Data Wrangling Report

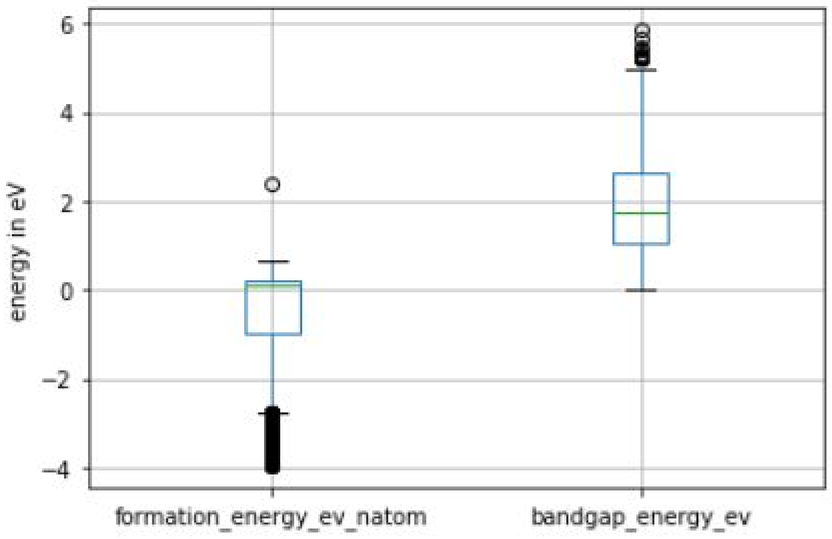
Meng Zhao

The goal of this capstone project is to use materials structural information to predict their formation energies and bandgaps. Data is acquired from two sources: one is from the Kaggle repository and data is collected by downloading the train.csv file which contains a set of materials with target values(the bandgap and formation energies). Train.zip is also downloaded and it contains geometry.xyz files saved under the directory of {train}/{id}/, storing spatial information about the material. When loading the train.csv file into a pandas dataframe, we have confirmed that this pandas dataframe has 2400 entries and 14 columns in total without missing values. Each entry represents a material. Within these 14 columns, 12 of them are materials features named as 'id', 'spacegroup', 'number_of_total_atoms', 'percent_atom_al', 'percent_atom_ga', 'percent_atom_in', 'lattice_vector_1_ang','lattice_vector_2_ang', 'lattice_vector_3_ang', 'lattice_angle_alpha_degree', 'lattice_angle_beta_degree', and 'lattice_angle_gamma_degree'. The rest of two columns, 'formation_energy_ev_natom', and 'bandgap_energy_ev', are the target values. Although the Cartesian coordinates (positions) for each atom and a list of elements in a chemical system can be derived from these 12 materials features, a more efficient way of retrieving comprehensive structural information is to transform the geometry.xyz files into pymatgen.core.structure.Structure objects. The structural information stored in the pymatgen.core.structure.Structure object can further be used for generating structure-based fingerprints to train the machine learning model. The processed Kaggle dataset contains three features of ‘formula’, ‘structure’, ‘space group’ and two target values of 'formation_energy_ev_natom', and 'bandgap_energy_ev'.

The Kaggle repository has a limited number of oxides containing various combinations of Al, Ga and In. To extend the space of the chemical systems, second data source from Materials Project database is employed. Materials Project database is an open-source materials genome database containing both experimental and theoretical information of materials throughout the periodic table. Instead of collecting a massive datasets of at least ~100K inorganic materials, a subset of 829 entries representing oxides of 'Al', 'Ga', 'In', 'Mo' ,'Zr' ,'W', 'Ta', 'Sb', 'Zn', 'Sn', 'Ti', and 'Ce' are collected using Materials Project API called MPRester and requests package. The JSON strings of API response stores target values and structural information in a string with CIF format. Crystallographic Information File (CIF) is the internationally agreed standard file format for information exchange in crystallography. Similar to the way of processing the Kaggle dataset, lattice vector, coordinates and a list of elements, are extracted. The data extraction is realized by the function called read_pymatgen_cif(). Using lattice vector, coordinates and a list of elements as inputs for read_pymatgen_cif(), pymatgen.core.structure.Structure objects are obtained for each material entry. The processed Materials Project dataset has the same columns as that of the processed Kaggle dataset, which includes three features of ‘formula’, ‘structure’, ‘space group’ and two target values. The complete dataset is from the concatenation of Kaggle dataset and Materials Project dataset, in which 3229 entries are available without missing values.

For initial data exploration, a boxplot is generated for two targets values shown as Figure 1. The target value of 'formation_energy_ev_natom' varies from -4 to ~2.5 eV, and bandgap_energy_ev ranges from 0 to 6 eV. Although both have a significant amount of data lie in the outlier region, those data points may not actually be outliers as both targets have wide ranges of formation energies and band gaps due to the materials diversity. The data points lie in the outlier region are going to be carefully treated with caution when training machine learning model.

Figure 1. boxplot of formation_energy_ev_natom and bandgap_energy_ev


energy in eV
formation_energy_ev_natom
bandgap_energy_ev