Spectrochimica Acta Part B 150 (2018) 49–58

Contents lists available at ScienceDirect

Spectrochimica Acta Part B

journal homepage: www.elsevier.com/locate/sab

Feature selection of laser-induced breakdown spectroscopy data for steel aging estimation

Shengzi Lu[a,b], Shen Shen[a,b], Jianwei Huang[a,b], Meirong Dong[a,b], Jidong Lu[a,b,*], Wenbing Li[a,b]

[a] School of Electric Power of South China University of Technology, Guangzhou 510640, China
[b] Guangdong Province Engineering Research Center of High Efficient and Low Pollution Energy Conversion, Guangzhou, 510640, China

ARTICLE INFO

Keywords:
Laser-induced breakdown spectroscopy
Feature selection
Aging grade
Layered interval wrapper

ABSTRACT

T91 and 12Cr1MoV are two representative heat-resistant steel widely used in industrial pressure-bearing equipment. During long-term operation, the safety and service life of the equipment will be affected by the change of metallographic structure and mechanical properties of the steel components (i.e. material aging). In this study, the spectral data of eight T91 steel specimens with different aging grades, and seven 12Cr1MoV steel specimens with different grain size grades were obtained by laser induced breakdown spectroscopy (LIBS). In order to construct steel aging estimation models, two classification models including logistic regression (LGR) and support vector machine (SVM) were used, and two representative feature selection methods including analysis of variance (ANOVA) and LGR filter were utilized to reduce the high dimensional LIBS data (12,281 initial variables) into fewer features for improving the performance of estimation models. Furthermore, a new layered interval wrapper (LIW) feature selection method was proposed for being more targeted toward LIBS data. The effects of different feature selection methods on model performance were compared and discussed. The characteristics of the variables selected by different feature selection methods were observed and analyzed. The results showed that the model performance of LGR and SVM can be improved to a certain degree by all three feature selections methods, and LIW showed a greatest improvement for classification prediction. For T91, the prediction accuracy of LGR/SVM coupled with LIW was improved to 0.92/0.94, while the prediction accuracy with full spectral input was 0.76/0.81; for 12Cr1MoV, the prediction accuracy using LIW was improved to 0.87/0.90, while the prediction accuracy with full spectral input was 0.69/0.69. This study demonstrates that LIW is a new effective feature selection method for high dimensional LIBS data.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS) is obtaining a unique position among other analytical techniques due to its advantages such as rapid real-time measurement, simultaneous multi-elemental detection, little or no sample pretreatments, and little destructive impact on the sample [1–3]. LIBS has been widely utilized in various fields as showed in a number of review papers [4–6]. However similar to other laser-ablation based analytical methods, LIBS is strongly influenced by the uncontrollable fluctuation of laser-matter interaction and experimental parameters in measurement [7], and these influences can often distort the relationship between measured value and true value. So when exploiting LIBS data in some regression or classification application, it is very difficult to directly use conventional univariate method to meet the complex requirements [8]. To overcome this problem, various multivariate chemometric methods, statistical methods and machine learning methods such as partial least squares (PLS), logistic regression, artificial neural network (ANN) and support vector machine (SVM) have already been exploited to analyze LIBS datasets [9–11]. Meanwhile with the advances in detection equipment, LIBS can provide larger number of spectral information by using high resolution spectrometer (e.g. echelle spectrograph). Sample with richer spectral features can be feed into multivariate models to seek a more reliable correlation between measured variables and dependent variables. However, the massive amount of spectral variable also brought another problem: dimensionality curse (i.e. the $p \gg n$ problem) [12,13]. The high dimensional data and relative small sample size can cause intractability of mapping the measured variables to dependent variables. In other words, the performance of multivariate model can be poor affected by the redundant variables (e.g. noise) and disturbing variables (e.g. background and overlap of spectral signatures) existing in whole spectrum. This issue will not only be an obstacle to train a model with

* Corresponding author at: School of Electric Power of South China University of Technology, Guangzhou 510640, China.
E-mail address: jdlu@scut.edu.cn (J. Lu).

https://doi.org/10.1016/j.sab.2018.10.006
Received 3 April 2018; Received in revised form 3 October 2018; Accepted 4 October 2018
Available online 05 October 2018
0584-8547/

strong prediction capability, but also lead to poor model generalization ability. The increase of model complexity resulting from high dimensional variable may obtain high prediction error even when the model bias is low, i.e. the overfitting problem. Therefore, it is necessary to select variables by appropriate method for improving estimation/prediction performance of model [14]. And it also can improve model interpretation and acquire a better understanding of variable contribution instead of feeding data in model blindly.

T91 and 12Cr1MoV steel are two representative heat-resistant steel widely used in pressure-bearing equipment such as boiler tube in power plants [15,16]. During long-term operation of the equipment, the irreversible change of mechanical properties and metallographic structure of steel will occur and lead to the degradation and structural failure (i.e. steel aging) [17]. In order to develop a faster, in-situ and non-destructive steel aging estimation technique, the feasibility of using LIBS for steel degradation analysis has been proved in our previous works [18–21], and several specific methodologies have been presented. In this paper, two independent datasets; the LIBS spectral data of T91 steel with different aging grades, and 12Cr1MoV steel with different grain size grades (which is a critical indicator for steel aging assessment [22]) were used, and different classification models combined with several feature selection methods were used to construct the discrimination models for steel aging grade/grain size grade. The purpose of utilizing feature selection methods is to reduce the dimensions of spectral data, which contains more than 12,000 variables, into fewer features for improving the performance of estimation models.

For reducing the variables number, the mainly strategies of feature selection for classification application can be concluded in three categories (as reviewed in other studies [23–25]): (i) down-selection method, (ii) statistics-based method and (iii) model-based method. The variable down-selection is a very common method utilized in LIBS studies as reported elsewhere [26,27]. Specific spectral line intensities or their ratios are selected manually based on theoretical or empirical knowledge to serve as independent inputs for classification model. This method has an obvious advantage in the interpretability of model result. However, when dealing with the spectrum consisting of complex and diverse emission lines (e.g. the spectra of steel specimens studied in this paper), it is very difficult to choose an optimal feature subset for building the model with best performance by this method. In this case, using a method which is capable of taking all spectral variable into consideration is more suitable. Therefore, this paper is aimed at the study of statistics-based and model-based feature selection method in LIBS application. It is worth mentioning that the influence of variable down-selection on the performance of classification model has been investigated in our recent publication [21].

Statistics-based method mainly works by utilizing statistical tests to estimate some numerical measure of the correlation between measured variables and dependent variables, and eliminates the measured variables with lower degree of statistical relevance. It has been used in various fields [28,29]. Larsson et al. [30] used the median of standard deviation of peak intensity to create a threshold for variable reduction in bio-sample classification. Dolgin et al. [31] utilized analysis of variance (ANOVA) to evaluate the most useful feature for rapid characterization of parchment by LIBS. In this paper, ANOVA were used as a representative statistics-based feature selection method and compared with other model-based methods. The model-based feature selection method is based on the analysis result of specific algorithm and model, and this type of method can be further divided into two categories based on their strategies: filter-method and wrapper-method. The filter-method is a relatively straightforward approach compared with wrapper. Its idea is to obtain certain numerical measure of the relevancy of each variable from the classification/regression model built with all variables (e.g. the variable importance in projection scores of PLS-based model and the coefficients of logistic regression), then introduce a threshold on these value to eliminate the variables that are considered lower importance. Zhu et al. used the variable importance in projection (VIP) scores of partial least squares discriminant analysis (PLS-DA) to estimate the importance of each variable, and found that the performance of model built with high VIP scores variables was better. Liu et al. [32] reported that the variables selected by the regression coefficients of PLS-DA can be used to build an optimal SVM model for fruit vinegars classification. However, there are few comparative studies between the filter method and other feature selection method in LIBS domain. In this paper, filter method with the coefficient of logistic regression (LGR) was used as a representative filter feature selection method for the comparison with other methods. Another approach of model-based method called wrapper is to use specific search algorithm to find a variable subspace that has best estimation/prediction performance of model among all variables. It has been confirmed as an effective feature selection method and utilized in a number of studies [33,34]. One representative of the wrapper method is the interval partial least-squares combined with genetic algorithms (GA-PLS) proposed by Hasegawa et al. [35] and further developed by Leardi and Norgaard [36,37] for Fourier transform-infrared (FT-IR) spectroscopic analysis. Besides, Fernández et al. [38] proposed a backward iterative variable selection method by using the root-mean-square error of prediction (RMSEP) as a selection criterion in near-infrared spectroscopy (NIR) application. Although it has proved that the wrapper methods mentioned above such as GA-PLS and SPA are effective approaches in other domains, it is infeasible to use them directly in LIBS application due to the different characteristics of spectral data. The variable number of LIBS spectrum is much larger (when using a 6-channel high resolution spectrometer, the variable number can reach more than ten thousand), it will increase the number of iterations exponentially even when using the heuristic or random search algorithms. Meanwhile, LIBS signals have different spectral characteristics (e.g. more inherently narrow emission peaks) compared with other spectrum technology such as NIR and FT-IR. Those feature selection methods designed for NIR and FT-IR might not be effective for LIBS spectrum. Therefore, a layered interval wrapper (LIW) feature selection method was proposed and discussed in this paper for accelerating the computation speed and being more targeted toward the characteristics of LIBS data.

In this work, we aim at investigating the influence of several different feature selection methods on the performance of classification models in LIBS application. For this purpose, ANOVA (used as a statistics-based method) and filter method with the coefficient of LGR were compared and discussed. Furthermore, a layered interval wrapper (LIW) method was proposed for accelerating the computation speed and being more targeted toward LIBS data. The spectral data were obtained from LIBS measurement of T91 steel with different aging grades and 12Cr1MoV steel with different grain size grades. SVM and logistic regression were selected as prediction models to evaluate the effectiveness of different feature selection methods. The overall accuracy was selected as the figure of merit of model performance [39]. And the effects of different feature selection methods on LIBS data were observed and discussed.

2. Experimental

2.1. Samples

Two sets of specimens, 7 12Cr1MoV steel specimens with different grain size grades, and 8 T91 (10Cr9Mo1VNbN) steel specimens with different aging grades were selected for the analysis. These two kinds of steel both are representative heatproof steels widely used in thermal power stations. The steel specimens used for analysis were acquired by a series of heat and load treatments to original steel tubes.

The 12Cr1MoV steel specimens with different grain size grades was acquired by quenching and tempering at different temperatures. Firstly, the specimens cut from the original 12Cr1MoV steel tube were heated to 970 °C for 10 min in a preheated muffle furnace, and placed in cold water for cooling. Then the quenched specimens were heated in muffle