マテリアルズインフォマティクス：スモールデータと転移学習

Materials Informatics: Small Data Problem and Transfer Learning

統計数理研究所（兼：物質・材料研究機構） 吉田 亮 [1,2]

The Institute of Statistical Mathematics [1], National Institute for Materials Science [2]

E-mail: yoshidar@ism.ac.jp

マテリアルズインフォマティクス（MI）の問題の多くは，順問題と逆問題の形式に帰着する [1,2]．順問題の目的は，系の入力Sに対する出力Yの予測である．物性予測の文脈では，入力は物質，出力は物性値に相当する．これに対し，逆問題では文字通り逆方向の予測を行う．すなわち，出力Yの値（例えば，目標物性）を設定した上で，それを達成する入力Sの状態（構造）を予測する．データ科学の観点において，これらの計算は，物質構造の“表現・学習・生成”を行うことに相当する．記述子と呼ばれる特徴ベクトルを用いて物質の構造を“表現”し，データのパターンから構造から物性の数学的写像を“学習”する．さらに，計算で所望の物性値を有する物質を“生成”し，有望な候補物質を炙り出す．

機械学習の他の応用領域に比べると，材料研究のデータ数は圧倒的に少ない．データ科学が本格的に導入されてから間もないこともあり，データベースの整備は発展途上の段階にある．とりわけ，対象が最先端に近づくにつれて，スモールデータの傾向はより顕著になる． スモールデータに対する解決策として，転移学習と呼ばれるアプローチが有望視されている．ヒトの脳には，少ない経験でも合理的な予測を行えるメカニズムが備わっている．例えば，小さい頃からピアノを学んでいた人は，音楽に関する一般的な知識を修得しているため，他の楽器の演奏技術を比較的容易に取得できる．このような推論の過程を模倣したものが転移学習である．

本講演では，物質構造の表現・学習・生成というコンセプトに沿って MI の幾つかの要素技術を解説する．特に，高分子インフォマティクスでの適用事例 [3] を紹介しながら，ベイズ推論に基づく物質設計，転移学習に基づくスモールデータ解析技術，ディープラーニングによる物質の表現等の話題を取りあげる．

References

1. Ikebata et al., Bayesian molecular design with a chemical language model, Journal of Computer-Aided Molecular Design, 31(4):379–391 (2017).
2. XenonPy: http://xenonpy.readthedocs.io/en/late
3. Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, npj Computational Materials, 5:5 (2019).