Prediction of Perovskites' Water-splitting Ability

Jimin Qian Xueqiao Zhang Yiming Sui

https://github.com/JiminQ/Perovskite

Abstract

Our study aims at predicting whether a given perovskite has water-splitting ability. Since materials' water-splitting ability is mainly determined by its band structure and heat of formation, our main task is to predict perovskite's conduction/valence band energy and heat of formation based on atom's easy accessible parameters like electronegativity, ionization energy, etc. Two methods are applied for predictions (i) kernel ridge regression (ii) neural network. We compare the predictions made by the two models and analyze the outliers. At last, a Graphical User Interface is setup which can read in a perovskite user input and return: (i) predicted values of the perovskite (ii) perovskite's water splitting ability.

Background

Searching for a clean energy is currently an urgent issue due to the over-consumption of the fossil fuels and the relevant environmental issue. One possible solution is to produce hydrogen and oxygen as clean fuels by water-splitting reaction. However, the suitable material that enables this reaction to happen should meet several strict requirements on its bandgap structure and heat of formation. Precise calculation of these parameters becomes the prerequisite of finding materials with water-splitting ability. Since conventional quantum mechanical computations(i.e. density functional theory) for properties like bandgaps are enormously computation-time intensive and thus impractical in high throughput studies, machine learning approaches can be a promising alternative. Here we demonstrate a systematic feature-based learning and predicting framework to predict perovskite's water splitting ability.

Reference

1.Pilania G, Mannodikanakkithodi A, Uberuaga B P, et al. Machine learning bandgaps of double perovskites[J]. Scientific Reports, 2016, 6.
2.Computational Materials Repository, https:// wiki.fysik.dtu.dk/cmr/ (Documentation) and https:// cmr. fysik. dtu.dk/(Database).

Method

Database of Perovskite:

Features: anion/cation, electronegativity, ionization energy, atom affinity, electron orbital radii,volume,mass, etc
Prediction: conduction/valence band energy, bandgap and heat of formation

Kernel Ridge Regression (KRR)

Features selection:
We select the least correlated 15 features combination from the 27 features we have. The 27 features are either directly from the database or derived from the linear combination of the original features. We find that more features cannot significantly reduce the mean squared error of prediction and less features may be weak in describing the data.

Kernel ridge regression:

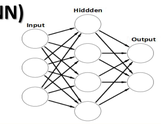
$$\min_{\alpha} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \cdot \sum_{i,j} \alpha_i k(x_i, x_j) \alpha_j$$

$$with \quad f(x) = \sum_{i=1}^{n} \alpha_i \underbrace{k(x_i, x)}_{\text{Kernel function}}$$

We try several kernel functions: polynomial function, Laplacian, radial basis function, sigmoid function. It's found that the regression model with polynomial kernel function has the best fitting result.

Neural Network (NN)

Figure schematic diagram for neural network:

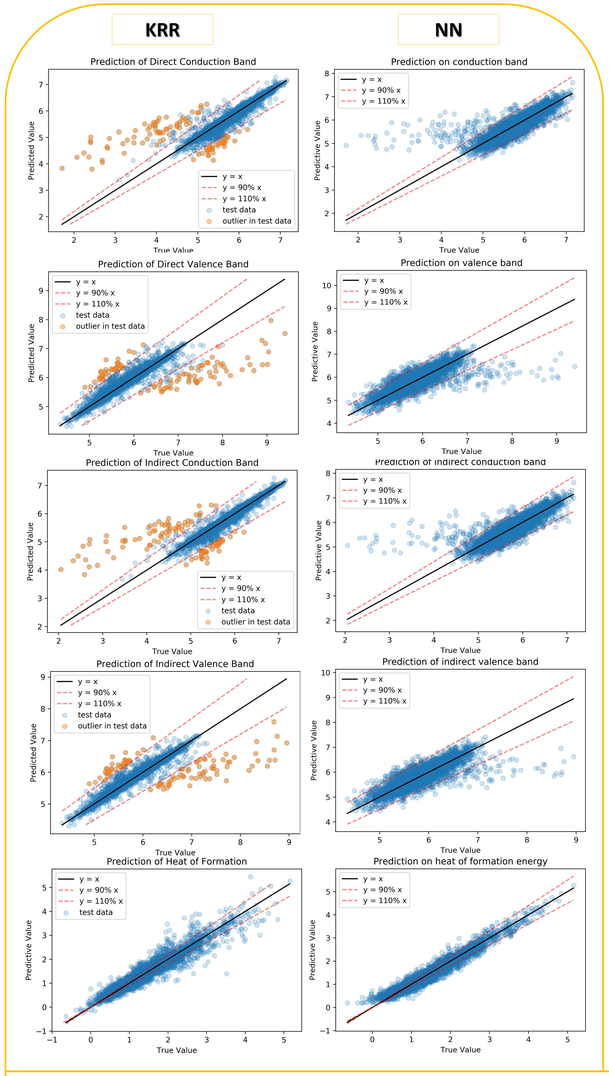


Model compilation:
Here we specify the loss function 'mean_squared_error' to be used to evaluate a set of weights and we use the optimizer 'adam' to search through different weights for the network due to its great efficiency.

Fit model:
The training process will run for a fixed number of iterations through the dataset called epochs. We try many epochs value to get the best overall performance. Too small amount of iterations will result in high loss value. But too many iterations will need more time on running iterations. Batch size corresponds to the number of instances that are evaluated before a weight update in the network is performed. Hence, we set a relative large value to involve as many instances as possible.

Evaluate and predictions:
The testing data will be input into the trained model and evaluated by calculating loss by comparing predicted y with the actual y.

Results and Discussion



		Direct Valence Band	Direct Conduction Band	Indirect Valence Band	Indirect Conduction Band	Heat of Formation	Direct Bandgap	Indirect Bandgap
mse of KRR	training	0.052	0.052	0.039	0.039	0.057	0.205	0.352
	testing	0.068	0.068	0.051	0.050	0.067	0.566	0.411
mse of NN	training	0.1	0.1	0.08	0.09	0.04	-	-
	testing	0.11	0.12	0.09	0.1	0.04	-	-

1. Both regression model and neural network model have 1. low mse(except the prediction of bandgap). The difference between mse of testing data and training data is only ~0.01, indicating there is no significant overfitting problem. The relatively large error of bandgap is because the predicted value of bandgap in our model is actually calculated from the difference of the predicted value of conduction band and valence band, thus errors accumulate on the prediction of bandgap.

2. The prediction result of Direct Valence Band prediction, Direct Conduction Band, Indirect Valence Band, Indirect Conduction Band from KRR and NN model are very similar: the predicted value of conduction band energy that is lower than 4.5 is significantly higher than the true value, while the predicted value of valence band energy that is higher than 7 is signigicantly lower than the true value.

3. Both the prediction result of KRR and NN have ~3% outliers (outlier is defined as the predicted value that is less than 90% or higher than 110% of the true value). From the prediction result of KRR, it is found that the data shown as outlier points in the prediction of Direct Valence Band, Direct Conduction Band, Indirect Valence Band, Indirect Conduction Band are highly overlapped. The recurring outlier data in the four predictions are marked as orange circle in the KRR's plot.

4. We analyze the recurring outliers in KRR prediction result. Given that the amount of each kind of anion are the same in the database, we find that the amount of each anion in the outlier data shows significant difference.

Anion	O_3	O_2N	ON_2	O_2F	O_2S	N_3	OFN
KRR's outliers	34	12	8	30	8	0	5

This result may help us determine what kind of input data can get more precise prediction in our model. The prediction for perovskite with N_3 anion are expected to have less error than the prediction for perovskite with O_3, O_2F anion.

Package use guide:

GUI: input a new perovskite → **in our database?**
- No → **predict its valence band, conduction band and formation energy** → **return wthether it can do water-splitting**
- Yes → **return whether it can do water-splitting**

DIRECT
Data Intensive Research Enabling Clean Technologies

CLEAN ENERGY INSTITUTE

MATERIALS SCIENCE & ENGINEERING
UNIVERSITY of WASHINGTON