International Journal of Robotics and Automation, Vol. 36, 2021

Special Issue: Artificial Intelligence in Environmental Automation Systems

AN IMPROVED BOOSTING-BIPLS MODELS BASED ON WEIGHT ADJUSTMENT FOR SOIL HEAVY METAL CONTENT PREDICTION

Dong Ren,*,** Jun Shen,*,** Shun Ren,*,**,*** Kai Ma,*,** and Xinting Yang*,**,***

Abstract

Heavy metal pollution in soil has got more and more attention, and X-ray fluorescence spectroscopy analysis is a widely used method for heavy metal content in soil. The establishment of accurate model is helpful for the rapid detection of heavy metal content. Firstly, eight spectrum pretreatment methods are used before modelling, and the pre-processing method of least squares which improved multi-scatter correction is chosen. Secondly, Boosting-backward interval partial least squares (Boosting-BiPLS) model is established which combines several basic models with different characteristics into a strong one to solving the "building nesting effect" of BiPLS. Then from bias-oriented model, an improved Boosting-BiPLS model is proposed, in which the weight of samples is adjusted on the basis of the relative deviation of the samples and the weight of base models is dynamically calculated by the spectral similarity. Finally, to prove the effectiveness of the improved model, the improved Boosting-BiPLS model is compared with the traditional Boosting-BiPLS model. The results show that the correlation coefficients of the five heavy metal elements of the improved Booting-BiPLS model are all about 0.99, and the average relative deviations are all <10%, with the prediction accuracy of Boosting-BiPLS improved by more than 50%. Moreover, the model is more stable.

Key Words

XRF, heavy metal, Boosting-BiPLS, spectral similarity

1. Introduction

A large number of quantitative models for soil heavy metal detection based on XRF spectroscopy have been studied, including one-dimensional linear regression model [1], [2], multiple linear regression model [3], partial least squares (PLS) regression model [4], and support vector machine (SVM) regression model [5]. Back Propagation neural network [6] is also used in a small amount in the heavy metal content detection model. Linear regression is simple to calculate, but it is easy to produce the problem of underfitting. PLS is widely used to solve the problem of multicollinearity among small sample variables, but the data information may be lost when the dimension is reduced. SVM has low generalization error, and easy to explain, but it is sensitive to the selection of parameters and kernel functions. BP neural network has the ability of nonlinear mapping, self-learning and self-adaptation, but its disadvantages are that it is easy to fall into local minimum value, the network convergence speed is slow, there are too many network structure parameters and so on [7]–[10].

The integrated learning model can organically combine multiple single learning models to obtain a more accurate, stable and robust model. In recent years, the integrated learning model has been widely used in near-infrared spectroscopy [11]–[13]. However, in the detection of heavy metal content based on X-ray fluorescence (XRF) spectroscopy, integrated learning method is currently used. The research of modelling is still relatively rare, especially the research on the fusion of integrated learning strategies for spectral variables.

At present, the research on integrated learning is mainly applied to classification. Experts and scholars have done a lot of research on the weight of classifiers in terms of weighing differences and accuracy. They can be roughly divided into the following categories: classifier weight optimization based on difference metrics [14]–[16], classifier weight optimization based on difference and accuracy [17]–[20], and weight optimization based on classifier credibility [21].

In this paper, XRF spectroscopy is used as the research object, and adopt the Boosting integration idea. According to the discomfort of the test sample and the base model, a spectral similarity measure on the basis of the similarity between the test sample and the training sample spectrum

* College of Computer and Information Technology, Three Gorges University, Yichang 443002, China; e-mail: rendong5227@163.com, shenjun@qq.com, {renshun, makai}@ctgu.edu.cn, yangxt@nercita.org.cn

** Hubei Engineering Technology Research Center for Farmland Environmental Monitoring, China Three Gorges University, Yichang 443002, China

*** National Engineering Laboratory for Agri-Product Quality Traceability, Beijing 100097, China

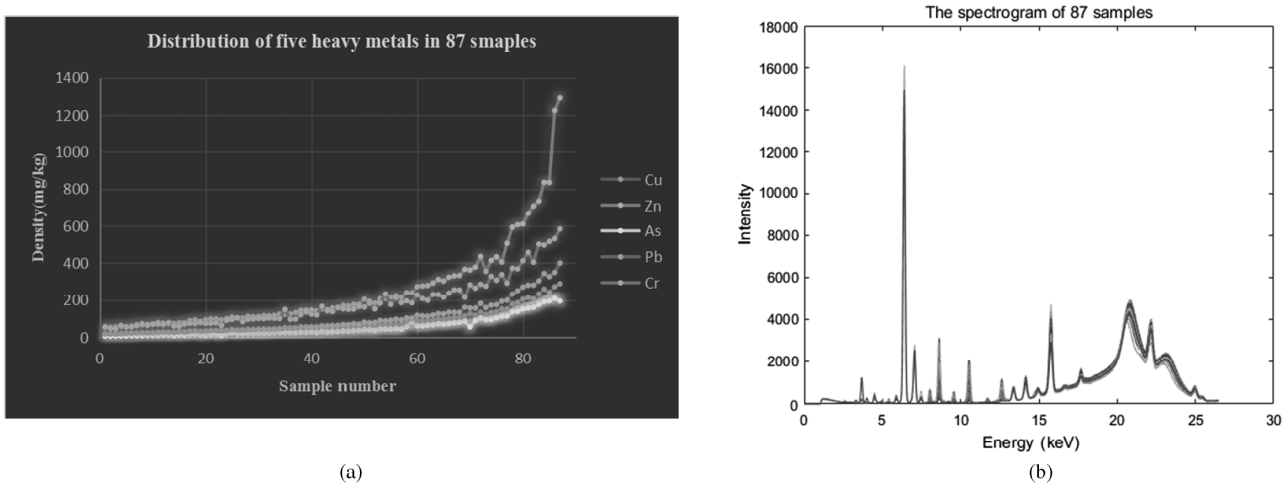
Corresponding author: Shun Ren

Recommended by Dr. Dong Ren

(DOI: 10.2316/J.2021.206-0615)

Table 1
Soil Samples Configuration (unit: 0.1‰)

Soil Sample Label	Heavy Metal Element				
	1–40	41–50	51–60	61–70	71–91
As element concentration	0.6	2×0.6	3×0.6	4×0.6	$J \times 0.6$
Cu element concentration	0.8	2×0.8	3×0.8	4×0.8	$J \times 0.8$
Cr element concentration	1.5	2×1.5	3×1.5	4×1.5	$J \times 1.5$
Pb element concentration	1.0	2×1.0	3×1.0	4×1.0	$J \times 1.0$
Zn element concentration	2.5	2×2.5	3×2.5	4×2.5	$J \times 2.5$



(a) (b)

Figure 1. AAS determination concentration profile (a) and average spectrum of 87 samples (b).

is proposed to improving the traditional Boosting-BiPLS model. Then the improved Boosting-BiPLS model is compared with the traditional Boosting-BiPLS in accuracy and stability.

2. Materials and Data

2.1 Sample Production

To improve the detection performance of heavy metals in the instrument, 91 samples were taken from farmland with no pollution source around 1,000 m. Then removing weeds, roots, stones, and other debris, using ceramic utensils for mixing, crushing, grinding, and pack with 100-mesh nylon sieve. According to the national standard [22], the content of each heavy metal in the first, second, and third grade soils is determined according to the certain content. Concentration gradients were added dropwise to different volumes of standard solutions of Cu, Pb, Zn, Cr, and As, and thoroughly stirred to make them evenly mixed. The specific soil samples concentration configuration scheme is shown in Table 1.

2.2 Data Collection

Considering the human error and instrument error in the actual configuration process, the concentration of five heavy metal elements was determined by atomic absorption spectroscopy (AAS) after samples configuration were completed. Then the prepared samples were placed on the instrument test bench for spectral scanning, and each sample was scanned three times. After removing four abnormal samples, the concentration distribution and average spectrum of the remaining 87 samples are shown in Fig. 1.

The training samples and test samples are divided by the concentration gradient method. The method divides the samples into two groups according to the chemical reference value of the measured samples, including 58 training samples and 29 test samples. The distributions are shown in Fig. 2.

2.3 Spectral Pretreatment

In addition to the useful chemical information, the spectrum also contains a large amount of interference information such as background noise and irrelevant information. Therefore, the pre-processing of the spectrum before the established calibration model can not only remove the influence of unrelated factors on the target spectrum but also improve the robustness and prediction accuracy of the model.

Eight spectrum pretreatment methods are used, including PLS improved multi-scatter correction (PLSMSC)

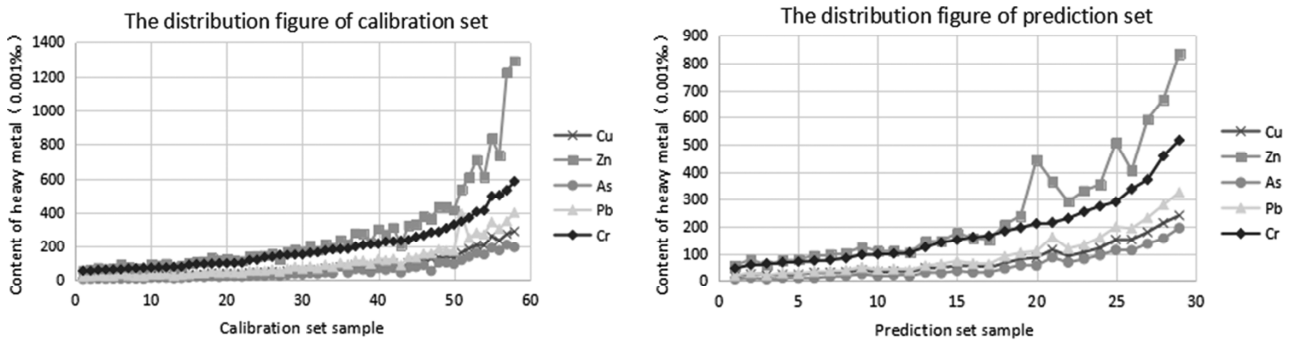


Figure 2. Training samples distribution (a) and prediction samples distribution (b).

Table 2
Spectral Pretreatment Results

Method	Unprocessed	SNV	WT	DT	MSC	SNV+DT	SG+1-Der	SG+2-Der	PLSMSC
R^2	0.988	0.988	0.986	0.692	0.990	0.772	0.986	0.978	**0.986**
RMSEP	20.809	20.76	21.44	90.18	19.55	79.58	21.81	27.22	**19.05**
MRD	0.166	0.143	0.143	0.73	0.138	0.658	0.136	0.197	**0.121**

[23], detrended processing (DT), standard normal variable (SNV) transform, multiple scattering correction (MSC), wavelet denoising (WT), SNV+DT, convolution smoothing (SG) + first derivative, convolution smoothing (SG) + second derivative, and BiPLS model is built and compared. The results are shown in Table 2. According to the analysis of Table 2, the correlation coefficient, root mean square error of prediction (RMSEP), and mean relative deviation (MRD) with PLSMSC method are 0.986, 19.051, and 0.121, respectively, which are better than other conventional pretreatment methods.

3. Adaptive Boosting Integration Method

3.1 Boosting Integration Framework

Boosting is an integrated learning framework, its core idea is to combine multiple weak classifiers with different strategies into a strong classifier. The key of constructing an integrated model is to adjust the weight of samples and weight of base models.

This paper is bias oriented, and the weight improvement methods of samples and base model are proposed. The weight of samples is adjusted on the basis of the relative deviation of the samples, and the weight of base models is dynamically calculated by the spectral similarity of the test sample and the training samples. The specific adaptive Boosting integration framework is shown in Fig. 3.

3.2 Sample Weight Adjustment Strategy

Through previous research and analysis, it is found that the deviations of the single PLS model and the variable interval selection model are somewhat large. Therefore, the sample

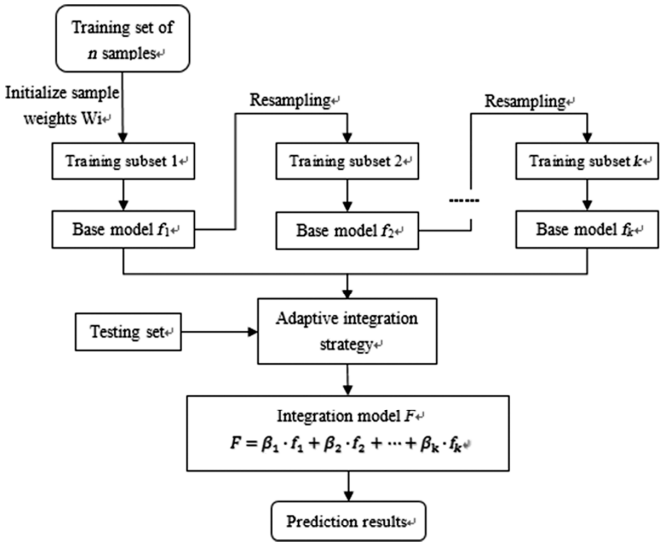


Figure 3. Adaptive boosting integration framework.

weight adjustment strategy in this study is adjusted on the basis of the relative deviation of the sample. The specific sample weight adjustment strategy is shown as follows.

Sample weight adjustment algorithm

1. The samples need to select: $L^{\mathrm{t}}=\{(x_1, y_1), \ldots, (x_N, y_N)\}$, $t = 1$
2. Initialization iteration number T, initialization sample weights $W_0 = 1/N$
3. While $t<=T$
4. Calculating sampling probability: $P_i^t = W_i^t \Big/ \sum_{i=1}^{N} W_i^t, t = 1, 2, \cdots, T$

5. Roulette sampling: $L' = resample(L^t, P^t)$
6. Taking L′ as a new sample, call the base model f_t
7. Using base model f_t to predicts training samples, calculate the prediction error of the sample:

$$L_i^t = \frac{\left|y_{real(i)}^t - y_{pred(i)}^t\right|}{\max\left|y_{real(i)}^t - y_{pred(i)}^t\right|}, i = 1, 2, \ldots N$$

8. Calculating weighted error sum: $\bar{L}_t = \sum_{i=1}^{N} L_i^t P_i^t$
9. Calculating common indicators: $\beta_t = \frac{\bar{L}_t}{1-\bar{L}_t}$
10. Calculating the new weight of the sample: $W_i^{t+1} = W_i^t \beta_t^{(1-\bar{L}_t)}$
11. Using the sample weight adjustment algorithm to get the next round of training samples L'
12. End

3.3 Base Model Weight Adjustment Strategy

The weight of the base model is a very important step in the Boosting-BiPLS. In this study, to reduce the deviation between the predicted value and the actual value, improve the model weight adaptive ability and the participation of effective information, an adaptive base model weight adjustment strategy on the basis of spectral similarity measure is proposed. First, a number of base models are trained according to the sample weight adjustment strategy sampling. Secondly, the training samples are clustered, and the weight matrix of the cluster is calculated for each class through the trained base model. Then, each test sample and class are calculated. The spectral similarity of the cluster centroids results in a similarity matrix. Finally, the base model weight matrix of each sample is obtained by multiplying the similarity matrix and the weight matrix and normalizing the matrix. The specific base model weight adjustment strategy is shown in Fig. 4.

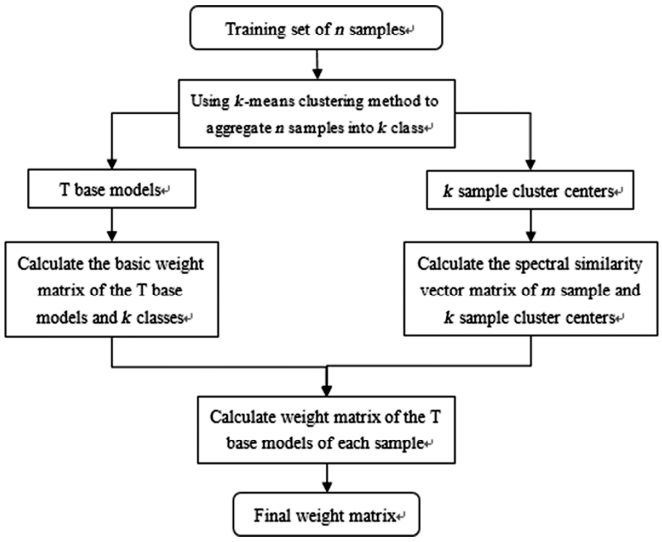


Figure 4. Base model weight adjustment strategy flow chart.

4. Experimental Results and Discussion

4.1 Selection Parameters

In this paper, the K-means clustering method on the basis of Euclidean distance is used to cluster 58 training samples into 3 categories. The clustering results are shown by Zn elements, the samples distribution are shown in Table 3. According to Table 3, the spectrum of the training samples is mostly concentrated in category 2. The concentration of Zn is concentrated between 53.1396 and 434.0840. The difference between category 1, category 2, and category 3 is very obvious. This shows that the clustering method can effectively divide the training samples into three categories.

The number of principal components and the number of base models in this paper are the Boosting-PLS model parameters that have been cross-validated in the previous period, *i.e.*, the number of principal components is set to 4, and the number of base models is set to 38.

4.2 Performance Analysis

In this paper, weighted average and weighted median are selected as traditional Boosting-BiPLS model to calculate heavy metal content, and the improved Boosting-BiPLS model is compared with the traditional model. The modelling results of the three models are shown in Fig. 5.

It can be seen from Fig. 5 that the RMSEP of five heavy metal elements calculated by the weighted average method is large, the relative average deviation is about 24%–50%, and the volatility is also fierce. The five heavy metal elements with the weighted median have been greatly improved in correlation coefficient, RMSEP, MRD, and standard deviation (STD) than weighted average.

The weight adjustment strategy proposed in this paper is performed best of the three models, the correlation coefficient of five heavy metal elements is all about 0.99, the MRD is reduced to below 10%, and the fluctuation of model deviation is less sexual. According to the data analysis and summary, the improved Boosting-BiPLS model on the basis of spectral similarity calculation solves the incompatibility of the test sample and the base model, and improves the utilization of global effective information and the accuracy of the model.

4.3 Stability analysis

Because the training samples of improved Boosting-BiPLS are obtained through roulette, the training samples collected in each round are not necessarily the same and the trained model is slightly different. To verify the stability of the model, the model runs 50 iterations for observation and analysis. The results of the five heavy metals generated in each iteration are shown in Figs. 6 and 7. It can be seen from the figure that Cu, As, and Pb are relatively stable, and due to the large concentration gradient, Zn and Cr have larger fluctuations than Cu, As, and Pb, but in general the five heavy metals are relatively stable. Combined with Figs. 6 and 7, the improved Boosting-BiPLS is a stable model.

Table 3
Cluster Data Distribution Table of Training Samples

Sample Category	Number of Samples	Density Range of the Zn	Average Value	Standard Deviation
Category 1	9	405.3560–1,292.9400	769.3733	305.0197
Category 2	38	53.1396–434.0840	166.1574	103.5492
Category 3	11	52.7382–272.9100	161.0032	69.4980

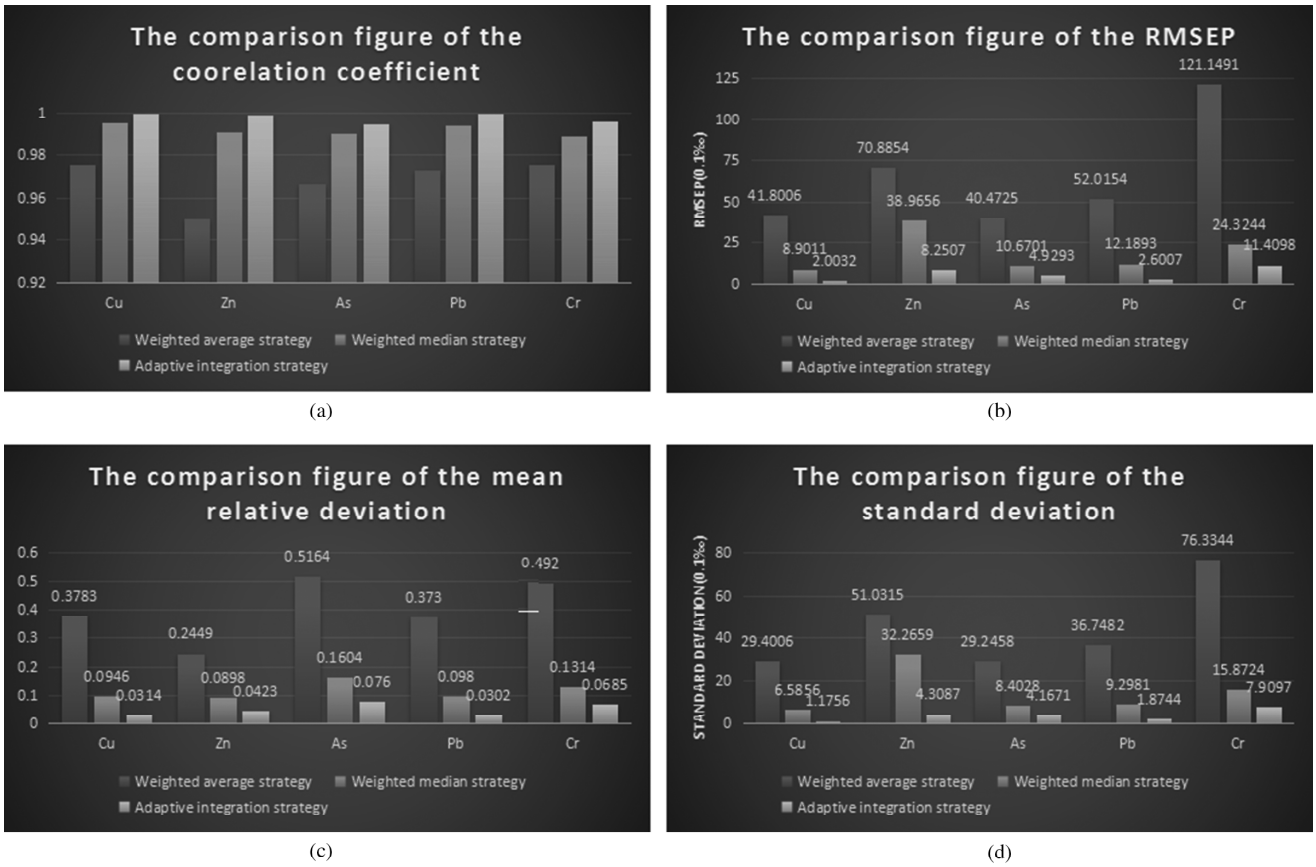


Figure 5. R (a), RMSEP (b), MRD (c), and STD (d) of the three models.

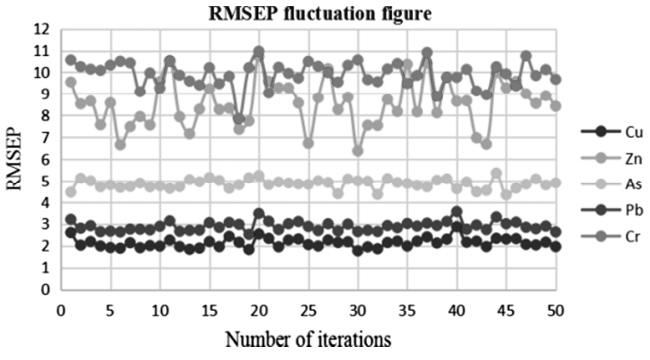


Figure 6. RMSEP fluctuation figure.

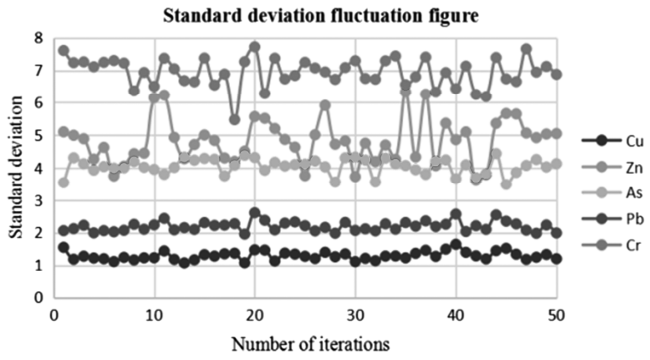


Figure 7. SD fluctuation figure.

4.4 Modelling Accuracy Analysis

To further verify the accuracy of the improved Boosting-BiPLS model on the basis of the spectral similarity, Fig. 8 is analysed from fitting results. Figure 8 shows the fit of the predicted and actual values of the five elements. It can be seen from the figure that the sample points of the five elements are all distributed near the regression curve, this illustrates that the predicted values and actual values are better fitted.

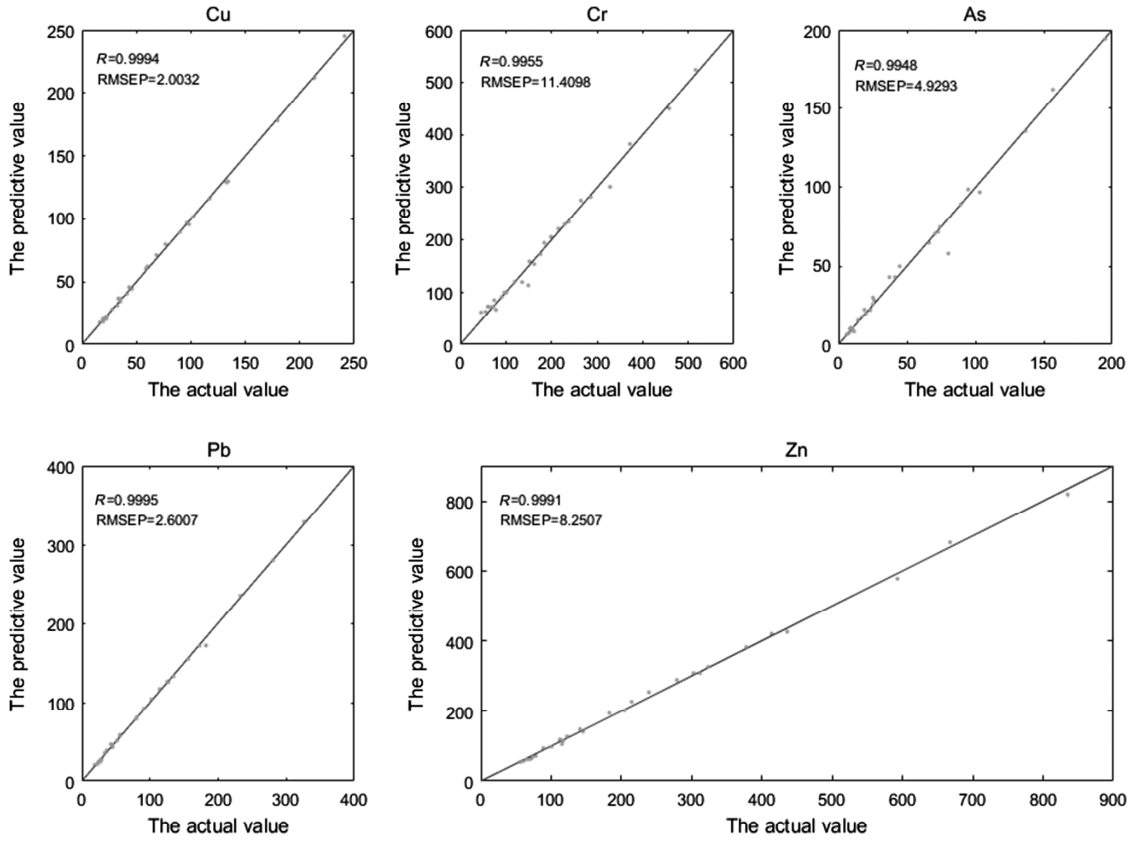


Figure 8. Fitting map of predicted and actual values of five heavy metal elements.

5. Conclusion

To solving the "building nesting effect" of BiPLS, taking Boosting integration idea, Boosting-BiPLS model is established. Then from bias-oriented model, an improved Boosting-BiPLS model is proposed, in which the weight of samples is adjusted on the basis of the relative deviation of the samples and the weight of base models is dynamically calculated by the spectral similarity. To prove the effectiveness of the improved model, weighted average and weighted median are selected as traditional Boosting-BiPLS model to calculate heavy metal content, and the improved Boosting-BiPLS model is compared with the traditional model. The experimental results show that the improved Booting-BiPLS model performs better than the traditional Boosting-BiPLS model, and the improved Boosting-BiPLS model on the basis of spectral similarity metric has higher prediction accuracy and stability, and can be used for online real-time detection of heavy metals in soil.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2016YFD0800902), Major Technology Innovation Hubei Province of China (Grant No. 2017ABA157), the Open Project Program of National Engineering Laboratory for Agri-product Quality Traceability (Grant No. GF-NAZS-2019002), and the Open Fund of the Hubei Engineering Technology Research Center for Farmland Environmental Monitoring (Grant No. 201603).

References

[1] S. Kotsiantis and D. Kanellopoulos, Combining bagging, boosting and random subspace ensembles for regression problems, *International Journal of Innovative Computing, Information and Control*, *8*(6), 2012 3953–3961.

[2] G. Yang, Z. Shang, X. Ni, and M. Zhang, Application of portable X-ray fluorescence spectrometry in rapid detection of soil heavy metals, *Applied Chemical Industry*, *45*(08), 2016, 1586–1591.

[3] L. Xiang and Z. Ma, Quantitative analysis of laser-induced metallic induced breakdown spectra of soil based on different chemometric methods, *Spectroscopy and Spectral Analysis*, *37*(12), 2017, 3871–3876.

[4] S. Wang, P. Han, J. Wang, A. Lu, and F. Li, Advances in the application of X-ray fluorescence spectrometry in the detection of heavy metals in soils, *Journal of Food Safety & Quality Testing*, *7*(11), 2016, 4394–4400.

[5] F. Li, A. Lu, and J. Wang, Establishment of X-ray fluorescence spectral heavy metal detection model based on support vector machine, *Analytical Instrumentation*, *47*(04), 2016, 68–73.

[6] G. Cheng, J. Li, and Z. Dai, Application of BP neural network in soil heavy metal pollution analysis, *Journal of Geology*, *41*(03), 2017, 394–400.

[7] D. Ren, C. Zhang, S. Ren, Z. Zhang, J.-H. Wang, and A.-X. Lu, An improved approach of cars for Longjing tea detection based on near infrared spectra, *International Journal of Robotics & Automation*, *33*(1), 2018, 97–103.

[8] P. Shao, W. Shi, P. He, M. Hao, and X. Zhang, Novel approach to unsupervised change detection based on a robust semi-supervised FCM clustering algorithm, *Remote Sensing*, *8*(3), 2016, 1–25.

[9] L. Meng, T. Dong, and W. Zhang, Drought monitoring using an Integrated Drought Condition Index (IDCI) derived from multi-sensor remote sensing data, *Natural Hazards*, *80*(2), 2016, 1135–1152.
[10] K. Ma, J. Wang, Z. Chen, X. Zhu, and L. Pan, A method for extending the geostatistical functions in spatial information processing, *Intelligent Automation and Soft Computing*, *22*(2), 2016, 261–266.
[11] Z. Chen, Z. Wu, X. Shi, B. Xu, N. Zhao, and Y. Qiao, A study on model performance for ethanol precipitation process of Lonicera japonica by NIR based on Bagging-PLS and Boosting-PLS algorithm, *Chinese Journal of Analytical Chemistry*, *42*(11), 2014, 1679–1686.
[12] D. Ren, F. Qu, K. Lv, Z. Zhang, H. Xu, and X. Wang, A gradient descent boosting spectrum modeling method based on back interval partial least squares, *Neurocomputing*, *171*, 2015, 1038–1046.
[13] H. Zhang, X. Li, W. Fan and Y. Liang, Fast measurement of protein content in milk powder by NIR combined with boosting-PLS, *Computers and Applied Chemistry*, *27*(9), 2010, 1197–1200.
[14] Y. Xu, Y. Wang, and Z. Zhao, Fast integration method of strong classifier based on instance, *Computer Applications*, *37*(04), 2017, 1100–1104.
[15] Y. Zhang, R. Dou, S. Zhao, and Z. Cao, Regularized linear discriminant analysis for face recognition based on ensemble learning, *Computer Engineering*, *36*(14), 2010, 144–146.
[16] M. Liu and H. Xie, An integrated collaborative training algorithm based on rotating forest, *Computer Engineering and Applications*, *47*(30), 2011, 172–175.
[17] T. Wang, Z. Duan, and W. Li, Adaptive integrated modeling of oil well surface based on improved AdaBoost, *Journal of Electronic Measurement and Instrument*, *31*(8), 2017, 1342–1348.
[18] Y. Wu and X. Yan, Integrated neural network based on instant learning and its dry point prediction, *Journal of East China University of Science and Technology: Natural Science Edition*, *42*(5), 2016, 696–701.
[19] X. Yao, X. Wang, Y. Zhang, and Y. Xing, A self-adaption ensemble algorithm based on random subspace and AdaBoost, *Chinese Journal of Electronics*, *41*(4), 2013, 810–814.
[20] X. Yao, X. Wang, Y. Zhang, and L. Lei, Selective ensemble algorithm based on AdaBoost and matching pursuit, *Control and Decision*, *29*(2), 2014, 208–214.
[21] S. Liu, T. Liu, and Z. Wang, Data stream ensemble classification based on classifier confidence, *Journal of Applied Sciences*, *35*(2), 2017, 226–232.
[22] Ministry of Environmental Protection of the People's Republic of China, GB 15618-1995, *Environmental quality standard for soil* (Beijing: Standards Press of China, 1995.1)
[23] D. Ren, J. Shen, S. Ren, J. Wang, and A. Lu, Study on X-ray fluorescence spectrometry pretreatment method for detection of heavy metal content in soil, *Spectroscopy and Spectral Analysis*, *38*(12), 2018, 3934–3940.

Biographies

Dong Ren received his Ph.D. degree in Jilin University. He is a Professor in the College of Computer and Information Technology at China Three Gorges University, Yichang, China. His research interests include artificial intelligence, pattern recognition, and 3S technology.

Jun Shen received his B.S. degree in China Three Gorges University. He is now a master's degree candidate of Institute of Collaborative Innovation Center for Key Technology of Smart Irrigation District in Hubei, China Three Gorges University. His main research interests are spectroscopy analysis and pattern recognition.

Shun Ren received his Ph.D. degree in Jilin University. Now he is working as a lecturer in the College of Computer and Information Technology at China Three Gorges University, Yichang, China. His research interests include artificial intelligence, Internet of things, and wireless sensor network.

Kai Ma received his Ph.D. degree in China University of Geosciences. Now he is working as an Associate Professor in the College of Computer and Information Technology at China Three Gorges University, Yichang, China. His research directions include spatio-temporal big data, knowledge mining, and intelligent environment monitoring.

Xinting Yang received his Ph.D. degree in Research Center for Eco-Environmental Science, Chinese Academy of Sciences. He is a Researcher in Beijing Research Center for Information Technology in Agriculture, Beijing, China. His research interests include agricultural information technology, agricultural products safety and traceability technology.