Determination of lattice structure of perovskites using ML

Kaling Vikram Singh
kalingvikram.singh@niser.ac.in
NISER

S Sunil Raja
sunilraja.s@niser.ac.in
NISER

Abstract

The crystalline structure is an important notion that controls various physical properties of a matter and thus, it is important to correctly estimate the crystal structure through various techniques. The various techniques are time consuming and costly. Thus, one alternative to correctly estimate the crystal structure is using a machine learning model. Perovskites are a special class of elements with $AB0_3$ type configuration. Several properties such as valency, atomic radii, bang gap, etc are used to predict the crystal structure. We have used various classification techniques, after oversampling the data to get the required classification model. We have utilized SVM, Light BGM and XgBoost to get our results. It was found that LBGM gives the best accuracy of 94.32% followed by Xgboost with 94.13% and weighted SVM with 92.03% accuracy. This accuracy is much higher than the results noted in the reference paper.

1 Introduction

Perovskites are a class of elements that have similar crystal structure as the compound calcium titanium, ABX_3. These are used in various industries and its main applications include creating solar panels that could be coated on various surfaces.These materials are lightweight and cheap and are used in photovoltaic industry. These have high variations in A,B and can be found in various structures such as cubic, monoclinic, orhthorhombic, tetrahedral, hexagonal or rhombohedral. The most significant reasons for the change is shapes include (i) displacement of the cation, (ii) distortion of the octahedra and (iii) tilting of the octahedra. The displacement and distortions are instability driven factors.

Good oxide ion conductivity, which is necessary in fuel cell applications, is an important property of a cubic perovskite structure's reduced distortion. In a cubic perovskite, the 3D framework leads to corner sharing of BO_6 octahedra and the A-cation is enclosed within 12 equidistant atoms. The coordination number of O is 2 and is low since the A-O distance is almost 1.4 times the B-O bond distance.

There is extensive work in progress that include Density functional Theory (DFT) calculations and tough techniques such as X-ray diffraction (XRD) to get the the crystalline structure of materials. These are power intensive and costly processes which can be reduced if Machine learning (ML) models are used.

1.1 Related works

Santosh and Taher et al.[1], provided the basis for this work, although the data they used included only 675 data entries and the data was biased towards orthorhombic structures. The methods used include XgBoost, SVM, Light BGM, and Random Forest (RF) with accuracy of 74.8%, 76.6% ,80.3% and 62.8% respectively. The paper accounts for tolerance factor (τ), derived from radius of a,b, which has

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

been neglected in current paper and model is built with the available data-set. They reported the best accuracy for RBF (which is explained using the density of states function which is also RBF) kernel in SVM and Light GBM to give the best overall accuracy. Also, the sampled data was cut down to just the ones that have each and every feature value.

Jarin et al.[2] reported various models through which they have found the crystal structures without oversampling with accuracy of 95% using genetic algorithm support vector regression. They also utilized various Neural networks to achieve the best possible accuracy. They have not taken into account the tolerance factor (τ) but also removed some features that are of less importance according to the importance matrix.

2 Baselines

We have used simple yet fast and reliable models such as Light BGM, XgBoost, SVMs to obtain the crystal structures. Boosting algorithms work on decision trees in which sequential tree growth using gradient boosting improves performance by correcting categorization errors made by earlier trees. This is fast and effective methods. Further, SVMs are simple machines that work on principle of support vectors along with kernel functions. The corresponding kernel functions are used to get the best maximum accuracies. Further, classifiers are used to get the importance matrix through which we can see the corresponding importance of each features.

The main aim is to build a classification model that correctly classifies a perovskite. The workflow is done as follows:

(i) Database collection

(ii) Feature selection and data-processing

(iii) Model selection

(iv) Hyper-parameter optimization

(v) Testing for accuracy.

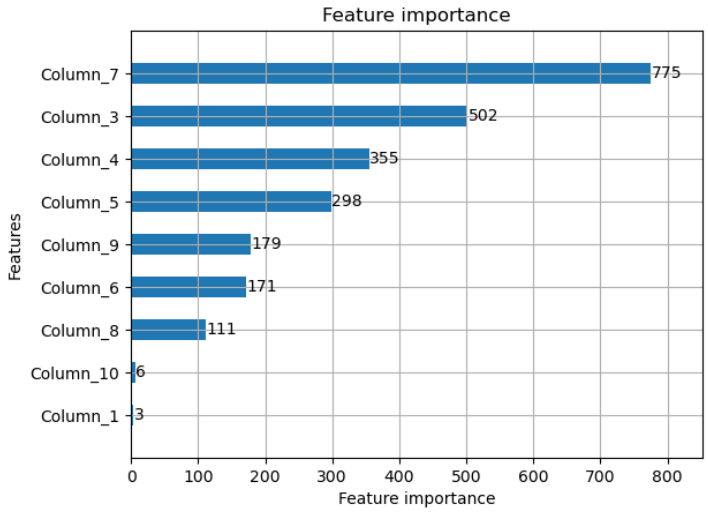

Figure 1: Feature importance

2.1 Model Environment

Python 3 was used for this project. Scikit-learn, Pandas, Imblearn, Numpy, etc librariers were used for data processing and implementation of different models such as XgBoost, SMOTE , SVM , and Light GBM.

2.2 Feature selection

Unlike others authors, we planned to give importance to each and every feature even if their importance was low. The lgb.classifier was used to get the relative feature importance (fig 1). The column 7 representing the ionic radii of A was most important feature. The correlation matrix was plotted using the data (fig 2).In case of weighted SVM, more importance was given to the instances which are found in nature and which have balanced atoms. Only, the compound names were omitted from the calculations as they did not show any importance. The bond angles and lattice edge length were omitted as they are precursors of crystalline structure.

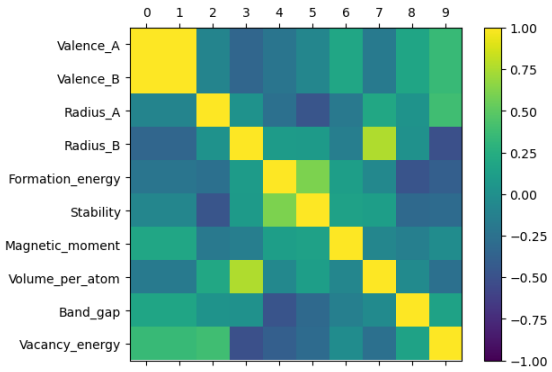

Figure 2: Correlation matrix

3 Experiment

GitHub repository: `https://github.com/Kalingvikramsingh/Machine-Learning`

3.1 Data-Pre processing

The data consists of 5329 datasets which include data that cannot be used for our model and thus, such rows were deleted. The data was obtained through DFT calculations and includes various chemical and physical properties of compounds. The dataset is highly imbalanced which can lead to overfitting of particular type of dataset. SMOTE was used to equalize the number of instances of each label. The features were selected beforehand. Since the values were not highly scattered, standardization and normalization techniques were not utilized. The data was divided in training and test before SMOTE to reduce bias in the data. After sampling, the new training data includes 5116 samples and test contains 528 samples.

We have used Boosting methods on decision trees that can be used to predict results on the test data. The boosting methods use sequential corrections in order to calculate the loss and get best possible predictions using decision tree. In case of SVM, we have utilized different kernels to study the best fit. Further, we used weighted SVM to calculate the predictions.

3.2 Light BGM

The parameters used were the height of the decision tree, learning rate. Validation dataset was prepared from the training dataset which was used to get the hyper-parameter values. The optimum height was found to be 7 and learning rate was 0.08. The losses were calculated on a logarithmic loss function. Finally, the hyper parameters were used to get the best accuracy of 94.32% using LBGM.

3.3 XgBoost

The parameters used were height of decision tree, learning rate, and number of epochs. After hyper parameter tuning, 20 epochs with 1.25 learning rate and tree height 2 was found to be the best fit. The loss was again logarithmic. The accuracy found through this method was 94.13%

3.4 SVM

The parameters used were C(Penalty parameter) and kernels(Linear,RBF,Sigmoid and Polynomial). After Hyper parameter tuning, RBF (fig 3) was chosen to be the kernel and the value of C as 209.5. Upon further tuning the value of gamma for RBF was chosen to be 0.01.

3.5 Weighted SVM

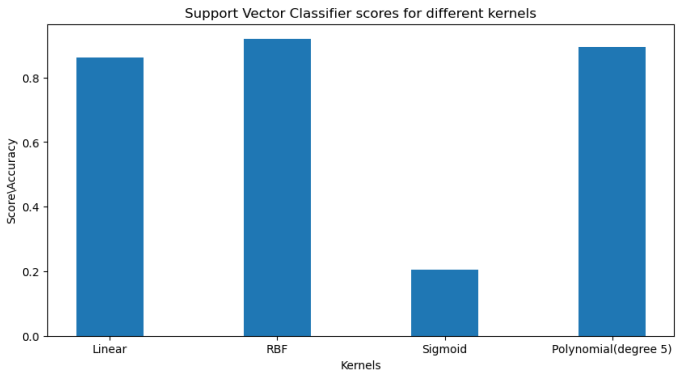

Figure 3: SVM kernels vs accuracy

The parameters used were C(Penalty Parameter), kernels and the weights of the instances. It was decided that instances which occur in nature have higher weights compared to those which were produced using SMOTE. The final parameter which were obtained are 208 for penalty parameter, kernel as RBF(with gamma as 0.1) and the value of weights as 5,1,0.5.

Table 1: Accuracy of different models

Accuracy vs Model	
Model	Accuracy
SVM	91.33%
Weighted SVM	92.03%
Light GBM	94.32%
XGBoost	94.13%

4 Conclusion

Out of the sample data we analyzed, the overall accuracy was found to be 91.33% for SVM and 92.03% for weighted SVM,94.32% for LBGM and 94.13% for XgBoost. The model ,thus can be used to classify the perovskites fairly accurately. The prediction accuracy was found to be as high as 94.32% which is fairly better than 80.3%[1] than one reported by Santost et al. Thus, this model can be fairly used for crystal structure predictions.

Future Scope

We utilized 4 models to predict the crystal structures of perovskites. Neural networks can be utilized to get the accuracy better than benchmark of 95%[2] reported in Jarin et al. Further, this model can be used to identify crystal structures of different halides and not necessarily oxide perovskites.

Dataset

The dataset was a part of a research by Emery et al.[3], licensed by MIT, and can be found at: `https://figshare.com/articles/dataset/Wolverton_Oxides_Data/7250417?file=13354619`

References

[1] Santosh Behara a, et al. "Crystal Structure Classification in ABO3 Perovskites via Machine Learning." Computational Materials Science, Elsevier, 1 Dec. 2020.

[2] Jarin, S.; Yuan, Y.; Zhang, M.; Hu, M.; Rana, M.; Wang, S.; Knibbe, R. Predicting the Crystal Structure and Lattice Parameters of the Perovskite Materials via Different Machine Learning Models Based on Basic Atom Properties. Crystals 2022, 12, 1570. https://doi.org/10.3390/cryst12111570

[3] Emery, Antoine Wolverton, Chris. (2017). High-Throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO 3 perovskites. Scientific Data. 4. 170153. 10.1038/sdata.2017.153.