Project: Capstone Project 1: Project Proposal

Predicting the key properties of novel transparent semiconductors

Meng Zhao

Machine learning and data analytics can accelerate materials innovation. Among the various categories of the materials, transparent conductors are an important class of compounds that are both electrically conductive and having a low absorption in the visible range.Transparent conductors have a wide variety of applications including solar cells, flat-panel displays, touch-screens, and energy-conserving windows due to their ability to reflect thermal infrared heat. Currently, only a limited number of compounds meet the criteria of high conductivity and low visible light adsorption displayed by transparent conductors. Conventional high-throughput theoretical calculations and experiments could provide guidelines for identifying new transparent conductors. However, they are computationally expensive and time-consuming, due to a large materials searching space generated by the numerous possible compositions and configurations. Therefore, building data-driven models to accurately predict materials stabilities and light adsorption properties becomes an alternative and helps with an efficient search on promising transparent conductors.

The structural-properties relationships and any promising transparent conductor candidates unrevealed from the model are valuable to the companies manufacturing electronic devices. They can use these information as a starting point/shortcut to conduct further advanced experiments, performance tests and manufacturing scale-up , which tremendously reduces their time of trial-and-error.

There are two data sources I will be using for this project: one is provided by kaggle competitions:Nomad 2018 Predicting Transparent Conductors(Link: https://www.kaggle.com/c/nomad2018-predict-transparent-conductors). This datasets of 3,000 promising transparent conductors are relatively clean which contains materials structural information(spacegroup, total number of atoms, compositions, lattice vectors and angles), and target properties of formation energies and bandgaps. Another datasets to be incorporated is from Materials Projects(a open-source materials library)(https://materialsproject.org/). I am going to acquire the relevant data(structural information, and target properties of formation energies and bandgaps) of ~300 materials using their API.

The goal of this project is to predict two chemical properties: formation energy and bandgap. Therefore, two regression models are probably needed as two targets are independent. The number of features describing a single chemical system is far more

than the size of the datasets. I am going to concentrate on feature engineering to avoid overfitting and try various regression algorithms mostly within scikit-learn package for the best predicting performance. The prediction accuracy is evaluated by root mean squared logarithmic error(RMSLE) and the equation is shown below:

$$\sqrt{(1/n)\sum_{i=1}^{n}(log(pi+1)-log(ai+1))^2}$$

Where n is the total number of observations, pi is the predicted value, ai is the actual value.

My deliverables would be a report, reproducible code on Github and client-facing slides summarizing the report.