Porosity Prediction of Ceramic Matrix Composites Based on Random Forest

Xiang Gao¹*, Lefan Wang¹ and Leijiang Yao²

¹School of Computer Science, Northwestern Polytechnical University, Xi’an 710129, China
²School of Laboratory of Science and Technology on UAV, Northwestern Polytechnical University, Xi’an 710072, China

*Corresponding author e-mail: gaoxg@nwpu.edu.cn

Abstract. Porosity is an important characteristic attribute of carbon fiber reinforced ceramic matrix composites (CMCs), which is closely related to material properties and directly affects the application range and prospect of composite materials. For the purpose of design optimization in Material Genetic Engineered (MGE), the porosity prediction method based on machine learning technology was proposed to provide material attribute data for propulsion material performance prediction. This study collect CMCs experimental data from papers and Materials Genome Engineering Databases. According to the composition of the materials and the characteristics of the preparation process, 9 influencing factors and 1674 experimental data were selected to establish random forest regression (RFR) and compared with support vector regression (SVR). Using R² scores and root mean squared error as model evaluation indicators, R² scores respectively are 0.751 and 0.927 in SVR and RFR, root mean squared error respectively are 0.0062 and 0.0019. The results of the study show that the RFR model predictions are well matched with the experimental values.

1. Introduction

Ceramic matrix composites (CMCs) has the characteristics of light weight, high strength, corrosion resistance and high temperature resistance, at the same time it has a performance of non-brittle fracture failure under stress, so widely used in the aerospace field [1]. Traditional materials science research mainly relies on the “trial and error” experimental method and iteratively follows the “propose hypothesis-experimental verification” method to continuously approach the target material. Material Genetic Engineering (MGE) explores the mapping between material structure and performance through data mining technology on experimental data, guiding material development and shortening the purpose of research and development cycle [2-4]. Porosity has a direct effect on the properties of material bending, compression, fatigue, interlaminar shear, fracture toughness, thermal expansion coefficient, etc. To predict material properties, porosity is an attribute that must be considered.

Due to its special manufacturing process, composite materials have different types of defects. The pores are the most common and most difficult to be detected [5]. Porosity, as a criterion for evaluating the pore content, refers to the percentage of the pore volume of the material to the total volume of the material in its natural state. The high porosity of the material corresponds to the low density of the material, and at the same time the strength of the material decreases [6]. At present, the most commonly used method for porosity measurement is ultrasonic method, and the ultrasonic method is divided into ultrasonic attenuation method and nonlinear ultrasonic detection method [7,8]. Ultrasonic
testing only characterizes the porosity when measuring, but does not give accurate values of porosity [9]. The absolute error of porosity predicted by random forest regression (RFR) model based on the existing material data [10] is less than the measurement error. Besides, all porosity detection techniques are based on existing materials, while prediction models based on data analysis can accurately estimate porosity during the material design phase.

2. Data and model

2.1. Data Sources

In order to ensure the safety and reliability of CMCs data, there are mainly two aspects to obtain experimental data. First, collect data from the papers and select relatively complete papers for data extraction. Second, collect data from Materials Genome Engineering Databases. Under the impetus of material genetic engineering (MGE), the material genetic engineering database has been initially established [11]. The data set (https://www.mgedata.cn/) is uploaded and reviewed by CMCs relevant experts, and the security and accuracy of the data can be guaranteed.

2.2. Feature extraction

Porosity is primarily determined by the complex process of the composite [12]. According to the knowledge of material preparation process, 9 kinds of influencing factors such as fiber tow size, weaving method, prefabricated type, fiber grade, fiber volume content, interface composition, interface process, matrix composition and matrix process were selected as independent variables [13].

2.3. Data preprocessing

There are two types of data for the selected features, which are numeric and textual. In this study, the data size of the tow, the volume of the fiber, and the porosity of the target variable were all normalized to [0,1] using the data preprocessing method of MinMaxScaler [14]. For the text attributes such as weaving method, matrix component, interface component, etc., need to be converted into digital types that can participate in modeling. This work uses One-Hot Encoding [15] to extend the discrete feature values to a certain point in the European space. The calculation of the distance between features is more reasonable.

2.4. Random forest regression model

The random forest is an extension based on the decision tree and can solve the problem of weak generalization of the decision tree [16]. The final prediction results are obtained by summarizing several independent sub-decision tree models internally. The most important parameters that need to be optimized in random forest regression are the number of decision trees and the maximum depth of the decision tree, and the number of random features. These three parameters are analyzed in turn to select the most suitable combination of parameters for modeling.

As shown in Figure 1, the methods and steps for predicting the porosity of random forest regression ceramic matrix composites are summarized as follows:

1) Using the Bootstrap method to sample the data set and randomly generate n training sets.
2) Each training set is constructed with a separate regression tree. Randomly extract F features from the input M different features as a subset of the current node split, and use the CART method to select the best split mode from this subset to split.
3) Each regression tree will output a predicted value of $Y_i$, $i=1,2,...,n$. The predicted value of the regression forest $\hat{y}$ is obtained by taking the average of n predicted values.
3. Experiment and optimization

For a total of 1674 data, 335 were randomly selected as test sets that did not participate in modeling for final evaluation of model accuracy. Ten-fold cross-validation is used in the model training process to avoid over-fitting.

Support Vector Regression (SVR) is one of the classic supervised type regression methods with good performance on smaller data sets [17]. For the penalty parameter C and the parameter gamma of the support vector regression model, use the grid search method to test. When the C is 3.5 and the gamma is 0.1, we got the best $R^2$ score which is 0.751 and root mean squared error on the test dataset which is 0.0062. The prediction accuracy of using SVR is not very good, probably because the data which comes from the paper has part of missing values in it. In addition, 70% of the features are text-type values, and the feature matrix is sparse after encoding, which is difficult to learn.

The main parameters that need to be set in the RFR model are the number of decision trees, the maximum number of features F used by a single decision tree, and the largest leaf node of the regression tree. The three variables are controlled separately in turn, and the performance of the RFR model is shown in the Fig2 below.
Figure 2. The relationship between the parameter and $R^2$ score, a) the number of decision trees b) the maximum number of features, c) the max leaf node, d) accuracy of the RFR.

Figure a shows the $R^2$ score change for each corresponding test set from 1 to 100 for the sub-regress tree contained in the RFR model. As the number of regression trees increases, the $R^2$ score gradually increases. When the number of regression trees increases to a certain extent, the $R^2$ score no longer rises and gradually stabilizes, but will slightly float above 0.858. Figure b shows the change in the $R^2$ score of the model on the test set when the maximum number of set features $F$ is from 5 to 100. Adding max features can improve model performance because there are more options to consider on each node, but it also means less tree diversity and slower the algorithm execution. The results show that when max features increased to about 20, the accuracy begins to decline. Although the initial selection of feature values is 9, the data set has 153 dimensions after encoding the textual features. Therefore, each sub-regression tree can use the square root of the total feature number to predict better. Figure c shows that the maximum number of leaf nodes in the initial stage is monotonous with $R^2$ score, and then no longer rises to be stable, so the maximum number of leaf nodes is set to be above 100. The RFR model parameters combination and evaluation are recorded in Table 1.

| Parameters | $n$ estimators | Max leaf nodes | Max features | $R^2$ | RMSE |
|------------|----------------|----------------|--------------|-------|------|
|            | 10             | 50             | sqrt         | 0.853 | 0.0035 |
|            | 20             | 50             | sqrt         | 0.857 | 0.0035 |
|            | 20             | 100            | sqrt         | 0.898 | 0.0026 |
|            | 50             | 100            | sqrt         | 0.927 | 0.0019 |

Some data with a range of porosity error were selected from the data set and predicted using SVR and RFR models respectively. The results are shown in Table 2.

| Measured value | SVR Predictive value | RFR Predictive value |
|----------------|----------------------|----------------------|
| Minimum       | Maximum              | Average              | Absolute error | 11.89 | 10.07 |
| 8             | 12                   | 10                   | 4              | 12.96 | 11.25 |
| 10.76         | 12.04                | 11.4                 | 1.28           | 10.78 | 10.82 |
| 10.1          | 11.3                 | 10.7                 | 3              | 7.65  | 6.98  |
| 5             | 8                    | 6.5                  | 3              | 16.23 | 15.26 |
| 14            | 17                   | 15.5                 | 0.4            | 11.3  | 10.67 |

It can be seen that the porosity predicted by the RFR model is within the measurement error range and is close to the average of the measured values. Table 3 is the absolute error distribution of the predicted values obtained on the test set using the RFR model.
Table 3 shows that the predicted value of less than the measured maximum absolute error of 4 is 98%, and the absolute error of 85% is less than 2.

4. Conclusion
This study used experimental data to establish a random forest regression model. The optimal RFR prediction model was obtained by adjusting the number of trees in the random forest, the depth of the tree, and the maximum number of features, and compared with the SVR model. The results show that the RFR can be well matched with the porosity in the experimental data, which provides reference value for material design.

Through this research, it is found that the experimental data extracted from the papers is not detailed enough and comprehensive, and some factors affecting the performance of the materials have not been explored.

5. Acknowledgments
This work was financially supported by the National Key R&D Program of China (2016YFB0700504) and the Science and Technology Project of Shaanxi Province (2019GY-048).

References
[1] Jinguang Du, Haizhen Zhang, Yongmiao Geng, Wuyi Ming, Wenbin He, Jun Ma, Yang Cao, Xiaoke Li, Kun Liu, A review on machining of carbon fiber reinforced ceramic matrix composites, Ceramics International, Volume 45, Issue 15, 2019, 18155-18166.
[2] Gao Hong, He Duanpeng, Dong Li. Research and Thoughts on "NASA Materials Genetic Engineering 2040 Planning" [J]. Spacecraft Environmental Engineering, 2019, 36(03): 205-210.
[3] Wu Miaomiao, Liu Limin, Han Yafang. Material genetic engineering - material design, simulation and top-level design of database [J]. Modern Science, 2018 (10): 53-58.
[4] Wang Hong, Xiang Xiaodong, Zhang Yuting. Data + artificial intelligence is the core of material genetic engineering [J]. Science and Technology Review, 2018, 36 (14): 15-21.
[5] Yue Liu, Tianlu Zhao, Wangwei Ju, Siqi Shi. Materials discovery and design using machine learning. Journal of Materiomics. 2017.
[6] Liu Ling, Lu Mingkun, Zhang Boming, et al. Effects of porosity on the ultrasonic absorption coefficient and mechanical strength of carbon/epoxy composites [J]. Acta Material Composite Silica, 2014, 21(5): 116-121.
[7] Chen Xianzhen, Deng Yanan, Ren Wenjian. Study on Ultrasonic Evaluation Method of Porosity of Carbon Fiber Composites [J]. Journal of Mechanical Strength, 2018, 40(06): 1330-1335.
[8] Fan Kai, Lu Xuefeng, Lu Kaiming, Qian Kun. Research Progress in Pore Structure of C/C Composites [J]. Materials Reports, 2019, 33(13): 2184-2190.
[9] Hong Yu, Wang Wujie, Liu Jiaqin, Tang Wenming, Wu Yucheng. Effect of Porosity and Interfacial Structure on Thermal and Mechanical Properties of High Volume Fraction SiC_p/Al Composites [J]. Transactions of Nonferrous Metals Society of China, 2019, 29 (05): 941-949.
[10] Raccuglia P, Elbert K C, Adler P D F. et al. Machine-learning-assisted materials discovery using failed experiments. Nature. 2013.
[11] Du Bin. Preparation and properties of SiOC/CBCF porous ceramic matrix composites and their oxidation resistant coatings [D]. Harbin Institute of Technology, 2018.
[12] Wuhong Lin, Jianfeng Huang, Ching Yee Suen, Lihua Yang. A feature extraction model based on discriminative graph signals [J]. Expert Systems With Applications, 2020, 139.
[13] Wang Yuefei. Research on molding process and pore control of carbon fiber reinforced composite HP-RTM [D]. Hunan University, 2017.
[14] Bai Fengwei. Research and implementation of several key technologies of data preprocessing system [D]. Beijing Jiaotong University, 2012.
[15] Beck, J.E. and B.P. Woolf. High-level student modeling with machine learning. in International Conference on Intelligent Tutoring Systems. 2000. Springer
[16] Random Forests[J]. Leo Breiman. Machine Learning. 2001 (1)
[17] WANG Ding-cheng, LU Yi-yi, ZOU Yong-jie Multi-output Intuitionistic Fuzzy Least Squares Support Vector Regression Algorithm Computer Science, 2019, 46(5)