**Nishika 材料の物性予測コンテスト　　　留野慎也**

**2023/5/22**分析期間(1/15~3/10)　開発環境:Google colab pro, Jupyter Lab

**背景**

材料開発において、理論計算による性質の予測や、現象のメカニズム解明は重要な役割を果

たしている。最もよく使われている手法の１つとして密度汎関数理論(DFT:density functional theory)に基づく計算が盛んにおこなわれているが、精確である一方で計算コストがかかる。

**目的**

機械学習を使い、DFT計算結果を高速で予測できれば材料開発のスピードアップやコストダウンが見込める。

**概要**

今回の分析では無機材料の第一原理計算の結果を公開しているデータベースである、Materials Projectからのトレーニング/テストデータを約８万/３万件使用した。

目的変数は物質の生成エネルギーであった。

生成エネルギーEfは、たとえば酸化鉄Fe2O3 の場合を考えると、

(<Fe2O3のエネルギー> -2 \* <Feのエネルギー＞ - 3/2 \* <O2のエネルギー＞) / 5

で表される。

生成エネルギーの値が分かると、物質の安定性や反応性を議論、予測することができるという点で役に立つ。説明変数としては物質の結晶構造における原子の三次元座標等の情報を使った。

手法としては2021年に発表されたモデルGeo-CGNNと著者のページに公開されたコード[1]を使った。

**結果と評価**

RMSEは0.084程度だった。この予測精度がどれほど良いものかについて知るため、実際のDFT計算で生じる誤差について調べた。

文献[3]を引用している文献[2]によると、生成エネルギーのシミュレーションの誤差(MAE)が0.081~0.134eVとの記述があった。

しかし、文献で使われていたMAEと自分の分析で使用したRMSEは指標として異なるので単純比較できないので自分の分析の精度の良しあしを判断できない(コンペのテストデータは正解データが非公開なためMAEの算出が不可能)。

**コンテストへの取り組み**

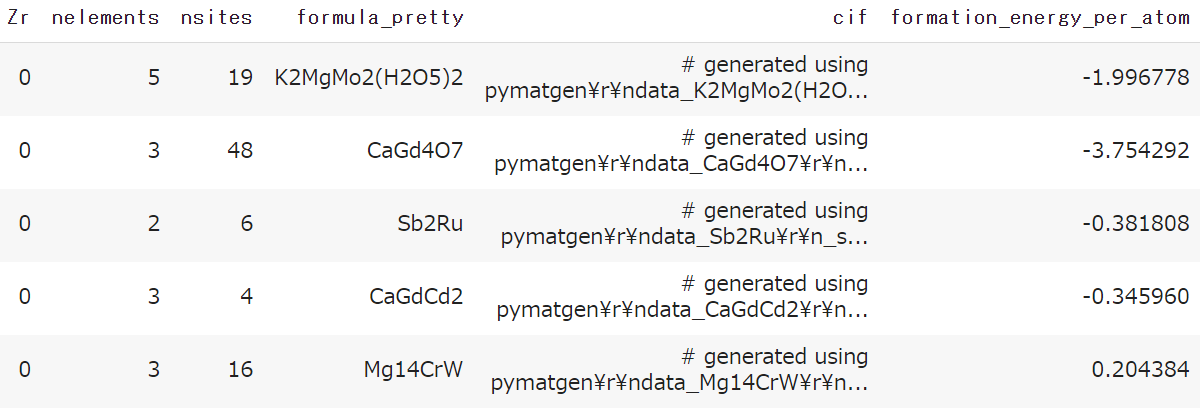


図１　訓練データのテーブル

組成式中の原子の数、結晶中の原子の個数、組成式、結晶構造情報(cif)

生成エネルギー(目的変数)などが格納されている。

上の図1のようなテーブルデータをつかって分析した。

当初は2017年に発表された結晶構造をグラフ表現した深層学習モデルCGCNN[4]をベースにしたnishikaのチュートリアルコード[5]で分析していた。しかし、損失関数をRMSE(平均二乗誤差)で評価した誤差は、畳み込み層の数や活性化関数、ハイパーパラメータをどんなに変えても出力された生成エネルギーが-1.5であり、損失関数が1.21から全く学習が収束せず、全く予測が進まず予測精度が非常に悪かった。

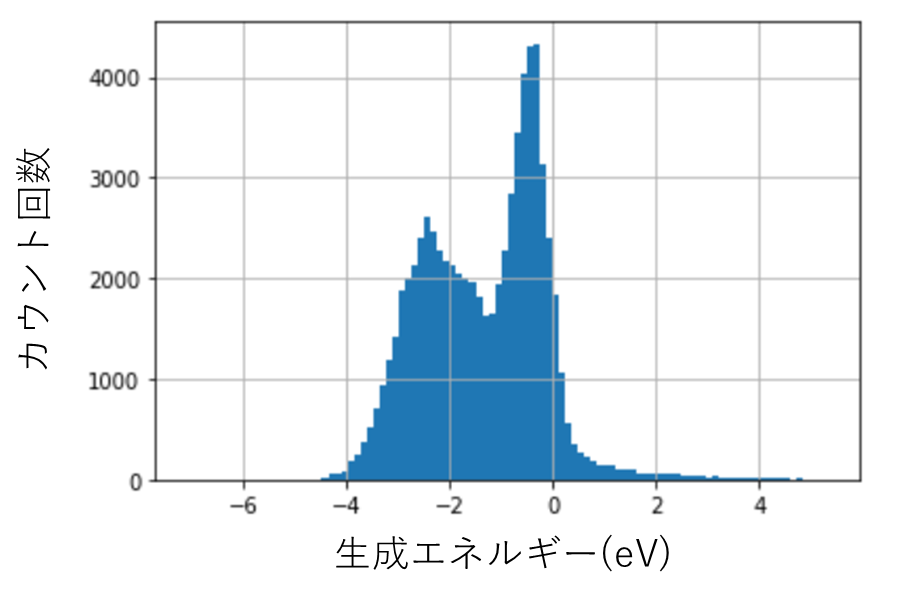


図２ 目的変数の分布

底軸は目的変数である生成エネルギー、縦軸が物質のカウント数。

赤の矢印は-1.5の位置を示している(自分のCGCNNの分析でほぼすべての物質から出力された値)。

目的変数の分布を図示すると上の図２になる、図2の全データの生成エネルギーの分布をみたときに、鋭い山が２つあるが、出力結果で良く得られた赤矢印の-1.5という値はその二つの山の間にある値であり、物質個々の特徴を無視し均一な値を出力していることがわかった。

学習が収束しない原因として考えられることの一つに、CGCNNは結晶構造をグラフ表現する際に、多くの情報をそぎ落としており、結果として個々の材料の結晶構造の特徴をとらえきれていないというのが考えられる。

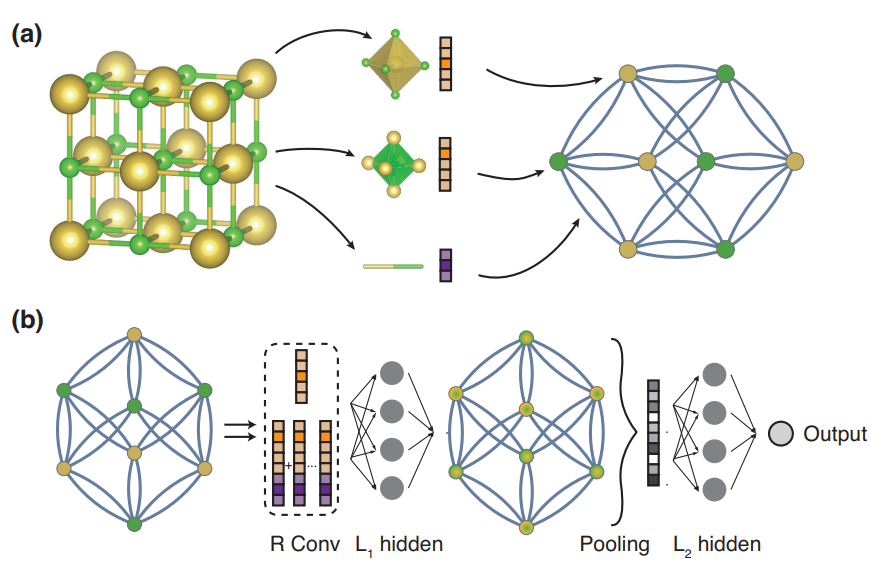


図2 CGCNNの構造[4]

(a)NaClの結晶構造のグラフ表現

緑のノードがNa,黄色のノードがClに対応している。エッジ特徴量は隣接している原子種とその距離。

(b)畳み込みグラフニューラルネットワークの構造

数回の畳み込み後、隠れ層を経て個別の結晶毎にプーリングされ、再度隠れ層を通し出力を得る。

実際に図3のように、CGCNNのノード特徴量は、結晶内の原子の種類をOne-Hotエンコーディングしたもので、エッジ特徴量は着目した原子の隣接原子と距離である。よって、結晶構造の三次元的な情報はないといえる。

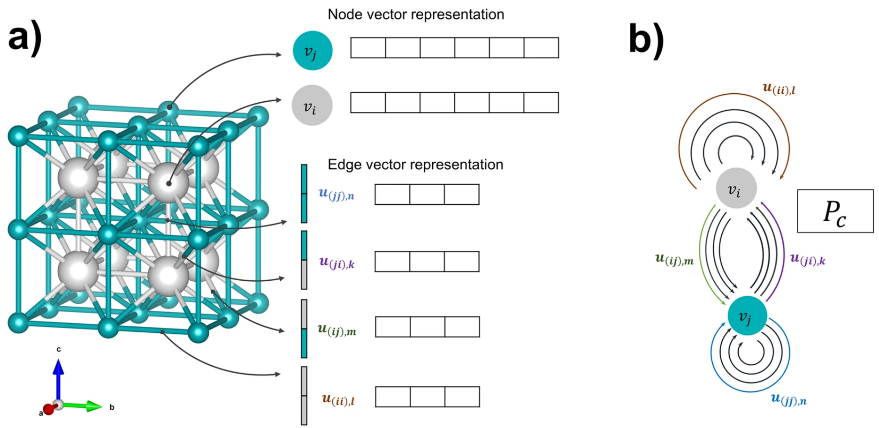


図４Geo CGNNでのNaClの結晶構造グラフ表現[1]

(a) NaClの結晶構造。ここでvi,vjはそれぞれノード特徴量で、それぞれの原子の原子番号をOne-hot encodingしたもの。uijはエッジ特徴量で原子ijの結晶構造内での三次元位置座標ベクトルの差。

(b)NaClのグラフ表現。Pcはグラフ全体の特徴量Pc={a,b,c,Ω}であり、結晶の単位格子の三辺a,b,cと体積Ω。

一方で、2021年に発表されたGeoCGNNは、図4のように原子の位置座標ベクトル(3次元)を入力に使っている点でCGCNNよりも多くの情報を使って予測ができる(他にも改良されている部分はある)ことが期待できるため、この手法で分析した。

実際に、誤差が0.08程度まで一気に下がり、CGCNNでの解析から誤差が1/15になり、より正確な予測ができるようになった。

**工夫したこと**

より精密に予測できるモデルを文献調査により見つけ、精度を向上させた。

**問題点と改善策**

訓練データに過学習しすぎた。

コンテスト締め切り２日前にGeoCGNNを知ったので、過学習しないラインを見極めて学習のエポック数を定めたりすることができず、訓練時に比べテストデータの推論時で大幅に精度を落とした。次はその点に注意して作業を行う。

**参考文献**

[1]Cheng, J., Zhang, C. & Dong, L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun Mater* **2**, 92 (2021).

著者のgithubページに公開されているコード

<https://github.com/Tinystormjojo/geo-CGNN>

[2] Kirklin, S., Saal, J., Meredig, B. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput Mater* **1**, 15010 (2015).

[3] 鶴田 博文, 桂 ゆかり,熊谷 将也

3次元メッシュで表現した結晶構造を用いた 材料物性の予測に向けた深層学習モデルの設計 2022年人工知能学会全国大会(第３６回)

[4]Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).

[5]nishikaの公開したチュートリアルコード

https://competition.nishika.com/competitions/bussei/topics/427