Transparent Conductor Prediction

Yan Sun, Yiyuan Xing, Xufan Xiong, Tianduo Hao

Abstract

Nowadays the development of innovative materials is one of the most challenges for physical application, which concerns about the development of heath equipment, new energy application and many other fields. In order to optimize the property of materials, it is crucial to get a deep understanding the relationship among properties, composition and internal energy condition. Specifically, transparent conductors are significant compounds that are electrically conductive and low absorption in the visible range, which is a special property of these conductors and could make them applied to sensors, transistors and laser equipment. However, one of the biggest problems is only a small portion of compounds is well understood that is able to be considered as transparent conductor. In order to find the optimum composition for transparent conductor, some basic principle should be the basic rule for computational approach. There exist one primary computational method for materials science named Density Functional Theory (DFT), which is able to get high accuracy result but requires much computing time even for supercomputers. In this way, the data-driven method will be an alternative way to improve the efficiency for the transparent conductor design process.

Motivation

Transparency
Wider Bandgap

- More photons with energies less than band gap value are not absorbed by these materials.
- Wider range of visible light passes through.

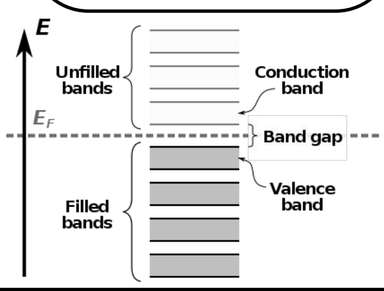
Competing Properties

Conductivity
Wider Bandgap

- More difficult to activate electron to conduction band.
- Less Conductivity.

Transparent Conductor

red	~ 625–740 nm	~ 1.77 eV
orange	~ 590–625 nm	~ 2.00 eV
yellow	~ 565–590 nm	~ 2.14 eV
green	~ 500–565 nm	~ 2.34 eV
cyan	~ 485–500 nm	~ 2.52 eV
blue	~ 450–485 nm	~ 2.64 eV
violet	~ 380–450 nm	~ 2.95 eV



Prediction through calculation:

- Density Functional Theory (DFT) calculation requires too much computing resource and time.
- Machine Learning could be alternative method for transparent conductor prediction.

Data Set

This is a data set containing 11 features of 3,000 different transparent conductor materials, including:

- Space group
- Total number in the unit cell
- Relative compositions of Al, Ga, In (3 features)
- Lattice vectors (3 features)
- Lattice angles (3 features)
- Coordinate information of each atom in each data sample (include basic vectors)

The purpose is to predict the 2 target properties of these materials:

- Formation energy - an important indicator of the stability of a material
- Bandgap energy - an important property for optoelectronic applications

Methods

1 Data preprocessing

- No Normalization.
- Drop one feature(relative component of In) to maintain the independence of features.
- Extract atom coordinates and apply principal component analysis (PCA).

2 Models

2.1 Linear Regression

- Use linear model to fit the dataset
- Simplest & fastest model
- k-fold cross-validation: k = 10

2.2 Neural Network

- 4 layers: Input -> 1024 -> 512 -> 64 -> 2 -> Output
- 6 layers: Input -> 1024 -> 512 -> 256 -> 128 -> 2 -> Output
- Higher Accuracy, Adam Optimization, Mini-Batch, RMLSE

2.3 Tree-based Model

Random Forest

- Combining several independent regression trees, less variance and reduction in overfitting.
- Applying k-fold cross validation k=15 Number of trees: 500 Maximum depth: 9

AdaBoost

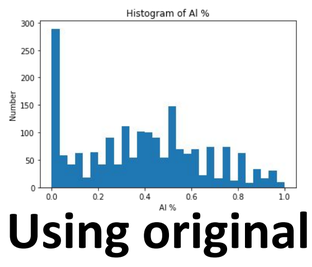
- Train the weak learner based on previous prediction error.
- Combined all weak learner into one strong learner.
- The outputs of weak learners are combined into a weighted sum that represents the final output of strong learner.
- k-fold validation: k=15 Estimator: regression tree Max depth: 3 Number of estimator: 50

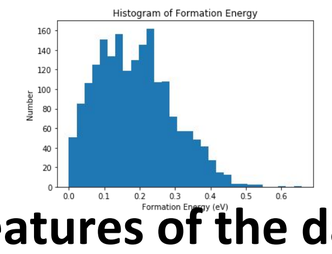
Gradient Boost

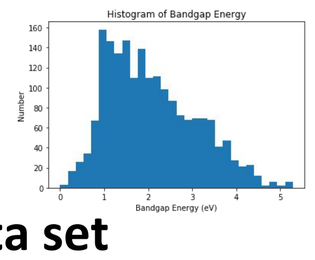
- Train the initial weak learner and get the residual error.
- Train the next weak learner aimed at fitting the residual error from previous learner.
- Ensemble all weak learner into one strong learner.
- k-fold validation: k=15 Estimator: regression tree Maximum depth: 3 Number of estimator: 90

Results

1 Dataset Distribution

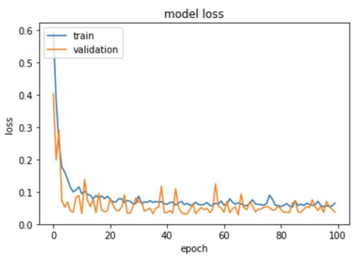






2 Using original features of the data set

2.1 Loss vs Epoch Curve (Before Optimization)



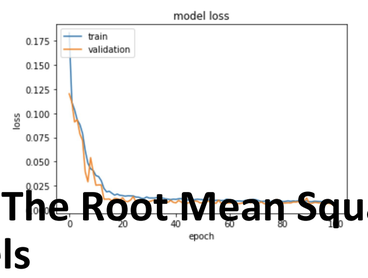
2.2 Comparison of The Root Mean Squared Logarithmic Error (RMSLE) among 5 models

Index	Model	Type	Formation Energy RMSLE	Band gap Energy RMSLE
1	Linear Regression	Linear	0.066825	0.178515
2	Artificial Neural Network	Neural Network	0.069342	0.104174
3	Adaboost + Regression Tree	Tree	0.046249	0.129517
4	Gradient Boost Tree	Tree	0.033199	0.101471
5	Random Forest	Tree	0.032237	0.090953

3 Advanced Optimization

- Combining existing features to create new ones (i.e. Volume, Density)
- Build deeper ANN architecture.
- Extract particle coordinates for all data samples.
- Split the coordinates by element (Ga, Al, In, O).
- Use PCA to get 2-dim feature (explain 99% ratio of variance) for x,y,z per element.
- Get 2(number of components after PCA) * 3(directions) * 4(elements) = 24 new features, indicating element distribution.

3.1 Loss vs Epoch Curve (After Optimization)



3.2 Comparison of The Root Mean Squared Logarithmic Error among the 5 models

Index	Model	Type	Formation Energy RMSLE	Band gap Energy RMSLE
1	Linear Regression	Linear	0.064418	0.171237
2	Artificial Neural Network	Neural Network	0.054654	0.094746
3	Adaboost + Regression Tree	Tree	0.046249	0.123980
4	Gradient Boost Tree	Tree	0.033199	0.092334
5	Random Forest	Tree	0.032237	0.090953

Conclusions

- The performance of tree-based models are better no matter after simply dropping out one feature and feeding the remaining 10 independent features into the 5 models or taking measures to further optimize the model. And the random forest model is always the one with lowest error
- Among the 5 models, the errors for predicting formation energy are much lower than that for predicting band gap energy.
- After optimization (described in the ‘result’ section), the error of the models slightly decreases for predicting the bandgap energy.