材料科学データに対する機械学習手法の応用

京都大学工学研究科材料工学専攻　世古 敦人

近年，材料科学の諸問題に機械学習手法を応用することで，高精度かつ効率的な材料探索が可能になってきている．このような材料探索においては，記述子（物性予測モデル構築のための説明変数）が予測精度の大部分を決定する．本研究では，機械学習による物性予測モデルに適した一般的な記述子を考案した．また，幾つかの物性データに対して考案した記述子を応用し，記述子の予測精度を評価した．考案した記述子をもとに新規化合物発見の指針を得る方法を開発した．

結晶は単位胞および単位胞中の原子により表現されるが，異なる化合物に対しては単位胞中の原子数が異なる．そのため，単位胞中の原子数によらない記述子が望ましい．本研究では，まず化合物中の原子を，その原子種や配位環境を表現する量へと変換し，化合物をその分布として考えた．その後，この分布の重心や分散共分散行列など代表的な量を記述子として採用した．元素を表現する量としては，電気陰性度など 21 種の物理量，結晶構造を表現する量としては，三角関数など 40 種を用いた．機械学習手法としては，物性予測にはカーネルリッジ回帰，新規化合物発見の予測モデルについては，ランダムフォレスト分類器などを用いた．物性予測モデル構築のためのデータとして，第一原理計算による凝集エネルギー（18096 化合物），格子熱伝導率（110 化合物），実験による融点（248 化合物）を用いた．新規化合物発見のための予測モデルについては，ICSD 収録の 33367 化合物を合成成功データとして，分類予測モデルを構築した．さらに約 130 万通りの化合物から，合成の可能性が高い化合物を予測した．

まず，凝集エネルギー予測モデルの構築を行った．予測モデルの精度は 0.042 eV/atom となり，広く用いられている動径分布関数を用いた場合（0.154 eV/atom）と比べると，大幅に精度が向上した．また，格子熱伝導率や融点についても，高精度なモデルが得られ，本研究の記述子が多くの材料科学データに対して有効であることが示された．さらに，新規化合物発見のための方法により，合成成功データに含まれない候補化合物から，合成可能な化合物を発見する効率について検証した．図には，候補化合物から選んだ構造の中に含まれる ICDD 収録の構造数を示す．ランダムフォレストによるモデルを使うことで，ランダムサンプリングに比べ，約 50 倍の合成可能な化合物を発見することができた．このように，本研究の方法により新規化合物の発見が大幅に加速されると期待される．

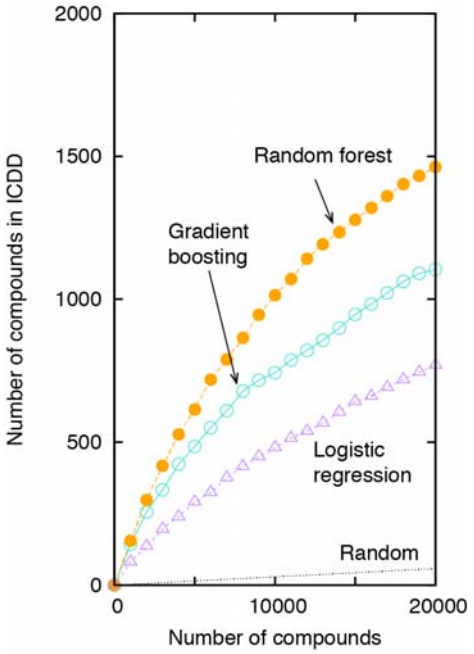

図　約 130 万通りの候補化合物から選択した構造の中に含まれる ICDD 収録の構造数．