Supporting Information

Machine Learning Prediction of Superconducting Critical Temperature through Structural Descriptor

Jingzi Zhang[a,b], *Zhuoxuan Zhu*[d], *X.-D. Xiang*[d*], *Ke Zhang*[a,b], *Shangchao Huang*[a,b], *Chengquan Zhong*[a,b], *Hua-Jun Qiu*[a,b], *Kailong Hu*[a,b*], *and Xi Lin*[a,b,c*]

[a] School of Materials Science and Engineering, Harbin Institute of Technology, Shenzhen 518055, P. R. China

E-mail: linxi@hit.edu.cn

E-mail: hukailong@hit.edu.cn

[b] Blockchain Development and Research Institute, Harbin Institute of Technology, Shenzhen 518055, P.R. China

[c] State Key Laboratory of Advanced Welding and Joining, Harbin Institute of Technology, Harbin 150001, P. R. China

[d] Department of Materials Science and Engineering & Department of Physics, Southern University of Science and Technology, Shenzhen 518055, P.R. China

E-mail: xiangxd@sustech.edu.cn

Machine Learning Model Details

Random Forest Regression (RFR) algorithm

The Random Forest Regression (RFR) model is an ensemble learning approach for regression that builds a large number of decision trees during training and outputs the class that is the mean/average prediction (regression) of the individual trees. [1] Random decision forests counteract the tendency of decision trees to overfit their training set. [2] The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1 \ldots x_n$ with responses $Y = y_1 \ldots y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

(a) For b = 1... B: Sample, with replacement, n training examples from X, Y; call these X_b, Y_b; Train a regression tree f_b on X_b, Y_b.

(b) After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x'

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{S1}$$

Xgboost Regression (XGBR) algorithm

The Xgboost Regression (XGBR) model is a highly effective and widely tree boosting in machine learning method. The scalability of XGBR is due to several important systems and algorithmic optimizations. A novel tree learning algorithm is for handling sparse data; a theoretically justified weighted quintile sketch procedure enables handling instance weights in approximate tree learning. [3-4]

For given dataset with n examples and m descriptors:

$$D = \{(x_i, y_i)\} (|D| = n, x_i \in R^m, y_i \in R) \quad \text{(S2)}$$

a tree ensemble model uses K additive functions to predict the output.

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad \text{(S3)}$$

$$F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^T) \quad \text{(S4)}$$

Where F is the space of regression trees. Here q represents the structure of each tree that maps an example to the corresponding leaf index. T is the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and leaf weights w.

Support Vector Regression (SVR) algorithm

The Support Vector Regression (SVR) model is a supervised learning model with related learning algorithms for regression analysis that analyzes data.[5] SVR training algorithm creates a model that assigns new instances to one of two categories, making it a non-probabilistic binary linear classifier, given a series of training examples that are individually designated as belonging to one of two categories. [6] An SVR model is a representation of the instances as points in space, mapped so that the examples of the different categories are separated by a large gap. New instances are then mapped into the same area and assigned to one of the categories based on which side of the gap they land on. SVR can do non-linear regression as well as linear regression by applying a technique known as the kernel trick, which involves implicitly mapping inputs into high-dimensional feature spaces. Training the original SVR to:

$$\text{Minimize}\frac{1}{2}||w||^2 \tag{S5}$$

$$\text{Subject to } |y_i - \langle w, x_i \rangle - b| \leq \varepsilon \tag{S6}$$

where x_i is a training sample with target value y_i. The inner product plus intercept $\langle w, x_i \rangle - b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold: all predictions have to be within a ε range of the true predictions. Slack variables are usually added into the above to allow for errors and to allow approximation in the case the above problem is infeasible.[7]

R^2 score, MAE, RMSE

The prediction accuracy of ML models is evaluated with R^2 score, MAE (Mean Absolute Error) and RMSE (Root Mean Square Error). Where y_i is the truth values, $f(x_i)$ is the prediction values of machine learning model, which are expressed as:

$$R^2 = 1 - \frac{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2}{\frac{1}{m}\sum_{i=1}^{m}(y_i - \overline{y_i})^2} \tag{S7}$$

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - f(x_i)| \tag{S8}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2} \tag{S9}$$

R^2 score ranges from 0 to 1, and the closer it is to 1, the better the prediction performance of the model is. MAE and RMSE represents the loss between the prediction and the ground truth. The loss is lower, the model performs better. [8-9]

Table S1. Detailed parameters of three ML models.

ML model name	Parameters
RFR	Criterion: mse N_estimators: 400 Max_depth: 24
XGBR	Booster: gbtree N_estimators: 400 Max_depth: 5
SVR	Gamma parameter: scale Kernel: RBF

TableS2. SOAP descriptor parameters

Descriptor name	Descriptor parameters
SOAP	Rcut: 6 Nmax: 1 Lmax :1 Average: inner

Table S3. The R^2 score, MAE and RMSE of three models on 10-fold cross validation.

	Train			Test		
	R^2	MAE	RMSE	R^2	MAE	RMSE
RFR	0.986	1.98	3.24	**0.929**	**4.11**	**7.61**
XGBR	0.997	0.98	1.43	0.922	4.15	7.54
SVR	0.979	2.23	3.46	0.913	4.27	7.87

REFERENCES

[1] Grömping U. Variable importance assessment in regression: linear regression versus random forest[J]. The American Statistician, 2009, 63(4): 308-319.

[2] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.

[3] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[4] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4): 1-4.

[5] Awad M, Khanna R. Support vector regression[M]//Efficient learning machines. Apress, Berkeley, CA, 2015: 67-80.

[6] Smola A J, Schölkopf B. A tutorial on support vector regression[J]. Statistics and computing, 2004, 14(3): 199-222.

[7] Gu B, Sheng V S, Wang Z, et al. Incremental learning for ν-support vector regression[J]. Neural networks, 2015, 67: 140-150.

[8] De Myttenaere A, Golden B, Le Grand B, et al. Mean absolute percentage error for regression models[J]. Neurocomputing, 2016, 192: 38-48.

[9] Wang W, Lu Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model[C]//IOP conference series: materials science and engineering. IOP Publishing, 2018, 324(1): 012049.