Spectrochimica Acta Part B 107 (2015) 1–10

Contents lists available at ScienceDirect

Spectrochimica Acta Part B

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

A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy

Thomas F. Boucher [a,*], Marie V. Ozanne [b], Marco L. Carmosino [a], M. Darby Dyar [b], Sridhar Mahadevan [a], Elly A. Breves [b], Kate H. Lepore [b], Samuel M. Clegg [c]

[a] School of Computer Science, University of Massachusetts Amherst, 140 Governor's Drive, Amherst, MA 01003, United States.
[b] Department of Astronomy, Mount Holyoke College, South Hadley, MA 01075, United States
[c] Los Alamos National Laboratory, P.O. Box 1663, MS J565, Los Alamos, NM 87545, United States

ARTICLE INFO

Article history:
Received 6 May 2014
Accepted 3 February 2015
Available online 12 February 2015

Keywords:
Laser-induced breakdown spectroscopy (LIBS)
Partial least squares (PLS)
Support vector regression (SVR)
Lasso
Principal component regression (PCR)

ABSTRACT

The ChemCam instrument on the Mars *Curiosity* rover is generating thousands of LIBS spectra and bringing interest in this technique to public attention. The key to interpreting Mars or any other types of LIBS data are calibrations that relate laboratory standards to unknowns examined in other settings and enable predictions of chemical composition. Here, LIBS spectral data are analyzed using linear regression methods including partial least squares (PLS-1 and PLS-2), principal component regression (PCR), least absolute shrinkage and selection operator (lasso), elastic net, and linear support vector regression (SVR-Lin). These were compared against results from nonlinear regression methods including kernel principal component regression (K-PCR), polynomial kernel support vector regression (SVR-Py) and *k*-nearest neighbor (*k*NN) regression to discern the most effective models for interpreting chemical abundances from LIBS spectra of geological samples. The results were evaluated for 100 samples analyzed with 50 laser pulses at each of five locations averaged together. Wilcoxon signed-rank tests were employed to evaluate the statistical significance of differences among the nine models using their predicted residual sum of squares (PRESS) to make comparisons. For MgO, SiO_2, Fe_2O_3, CaO, and MnO, the sparse models outperform all the others except for linear SVR, while for Na_2O, K_2O, TiO_2, and P_2O_5, the sparse methods produce inferior results, likely because their emission lines in this energy range have lower transition probabilities. The strong performance of the sparse methods in this study suggests that use of dimensionality-reduction techniques as a preprocessing step may improve the performance of the linear models. Nonlinear methods tend to overfit the data and predict less accurately, while the linear methods proved to be more generalizable with better predictive performance. These results are attributed to the high dimensionality of the data (6144 channels) relative to the small number of samples studied. The best-performing models were SVR-Lin for SiO_2, MgO, Fe_2O_3, and Na_2O, lasso for Al_2O_3, elastic net for MnO, and PLS-1 for CaO, TiO_2, and K_2O. Although these differences in model performance between methods were identified, most of the models produce comparable results when $p \leq 0.05$ and all techniques except kNN produced statistically-indistinguishable results. It is likely that a combination of models could be used together to yield a lower total error of prediction, depending on the requirements of the user.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

A laser-induced breakdown spectrometer (LIBS), along with a remote microscopic imager, comprises ChemCam [1,2], a payload instrument on the Mars Science Laboratory (MSL) rover *Curiosity*. This LIBS instrument records emission spectra in the ultraviolet (UV), violet (VIO), and visible to near-infrared (VNIR) ranges. The laser can be focused on a small location size of roughly <0.5 mm from a standoff distance of up to 7 m. ChemCam is being used to determine chemical compositions of dust, rocks, and minerals on the Martian surface. To aid in such quantitative analyses, a broad training set of LIBS spectra of geological standards with known compositions is being developed for calibration [3]. The goal of ChemCam is to produce robust, accurate chemical analyses of minerals, rocks, and soils on the Martian surface.

However, producing quantitative chemical analyses from LIBS data is a challenging task due to the wide variety of chemical compositions found on Mars. Ionization states from the many different elements found in geological materials may interact in the LIBS plasma, causing variations in line intensities that defeat univariate analysis techniques using single-peak calibrations of intensity vs. concentration. Multivariate analysis techniques are thus needed to account for the covariate interactions that occur within the LIBS plasma. They are designed to provide stable models when the data suffer from multicollinearity, and are better suited to LIBS data analysis.

* Corresponding author.
E-mail address: boucher@cs.umass.edu (T.F. Boucher).

http://dx.doi.org/10.1016/j.sab.2015.02.003
0584-8547/© 2015 Elsevier B.V. All rights reserved.

Thus, this paper explores a variety of current machine learning algorithms for regression problems and compares their performance on a suite of 100 spectra from igneous and meta-igneous rocks. LIBS spectral data are analyzed here using linear methods including partial least squares (PLS-1 and PLS-2), principal component regression (PCR), least absolute shrinkage and selection operator (lasso), elastic net, and linear support vector regression (SVR-Lin). These were compared against results from nonlinear methods including kernel principal component regression (K-PCR), polynomial kernel support vector regression (SVR-Py) and *k*-nearest neighbor (*k*NN) regression to discern the most effective models for interpreting elemental concentration from LIBS spectra of geological samples. Ten-fold cross-validation was used to train the parameters and tune the hyperparameters of each model using 70 samples while 30 samples were held out for use as a test set. Wilcoxon signed-rank tests were employed to evaluate the statistical significance of differences among the nine models using their predicted residual sum of squares (PRESS) to make comparisons. Results show the advantages of linear models for this application, and lend insights into best practices for interpretation of data from ChemCam and other LIBS studies of geological samples elsewhere in the solar system.

2. Background

LIBS relies on quantized valence-electron transitions that occur when the electrons move to an excited state in the presence of an excitation source and subsequently decay back down to their ground states, emitting photons. When these transitions are detected by a spectrometer, emission lines are observed at wavelengths that are specific to the elemental or ionic electron source.

LIBS is challenging to use for geological sample analysis because peak intensities and areas are influenced by interactions in the plasma that are partially a function of the sample's chemical composition. These interactions are collectively referred to as matrix effects; they are chemical properties of a material that influence the extent to which a given wavelength emission is detected compared to the true abundance of the parent element. The matrix effects are related to the relative abundances of neutral and ionized species within the plasma, collisional interactions within the plasma, laser-to-sample coupling efficiency, and self-absorption [4]. Fortunately, advanced statistical analysis techniques can tease out relationships that may be obscured by matrix effects.

Multivariate analyses have been used increasingly for LIBS over the last decade, starting with the applications of principal components [5] and partial least squares [4–10]. A few other methods have been investigated, such as artificial neural networks (ANN) [11]; however, results showed that PLS was equivalent or superior to ANN. A few forays have been made into the sparser models (lasso) [12] and intelligent selection or rejection of training set spectra based on clustering methods [11]. Both of these show promise in improving results, particularly with clustering, and in more closely connecting the models with physical details, i.e., with lasso predominantly using the emission lines of the element of interest. Here we follow up on these works by comparing and contrasting additional methods for providing sparseness to the data.

An ideal regression model for LIBS should be sparse, interpretable, and well predicting. The property of sparsity, in which a small subset of predictor variables drives the prediction results, can be critical to instrument design because it may enable improved count rates and higher-resolution spectra by guiding sampling to fewer channels more frequently. It may not, however, enable model interpretation, because the chosen features are dependent upon a complicated convolution of end-member oxide spectra, experimental conditions, and measurement errors [13–15]. In this paper, several multivariate analysis techniques that meet these criteria to varying degrees are utilized and compared to assess the effectiveness of each model and the effects of training set size on the resultant predictions.

Table 1 provides a summary of the methods considered in this study. The following discussion provides some background on the techniques to be compared.

Table 1
Summary of models used.

Method	Summary	Tuning parameters	Advantage(s)	Disadvantage(s)	Other
PLS	Projects explanatory matrix, **X**, into a subspace of latent components that maximize the covariance of **X** and the response matrix, **Y**.	*k*, # of components	Used when **X** has many collinear features and when p> > N. Provides a stable multivariate model that can account for all oxides (PLS-2).	Provides a complex model in which all coefficients are linear combinations of the original channels. Involves a complex optimization problem with no simple, closed-form representation.	Linear, uses all channels (not sparse)
Lasso	Shrinks some coefficients and sets others equal to zero in accordance with shrinkage parameter. Provides a sparse model that can be used for both feature selection and composition predictions.	α, sparsity weight	Provides an interpretable model, selects subset of predictors with the strongest effects on the response variable. Can be used for feature selection when less data are available.	Arbitrarily chooses one covariate from a group of highly collinear covariates to use in the model and discards the rest [18].	Linear, sparse, eliminates noisy channels
Elastic net	Extends the lasso. Shrinks some coefficients and sets others equal to zero; averages highly correlated features and shrinks averages. Provides a sparse model that has more terms than the lasso and can be used for feature selection and composition predictions.	α and l_1 ratio	Performs well in the $p >> N$ case. Provides an interpretable model that is more stable than the lasso. Useful for feature selection.	Cannot be used for feature selection in situations when less data are available because it overwhelms the data with too many model variables.	Linear, sparse, eliminates noisy channels
PCR	Projects data to a low-dimensional uncorrelated subspace, then uses ordinary least squares to regress in the latent space.	*k*, # of components	De-correlates the data and reduces its dimensionality, combating the "curse of dimensionality"	Higher order polynomial kernels tend to over-fit the training set and poorly predict the testing set in this application.	May be linear or nonlinear; both use all channels
SVR	Uses only a subset of the training data (support vectors) to construct a model that is most generalizable. Can be linear or nonlinear depending on the kernel function used.	ϵ, sensitivity	Performs well with a linear kernel. Can be either linear or nonlinear depending on the kernel.	As above, polynomial kernels tend to over-fit the training set and poorly predict the testing set in this application.	May be linear or nonlinear; either uses all channels
*k*NN	A nonlinear regression model that predicts samples using a weighted interpolation of the *k* nearest training samples.	*k*, # of neighbors	Requires no model training other than choosing the number of neighbors, reducing run time and making it scale well to large data sets.	Tends to over-fit the training data and is only as effective as the distance metric used to compare samples.	Nonlinear, uses all channels

Download English Version:

https://daneshyari.com/en/article/1239615

Download Persian Version:

https://daneshyari.com/article/1239615

Daneshyari.com