状態密度に基づく特徴量合成アプローチの信頼度評価

東大理 [A], 学術振興会 [B], 筑波大シス情 [C], NIMS[D], 東大新領域 [E]
大日方孝輝 [AB], 五十嵐康彦 [C], 永田賢二 [D], 袖山慶太郎 [D], 岡田真人 [E]

Confidence evaluation for feature selection in composed feature space based on density of states

[A]*Grad. School of Science, Univ. Tokyo,* [B]*JSPS,* [C]*Faculty of EIS, Univ. of Tsukuba,*
[D]*NIMS,* [E]*Grad. School of FS, Univ. Tokyo*
K. Obinata[AB], Y. Igarashi[C], K. Nagata[D], K. Sodeyama[D], M. Okada[E]

本講演では合成された特徴量空間における特徴量選択の信頼度評価手法を取り扱う．少数の特徴量に非線形な操作を繰り返し適用することで高次元空間を合成し，その中から特徴量選択によって低次元モデルを構築する研究が材料科学分野で進められている ([1] Ghiringhelli et al., 2015). 合成された高次元特徴量空間ではモデル候補の冗長性が高いため，得られたモデルから物理学的な知見の考察を行う上で，特徴量選択結果の信頼度を評価する必要がある．そこで本研究では評価指標に関する状態密度推定を行うことで，選ばれたモデルの有意性を評価する．モデル候補の数が多いため，交換モンテカルロ法によるサンプリングとマルチヒストグラム法を用いることで状態密度推定を行う．

先行研究 [1] で用いられたデータセットに対して推定した状態密度が図 1 である．先行研究で得られた最良モデルについて，推定される精度の揺らぎの範囲内（図の青領域）におよそ 1.8×10^4 個のモデルが存在しており，有意に優れているとは言えないことが明らかとなった．また，先行研究で行われた LASSO による絞り込み後の特徴量空間に対する状態密度（薄灰色）を調べることで，探索空間の縮小が悪影響を及ぼしていることが明らかとなった.

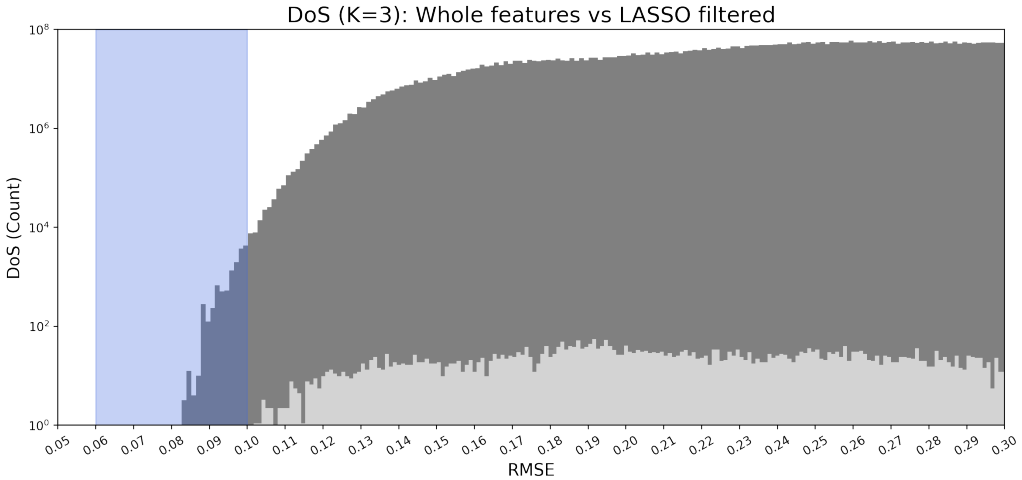

図 1　平均二乗誤差の平方根 (RMSE) を評価指標として用いた場合における 3 次元モデルの状態密度 (DoS) プロット．縦軸は対数スケールとしている．濃灰色は全特徴量を対象とした場合の DoS であり，薄灰色は LASSO により選択された特徴量のみを対象とした場合の DoS である．青い領域は先行研究で報告された最良モデルについて，クロスバリデーションから推定される RMSE の 1σ の範囲を表す.