4J1-2in2

材料物性予測のための原子間距離に基づくカーネル設計

Interatomic-Distance-Based Kernels for Material Property Prediction

秋田 大空 *1 Hirotaka Akita　馬場 雪乃 *1 Yukino Baba　鹿島 久嗣 *1 Hisashi Kashima　世古 敦人 *2 Atsuto Seko

*1京都大学大学院情報学研究科知能情報学専攻
Department of Inteligence Science and Technology, Graduate School of Informatics, Kyoto University

*2京都大学大学院工学研究科材料工学専攻
Department of Materials Science and Engineering, Graduate School of Engineering, Kyoto University

Comprehensive search over various candidate materials is an important step to discover novel materials with desirable physical properties; however its search space is quit vast so that it is limited to perform exhaustive experiments to check all of the candidates. Even if the chemical composition is the same, the properties of the material may differ greatly depending on the crystal structure, and therefore the number of possible combinations significantly increases. Recently, machine learning methods are successfully applied to material discovery that estimates prediction models using existing databases to predict physical properties of unknown substances. In this research, we propose a novel kernel function between compounds, which directly uses structure information for the prediction of physical properties of inorganic crystalline compounds based on the crystal structures. We carry out evaluation experiments on each and show that the structure information improves prediction accuracy.

1. 序論

1.1 背景

硬さや弾性，そして半導体としての性質など，望ましい物性を持った材料を見つけるためには様々な材料を網羅的に探索する必要がある．望ましい物性を持った材料を見つけるためには様々な材料を網羅的に探索する必要がある．このような，既に存在する材料の物性よりも優れた物性を持つ材料を見つける問題を材料探索という．

しかし，そのような化合物を見つけようとした場合，元素や組成，結晶構造など様々なパラメータが存在するため，探索するべき化合物の数はとても多くなってしまう．そのため，化学者が経験と勘によって化合物を精製してから，化合物を物理的に叩いて延性を調べたり，熱して融点を調べたりする実験的な手法には限界がある．また，近年ではスーパーコンピュータを用いて原子一つ一つについてモデル化をし，シュレディンガー方程式を数値的に解くことで物性を予測する (第一原理計算と呼ばれる) といった手法も使われているが，それでもなお計算量を考えると時間的・計算資源的なコストが非常に高く，網羅的に探索を行う上では未だに非常に効率が悪いため，より高速に，かつ無駄なく材料探索を行う方法が必要とされている．

そこで，既存の化合物のデータベースを用いて物性値を予測するモデルを構築し，構築したモデルを用いることで，生成するのが難しい化合物や未知の化合物の物性を予測するという機械学習手法の応用が，材料探索を網羅的に行う上で重要となってきている．

1.2 研究の概要

材料の性質を機械学習によって予測する場合には，物質の特徴を数値で表して特徴量を生成する必要がある．より良い予測結果を出すためには，材料の物性値を知るために必要な情報が特徴量に含まれていなければならない．化合物の性質を予測する場合，まず最初にその化合物のにどのような元素がどれだけの割合で含まれているかを特徴量とし，各元素の周期表の族や融点・沸点，電気陰性度などの元素情報を用いて回帰を行うのが自然な発想である．しかし，元素の情報を用いるだけでは，同一の組成から成り立っているにもかかわらず，それぞれが互いに全く異なる物理的な性質を示す材料に対して，異なった値を導き出すことができないこのようなケースに対応できるよう，化合物の物性を正確に予測するためには，化合物のもう一つの重要な要素である「結晶構造」も用いて予測モデルを構築する必要がある．

含有されている元素の情報については，電気陰性度や沸点・融点，イオン半径など，既に観測され，数値化されているデータがあるため，ベクトルとして表すことが容易だが，結晶構造については，3 次元空間における原子配置の幾何学的情報が含まれているため，簡単に数値化することができない．そのため，構造情報をいかに予測モデルに組み込むかということが，予測モデルの精度を向上させる上で重要になってくる．本研究では，構造情報のうち，無機結晶化合物の最小の繰り返し単位であるユニットセル内のある原子と，そこから一定のしきい値の距離までにある原子との距離に着目し，「距離情報を予測モデルに組み込めるようにする」を目標として，構造情報を数値化できるような特徴の抽出方法を検討する．材料は「組成」と「立体構造」の二つによって決定されるため，岩塩型構造やスピネル型構造などといった，既に調べられ，あるいは計算されてデータベースとしてまとめられている立体構造を用いることで，全く新しい結晶構造をもつものでなければ，未知の材料であっても表現することが出来る．

1.3 提案する解決法

本研究では，ユニットセル内に存在する原子から構成される原子のペアの原子間距離と，両端の原子の元素的性質を考慮した適切なカーネルを用いることにより，立体構造の情報および元素情報を保持したままでカーネルを用いた回帰を行って，材料の物性値を予測することを提案する．古典的な機械学習手法では，扱えるデータの形はベクトルに限定されているため，先行研究においては，化合物に含まれている元素の情報に加えて，

連絡先: 秋田大空，京都大学情報学研究科知能情報学専攻，
h_akita@ml.ist.i.kyoto-u.ac.jp

ある原子から見てどの距離にどのような原子が存在するかといった立体構造情報を，分布のヒストグラム化やカーネル密度推定などを行って全てベクトルに変換して特徴量とし，それをRBF カーネルやシグモイドカーネル等を用いてサポートベクター回帰やカーネルリッジ回帰を行っている [Nakayama 15]．立体構造のように内部構造を持ったデータに対する特徴ベクトルの構成は自明でなく，このような従来の手法では，ベクトルに変換する過程で構造情報の一部が欠落してしまうことが考えられる．そのため，ベクトルへの変換を介さず化合物の構造から直接カーネルを設計することができれば，さらに精度を向上させることが出来ると考えられる．

2. 関連研究

2.1 有機化合物への機会学習応用

有機化合物は複数の原子が結合して電気的に中性となった分子を最小の構成単位とする．Rupp らは，分子の表現として，原子群の単純な行列表現であるクーロン行列を導入した [Rupp 12]．Hansenn らはクーロン行列を拡張し，線形回帰やカーネルリッジ回帰等様々な機械学習手法を用いて，有機化合物の物性を予測するモデルを構築し，クーロン行列による分子表現の有用性を実証し，用いるべき機械学習手法について議論した [Hansen 13]．Montavon らは，ニューラルネットワークを用いて，有機化合物の分子構造から，原子化エネルギー，分極率，フロンティア軌道固有値，イオン化ポテンシャル，電子親和力および励起エネルギーを同時に求めるモデルを構築した [Montavon 13]．

2.2 無機化合物への機械学習応用

有機化合物が複数の原子が結合して構成する分子の形で表現されるのに対し，無機化合物の多くは同一の結晶構造が無限に続く形で表現される．Schutt らは，今まで機械学習による化合物の様々な性質の予測に用いられてきたクーロン行列が，同様の構造が一様に無限に続く無機化合物の性質予測には有用でないことを示し，X 線粉末回折パターンやテキストマイニングに用いられている部分動径分布関数を用いて，結晶構造の特徴量の表現を提唱し，電子特性を予測するモデルを構築した [Schütt 14]．また，Seko らは，第一原理計算による時間的コストが非常に高い無機材料の格子熱伝導率 (Lattice Thermal Conductivity :LTC) を，ベイズ最適化を用いたガウス過程回帰によってバーチャルスクリーニングを行った [Seko 15]．また，中山らは，化合物の性質をそれに含まれる元素の情報の平均や分散，そして原子間の距離および存在する原子の電気陰性度などに着目して特徴ベクトルを構成する方法について検討した [Nakayama 15]．これらの研究は，化合物の情報から特徴ベクトルを生成することで機械学習手法を用いている．同様に無機結晶化合物の物性予測を取り扱う本研究では，中山らによる化合物の特徴生成の成果を参考に，化合物の構造情報を特徴ベクトルに変換するのではなく，二つの化合物の間のカーネル関数を定義し，直接化合物同士の類似度を比較するアプローチを提案する．

3. 問題設定

本研究で取り組む問題は，構造情報を含んだ化合物のデータベースを用いて，新たな化合物の物性を予測するモデルを構築することであり，化合物の集合 $\mathcal{X}$ および物性値の集合 $\mathcal{Y}$ とデータ集合 $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^{N}, \boldsymbol{X}_i \in \mathcal{X}, y_i \in \mathbb{R}$ が与えられた場合に，物性値が未知の $\boldsymbol{X} \in \mathcal{X}$ に対して予測値 y を出力する写像 $f : \mathcal{X} \to \mathbb{R}$ を学習する教師あり回帰問題の枠組みで定式化することが出来る．本研究では，化合物 $\boldsymbol{X}_i$ を，式 (1) のような L 個の要素からなる集合で表す．

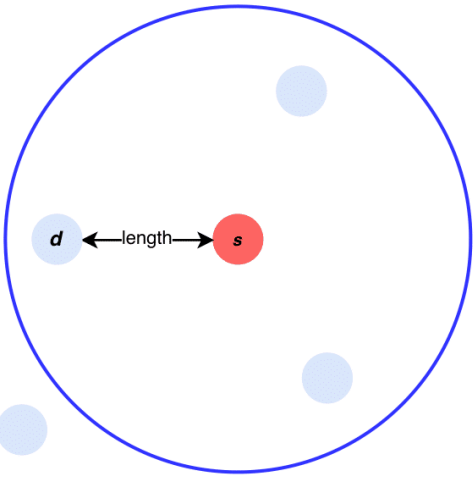

図 1: 図のような原子の分布が存在する時，赤色の原子 s と s から length の距離に存在する原子 d から，始点を s，終点を d，長さを length とする近傍原子ユニットが構成される．また，周囲の青線の円はしきい値を表し，s を起点とした場合はこの円の外に存在する原子からは近傍原子ユニットを構成しない．

$$\boldsymbol{X}_i = \{x_{ik} \mid k = 1, \ldots, L\} \tag{1}$$

各 x_{ik} を**近傍原子ユニット**と呼ぶことにする．近傍原子ユニットは，次のようにして構成される．まず，ユニットセル内のある原子 s に着目する．これを近傍原子ユニットの**始点** (source) と呼ぶことにする．始点の原子から見て，あるしきい値内の距離に原子が存在すれば，そのうちの一つを選択する．これを近傍原子ユニットの**終点** (destination) と呼ぶことにする．そして，始点の原子から終点の原子までの距離を，近傍原子ユニットの**長さ** (length) と呼ぶことにする．このようにして，始点，終点，長さの三つの要素を持った近傍原子ユニットを構成する．構成の模式図を図 1 に示す．また，化合物内には同一の種類の原子も含まれているため，区別できるように，ユニットセル内の各原子には 1 から順番にインデックスを振っておき，その情報も近傍原子ユニットに追加する．この作業を，原子 s から見てしきい値内の距離に存在する原子全てについて行うことで，複数の近傍原子ユニットが構成される．ユニットセル内の全ての原子について同一の作業を行うことで，化合物は上記の近傍原子ユニットの集合に変換される．

以上より，各 x_{ik} は式 (2) に示すような 4-tuple で表現される．

$$x_{ik} = (n_{ik}, \boldsymbol{s}_{ik}, \boldsymbol{d}_{ik}, l_{ik}) \tag{2}$$

各要素の詳細を以下に示す．

- n_{ik}
  近傍原子ユニットの始点の元素の，ユニットセル内でのインデックスを表す整数値である．1 から始まる．

- $\boldsymbol{s}_{ik}$
  近傍原子ユニットの始点の元素の特徴量を格納したベクトルである．
- $\boldsymbol{d}_{ik}$
  近傍原子ユニットの終点の元素の特徴量を格納したベクトルである．
- l_{ik}
  近傍原子ユニットの長さを表すスカラーである．

4. 原子間距離に基づく化合物カーネル

4.1 近傍原子ユニット同士の類似度

先行研究においては，化合物の立体構造情報を，分布のヒストグラム化やカーネル密度推定などを行ってベクトルに変換して特徴量として用いていた [Nakayama 15]．しかし，立体構造のように内部構造を持ったデータに対する特徴ベクトルの構成は自明でなく，このような方法では，ベクトルに変換する過程で構造情報の一部が欠落してしまうことが考えられる．そこで我々は，ベクトルへの変換を介さず化合物の構造から直接カーネルを設計することで，情報欠落に対処する．二つの近傍原子ユニット x_{ik} および $x_{jk'}$ の間の 3 種類のカーネルを，(3)，(4)，(5) に示す式でそれぞれ定義する．

$$D_1(x_{ik}, x_{jk'}) := \exp(-\lambda_1(l_{ik} - l_{jk'})^2) \tag{3}$$

$$D_2(x_{ik}, x_{jk'}) := \exp(-\lambda_2||\boldsymbol{s}_{ik} - \boldsymbol{s}_{jk'}||^2) \times \exp(-\lambda_2||\boldsymbol{d}_{ik} - \boldsymbol{d}_{jk'}||^2) \tag{4}$$

$$D_3(x_{ik}, x_{jk'}) := D_1(x_{ik}, x_{jk'})D_2(x_{ik}, x_{jk'}) \tag{5}$$

D_1 は，距離情報のみ，D_2 は始点と終点の元素の性質のみ，D_3 は両方を考慮するカーネルである．また，λ_1 および λ_2 はパラメータである．

4.2 化合物同士の構造類似度

近傍原子ユニット同士のカーネルを用いて，化合物 $\boldsymbol{X}_i$ と化合物 $\boldsymbol{X}_j$ に対するカーネル $K_v(\boldsymbol{X}_i, \boldsymbol{X}_j), v = 1, 2, 3$ は，化合物内の近傍原子ユニットそれぞれについて全組み合わせを取ってそれぞれのカーネルを計算し，その平均値と定義する．

$$K_v(\boldsymbol{X}_i, \boldsymbol{X}_j) := \frac{1}{|\boldsymbol{X}_i||\boldsymbol{X}_j|}\sum_{k=1}^{M}\sum_{k'=1}^{N} D_v(x_{ik}, x_{jk'}), \quad v = 1, 2, 3 \tag{6}$$

4.3 元素情報の類似度

以上に加えて，必要に応じて元素情報の類似度も定義する．化合物 $\boldsymbol{X}_i$ と化合物 $\boldsymbol{X}_j$ の，元素情報に基づいて計算されるカーネル T を式 (7) で定義する．

$$T(\boldsymbol{\phi}_a(\boldsymbol{X}_i), \boldsymbol{\phi}_a(\boldsymbol{X}_j)) := \exp(-\lambda||\boldsymbol{\phi}_a(\boldsymbol{X}_i) - \boldsymbol{\phi}_a(\boldsymbol{X}_j)||^2) \tag{7}$$

ここで，λ はパラメータであり，$\boldsymbol{\phi}_a(\boldsymbol{X}_i)$ および $\boldsymbol{\phi}_a(\boldsymbol{X}_j)$ はそれぞれ元素情報の平均から構成される特徴ベクトルである．また，このカーネルは RBF カーネルである．

5. 評価実験

5.1 データセット

化合物 500 件の構造情報データと，予測値である凝集エネルギーの値，および 32 種類の元素の性質をまとめたテーブルを用いて実験を行った．テーブルに記載される元素情報としては，原子番号，原子質量，周期，族，第 1 イオン化エネルギー，第 2 イオン化エネルギー，電子親和力，ポーリングの電気陰性度，アレンの電気陰性度，ファンデルワールス半径，共有半径，原子半径，s 軌道の擬ポテンシャル半径，p 軌道の擬ポテンシャル半径，融点，沸点，密度，モル体積，融解熱，気化熱，熱伝導率，比熱の 22 種類を用いた．

5.2 実験手順

4 章において定義した 3 種類のカーネル K_1 (近傍原子ユニットの長さのみ)，K_2 (近傍原子ユニットの始点と終点)，K_3 (近傍原子ユニットの始点と終点の長さ) について，近傍原子ユニットを構成する際のしきい値を 6 Å に設定し，カーネル関数の値を計算した．また，カーネル K_2,K_3 は元素の情報を加味しているが，しきい値を化合物のユニットセルの大きさとは無関係に固定した値にしているため，ユニットセル外の隣のユニットセルに存在する原子の元素情報を取り込んでしまったり，単一のユニットセル内の全ての元素の情報を取り込むことができない等の問題が考えられる．そのため，カーネルによって表現されたユニットセルの組成とユニットセルの実際の組成との間にはズレが生じてしまう．この問題を解決するため，ユニットセル内の元素の正しい組成を反映できるよう，同じく 4 章で定義した元素情報のみを用いたカーネル T を計算した．ここで，ニつのカーネルの積はまたカーネルになるため，二つの情報を統合するためにカーネルの積を用いることにした．そのためにカーネル T を K_1,K_2,K_3 とそれぞれ要素同士の積を計算することで，新たなカーネルを三つ作成し，合計 6 個のカーネルを計算した．その後にそれらを用いてサポートベクター回帰を行い，得られたモデルの精度を評価した．

5.3 評価指標

データ全体で 4-fold のクロスバリデーションを行い，その時の RMSE の平均値を評価指標として用いた．

5.4 比較手法

各手法が識別能力を有するかを確認するためのベースラインとして，訓練データ全体の凝集エネルギーの平均値を，全テストデータの予測値とした場合の誤差も算出した．この手法によって算出された値を上回る誤差を算出するモデルは，予測能力がないことが分かる．

また，提案手法と比較するため，先行研究 [Nakayama 15] による次のような手法についても，同様の実験を行った．

- 手法 1: 元素情報のみ
  化合物内に含まれる元素情報の平均を用いたベクトル $\boldsymbol{\phi}_a(\boldsymbol{X}_i)$ を特徴ベクトルとした．
- 手法 2: 元素情報+ヒストグラム
  元素情報ベクトル $\boldsymbol{\phi}_a(\boldsymbol{X}_i)$ と，各化合物の原子間距離の分布をヒストグラム化したベクトル $\boldsymbol{\phi}_h(\boldsymbol{X}_i)$ を結合して構成したベクトル $\begin{pmatrix} \boldsymbol{\phi}_a(\boldsymbol{X}_i) \\ \boldsymbol{\phi}_h(\boldsymbol{X}_i) \end{pmatrix}$ を特徴ベクトルとした．
- 手法 3: 元素情報+カーネル密度推定
  手法 1 による元素情報ベクトル $\boldsymbol{\phi}_a(\boldsymbol{X}_i)$ と，各化合物の原子間距離の分布をガウス関数の線型和で表したときの係数ベクトル $\boldsymbol{\phi}_a(\boldsymbol{X}_e)$ を結合して構成したベクトル $\begin{pmatrix} \boldsymbol{\phi}_a(\boldsymbol{X}_i) \\ \boldsymbol{\phi}_e(\boldsymbol{X}_i) \end{pmatrix}$ を特徴ベクトルとした．

表 1: 各手法の RMSE．ユニットの長さと始点終点を両方考慮することで（K_3），元素情報の平均だけを用いる手法よりも高い予測精度を達成できた

手法	カーネル	RMSE
訓練データの平均	-	1.178
元素情報	(RBF)	0.301
元素情報&ヒストグラム	(RBF)	0.280
元素情報&カーネル密度推定	(RBF)	**0.248**
長さのみ	K_1	1.195
始点と終点	K_2	0.320
長さ&始点と終点	K_3	**0.298**
元素情報&長さ	$T \times K_1$	0.301
元素情報&始点と終点	$T \times K_2$	0.310
元素情報&長さ&始点と終点	$T \times K_3$	0.305

以上の 3 つの既存手法について，特徴ベクトルから RBF カーネルを算出し，提案手法と同様にサポートベクター回帰を行って比較対象とした．

5.5 結果

500 件のデータセットについて，使用したカーネル行列と，それを用いたときの評価誤差の組み合わせのを表 1 に示す．

近傍原子ユニットの長さと始点及び終点を考慮した場合，元素情報の平均だけを見るよりも僅かに精度が向上することがわかった．元素の平均の情報と提案手法であるカーネルを組み合わせて使った場合，近傍原子ユニットの長さ単体と組み合わせたケースが一番精度が高くなった．近傍原子ユニットの長さのみを用いたカーネルは，訓練データの平均を全テストデータの予測値とした場合の誤差よりも大きい誤差を算出したため，単体では予測能力がないことが分かった．

5.6 分析

500 件のデータを用いた実験において，構造情報単体を用いた場合は，「近傍原子ユニットの長さのみ」，「始点と終点」，「始点と終点&近傍原子ユニットの長さ」の順に誤差が小さくなり，近傍原子ユニットの長さが精度の向上に寄与することを確認できた．一方，ユニットセルの組成に合わせた元素情報の平均を共に用いた場合は，「元素情報&近傍原子ユニットの始点と終点」，「元素情報&近傍原子ユニットの長さ&始点と終点」，「元素情報&近傍原子ユニットの長さのみ」の順に誤差が小さくなった．また，始点と終点の元素情報とユニットセル内の元素情報を用いた場合については，ユニットセル内の元素情報のみを用いた場合よりも精度が上がらず，逆に 0.305 に低下する結果となった．これは，「近傍原子ユニット始点および終点の情報」と「元素情報の平均」は両方元素の情報を加味したデータの表現であるため，その両者の積を取っていることで元素の情報を重複して使っているために生じたと考えられる．

6. 結論

本研究では，「無機結晶化合物の結晶構造から化合物の性質を予測する」という問題に対して，データを特徴ベクトルに変換するのではなく，データ間のカーネルを直接定義して解くアプローチを提案した．

「近傍原子ユニットの長さのみ」「近傍原子ユニットの始点と終点の元素情報」「近傍原子ユニットの長さおよび始点と終点の元素情報」をそれぞれ考慮するカーネルを設計し，精度を比較することで，構造データをベクトル化せずにカーネルを用いて直接類似度を求める手法が，識別能力を有することを確認できた．また，設計したカーネルは，既存手法のように基底関数を考えることなく，化合物同士の類似度を算出することが可能であり，この点で既存手法との差別化を図れていると考えられる．

設計したカーネルは，現段階では元素情報の平均を用いるナイーブな手法の精度とあまり変わらない結果となっている．改善すべき点としては，現段階のカーネルは，二つの化合物間の類似度を，各化合物内の全近傍原子ユニットについて計算しているため，本来比較する必要のない部分を比較している可能性がある点が挙げられる．これによって，本来必要な情報より多くの情報を得ようとするため，精度にも悪影響を与えていると考えられる．そのため，どのペアが比較すべきでどのペアが比較すべきでないのかなどを考慮して優先的に比較するアルゴリズムを考案することで，精度の向上が可能であると考えられる．

参考文献

[Hansen 13] Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O. A., Tkatchenko, A., and Müller, K. R.: Assessment and validation of machine learning methods for predicting molecular atomization energies, *Journal of Chemical Theory and Computation*, Vol. 9, No. 8, pp. 3404–3419 (2013)

[Montavon 13] Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K. R., and Anatole von Lilienfeld, O.: Machine learning of molecular electronic properties in chemical compound space, *New Journal of Physics*, Vol. 15, (2013)

[Nakayama 15] Nakayama, K.: 機械学習に基づいた材料探索のための結晶構造記述, Master's thesis, 京都大学工学研究科材料工学専攻 (2015)

[Rupp 12] Rupp, M., Tkatchenko, A., Müller, K.-R., and Lilienfeld, von O. A.: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, *Physical Review Letters*, Vol. 108, No. 5, p. 58301 (2012)

[Schütt 14] Schütt, K. T., Glawe, H., Brockherde, F., Sanna, A., Müller, K. R., and Gross, E. K. U.: How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, *Physical Review B: Condensed Matter and Materials Physics*, Vol. 89, No. 20, pp. 1–5 (2014)

[Seko 15] Seko, A., Togo, A., Hayashi, H., Tsuda, K., Chaput, L., and Tanaka, I.: Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization, *Physical Review Letters*, Vol. 115, No. 20, pp. 1–5 (2015)