Machine Learning Regression Algorithm Predicts Multi-component Crystal Configuration Energy

Machine Learning Regression Algorithm Predicts Multi-component Crystal Configuration Energy

Peng Wang[1], Jinshuo Mei[1],Yingjie Lang[1]and Shu Li[1,*]

[1]School of Science, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China

*Corresponding author e-mail: lishu@hrbust.edu.cn

Abstract. Some machine learning algorithm tools, such as neural networks and Gaussian process regression, are increasingly being applied to the exploration of materials. Here, we have developed a form to use this nonlinear interpolation tool to describe properties that depend on the degrees of freedom in multi-component solids. A symmetrically adapted clustering function is used to distinguish different atomic order degrees. These features are used as the input of neural networks, Gaussian process regression and other algorithmic models, and some inherent properties of materials, such as formation energy, can be reproduced by the trained machine algorithm model. We use this technique to reproduce the expansion Hamiltonian of a synthetic cluster with multi-body interaction, and calculate the formation energy of ZrO based on first principles. The form proposed in this paper and the results shown that complex multi-body interactions can be approximated by nonlinear models involving smaller clusters. The training models used in this paper to predict energy include neural networks, Gaussian process regression, random forests, and support vectors regression, using MSE and coefficient of determination to evaluate the prediction results, and adding genetic algorithms in the feature selection process can remove some redundant features and improve the prediction efficiency and accuracy. The results show that the neural network is the best algorithm model which selected in this article, the prediction effect of support vector regression is relatively inferior.

1. Introduction

In recent years, machine learning tools have increased significantly in the research of materials science[1]. Combining large-scale databases and high-throughput calculations[, it has become a global trend to explore the chemistry of new materials. At the same time, machine learning tools are increasingly used to construct interatomic force-fields.

The lattice model Hamiltonian plays a central role in the first-principles statistical mechanics scheme to predict the thermodynamic potential and diffusion coefficient[2]of alloys and off-stoichiometric compounds. Sanchez et al. proposed the implementation framework of Connolly and Williams6 in 1984. The basis of this formalism is to construct a complete and orthonormal cluster base in the configuration space, which allows the cluster expansion method to describe the configuration function according to the expansion coefficient, that is, the CE method, which has been widely used in First-principles calculation of the physical properties of metals and semiconductor alloys. Then, Sanchez et al. built it on a solid theoretical basis[3]. In the case of strictly derived cluster expansion,

the effective Hamiltonian can be expressed by the orthogonal basis function of the configuration occupying variable. The cluster expansion form constructs a natural mathematical framework by which the characteristics of the crystal can be expressed as a function of the degree of freedom[3,4]of the site.

Cluster expansion is formulated as a linear cluster basis function multiplied by a constant expansion coefficient, which depends on the basic chemical properties and crystal structure of the multi-component solid. The cluster expansion is precise in form, but in fact it must be truncated. Many advanced and efficient methods have been developed to help accurately and effectively parameterize truncated cluster expansion. These methods include the genetic algorithm[5]that selects the cluster basis set, the use of cross-validation[6]and regularization to reduce over-fitting schemes,

Here, we use the cluster expansion method as the basis, but relax the linear constraints[7], and use advanced machine learning tools (such as neural networks and Gaussian process regression) to express the crystal properties that depend on the alloy configuration (depending on the symmetry Invariant sequence descriptor).

This article is based on the GitHub open source computing software CASM[8], to obtain the configuration set of the ZrO material required to train the machine learning model and the corresponding input features and real energy values. The material is hcp Zr with octahedral interstitial O, after obtaining the above data set, this article selects neural network (NN), Gaussian process regression (GPR), support vector regression (SVR), random forest (RF) four algorithms for training, and then energy on the test set Forecast, and finally, a horizontal comparison and analysis of the forecasting effects of the above-mentioned machine learning algorithms. The work of this paper has certain reference value and practical significance for calculating the inherent properties of materials using machine learning algorithms.

2. Input features and feature selection methods

2.1. Cluster Expansion Method

According to Sanchez et al. [3] showed that any scalar properties of binary alloys depend on $\vec{\sigma}$, $\vec{\sigma} = \{\sigma_1, \ldots, \sigma_i, \ldots, \sigma_N\}$ .It is used to represent the specific order of the composition of the N-site meta-crystal σ_i Whether it is +1 or -1 depends on the occupancy of position i. For example, its completely relaxed formation energy can be expressed as a basis:

$$E(\vec{\sigma}) = NV_0 + \sum_{\alpha} V_\alpha \phi(\vec{\sigma}) \tag{1}$$

The sum extends to all clusters at site α in the crystal, and

$$\phi_\alpha(\vec{\sigma}) = \prod_{i \in \alpha} \sigma_i \tag{2}$$

The cluster function is defined as the product of the occupied variables belonging to the cluster α. .

Symmetry that imposes a constraint on the unmodified parent crystal structure of the expansion coefficient V_β . Any two cluster functions $\phi_\alpha(\vec{\sigma})$ with $\phi_\beta(\vec{\sigma})$ Having the same expansion factor (i.e. $V_\alpha = V_\beta$) Can be mapped to each other through spatial group operations of crystals. All can pass the symmetry operation of the crystal with the prototype cluster function $\phi_\alpha(\vec{\sigma})$ Related cluster functions $\phi_\beta(\vec{\sigma})$ Orbitals that can be combined into a cluster function $\Omega_\alpha = \{\phi_\alpha(\vec{\sigma}), \ldots, \ \phi_\beta(\vec{\sigma}), \ldots\}$ For example, all the cluster functions of the nearest neighbor pair clusters related to the prototype nearest neighbor pair cluster through the symmetry operation belong to the same orbit. However, in binary alloys, for each cluster type with different symmetry, there is a cluster function orbit, they belong to different cluster orbitals.Form a set of tracks.

In this way, equation (1) can be rewritten as:

$$E(\vec{\sigma}) = NV_0 + \sum_{\Omega_\alpha \in \Lambda} V_\alpha \sum_{\delta \in \Omega_\alpha} \phi_\delta(\vec{\sigma}) \tag{3}$$

The above formula can be normalized by the number of atoms in the crystal and rewritten as

$$\frac{E(\vec{\sigma})}{N} = V_0 + \sum_{\Omega_\alpha} V_\alpha \mathrm{m}_\alpha < \phi_\alpha(\vec{\sigma}) > \tag{4}$$

among them m_α Is the diversity of clusters at each site. related functions $< \phi_\alpha(\vec{\sigma}) >$ Is for a specific order $\vec{\sigma}$ track Ω_α Cluster letter $\phi_\alpha(\vec{\sigma})$ average value. Definition of related functions:

$$< \phi_\alpha(\vec{\sigma}) >= \frac{(\sum_{\delta \in \Omega_\alpha} \phi_\delta(\vec{\sigma}))}{N m_\alpha} \tag{5}$$

For binary alloys, each cluster type with different symmetry (for example, nearest neighbor pair cluster, second nearest neighbor pair cluster, nearest neighbor triplet cluster, etc.) has a correlation function $< \phi_\alpha(\vec{\sigma}) >$ related to. Since the correlation functions are the average of all symmetric equivalent cluster functions, they are unchanged for any space group operations applied to the parent crystal of the $\vec{\sigma}$ crystal in a specific order. Therefore, they are a measure of the ordered state of a particular configuration on a crystal that is not affected by the space group operation of the underlying crystal.

2.2. Feature Selection

The solution we propose in this article is to weaken the linear relationship between the correlation function and the formation energy in equation (4), but use machine learning algorithms to establish a nonlinear mapping relationship characterized by the correlation function and energy as the output value, achieve the prediction function. For example, in this article, 336 configurations of ZrO and 74 input features corresponding to each configuration were obtained through CASM software. The above four machine learning algorithms were first performed without feature optimization Model training and prediction, and then use genetic algorithm to screen 74 input features, select 25 of them as input, as shown in Table 1,Table 1 shows a small portion of the 336 configurations, as well as a portion of the 25 features screened out by feature optimization for each configuration.again use neural network, Gaussian process regression, random forest and support vector regression to train and predict, and get the results for comparison .

Table 1. Partial features of part training set after optimization.

	Corr1	Corr2	Corr3	Corr5	Corr6	Corr7	Corr8	Corr13
SCEL1_1_1_1_0_0_0/0	0	0	0	0	0	0	0	0
SCEL1_1_1_1_0_0_0/1	0.5	0	0.5	0.5	0.5	0.5	0	0
SCEL1_1_1_1_0_0_0/2	1	1	1	1	1	1	1	1
SCEL2_1_1_2_0_0_0/0	0.25	0	0.25	0	0.25	0	0	0
SCEL2_1_1_2_0_0_0/2	0.75	0.5	0.75	0.5	0.75	0.5	0.5	0.5
SCEL2_1_2_1_0_0_0/3	0.5	0	0.167	0.5	0.167	0.167	0.333	0
SCEL3_1_1_3_0_0_0/1	0.333	0	0.333	0.167	0.333	0.167	0	0
SCEL3_1_1_3_0_0_0/9	0.667	0.333	0.667	0.333	0.667	0.333	0.333	0.667
SCEL4_1_1_4_0_0_0/10	0.5	0.375	0.5	0.25	0.5	0.25	0.375	0.125
SCEL4_1_2_2_1_0_0/8	0.25	0	0.833	0	0.833	0	0.833	0.125

3. Formation energy prediction

After obtaining the labels and input features required for the above training, we can choose some machine learning models for training. For example, several models were chosen, based on the 336 configuration data set I got above,We use CASM to get the true value for the formation of each configuration case and each label contains 74 Correlation function (corr) is used as input features, 75% of which are used as the training set and 25% as the test set. The convex hull images used in the training set to form real values are shown in Figure 1.The Bootstrap method is used to randomly select the test set during testing, and MSE (mean square error) and R^2coefficient of determination) It is used

to evaluate the degree of fit between the predicted value and the true value. The smaller the MSE, the better the fit, and the larger R^2, the better the fit.

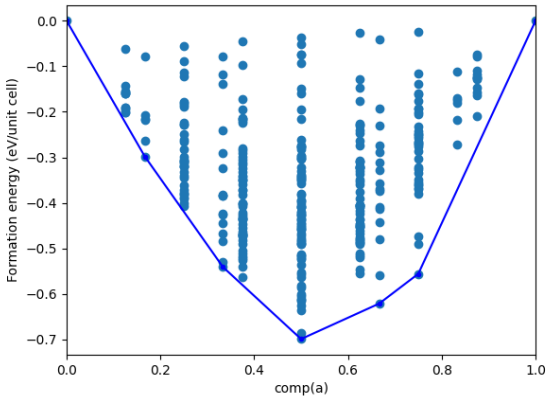


Figure 1. The convex hull of formation energy each configuration

In the process of machine learning model training, we respectively used all the features and the better features screened by genetic algorithm as input, and then tested the prediction results using the trained model. In the neural network algorithm, we used The hidden layer of the neural network used in this article uses a three-layer 4, 4, 2 node mode to activate Function choices at each node include rectified linear units (ReLU), sigmoid and hyperbolic tangents[9].As shown in Figure 2.

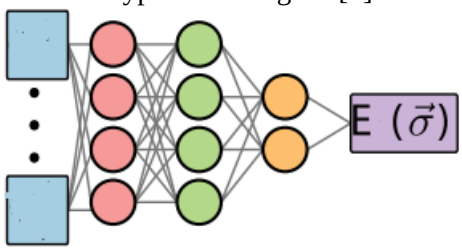
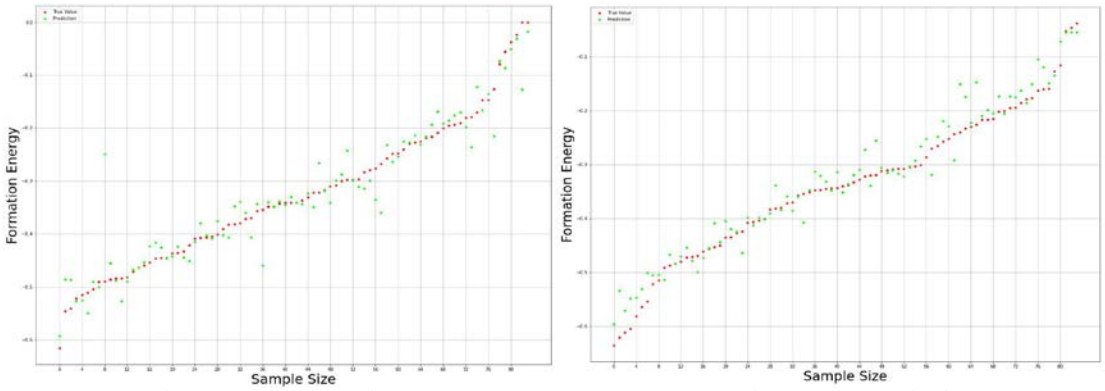


Figure 2. A neural network model was used to calculate configuration energy

We use advanced gradient descent techniques such as ADAM[10] that adaptively change the learning rates for each weight parameter, loss function adopts MSE:

$$\Gamma = \frac{1}{M}\sum_{\vec{\sigma}}\left(E(\vec{\sigma}) - E_{DFT}(\vec{\sigma})\right)^2$$

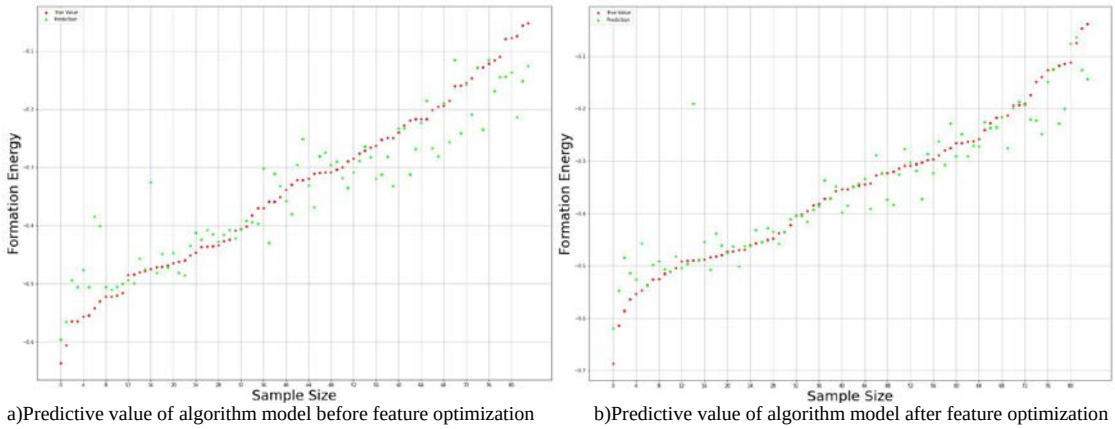
The energy prediction results obtained by the neural network are shown in Figure 3:



a)Predictive value of algorithm model before feature optimization b)Predictive value of algorithm model after feature optimization

Figure 3. The fitting effect of neural network on formation energy prediction results and real values

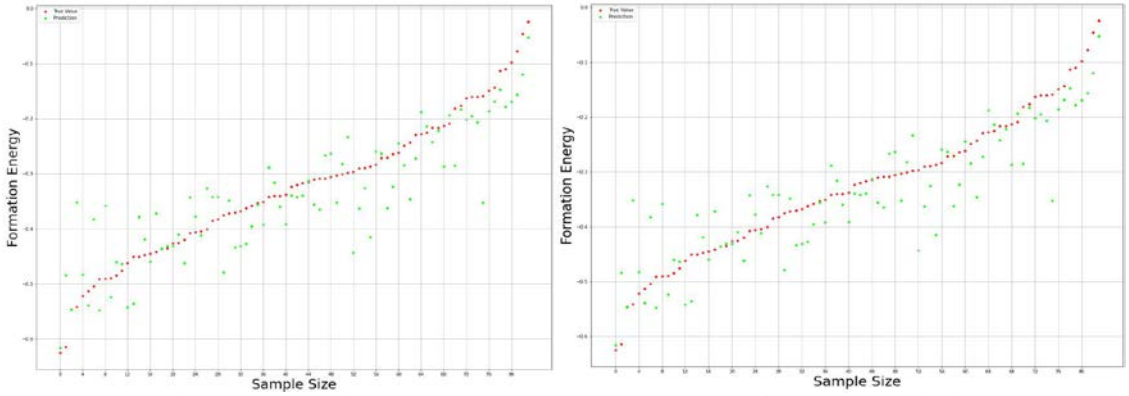
The energy prediction results obtained by random forest (RF) are shown in Figure 4:



a)Predictive value of algorithm model before feature optimization b)Predictive value of algorithm model after feature optimization

Figure 4. The fitting effect of random forest (RF) on formation energy prediction results and real values

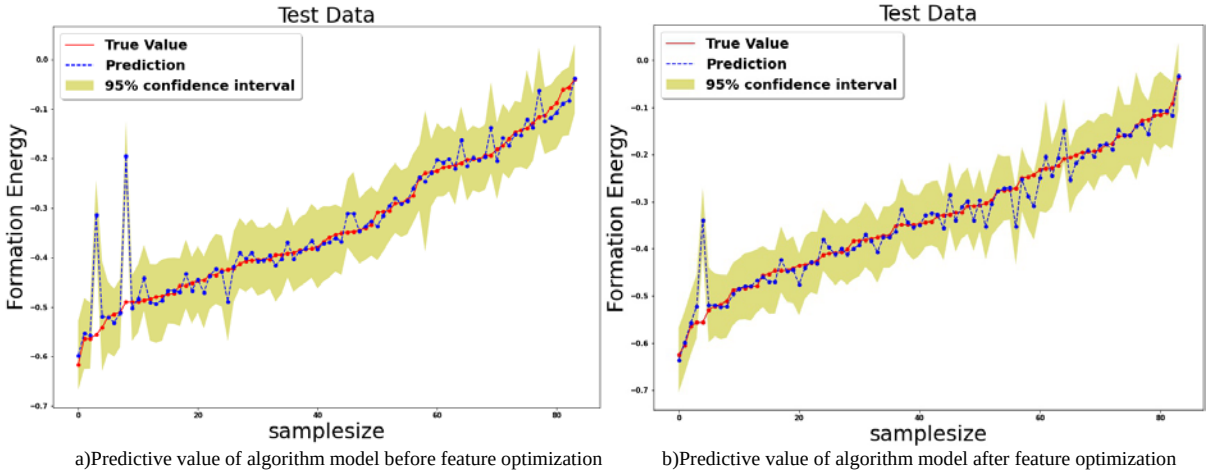
The energy prediction results obtained by support vector regression (SVR) are shown in Figure 5:



a)Predictive value of algorithm model before feature optimization b)Predictive value of algorithm model after feature optimization

Figure 5. The fitting effect of support vector regression (SVR) on formation energy prediction results and real values

The energy prediction results obtained by Gaussian process regression (GPR) are shown in Figure 6:



a)Predictive value of algorithm model before feature optimization b)Predictive value of algorithm model after feature optimization

Figure 6. The fitting effect of Gaussian process regression (GPR) on formation energy prediction results and real values

After training and forecasting with the above four machine learning models respectively, mean square error and determination coefficient are obtained respectively to reflect the prediction effect of each machine learning. Their values are shown in Table 2.

Table 2.MSE and R^2 values generated by the four machine learning models predicting .

	before feature optimization		after feature optimization	
	MSE	R^2	MSE	R^2
NN	0.00177394349472385	0.9040089343592	0.000960273988933	0.9471519720615
GPR	0.00208119874259705	0.8929144486235	0.001029807234162	0.9418407291903
RF	0.00279865619002893	0.8704372005930	0.002637650895562	0.8700479154688
SVR	0.00406691300430967	0.7558093384159	0.002577661640719	0.8470975203675

4. Conclusion

According to the prediction results of the last four machine learning algorithms, on the whole, all the above regression algorithms have good prediction effect and small mean square error, but the best performance is that of neural network, and the less good is that of support vector regression.Moreover, in the calculation process, from the perspective of the final mean square error and determination coefficient, the excellent features screened by machine learning algorithm are used as the input of the training model, and the predicted results and real values are more effective in fitting.

To sum up, when appropriate physical features are selected as input features of machine learning model, the machine learning algorithm performs well and is more efficient on the whole when computing crystal related attributes

5.Acknowledgments

The authors acknowledge the support of National Natural Science Foundation of China (No. 51671075), Heilongjiang Postdoctoral Fund for Scientific Research initiation (No. LBH-Q16118) and Fundamental Research Foundation for Universities of Heilongjiang Province (No. LGYC2018JC004).

References

[1] T. Mueller, A. G. Kusne, R. Ramprasad, Machine learning in materials science: recent progress and emerging applications, Rev. Comput. Chem. 29(2016) 186–273.
[2] A. Van der Ven, J. C. Thomas, B. Puchala, A. R. Natarajan, First-principles statistical mechanics of mult-component crystals, Annu. Rev. Mater. Res. 48 (2018) 27 – 55.
[3] J. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems, Phys. A: Stat. Mech. Appl. 128 (1984) 334 – 350.
[4] D. De Fontaine, Cluster approach to order-disorder transformations in alloys, Solid State Phys. 47 (1994) 33-46 .
[5] V. Ozoliņš, C. Wolverton, A. Zunger, Cu-Au, Ag-Au, Cu-Ag, and Ni-Au intermetallics: first-principles study of temperature-composition phase diagrams and structures, Phys. Rev. B 57 (1998) 6427.
[6] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilibria 23 (2002) 348 .
[7] A. R. Natarajan, A. Van der Ven, Machine-learning the configurational energy of multicomponent crystalline solids, npj Comput. Mater 4 (2018) 1 .
[8] CASM Developers. CASM: A clusters approach to statistical mechanics (2016).
[9] V. Nair, E. Hinton, Rectified linear units improve restricted boltzmann machines, In Proceedings of the 27th International Conference on Machine Learning, edited by J. Furnkranz et al. Haifa, Israel, 2010, pp. 807 – 814 .
[10] D. P. Kingma, J. Ba, ADAM: A Method for Stochastic Optimization (2015).