JAAS

Accepted Manuscript

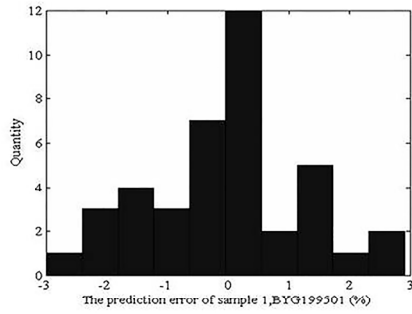Frequency histogram of error variable e_k

Different schemes of the segmented weighting function of WLS-SVM and RLS-SVM
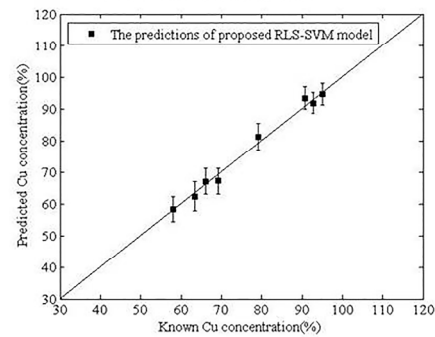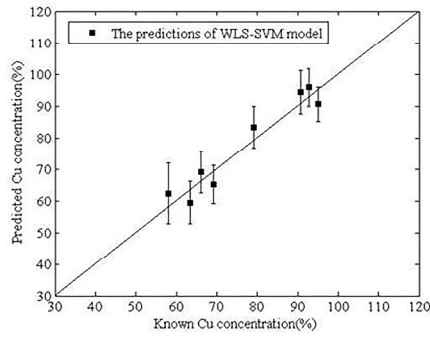
Prediction results of different regression models

Graphical Abstract
128x96mm (600 x 600 DPI)
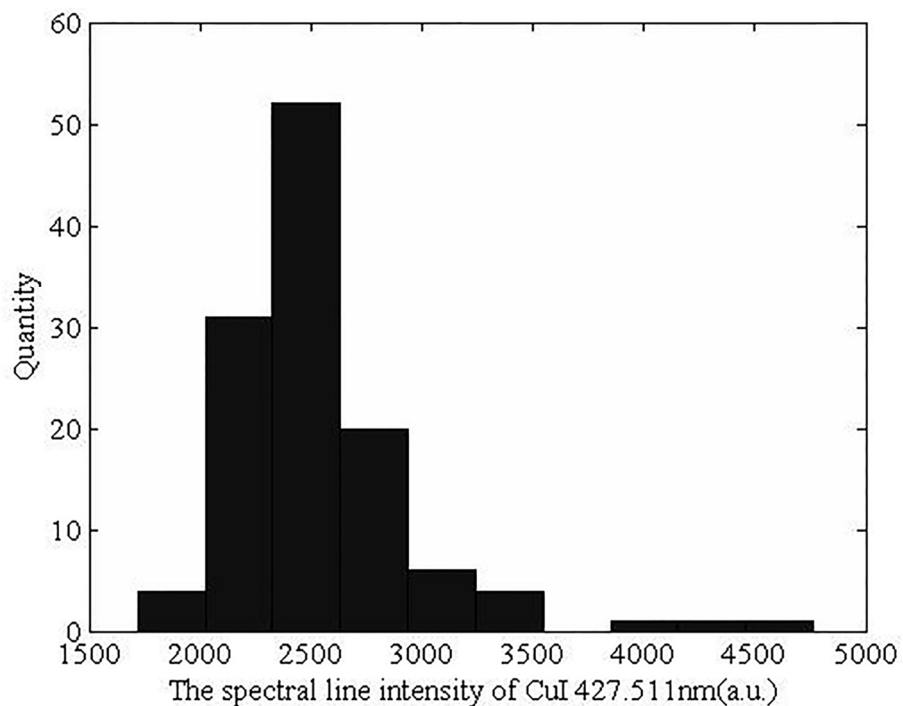
Fig. 1 Data fluctuation of spectral intensities in 120 times of LIBS measurements at a same position of the surface of 6# sample (BYG199506)
(a) Frequency histogram of spectral intensities in 120 repetitions of measurement
62x46mm (600 x 600 DPI)
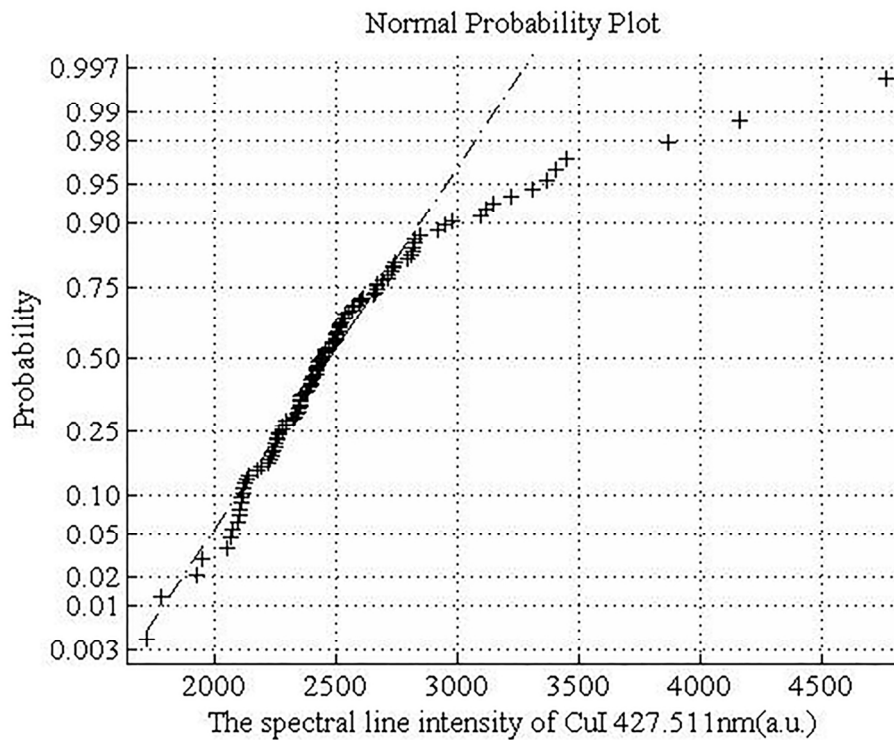
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Fig.1 Data fluctuation of spectral intensities in 120 times of LIBS measurements at a same position of the surface of 6# sample (BYG199506)
(b)The normal probability plot of the intensities of multiple measurements
62x46mm (600 x 600 DPI)

(a) The weighting function of WLS-SVM

(b) The weighting function of RLS-SVM
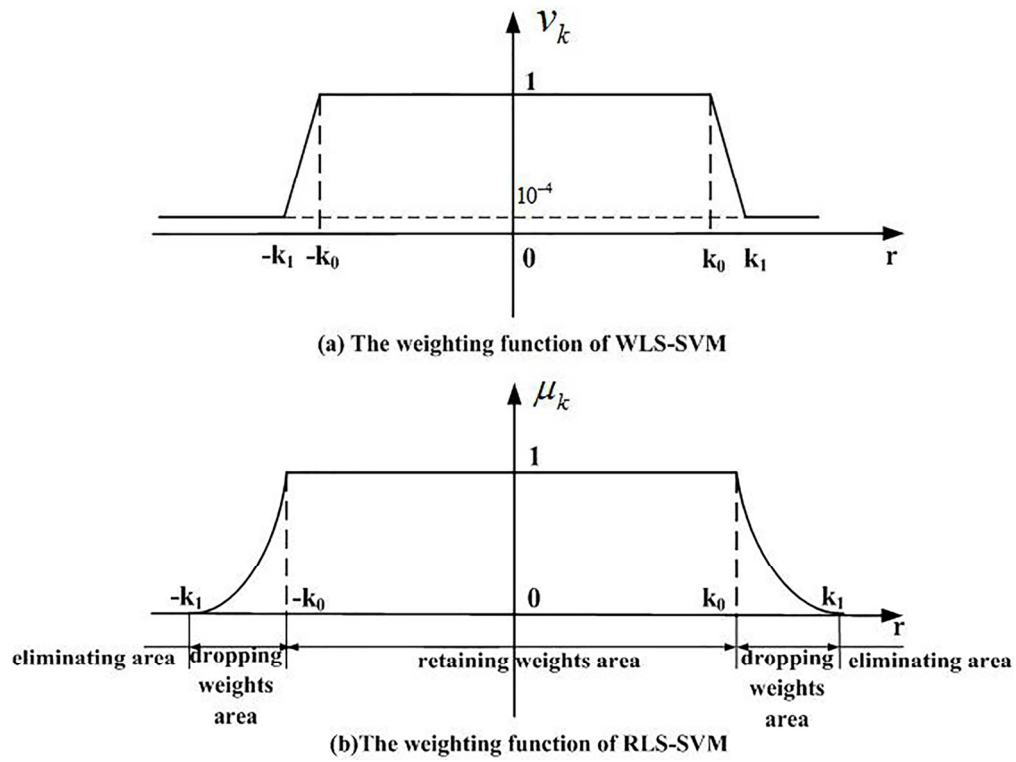
Fig. 2 Different schemes of the segmented weighting function of WLS-SVM and RLS-SVM
64x49mm (600 x 600 DPI)
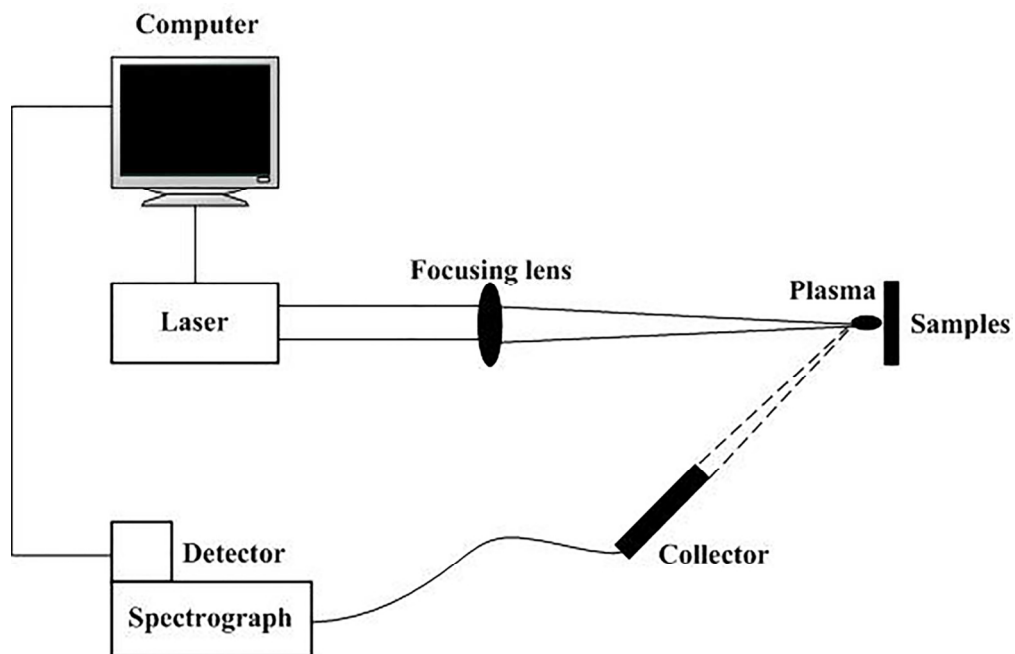
Fig. 3 Schematic experimental setup of LIBS measurement
54x35mm (600 x 600 DPI)
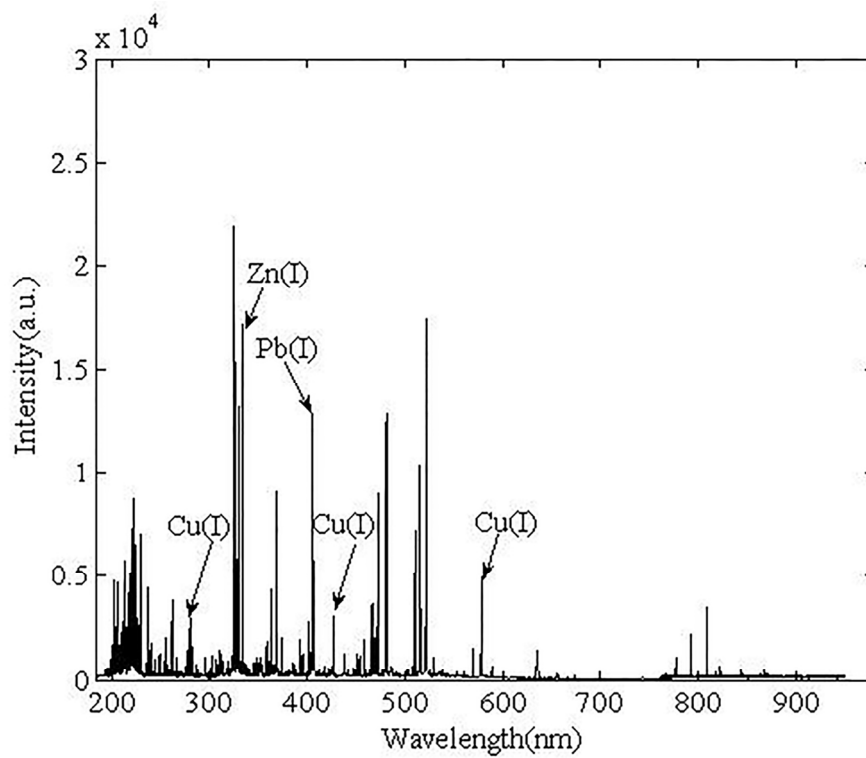
Fig. 4 A typical LIBS spectrum of brass samples
63x48mm (600 x 600 DPI)
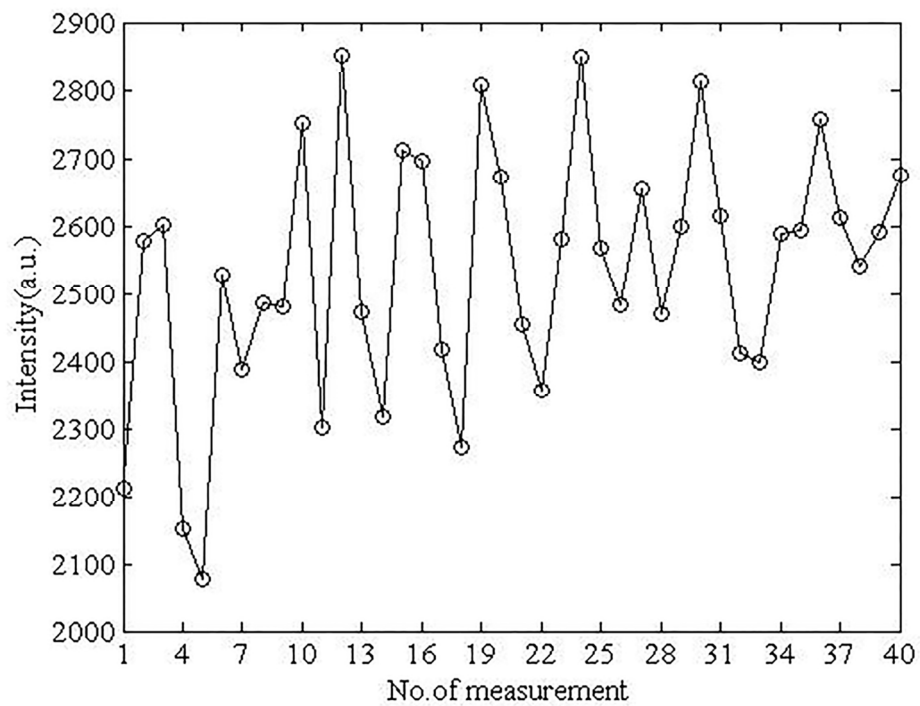
Fig. 5 Pulse-to-pulse spectra fluctuations of the intensity (BYG119504 427.511nm)
62x46mm (600 x 600 DPI)
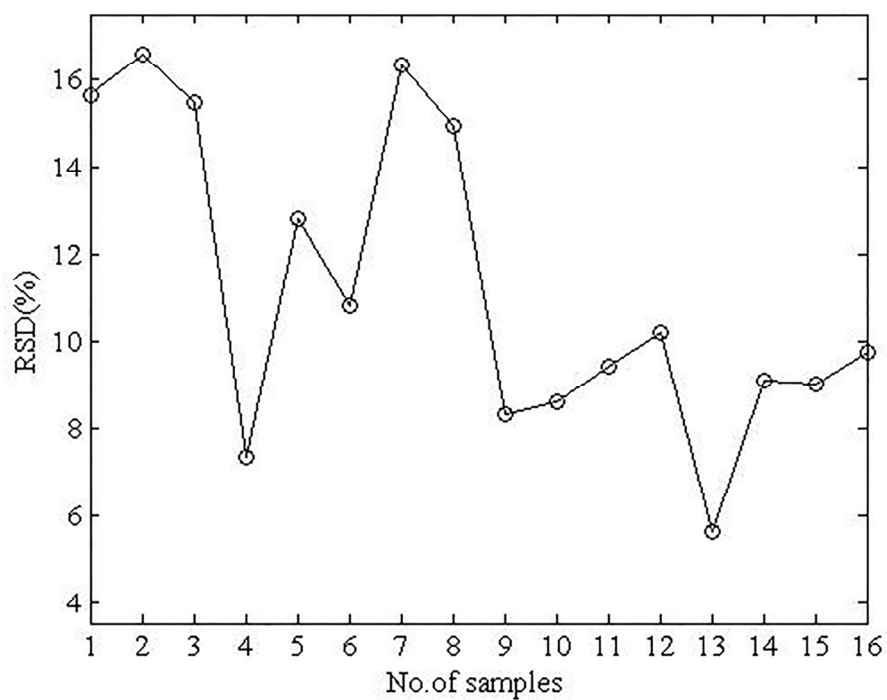
Fig. 6 RSD values of the characteristic line intensity at 427.511nm of Cu normalized for all 16 brass samples
62x46mm (600 x 600 DPI)
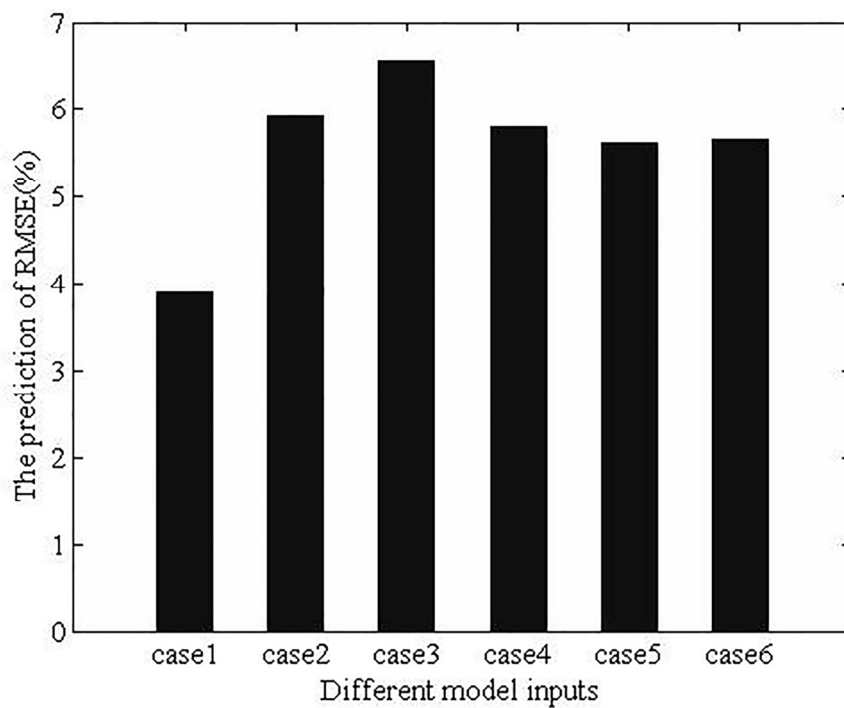
Fig. 7 The RSD values of predictions by different input variables
62x46mm (600 x 600 DPI)
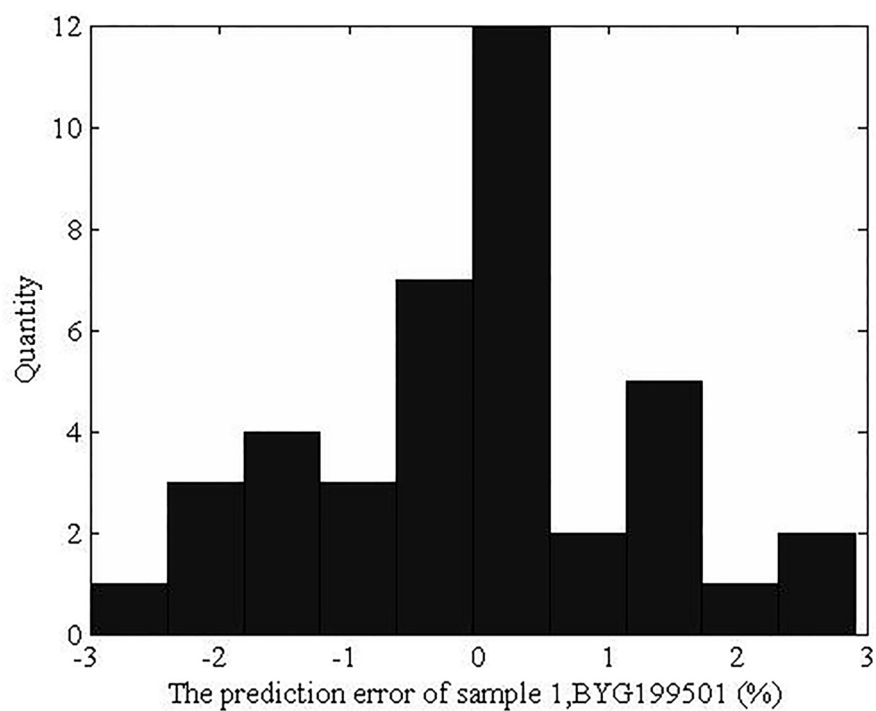
Fig. 8 The frequency histogram of error variable and normal probability plot
(a) Frequency histogram of error variable e_k
62x46mm (600 x 600 DPI)
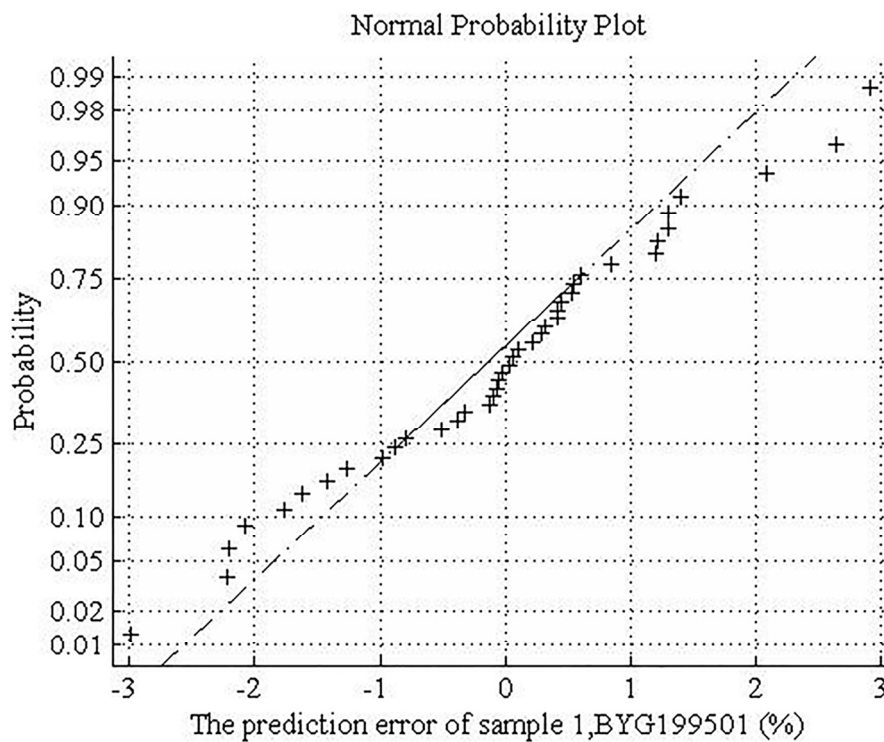
Fig. 8 The frequency histogram of error variable and normal probability plot
(b)The normal probability plot of error variable e_k
62x46mm (600 x 600 DPI)
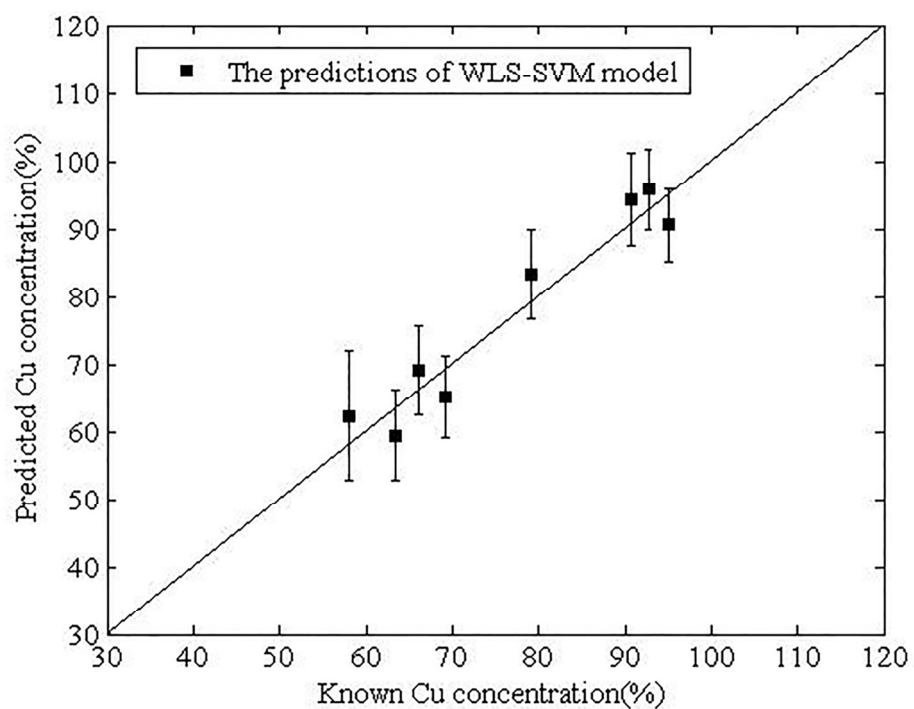
Fig. 9 Prediction results of different regression models
(a)Prediction results of WLS-SVM
62x46mm (600 x 600 DPI)
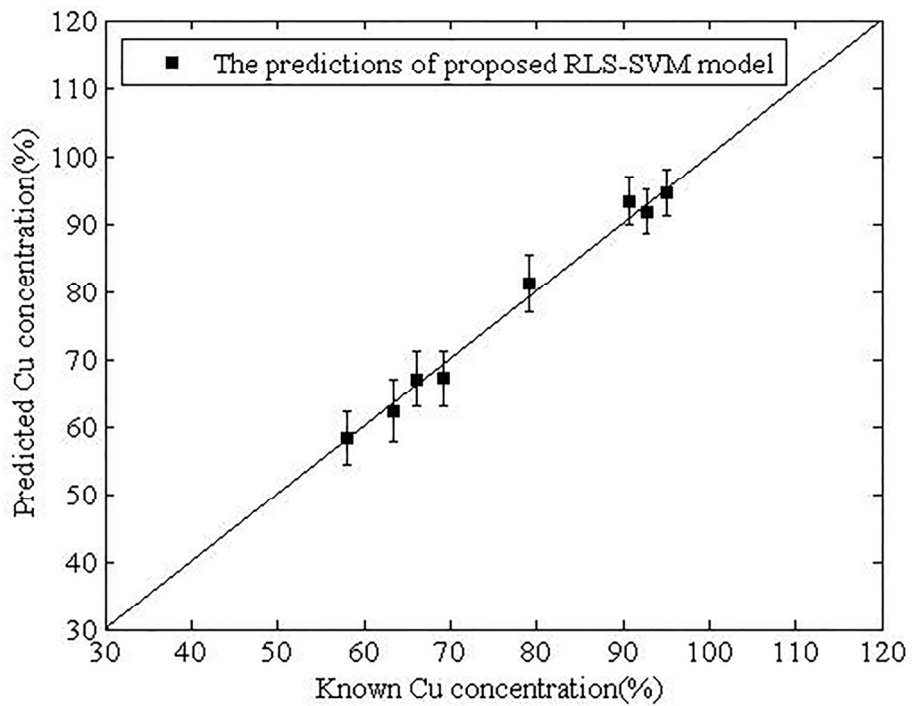
Fig. 9 Prediction results of different regression models
(b) Prediction results of proposed RLS-SVM
62x46mm (600 x 600 DPI)
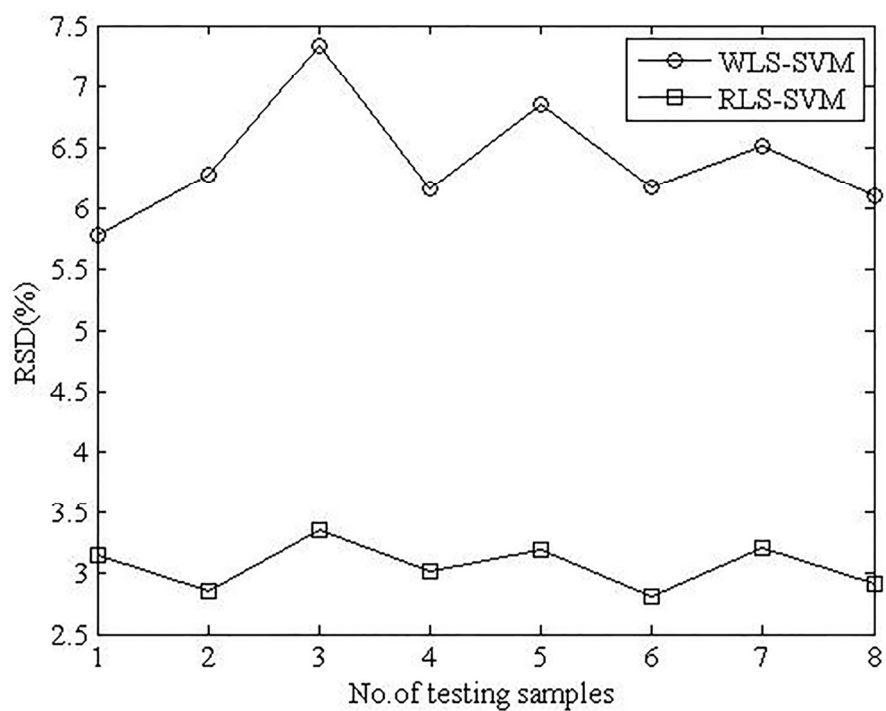
Fig .10   RSD values of predictions of different regression models
62x46mm (600 x 600 DPI)
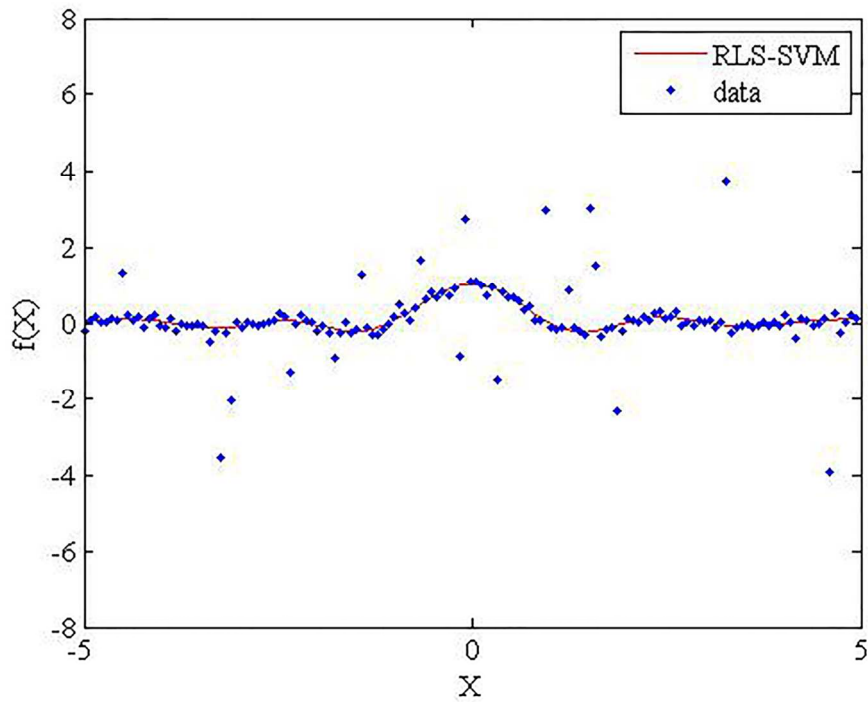
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(a) 15% outliers(143 samples)
62x46mm (600 x 600 DPI)
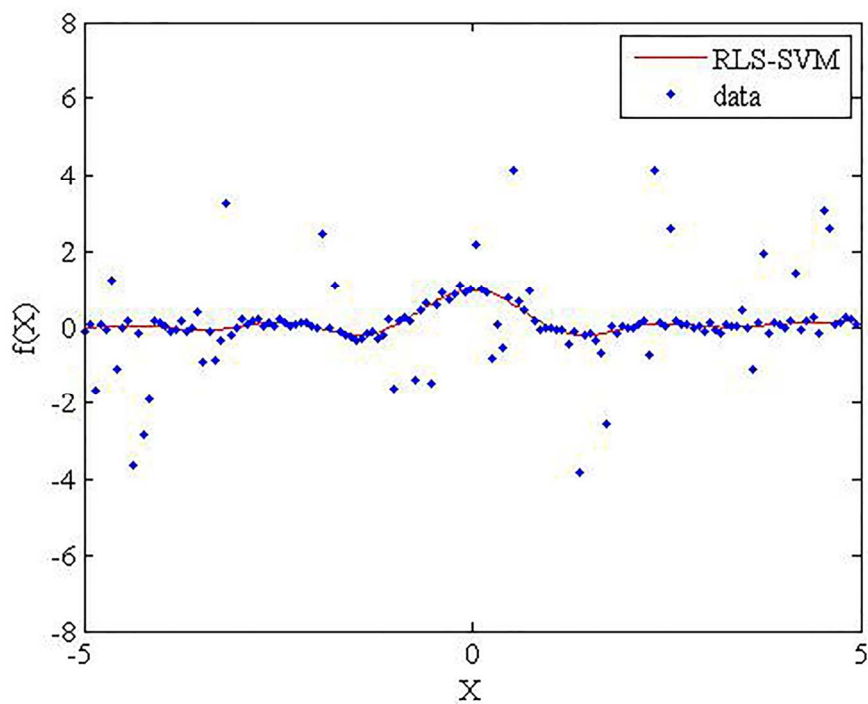
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(b)30% outliers(143 samples)
62x46mm (600 x 600 DPI)
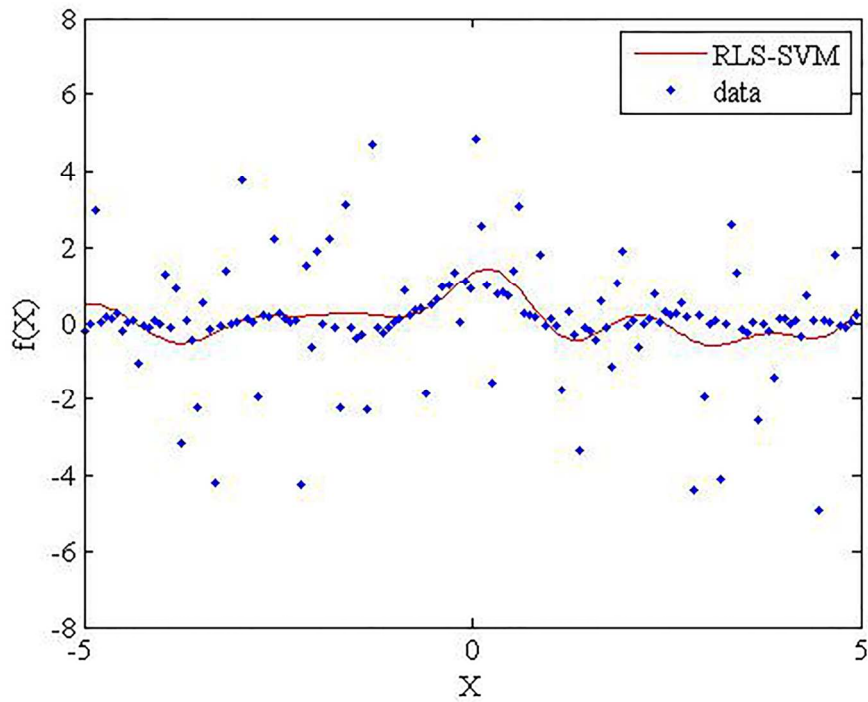
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(c) 45% outliers(143 samples)
62x46mm (600 x 600 DPI)
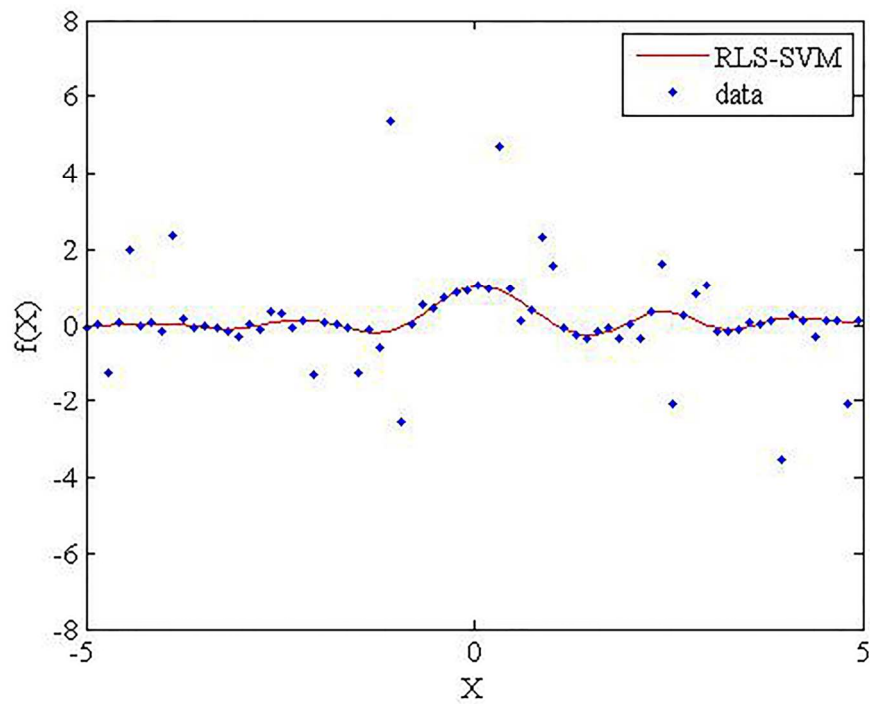
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(d) 72 samples(25% outliers)
62x46mm (600 x 600 DPI)
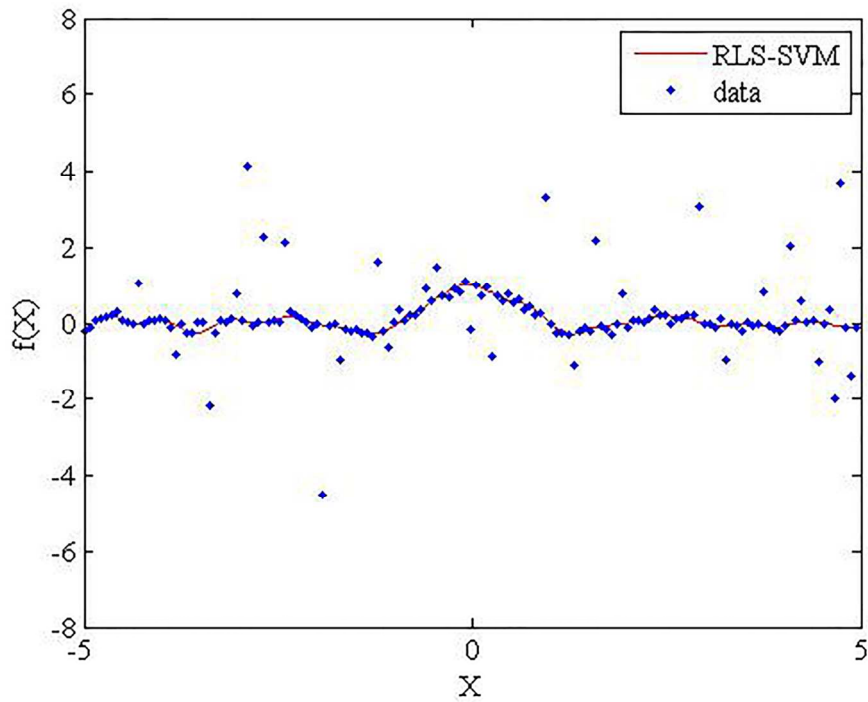
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(e) 143 samples(25% outliers)
62x46mm (600 x 600 DPI)
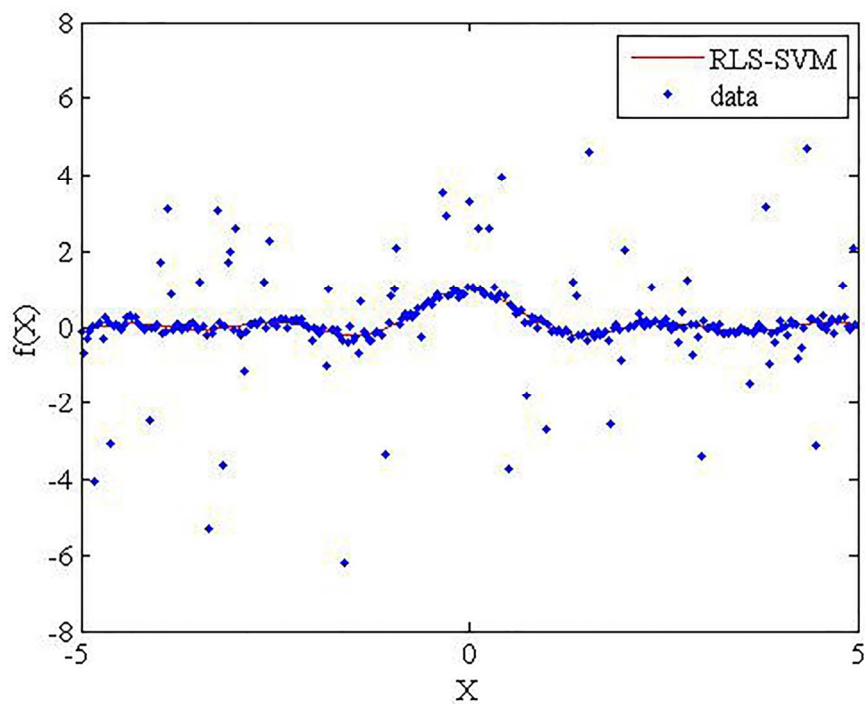
Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise
(f) 286 samples(25% outliers)
62x46mm (600 x 600 DPI)

Table 1 The elemental concentrations of the samples

No. of sample	No. of Steel Grade	Cu(%)	Zn(%)	Pb(%)	Fe(%)
1#	BYG199501	96.86	3.06	0.008	0.024
2#	BYG199502	95.1	4.78	0.0236	0.066
3#	BYG199503	94.46	5.26	0.05	0.182
4#	BYG199504	92.7	6.81	0.098	0.336
5#	BYG199505	89.97	9.83	0.0301	0.124
6#	BYG199506	90.76	9.15	0.012	0.051
7#	BYG199507	85.49	14.41	0.0283	0.097
8#	BYG199508	79.1	20.74	0.029	0.098
9#	BYG199509	70.44	29.04	0.132	0.182
10#	BYG199510	69.25	30.66	0.0105	0.016
11#	BYG199511	67.59	32.17	0.06	0.101
12#	BYG199512	66.11	33.72	0.026	0.0353
13#	BYG199513	64.32	35.51	0.0697	0.067
14#	BYG199514	63.42	36.18	0.163	0.140
15#	BYG199515	60.81	38.59	0.294	0.236
16#	BYG199516	57.98	41.04	0.591	0.427

Table 2   The selected analytical lines of the Cu

Number	Atom/ion	Wavelength (nm)
1	CuI	261.837
2	CuI	282.437
3	CuI	296.116
4	CuI	427.511
5	CuI	458.695
6	CuI	570.024
7	CuI	578.213
8	CuII	204.38
9	CuII	221.027
10	CuII	224.7

Table 3   Prediction results of different models for Cu concentration

Model	RMSE(%)	MRE(%)
PLS	10.569	9.78
SVM	8.237	8.95
LS-SVM	6.164	7.59
WLS-SVM	3.895	5.21
RLS-SVM	1.537	1.73

Table 4   Comparison of different weighting scheme for LIBS predictions

Weighting schemes	The average values of RSD (%)	RMSE (%)	Iteration times
Huber	6.62	4.172	20
Hampel	6.39	3.895	4
Logistic	3.85	2.231	10
Myriad	3.02	1.556	26
The proposed function	3.06	1.537	5

Table 5 The influence of different proportions of outliers on the regression results

Outliers proportion	RMSE	MRE	SKE
15%	0.7887	1.4685	0.3828
30%	1.0088	1.6171	0.5591
45%	1.3721	3.8815	1.6033

Table 6 The effect of the number of samples against the regression results

Number of samples	RMSE	MRE	SKE
72	0.8261	1.6737	0.5122
143	0.8502	1.5365	0.3295
286	1.1459	1.6287	0.4097

Laser Induced Breakdown Spectroscopy quantitative analysis method based on Robust Least Squares Support Vector Machine regression model

Jianhong Yang[a,*], Cancan Yi[a], Jinwu Xu[a], Xianghong Ma[b]

[a] School of Mechanical Engineering, University of Science and Technology Beijing, Beijing, 100083, China

[b] School of Engineering & Applied Science, Aston University, Birmingham, B4 7ET, UK

Abstract: Data fluctuation in multiple measurements of Laser Induced Breakdown Spectroscopy (LIBS) greatly affects the accuracy of quantitative analysis. A new LIBS quantitative analysis method based on Robust Least Squares Support Vector Machine (RLS-SVM) regression model is proposed. The usual way to enhance the analysis accuracy is to improve the quality and consistency of the emission signal, such as averaging the spectral signals or spectrum standardization over a number of laser shots. The proposed method focuses more on how to enhance the robustness of the quantitative analysis regression model. The proposed RLS-SVM regression model originates from the Weighted Least Squares Support Vector Machine (WLS-SVM) but has an improved segmented weighting function and residual error calculation according to the statistical distribution of measured spectral data. Through the improved segmented weighting function, the information of the spectral data in the normal distribution will be retained in the regression model while the information of the outliers will be restrained or removed. Copper elemental concentration analysis experiments of 16 certified standard brass samples were carried out. The average value of relative standard deviation obtained from the RLS-SVM model was 3.06% and the root mean square error was 1.537%. The experiment results showed that the proposed method achieved better prediction accuracy and better modeling robustness compared with the quantitative analysis methods based on Partial Least Squares (PLS) regression, standard Support Vector Machine (SVM) and WLS-SVM. It was also demonstrated that the improved weighting function had better comprehensive performance in model robustness and convergence speed, compared with the known four weighting functions.

Keywords: Laser Induced Breakdown Spectroscopy; Quantitative Analysis; Robust Regression; Least Squares Support Vector Machine

1 Introduction

Laser Induced Breakdown Spectroscopy (LIBS) is a type of atomic emission

spectroscopy which has well-known advantages of low requirement of sample preparation, micro-destructiveness, short measurement time and suitability of multi-element analysis in situ [1-3].

Although LIBS quantitative analysis technique has a wide application [4-6], the accuracy of LIBS quantitative analysis is affected by the fluctuation of spectral data due to poor repeatability of LIBS measurement [7-8]. The spectral data fluctuation problem in the LIBS analysis is caused by a few complicated factors such as laser pulse fluctuation, inhomogeneity of sample composition, instability of plasma position, variation of focus depth, environmental noise, instrumental errors and operational errors. In order to enhance the accuracy of the quantitative regression analysis, the simplest and most frequently applied method is to improve the consistency of spectral data by averaging the spectral intensities of multiple measured emission signals [9-10]. Ideally, the spectral intensities of the LIBS spectra collected from a number of laser shot on the surface of one same sample should strictly obey normal distribution. In reality, the spectral intensities only follow an approximate normal distribution, in which some outliers with particularly large or small values may exist. Thus, the averaging results will be inaccurate especially when the outliers deviate severely from the normal values. Currently, the common method to deal with the outliers in the quantitative regression analysis is to preprocess the spectral data before building the regression model. The common preprocessing techniques include the normalization and standardization of the spectrum [11-13], the optimization of the plasma characteristic parameters [14]. However, a great deal of priori knowledge is required and some parameters have to be determined by numerous trial-and-error tests. A different and more feasible way is to build a robust regression model to tolerate the fluctuation [15-17] of spectral data. In this paper, a LIBS quantitative regression analysis method based on RLS-SVM is proposed in order to reduce the negative effects caused by the abnormal data through enhancing the robustness and error tolerance of the regression model itself.

The regression model is the core of LIBS quantitative analysis method with calibrated curves of standard samples. The regression model is built or trained to describe the relationship between the elemental concentrations and the analytical spectral intensities of emission spectra collected from the plasma. The multivariate techniques of principal component regression (PCR) [18], partial least squares regression (PLS) [19], random forest regression (RFR) [20] have been used in LIBS

quantitative analysis. The methods of both PCR and PLS are linear regression models which may not be suitable for describing the nonlinear relationship between the spectral line intensities and the elemental concentrations, especially for the samples with high elemental concentrations. Zhang et al. [20] reported that RFR had a better performance potential in LIBS quantitative analysis compared with the methods of SVM and PLS when a feature spectral band was regarded as the input variables and a preprocessing of smoothing and de-noising of the spectra was made before modeling. However, the predictive ability of RFR method highly depends the two parameters, the number of the trees in the forest and the number of the peaks randomly selected as the candidates for splitting at each node [20]. A desired regression model not only should be robust since the measured spectral intensities may fluctuate even under the same condition, but also should have nonlinear fitting ability [21-22] because the relationship between the elemental concentrations and the analytical spectral intensities usually exhibits strong nonlinearity due to the matrix effect, the absorption effect and the influence of background radiation. The Support Vector Machine (SVM) [23] and its improved variations, such as Least Squares Support Vector Machine (LS-SVM) [24] and Weighted Least Squares Support Vector Machine (WLS-SVM) [25], are the most commonly used nonlinear regression models for limited samples based on structural risk minimization. Structural risk minimization (SRM) is an inductive principle of use in machine learning. Commonly in machine learning, a generalized model must be is usually established selected from a finite training data set, with the consequent problem of overfitting, that is, the model becoming too strongly tailored to the particularities of the training set and generalizing poorly to new data. The SRM principle addresses this problem by balancing the model's complexity against its success at fitting the training data. The SVM method has already been widely used in LIBS quantitative analysis [26-30]. A spectral algorithm based on SVM was used as a classifier to automatically identify areas with aluminum presence [26]. Pokrajac et al.[27] utilized the SVM method in automatic classification of LIBS data of protein biomarker solutions. SVM was used in advanced statistical analysis of LIBS data to discriminate sedimentary rocks [28]. SVM classification of suspect powders using LIBS spectral data was proposed by Cisewski et al.[29]. Yu et al. proposed an approach to polymer identification by LIBS with adjusting spectral weightings and SVM aiming at improving the identification accuracy [30]. In the current applications in LIBS, SVM is usually used as a statistical tool. But, the

predictions of SVM model are easy to be influenced by abnormal data [31]. Therefore, more robust statistical model should be considered in LIBS quantitative analysis. The WLS-SVM has a better model robustness than SVM. However, the weighting strategy of WLS-SVM is rigid and does not consider the actual data distribution. The predictions of WLS-SVM regression model will be poor when the percentage of the abnormal data is large. Traditional robust learning is to find ways of reducing the effects of outliers. The idea of such an approach is to minimize the values of weights during the process of minimizing errors. Thus, the weighting function has a great importance in improving the model robustness. Brabanter et al. [32] compared four commonly used weighting functions, the Myriad, Huber, Hampel and Logistic weighting function, according to their performance in providing model robustness and fast convergence. Based on WLS-SVM, an improved regression model named Robust Least Square-Support Vector Machine (RLS-SVM), which has a new segmented weighting scheme, is proposed for LIBS quantitative analysis, to handle with the abnormal data according to the actual spectral intensity distribution.

The proposed RLS-SVM regression model was tested in the copper concentration analysis experiments of 16 certified standard brass samples. The experimental results showed that the RLS-SVM model had a good performance in suppressing the negative effect caused by the data fluctuation of multiple measurements and the average values of Relative Standard Deviation (RSD) obtained from the RLS-SVM model was 3.06%, the root mean square prediction error (RMSE) was 1.537%. All of which were better compared with the WLS-SVM and the LS-SVM regression model. Also, it was demonstrated that the improved weighting function had better comprehensive performance in model robustness and convergence speed, compared with the known four weighting functions.

2   Theory

2.1 Data fluctuation in LIBS measurement

LIBS technology is sensitive to the measurement parameters, such as the laser energy, the plasma volume, the background radiation and the instrumental error etc. The distribution of spectral lines intensity of CuI427.511nm in the LIBS spectra collected from 120 times laser shot on the same position of the standard sample BYG199506 in which the concentration of Cu is 90.76%, is shown in Fig.1. The histogram of spectral line intensity distribution is shown in Fig.1(a) and the normal probability plot is drawn in Fig.1(b). In the practice of quantitative analysis of LIBS, a

number of measurements are usually implemented in order to obtain a relatively stable and correct spectrum. Averaging is the most often used technique to deal with the data fluctuation problem in the multiple measured spectra. However, the distribution of analytical line intensities in the measured spectra from multiple tests forms only an approximate normal distribution. In a normal probability plot (also called a "normal plot"), the sorted data are plotted vs. values selected to make the resulting image look close to a straight line if the data are approximately normally distributed. Deviations from a straight line suggest departures from normality. From Fig.1 we can find that there are some outliers whose values are relatively large or small and do not fit the normal distribution. It is a significant task to determine how to select or prepare the model training dataset from the collection of the spectra of multiple measurements. The existence of outliers brings challenges to the averaging technique because before applying the averaging technique, the outliers must be dropped by proper statistical methods, in which careful consideration should be taken when determining the threshold parameters of outliers. In this paper, a feasible alternative, that is to put all the measured results into the training dataset but give each value a weight according to its credibility in the whole distribution, is proposed and developed.

[Fig.1   Data fluctuation of spectral intensities in 120 times of LIBS measurements at a same position of the surface of 6# sample (BYG199506)]
[(a)Frequency histogram of spectral intensities in 120 repetitions of measurement]
[(b)The normal probability plot of the intensities of multiple measurements]

2.2 The regression model based on the Weighted Least Squares Support Vector Machine

The main idea of quantitative analysis method with calibration curves of standard samples is to establish a statistical regression model between the elemental concentrations and spectral intensities. In the quantitative analysis of LIBS, for the j-th training standard sample, a set of input-target pairs $\left\{x_k, y_j\right\}_{k=1}^{N}$, $x_k \in R^d$, $y \in R$, $j = 1 \cdots M$, where x_k and y_j denote the analytical line intensities and the corresponding elemental concentrations respectively, N is the number of measurements for each sample, M is the number of samples, d is the number of the analytical lines, will be used to train the regression model.

In the Weighted Least Squares Support Vector Machine, the regression model

can be considered as the following optimization problem in primal weight space[25]:

$$\min J = \frac{1}{2}ww^{\mathrm{T}} + \frac{1}{2}\gamma\sum_{k=1}^{N}v_k e_k^{\ 2} \tag{1}$$

$$s.t \quad w^{\mathrm{T}}\Phi(x_k) + b + e_k = y_k$$

where J is the cost function, $\Phi(x)$ is a function which maps the input space into a so-called higher dimensional feature space, w is the weight vector in primal weight space, e_k is the error variable , b is the bias term, γ is a regularization term and v_k is the weight.

With a Lagrangian multipliers α_k, Eq. (1) can be rewritten as

$$L = \frac{1}{2}ww^{\mathrm{T}} + \frac{1}{2}\gamma\sum_{k=1}^{N}v_k e_k^{\ 2} - \sum_{k=1}^{N}\alpha_k[w\Phi(x_k) + b + e_k - y_k] \tag{2}$$

The conditions for optimality are given by

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 & \rightarrow w = \sum_{k=1}^{N}\alpha_k\Phi(x_k) \\[2mm] \dfrac{\partial L}{\partial b} = 0 & \rightarrow \sum_{k=1}^{N}\alpha_k = 0 \\[2mm] \dfrac{\partial L}{\partial e_k} = 0 & \rightarrow \alpha_k = \gamma v_k e_k \\[2mm] \dfrac{\partial L}{\partial \alpha_k} = 0 & \rightarrow w^T\Phi(x_k) + b + e_k - y_k = 0 \end{cases} \tag{3}$$

After elimination of w, e_k, one obtains the solution[25]

$$\begin{bmatrix} 0 & 1 \\ 1 & \Omega + V_\gamma \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{4}$$

where $y = [y_1,...,y_N]^{\mathrm{T}}$, $1 = [1_1,...,1_N]^{\mathrm{T}}$, $\alpha = [\alpha_1,...,\alpha_N]^{\mathrm{T}}$, $V_\gamma = diag\left\{\dfrac{1}{\gamma v_1},...,\dfrac{1}{\gamma v_N}\right\}$,

$\Omega = \{\Omega_{kl} \mid k,l = 1,...,N\}$, $\Omega_{kl} = \Phi(x_k)\Phi(x_l) = K(x_k,x_l)$. Therefore, the resulting

WLS-SVM model for function estimation becomes $y = \sum_{k=1}^{N}\alpha_k K(x_i,x) + b$ .Where

$K(\cdot)$ is the kernel function and the Gaussian kernel function defined by Eq.(5) is

normally used, where δ is the kernel parameter:

$$K(x_k, x_l) = \exp(-\|x_k - x_l\|^2 / \delta^2) \qquad (5)$$

The choice of the weight v_k is determined based upon the error variable

$e_k = \alpha_k / \gamma$ and the value is obtained by taking:

$$v_k = \begin{cases} 1 & |r| < c_1 \\ \dfrac{c_2 - |r|}{c_2 - c_1} & c_1 < |r| < c_2 \\ 10^{-4} & other \end{cases} \qquad (6)$$

where $r = e_k / s$, $s = 1.483 MAD(e_k)$ is a robust estimate of the standard

deviation of the WLS-SVM error variables, $MAD(e_k) = med(|e_k - med\{e_k\}|)$

stands for the median absolute deviation, med stands for median. The constants

c_1 and c_2 are typically chosen as $c_1 = 2.5$ and $c_2 = 3$ [25].

The WLS-SVM model has two shortcomings. First, the result can't converge by

means of iterative calculation because the residual vector is defined as

$e_k = \alpha_k / \gamma$ where v_k was not involved in calculation. Second, the model's prediction

ability is poor when the quantity of abnormal data among the training dataset is large.

2.3 The regression model based on Robust Least Squares Support Vector Machine

Based on the WLS-SVM, the main improvements of the proposed RLS-SVM

include two aspects. First, different from the WLS-SVM, the new calculated residual

vector is $e_k = \alpha_k / v_k \gamma$. Thus, the information of the weight variable v_k could be

taken into consideration during the iterative optimization process and the residual

vector could be updated according to the adjustment of the weight variable. Second,

when the error variable e_k does not satisfy the normal distribution, a novel segmented

weighting function μ_k defined in Eq. (7), which is originated from an adaptive

factor of robust filtering for kinematic geodetic positioning [33], is put forward to

replace the old function v_k in Eq. (6).

$$\mu_k = \begin{cases} 1 & |r| < k_0 \\ \dfrac{k_0}{|r|}\left(\dfrac{k_1 - |r|}{k_1 - k_0}\right)^2 & k_0 \le |r| \le k_1 \\ 0 & |r| > k_1 \end{cases} \tag{7}$$

In Eq.(7), $r = e_k / s$, $s = 1.483 MAD(e_k)$ is a robust estimate of the standard deviation of the RLS-SVM error variables. The parameters are typically chosen as $k_0 = 1.5$ and $k_1 = 2.5$, on account that the probabilities of residual errors beyond the scope of $\pm 1.5\sigma$ and $\pm 2.5\sigma$ are 0.13 and 0.01 respectively, where σ stands for the standard deviation. The segmented weighting function is divided into three parts: retaining weights area corresponding to the credible data, dropping weights area corresponding to the suspicious data and eliminating area corresponding to the outliers. The weighting function is bounded, segmented, continuous and efficient. Schematic diagrams of different weighting functions are shown in Fig.2. Fig.2(a) illustrates the weighting function of WLS-SVM and Fig.2(b) demonstrates the segmented weighting function of RLS-SVM. For a series of spectra collected from the multiple LIBS measurements, draw the distribution of the analytical line intensities firstly, and then assign the weights for each spectrum according to Eq. (7). For the main body of the measurements whose spectral intensities are within the range of $\pm 1.5\sigma$, the weights should be set as 1; for the measurements whose spectral intensities are within the range between $\pm 1.5\sigma$ and $\pm 2.5\sigma$, the weights should be set as dropping weights according to Eq. (7); and for the outliers which are out of the range of $\pm 2.5\sigma$, the weights should be set as zero which means the outliers should be removed from the training dataset. Compared with weighting function defined in Eq.(6), the segmented weighting scheme in RLS-SVM has slower dropping weights for the suspicious measured data and directly eliminates the outliers other than assigns a relatively small weight. The parameters k_0 and k_1 are selected according to the characteristic of norm distribution.

[Fig.2 Different schemes of the segmented weighting function of WLS-SVM and RLS-SVM]

The calculation steps of the proposed RLS-SVM regression model in LIBS quantitative analysis are as follows:

Step1: For N measurements of M different spectral standard samples, a set of total input-target pairs $\{x_k, y_k\}_{k=1}^{P}$, $P = M \cdot N$, $x_k \in R^d$, $y \in R$, are used to train the regression model according to Eq. (1) to Eq. (4).Where x_k and y_k denote the analytical line intensities and the corresponding elemental concentrations respectively, d is the number of the analytical lines. The appropriate regularization parameter γ and kernel parameter δ can be achieved by cross validation method. The initial weight is $\mu_k = 1$.

Step2: Calculate the residual vector $e_k = \alpha_k / [\mu_k \cdot \gamma]$ and the robust estimate of the error variables $s = 1.483 MAD(e_k^{(i)})$ for each standard samples, where i is the number of iteration.

Step3: $r^{(i)}$ can be calculated by $r^{(i)} = e_k^{(i)} / s$ and the weighting vector can be obtained from Eq. (7) for each standard samples.

Step4: The fitted regression model is $g^{(i)}(x) = \sum_{k=1}^{P} \alpha_k^{(i)} K(x, x_k) + b^{(i)}$, whose solution can be obtained according to Eq. (4).

Step5: The termination condition of the iterative optimization is $max\left(\left|\alpha_k^{(i)} - \alpha_k^{(i-1)}\right|\right) \leq \varepsilon$, where ε is a small enough positive number. If the condition is satisfied, stop the iteration and the robust regression model obtained in Step 4 is the final quantitative analysis model; otherwise, return to step2 and continue the optimization iteration.

Step6: The elemental concentration predictions for a given x_* can be computed by $y_* = \sum_{k=1}^{P} \alpha_k^{(i)} K(x_*, x_k) + b^{(i)}$.

3   Experimental Setup

The schematic drawing of the LIBS experimental setup used in this investigation is shown in Fig. 3. The LIBS device used in the experiment was the Spectrolaser 4000(XRF, Australia). The laser source employed was a Q-switched Nd:YAG laser with a wavelength of 532 nm, energy of 0-190 mJ per pulse, a pulse duration of 5 ns and a maximum repetition rate of 10 pulse/s. The laser source was focused on the

sample through a 50mm focal lens. The integration time of the detection system is fixed at 1ms. The distance between the focal lens and the sample surface is about 49.5mm. The plasma-emitted radiation was collected with a condenser, which was connected to the detection system. The detection system was composed of 4 Czerny-Turner spectrometers and CCD detectors which covered the spectral range from 190 to 940nm, giving a nominal resolution of 0.09nm. The optimal laser energy and delay time were set as 100 mJ/pulse and 2us respectively, to ensure the measured spectrum could have a high signal-to-background ratio.

[Fig3   Schematic experimental setup of LIBS measurement]

The samples investigated in this paper were 16 pieces of standard certified brass samples from CISRI (Central Iron & Steel Research Institute, China). The main elements contained in the samples are Cu, Zn, Pb and Fe, etc. The compositions of samples are listed in Table 1 in the descending order of the weight percentage concentrations of Cu. The element of cuprum was taken as the main analytical target in the experiments. Among the samples, the highest concentration of Cu is 96.86% and the lowest is 57.98%. The even-numbered samples were used to train the LIBS quantitative analysis model. The rest samples with odd serial numbers were used to test the quantitative analysis model. Before data collection, the sample surface was carefully cleaned using ethanol and dried in air. The main purpose of the experiments was to verify whether the proposed LIBS quantitative analysis method based on RLS-SVM could effectively reduce the influence caused by the fluctuations of spectral dada. Therefore, for each standard sample, the beam of the pulsed laser was focused on the sample surface successively for multiple 40 times and the emission radiation were collected for each repetition. Fig.4 shows one LIBS spectrum measured in the experiments.

Table 1 The elemental concentrations of the samples

No. of sample	No. of Steel Grade	Cu(%)	Zn(%)	Pb(%)	Fe(%)
1	BYG199501	96.86	3.06	0.008	0.024
2	BYG199502	95.1	4.78	0.0236	0.066
3	BYG199503	94.46	5.26	0.05	0.182
4	BYG199504	92.7	6.81	0.098	0.336

5	BYG199505	89.97	9.83	0.0301	0.124
6	BYG199506	90.76	9.15	0.012	0.051
7	BYG199507	85.49	14.41	0.0283	0.097
8	BYG199508	79.1	20.74	0.029	0.098
9	BYG199509	70.44	29.04	0.132	0.182
10	BYG199510	69.25	30.66	0.0105	0.016
11	BYG199511	67.59	32.17	0.06	0.101
12	BYG199512	66.11	33.72	0.026	0.0353
13	BYG199513	64.32	35.51	0.0697	0.067
14	BYG199514	63.42	36.18	0.163	0.140
15	BYG199515	60.81	38.59	0.294	0.236
16	BYG199516	57.98	41.04	0.591	0.427

[**Fig.4   A typical LIBS spectrum of brass samples**]

4 Experimental analysis results

The correct selection of analytical lines is an important step to guarantee the accuracy of quantitative analysis. The rules of the analytical line selection adopted here are as follows. First, the lines which have interference with the adjacent lines of other emission species must be avoided; Second, the lines which have larger possibilities of self-absorption, such as resonant lines, should not be selected. The lines which were clearly separated from other adjacent lines and have high signal-to-background ratios were chosen as analytical lines in consideration of the larger elemental concentration listed in Table1. Finally, 10 analytical lines of Cu element were determined as shown in the Table 2 according to the above rules. The wavelengths described in the paper are referred to the identification results by comparing the measured spectra with the reference values in the NIST database [34].

Table 2   The selected analytical lines of the Cu

Number	Atom/ion	Wavelength (nm)
1	CuI	261.837
2	CuI	282.437
3	CuI	296.116
4	CuI	427.511
5	CuI	458.695
6	CuI	570.024

7	CuI	578.213
8	CuII	204.38
9	CuII	221.027
10	CuII	224.7

After the determination of the analytical lines, the fluctuations of analytical lines spectral intensities need to be considered. The main index used to evaluate the degree of the data fluctuation is the RSD. A greater RSD value stands for greater data fluctuations. Fig. 5 describes the fluctuations of spectral intensity of the line at Cu I 427.511nm in the ahead of 19 times measurement from a same place of the standard sample BYG119504. The horizontal axis represents the different measurements and vertical axis represents the spectral line intensity. It can be seen that the spectral intensities fluctuate within a certain range during the process of continuous measurements. As seen in Fig. 6, the RSDs of the spectral intensity of the line at Cu I 427.511nm for all the 16 standard samples fluctuate within the range from 16.59% to 5.62%.

[Fig.5 Pulse-to-pulse spectra fluctuations of the intensity (BYG119504 427.511nm)]

[Fig.6 RSD values of the characteristic line intensity at 427.511nm of Cu normalized for all 16 brass samples]

The spectral intensities of the selected 10 analytical lines listed in Table 2 were regarded as the input variables for each sample in the quantitative analysis model based on RLS-SVM. For each sample, multiple measurements were carried out and a series of spectra were obtained. The 10 analytical line intensities in all the measured spectra formed spectral data matrix for each sample. The spectral data matrices of all the samples were used to train the regression model. In the experiments, the Kernel parameter was 5.75 and regularization parameter γ was 1.05 by leave-one-out cross-validation techniques. The maximal iterations number was set as 100 by the convergence requirement. In order to evaluate the effectiveness the proposed method, the root mean square error ($RMSE$) and the mean relative error (MRE) were used as the evaluation indicators.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(C_i - \mu_i)^2}{n}} \qquad (8)$$

$$MRE = \frac{\sum_{i=1}^{n}|C_i - \mu_i|}{\sum_{i=1}^{n}|C_i|} \tag{9}$$

In Eq. (8) to Eq. (9), C_i is the certified elemental concentration of Cu of the i-th sample, μ_i is the predicted elemental concentration of the i-th sample by the quantitative analysis model, n is the number of the samples.

4.1 Experiments on different model inputs

In the experiments, the training samples consisted of 8 even-numbered samples and the samples for model testing included the other 8 odd-numbered samples. For each sample, 40 emission spectra were collected from 40 times of focusing laser on a same spot on the sample surface. In order to compare the influence of different inputs on the predicting results of the RLS-SVM regression model, 6 cases of different form of inputs were considered in the experiments. In case 1, the analytical line intensities of all the measured spectra from multiple measurements were taken as the input variables. In case 2, the analytical line average intensities of all the 40 measured spectra were regarded as the input variables. In case 3 to case 6, the analytical line average intensities of the first to 10th, the 11th to 20th, the 21th to 30th, the 31th to 40th measured spectra were regarded as the input variables, respectively. Fig. 7 demonstrates the RMSE of the regression model predictions in the above 6 cases. It can be seen that in case 1 the RMSE has the smallest value which means that taking the analytical lines of all the measured spectra as the input variables of the model is helpful to enhance the accuracy of prediction. It can be explained from two aspects. First, all the measured spectra contain more useful information than the average value. Second, the average value is more vulnerable to abnormal data.

[Fig.7   The RSD values of predictions by different input variables]

4.2 The distribution of error variable e_k in LIBS quantitative analysis

The multiple measurements of standard samples #1 is utilized to test the distribution of error variable e_k and the analytical lines are determined as shown in Table2 . The LS-SVM regression model is chosen according to Eq. (1) to Eq. (4) and the initial weight is $\mu_k=1$. The histogram of error variable e_k distribution is shown in Fig.8(a) and Fig.8(b) demonstrates the test of normal distribution. From Fig.8 we can

find that the error variable e_k doesn't from normal distribution strictly and there are some abnormal data. Therefore, the weighting function is proposed for the purpose of reducing the weights of abnormal data and improving the model robustness.

[Fig.8   The frequency histogram of error variable and normal probability plot]

[(a) Frequency histogram of error variable e_k]

[(b) The normal probability plot of error variable e_k]

4.3 Experiments on different regression models

The effectiveness of the proposed quantitative analysis method based on RLS-SVM model was investigated by a comparative study of the methods based on different models including the PLS model, the standard SVM model, the LS-SVM model and the WLS-SVM model. In the experiments, the analytical line intensities of all the 40 measured spectra were taken as the input variables of all the regression models. Fig. 9 shows the predictions of the WLS-SVM regression model and the RLS-SVM regression model respectively for pulse-to-pulse measurement spectral data. As shown in Fig. 9, the RLS-SVM model has better accuracy of prediction than the WLS-SVM model. Also, the smaller distribution range of predictions also shows that the RLS-SVM model has better reliability and robustness than the WLS-SVM model. Figure 10 shows the RSD values of the predictions of the WLS-SVM model, and the RLS-SVM model. The average values of RSD of the WLS-SVM model and the RLS-SVM model are 6.39% and 3.06% respectively, which indicates the RLS-SVM model is able to improve the predicting capability by reducing the negative influence of the spectral data fluctuation.

[Fig.9   Prediction results of different regression models]

[(a) Prediction results of WLS-SVM]

[(b) Prediction results of proposed RLS-SVM]

[Fig .10   RSD values of predictions of different regression models]

It is necessary to further evaluate the accuracy of quantitative analysis by $RMSE$ and MRE . The comparison results are listed in Table 3 from which we also can conclude that the proposed new quantitative analysis method RLS-SVM has better prediction accuracy and generalization ability compared with the other regression models. The possible reasons include that a)the PLS model belongs to linear regression methods and has a poor ability to deal with nonlinear problem, b) both the SVM and LS-SVM model lack robustness relative to the abnormal data, and c) the scheme of weighting function in the WLS-SVM does not consider the actual

distribution of the measured spectral data.

Table 3   Prediction results of different models for Cu concentration

Model	RMSE(%)	MRE(%)
PLS	10.569	9.78
SVM	8.237	8.95
LS-SVM	6.164	7.59
WLS-SVM	3.895	5.21
RLS-SVM	1.537	1.73

The weighting scheme is a critical procedure of robust regression model. As mentioned in ref [32], there are four weight functions often used in regression. According to the conclusions in ref [32], the Logistic and Myriad weights are suitable reweighting schemes when outliers are present in the data and the Myriad shows better performance over the others in the presence of extreme outliers. However, the effectiveness of Myriad weighting function was restricted by the selection of function parameter, which will cost a heavy workload. In ref [32], the simulated example and real life data sets demonstrated that the Myriad weight function is highly robust against (extreme) outliers but has a slow speed of convergence. A comparative study of different weighting schemes for LIBS quantitative analysis was carried out and the results were listed in Table 4. It is noteworthy that the Hampel function is equivalent to the original weighting function of WLS-SVM defined in Eq.6 when the parameter was set as $b_1 = 2.5$, $b_2 = 3$. The experimental results showed that although the Myriad function had a smallest RSD value, its number of iteration times for algorithm convergence was the largest, which meant the largest workload of computation. From Table 4, it can be found that a good compromise among speed of convergence, robustness and accuracy can be achieved by using the proposed weighting function, which indicates that the proposed RLS-SVM model has a better potential in LIBS quantitative analysis, especially in the situation of real-time analysis.

Table 4   Comparison of different weighting scheme for LIBS predictions

Weighting schemes	The average values of RSD (%)	RMSE (%)	Iteration times
Huber	6.62	4.172	20
Hampel	6.39	3.895	4

Logistic	3.85	2.231	10
Myriad	3.02	1.556	26
The proposed function	3.06	1.537	5

5 Discussions

When predicting the concentrations of the test samples, the variables input into the trained regression models in the experiments listed above were also set as the analytical line intensities of all the 40 measured spectra of each test sample. However, it must be claimed that in practical applications it is not easy to determine how many times of measurements should be carried out to achieve the best prediction accuracy. It is beneficial to improve the efficiency of the online LIBS quantitative analysis in industrial situations with the least necessary number of the multiple measurements. Therefore, developing a robust regression model is important to LIBS quantitative analysis. This is the aim of the proposed RLS-SVM regression model. In the training stage, multiple LIBS measurements of each sample could be performed to train the regression model by the spectral data with fluctuations. Once the regression model is trained, in the predicting stage, a reduced number of measurements are needed because the regression model is robust and can guarantee the prediction accuracy by tolerating the fluctuations of the spectral data.

The weighting scheme selection is important in building LIBS robustness regression model. The mentioned Myriad function also has a good robustness and prediction accuracy. However, it is worthy mentioning that the selection of function parameter is too complex and its iteration times are larger than the others.

Cui et al. also proposed a method of adaptive weighted least square support vector machine regression integrated with outlier detection and achieved successful application in quantitative structure–activity relationship (QSAR) [35]. However, outlier detection is an important preprocess of Cui's method. In the proposed method, it is not necessary to remove outliers before building the regression model, because the segmented weighting function and iterative weighting scheme can handle with the outliers automatically. Therefore, the proposed method is more appropriate and practical for LIBS quantitative analysis in industrial environment, especially in the situation of real-time analysis.

The RLS-SVM method is improved based on WLS-SVM. In order to assess the

performance of the algorithm against data variability, datasets generated from the sinc function with Gaussian noise are used for evaluation. First, three datasets which contain 15%, 30% and 45% outliers respectively are tested. In the three datasets, the number of samples is all 143. The Fig.11 (a), Fig.11 (b), Fig.11(c) and Table5 demonstrated the influence of different proportions of outliers on the regression results. Second, another three datasets in which the numbers of sample are 72, 143 and 286 respectively are used in the experiment. In the three datasets, the proportion of outliers is all 25%. The Fig.11 (d), Fig.11 (e), Fig.11 (f) and Table6 illustrated the effect of the number of samples against the regression results. It is worthy mentioned that the value SKE in Table5 and Table 6 stands for the symmetry of the distribution of the sample which can be used for evaluating how skewed the data will be. The experimental results showed that the effectiveness of RLS-SVM model will be reduced with the increase of the proportion of outliers and have small fluctuation when the number of samples vary. Also, it can be found from Table 5 and Table6 that a larger SKE of data indicates a larger predictive error.

(a) 15% outliers(143 samples)	(b) 30% outliers(143 samples)	(c) 45% outliers(143 samples)

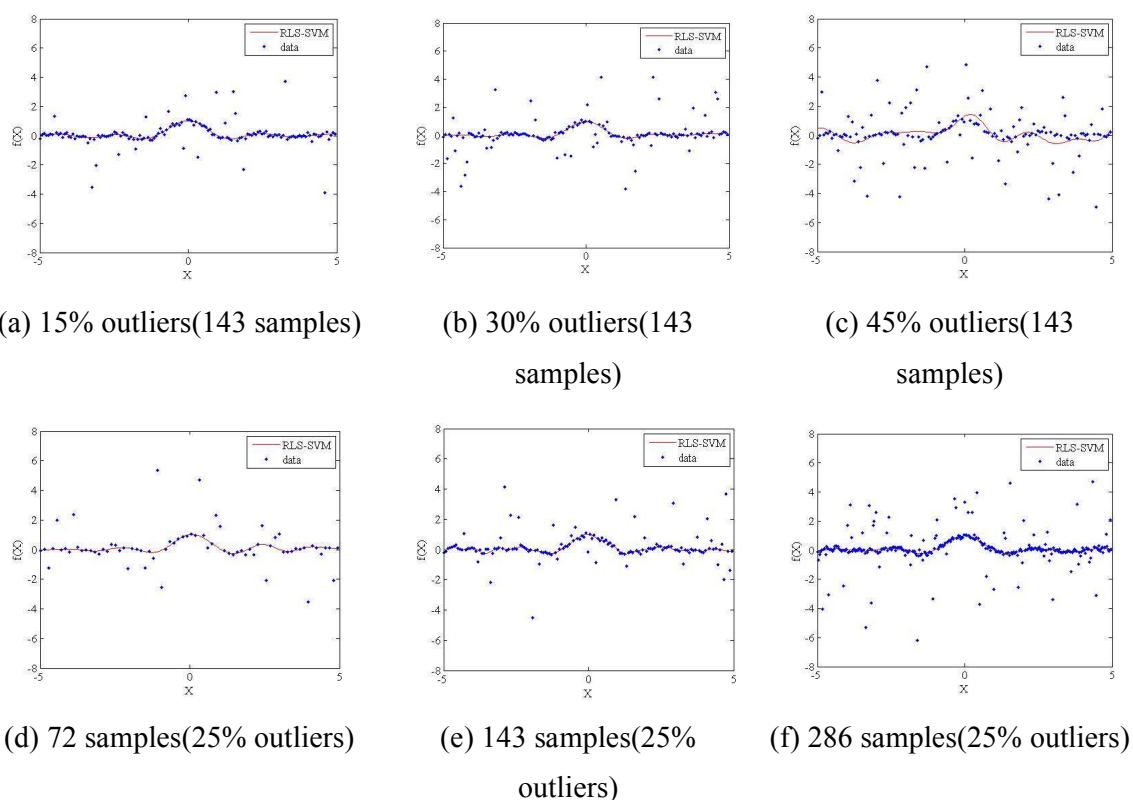| (d) 72 samples(25% outliers)	(e) 143 samples(25% outliers)	(f) 286 samples(25% outliers)

Fig. 11 RLS-SVM regression results of different datasets generated from sinc function with Gaussian noise

Table 5 The influence of different proportions of outliers on the regression results

Outliers proportion	RMSE	MRE	SKE
15%	0.7887	1.4685	0.3828
30%	1.0088	1.6171	0.5591
45%	1.3721	3.8815	1.6033

Table 6 The effect of the number of samples against the regression results

Number of samples	RMSE	MRE	SKE
72	0.8261	1.6737	0.5122
143	0.8502	1.5365	0.3295
286	1.1459	1.6287	0.4097

6 Conclusions

This paper proposed a robust quantitative analysis method based on RLS-SVM regression model in order to reduce the negative influence of the spectral data fluctuation and enhance the prediction accuracy of LIBS quantitative analysis method with calibration curves of standard samples. The significance of this article reflects on two aspects. On the one hand, the concept of improving the model robustness is proposed to handle with the abnormal data and outliers in LIBS quantitative analysis. On the other hand, an improved segmented weighting function according to the actual statistical distribution of spectral data in multiple measurements is introduced, which can keep the spectral information of the credible measurements to the maximum and suppress or eliminate the spectral information of the suspicious or abnormal measurements.

The effectiveness of the proposed method was verified by the quantitative analysis experiments of 16 certified standard brass samples. The experimental results demonstrated that taking the analytical line intensities of multiple measurements as the model input variables can effectively improve the robustness and prediction accuracy of the regression model. In the experiments, with the proposed RLS-SVM method, the root mean square error of predicted Cu concentration was 1.537% and the average relative error of the prediction was 1.73% which was better than the results obtained by the quantitative analysis methods based on PLS, SVM, LS-SVM and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

WLS-SVM. The RSD value of the predictions of the proposed RLS-SVM was smaller than the WLS-SVM model, which indicated that the proposed method had better robustness to spectral data fluctuations. Also, the results of the comparative study demonstrated that the proposed weighting scheme had better performance with a compromise among speed of algorithm convergence, model robustness and prediction accuracy.

Acknowledgements

References

[1]A. W. Miziolek,V. Palleschi, I. Schechter, Laser-Induced Breakdown Spectroscopy (LIBS): Fundamentals and Applications. Cambridge University Press, Cambridge, UK, 2006.

[2] A. M. Popov, T .A. Labutin, S .M. Zaytsev, I.V. Seliverstova, N.B. Zorov , I. A. Kal'ko, Y.N. Sidorina, I.A. Bugaev, Y. N. Nikolaev, J. Anal. At. Spectrom., 2014,29, 1925-1933.

[3] I.Y. Elnasharty, F.R. Doucet, J.Y. Gravel, P.Bouchard , M.Sabsabi , J. Anal. At. Spectrom., 2014,29, 1660-1666.

[4] D. Syvilay , A. Texier , A. Arles , B. Gratuze, N. Wilkie-Chancellier, L. Martinez, S. Serfaty, V.Detalle, Spectrochim.Acta,Part B, 2015,103,34-42.

[5] M. Kraushaar, R. Noll, H.-U. Schmitz, Appl. Spectrosc., 2003, 57, 1282-1287.

[6] E.M. Rodriguez-Celis, I.B. Gornushkin, U.M. Heitmann, J. R. Almirall, B. W. Smith, J. D. Winefordner, N. Omenetto, Anal. Bioanal. Chem., 2008,391, 1961-1968.

[7] Anna P.M. Michel, Spectrochimica.Acta,Part B,2010,65, 185-191.

[8] J.E. Carranza, D.W. Hahn, Spectrochimica.Acta,Part B, 2002,57,779-790.

[9]S. Yao, J. Lu, J. Li, K. Chen, J. Li, M. Dong, J.Anal.At.Spectrom.,2010, 25,1733-1738.

[10]B. C. Windom, D. W. Hahn, J.Anal.At.Spectrom., 2009,24,1665-1675.

[11] L. Li, Z. Wang, T. Yuan, J. Anal. At. Spectrom., 2011, 26, 2274-2280.

[12] Z. Wang, L. Li, L. West, Z. Li, W. Ni, Spectrochim.Acta, Part B, 2012,68, 58-64.

[13]N. B. Zorov, A. A. Gorbatenko, T. A.Labutin, A.M.Popov, Spectrochim.Acta,Part B, 2010, 65,642-657.

[14] J. Feng, Z. Wang, Z. Li, W. Ni, Spectrochim.Acta ,Part B, 2010,65 549-556.

[15] P.J. Huber, Robust Statistics, Wiley, New York, 1981.

[16] D.F. Andrews, P.J. Bichel, F.R. Hampel, Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, NJ, 1972.

[17] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel, Robust Statistics: The Approach Based on Influence Functions(Vol. 114), John Wiley & Sons, 2011.

[18] S. M. Zaytsev, A. M. Popov, E. V. Chernykh, R. D. Voronina, N. B. Zorov, T. A. Labutin, J.Anal.At.Spectrom.,2014,29, 1417-1424.

[19] C. D. Quarles, J. J. Gonzalez, L. J. East, J. H. Yoo, M. Morey, R. E. Russo, J.Anal.At.Spectrom., 2014, 29: 1238-1242.

[20] T. Zhang, L. Liang, K. Wang, H. Tang, X. Yang, Y. Duan, H. Li , J. Anal. At. Spectrom., 2014,29, 2323-2329.

[21] P. Inakollu, T. Philip, A.K. Rai , F. Y. Yueh, J.P. Singh, Spectrochim.Acta, Part B, 2009,64, 99-104.

[22] J. B. Sirven, B. Bousquet, L. Canioni , Anal.Bioanal.Chem.,2006,385, 256-262.

[23] C. Cortes, V. Vapnik , Machine Learning, 1995, 20, 273-297.

[24] J. A. K. Suykens, J. Vandewalle, Neural Processing Letters,1999, 9,293-300.

[25]J.A.K. Suykens, J. D. Brabanter, L. Lukas, J. Vandewalle, Neuralcomputing, 2002, 48,85-105.

[26] F. Anabitarte, J. Mirapeix, O.M. C.Portilla, J. M. Lopez-Higuera, A. Cobo ,IEEE Sens. J. ,2012,12, 64-70.

[27]D. Pokrajac, A. Lazarevic, V. Kecman, A. Marcano, Y. Markushin, T. Vance, N. Melikechi, Appl. Spectrosc., 2014, 68, 1067-1075.

[28] X. Zhu, T. Xu, Q. Lin, L. Liang, G. Niu, H. Lai, M. Xu, X. Wang, H. Li, Y. Duan ,Spectrochim. Acta Part B , 2014, 93, 8-13.

[29]J. Cisewski, E. Snyder, J. Hannig ,L. Oudejans, J. Chemometrics, 2012,26, 143-149.

[30]Y. Yu , L. Guo , Z. Hao , X. Li, M. Shen, Q. Zeng, K. Li, X. Zeng, Y. Lu, Z. Ren,Opt.Express, 2014, 22, 3895-3901.

[31] C. C. Chuang, S. F. Su, J. T. Jeng, C. C. Hsiao, IEEE Transactions on Neural Networks, 2002,13, 1322-1330.

[32] K. D. Brabanter, K. Pelckmans, J. D. Brabanter, M. Debruyne, J. A. K. Suykens, M. Hubert, B. D. Moor, Artificial Neural Networks–ICANN 2009. Springer Berlin

Heidelberg, 2009,100-110.

[33] Y. Yang, H. He, G. Xu, Journal of geodesy, 2001, 75, 109-116.

[34] NIST (National Institute of Standards and Technology), http://physics.nist.gov/ PhysRefData/ASD/levels_form.html.

[35] W. Cui, X. Yan, Chemometrics and Intelligent Laboratory Systems, 2009, 98, 130-135