カーネル平均埋め込みによる材料の表現

草場　穫　　ものづくりデータ科学研究センター　特任研究員

1. 研究背景

機械学習で材料特性を予測するためには、モデルに与える材料の特徴表現が基本的な役割を果たす。モデルは、しばしば記述子と呼ばれる固定長の数値ベクトルとして表現される任意の与えられた材料系の関数として材料特性を記述する。しかしながら、材料を固定長のベクトルに符号化することは本質的に非自明である。なぜなら、材料は一般に様々な数の構成要素（原子、分子、化合物など）から構成されているからである。例えば、二元系化合物（e.g. SiO_2）と三元系化合物（e.g. Na_2SiO_3）では、元素の数が異なる。従来、このような多成分系を固定長の記述子ベクトルに変換するために、予め定義された成分特徴量の分布を、平均や分散などの少数の要約統計量にまとめることが行われてきた。この圧縮操作の欠点は、高次モーメントやマルチモーダリティといった成分特徴量の分布情報の一部が、ベクトル化の過程で失われてしまうことである。本研究では、機械学習理論であるカーネル平均埋め込みに基づいた、一般的な材料記述子のクラスを提案する。一般的な記述子とは異なり、カーネル平均埋め込みは、ベクトル化処理において、成分特徴量の分布に関するすべての情報を保持することができる。さらに、カーネル平均記述子は、元の材料空間への逆写像を一意に決定できることが保証されている。本研究では、無機化合物の形成エネルギー予測、準結晶材料を形成するための化学組成の予測、高分子材料の特性評価における力場パラメータの使用、以上3種の用途においてカーネル平均記述子の表現力と汎用性を実証した。

2. 提案手法

前項で説明したように、一般に材料は混合物であり、各特徴量ごとの構成要素の分布として情報が与えられている。よって、材料記述子の生成は確率分布の固定長ベクトル化問題であると見なすことができ、提案する手法ではカーネル平均埋め込みによるベクトル化を行う。カーネル平均埋め込みとは、任意の与えられた確率分布 $P_X(\lambda)$ を、カーネル関数 k で定義される再生核ヒルベルト空間 $\mathcal{H}_k$ 上の点 $m_X(\cdot)$ に写像する手法である。ここで k が特性的なカーネルである場合、$m_X(\cdot) = mY(\cdot) \Leftrightarrow P_X(\cdot) = P_Y(\cdot)$ が成り立つため、この埋め込み操作による情報損失は起きないと言える。本研究では、特性的なカーネルであるガウジアンカーネルを k として採用した。$m_X(\cdot)$ は無限次元のベクトルなので、特殊なケースを除いて、$m_X(\cdot)$ は特徴量空間 λ 上で均等に配置された格子点で離散化される必要がある。図1に、ある特徴量空間上でのカーネル平均埋め込みと要約特徴量をそれぞれ用いた記述子生成の模式図を示す。このように各特徴量ごとに計算された記述子を、全ての特徴量で結合することで最終的な材料記述子を得る。行列の要素が$G^k_{i,j} = k\left(\lambda^k_i, g^k_j\right)$ である $N \times d$ 行列 G^k を用いて k 番目の特徴量 λ^k に対するカーネル平均記述子は$\phi_{\lambda^k} = {G^k}^{\mathrm{T}} w$ と示すことができ、最終的な記述子 $\phi(X)$ は式 (1) のように書ける。任意の記述子ϕ^* から元の材料空間へ逆翻訳作業は式 (2) のように定式化できる。式 (2) は式 (3) の二次計画問題の形式に変形することができる。ここで行列 $H^{\mathrm{T}}H$ がフルランクであれば目的関数は狭義に凸になるので、一意的な最適解の存在が保障される。格子点数である d はユーザーが調整可能な値であるため、$rank(H) = N$ を満たすまで d を増やすことで、行列 $H^{\mathrm{T}}H$ がフルランクになるように制御することができる。本手法のコードは既に公開されており（https://github.com/Minoru938/KmdPlus）、材料科学に限らず任意の混合物系でカーネル平均記述子生成や逆翻訳を実行することができる。

$$\phi(X) = \begin{pmatrix} \phi_{\lambda^1}(X) \\ \vdots \\ \phi_{\lambda^K}(X) \end{pmatrix} = \begin{pmatrix} {G^1}^{\top} \\ \vdots \\ {G^K}^{\top} \end{pmatrix} w = Hw. \quad (1)$$

$$\min_w \|\phi^* - Hw\|^2 \quad \text{s.t.} \quad \mathbf{1}^{\top} w = 1, \; w \geq \mathbf{0} \quad (2)$$

$$\min_w \frac{1}{2} w^{\top} H^{\top} H w - {\phi^*}^{\top} H w \quad \text{s.t.} \quad \mathbf{1}^{\top} w = 1, \; w \geq \mathbf{0}. \quad (3)$$

当日ポスター発表

図1: カーネル平均埋め込み（左）と要約統計量（右）を用いた記述子生成の概要

3. 実験結果

図2に化学組成に対するカーネル平均記述子の生成例を示す。図2(a)は、形成エネルギーが小さい順に並べられた35,463化合物のカーネル平均記述子のヒートマップである。視覚化を容易にするため、ここでは6つの特徴量のみをプロットしている。図2(b)は、35,463化合物のカーネル平均記述子（左）と要約統計量記述子（右）を、第1主成分軸と第2主成分軸に投影したものである。化合物は、形成エネルギーの大きさによって色分けされている。カーネル平均記述子が従来の要約統計量記述子に比べて高い表現力を持つことを実証するために、第1項で記述した3種の用途においてそれぞれの記述子を使用した際の機械学習モデルの予測精度を比較した。その結果ほとんどの場合においてカーネル平均記述子が優れた性能を示した（詳細は図3を参照）。上記のように、カーネル平均記述子は高い表現力、一意的な逆翻訳可能性といった材料記述子として優れた性質を持つことが確かめられた。本研究は既に論文化しており、近日中に投稿予定である。

当日ポスター発表

図2: 化学組成に対するカーネル平均記述子の生成例

当日ポスター発表

図3: (a) 無機化合物の形成エネルギー予測、(b) 準結晶材料を形成するための化学組成の予測、(c) 高分子材料の特性評価における力場パラメータの使用 におけるカーネル平均記述子と要約統計量記述子の予測精度の比較