SCIENCE CHINA
Information Sciences

• LETTER •

January 2022, Vol. 65 119202:1–119202:3
https://doi.org/10.1007/s11432-019-2640-y

Modeling for coke quality prediction using Gaussian function and SGA

Yan YUAN[1,2*], Qilin QU[1,2], Weihua CAO[1,2] & Min WU[1,2]

[1]*School of Automation, China University of Geosciences, Wuhan* 430074*, China;*
[2]*Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan* 430074*, China*

Received 5 April 2019/Revised 28 June 2019/Accepted 13 August 2019/Published online 1 February 2021

Citation Yuan Y, Qu Q L, Cao W H, et al. Modeling for coke quality prediction using Gaussian function and SGA. Sci China Inf Sci, 2022, 65(1): 119202, https://doi.org/10.1007/s11432-019-2640-y

Dear editor,

Recently, we made some new attempts at predicting coke quality. Coke quality prediction provides important guidance for coal blending, improving the quality of coke, and reducing the cost. Prediction usually involves selection of the coal blending parameters and selection of the modeling methods. In recent years, traditional coal quality indicators such as coal impurity (ash A_d, sulfur $S_{t.d}$), coal degree (volatile fraction V_{daf}), and bond performance (bond index G, colloid layer index X, and colloid layer thickness Y) have been augmented with coal petrography data, which have made much progress [1,2].

The vitrinite reflectivity (VR) is an important index of the coal microstructure. Pusz et al. [3] showed that the average maximum VR, average minimum VR, and double VR are significantly related to the coke reactivity index (CRI) and the post-reaction strength (CSR). Dash et al. [4] combined the coal and rock indices with traditional indices, and improved the prediction results by a data-driven intelligent modeling method.

Combining the VR with traditional coal quality indices promises a new direction for the prediction and modeling of coke quality. However, as the distribution data of VR are very complex, the appropriate use of VR needs further discussion. Today, many numerical optimization and intelligent optimization algorithms have become available for scientific research [5]. The problem of applying these optimization methods to feature selection should also be resolved.

Against this background, we mainly focus on two problems: analysis of the coking process, and comprehensive extraction of the VR distribution characteristics using a Gaussian function. The well correlated high-dimensional features are detected by a selection genetic algorithm (SGA), and the features that best reflect the coke quality are selected. The effectiveness of the purposed method is verified on actual process data.

Gaussian function. The VR of coal refers to the reflectivity R (%), defining the percentage of the reflected light intensity relative to the incident light intensity on the surface of the illuminated microscopic components. The reflectivity differs among different micro-components, and the VR depends on the degree of metamorphism. The microscopic components of coal species, with their obvious regularity, are mainly assigned to the mirror-mass group. As shown in Figure 1(a), the VR values of coal are distributed in the 0%–3% range. The percentage (height) and position of the specular peak depend on the coal species. At present, VR values are determined by a coal-rock analyzer, and plotted as a histogram to guide the coal blending. The average maximum of the VR ($R_{\max}$) is adopted as the main index of the coal microscopic composition.

If all VR values are directly included as modeling parameters in the prediction model, the large dimensional space of the VR distribution data will hamper the operation. In contrast, if only the representatively high distribution values are included in the modeling, important information will be ignored. In fact, the VR curve has certain physical meanings; in particular, the coal blending effect improves as VR more closely resembles a Gaussian function. Therefore, the VR distribution can be characterized using the characteristics of a Gaussian function, namely, the peak position, peak height, and bell-shaped width. These three features can be calculated as shown below.

The Gaussian functions for a set of VR distribution datasets are defined as

$$y = R_h \times \mathrm{e}^{-((x-R_p)/R_w)^2}, \tag{1}$$

where R_h, R_p, and R_w represent the peak height, the peak position, and the half width at half maximum of the VR, respectively. Taking the logarithm of (1), we get

$$Y = Ax^2 + Bx + C, \tag{2}$$

where Y, A, B, and C are respectively given by

$$\begin{cases} Y = -\ln(y), \\ A = 1/{R_w}^2, \\ B = -2R_p/{R_w}^2, \\ C = {R_p}^2/{R_w}^2 - \ln(R_h). \end{cases} \tag{3}$$

* Corresponding author (email: yuanyan@cug.edu.cn)

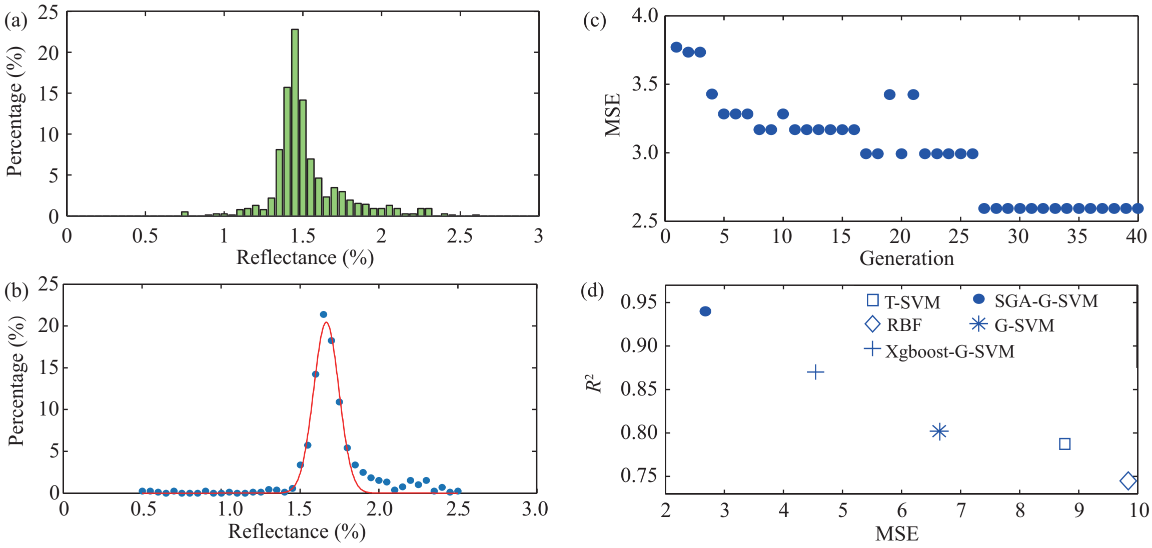

Figure 1 (Color online) (a) VR histogram of coal; (b) Gaussian curve after fitting to (a); (c) iteration process of SGA; (d) comparison of the prediction results of M_{40}.

According to the least-squares principle, the features of the VR distribution are calculated as

$$\begin{cases} R_h = \mathrm{e}^{(B^2/4A-C)}, \\ R_p = -B/2A, \\ R_w = \sqrt{1/A}. \end{cases} \tag{4}$$

Another important VR distribution feature is $R_{\max}$, denoting the average maximum VR as mentioned above. Note that $R_{\max}$ and R_p are not generally equal because the data are asymmetric. The two features can be considered as complementary features that together represent the VR distribution characteristics more completely than either feature alone. After fitting the data shown in Figure 1(a), we get the Gaussian curve shown in Figure 1(b).

Model and methodology. After merging the VR features with the traditional indexes, we obtain 10 characteristics for predicting the coke quality: A_d, $S_{t.d}$, V_{daf}, G, X, Y, R_h, R_p, R_w, and $R_{\max}$. However, some of these parameters are highly correlated. The highly correlated input features will increase the complexity of the data-driven intelligent model and reduce its performance. To reduce the negative effects of redundant features, this study proposes a feature selection method based on SGA.

The numerical optimization is performed by traditional GA, which is based on biological evolution theory. The binary string of a GA represents the corresponding decimal numeric encoding. Obviously, a decimal number cannot represent whether a feature is chosen or not, so the GA is not directly applicable to feature selection. We therefore improve the traditional GA by (1) determining whether to choose or discard a feature by checking for 1 or 0 in the corresponding position of the binary string, and (2) defining the same-parent homologous gene as the dominant gene. During crossover operations, the dominant genes are retained, and the non-dominant genes are produced by mutation. Designed in this way, the SGA can be combined with machine learning algorithms for feature selection. Here, the machine learning algorithm is a support vector machine (SVM [6]), and the feature selection mode is decided by the best individual. To realize the most accurate model, we evaluate the fitness function of the population by a precision index (mean squared error, MSE).

Simulations and results. The proposed method is verified on 108 groups of actual process samples. We select crushing strength (M_{40}) as the coke quality index. After extracting the VR features, we select 10% of the samples as the test set and use the remaining 90% samples as the training set. As 108 groups of data are relatively small for modeling, the reliability of the results is validated by 10-fold cross verification.

For comparison, we also establish a radial basis function neural network model. The features to be compared with the selective GA (SGA-G-SVM) are selected by the limit gradient lifting algorithm Xgboost (Xgboost-G-SVM). Traditional SVM parameters (T-SVM) and fusion SVM parameters (G-SVM) are also selected for further comparison.

To improve the estimation, the model performance is evaluated by both the MSE and the determination coefficient R^2. The MSE and R^2 are respectively calculated as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{5}$$

$$R^2 = \frac{\left(N\sum_{i=1}^{N}\hat{y}_i y_i - \sum_{i=1}^{N}\hat{y}_i \sum_{i=1}^{N} y_i\right)^2}{\left(N\sum_{i=1}^{N}\hat{y}_i^2 - \left(\sum_{i=1}^{N}\hat{y}_i\right)^2\right)\left(N\sum_{i=1}^{N}y_i^2 - \left(\sum_{i=1}^{N}y_i\right)^2\right)}, \tag{6}$$

where N is the scale of test set, y_i is the actual value, and $\hat{y}_i$ is the model-predicted value.

As shown in Figure 1(c), the MSE reduces as the SGA iterations proceed. The best feature selection method is chosen for the modeling. Appendix A shows the MSE of every generation and the selected features. In Figure 1(d), the MSE of SGA-G-SVM is 2.68, 3.92 lower than that of G-SVM, and 1.90 lower than that of Xgboost-G-SVM. Meanwhile, the R^2 of SGA-G-SVM is 0.94, 0.14 higher than that of G-SVM, and 0.07 higher than that of Xgboost-G-SVM. Overall, the SGA-G-SVM outperforms the other methods. This result can be attributed to extracting the features of the Gaussian function besides the feature selection of SGA. However, the computational cost of SGA-G-SVM is increased by the large number of SGA iterations. In practical applications, our model (which consumes only 0.001 s per calculation) can meet the requirements of industrial processes. Appendix B describes the datasets used in the modeling and shows the results of each method.

Conclusion. To improve the precision of coke quality, we analyzed the coal quality parameters. The characteristic VR parameters are difficult to extract from the distributed VR data, so the data were fitted by a Gaussian function. We then designed an SGA to select the most effective features, and to eliminate the strong correlation interferences between pairs of features. Finally, we verified the new data-driven coke quality prediction model in a reliable experiment on real production data. The results confirmed the effectiveness of our approach.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61773354), Hubei Provincial Natural Science Foundation (Grant No. 2015CFA010), and the 111 Project (Grant No. B17040).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Steller M, Arendt P, Kühl H. Development of coal petrography applied in technical processes at the Bergbau-Forschung/DMT during the last 50 years. Int J Coal Geol, 2006, 67: 158–170

2 Gupta S, Ye Z Z, Kim B, et al. Mineralogy and reactivity of cokes in a working blast furnace. Fuel Processing Tech, 2014, 117: 30–37

3 Pusz S, Buszko R. Reflectance parameters of cokes in relation to their reactivity index (CRI) and the strength after reaction (CSR), from coals of the Upper Silesian Coal Basin, Poland. Int J Coal Geol, 2012, 90–91: 43–49

4 Dash P S, Guha M, Chakraborty D, et al. Prediction of coke CSR from coal blend characteristics using various techniques: a comparative evaluation. Int J Coal Prep Util, 2012, 32: 169–192

5 Luo Y Q, Wang Z D, Wei G L, et al. Nonfragile l_2-l_∞ fault estimation for markovian jump 2-D systems with specified power bounds. IEEE Trans Syst Man Cybern Syst, 2020, 50: 1964–1975

6 Mia M, Dhar N R. Prediction and optimization by using SVR, RSM and GA in hard turning of tempered AISI 1060 steel under effective cooling condition. Neural Comput Appl, 2019, 31: 2349–2370