X-ray fluorescence spectra quantitative analysis based on characteristic spectra optimization of partial least-squares method

Wei Zhang (章 炜)[1*], Lianfei Duan (段连飞)[1], Luozheng Zhang (张罗政)[1], Yujun Zhang (张玉钧)[2], Liuyi Ling (凌六一)[2], and Yunjun Yang (杨允军)[1]

[1]*New Star Research Institute of Applied Technology, Hefei 230031, China*
[2]*Key Laboratory of Environment Optics and Technology, Anhui Institute of Optics and Fine Mechanics, the Chinese Academy of Sciences, Hefei 230031, China*
[*]*Corresponding author: wzhang_ly@hotmail.com*

Received January 15, 2014; accepted March 15, 2014; posted online November 7, 2014

The quantitative analysis of X-ray fluorescence (XRF) spectra is studied using the partial least-squares (PLS) method. The characteristic variables of spectra matrix of PLS are optimized by genetic algorithm. The subset of multi-component characteristic spectra matrix is established which is corresponding to their concentration. The individual fitness is calculated which combines the crossover validation parameters (prediction error square summation) and correlation coefficients (R^2). The experimental result indicates that the predicated values improve using the PLS model of characteristic spectra optimization. Compared to the nonoptimized XRF spectra, the linear dependence of processed spectra averagely decreases by about 7%, root mean square error of calibration averagely increases by about 79.32, and root mean square error of cross-validation averagely increases by about 14.2.

OCIS codes: 300.6560, 020.1335.
doi: 10.3788/COL201412.S23001.

The essence of X-ray fluorescence (XRF) spectra quantitative analysis is that the concentration C_i of measured elements can be calculated by the intensity I_i of XRF characteristic spectra. The conversion relation of I_i and C_i is that real concentration = apparent concentration * correction factor[1]. Correction factor is a key to confirm analysis elements concentration accurately. There are two methods to determine the correction factor, namely, experimental correction and mathematic correction methods[2–4]. The purpose of these two methods is to eliminate the complex absorption and enhancement effect existing in components. However, the absorption and enhancement effect of components can be directly described by the overlapping spectra. So influence factors must be considered in XRF spectra quantitative analysis.

Here the correction factor can be calculated using the partial-least squares (PLS) method. The merit of this method is that it can complete the mixed multi-components analysis, characteristic spectra extraction, and regression modeling at the same time[5,6]. During this process, two aims would be achieved which are the characteristic spectra analysis and the correction of elements absorption and enhancement effect. The nonlinear influence between elements concentration and their intensity could be compensated and corrected by extracting proper latent variable[7].

Generally, integrated characteristic spectra should be introduced while XRF spectra are analyzed using the PLS regression method. So there are too many XRF spectra variables in the PLS model, when multi-components are analyzed at the same time. These spectra increase the operational complexity and calculation workload. And these are disadvantages to achieve the results quickly. Then the structure of PLS model should be predigested and spectra variable should be reduced. It is well-known that the weight matrix W of spectra expresses the importance between the each spectrum and its concentration of corresponding component. So partial spectra can be removed, which contributes to the smaller element concentration. This is a very effective method to predigest PLS regression model by removing partial characteristic spectra.

The improved individual fitness of genetic algorithm (GA) is applied to variable optimization of PLS XRF spectra matrix. These characteristic spectra that have higher correlation with the component concentration can be extracted. In order to improve the rationality of PLS concentration matrix of many samples, Kennard–Stone uniformity method can be applied to divide the training set and prediction set scientifically.

Figure 1 shows the principle of experiment system. The target of the mini X-ray tube was made using silver. Al filter was used as the filter of the target. In order to effectively excite heavy metal elements, working voltage and current were respectively set 40 kV and 150 μA. The excitation time was set 120 s. The silicon drift detector was used and energy detection ranged from 1.5 to 25 keV and the resolution ratio was from 145 to 260 eV (Mean: 5.90 keV).

A certain quantity of three kinds of chemical reagents of $Ni(NO_3)_2 \cdot 6H_2O$, $Cu(NO_3)_2 \cdot 3H_2O$, and $ZnCl_2$ were weighted by electronic scales. These three reagents were mixed symmetrically and added into the purified water having

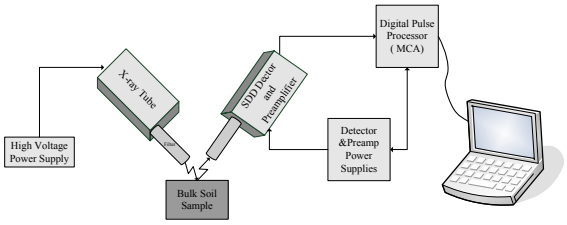

Fig. 1. Experimental setup.

a volume of 30 mL. Mixed solutions were obtained with a certain concentration of three elements. Then different volume solutions were measured by the micro-pipette enriched on the fiber glass films. These 15 fiber glass films were quickly dried by electrical heating. The concentrations of Ni, Cu, and Zn element are listed in Table 1.

GA is a kind of random searching algorithm which simulates the natural selection and genetic mechanism in living system. It is suitable for dealing with the complicated and nonlinear optimization problems. It is an ideal and overall nonlinear optimum algorithm. The algorithm flow is shown in Fig. 2.

Different from traditional searching algorithm, GA begins to search based on the initial solution which is randomly generated. The new solution can be achieved by iteration operation step by step. The operation mainly includes selection, crossover, and mutation. Each individual in the population represents a solution of the problem (namely chromosome). The fitness is used to evaluate the stand or fall of the chromosome. A partial excellent individual can be selected from their father generation according to the fitness and the filial generation can be formed by crossover and mutation. After several generations evolve, the algorithm would converge at the best chromosome which is the optimal solution or second best solution[8].

The advantage of GA is that it lets the specific problem be coded into chromosome and be optimized. This algorithm does not refer to the parameters themselves. So it is not restricted by constraint conditions. The searching process starts from one set of solution and has the characteristic of latent and parallel searching. It can greatly reduce the possibility of trapping in local minimum. Optimization computation of the algorithm does not depend on the gradient information. The target function does not require continuum and derivable functions. So it can resolve combinational optimization

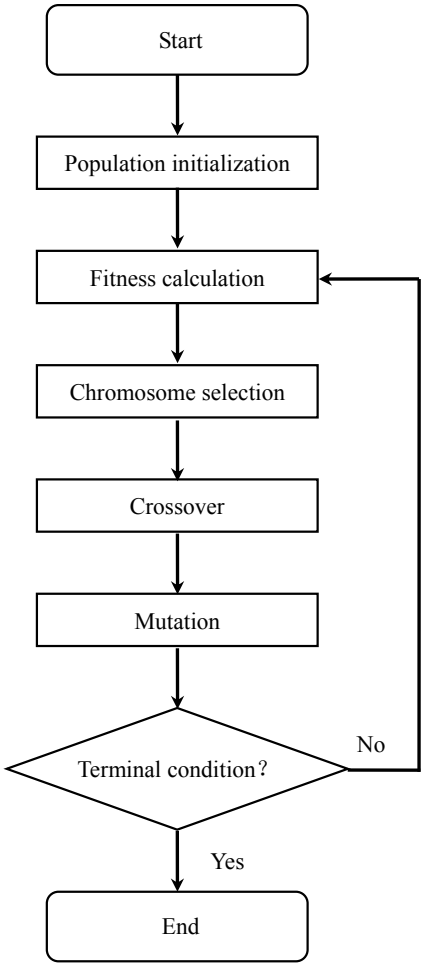

Fig. 2. GA flow diagram.

problems of large scale and nonlinear which the traditional searching method cannot resolve[9].

It can be seen from Table 1 that the concentration change in Ni, Cu, and Zn is inconformity. And with the increase in sample numbers and the enlargement of concentration change in elements, it is difficult to divide the training set and prediction set reasonably. In order to improve the accuracy of the established PLS concentration matrix, the Kennard–Stone uniformity method was used to divide the combination of 15 samples. The partition result of the training set and prediction set is shown in Table 2, when the sample number of the training set is 12.

Table 1. Concentration of Ni, Cu, and Zn Elements (Unit: μg cm^{-2})

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ni	103.16	153.26	191.58	290.96	268.21	429.55	839.32	230.25	2142.01	2276.90	2369.43	3046.41	758.96	1713.61	344.84
Cu	136.48	202.75	253.43	384.90	354.81	568.25	1110.32	280.47	2833.64	3012.09	3134.49	4030.06	1004.03	2266.91	456.18
Zn	122.58	182.11	227.63	345.72	318.69	510.41	997.29	270.29	2545.19	2705.48	2815.42	3619.83	901.83	2036.15	409.74

Table 2. Result of the Training Set and the Prediction Set

Training Set	1	3	5	6	7	9	10	11	12	13	14	15
Prediction Set	2	4	8										

Table 2 also shows that the concentration range of the training set covers the whole concentration range of the prediction set.

The initialization of prediction variables model is completed through random establishing of a series of character string of binary coding. Each character string delegates one predicted variable of XRF spectra intensity. Here 1 denotes that these prediction spectra variables are selected by the model and 0 denotes that these prediction spectra variables are not selected. The number of prediction variables could be confirmed through the statistics of the number of string 1 in the character strings.

Fitness describes the stand or fall of model performance of the corresponding individual. When the fitness is higher, the probability of the reserved individual is higher and it could be copied to the next generation. However, when the fitness is lower, the probability of the deleted individual is higher. Therefore, the evaluation of fitness decides the searching direction of GA and directly determines the performance of the algorithm. Here, the calculation method of fitness is modified. The evaluation function of fitness can be obtained according to the characteristic of multi-collinearity regression analysis of PLS.

The essence of crossover is inheritance. Two individuals are randomly selected from the population of father generation. According to a certain regulation and probability, character strings are exchanged among bits. The exchange manner of two points is applied. The crossover probability is set 50%. The bit of parameter strings is randomly forcibly exchanged during the mutation. Then the searching direction is changed and the searching space is enlarged. The probability of mutation should not be too large and is 0.005% in this experiment.

As we know, GA can achieve the global optimum searching and the PLS method can extract the principal component which has the multi-collinearity question. So the individual fitness can also be calculated using the index which expresses the performance of the PLS model during the process of the GA model. Hasegawa *et al.*[10] calculated the individual fitness using the maximum of correlation coefficients. Generally during the PLS modeling, the ability of fitting would be enforced by increasing the principal component and the correlation coefficient (r^2_{pred}) would approach 1. But this method only fits to the modeling of the selected variables. The phenomenon of excessive fitting would appear by increasing the principal component when the modeling variables are not optimized. At the same time, this method can only calculate the predicted correlation coefficients of the single component and does not consider the result of interaction of multi-components operation.

In order to avoid the phenomenon that the model would be excessively fit by increasing the principal component and considering the multi-components together, the calculation method of fitness is modified. The cross-validation parameter (prediction error square summation, PRESS) and correlation coefficients (R^2) are banded together to calculate individual fitness.

$$Q_k = \sqrt{P_k^2 * R_{\text{pred}}^2}, \tag{1}$$

where k is the number of principal components. P_k^2 can be obtained according to PRESS:

$$P_k^2 = 1 - \text{PRESS}_k/\text{PRESSMAX}. \tag{2}$$

PRESS_k is the PRESS of k principal components of multi-elements.

$$\text{PRESS}_k = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ijk} - \hat{y}_{ijk}\right)^2, \tag{3}$$

where M is the number of the elements and N is the number of the samples.

PRESSMAX is the maximum of PRESS_1, PRESS_2, PRESS_3,..., PRESS_N.

R^2_{pred} is the fitting correlation coefficient of multi-components.

$$R^2_{\text{pred}} = 1 - \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij,\text{obs}} - \hat{y}_{ij,\text{pred}}\right)^2 / \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij,\text{obs}} - \overline{y}_j\right)^2, \tag{4}$$

$y_{ij,\text{obs}}$ is the element concentration of training set, $\hat{y}_{ij,\text{pred}}$ is the element concentration of prediction set and $\overline{y}_j = \sum_{i=1}^{N} y_{ij}/N$.

The number of the principal components can be confirmed by the method of cross-effective validation[11]. The number would be increased if $\text{PRESS}_k/\text{PRESS}_{k-1} \leq 0.95^2$.

In order to decrease the population size and shorten the iteration time in each inheritance, parameters are set as follows: maximum generation is 11, population size is 32, crossover probability is 50%, and mutation probability is 0.5%. "Leaving three" cross-validation method of PLS regression is used to setting fitness. The maximum principal components are 12. Iteration operation is applied once in PLS regression. The characteristic spectra of Ni, Cu, and Zn elements of training set in Table 2 are extracted by means of the above genetic parameters. The result of fitness is shown in Fig. 3.

Figure 3 shows the variable number and fitness of each population in 32 populations, when the generation reaches 11. The number of selected spectra concentrates between 63 and 80 in these 32 populations and the minimum fitness can reach 165.52.

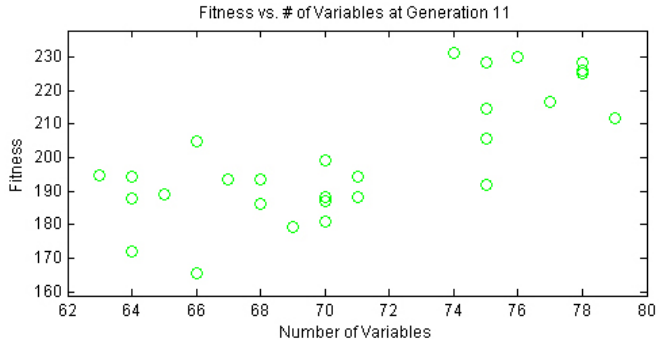

Fig. 3. Fitness at generation 11.

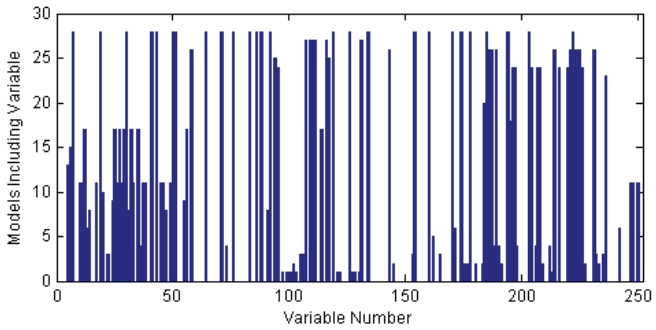

Fig. 5. Models with variable at generation 11.

Figure 4 shows that using variable number of different generation has a certain fluctuation. As a whole, it can be seen that the variable number presents downtrend when the reproduction generation exceeds eight. The number reaches minimum when the reproduction generation get to 11.

Figure 5 shows that GA could finally confirm the selection time of each XRF spectra according to their importance. The more a spectrum is selected, the more it is important, vice versa.

Although characteristic spectra can be optimized by increasing the number of generation, the maximum genetic algebra should not be too big. Otherwise, the phenomenon of excessive fitting would happen which will make partial useful XRF spectra lose. It can be seen from Figs. 6 and 7 that the information reflecting the character of spectra peak has lost more when the spectra have eliminated more from generation 51 to 101, which is the disadvantage to construct PLS prediction model.

PLS regression model can be constructed and elements concentration of prediction set can be calculated after the XRF characteristic spectra are optimized. Figure 8 shows that the characteristic spectra information extracted from the elements of training set takes up 99.45% of whole spectra when the principal components are 3. Then it can meet the requirement when the principal components are 3. According to the PLS model, the concentration of prediction set is calculated and the results are shown in Table 3. There is a phenomenon that the error between predicted value and standard value of No. 2 sample is largest among three samples. The phenomenon is mainly caused by uneven sample enrichment. The linearity fitting relationship of Ni, Cu, and Zn elements can be obtained according to their real values and predicted values. The results are shown in Fig. 9. It can be seen that the correlation coefficient is bigger than 0.99 after spectra correction and optimization. This explains that they have good correlation.

The score figure of the first and second principal components in Fig. 9 also expresses that there are no abnormal values during the 95% confidence level in the whole data set (including training samples and prediction samples).

When XRF spectra are not optimized, the calculation results of Ni, Cu, and Zn are shown in Fig. 10. Comparing with Fig. 8, the correlation of three elements averagely decreases by about 7%, root mean square error of calibration (RMSEC) averagely increases by

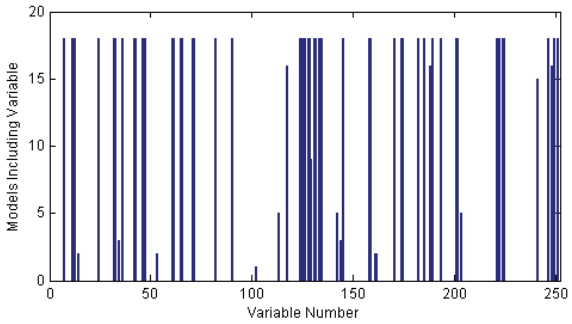

Fig. 6. Models with variable at generation 51.

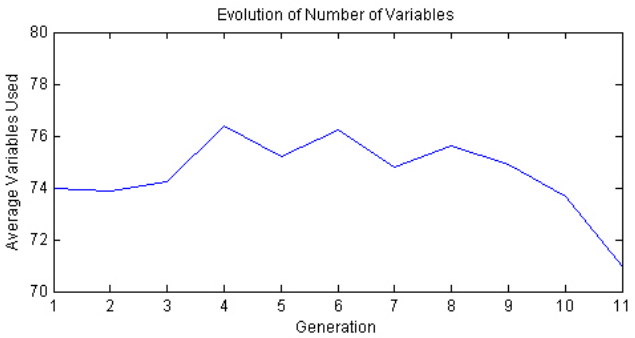

Fig. 4. Average variables used in each generation.

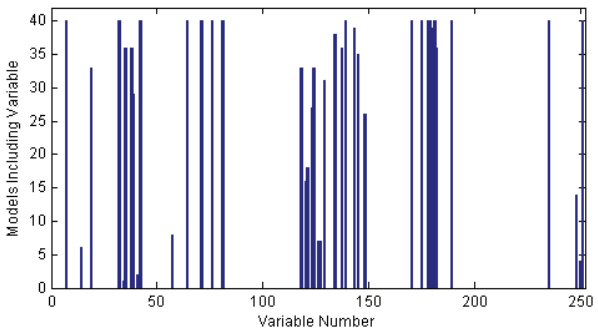

Fig. 7. Models with variable at generation 101.

Table 3. Prediction Results and Errors of Ni, Cu, and Zn

	Index	Ni	Cu	Zn
	Standard Value	153.26	202.75	182.11
No. 2	Predicted Value	102.46	153.28	139.68
	Error	50.80	49.47	42.43
	Standard Value	290.96	384.90	345.72
No. 4	Predicted Value	267.58	354.35	315.77
	Error	23.38	30.55	29.95
	Standard Value	230.25	280.47	270.29
No. 8	Predicted Value	213.05	261.92	249.05
	Error	17.20	18.55	21.24

about 79.32, and root mean square error of cross-validation (REMSECV) averagely increases by about 14.2. This also explains that optimized XRF characteristic spectra are benefits to improve the prediction ability of PLS regression model.

In conclusion, in order to decrease the influence of elements absorption and enhancement effect during XRF spectra quantitative analysis, PLS method is applied to analyze the concentration of metal elements enriched on fiber glass film. The aim of this method is to extract partial characteristic spectra which have the best relevance with the elements concentration. By this means, the multi-elements characteristic spectra matrix is predigested during the process of PLS modeling. The fitness is modified according to the cross-validation parameter (PRESS) and correlation coefficients (R^2). This method can effectively avoid the phenomenon of excessive fitting by increasing the principal component.

The experimental result indicates that optimized XRF characteristic spectra are benefits to improve the prediction ability of PLS regression model. This method is best fit for XRF overlapping spectra analysis.

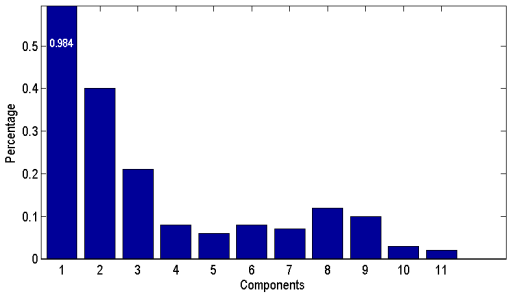

Fig. 8. Principal components and characteristic information proportion.

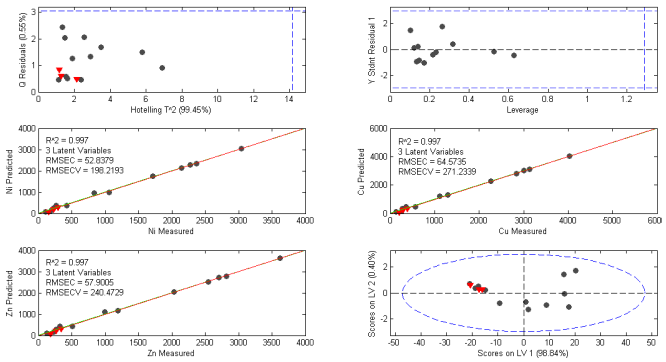

Fig. 9. Contrast of predicted values and real values of three elements and the first and the second principal components of PLS model for optimized XRF spectra.

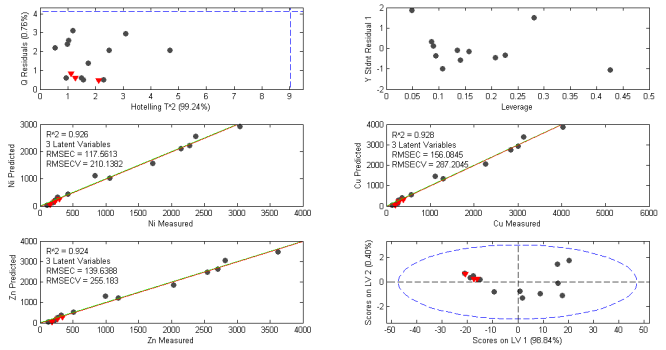

Fig. 10. Contrast of predicted values and real values of three elements and the first and the second principal components of PLS model for non-optimized XRF spectra.

This work was supported by the Project of the Academic Fund (No. 2013XYJJ-008), the Science and Technology Program of Anhui province (No. 1206c0805012), and the National "863" Program (No. 2013AA065502).

References

1. A. Ji, G. Y. Tao, S. J. Zhuo, and L. Q. Luo, *X-ray Fluorescence Spectra Analysis* (Science Press, 2003).
2. R. Jenkins, *X-Ray Fluorescence Spectrometry* (Wiley, 1999).
3. N. Tsoulfanidis, *Measurement and Detection of Radiation* (Taylor & Francis, 1995).
4. Y. Liang, *Spectral Analysis Foundation of XRF* (Science Press, 2007).
5. H. Wold, *Path Models with Latent Variables: The NIPALS Approach, Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building* (Academic Press, 1975).
6. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, SIAM J. Sci. Stat. Comput. **5,** 735 (1984).
7. D. M. Haaland and E. V. Thomas, Am. Chem. Soc. **60,** 1192 (1988).
8. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, 1989).
9. A. Popov, *Genetic Algorithms for Optimization, User Manual* (Hamburg, 2005).
10. K. Hasegawa, Y. Miyashita, and K. Funatsu, J. Chem. Inf. Comput. Sci. **37,** 306 (1997).
11. V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang, *Handbook of Partial Least Squares* (Springer-Verlag, 2010).