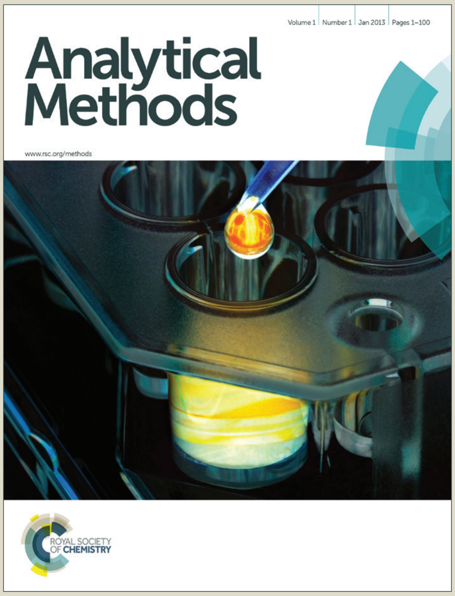
Analytical Methods

Accepted Manuscript

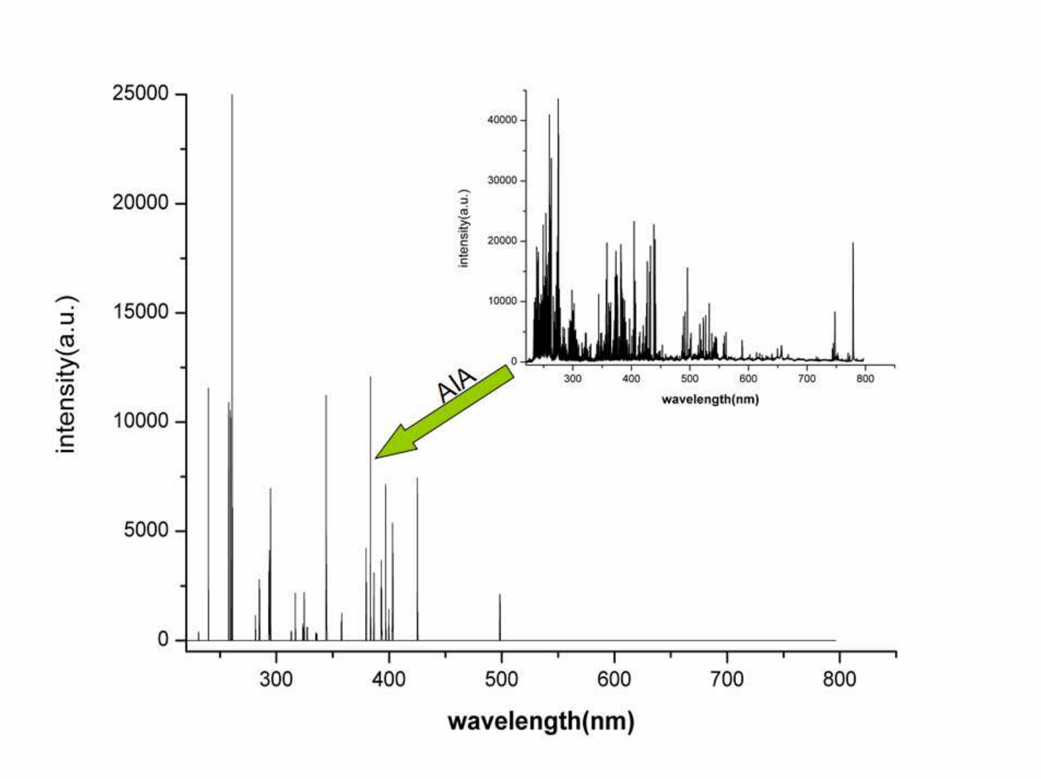

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analytical Methods
www.rsc.org/methods
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

## Graphical Abstract

A novel method of improving classification precision, Accuracy Influence Analysis (AIA), is proposed to combined with Support Vector Machines (SVM) for selecting informative variables of Laser-induced Breakdown Spectroscopy (LIBS) spectra. Based on Model Population Analysis (MPA), AIA could reveal informative variables which have statistically significant influence on the prediction accuracy of SVM sub-models. Support Vector Machine is then employed to build more robust model and classify nine types of round steel based on the selected spectral variables. In this way, the classification performance of SVM is further improved and the computation time is reduced greatly. It is demonstrated that AIA is a good alternative for variable selection of high-dimensional LIBS dataset.



# A method of improving classification precision based on Model Population Analysis of steel material for Laser-induced Breakdown Spectroscopy

Lin Xu[a], Long Liang[a], Tianlong Zhang[a], Hongsheng Tang[a], Kang Wang[b], Hua Li*[a]

a. *Institute of Analytical Science, College of Chemistry & Materials Science, Northwest University, Xi'an, 710069, P. R. China*

b. *College of Science, Chang'an University, Xi'an, 710064, P. R. China*

Author to whom correspondence should be addressed:

Dr. Hua Li

Institute of Analytical Science, Northwest University

Tel: 86-29-88302635

Fax: 86-29-88303527

*E-mail:* nwufxkx2012@.126.com

## Abstract

A novel method of improving classification precision, Accuracy Influence Analysis (AIA), is proposed to combined with Support Vector Machines (SVM) for selecting informative variables of Laser-induced Breakdown Spectroscopy (LIBS) spectra. Based on Model Population Analysis (MPA), AIA could reveal informative variables which have statistically significant influence on the prediction accuracy of SVM sub-models. Support Vector Machine is then employed to build more robust model and classify nine types of round steel based on the selected spectral variables. In this way, the classification performance of SVM is further improved and the computation time is reduced greatly. It is demonstrated that AIA is a good alternative for variable selection of high-dimensional LIBS dataset.

**Keywords:** Accuracy Influence Analysis; LIBS; SVM; informative variables

## 1. Introduction

Laser-induced Breakdown Spectroscopy (LIBS) is an atomic emission spectroscopy used for qualitative and quantitative analyses of solid, liquid, and gaseous materials.[1,2] There are many advantages of the LIBS-technique, such as no sample preparation, minimum destruction to sample, and allowing for in situ and real-time analysis.[3-10] LIBS has been successfully used to analyze biological tissues,[11-13] archeological samples,[14-16] geological materials,[17,18] polymers,[19-21] pharmaceuticals[22,23] and metallic materials.[24-27] In the metallurgy industry, LIBS has been extensively applied for industrial process control, characterization of hot and molten metal, on-line measurement of coating thickness and composition, analysis of slag, and mostly focused on quantitative determination of elements. R. Fantoni et al. had given a review of methodologies for laboratory LIBS semi-quantitative and quantitative analysis. Moreover, Z. Wang et al. had established a new approach of PLS method combined with a physical principle based on dominant factor which had been applied to quantitative measurements for LIBS.[28,29] In addition, based on the intensities and ratios of elemental emission lines of the whole spectrum, LIBS could be applicable for qualitative analysis, such as identification and discrimination of the different types of steel and recycling of metallic scrap. We investigate the application of LIBS and chemometrics techniques for classification of different types of round steels based on statistical learning theory ( SLT ).

LIBS has been used for identification and classification in combine with different chemometrics methods such as Principal Component Analysis (PCA), Soft Independent Modeling of Class Analogy (SIMCA), Artificial Neural Networks (ANN) and Support Vector Machines (SVM).[30-32] Traditional approaches are often univariate, thus the results of classification are not precise enough. To compare with them, multivariate statistical methods were used to select informative variables from the whole spectrum and evaluate interactions among variables. However, it is worth noting that the majority of the variables in the whole spectrum are unrelated to the elemental composition, and thus do not necessarily contribute significantly to the qualitative and quantitative analyses. Furthermore, the redundant variables are more likely to incorporate spurious correlations and noises, and hamper the accuracy and precision of analysis result seriously. In addition, for the high dimensionality of the LIBS dataset, the computing time will greatly extend if the whole spectrum is used as the input variable .

For all of the reasons above, it is necessary to introduce appropriate feature extraction procedures to eliminate the influence of excess variables or spectrum regions during LIBS analysis. Traditional methods manually selected emission intensity peaks according to characteristic elements of the samples as analyzed variables. However, it requires a large amount of experiments datasets, and as a result it is time consuming; moreover, additional chemical relationships among variables are not evaluated. What’s more, it will become extremely difficult to acquire the detailed constituent elements of the samples.

Recently, H. D. Li et al. had developed a new variable selection method, Margin Influence Analysis (MIA), to work with Support Vector Machines specifically.[33,34] MIA, based on Model Population Analysis (MPA), could select variables which have statistically significant influence on the margin by using Mann-Whitney U test. Using two publicly available cancerous microarray data sets, it is demonstrated that MIA could typically select a small number of margin-influencing genes and further achieves outstanding performance. In the present work, Accuracy Influence Analysis(AIA), an improved variable selection method based on MPA, is presented to identify informative LIBS spectral variables by statistically analyzing the prediction accuracy. Support Vector Machine is then employed to obtain a classification model based on the selected spectral variables. It is different from traditional methods that informative variables extracted by AIA are actual spectral bands, they are helpful for spectral analysis. The results prove AIA is a variable, competitive and robust selection method, which obviously improve the prediction ability of SVM model based on high-dimensional LIBS spectrum.

## 2. Experimental Methods

### 2.1. Materials

Nine common types of round steel from steel market were analyzed: 20#(Φ20×900mm), 20Cr(Φ20×900mm), 20CrMnTi(Φ30×900mm), 20CrMo(Φ20×900mm), 20CrNiMo(Φ20×900mm), 35#(Φ20×900mm), 35CrMo(Φ20×900mm), 40Cr(Φ20×900mm), 42CrMo(Φ25×900mm) (XINING SPECIAL STEEL CO., LTD.). A 6 mm-height steel column was cut separately from three different regions of each type of sample. The two cross sections of every steel column should be measured because of heterogeneous composition.

### 2.2. Experimental Setup

The LIBS data presented in this study were acquired by the Nd: YAG laser (LOTIS, TII2131, Belarus) operating at 1064 nm and producing 61 mJ (4.3 GW/cm$^2$) energy with a repetition rate of 10 Hz, which was focused onto the surface of samples with a 5 cm focal length lens. The LIBS spectra were recorded using an echelle spectrometer (ARYELLE-UV-VIS, LTB150, German), which provided a Constant Spectral Resolution (CSR) of 6000 over a wavelength range 220-800 nm displayable in a single spectrum. An Electron-Multiplying CCD camera (UV enhanced, 1004 x 1002 Pixels, USA), coupled with the spectrometer was used for detection of the dispersed light. The overall linear dispersion of the spectrometer camera system ranges from 37 pm (at 220 nm) to 133pm/pixel (at 800 nm). A mechanical chopper was used in front of the entrance slit to prevent the CCD from detecting the early plasma continuum. All spectra were acquired after a 1.5μs delay from the laser pulse.

### 2.3. Data Acquisition

For each surface of every sample from one class, 50 locations are randomly selected for measuring. A measured spectrum was collected as an accumulation of 20 laser shots in per location in purpose of improving the signal-to-noise ratio. In order to reduce the influence from sample heterogeneity and other fluctuations, every 5 measured spectra at different locations were averaged into an analytical spectrum. Then the other eight classes of samples were taken in the same way. As a result, a total of 540 analytical spectra were acquired from 9 classes of round steel (6 surfaces per classes, 10 average spectra per surface). Additionally, in order to prevent any spurious effects on our classification models, chemometrics analysis is performed without the application of any additional preprocessing methods.

The broadband LIBS spectrum consists of 29888 intensity channels ranging from 220.433 to 796.352 nm. If each channel was used as an input variable, it is needlessly and the computation cost can be very large. In order to facilitate the calculation, the entire spectrum is divided into 977 segments (30 wavelength channels per segment, insufficient padded with 0). Therefore, in this work, one informative variable contains 30 wavelength points.

The MATLAB (version 2007a, MathWorks) software equipped with the LIBSVM Toolbox was used for all computations reported in this study.

## 3. Statistical Analysis

### 3.1. Support Vector Machine

Support Vector Machine (SVM), based on statistical learning theory, is a promising kernel-based method for data mining and pattern recognition. The main idea for SVM classification is to construct a separating hyperplane that splits the data into two separate regions and maximizes the margin between the closest points in each separate region. The data points closer to the hyperplane are called the support vectors. Different from the traditional machinelearning methods which are based on empirical risk minimization, the learning discipline of SVM is to minimize the structural risk, and thus the better generalization ability is guaranteed. It also has special advantage in solving small sample, non-linear and high dimension mode recognition. An overview of the methods can be found in monographs of Shawe-Taylor and Cristianini or Hastieet al.[35-37]

SVM is originally designed for binary classification. Thus the multi-class problem can be transformed to the combination of several two-class problems. The most commonly used implementations for SVM multi-class classification are "one-against-all" and "one-against-one".[38-40] In this work, a modified combination model strategy that combines "one-against-all" and "one-against-one" is applied to construct multi-class SVM model to identify the steel materials. The combination model includes two stages: fuzzy classification and fine classification. For the fuzzy classification, suppose that a total of $k$ classes of round steels should be discriminated, the class memberships are stored in a matrix by binary coding, where the columns and rows are equal to the number of classes and spectrum, respectively. Each column in this matrix consists of ones ("1") for the corresponding class; otherwise it consists of zeroes ("0"). Then membership matrix and all training spectrum are used to constructs $k$ SVM binary classifiers. The $i$ th SVM classifier is trained with all of the spectrum data in the $i$ th class with positive labels, and all other data with negative labels. The test spectrum is put into all classifiers for discriminating and the prediction result is also a $1 \times k$ vector by binary coding. If its prediction vector has only one column consists of "1", the test spectrum is considered belong to the class corresponding to the "1" and whole prediction processes end early. Otherwise, all the classes corresponding to the "1" are considered as candidate classes. Specially, when there is no one "1" in prediction vector, all $i$ classes are also considered as candidate classes. Then the test spectrum would be finely classified again by SVM again based on one-against-one model which fully considers the class-to-class variability. Suppose that there are m candidate classes, thus the membership matrix becomes a vector contains m kinds of class identifier. The training spectra in candidate classes and membership vector are used to construct $m(m-1)/2$ classifiers where each one is trained with data from two classes. Then the test spectrum is discriminated by all classifiers and the prediction result is decided by the following voting strategy (called "Max Wins").[41] If the classifier ($i$ th class and $j$ th class) predicts the spectrum belong to the $i$ th class, then the vote for the $i$ th class is added by one. Otherwise, the $j$ th is increased by one. In the end, the spectrum we predict belongs to the class with the largest vote. If the number of the classes with the largest vote is two or more, the test spectrum is predicted again by "Max Wins" strategy involving the largest vote classes and this step loop runs until two consecutive identical results are obtained. If the final result is still not unique, the test spectrum is considered as "unclassified". The introduction of "unclassified" is based on the consideration of component discrepancies existing within different regions of same type steel in the application for steel materials identification. The spectrum from some measurement locations may be unable to obtain enough composition information which will lead to be ambiguously classified. These confused data should be identified to avoid misclassification.

### 3.2. Margin Influence Analysis for Variable Selection Based on MPA

Like the Margin Influence Analysis（MIA）, the AIA is also applied by implementing the Model Population Analysis (MPA). Briefly, a standard MPA is described by the following steps:

(1) Obtain $N$ sub-data sets by Monte Carlo sampling (MCS), where $N$ is the number of MCS

(2) Construct a sub-model for each sub-dataset.

(3) Statistically analyze the interesting outputs of all the sub-models.

However, the main interesting output of AIA is the prediction accuracy of SVM sub-models instead of the margins.[31, 32, 42-44]

### 3.2.1. Monte Carlo Sampling

In each Monte Carlo Sampling, $Q$ variables are randomly picked out from whole spectral variables without any replacement, thus a sub-dataset can be obtained. Repeat this procedure $N$ times, and we can get $N$ sub-datasets.

### 3.2.2. Constructing SVM Sub-model

The each sub-dataset is randomly split into training subset and test subset. The training subset is used to build the SVM sub-model based on the aforementioned combination model and then the test subset is predicted by the SVM sub-model. Finally, we get $N$ sub-models and correspond to prediction accuracy.

### 3.2.3. Statistical Analysis by the Mann-Whitney U test

The computing procedure is illustrated by the case of the $i$ th variable. For the $N$ times of samplings, the prediction accuracy of SVM sub-modes which include the $i$ th variable are subsumed into Group A and others are subsumed into Group B. The mean values of Group A and Group B are denoted by $\mathrm{MEAN}_{i,A}$ and $\mathrm{MEAN}_{i,B,}$ respectively. And $\mathrm{DMEAN}_i = \mathrm{MEAN}_{i,A} - \mathrm{MEAN}_{i,B}$. If $\mathrm{DMEAN}_i >$ 0, the $i$ th variable is treated as a candidate of informative variables that increases the predictive performance of the sub-model. Otherwise, the $i$ th variable is treated as the uninformative variable that reduce the predictive performance and removes from the variable space.

Next, the Mann-Whitney U test[45] is applied to check whether each variable in candidates of informative variables could significantly increase the predictive performance. Only the regions with

$p$ value are smaller than the predefined threshold (0.01 in this work), can they be defined as identified informative variables.

## 4. Result and Discussion

In order to avoid over fitting, the spectra dataset should be split into training and test sets. Informative variables were selected only from the training sets, and the test sets were used to test the effectiveness of informative variables. Taking into account a realistic scenario, steel materials have the characteristic of local uniform composition and overall uneven composition, if entire spectra are randomly split into training sets and test sets, the local similar spectra will be overfitted to model and lead to less unrealistically high predictive accuracy. Therefore, the spectra from one surface in total were conducted as analysis object. Thus, the spectra of 4 randomly selected surfaces from each class are training sets (360 spectra, 40 spectra per class) and the remaining spectra compose the test sets (180 spectra, 20 spectra per class). The penalizing factor $C$ and the polynomial parameter $d$ were optimized by performing a grid search over the range of $10^{-5}$ to $10^{5}$ ($C$) and 1 to 10 ($d$), respectively. In the final, $C = 1$, $d = 1$ were selected as the optimal value. Fig. 1 presents the representative spectra of nine round steel classes measured on the LIBS system. It was observed that there are lots of uninformative regions and background spectrum in the range of approximately 220–796 nm, discriminating the nine types of round steel on the basis of LIBS measurements is not straightforward. Owing to the complexity of the spectra, the use of the AIA is required. When the program of AIA was run, informative variables could be obtained. Fig. 2 presents that the spectrum only contains informative variables.

**“Here Fig. 1”**

**“Here Fig. 2”**

As discussed in Section 2, before implementing Monte Carlo Sampling(MCS), three parameters are determined: $N$ -the number of MCS; $S$ -the number of spectra selected in each sampling; $Q$ -the number of variables selected in each sampling. In order to get as many variables combinations as possible, parameter $N$ should be large enough. But considering the computational cost, we choose $N = 5000$ in the present work. $S$ should be 1/2-2/3 as large as the total number of spectra.

Then parameter $Q$ is set to 5, 10, 30, 50, 100, 150, 200, respectively, to investigate the influence on the SVM results of the number of variable selection for the model.

**"Here Fig. 3A"**

**"Here Fig. 3B"**

In Figure 3, two types of variables lead to different accuracy distribution. Plot A: informative variable (the 303th region, $Q = 30$ , Plot A). Plot B: uninformative variable (the 544th region, $Q = 30$ , Plot B). The red bars represent the accuracy distribution of models including a given variable, while the blue bars represent the accuracy distribution of models without the given variable.

A typical informative variable (the 303th region, $Q = 30$ , Plot A) and an uninformative variable (the 544th region, $Q = 30$ , Plot B) are illustrated in Figure 2. In Plot A, it is clear that the accuracy distribution of models including the 303th region is significantly right-shifted ( $p$ =$4.28\times10^{-21}$). This means that this region can improve the classification performance if it is included. By contrast, for the 544th region, the accuracy distribution is unchanged while the 544th region is included. Thus, this region can be considered as uninformative variable and should be removed from the model.

We note that the results of AIA could not be exactly reproduced due to the embedded Monte Carlo strategy. Therefore, we investigated the variation of the classification accuracy of the informative spectra model by running AIA on the same data 20 times. The average prediction accuracy and the standard deviations are shown in Figure 4. Table 1 lists more details and computation time of the informative spectra models by AIA.

**"Here Fig. 4"**

Figure 4 shows the average prediction accuracy and the standard deviations of running the AIA program 20 times. The highest average prediction accuracy and the least standard deviation were obtained when $Q = 50$ .

**"Here Table 1"**

As shown in Figure 4, the average prediction accuracy first increases until it reaches the maximum and then, as more variables are selected, the average prediction accuracy decreases gradually. By contrast, the standard deviation shows an opposite trend of variation against the average prediction accuracy. We note that the value of $Q$ greatly affect the sensitivity and robustness of AIA. Insufficient sampled variables in sampling cannot provide the requisite information and results in poor discrimination between classes, while excessive sampled variables affect the selection of informative variable and even worsen the classification performance. It is clear that the highest average prediction accuracy (approached 100%) can be obtained while 50 variables are included, and the standard deviation affords no significant change. Implementing Monte Carlo sampling, contribution of target variable to classification model was comprehensively considered by combining target variable and other different variables. By eliminating the redundant (uninformative) regions of the spectrum, the separation between classes is enhanced and the classification model, based solely on the informative features, becomes more sensitive for distinguishing classes with similar constituent elements.

On the whole, it might be concluded that AIA is a good alternative for informative variable selection and the AIA-based SVM classifier is a promising predicted tool for steel materials discrimination.

## 5. Conclusion

In summary, AIA, a new statistical variable selection algorithms, has been applied to obtain informative variable of the LIBS spectrum. The AIA works similarly to MIA, but it can identify informative variables by statistically analyzing the distribution of the prediction accuracy instead of the margins, which has shown better generalization ability. Implementing Monte Carlo sampling, contribution of target variable to classification model was comprehensively considered by combining target variable and other different variables. Support vector machine is then employed to classify nine types of round steel based on the selected spectral variables. It is demonstrated that AIA is a competitive and robust variable selection method that obviously improve the prediction ability of SVM model based on high-dimensional LIBS spectrum. Furthermore, compared with the whole spectra model, the informative spectra model can shorten the computation time remarkably, which makes the LIBS-SVM technique suitable for in-field fast and real-time on-line analysis.

## Acknowledgements

This work was financially supported by the National Special Fund for the Development of Major Scientific Instruments and Equipment (No. 2011YQ030113) of China, National Natural Science

Foundation of China (No. 21375105).

## Reference

1 D. A. Cremers and L. J. Radziemski, *Handbook of Laser-Induced Breakdown Spectroscopy,* Wiley, 2006.

2 A. W. Miziolek, V. Palleschi, and I. Schechter, *Laser Induced Breakdown Spectroscopy,* Cambridge University Press, 2006.

3 D. W. Hahn and N. Omenetto, *Applied Spectroscopy*, 2010, **64**, 335A–366A .

4 D. W. Hahn and N. Omenetto, *Applied Spectroscopy*, 2012, **66**, 347–419 .

5 F. J. Fortes, J. Moros, P. Lucena, L. M. Cabalin and J. J. Laserna, *Analytical Chemistry*, 2013, **85**, 640–669 .

6 E. G. Snyder, C. A. Munson, J. L. Gottfried, F. C. De Lucia Jr, B. Gullett and A. Miziolek, *Applied Optics*, 2008, **47**, G80–G87.

*7* F. C. De Lucia Jr, R. S. Harmon, K. L. McNesby, R. J. Winkel Jr and A. W. Miziolek, *Applied Optics*, 2003, **42**, 6148–6152.

8 J. B. Sirven, B. Sallé, P. Mauchien, J. L. Lacour, S. Maurice and G. Manhès, *Journal of Analytical Atomic Spectrometry*, 2007, **22**, 1471–1480.

9 S. Palanco and J. J. Laserna, *Journal of Analytical Atomic Spectrometry*, 2000, **15**, 1321–1327.

10 R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Moench, L. Peter and V. Sturm, *Spectrochimica Acta Part B-Atomic Spectroscopy,* 2001, **56**, 637–649.

11 J. D. Hybl, G. A. Lithgow and S. G. Buckley, *Applied Spectroscopy*, 2003, **57**, 1207-1215.

12 A. C. Samuels, F. C. De Lucia Jr, K. L. McNesby and A. W. Miziolek, *Applied Optics*, 2003, **42**, 6205-6209.

13 P. B. Dixon and D. W. Hahn, *Analytical Chemistry*, 2005, **77**, 631-638.

14 A. Giakoumaki, K. Melessanaki and D. Anglos, *Analytical and Bioanalytical Chemistry*, 2007, **387**, 749-760.

15 A. Brysbaert, K. Melessanaki and D. Anglos, *Journal of Archaeological Science*, 2006, **33**, 1095-1104.

16 A. Ramil, A . J. López and A. Yáñez, *Applied Physics A*, 2008, **92**, 197-202.

17 S. Béatrice, A. C. David, M. Sylvestre, C. W. Roger and F. Pascal, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2005, **60**, 805-815.

18 R. D. Harris, D. A. Cremers, C. Khoo and K. Benelli, *36th Annual Lunar and Planetary Science Conference, in League City, Texas, abstract*, 2005, **1796**, 14-18.

19 R. J. Lasheras , C. Bello-Gálvez and J. Anzano, *Polymer testing*, 2010, **29**, 1057-1064.

20 J. Juraj, H. Johannes, D. P. Johannes and V. Pavel, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2009, **64**, 1128-1134.

21 R. Viskup, B. Praher, T. Linsmeyer, H. Scherndl, J.D. Pedarnig and J. Heitz, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2010, **65**, 935-942.

22 C. S-C. Yang, E. E. Brown, E. Kumi-Barimah, U. H. Hommerich, F. Jin, S. B Trivedi, A. C. Samuels and A. P. Snyder, *Applied Spectroscopy*, 2014, **68**, 226-231.

23 N. Lewen, *Journal of Pharmaceutical and Biomedical Analysis*, 2011, **55**, 653-661.

24 M. Gaft, L. Nagli, I. Fasaki, M. Kompitsas and G. Wilsch, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2009, **64**, 1098-1104.

25 D. Bulajic, G. Cristoforetti, M. Corsi, M. Hidalgo, S. Legnaioli, V. Palleschi, A. Salvetti, E. Tognoni, S. Green, D. Bates, A. Steiger, J. Fonseca, J. Martins, J. McKay, B. Tozer, D. Wells, R. Wells and M.A. Harith, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2002, **57**, 1181-1192.

26 R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Mönch, L. Peter and V. Sturm, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2001, **56**, 637-649.

27 V. Sturm, L. Peter and R. Noll, *Applied Spectroscopy*, 2000, **54**, 1275-1278.

28 R. Fantoni, L. Caneve, F. Colao, L. Fornarini, V. Lazic and V. Spizzichino, *Spectrochimica Acta Part B-Atomic Spectroscopy*, 2008, **63**, 1097-1108.

29 Z. Wang, J. Feng, L. Z. Li, W. D. Ni and Z. Li, *Journal of Analytical Atomic Spectrometry*, 2011, **26**, 2289-2299.

30 J-B Sirven, B. Bousquet, L. Canioni, L. Sarger, S. Tellier, M. Potin-Gautier and I. L. Hecho, *Analytical and Bioanalytical Chemistry*, 2006, **385**, 256-62.

31 S. Schröder, S. G. Pavlov, I. Rauschenbach, E. K. Jessberger and H. W. Hübers, *ICARUS*, 2013, **223**, 61-73.

32 C. W. Hsu and C. J. Lin, *Neural Networks, IEEE Transactions on*, 2002, **13**, 415-425.

33 H. D. Li, Y. Z. Liang, Q. S. Xu, D. S. Cao, B. B. Tan, B. C. Deng and C. C. Lin, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2011, **8**, 1633-1641.

34 H. D. Li, M. M. Zeng, B. B. Tan, Y. Z. Liang, Q. S Xu and D. S. Cao, *Metabolomics*, 2010, **6**, 353-361.

35 N. Cristianini and J. Shawe-Taylor, *Cambridge university press*, 2000.

36 J. Shawe-Taylor and N. Cristianini, *Cambridge university press*, 2004.

37 T. Hastie, S. Rosset, R. Tibshirani and J. Zhu, *Journal of Machine Learning Research,* 2004, **5**, 1391-1415.

38 C. W. Hsu and C. J. Lin, *IEEE TRANSACTIONS ON NEURAL NETWORKS,* 2002, **13**, 415-425.

39 M. N. Nguyen and J. C. Rajapakse, *Genome informatics. International Conference on Genome Informatics*, 2003, **14**, 218-227

40 C. Tan, T. Wu and X. Qin, *ASIAN JOURNAL OF CHEMISTRY,* 2013, **25**, 3668-3672

41 J. H. Friedman, *Technical Report, Department of Statistics, Stanford University*, 1996.

42 M. H. Xie, F. F. Deng, X. Y. Zhang, Y. L. Tian, P. Z. Li and H. L. Zhai, *Chemometrics and Intelligent Laboratory Systems*, 2014,**132**, 124-132.

43 J. H. Huang, J. Yan, Q. H. Wu, M. D. Ferro, L. Z. Yi, H. M. Lu, Q. S. Xu and Y. Z. Liang, *Talanta*, 2013, **117**, 549-555.

44 H. D. Li, Y. Z. Liang, D. S. Cao and Q. S. Xu, *Trends in Analytical Chemistry*, 2012, **38**, 154-162.

45 B. Rosner and D. Grove, S*tatistics in Medicine*, 1999, **18**, 1387-1400.

**Captions**

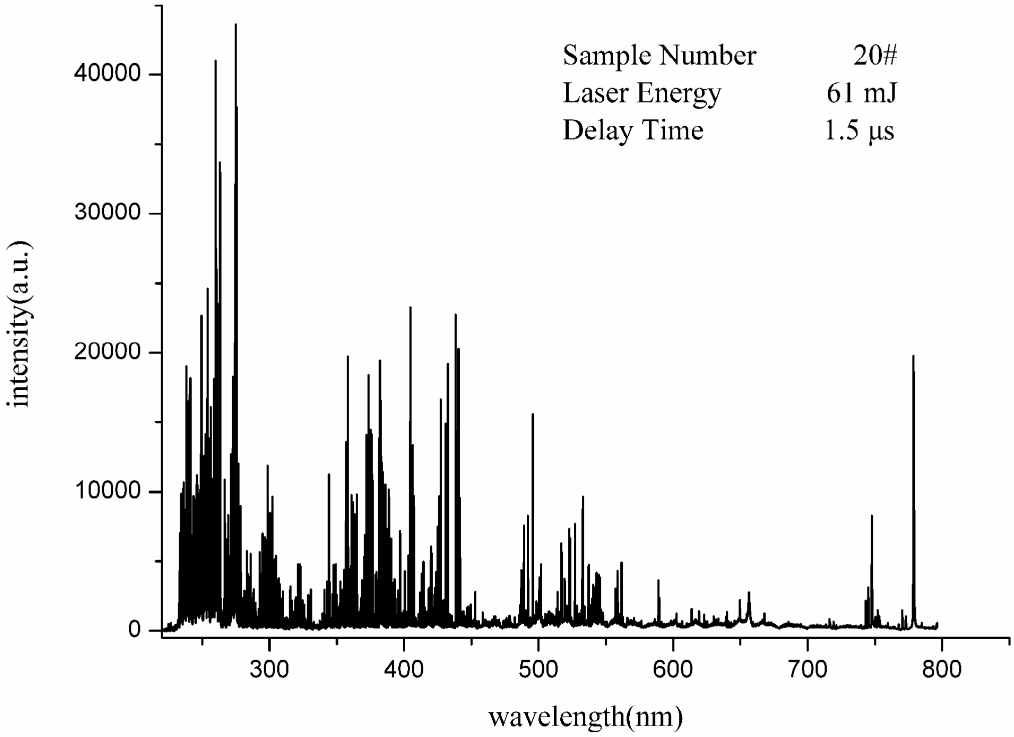


**Fig. 1** The spectra of 20# round steel measured on the LIBS system.

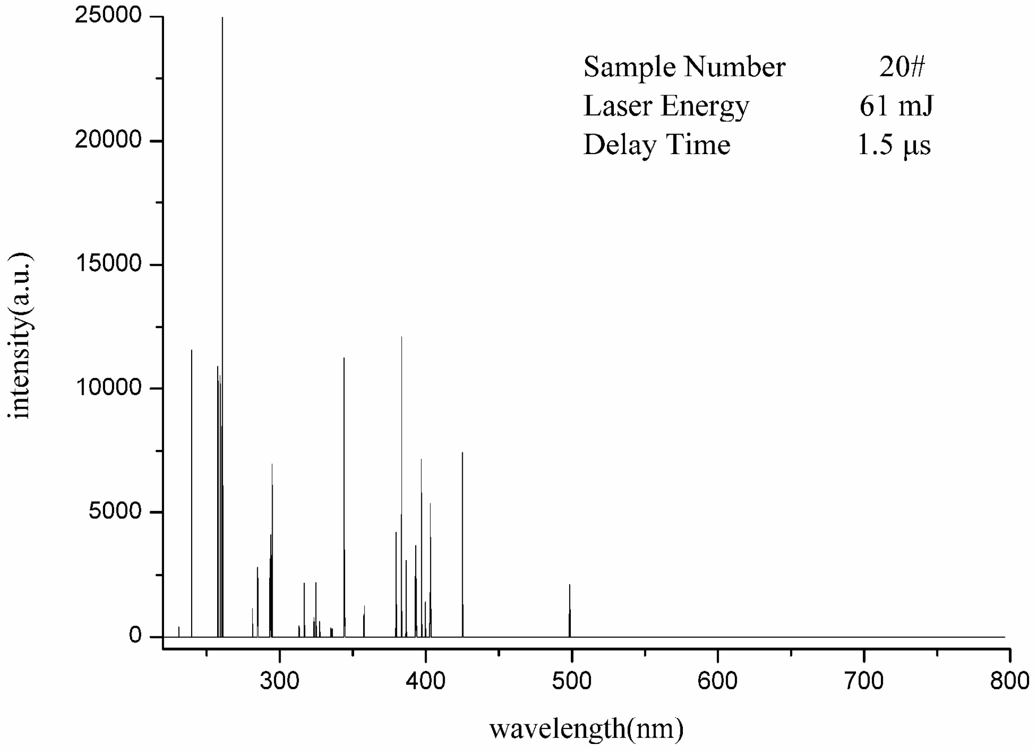


**Fig. 2** The spectrum only contains informative variables of 20# round steel.

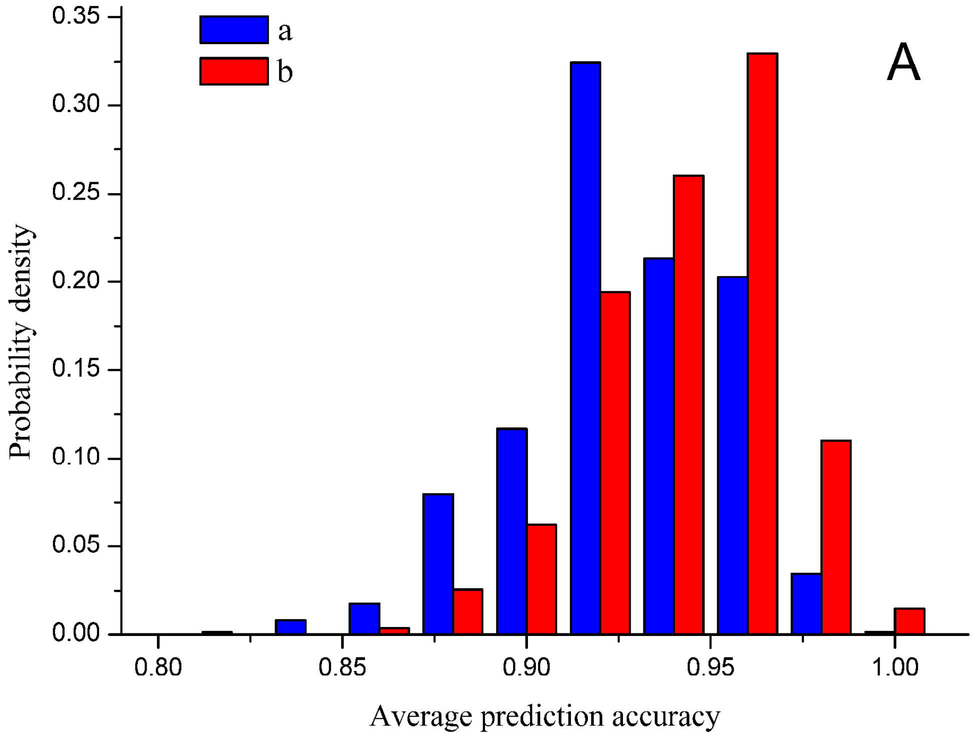


**Fig. 3(A) A**ccuracy distribution of informative variable (the 303th region).

a. Without a given variable; b. With a given variable.

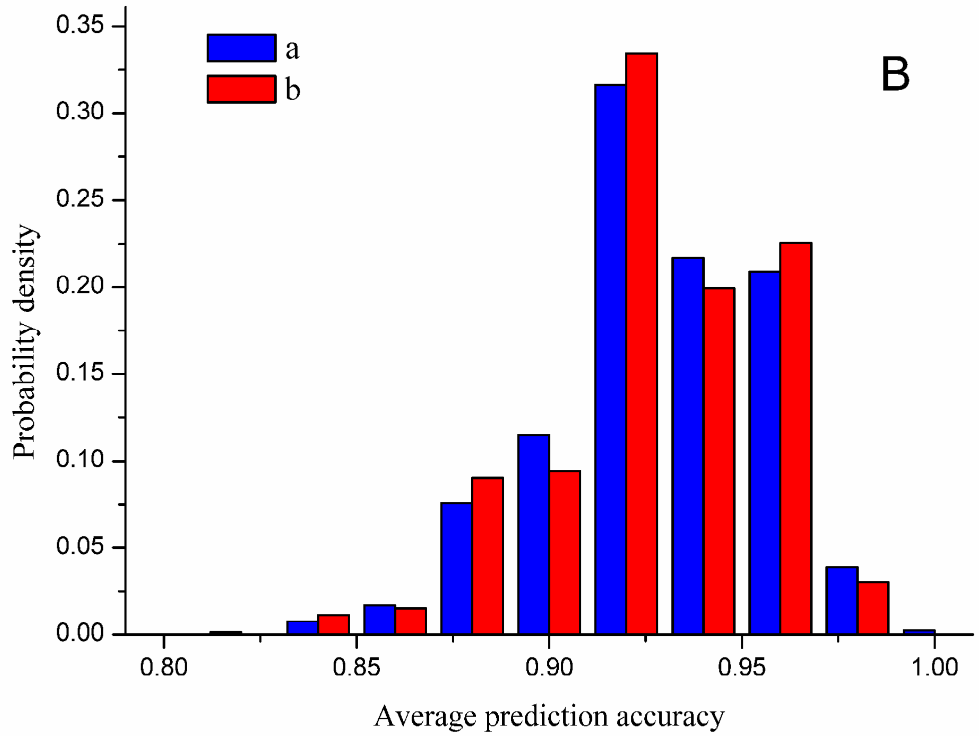


**Fig. 3(B) A**ccuracy distribution of uninformative variable (the 544th region).

a. Without a given variable; b. With a given variable.

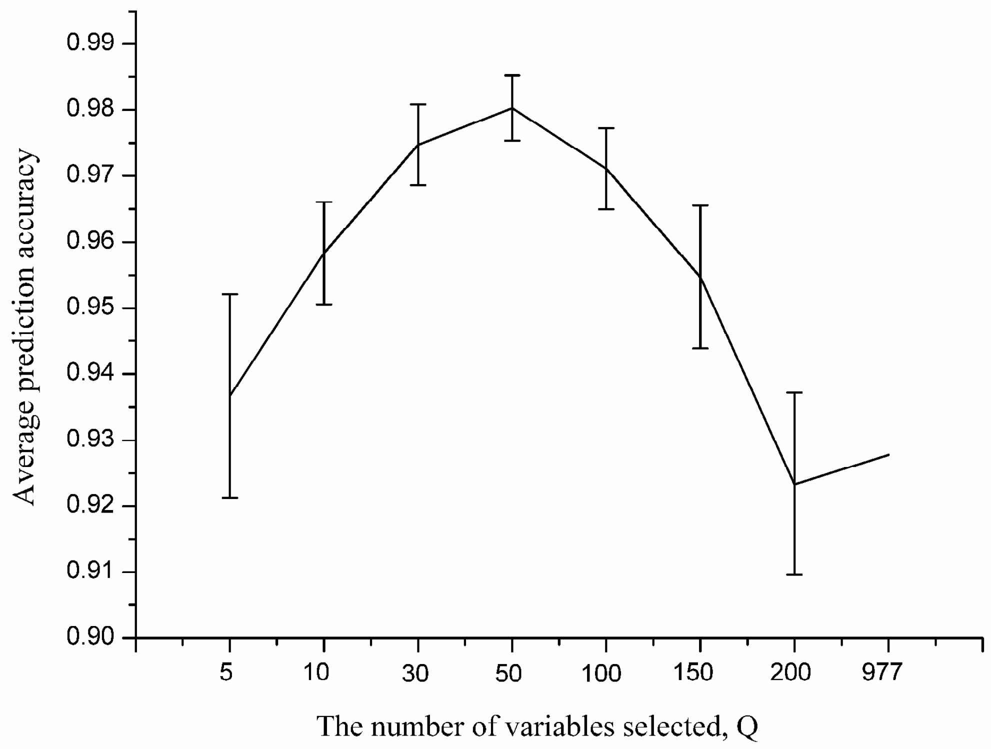


**Fig. 4** The average prediction accuracy and the standard deviations of running AIA program 20 times.

**Table 1** The result details and computation time of the informative spectra models by AIA.

| Q | Average prediction accuracy | NO. of informative variables | Training time(s) (360 spectra) | Testing time(s) (180 spectra) |
|---|---|---|---|---|
| 5 | 0.9366 | 132 | 2.0774 | 1.3163 |
| 10 | 0.9583 | 94 | 0.1912 | 0.6260 |
| 30 | 0.9747 | 52 | 0.6032 | 0.2200 |
| 50 | 0.9802 | 35 | 0.4048 | 0.1520 |
| 100 | 0.9711 | 26 | 0.3238 | 0.2260 |

| | | | | |
|---|---|---|---|---|
| **150** | **0.9547** | **22** | **0.2808** | **0.1850** |
| **200** | **0.9233** | **18** | **0.2580** | **0.2155** |
| **Whole spectra** | **0.9278** | **997** | **14.86** | **21.65** |