JAAS

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/jaas

JAAS

RSCPublishing

PAPER

Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF)

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Liwen Sheng,[a] Tianlong Zhang,[a] Guanghui Niu,[c] Kang Wang,[b] Hongsheng Tang,[a] Yixiang Duan[d] and Hua Li[*a]

Laser-induced breakdown spectroscopy (LIBS) integrated with random forest (RF) was developed and applied to the identification and discrimination of ten iron ore grades. The classification and recognition of the iron ore grade was completed by their chemical properties and compositions. In addition, two parameters of the RF were optimized by out-of-bag (OOB) estimation. Finally, support vector machines (SVM) and RF machine learning methods were evaluated comparatively on their ability to predict unknown iron ore samples using models constructed from a predetermined training set. Although results show that the prediction accuracies of SVM and RF models were acceptable, RF exhibited the better predictions of classification. The study presented here demonstrates that LIBS–RF is a useful technique for the identification and discrimination of iron ore samples, and is promising for automatic real-time, fast, reliable, and robust measurements.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS), an emerging atomic emission spectroscopic (AES) technique, was developed in the late 20th century. This technique uses a pulsed laser with a high peak power (>1 GW/cm^2) which generates laser-induced breakdown plasma that contains electronically excited atoms, ions, and small molecules to ablate material from the surface of the sample. Because the excited species emit unique spectral light peaks as they relax to lower electronic states, the emitted elements (even many minor and trace elements) can be identified by spectral light peaks. The emission lines are exploited as spectral signatures that identify the existence of characteristic elements in the sample.[1] Thus, each material with different chemical compositions has a unique LIBS signature can be recognized from its broadband spectrum. The technique can be applied to samples in any physical state, such as solids, liquids, and gases. In addition, compared with the conventional analytical techniques, LIBS technique bears some advantages[2-4]: (1) highly advanced analysis of all types of samples, (2) simple operation procedure, (3) without any sample preparation, (4) rapid and real-time measurement, (5) simultaneous multi-element assay detection, and (6) remote detection. Thus, LIBS technique has been widely used in various fields, cultural heritage,[5] industrial analysis,[6,7] environmental monitoring,[8,9] security and forensics,[10] biomedical analysis,[11,12] space exploration,[13] and mineral analysis.[14-17] The capacity for little sample preparation is one of the most important advantages of LIBS in beneficiation technology applications.

Iron ore is the most significant raw material in the field of iron and steel industry. In recent years, quality control and process analysis is an indispensable step in the iron and steel industry. It is a significant way to control quality by identification and classification of iron ore grades. Also, it is conducive to beneficiation and mineral separation. Iron ores are identified in accordance with some measurements as follows: optical image analysis,[18] X-ray diffraction (XRD),[19] etc. The classification by chemical composition of the iron ore is a familiar method for quality control and process analysis. The traditional identification methods are time-consuming and require complicated sample preparation; therefore they cannot provide and feedback quality information of the iron ore in steelmaking in time. However, the iron ore grade can be rapidly classified using the LIBS technique because of its advantages.

LIBS technique produces several thousands to tens of thousands of variables per spectrum.[20] Multivariate analysis as a part of chemometrics can make high dimensional data to lower dimensional factors that describe the variance among samples. The application of chemometrics methods coupled with LIBS of iron ores has drawn attention over the years. The traditional chemometrics approaches, such as partial least squares discriminant analysis (PLS-DA),[21] soft independent modelling of class analogy (SIMCA),[22, 23] principle component analysis (PCA),[24] artificial neural network (ANN)[25] and independent component analysis (ICA), [26] etc. have been used for classification. Thus, these approaches are suitable for the rapid analysis and classification. Random Forest (RF) as a new classification algorithm based on multiple classifiers was originally developed by Leo Breiman[27] in 2001. It can overcome the drawback of low accuracy and excessive fitting compared with the traditional classification algorithms. It has been proved that RF classifier has a good tolerance for the noise through a variety of ways by many researchers, such as blasting engineering,[28] tea identification,[29] cut

tobacco classification,[30] etc. In addition, RF was used to apply to LIBS technique. Jeremiah Remus et al[31] proposed an approach of LIBS and RF to classify five different materials (four rock samples and one pen ink sample). Recently, Zhang et al[32] reported that LIBS combined with random forest regression (RFR) was proposed for the quantitative analysis of multiple elements in fourteen steel samples.

In the present study, RF was employed to differentiate and classify iron ore samples based on LIBS spectra. 300 LIBS spectra of iron ore samples were randomly divided into training sets (200 spectra) and test sets (100 spectra). Training sets was used to construct the RF classification model; while test sets was selected as the predict samples to verify the performance of the constructed model. Two parameters (*ntree*-number of trees and m_{try}-random variables) of the RF algorithm were optimized using out-of-bag (OOB) estimation. Finally, support vector machines (SVM)[33] was compared with RF to classify the iron ore samples.

2. Experimental

2.1 Experimental set-up

A schematic diagram of the complete experimental system used in this study is shown in Fig. 1. The LIBS datasets presented in the study were acquired using the Nd: YAG laser (Litron, NANOSG120-20, UK) operating at 1064 nm and producing 21 mJ detected by energy meter (ES220C, Thorlabs) with a repetition rate of 5 Hz, pulse widths of 10 ns, which irradiated the samples with a 50 mm focal length lens. The focal area on the sample surface was evaluated as about 2.0×10^{-3} cm^2. A fused silica optical fiber was mounted on a micro-auto xyz-translation stage and used to collect the plasma emissions, before feeding them to a three-channel spectrometer (Avantes, AvaSpec, Netherlands) with broadband that covered a range of 200-940 nm (0.15 nm resolution). The integration time used on the spectrometer was 1.050ms. The detector was a compact charge-coupled device (CCD). A Stanford DG535 pulse generator (Stanford Research Systems, Inc., USA) was used to control the acquisition time settings. To eliminate continuous emission, all of the spectra were recorded with a delay of 1.8 μs after the laser pulses.

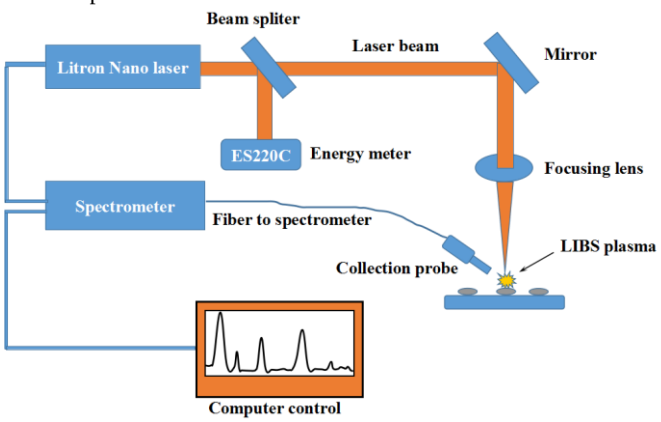

Fig. 1 LIBS experiment set-up for iron ore analysis

2.2 Materials

The ten grades of iron ore samples were listed on Table 1. A total of 10 typical iron ore samples from ten iron ore grades were provided by Pangang group Chengdu ore & steel CO., LTD (China) which provided in the form of ore powder. Each iron ore powder was 0.60g, and polyethylene (PE) was 0.40g. PE was put in the pelleting mould, and made it flat. Then, we put iron ores powder on the top of PE and compressed it. PE was used to pack iron ore powder in order to avoid the pellet scattering by pulse laser, which was compressed to form a pellet using a manual pellet presser with sufficient pressure (200Kgf/cm^2) lasting 4 minutes.

Table 1 Certified element composition of ten grades of iron ore samples (wt%)

Sample name	Composition wt%								
	TFe	SiO_2	TiO_2	MnO_2	V_2O_5	P	CaO	MgO	Al_2O_3
H140115-014	53.32	10.63	0.914	0.58	0.07	0.132	1.87	1.19	3.43
H140122-036	52.56	11.2	0.729	0.542	0.09	0.123	1.97	1.04	3.56
H140121-041	55.76	2.93	11.47	0.318	0.68	0.011	0.95	2.25	3.15
H140115-042	60.32	4.97	0.33	0.282	0.05	0.042	1.83	4.79	0.53
H140123-065-PT	55.64	9.37	0.69	0.485	0.08	0.104	2.25	1.04	3.05
H140122-073-PT	59.55	4.71	0.785	0.818	0.41	0.162	6.45	2.05	1.36
H140123-080-PT	63.67	1.91	0.035	0.153	0.03	0.043	0.08	0.68	0.53
H140119-092	55.74	8.91	0.583	0.516	0.08	0.111	1.78	0.89	2.94
H140123-097-PT	60.32	5.83	0.261	0.254	0.05	0.009	0.61	6.05	0.59
H140118-080	53.72	5.13	10.61	0.295	0.56	0.005	1.66	2.28	4.1

2.3 Data Acquisition

LIBS spectra were randomly collected by measurement at 150 different locations from one sample. A measured spectrum was collected as an accumulation of 20 laser shots per location for the purpose of improving the signal-to-noise ratio. To minimize the influence from sample heterogeneity and other fluctuations, every 5 measured spectra at 150 different locations were averaged into an analytical spectrum; finally we got 30 spectra for one grade sample. Then, the other nine classes of samples were analyzed in the same way. As a result, a total of 300 analytical spectra were acquired from 10 classes of iron ores (every class has 30 analytical spectra). Fig. 2 shows the original LIBS spectra for ten ore grades. The classification of iron ore samples grades via RF and SVM algorithm were both completed under MATLAB version 2007a (Mathworks).

2.4 Random Forest

RF is an advanced algorithm of machine learning. It is a classifier consisting of a collection of tree-type classifiers. Each tree-type classifier utilizes a unique training set constructed by bootstrap.[27] A resampling technique based on bootstrap method is used to continuously generate training and test sets; the training sets generate multiple classification tree form with RF. The final predictions results based on the combination are received by a simple majority voting of the single classification tree.

The classification was completed by constructing an ensemble of randomized classification and regression tress (CART) algorithm.[34] Assuming that $A=\{(x_i, y_i)\}_{i=1}^{N}$ as a set of features (here, peaks) is given as a train set, single data (x_i, y_i) is the input variable, and $y_i \in \phi_i = \{1,2,\cdots,J\}$ (J is the number of total class). There are two significant parameters for RF: *ntree* – the number of trees in the forest; m_{try} – the number of different descriptors tried at each split. The main steps of RF algorithm can be described as follows.[35]

Step 1 From the training data of *n* spectra, a bootstrap sample is drawn from the original spectrum (i.e., a randomly selected sample with replacement).

Step 2 For each bootstrap sample, the process of growing a tree is as follows: at each node, choose the best split among a randomly selected subset of m_{try} descriptors. The m_{try} is an essential and the

only tuning parameter in the algorithm. The tree is grown to the maximum size until no further splits are possible.

Step 3 Repeat the mentioned steps above until trees are grown large.

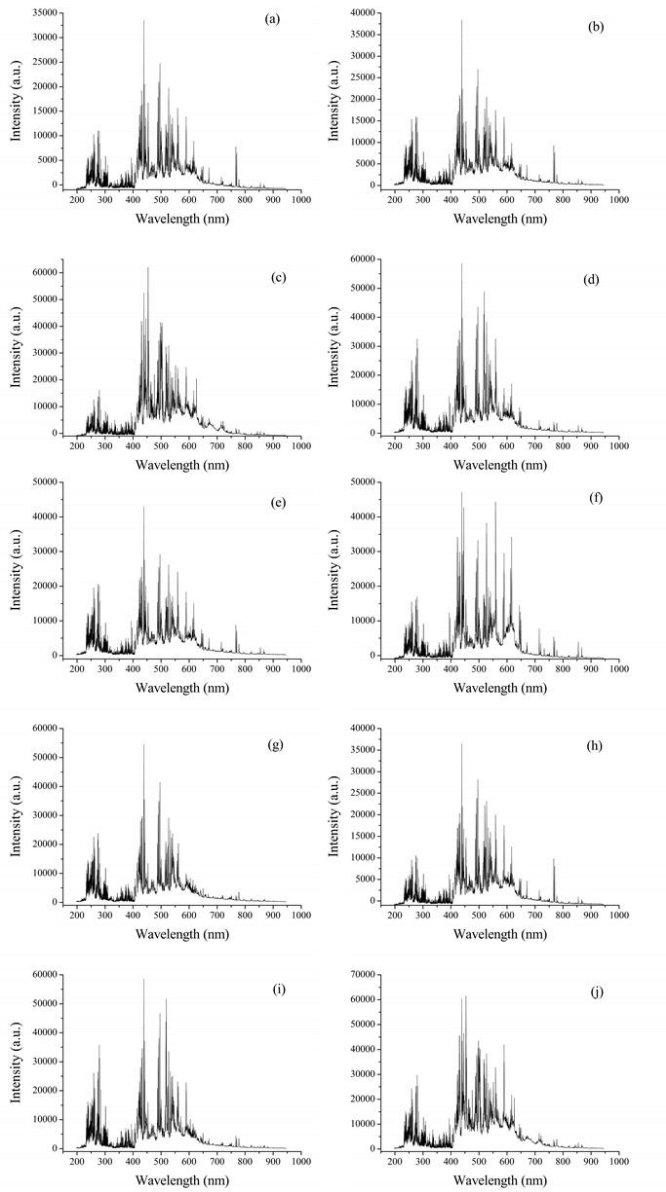

Fig. 2 Representative LIBS spectra of the iron ore samples, (a)H140115-014, (b)H140122-036, (c)H140121-041, (d)H140115-042, (e)H140123-065-PT, (f)H140122-073-PT, (g)H140123-080-PT, (h)H140119-092, (i)H140123-097-PT, (j)H140118-080.

3. Results and discussion

3.1 Parameter optimization

There are two important parameters in RF: the number of the trees in the forest called *ntree*, the number of the variables randomly selected as the candidates for splitting at each node called m_{try}. The out-of-bag error (OOB error) used to evaluate the effect of different settings of *ntree* and m_{try}, and OOB error is calculated by an estimate of the error rate (ER) for classification using Eq. (1) as follows:[36]

$$ER^{OOB} = n^{-1}\sum_{i=1}^{n} I(\hat{Y}^{OOB}(X_i) \neq Y_i) \qquad (1)$$

since each training feature X_i is in an OOB sample, we can calculate an ensemble prediction $\hat{Y}^{OOB}(X_i)$ by aggregating only its OOB predictions, where $I(\cdot)$ is the indicator function, which means that a function defined $\hat{Y}^{OOB}(X_i)$ does not belong to Y_i. Y_i is the observed output and n represents the total number of out of bag samples.

3.1.1 Select the number of trees

In the present work, selecting the number of trees is the first step to construct the RF model from the tested with 500 trees. The plotted OOB classification error vs. the number of grown trees shows in Fig. 3. As it could be seen in Fig. 3, the OOB error could not decrease after the number of trees that means the model does not over-fit when the error reaches 72, no matter how many trees are grown. Though a lower number of trees (closer to 50) seems to produce a lower mean square error, 72 could achieve a stable and relative low MSE. The optimal number of trees was determined to be the one that reached a relatively stable trend at the lowest OOB error. 72 was chosen as the optimal parameter for models, when the OOB error was at the lowest as the red line pointed.

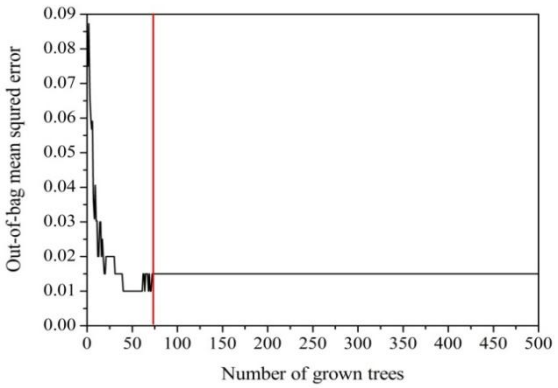

Fig. 3 Selection of tree number in RF

3.1.2 Determine the number of the variables randomly selected

m variables (m_{try}) are selected at random from all M variables (wavelengths) ($m_{try} \leqslant M$) and the best split of all m_{try} is used at each node. Each tree is grown to the largest scale (until no further splitting is possible) and no pruning of the trees occurs. It was assumed that there are M attributes in the training sample, and m_{try} attributes were extracted randomly as candidate attribute between each of the internal nodes in the decision tree ($m_{try} \leqslant M$). The effects of different *ntree* and m_{try} for the classification model were investigated by OOB error estimate. (show as Fig. 4).

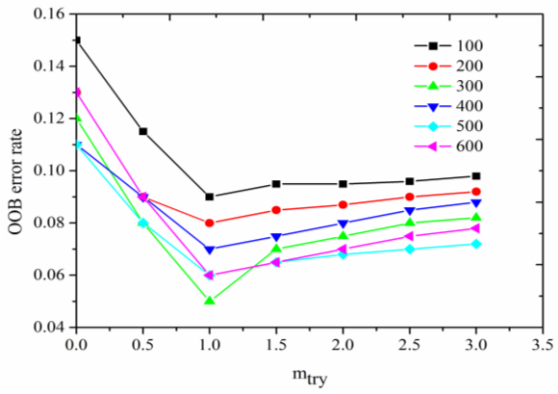

Fig. 4 Relationship of OOB error rate with *ntree* and m_{try}.The value of the abscissa is the coefficient of $\sqrt{M}$, where *n* is the number of peaks. If the coefficient is 0, the number of peaks tried at each split is 1.[30] *ntree* represents the number of the trees in the random forest.

Fig. 4 shows the relationship of the OOB error rate between the two parameters. *ntree* were used to set 100, 200, 300, 400, 500, and 600, respectively. As it can be seen, Fig. 4 shows that the OOB error of RF model is relatively high when *ntree* is less than 300; the OOB error of RF model reaches minimum when *ntree* was 300. When *ntree* was large enough, the OOB error tended to be limited by an upper bound. m_{try} were used to test $0.5\sqrt{M}$, $\sqrt{M}$, $1.5\sqrt{M}$, $2\sqrt{M}$, $2.5\sqrt{M}$ and $3\sqrt{M}$, respectively. It was the best choice based on the OOB error rate when $m_{try} = \sqrt{M}$.

3.2 Comparison classification results with SVM

As for RF method, we compared all broadband spectra with the characteristics emission spectral lines of Si and Ti as input data. Si emission lines include Si I (221.058nm, 251.621nm 252.921nm and 288.136nm) and Si II (385.653nm); Ti emission lines include Ti I (363.608nm, 365.383nm) and Ti II (307.843nm, 308.798nm and 323.503nm), which are the values for the peaks on the spectrometer. 300 LIBS spectra of iron ore samples were randomly divided into training sets (200 spectra) and test sets (100 spectra). Training sets was used to construct the RF classification model; while test sets was selected as the predict samples to verify the performance of the constructed model. Training set accuracy of all spectral data as input was 97.5% and the prediction accuracy of iron ore samples was 100%. It has no problems with overfitting due to the use of the Strong Law of Large Numbers[27]. In L. Breiman's paper, author proved RF converges. However, training set accuracy of characteristics emission spectra of Si and Ti as input was 95.0% and the prediction accuracy of iron ore samples was 90%. Therefore, all broadband spectra could provide more differences of features of spectra rather than several characteristic spectral emission lines only.

SVM are supervised learning models for classification and regression analysis in machine learning. SVM are based on empirical risk minimization; the learning discipline of SVM minimizes structural risk. The high-dimensional dataset can be efficiently dealt with SVM, and SVM can be flexible to decide the boundary in the high-dimensional feature space, since it has strong ability of global convergence. In the present study, the iron ore samples data obtained from LIBS contain a large number of variables. For the characteristic of LIBS data, SVM were utilized to identify and classify the 10 grades of iron ore samples to construct the SVM model. Although traditional SVM usually solve binary classification problem, it also can rebuild multi-class model.[37] In this work, 200 out of 300 datasets were randomly selected as the training set, and the rest of 100 datasets were used as test set.

In SVM, it is thought whether to create a hyperplane that allows linear separation in the higher dimension. It is solved by a transformation function $\Phi(x)$ that converts the data from an input space to feature space. A kernel function is used to perform. Two advantages of a kernel include reducing the computation load and retaining the effect of higher-dimensional transformation. The kernel function $K(x_i, y_i) = \Phi(x_i) \cdot \Phi(y_i)$ is defined, where Φ is a function to project the data into feature spaces. The more popular kernel function is the radial basic function (RBF) in Eq. (2) as follows:[29]

$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}) \qquad (2)$$

where parameter σ of the kernel defines implicitly the non-linear mapping from input space to feature space. The RBF is as kernel functions in SVM, which possessed 10-fold cross-validation accuracy of 98.5%, training set accuracy of 98.5% and test set accuracy of 96.0% used 68 support vectors. As can be seen in Fig. 5, the SVM for classification (SVC) parameter selection result using grid search method calculated the best cost (*c*) of 8, best gamma (*g*) of 0.088388, where *c* is penalty factor, showing the penalty degree to samples of excessive error; g is $1/2\sigma^2$ in RBF function (default the reciprocal of features numbers).

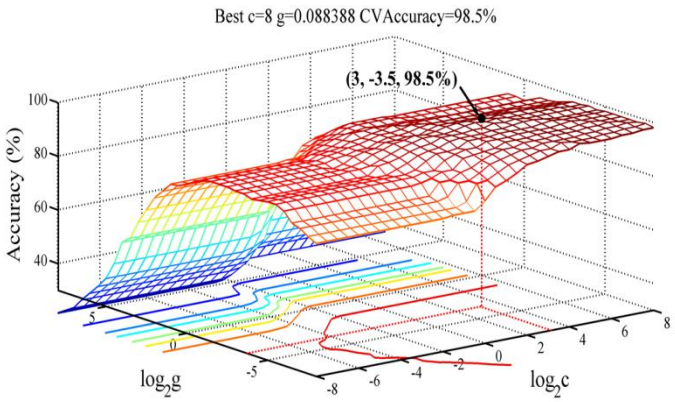

Fig. 5 The SVC parameter selection result

Based on the differences between each iron ore grade, all of the ore samples can be identified and classified. Table 2 shows the prediction accuracy of iron ore samples calculated by SVC and RF.

Table 2 The prediction accuracy of iron ore sample calculated by SVM and RF

Sample Type	SVM		RF	
	Average Rate of Correct Classification	Class (No.) of Misclassified Spectra	Average Rate of Correct Classification	Class (No.) of Misclassified Spectra
H140115-014	0.90(9/10)	H140119-092 (1)	1.00(10/10)	——
H140122-036	1.00(10/10)	——	1.00(10/10)	——
H140121-041	1.00(10/10)	——	1.00(10/10)	——
H140115-042	1.00(10/10)	——	1.00(10/10)	——
H140123-065-PT	1.00(10/10)	——	1.00(10/10)	——
H140122-073-PT	1.00(10/10)	——	1.00(10/10)	——
H140123-080-PT	1.00(10/10)	——	1.00(10/10)	——
H140119-092	0.70(7/10)	H140115-042 (3)	1.00(10/10)	——
H140123-097-PT	1.00(10/10)	——	1.00(10/10)	——
H140118-080	1.00(10/10)	——	1.00(10/10)	——
Average	0.96		1.00	

Here, we observed unstable performances for two classes. One spectrum of Grade H140115-014 was misclassified as H140119-092,

and three spectra of Grade H140119-092 were misclassified as H140115-042, while others were almost perfectly classified. Due to features extracted from SVs may similar with others, it would be happened to misclassification.

4. Conclusion

In this paper, LIBS based on chemometrics was performed in order to identify and classify iron ore samples. This represented a significant challenge since the difference in the iron ore samples with a high dimensional data produced by LIBS. SVM and RF were comparatively evaluated as methods for processing the LIBS for classification and prediction of the different iron ore samples. In case of SVM, RBF produced the best accuracy, which possessed 10-fold cross-validation accuracy of 98.5%, training set accuracy of 98.5% and test set accuracy of 96.0% used 68 support vectors. RF exhibits the better predictive power than SVM in classifying the iron ore samples in the test set. The prediction results of both training and tested samples demonstrate that the proposed RF model is an effective and efficient approach for classification of iron ore samples grade. 72 trees and 300 random variables were optimized and selected as the best parameters for the classification of iron ore grade. The average predicted accuracy rate of the RF method is 100%, which shows more perfect performance than that of SVM method. Overall, LIBS–RF is a useful technique for the identification and discrimination of iron ore samples, and is promising for automatic real-time, fast, reliable, and robust measurements.

Acknowledgement

We gratefully acknowledge the support of the National Major Scientific Instruments and Equipment Development Projects of China (2011YQ030113), National Natural Science Foundation of China (NO.21175106 and NO.21375105), Research Fund for the Doctoral Program of Higher Education of China (NO.20126101110019), the Natural Science Foundation of Shaanxi Province of China (NO.2014JM2045).

Notes and References

[a]*Institute of Analytical Science, College of Chemistry &Materials Science, Northwest University, Xi'an, 710069, China Email: nwufxkx2012@126.com*
[b]*College of Science, Chang'an University, Xi'an, 710064, China*
[c]*Research Center of Analytical Instrumentation, College of Chemistry, Sichuan University, Chengdu 610064, China*
[d]*College of Life Sciences, Sichuan University, Chengdu, 610064, China*

1 F. J. Fortes, J. Moros, P. Lucena, L. M. Cabalín and J. J. Laserna, *Anal. Chem.*, 2012, **85**, 640-669.
2 D. A. Cremers and R. C. Chinni, *Appl. Spectrosc. Rev.*, 2009, **44**, 457-506.
3 F. J. Fortes and J. J. Laserna, *Spectrochim. Acta Part B*, 2010, **65**, 975-990.
4 D. W. Hahn and N. Omenetto, *Appl. Spectrosc.*, 2012, **66**, 347-419.
5 M. Brai, G. Gennaro, T. Schillaci and L. Tranchina, *Spectrochim. Acta Part B*, 2009, **64**, 1119-1127.
6 J. Gurell, A. Bengtson, M. Falkenstrom and B.A.M. Hansson, *Spectrochim. Acta Part B*, 2012, **7**, 446-450.
7 L. M. Cabalin, A. Gonzalez, J. Ruiz and J. J. Laserna, *Spectrochim. Acta, Part B*, 2010, **65**, 680-687.
8 L. C. Nunes, G. A. da Silva, L. C. Trevizan, D. Santos Júnior, R. J. Poppi and F. J. Krug, *Spectrochim. Acta Part B*, 2009, **64**, 565-572.
9 Q. Lin, Z. Wei, M. Xu, S. Wang, G. Niu, K. Liu, Y. Duan and J. Yang, *RSC Adv.*, 2014, **4**, 14392-14399.
10 J. L. Gottfried, *Anal. Bioanal. Chem.*, 2011, **400**, 3289-3301.
11 J. P. Singh and S. N. Thakur, *Laser-Induced Breakdown Spectroscopy*, Elsevier, Amsterdam, 2007.
12 D. W. Hahn and N. Omenetto, *Appl. Spectrosc.*, 2010, **64**, 335-366.
13 D. W. Hahn and N. Omenetto, *Appl. Spectrosc.*, 2012, **66**, 347-419.
14 C. Aragon and J. A. Aguilera, *Spectrochim. Acta Part B*, 2008, **63**, 893-916.
15 I. Gornushkin and U. Panne, *Spectrochim. Acta Part B*, 2010, **65**, 345-359.
16 N. Konjević, M. Ivković and S. Jovićević, *Spectrochim. Acta Part B*, 2010, **65**, 593-602.
17 E. Tognoni, G. Cristoforetti, S. Legnaioli and V. Palleschi, *Spectrochim. Acta Part B*, 2010, **65**, 1-14.
18 F. Nellros and M. J. Thurley, *Miner. Eng.*, 2011, **24**, 1525-1531.
19 R. Z. Abd Rashid, H. Mohd Salleh, M. H. Ani, N. A. Yunus, T. Akiyama and H. Purwanto, *Renew. Energ.*, 2014, **63**, 617–623.
20 J. J. Remus, R. S. Harmon, R. R. Hark, G. Haverstock, D. Baron, I. K. Potter, S. K. Bristol and L. J. East, *Appl. Opt.* 2012, **51**, B65–B73.
21 S. M. Clegg, E. Sklute, M. D. Dyar, J. E. Barefield and R. C. Wiens, *Spectrochim. Acta Part B*, 2009, **64**, 79–88.
22 J. B. Sirven, B. Sallé, P. Mauchien, J. L. Lacour, S. Maurice and G. Manhès, *J. Anal. At. Spectrom.*, 2007, **22**, 1471–1480.
23 M. Hoehse, A. Paul, I. Gornushkin and U. Panne, *Anal. Bioanal. Chem.*, 2012, **402**, 1443-1450.
24 N. L. Lanza, R. C. Wiens, S. M. Clegg, A. M. Ollila, S. D. Humphries, H. E. Newsom and J. E. Barefield, *Appl. Opt.*, 2010, **49**, C211–C217.
25 A. Ramil, A. J. López, and A. Yáñez, *Appl. Phys. A-mater.*, 2008, **92**, 197–202.
26 O. Forni, S. Maurice, O. Gasnault, R. C. Wiens, A. Cousin, S. M. Clegg, J. B. Sirven and J. Lasue, *Spectrochim. Acta Part B*, 2013, **86**, 31–41.
27 L. Breiman, Random forests, *Mach. Learn.*, 2001, **45**, 5-32.
28 L. Dong, X. Li, M. Xu and Q. Li, *Pro. Eng.*, 2011, **26**, 1772-1781.
29 L. Zheng, D. G. Watson, B. F. Johnston, R. L. Clark, R. Edrada-Ebel and W. Elseheri, *Anal. Chim. Acta.*, 2009, **642**, 257-265.
30 X. Lin, L. Sun, Y. Li, Z. Guo, Y. Li, K. Zhong, Q. Wang, X. Lu, Y. Yang and G. Xu, *Talanta*, 2010, **82**, 1571-1575.
31 J. Remus and K. S. Dunsin, *Appl. Optics*, 2012, **51**, B49-B56.
32 T. Zhang, L. Liang, K. Wang, H. Tang, X. Yang, Y. Duan and H. Li, *J. Anal. At. Spectrom.*, 2014, **29**, 2323-2329.
33 C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273–297.
34 L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and regression tree*, CRC Press, 1984.
35 A. Liaw and M. Wiener, *R News*, 2002, **2**, 18–22.
36 S. Adusumilli, D. Bhatt, H. Wang, P. Bhattacharya and V. Devabhaktuni, *Expert Syst. Appl.*, 2013, **40**, 4653-4659.
37 L. Liang, T. Zhang, K. Wang, H. Tang, X. Yang, X. Zhu, Y. Duan and H. Li, *Appl. Optics*, 2014, **53**, 544-552.