

# SHAP 主成分分析を用いた目的変数指向化合物マップの作成

## A Target Oriented Compound Map Using SHAP-PCA

物材機構 ○小山 幸典

NIMS °Yukinori Koyama

E-mail: KOYAMA.Yukinori@nims.go.jp

化合物群を俯瞰的に捉えるための試みの一つとして、化合物マップの作成が挙げられる。AB 化合物の構造マップのような限定的な化合物群に対する化合物マップの作成は古くから行われているが、より広範な化合物群に対する化合物マップ作成の試みは限られている。マテリアルズ・インフォマティクスで使用されている化合物の特徴量を次元圧縮して化合物マップとすることは可能であるが、何らかの目的変数との関係に着目した場合に、特徴量を次元圧縮して得られる化合物マップが有用であるとは限らない。様々な化合物群に対して、目的変数との関係を考慮した化合物マップの作成方法を構築することが必要である。

本発表では、典型元素酸化物のバンドギャップの DFT 計算値を例とし、SHAP<sup>[1]</sup>と主成分分析を用いた化合物マップ(以下、SHAP-PCA マップ)を提案する。SHAP は機械学習モデルの予測結果を解釈する手法の一つであり、予測値を各特徴量の寄与に分割する。SHAP により、特徴量を目的変数に指向して変換したと考えることができる。得られた特徴量の寄与を主成分分析することで、その傾向を少数の主成分で表現することができる。データは DFT 計算データベースである Materials Project から取得した。安定相で、バンドギャップが 0.5 eV 以上の約 2,800 件の化合物を用いた。化合物の特徴量には XenonPy の組成記述子を用い、データ全体で値がほぼ一定である特徴量は除外した。SHAP で使用する機械学習モデルにはランダムフォレスト回帰を用いた。

第 1 主成分と第 2 主成分を用いた SHAP-PCA マップを Fig. 1 に示す。マップの各点は化合物を表し、その色でバンドギャップを表している。SHAP を用いたことにより、バンドギャップが大きな化合物ほど図の右上に位置するように図示されている。今回使用したデータでは第 1 主成分の寄与が非常に大きかった。SHAP-PCA マップから、第 1 主成分の大小で 2 つのクラスターに分けられることが見て取れる。第 1 主成分に対する寄与が最も大きな特徴量は「価電子数の分散」であったが、詳細な分析から、周期表で Zn より右下の元素を含むか否かで、2 つのクラスターのいずれに属するかを説明できることがわかった。

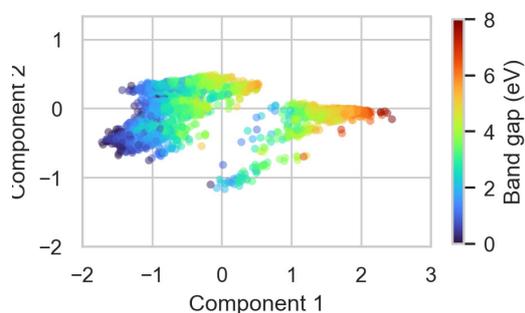


Figure 1. A band-gap oriented SHAP-PCA map of main-group oxides.

[1] S. M. Lundberg, S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", 31<sup>st</sup> Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.