2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)

Machine Learning Methods for Predicting the Lattice Characteristics of Materials

Filanovich A. N.
physics department
Ural Federal University
Ekaterinburg, Russia
a.n.filanovich@urfu.ru

Povzner A. A.
physics department
Ural Federal University
Ekaterinburg, Russia
a.a.povzner@urfu.ru

***Abstract*—Data on 5244 crystalline compounds from the open AFLOWlib repository are used to build machine learning models, which enable to predict important features of phonon spectrum of a material (Debye temperature and Gruneisen parameter) required for simulation of its lattice properties. We build two types of descriptors: the first one contains data solely on the chemical composition of a compound and the second one incorporates information on the elemental properties of atoms that make up the compound and additionally contains several features regarding its crystal structure. The regression models are built using four popular approaches – gradient boosting (GB), random forests (RF), artificial neural networks (ANN) and support vector machines (SVM). Prior the regression a search for the best values of hyperparameters has been performed for each of the model supplemented with a 5-fold cross validation. We compare prediction accuracy of models based on different methods as well as trained on each of the descriptors.**

Keywords—machine learning, regression, thermal properties

I. Introduction

For effective and safe application of materials, information on their thermal properties is needed, such as heat capacity, coefficient of thermal expansion, etc., at different temperatures and pressures. This information can be obtained either from experimental measurement or by using computer simulations. An experiment is often expensive both financially and in terms of the long time required for the synthesis and processing of samples, their certification, etc. In order to conduct sufficiently accurate simulations of the properties, information on the phonon spectrum of a substance is required, the calculations of which from the first principles are often complicated, especially in cases when the unit cell of the compound contains a large number of atoms.

On the other hand, it has been shown in [1] that, based on relatively simple descriptors, such as the chemical composition of a compound, it is possible to train a machine learning (ML) algorithm that predicts the properties of a substance. The disadvantage of such "direct" application of ML is that the prediction is carried out at a single temperature and pressure. If instead we will predict the quantities characterizing the phonon spectrum of a substance and its dependence on volume, such as the Debye temperature θ_D and Grüneisen parameter Γ [2], it becomes possible to quickly simulate the properties at various external conditions. The AFLOWlib open database [3] (available online at http://aflowlib.org) contains information on θ_D and Γ of several thousand compounds calculated from first principles. In this paper, we apply various ML algorithms (random forests, artificial neural nertworks, gradient boosting, etc.) to perform regression on the values of θ_D and Γ from the AFLOWlib base. We analyze the effect of various algorithms and two types of descriptors on the efficiency of predicting target properties θ_D and Γ.

II. Dataset Structure And Peculiarities Of The Algorithms

The complete AFLOWlib repository contains information on tens of thousands of compounds, however, not for all of them data both on the Debye temperature θ_D and on the Gruneisen parameter Γ are available. By querying with The LUX materials search API [4] only those compounds for which both values are available, we obtained a dataset for 5244 compounds that contains information on the Debye temperature θ_D with values ranging from 23 K to 2145 K (average value is 363 K and standard deviation is 219 K) and the Gruneisen parameter with an interval of values from 1.0 to 3.5 (average value is 2.1 and standard deviation is 0.24). Further, based on the chemical composition of the compounds, we formed two types of descriptors used to train the regression models.

The first of the descriptors under consideration – let us call it pseudo-binary was constructed following the ideas of [1,5]: each column corresponds to a certain element of the Periodic System from the entire list of elements in all the compounds that make up the dataset. The numerical value of each of the columns for a given row (corresponding to a particular compound) is equal to the number of atoms of a given element in the chemical formula of the compound.

The second descriptor contains data on the properties of the atoms that make up the compound, as well as data on its crystal structure. Of the atomic properties, information on the mass, electrical charge of the nucleus (which is equal to the ordinal number of the element in the Periodic system), covalent radius and electronegativity was included borrowed from the WebElements resource (available online at http://webelements.com). The features of the crystal structure were taken into account through information on the distances between the nearest neighbors and the average volume per atom. This information, along with the values of the predicted values – the Debye temperature and the Gruneisen parameter, is contained in the Aflowlib database. The record for each compound in this descriptor contains the maximum and minimum value for each of the properties among the elements in this compound, as well as the average value. Prior the regression, the obtained numbers were processed using the "preprocessing" module in scikit-learn so that the array of values of each of the features had zero mean and unit variance.

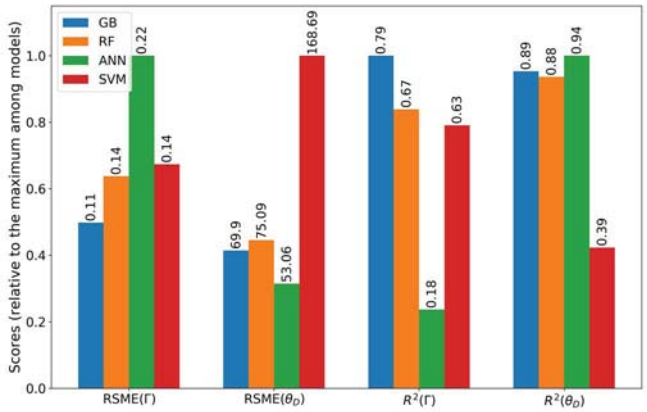
The regression was carried out within the framework of four popular techniques – gradient boosting (GB), random forests (RF), artificial neural networks (ANN) and support vector machines (SVM), implemented using Python scikit-learn libraries [6]. For each of the models, optimization of

978-1-7281-3165-8/20/$31.00 ©2020 IEEE

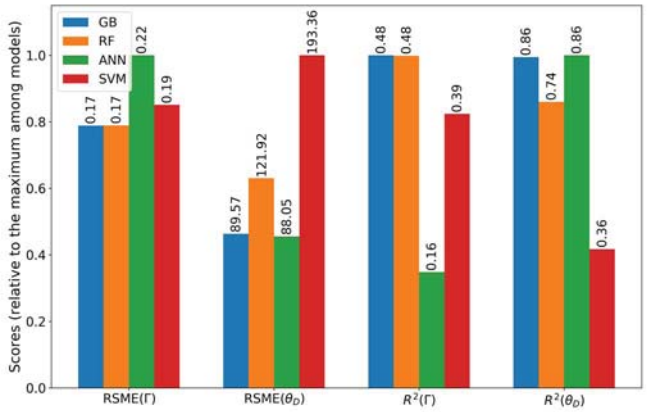
hyperparameters was performed using the RandomizedSearchCV method implemented in scikit-learn. This method forms *n* random samples from the set of hyperparameters of this model and determines which of the sets corresponds to the model that best describes the input data. We have chosen $n = 50$, and for each iteration, 5-fold cross-validation was performed in order to avoid overfitting. The initial dataset was divided into two parts – 95% was used to train the models (these 95% were, in turn, divided into train and test parts during the 5-fold cross-validation) and 5% of the data was used exclusively for model evaluation. After the optimal values of the hyperparameters were determined, each of the models was used to predict the target properties of the test set of compounds. As the evaluation metrics we selected the root-square-mean error (RSME), which characterizes how numerically the data predicted by the model differ from the true values, and the coefficient of determination R^2, which determines how well the model describes the data as compared to the horizontal line (the closer this coefficient to unity, the better is description given by the model).

III. Results And Discussion

The regression results for the training and test datasets using a pseudo-binary descriptor are presented in Fig. 1. Since the predicted values have different dimensions and magnitudes, for convenience, each of the characteristics was divided by the maximum value among the four models (thus we get relative units), while the absolute value is shown in numerical form next to the corresponding column of the diagram.



a)

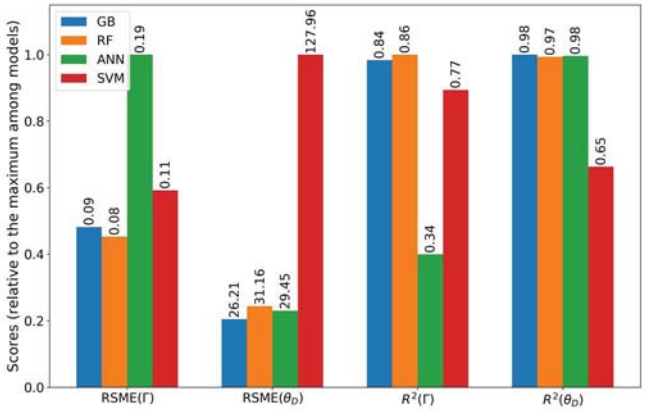


b)

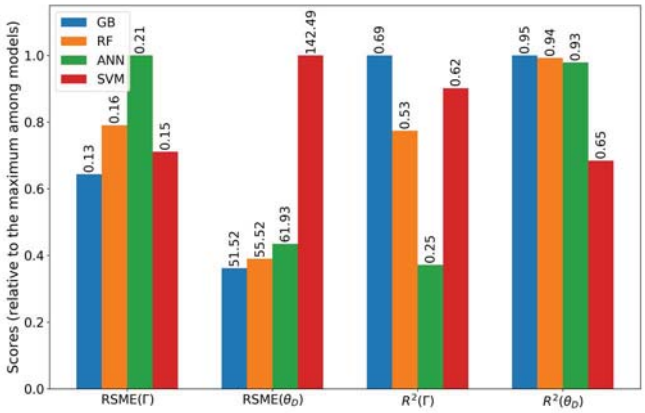
Fig. 1. Evaluation metrics for the ML models trained with the pseudo-binary descriptor: a) train set b) test set

As might be expected, RSME is slightly lower, and R^2, on the contrary, is slightly higher in the case of training dataset compared to the test one. It can be seen that in the case of the Debye temperature θ_D, the best description is provided by the models based on GB and ANN, and in the case of the training set ANN is even slightly ahead of GB. On the other hand, the ANN based model is significantly inferior to the other models (even to the SVM-based model that demonstrates significantly lower performance on θ_D) in predicting the Grüneisen parameter Γ. It is worth noting that, for all models, predicting Γ is a more difficult task compared to θ_D: if in the case of the latter, R^2 reaches 0.86 for the test data set, for the former it does not exceed 0.48.

Such low values of R^2 in the case of Grüneisen parameter indicate the need to use more advanced descriptors that take into account not only the formal composition of the compound, but also the physicochemical characteristics of compounds with different composition. This is exactly what was implemented in the second descriptor that we designed, the regression results for which are presented in Fig. 2.



a)



b)

Fig. 2. Evaluation metrics for the ML models trained with the descriptor that takes into account the physicochemical characteristics: a) train set b) test set

One can see that this descriptor enables to significantly improve the accuracy of prediction of both target properties, which is especially noticeable for the SVM-based model: training it with the descriptor that takes into account physicochemical properties, improves R^2 for the test data by more than two tenths. As in the case of the pseudo-binary descriptor, all models demonstrate worse accuracy predicting

the Gruneisen parameter compared to Debye temperture, especially the model based on neural networks (ANN), which, nevertheless, practically doesn't inferior to the models based on GB and RF in predicting θ_D.

Overall, for the test data, the GB-based model gives the best description, therefore, in figures 3 and 4 we show a comparison between the true (from the AFLOWLIB database) and predicted values for this model. As could be expected from the said above, in the case of the Debye temperature, a much smaller deviation of the points from the straight line is observed, which corresponds to the equality of the predicted and true values.

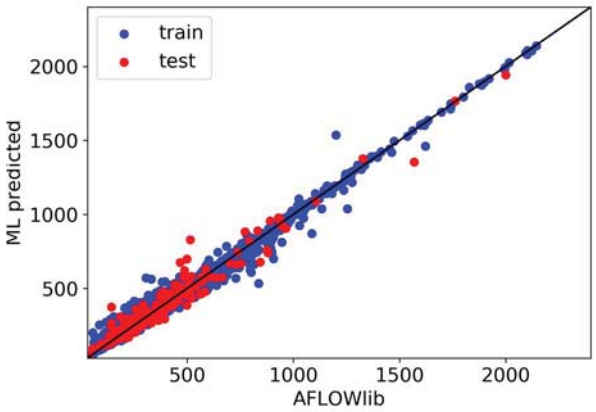


Fig. 3. The results of prediction of Debye temperature by the GB-based model

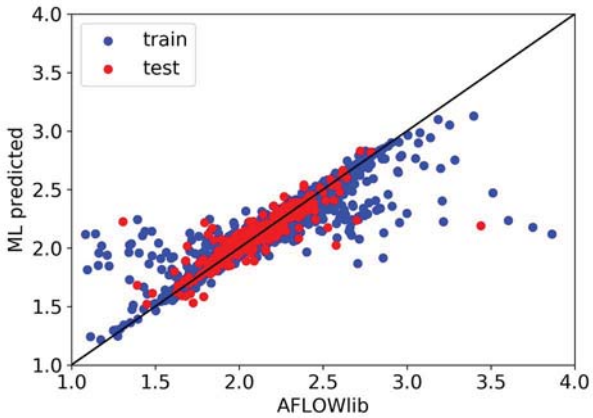


Fig. 4. The results of prediction of Gruneisen parameter by the GB-based model

IV. Conclusion

In the present work, it has been confirmed that using machine learning algorithms it is possible to predict such characteristics of the phonon spectrum of solids as the Debye temperature and the Gruneisen parameter, which enable calculation of the lattice properties. The prediction is possible even with a simple descriptor containing only formal information about the chemical composition of the compound. Nevertheless, the descriptor that contains information on the physicochemical properties of the elements entering the compound, as well as its crystal structure, provides better accuracy of prediction. It is demonstrated that the choice of machine learning algorithm depends on the target property. While the ANN-based model demonstrates higher efficiency in predicting the Debye temperature in comparison with the SVM-based model, in the case of the Grüneisen parameter it is exactly the opposite. Overall, the most accurate predictions can be obtained using a model based on the gradient boosting (GB) algorithm. The information obtained in this study can be useful for a quick evaluation of the Debye temperature and Gruneisen parameter, e.g. for high throughput calculations and the search for new materials with desired properties.

References

[1] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, "How chemical composition alone can predict vibrational free energies and entropies of solids", Chem. Mater., vol. 29, pp. 6220-6227, 2017.

[2] Charles Kittel, Introduction to Solid State Physics, 8 ed. NJ: Wiley, 2005.

[3] S. Curtarolo, W. Setyawan, S. Wang, et al. "AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations", Comput. Mater. Sci., vol. 58, pp. 227-232, 2012.

[4] F. Rose, C. Toher, E. Gossett, et al., "AFLUX: The LUX materials search API for the AFLOW data repositories", Comput. Mater. Sci., vol. 137, pp. 362-370, 2017.

[5] J. J. Moller, W. Korner, G. Krugel, D. F. Urban, et al., "Compositional optimization of hard-magnetic phases with machine-learning models", Acta Materialia, vol. 153, pp. 53-61, 2018.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: machine learning in Python", J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.