Applied Physics B (2021) 127:19
https://doi.org/10.1007/s00340-020-07556-8

Applied Physics B
Lasers and Optics

A hybrid feature selection combining wavelet transform for quantitative analysis of heat value of coal using laser-induced breakdown spectroscopy

Peng Lu[1,2] · Zhuang Zhuo[1] · Wenhao Zhang[1] · Jing Tang[3] · Yan Wang[4] · Hongli Zhou[5] · Xiaole Huang[6] · Tengfei Sun[7] · Jingqi Lu[1]

Received: 16 September 2020 / Accepted: 29 November 2020 / Published online: 15 January 2021

Abstract

A hybrid feature selection method combining with wavelet transform (WT) was proposed to analyze the heat value of coal using laser-induced breakdown spectroscopy (LIBS). The hybrid feature selection method consisted of distance correlation (DC) method and recursive feature elimination with cross-validation (RFECV) method, which combined the advantages of DC-based filter method and RFECV-based wrapper method. First, WT method was used to filter noise signal from LIBS spectra of coal samples, and the de-noised wavelet coefficients were obtained. Second, the de-noised wavelet coefficients were further eliminated by the hybrid feature selection method. Finally, the retained wavelet coefficients were used directly as input variables to establish a prediction model for heat value determination of coal. 28 powdery coal samples were used in this experiment, of which 21 were calibration set and 7 were validation set. The effectiveness of the hybrid model was studied. Compared with several other models, the proposed hybrid model showed the greatest improvement in predictive accuracy and precision, and the computing time has been greatly reduced. The experimental results demonstrated that the hybrid model can effectively reduce the calculation time and improve the performance of the model.

1 Introduction

Coal is the most abundant fossil energy on earth, which is widely used in power generation, chemical industry, metallurgy and other fields. As one of the most important properties of coal, heat value has a great impact on the work of coal-fired power station boilers [1, 2]. It is very important to realize the rapid and accurate quantitative analysis of heat value of coal. LIBS detection method has the advantages of rapid, in suit test, and simultaneous detection of multiple elements. Therefore, LIBS is a promising online measurement technology [3, 4]. In recent years, LIBS has also been widely used in the coal industry, and some works have been reported on the heat value analysis of coal [5–8].

However, coal is a complex mixture and the spectra of LIBS contain a great deal of noise information and irrelevant information. In recent years, many methods have been designed for removing noise from raw signal to reduce the influence of noise on the detection accuracy [9, 10]. Wavelet transform is a signal decomposition method, which has been applied as a powerful analysis tool. There have been many reports on the application of wavelet threshold de-noising (WTD) in LIBS [11, 13]. The wavelet threshold

✉ Jingqi Lu
Lu618@sdu.edu.cn

[1] Key Laboratory of Laser and Infrared System Integration of Ministry of Education, School of Information Science and Engineering, Shandong University, Qingdao, China

[2] School of Physics, National Demonstration Center for Experimental Physics Education, Shandong University, Jinan 250100, China

[3] Shandong Huatow Environmental Protection Technologies, Jinan 250100, Shandong, China

[4] Huadian Longkou Power Generation Co. LTD, Longkou 265700, Shandong, China

[5] Huadian Zouxian Power Generation Co. LTD, Zouxian 273522, Shandong, China

[6] State Grid JiBei Electric Power Co. LTD, Beijing 100054, China

[7] Centre for Advanced Laser Manufacturing (CALM), School of Mechanical Engineering, Shandong University of Technology, Zibo 255000, Shandong, China

de-noising method was used to remove the LIBS noise and the satisfactory de-noising effect was obtained. Wavelet coefficients contain full information about the line intensity and can be used for establishing a prediction model [14, 15]. In addition, the signal representation in the wavelet domain is sparse, which is convenient for feature selection and de-noising.

LIBS spectra contain a large amount of irrelevant information. If all the spectral information is used for the establishment of the prediction model, the robustness and accuracy of the prediction model will be reduced. Therefore, feature selection as an important preprocessing step in data mining is necessary to be used. Generally speaking, feature selection methods usually have three strategies: filter method, embedded method, and wrapper method [16, 17]. The filter method has the advantage of low computational cost for its model-independence, but its results are not always satisfactory. In contrast, the embedded and wrapper approaches determine features by the model performances, which are more effective than the filter method, but it is time-consuming. Huang et al. [18] proposed a hybrid model based on WTD and *k*-fold recursive feature elimination (RFE) to estimate the indicators of aging and hardness grades for steel using LIBS. However, there are two defects with *k*-fold RFE. One is that it is very time-consuming in the elimination process. Another defect is that the ranking of features (weight coefficient vector) is calculated without cross-validation. Yan et al. [19] proposed a hybrid method based on kernel extreme learning machine (K-ELM) and particle swarm optimization (PSO) for coal properties analysis. As we all know, genetic algorithm (GA) and PSO method are time-consuming and require parameter optimization. Zhang et al. [20] combined PCA with SVR to establish a nonlinear model of heat value, ash, and volatile content. However, PCA method only reduces dimensions, and researchers do not know which spectra are more important for the prediction model. Therefore, to further improve the efficiency and accuracy of LIBS coal properties quantitative analysis, it is necessary to propose a fast and reliable feature selection method.

In this paper, we proposed a hybrid feature selection based on distance correction (DC) and recursive feature elimination with cross-validation (RFECV). It combines the advantage of the DC-based filter method and the advantage of the RFECV-based wrapper method [21–24]. To our knowledge, this is the first time that the DC method has been used for feature selection of LIBS spectra. Both DC and RFECV methods have their own advantages and disadvantages. The DC method has the advantage of low computational cost. The RFECV method is more effective and reliable than the DC methods, but it is more time-consuming in processing high-dimensional data sets. To obtain high-precision input features quickly, we combine DC with RFECV method to form the hybrid feature selection method (DC + RFECV). The hybrid feature selection can achieve characteristics of the high efficiency of DC method and high accuracy of RFECV method.

In this study, a hybrid model based on WT and hybrid feature selection was applied to improve the accuracy and efficiency of heat value determination by LIBS. LIBS spectra were transformed to wavelet coefficients and wavelet threshold de-noising was used to filter the noisy information. Then, a hybrid selection method combining the advantages of the filter method (DC) and wrapper method (RFECV) was proposed to select the optimal features from the de-noised wavelet coefficients. Finally, the retained coefficients were used directly as input variables to establish the support vector regression (SVR) prediction model for heat value determination of coal. The results of the proposed model were compared with several other models and the root mean square error of cross-validation (RMSECV), determination coefficient of cross-validation (R^2_{cv}), root mean square error of prediction (RMSEP), determination coefficient of prediction (R^2_p), relative standard division (RSD) and average relative error (ARE) were used as criteria to evaluate performances of models.

2 Experimental setup

2.1 LIBS set-up

The experimental setup is shown in Fig. 1. The LIBS measurement system consists of a spectrometer, a pulsed laser, a digital delay generator, a precision three-dimensional platform, and a computer. The data acquisition software was written by the author, which can realize instrument control, spectrum data storage, and data analysis. The laser source is a Q-switched Nd:YAG pulse laser (Qsmart450, Quantel,

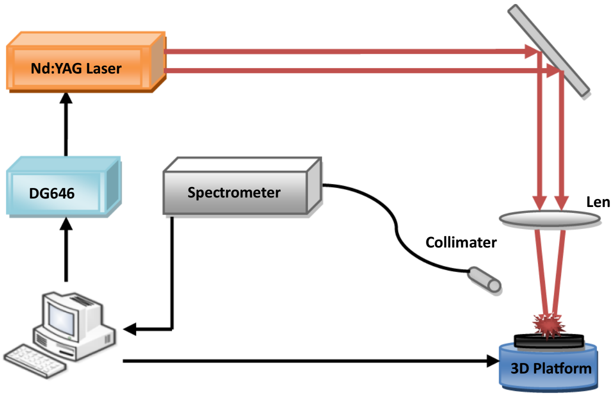

Fig. 1 Experimental setup used in this study

France), at 1064 nm, 6 ns duration, the pulse repetition rate is 0.5HZ. The pulse delay between two lasers was adjusted by the digital delay generator (DG646, SRS, USA). The eight-channel spectrometer (AvaSpec-2048, Avantes, Netherlands) can record the plasma emission spectrum in the range of 180–1060 nm. Each channel has 2048 pixels, and the spectrometer can record 16,384 variables at a time, and the spectral resolution is 0.06–0.1 nm. The laser energy and delay time were optimized to 93 mJ and 1.2 µs, respectively, and the integration time was 1.1 ms. In this setting, the signal-to-noise ratio performed best without spectral line saturation. The samples were placed on a 3-D platform. The spectrum data of 6*6 different target positions were obtained by moving the sample stage. The laser ablation was repeated 10 times at each position, and 360 sets of spectral data were collected for each sample. The angle between the collimater and the laser beam is 45 degrees. The focus position of this experiment is 1 mm below the sample surface. At the same time, to reduce the impact of the evaporated material caused by laser ablation, we installed a ventilation system.

2.2 Samples

The samples used in the experiment were provided by Huadian International Power Co. LTD. The coal was ground and then sieved to a particle size less than 200 µm. The heat value of the coal samples was obtained on an air-dried basis using the national standard laboratory test by Shandong Huatow Environmental protection technologies [25]. To reduce the instability of the LIBS spectrum, the coal powder was pressed into a 2.5 cm diameter flat disc with a pressure of 45 t and maintained for about 300 s. To test the robustness of the calibration model, the validations were randomly selected from each heat value gradient. 21 samples (C1-C21) were chosen as the calibration sets, and 7 samples (V22-V28) were selected as validation sets. The averaged measurement from 6 different locations of 36 positions on the sample was used to determine the statistics of the measurements, so as to get 6 repeated measurements for the same sample as shown in Table 1.

3 Methods

3.1 Wavelet transform and de-noising

Wavelet transform (WT) is a powerful transform analysis method [27]. A raw spectral signal can be decomposed by WT to explore the characteristics in time-domain and frequency-domain. The detailed transformation can be expressed as Eq. (1).

$$W(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \times \psi\left(\frac{t-\tau}{a}\right) dt \tag{1}$$

where$W(a,\tau)$ is the wavelet coefficient, $\psi(a,\tau)$ is the wavelet basis function, α and τ are scale and translation parameters, respectively, and $f(t)$ is the signal function to be analyzed. In the multi-resolution analysis (MRA) process, a signal can be decomposed in terms of an approximation coefficient part and several detailed coefficient parts. A raw signal can be described as

$$f(t) = \sum_{KZ} \alpha_{J,K}\varphi_{J,K}(t) + \sum_{j\leq J}\sum_{KZ} d_{j,k}\psi_{j,k}(t) \tag{2}$$

where the first term is the approximation part of the signal ($\alpha_{J,K}$ is the approximation coefficients and $\varphi_{J,K}(t)$ is the scaling function), and the second term is the detailed part of the signal ($d_{j,k}$ is the detailed coefficients and the $\psi_{j,k}(t)$ is the mother wavelet). And j is the parameter of dilation and k is the parameter of the position [28]. The wavelet coefficients contain the full information of the spectra. The approximation coefficient part is the low-frequency components of signal with high amplitude, which represents the trend of characteristics of the emission signal. And the detailed coefficient parts are the high-frequency components of signal with low amplitude, which include more noise information. Therefore, according to the difference between the real signal coefficient and the noisy signal coefficient, we can separate the real signal and noisy signal using a proper threshold.

Table 1 Coal samples used in this work

Sample no	Heat value (MJ/Kg)	Sample No	Heat value (MJ/Kg)	Sample No	Heat value (MJ/Kg)	Sample no	Heat value (MJ/Kg)
C1	16.69	C8	22.23	C15	28.91	V22	18.33
C2	17.52	C9	23.72	C16	29.50	V23	21.51
C3	18.15	C10	24.03	C17	30.18	V24	24.53
C4	19.02	C11	24.86	C18	30.73	V25	26.79
C5	19.87	C12	25.84	C19	31.29	V26	28.83
C6	20.65	C13	26.35	C20	32.86	V27	30.46
C7	21.32	C14	27.32	C21	33.60	V28	32.22

The soft threshold function of the wavelet coefficient is chosen to be used in the present work.

In this study, the de-noised wavelet coefficients $\widehat{w_s}$ instead of the reconstructed spectral intensities are proposed as input variables. The main reason is that wavelet coefficients can also represent the energy of the signal and contain the full information of the spectral line intensity. Another reason is that there is an invariably constant deviation between the reconstructed signal and raw signal after the reconstruction process. Furthermore, ignoring signal reconstruction can save calculation time. Therefore, the information in the LIBS spectra can be analyzed separately using relevant wavelet coefficients.

3.2 Distance correlation filter method

Szekely et.al first proposed the distance correlation (DC) in 2007. Unlike the classical Person correlation coefficient method (PC), DC allows for arbitrary regression relationship of predictor variables *(X)* and dependent variables *(Y)*, regardless of whether it is linear or nonlinear. The predictor variables *(X)* also permit univariate and multivariate, whether it is continuous or discrete [21]. Distance correlation takes values in [0 1], and is equal to zero only if independence holds. The distance correlation is introduced and defined in [29], $\mathcal{R}$ denotes distance correlation, which is the square root of

$$\mathcal{R}^2(X,Y) = \frac{v^2(X,Y)}{\sqrt{v^2(X)v^2(Y)}} \tag{3}$$

where the $v(X, Y)$ is the distance covariance between random vectors X and Y, and the $v(X)$ and $v(Y)$ are the distance variance. The definition of $\mathcal{R}$, $v(X)$, $v(X)$ and $v(X, Y)$ can be seen in [30]. The steps of DC filter are as follows:

Step 1. Set an initial coefficient threshold.

Step 2. Calculates the distance correlation $\mathcal{R}$ between variables x_i and target Y.

Step 3. Eliminate the variables below the threshold and only retain the variables whose $\mathcal{R}$ are greater than the threshold.

3.3 Hybrid feature selection method

The RFECV method, which is one of the wrapper approaches, determines the feature subspace during each recursion process by an estimator. In this work, the estimator used in RFECV is SVR. The estimator used in RFECV is to obtain the weight coefficients of variables and the cross-validation results of the prediction model at each recursion. At each recursion, the ranking of features is obtained according to the value of their weight coefficients, which represent the importance of variables to the model. And features with the smallest weight coefficients will be eliminated. At the end of the recursion, the optimal feature subset will be determined accord to the cross-validation results. Compared with a signal validation, cross-validation can improve the accuracy and reliability of the model evaluation. In this study, the determination coefficient of cross-validation ($\mathrm{R}^2_{\mathrm{cv}}$) is used as an evaluation index in the RFECV method.

Algorithm DC + RFECV:

Input: Traning set X_{n*p}. training label Y_n. The number of folds K.

Step 1. Calculate the $\mathcal{R}$ of variables ([$\mathcal{R}_{1,}\mathcal{R}_2, \mathcal{R}_3 \ldots\ldots \mathcal{R}_p$]), and variables with $\mathcal{R}$ higher than the pre-set threshold are retained ($X_{n*p} \rightarrow X_{n*p'}$).

Repeat:

Step 2. The calibration samples are randomly divided into K equal subsets, one subset is used as the validation set and other subsets are used as the training set. Then, each of the K subspaces takes turns as the testing set. Record the average CV results.

Step 3. In each cross-validation, the SVR model is used as an estimator and the weight coefficients of each variable are obtained. Rank the weight coefficients ([w_1 w_2 w_3 $\ldots\mathrm{w}_{\mathrm{p'}}$]).

Step 4. Eliminate the features with the smallest weight coefficients in a certain percent. Update the remaining variables.

Until: the number of variables is reduced to a small number.

At the end of the step: the feature subspace with the best CV result is selected as the final feature subspace ($X_{n*p'} \rightarrow X_{n*p''}$).

At step 4, 30% of variables with the smallest ranking will be eliminated at each recursion cycle and we find that the condition of 30% is a good option for the balance of accuracy and efficiency. At each recursion, the value of $\mathrm{R}^2_{\mathrm{cv}}$ is preserved. Therefore, the optimal feature subset can be determined by the value of $\mathrm{R}^2_{\mathrm{cv}}$ at the end of recursion. Standardization has been extensively used in the pre-processing of LIBS spectra data. In this work, standardization of wavelet coefficient data (standardized in a column scaled between [0, 1]) was applied to reduce the uncertainty of the measurement.

To make the proposed hybrid model more clear and understandable, we completely describe it in Fig. 2.

3.4 Support vector regression

Support vector regression (SVR) is a regression algorithm based on the principle of support vector machine (SVM) [32]. SVR as a multivariate calibration method has been widely used for quantitative analysis of coal property. SVR

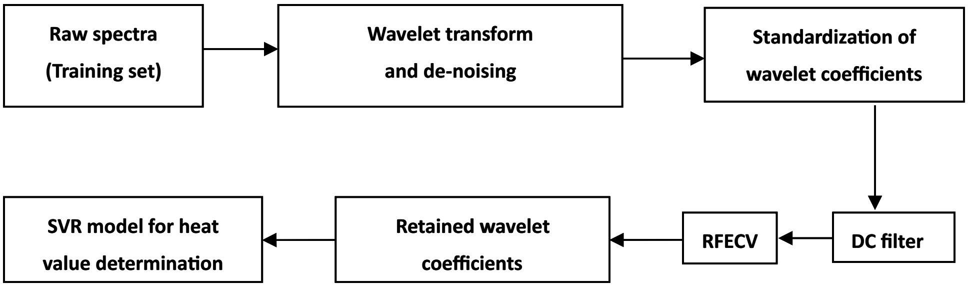

Fig. 2 Overview the proposed model

has better performance in nonlinear multivariate regression modeling and is easy to be implemented. In this work, SVR is used as an estimator of RFECV and the prediction model of heat value. SVR seeks the optimal hyper-plane to minimize the total deviation of all sample points from the hyper-plane. So, each point in the calibration sample set can be fitted into a linear model as much as possible. Its formal equation is

$$\mathrm{f(x)} = \sum_{\mathrm{i=1}}^{\mathrm{m}} \left(\mathrm{a_i} - \mathrm{\hat{a}_i}\right) \emptyset\left(\mathrm{x_i}\right)\mathrm{x} + \mathrm{b} \tag{4}$$

where the a_i are the Lagrange multipliers, the $\varnothing\left(x_i\right)$ is the kernel function, the *f(x)* is the prediction result. SVR accepts that the prediction is correct as long as *f(x)* deviates from y to a small degree. The kernel function used in this work is the 'RBF' function (Radial Basis Function). The parameter of SVR should be optimized before used to establish the calibration model. At the kernel function of RBF, the penalty parameter *c* and the parameter *g* are chosen from the exponent of 2, ($\mathrm{n} \in (-10,10)$), using the grid search method combined with RMSECV.

4 Results and discussion

In this work, leave-one-out cross-validation (LOOCV) on the calibration set is used as the criterion for the setup of the hybrid model. As described above, the hybrid model has three steps: de-noising the wavelet coefficients of LIBS spectra, selecting the optimal features subset and establishing the prediction model of heat value. The details of each intermediate step are described and the results are discussed separately in the following sections. The final results of the proposed model are compared with those of several other models.

4.1 Wavelet transform and de-noising

Wavelet basis function, decomposition level, and the threshold value are the three paramount parameters in WTD. Hence, it is necessary to determine these parameters first. The mother wavelet based on 'Daubechies' and 'Symlets' was widely used in LIBS and were explored in this study. The results obtained from the 'db4′ wavelet basis function combined with the soft threshold were somewhat better. Therefore, the 'db4′ wavelet basis function and universal threshold with soft threshold function are used to deal with all the wavelet coefficients. The decomposition level should also be optimized, and the RMSECV is selected as the optimization goal. The RMSECV results at different decomposition levels are present in Fig. 3. The values initially decrease and then increase slightly after level 5. When the DL is too small (level 1 to 3), the signal still retains a large amount of noise, and the results of RMSECV are not satisfactory. With the increase of DL value, better results can be obtained during level 4 to 6. When DL is too high (level 7 to 10), the result of RMSECV deteriorates again, because the significant compression of useful signal leads to the loss of useful information. As we can see from Fig. 3, the RMSECV value of the 5 decomposition levels is the minimum. Therefore, level 5 is selected as the final decomposition level.

The effectiveness of de-noising pretreatment is shown in Fig. 4. The black line is the raw LIBS spectrum of the C1 sample, the red line is the reconstructed spectrum of de-noised wavelet coefficients. In the range of 180–1060 nm, the real peaks have retained, and the de-noised spectrum matches well with the original spectrum at all wavelengths. The enlarged image in the range of 700–750 nm (Fig. 4) shows that the noise signal is smoothed perfectly in the

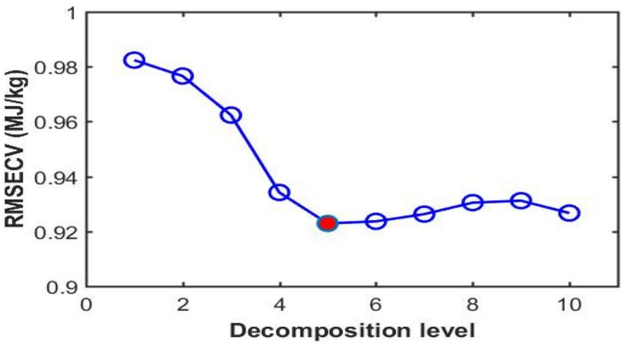

Fig. 3 RMSECV at different decomposition levels

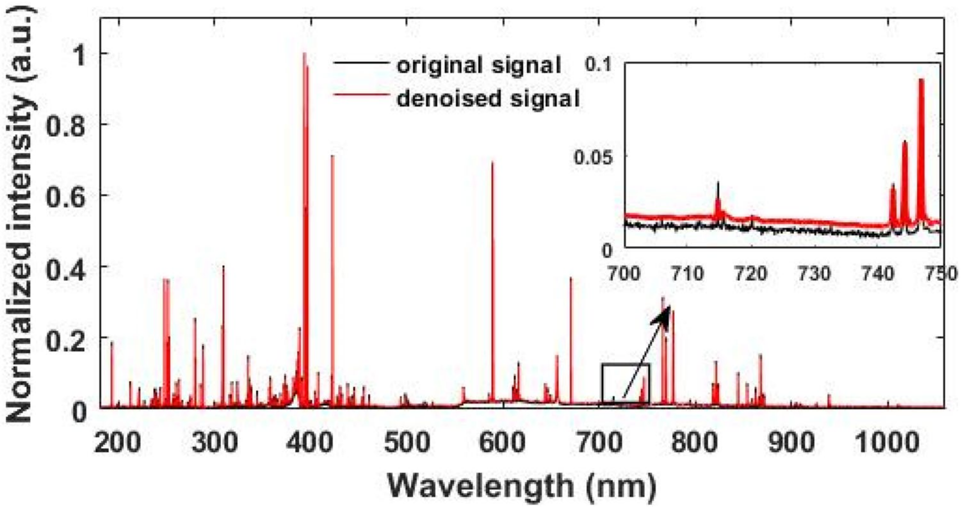

Fig. 4 Comparison of the original spectrum and de-noised spectrum of sample C1

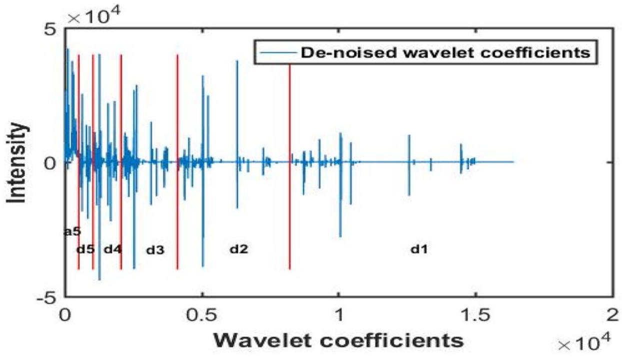

Fig. 5 The de-noised wavelet coefficients with 5 decomposition levels

reconstructed spectrum, which is inversely transformed by the de-noised wavelet coefficients.

The de-noised wavelet coefficients are obtained by 'db4′ wavelet function with 5 decomposition levels (Fig. 5). Figure 5 presents the amplitude and distribution of wavelet coefficients. It can be seen that the lager energy values are mainly distributed in the approximation coefficients of a5 band and the detailed coefficients of d5, d4, d3 bands. The coefficients of d2 and d1 bands are very sparse. The energy of the wavelet coefficient in a5, d5, d4, and d3 bands takes up about 90 percent of the total energy. As mentioned above, the coefficients of useful information usually have high amplitudes, and it indicates that the coefficients in a5, d5, d4, and d3 bands contain the primary information of the signal. In some previous work [33], only approximation coefficients were selected as input variables for modeling. However, as we can see, the detailed coefficients (d5 to d1) also contain a lot of high amplitude information (useful signal). Therefore, we use the whole de-noised wavelet coefficients as input variables for the next process of feature selection and modeling.

Table 2 The number of features, RMSECV and R^2_{CV} with different thresholds

Threshold	Number of features	RMSECV (MJ/kg)	R^2_{CV}
0.65	2738	0.7912	0.9675
0.7	1252	0.7817	0.9702
0.75	817	0.7538	0.9768
0.8	274	0.7866	0.9690
0.85	59	1.453	0.9482

4.2 Feature selection using DC + RFECV method

In this section, the DC + RFECV method is used to select features of wavelet coefficients. First, features with low DC values are eliminated by the DC method. As described, the value of the DC ranges from 0 to 1, zero indicates complete independence between the variable and the target, and 1 indicates complete correlation. Therefore, the filter threshold should be determined first. To obtain the optimal threshold, the value of the threshold is determined based on leave-one-out cross-validation of the calibration model, and the RMSECV and R^2_{CV} are used as criteria. The number of features and the values of the RMSECV and R^2_{CV} at different thresholds are presented in Table 2. The threshold value ranges from 0.65 to 0.85, and the interval value is 0.05. From Table 2, we can clearly see that when the threshold was 0.75, the best results of LOOCV are obtained, with the RMSECV of 0.7538 (MJ/kg) and R^2_{CV} of 0.9768. So the threshold is set to 0.75. That is, only 5% of variables (817) which are most relevant to the target variable have been retained. As we can see, as the number of features continues to decrease, the model performs better. That is because too many variables could worsen the accuracy of the SVR model, which may lead to over-fitting when establishing the SVR model.

The DC method improves model performance by eliminating irrelevant features. After selection by DC method, the retained wavelet coefficients are only 817. The results show that the DC method not only reduces the number of features but also improves the performance of the SVR model.

The selected wavelet coefficients are further eliminated by the RFECV method. There are 817 wavelet coefficients retained after selection by DC method. Therefore, RFECV algorithm only needs 12 recursion times, and the number of features is reduced from 817 to 133 (Fig. 6 black line). In each RFECV recursive process, the SVR model is established with different feature subsets, and CV results are obtained. The effects of the WT + DC + RFECV method and the WT + RFECV method were compared. As shown in Fig. 6, for WT + RFECV model (red line), the optimal number of features is 161, and the optimal value of R^2_{cv} is 0.9861. The trend of the red line increases slowly with the decreasing of the variable number, then, the curve goes down quickly when the variable number decreases to 55. The black line is the CV result curve of the WT + DC + RFECV model. Compared with the red line, the curve of the WT + DC + RFECV model is very different. During the first two points of the black curve, the curve rises rapidly. This is due to the change from using the DC method to using the RFECV method, which significantly improves the accuracy of feature selection. After the selection by the DC method, the CV result is further improved by the RFECV method. As we can see in Fig. 6, the optimal CV results of the two selection methods are very close (0.9868/0.9861), but the size of the optimal feature subset of the WT + DC + RFECV model is smaller (133/161). The final optimal feature subset is determined by the RFECV algorithm in both selection methods, which is the reason for the very close CV results of them. Due to the application of the RFECV method, both of the two selection methods can achieve great improvement in the prediction model. Moreover, model performance is further improved by the RFECV method based on the improvement of DC method. It indicates that RFECV method is more effective than the DC selection method, and its performance is stable when combining with the DC method.

To further verify the precision of the feature selection of the WT + DC + RFECV method, the reconstructed spectrum inverse transformed by retained wavelet coefficients (the removed wavelet coefficients are replaced with zero) is presented in Fig. 7. The overall trend of the reconstructed spectrum is basically consistent with that of the original spectrum. As we can see that the spectral lines of C(I) 193.1 nm, C(I) 247.8 nm, Si(i) 251.4 nm, Si(I) 288.1 nm, Al(I) 309.2 nm, Ca(I) 422.7 nm, Na(I) 588.9 nm, H(I) 656.2 nm, N(I) 746.8 nm, O(I) 777.19 nm and other characteristic lines can be identified both in original spectrum (Fig. 7a) and reconstruction spectrum (Fig. 7b). As we all know, the heat value mainly results from the combustion of organic elements (C, H, O, N). In addition, the ash content also absorbs part of the heat during the combustion, which is composed of various mineral elements, such as Ca, Si, Na, and Fe. As shown in Fig. 7b, the characteristic lines related to heat value are still retained in the reconstructed spectrum, which proved the effectiveness of the proposed feature selection method.

In addition, there are also many differences between Fig. 7a and b. Compared with the original spectrum, many irrelevant lines are eliminated by the WT + DC + RFECV method in Fig. 7b. Due to the reconstructed spectrum using

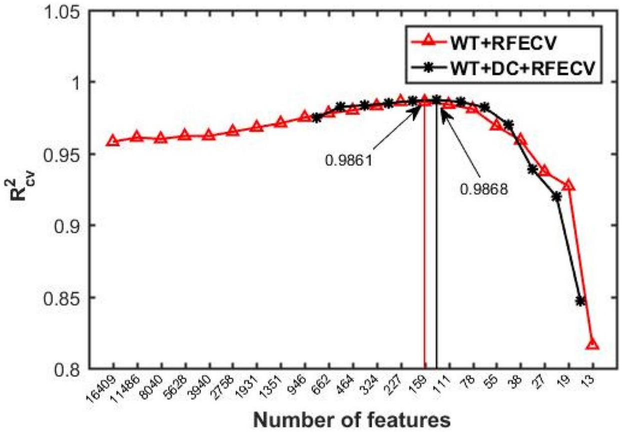

Fig. 6 Performance of model with different number of features under the WT + RFECV method and the WT + DC + RFECV method

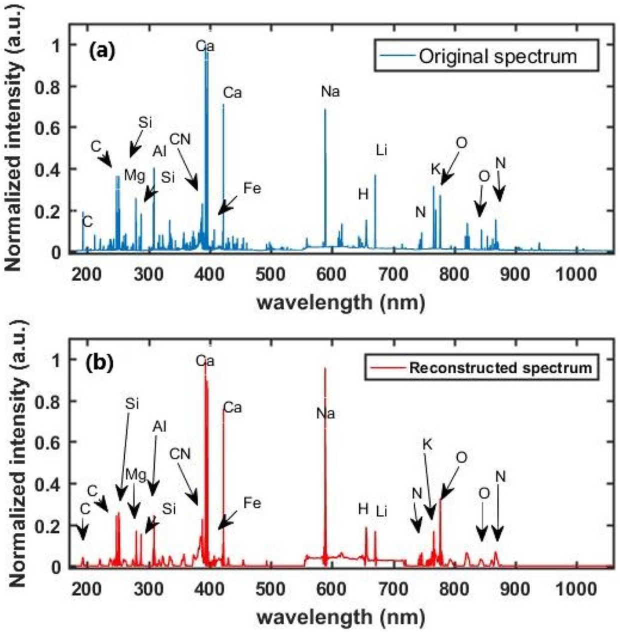

Fig. 7 **a** The original spectrum and **b** the reconstruction spectrum with retained 130 wavelet coefficients by WT + DC + RFECV method

only 133 coefficients, the normalized intensity of spectral lines in Fig. 7a ($I_{C\ 193.1}=0.19$, $I_{C\ 247.8}=0.365$, $I_{Si\ 288.1}=0.182$, $I_{Al\ 309.2}=0.403$, $I_{Ca\ 422.7}=0.712$, $I_{Na\ 588.9}=0.687$, $I_{H\ 656.2}=0.152$, $I_{N\ 746.8}=0.091$, $I_{O\ 777.19}=0.252$) are different from that in Fig. 7b ($I_{C\ 193.1}=0.0288$, $I_{C\ 247.8}=0.176$, $I_{Si\ 288.1}=0.155$, $I_{Al\ 309.2}=0.246$, $I_{Ca\ 422.7}=0.761$, $I_{Na\ 588.9}=0.958$, $I_{H\ 656.2}=0.185$, $I_{N\ 746.8}=0.061$, $I_{O\ 777.19}=0.326$). Moreover, the relative intensity of the spectral line is changed in the reconstruction process. Take C(I) 247.8 nm, C(I) 193.1 nm and Al(I) 309.2 nm, Na(I) 588.9 nm, for example, the relative intensity ratios of the spectral lines are 1.92 ($I_{C\ 247.8}/I_{C\ 193.1}$) and 1.7 ($I_{Na\ 588.9}/I_{Al\ 309.2.}$), respectively, in the original spectrum, while, which are 6.08 and 3.89 in the reconstructed spectrum. Therefore, in this study, the wavelet coefficients instead of the reconstructed spectrum are used as input variables for modeling. The using of wavelet coefficients rather than the reconstructed spectral intensity for modeling can avoid the changes of spectral intensity caused by the spectral reconstruction, and also can reduce the calculation time. As we can see from Fig. 7b, molecular CN emission lines can be identified. In some previous works, researchers found that the molecular CN emission lines could also be measured via carbon atomic emission from plasma under atmospheric conditions, and the combination of carbon atomic and molecular emissions can improve the accuracy of quantitative analysis of carbon in coal [34]. The composition and molecular structure of coal is complicated, Fig. 7b reveals that not only characteristic lines but also some molecular lines or background contribute to the model of heat value.

4.3 Evaluation the performance of the hybrid model

To verify the performance of the proposed hybrid model (WT + DC + RFECV), the results obtained by WT + DC + RFECV model are compared with several other models, including the original signal model, WT, WT + PC (Person correlation coefficient), WT + DC, WT + RFECV models. The WT + PC model refers to the previous work of other researchers [35]. As mentioned above, the WT model refers to the wavelet coefficients de-noised by the universal threshold method, and the removed noise wavelet coefficients are replaced with zero, so, the number of wavelet coefficients is unchanged (16,409). Figure 8 shows the LOOCV results obtained by different models. Compared with the feature selection methods, the improvement effect of the WT model is not as good as the feature selection models. The LOOCV results are further improved by feature selection models based on WT. This indicates that those feature selection models are very effective to improve the model performances. As shown in Fig. 8, the WT + DC + RFECV model performed best, the RMSECV, R^2_{cv} are 0.5768 (MJ/kg) and 0.9868, respectively. The results of the WT + DC + RFECV model are very close to those of the WT + RFECV model. In the WT + DC + RFECV model, the RFECV method plays a more important role in feature selection and determines the final optimal feature subset. And that is why its performance is very close to that of WT + RFECV model. As shown in Fig. 8, the performances of two RFECV models (WT + RFECV, WT + DC + RFECV) are better than WT, WT + PC and WT + DC models, indicating that the wrapper methods are more superior to filter methods for extraction of useful information from coal LIBS spectra. Those filter methods which perform independently of model establishment do not consider the effects of the selected features on the performances of the model. Features selected by those filter methods may not necessary for the SVR model. As a result, the performances of those filer methods are worse than those wrapper methods.

Validation samples are used to test the effectiveness and robustness of those models, as shown in Fig. 9. We can see that the WT + DC + RFECV model has the best performances. The accuracy and precision of prediction results have been improved using the hybrid model. Compared to the original model, the RMSEP decreases from 0.9246 MJ/

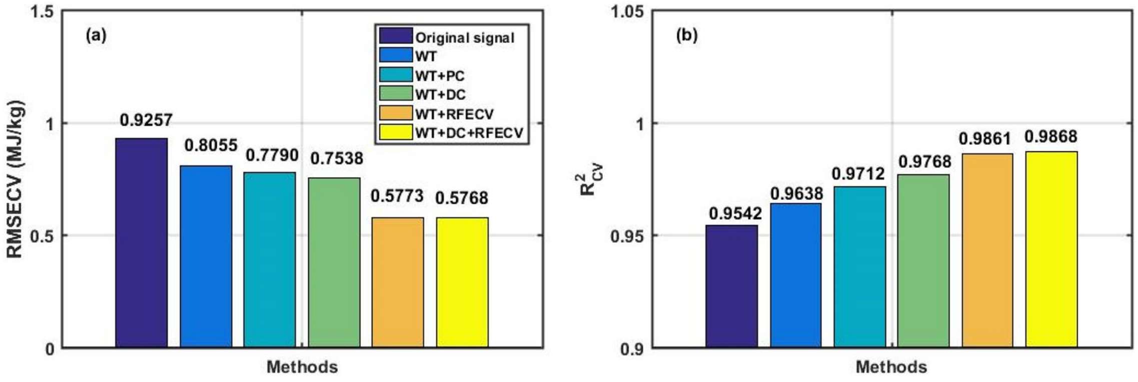

Fig. 8 The results of different calibration models coupled with different methods

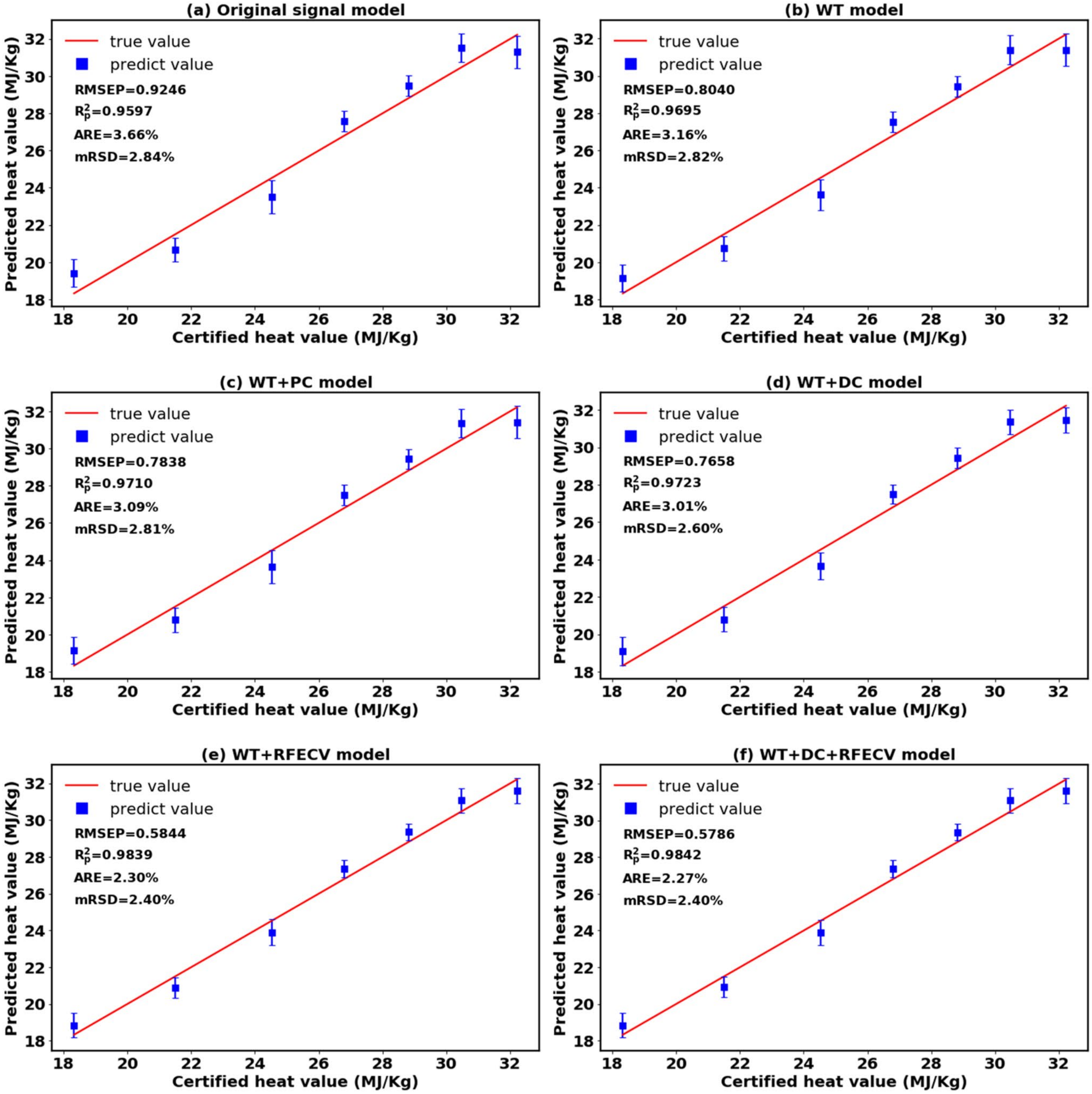

Fig. 9 The predicted results obtained by different models. **a** Original signal model, **b** WT model, **c** WT + PC model, **d** WT + DC model, **e** WT + RFECV model, **f** WT + DC + RFECV model. The mRSD means the average RSD of 7 validation samples

Kg to 0.5786 MJ/Kg, the ARE decreases from 3.66% to 2.27% (approximately 58% reduction), and the R_p^2 increases from 0.9597 to 0.9842. In addition, the mRSD of the predicted values by the hybrid model is smaller than any other model, which indicates that the precision of validation has been improved. Therefore, this hybrid model indeed improved the accuracy and precision of LIBS quantitative measurement on coal heat value.

Furthermore, the running time and the number of selected features of different feature selection methods are presented in Fig. 10. The WT + PC and WT + DC methods are faster than the methods based on RFECV method (WT + RFECV, WT + DC + RFECV) because filter method is independent of the learning algorithm. The processing time for the feature selections of RFECV method is much longer than other methods. If the full wavelet coefficients are used directly as the input variable for RFECV method, it takes more

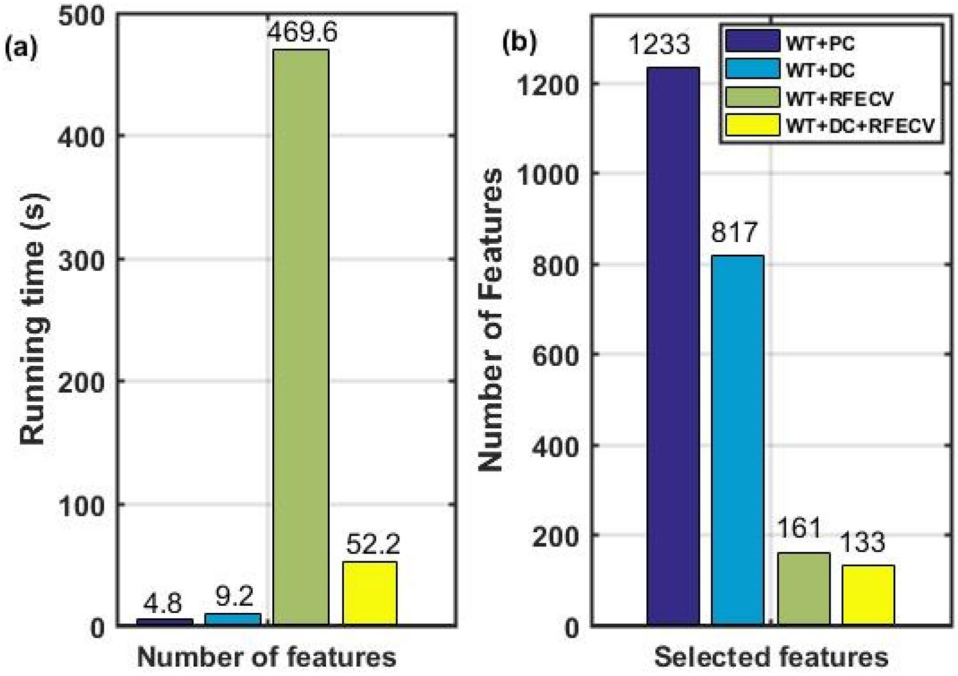

Fig. 10 The running time and selected features of different feature selection methods

recursion cycles to reduce the number of features to a small number. However, as for the WT + DC + RFECV method, after selection by DC, only 5% retained variables are used as input set for further selection by the RFECV method. The computer processing time of the WT + DC + RFECV method is only about 11% of the WT + RFECV method costs. As we can see from Fig. 10, the WT + DC + RFECV method is much faster than the WT + RFECV method (52.2 s/469.6 s). Moreover, the performances of WT + RFECV method and WT + DC + RFECV method are very close. Hence, the proposed method (WT + DC + RFECV) is a rapid and precise method and shows the greatest improvement of model performance over other methods.

5 Conclusion

In this study, a hybrid model was proposed to reduce the computation time and improve the quantitative measurement of the heat value of coal. The proposed hybrid model consisted of the wavelet transform and a hybrid feature selection method based on the DC method and RFECV method, which combined the advantages of DC-based filter method and RFECV-based wrapper method. The interference of noisy information was firstly de-noised by the WT method, and the de-noised wavelet coefficients were selected by DC method, and then the optimal feature subset was determined by the RFECV method. Finally, the retained wavelet coefficients were used as input variables to establish SVR model for heat value determination. Compared to the original model, the RMSEP decreases from 0.9246 MJ/Kg to 0.5786 MJ/Kg, the ARE decreases from 3.66% to 2.27%, and the R_p^2 increases from 0.9597 to 0.9842. In addition, the mRSD of the predicted values by the hybrid model are smaller than any other models and computer processing time of the WT + DC + RFECV method is only about 11% of the WT + RFECV method costs. Therefore, the proposed WT + DC + RFECV model is a rapid and precise tool of data processing in LIBS quantitative analysis and has broad application prospects.

Acknowledgements This work was supported by the Shandong province major science and technology innovation project (No. 2018CXGC0806).

References

1. R. Yan, H.J. Zhu, C.G. Zheng, M.H. Xu, Energy. **27**, 485 (2002)
2. T.B. Yuan, Z. Wang, S.L. Lui, Y.T. Fu, Z. Li, J.M. Liu, W.D. Ni, J. Anal. At. Spectrom. **28**, 1045 (2013)
3. T.L. Zhang, H.S. Tang, H. Li, J. Chemom. **32**(11), 2983 (2018)
4. Y.Q. Zhang, C. Sun, L. Gao, Z.Q. Yue, S. Shabbir, W.J. Xu, M.T. Wu, J. Yu, Spectrochim. Acta B. At. Spectrosc. **166**, 105802 (2020)
5. M.R. Dong, L.P. Wei, J.D. Lu, W.B. Li, S.Z. Lu, S.S. Li, C.Y. Liu, J.H. Yao, J. Anal. At. Spectrom. **34**, 480 (2019)
6. J. Feng, Z. Wang, Z. Li, W.D. Ni, Spectrochim. Acta B. At. Spectrosc. **88**, 180 (2013)
7. W.H. Zhang, Z. Zhuo, P. Lu, J. Tang, H.L. Tang, J.Q. Lu, T. Xing, Y. Wang, J. Anal. Atom Spectrom. **35**, 1621 (2020)
8. Z.Y. Hou, Z. Wang, J. Anal. At. Spectrom. **31**, 722 (2016)
9. A. Bogaerts, Z.Y. Chen, D. Bleiner, J. Anal. Atom Spectrom. **21**, 384 (2016)
10. P.K. Diwakar, S.S. Harilal, J.R. Freeman, Spectrochim. Acta Part B. **88**, 65 (2013)

11. P. Lu, Z. Zhuo, W.H. Zhang, J. Tang, H.L. Tang, J.Q. Lu, Appl。Optics. 59(22), 6443 (2020).
12. R.C. Wiens, S. Maurice, J. Lasue, O. Forni, D. Vaniman, Spectrochim. Acta B. At. Spectrosc. **82**, 1 (2013)
13. X.H. Zou, L.B. Guo, M. Shen, X.Y. Li, Z.Q. Hao, Q.D. Zeng, Y.F. Lu, Z.M. Wang, X.Y. Zeng, Opt. Express. **22**(9), 10233 (2014)
14. T.B. Yuan, Z. Wang, Z. Li, W.D. Ni, J.M. Liu, Spectrochim. Acta B. At. Spectrosc. **807**, 29 (2014)
15. V.D. Hoang, Tr Trend. Anal. Chem. **62**, 144 (2014)
16. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Mach. Learn. **46**, 389 (2002)
17. M.J.C. Pontes, J. Cortez, R.K.H. Galvao, C. Pasquini, M.C.U. Araujo, R.M. Coelho, M.K. Chiba, M.F.D. Abreu, B.E. Madari, Anal. Chim. Acta. **642**, 12 (2009)
18. J.W Huang, M.D, S.Z Lu, Y.S Yu, C.Y Liu, J.H. Yoo, J.D Lu, Analyst. 144, 3736 (2019).
19. C.H. Yan, J. Liang, M.J. Zhao, X. Zhang, T.L. Zhang, H. Li, Anal. Chim. Acta. **1080**, 35 (2019)
20. L. Zhang, Y. Gong, Y. Li, X.Wang, et al. Spectrochim. Acta B At. Spectrosc. 113(3), 167 (2015).
21. R.Z Li, W Zhong, L.P. Zhu, J. Anal. Atom Spectrom. 107, 1129 (2012).
22. G.J. Szekely, M.L. Rizzo, Ann. Appl. Stat. **3**, 1233 (2009)
23. S.Z. Lu, S. Shen, J.W. Huang, M.R. Dong, J.D. Lu, Spectrochim. Acta B. At. Spectrosc. **150**, 49 (2018)
24. F.J. Duan, X. Fu, J.J. Jiang, L. Ma, C. Zhang, Spectrochim. Acta B. At. Spectrosc. **143**, 12 (2018)
25. Standardization Administration of the People's Republic of China (SAC), GB/T 213–2008, Determination of Calorific Value of Coal, China, 2008.
26. S.G. Mallat, S. Zhong, IEEE Trans. Pattern. Anal. Mach. Intell. **14**, 710 (1992)
27. P. Angel, C. Morris, Comput. Vis. Image Und. **80**, 267 (2000)
28. D.L. Donoho, M. Nussbaum, J. Complex. **6**, 290 (1990)
29. G.J. Szekely, M.L. Rizzo, N.K. Bakirov, Ann. Statist. **35**(6), 2769 (2007)
30. G.J. Szekely, M.L. Rizzo, Journal of Multivariate Analysis. **117**, 193 (2013)
31. N.C. Dingari, I. Barman, A.K. Myakalwar, S.P. Tewari, M.K. Gundawar, Anal. Chem. **84**, 2686 (2012)
32. I. Guyon, J. Weston, S. Barnhill, V. Vapni, Mach. Learn. **46**, 389 (2002)
33. C.H. Yan, T.L. Zhang, Y.Q. Sun, H.S. Tang, H. Li, Spectrochim. Acta B. At. Spectrosc. **154**, 75 (2019)
34. S.C. Yao, Y.L. Shen, K.J. Yin, G. Pan, J.D. Lu, Energy Fuels. **29**, 1257 (2015)
35. C.H. Yan, J. Qi, J. Liang, T.L. Zhang, H. Li, J. Anal. At. Spectrom. **33**, 2089 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.