推定する下位モデルの結果を利用することで、ハイブリッド法バンドギャップを予測することができることを確認した。

このように階層的な機械学習モデルがスモールデータ問題に有効であることが示唆された一方で、本手法の課題が明らかになった。コクリギング法では、上位モデルのデータ推定において、下位モデルのデータが欠測していた場合は推定を重ねることができ、これがスモールデータ問題に対して有力な手段となる。一方で、機械学習の学習段階で使用するデータセットでは、上位モデルのデータには下位モデルのデータが必ず存在する必要がある。このようなデータパターン（欠測パターン）を「単調である」という。GGA レベルとハイブリッド法の第一原理計算の場合、ハイブリッド法のみの計算結果しかないデータに対して GGA レベルの計算を追加し、欠測パターンを単調にすることは容易である。しかし、実験データを目的変数とし、その下位レベルとして第一原理計算のデータを当てはめようとすると、実験データが観測されている全てのデータに対して第一原理計算が必ずしも実行可能ではないことから、欠測パターンは必ずしも単調とはならない。コクリギング法では単調な欠測パターンのデータしか使用することができないことから、第一原理計算のデータがない場合は実験データを使用しないことで欠測パターンを単調にする必要があった。このように修正したデータセットに対してコクリギング法を適用することができたが、元々少ない実験データの一部が使用できなくなるという点が課題となった。また、「研究の方法」で挙げたケース②のように物理モデルや経験則に基づいた階層モデルを構築しようとしたところ、個々の下位データのデータ数は少なくないものの、全ての下位データがそろっている上位データの数が非常に少ないことが分かった。コクリギング法を適用する場合は、データの欠測パターンが単調であることの必要性から、下位データの一部が欠けている上位データを使用することができず、利用可能なデータ数が非常に少なくなってしまうという問題が明らかになった。

そこで、非単調な欠測パターンのデータに適用可能であり、全てのデータを使用することが可能な、多重信頼度モデル以外の機械学習手法を検討した。予備検討の結果、データ補完手法の一つである多重代入法が、非単調な欠測パターンに適用可能であり、また、データに質的な上位・下位の概念はなく、データ間に相関があれば適用可能とされており、有望であることがわかった。「研究の方法」で挙げたケース②、および、ケース③に対して、融点を代表例として多重代入法を検討したところ、物性同士や目的物性と基本的な特徴の間に線形相関のような単純な関係が

見られる場合は高い推定精度が得られた。対数をとるなど適切なデータ変換を行うことで推定精度が大きく向上することがあり、物性・記述子間の関係が重要であることを示唆する結果となった。欠測パターンに対しては、人工的に欠測を生じさせたデータセットを検討したが、推定精度に明確な傾向は得られなかった。欠測パターンよりもデータセットの特徴の方が推定精度に強く影響していると考えられる。データセットの大きさに関しては、データ数や記述子・物性値の数が多いほど推定精度が高くなる傾向があるが、数が少ない場合は推定精度が低くなったが、多ければ多いほど良いというわけではなかった。これらの結果は機械学習における一般的な傾向と合致する。物性同士や目的物性と基本的な記述子の間に線形相関のような単純な関係が容易に得られない場合の推定精度は低かった。多重代入法で用いる予測モデルとして非線形な機械学習モデルを使用することは原理的には可能であるが、データを 1 点推定するごとに予測モデルの再学習が必要となり、現状では現実的な計算時間で満足のいく推定精度を得ることができなかった。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|