Machine-Learning-Assisted the Design of Resin Matrix Composites Coating with Ablation Resistance

D An*1,2, Z Ma1,2, C Y Li1,3, C Ma1,2, W Z Li1,2

1School of Material Science and Engineering, Beijing Institute of Technology, Beijing, 100081, China

2National Key Laboratory of Science and Technology on Material Under Shock and Impact, Beijing, 100081, China

3School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, 100081, China

andong@bit.edu.cn

Abstract. Traditional experimental methods always cost a lot but produce little when designing and developing new kinds of materials, especially for coating materials. However, with the assistance of machine learning, it is possible to predict the performance of a specific coating without preparation or simulation, which makes the design of material more efficient. In this study, machine learning was introduced to assist the design of resin matrix composites coating with ablation resistance. A structured method for engineering data in the material field was approved. Based on this method, the data from laboratory records and Lange’s Chemistry Handbook were collated into one operational database. All the 190 sets of data were used to train the artificial neural network (ANN) regression model to predict the back-surface temperature of the substrate for specific coating under given ablation condition. The mean absolutely percentage error (MAPE) of the model is 7%. Concerning the characteristics of the material database, a feature engineering method, which combines the Pearson correlation coefficient and random forest (RF) algorithm was performed to identify the main controlling factors of the service performance of the coatings.

1. Introduction

Material is an important basis for the development of human society. Each major breakthrough in industry and manufacture is inseparable from the renewal of materials. As an applied science integrating physics, chemistry, and metallurgy, the development of material has relied heavily on a large number of experiments and experience summary. With the development of life sciences, environmental sciences, aerospace, chemical engineering and materials science itself, the demand for new advanced material is unprecedented urgency[1]. The traditional trial and error method, which repeats the process of preparation, testing, and analysis, always takes lots of time and cannot meet the requirements of modern society. Comparatively, computer-assisted the design of new material is more efficient and it is also more theoretically predictive, which is an effective supplement to the traditional method.

Computer-assisted the design of new materials can be achieved in different methods. Since the 20th century, the First-Principles, Density Functional Theory, Molecular Dynamics[2-4], and other theories...
have been greatly developed, which can simulate the relationship between the composition, structure and properties of materials at different scales. However, this method still has great difficulties in cross-scale design.

Another way, which is also one of the latest and most potent ways, is artificial intelligence. It combines database technology with machine learning algorithms (such as Artificial Neural Network, Support Vector Machine, etc.) to optimize material preparation parameters and properties by using the data mining method[5,6]. Up to now, machine learning methods have gained lots of achievements in the field of material design. Oliynyk et al. trained a model for identifying Heusler and non-Heusler alloys by machine learning, with a high accuracy in 0.94, which is faster and more accurate than human experts[7]. Raccuglia et al. used machine learning methods to train a model to predict the experimental results of vanadium selenite crystallization in MOFs material synthesis [8]. The experimental results show that the success rate of predicting the product formation conditions using machine learning models is as high as 89%, which is far superior to the traditional synthetic strategy. Segler et al. used the Monte Carlo tree search and inverse synthesis analysis to increase the synthesis rate of organic compounds to 30 times of the traditional route[9].

Resin matrix composites coating is an important functional coating. Recent research shows that this kind of coating has a broad application prospect in the field of thermal protection. It is prepared by using a high char yield-based organic material as a binder and fillers such as metals and ceramics. When in high-temperature environment, the coating consumes heat by the physical and chemical changes of its components[10-14]. However, the design of resin matrix composites coating following traditional methods experiment the combination of each component constantly, this process is extremely complicated and resource-intensive. Improvement should be made to find the optimal composition ratio in a more efficient way. To authors’ best knowledge, machine learning assisted the design of functional coating has been never seen yet. In this paper, data structured method, which specific to the material design field was studied. The applicability of machine learning algorithm on the database of materials science is analyzed through validation experiments. The prior knowledge of material science is introduced into the optimization of the prediction model gained from machine learning.

2. Experimental

2.1. Data preparation

In the field of machine learning, the quality of data determines the upper limit of the model accuracy[15]. On the other hand, the materials data obtained from engineering practice come from various sources and in different formats. Therefore, it is very significant to design features reasonable and integrate all the data into one standardized database.

The ablation condition and the back-surface temperature of the substrate are always used as the main parameters to characterize the service performance for the thermal protection application of resin matrix composites coating. Besides, parameters used to describe the energy consumption ways during ablation were also chosen as features.(See Figure 1)
During the laser ablation, laser keeps heating up materials due to opto-thermal transformation and could result in the failure of the material. The resin matrix composites coating could effectively protect the substrate because of the following designed function. The organic components decompose and form a residual char layer, which can covers the surface of the substrate and helps insulate heat. The inorganic components in the coating can not only insulate heat but also take away heat energy by its physical changes when the temperature of reaches the critical point. Components in the coating may also undergo chemical reactions during ablation, which bring away heat energy and generate high-melting-point phase. The new high-melting-phase covering the substrate’s surface, insulates heat furtherly. According to the process happening during laser ablation, the following features and target listed in Table 1 and Table 2 are selected, where the most left letter in remark represents data type, c means continuous, l means logistic data. The middle letter represents data resource, e means collected from experiment while h means from Lange’s Handbook of Chemistry. The most right represents the description mode, A means using the maximum, minimum arithmetic and geometric means to describe the feature. For A+, mass weighted mean is also used.

| Table 1. | Feature design for resin matrix composite coating |
|----------|-----------------------------------------------|
| Feature name | Remark |
| Ablation laser power | c,e,- |
| ablation time | c,e,- |
| flame retardant components | l,e,- |
| number of organic components | c,e,- |
| total mass ratio of organic components | c,e,- |
| number of inorganic components | c,e,- |
| total mass ratio of inorganic components | c,e,- |
| coating thickness | c,e,- |
| reflectivity of coating | c,e,- |
| mass ratio of each inorganic component | c,e,- |
| thermal coefficient conductivity | c,h,A+ |
| melting point | c,h,A+ |
| boiling point | c,h,A+ |
| enthalpy of physical reaction which occurs during ablation | c,h,A+ |
| mass ratio of each organic component | c,e,- |
| char yield of organic component | c,h,A+ |
| number of chemical reaction | c,e,- |
| enthalpy of chemical reaction | c,h,A |
| new high-melting-point phase | l,e,- |
| melting-point of new phase | c,h,- |

| Table 2. | Target design for resin matrix composite coating |
|----------|-----------------------------------------------|
| Target name | Remark |
| back-surface temperature of the substrate after ablation | c,e,- |

2.2. Learning Stage

2.2.1. Traditional regression model. The linear model is the most primitive regression model and is intended to describe the target in a linear combination of features:
\[ \hat{y} = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_d x_d + b \]  

where \( \omega \) and \( b \) are learning objectives learned by minimizing the loss function:

\[ J(\omega) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]  

(2)

Although the linear model has advantages of strong interpretability and convenient operation, it is easy to result in over-fitting in practical applications. Introducing regularization into the loss function can alleviate this problem. When the penalty term takes the L1-norm, it is Lasso regression. When the penalty term takes the L2-norm, it is Ridge regression[16]:

\[ J(\omega) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\omega_j| \]  

(3)

\[ J(\omega) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \frac{\lambda}{2} \sum_{j=1}^{n} (\omega_j)^2 \]  

(4)

From the prospect of Bayesian[17], the target is not considered as a single value to be estimated but is assumed to be extracted from a normal distribution. At the same time, the parameter \( \omega \) of the linear regression model is also derived from the probability distribution. Based on this idea, a Bayesian regression model was developed for determining the posterior distribution of model parameters.

In addition, researchers developed the k-Nearest Neighbor algorithm (KNN) by referring to the idea that "Birds of a feather flock together". In this algorithm, the mean target value of \( k \) points closest to the unknown point used as the predict target value of that point. \( k \) is the learning objective[18].

\[ \hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \]  

(5)

2.2.2. Decision tree. Comparing with other algorithms, decision tree (DT) regression model is the only one that can simulate the human decision-making process and provides a set of if-then decision law. It divides the feature space into several units and takes the average of the target value of all samples in each unit as the output value of the unit[20]. For new data, as long as it is attributed to a unit according to the features, the corresponding output value is obtained. The partitioning of the feature space is based on minimizing the squared error between the real target value and the predict target value of all known samples in each unit:

\[ \min_{(i,d)} \left[ \min_{x \in R_1} \sum (y_i - c_i)^2 + \min_{x \in R_2} (y_i - c_2)^2 \right] \]  

(6)

2.2.3. Artificial neural network. ANN algorithm simulates the working pattern of mammalian neurons. The connection weight and activation threshold of each neuron are learning objectives (Figure 2). It is iteratively updated by back propagation and gradient descent method. Beneficial from the existence of hidden layers, ANN can regress any linearity theoretically, especially nonlinear relationships with arbitrary precision[19].

2.3. Model Evaluation

The ideal regression model is expected can not only regress known features and its target, but also predict the unknown target for known features. Better generalization ability is required. Therefore, machine learning generally divides the whole data set into two parts, one for training the model while another for testing. The error of the model on the testing set is taken as the evaluation result. In this study, MAPE is taken as the loss function.
The 11th International Conference on High-Performance Ceramics

IOP Conf. Series: Materials Science and Engineering 678 (2019) 012160 doi:10.1088/1757-899X/678/1/012160

Figure 2. Schematic for ANN: (a) schematic for single neuron; (b) structure of ANN.

Table 3. The contrast of working pattern between artificial neural and mammalian neurons

| Mammalian Neuron | Artificial Neuron |
|------------------|------------------|
| Receiving signals from upper neurons | input = \( \sum_{i=1}^{n} w_i x_i \) |
| Comparing the received signal with its action potential | compare = \( \sum_{i=1}^{n} w_i x_i - \theta_j \) |
| Transmit signals to the next layer of neurons, if the received signal is higher than the action potential | output = \( g(\text{compare}) \) |

\[
\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\] (7)

2.4. Main controlling factors identification

The intrinsic properties of materials have complex coupling effects on its performance. Finding the main controlling factors can effectively improve the efficiency of material design. This paper achieved main controlling factors identification by drawing helps from the methods of feature selection, including the Pearson correlation coefficient and RF algorithm. The Pearson correlation coefficient is used to measure the linear relationship between the two variables in statistics. The result changes from -1 to 1, where the numerical value indicates the magnitude of correlation and symbols indicate the positive and negative of correlation[21]. RF is a kind of ensemble algorithm, which integrates the process of feature selection with the training process of the DT. It estimates the importance of features by comparing the impact on model accuracy by adding random noise to specific feature [22].

However, specific to the field of materials science, due to the complex nonlinear relationship between material properties and their performance, the Pearson coefficient between feature and target cannot be relied on as the only criterion for feature selection. On the other hand, the selection results, which gained by RF algorithm separately, are lack materialistic significance. So this study coupling the Pearson and RF algorithm and perform main controlling factors step by step. For a couple of highly correlated features, retain the one with a higher score in the RF algorithm. This method called Correlation-Importance based Feature Selection (CIFS).
3. Results and discussion

3.1. Dataset
Each feature in each sample is checked against the laboratory notebook and Lange’s Chemistry Handbook. This operation aims to avoid errors caused by input mistakes. This fitter resulted in a dataset of 190 unique, complete samples. 49 features are used to describe each sample. The Pearson correlation coefficient was calculated to overview the dataset. As shown by figure 3, both x and y-axis represents the feature name, and the color of each lattice represents the Pearson coefficient between a couple of features. Collinearities are ubiquitous between features. It is caused by various reasons, including the interaction between intrinsic properties and the corresponding relationship between features.

![Figure 3. Pearson correlation coefficient thermogram](image)

![Figure 4. Histogram of target distribution: (a) histogram of the original target; (b) histogram of the target after transformation.](image)
For target, the temperature of back surface substrate after laser ablation obeys the skewed distribution. If it is used directly to train the regression model, the skewed distribution of samples would lead to the low accuracy in high temperature region[23]. In this case, the transformation of target using ln function with decreasing growth rate of dependent variables is performed to reduce the absolute value and make the trend more smoothly. After transformation, the new target is close to a normal distribution. (See Figure 4)

3.2. Model selection

Common regression algorithms were run on the dataset, and the results are shown in Figure 5. Due to the multiple collinearities, Ridge and Lasso regression models perform poorly. In addition, the dataset is on a small scale with high dimension feature space, so the predicted accuracy of the KNN and the Bayesian model is relatively low. Based on preliminary verification, ANN algorithm was selected and further model optimization was performed in it.

![Figure 5. MAPE in different regression model.](image)

3.3. Regression model base on ANN

The learning process of ANN is shown in figure 6(a). During iteration, the ANN model updates the learning objectives along the error gradient. Both the regression MAPE in the training set (blue line)
and the prediction MAPE in the testing set decreased. However, after 500 epoch, the prediction MAPE in the training set shows an upward trend, that is, the ANN model learned the change pattern of the data itself instead of the potential rules of material science. Regularization was performed to solve this problem, which means make the neurons in each layer stop working in a certain proportion during training. Through this operation, it can prevent the model from over-fitting due to over-reliance on local features[24]. As shown in Figure 6(b), the MAPE in the testing set was decreased to 7%.

3.4. Main controlling factors identification
Main controlling factors identification result based on CIFS method is shown in Table 4. Associated with this decision pattern of human, when facing complex problems, the method of controlling variables always be adopted. The rank of main controlling factors can be referred when design the resin matrix composites coating with specific service performance. Considering the 10 dimensions main controlling factors instead on whole feature space is much more convenience in decoupling, which make the precise design of coating composition possible.

| Feature name | Score     | Rank |
|--------------|-----------|------|
| total mass ratio of organic components | 0.32962 | 1 |
| ablation power | 0.222296 | 2 |
| coating thickness | 0.207954 | 3 |
| ablation time | 0.122586 | 4 |
| maximum of enthalpy of physical reaction which occurs during ablation | 0.09826 | 5 |
| mean thermal coefficient conductivity of inorganic component | 0.08587 | 6 |
| mass weighted mean of boil point | 0.08497 | 7 |
| geography mean of melting point | 0.022251 | 12 |
| minimum char yield of organic component | 0.015073 | 16 |
| maximum enthalpy of chemical reaction occurs between components in coating | 0.013513 | 19 |

4. Conclusion
A novel method with engineering significance for structurization of the database was approved. The ablation condition and macroscopic properties according to energy consumption ways of the resin matrix composites were selected as the features, and the back-surface temperature of the substrate after ablation was taken as the target. The MAPE for prediction was controlled as low as 7% by Regularized back propagation ANN. Eventually, 10 dimensions main controlling factors were screened and ranked by CIFS method. The total mass ratio of organic components was evaluated as the most significant parameter for the ablation resistance coating. The selected features could be really helpful for the further new kinds of coating design.

References
[1] Howe A A and Farrugia D C J 1999 Mater. Sci. Tech-lond. 15 15-21
[2] Haid S, Benstaali W, Abbad A, Bouadjemi B, Bentata SandAziz Z 2019 Mat. Sci. Eng. B 245 68-74
[3] Dung N T, Son T C, Cuong T M, Tan N V and Quynh T N 2019 ACTA. MATH. SCI. 39 955-70
[4] Khan M A and Herboldt M C 2011 J. Comput. Phys. 230 6563-82
[5] Service R F 1997 Science 277 474-5
[6] Wicker J G P and Copper R I 2015 Cry. Eng. Comm. 17 1927-24
[7] Oliynyk A O, Antono E, Sparks T D, Ghadbeigi L, Gaultois M W, Meredig B and Mar A 2016 Chem. Mater. 28 7324-31
[8] Raccuglia P, Elbert K C, Adler P D F, Falk C, Wenny M B, Mollo A, Zeller A, Friedler S A, Schrier J and Norquist A J 2016 *Nature* **533** 73-7
[9] Segler M H S, Preuss M and Waller M P 2018 *Nature* **555** 604-10
[10] Ma C, Ma Z, Gao L H, Liu Y B, Wu T T, Zhu Y X and Wang F C 2018 *Ceram. Int.* **44** 21374-80
[11] Ma C, Ma Z, Gao L H, Liu Y B, Wang J W, Song M Y, Wang F C and Ishida H 2018 *Compos. B* **155** 62-8
[12] Ma C, Ma Z, Gao L H, Liu Y B, Li H Z, Lv Y W and Wang F C 2019 *Mater. Res. Express.* **6** 085315
[13] Ma C, Ma Z, Gao L H, Liu Y B, Wu T T, Wang L J, Wei C H and Wang F C 2018 *Mater. Design.* **160** 1273-80
[14] Ma C, Ma Z, Gao L H and Wang F C 2017 *Chin. Opt.* **10** 249-55
[15] Xing E, Jordan M and Karp R 2001 *Proc. of 18th Int. Conf. on Machine Learning table of contents* (Berkshires) (Piscataway: IEEE) pp 601-608
[16] Diamond P and Korner R 1997 *Comput. Math. Appl.* **33** 15-32
[17] Bibi S and Stamelos I 2006 *Artificial Intelligence Applications and Innovations* (New York: Springer-Verlag)
[18] Sehgal M S B Gondal I and Dooley L 2005 *AI 2005: Advances in Artificial Intelligence* (Berlin: Springer-Verlag)
[19] Zurada J 1992 *Introduction to artificial neural system* (New York: West Publishing Co)
[20] Elder J, University R and Pregibon D 1995 *Proc. of the 1st Int. Conf. on Knowledge Discovery and Data Mining* (Montreal) (California: AAAI) pp 87-93
[21] Fisher R A 1991 *Journal of the Royal Statistical Society* **40** 462-3
[22] Dash M and Liu H 1997 *Intell. Data. Anal.* **1** 131-56
[23] Fan C C, Cui Z Z and Zhong X F 2018 *The 17th Int. Conf. on Machine Learning and Cybernetics* (Chengdu) (New York: ACM) pp 6-10
[24] Kovács G, Tóth L, Compernolle D V and Ganapathy S 2017 *Pattern. Recogn. Lett.* **100** 44-50