Chemometrics and Intelligent Laboratory Systems 167 (2017) 139–151

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Detection of nonlinearity in soil property prediction models based on near-infrared spectroscopy

Lu Yan, Matheus S. Escobar, Hiromasa Kaneko, Kimito Funatsu *

Department of Chemical System Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Keywords:
NIR spectroscopy
Soil properties
Regression analysis
Nonlinearity detection

ABSTRACT

Soil property analysis is indispensable in precision agriculture, an advanced field regarding site-specific management for crop production enhancement and environmental sustainability. Because of the difficulties in soil sample collection and measurement of soil properties, such as moisture content, total carbon, total nitrogen, electricity, and pH, near-infrared (NIR) spectroscopy is a useful technique to predict soil properties by using statistical learning methods. However, the prediction of soil properties without any knowledge about how different variables might influence their behavior is not adequate. Soil properties differ depending on location and environment. The variability within the same area could cause nonlinearity on a global scale. Therefore, to determine which method and strategy are suitable for this task, the detection of nonlinearity between NIR spectroscopy and soil properties is the main purpose of this study. Various numerical tools and graphical methods were applied to this soil property dataset, such as variable selection, sample splitting, applicability domain evaluation, and residual inspection. Global nonlinearity for all five soil properties was confirmed, and the strength of such nonlinearities was found to be property dependent.

1. Introduction

Precision agriculture is an advanced concept used to accomplish site-specific management based on developed decision support system [1]. It includes observing, measuring, and collecting information from crops, soils, and other conditions. By using site-specific knowledge, one can precisely apply fertilizer, water, and other chemicals to a particular location [2]. The goal of precision agriculture is to improve the production rate and quality of crop yields and to keep it sustainable and environment friendly [3]. One main aspect of precision agriculture is gathering soil samples and making a decision support map [2]. Because soil sample collection consumes a large cost, an alternative approach that indirectly predicts soil properties reveals itself a promising research topic.

near-infrared (NIR) spectroscopic measurement is quite convenient, because no special preparation of samples is needed. Thus, NIR data collection for soil samples can be done on-site. In the present study, a framework for soil property prediction based on NIR spectra was developed using statistical learning methods. Various studies have shown that the use of NIR spectra is feasible to predict the following soil properties: moisture content, total carbon, total nitrogen, electrical conductivity, and pH [4–6]. Linear regression and partial least squares (PLS) methods were already applied to determine the soil properties. Additionally, variable selection is an important part of NIR spectra calibration. Knowledge-based selection is a manual approach. PLS could be combined with other various variable selection methods such as stepwise, bootstrap, moving window, and forward/backward interval methods [7]. Genetic algorithm (GA) is another commonly used approach [7]. Removal of uncorrelated variables promotes the investigation of corresponding chemical substances based on optimized variables. Improving prediction ability is another purpose of using NIR spectroscopy.

There are other obstacles, however, when using this technique for real soil property prediction. We still do not know the inner deep relationship between NIR spectra and soil properties, because soil property differs depending on the area. More specifically, linearity or nonlinearity is the first issue that should be considered. For real case, it is unwise to use a linear method to predict nonlinearity data. Therefore, a deeper insight into the relationship approximated by models is the main topic of this study.

In addition to PLS, a well-known nonlinear regression method called support vector regression (SVR) can be applied to NIR spectroscopy, as reported by Thissen et al. [8]. Local strategy is another approach when dealing with nonlinearity [9,10]. Constructing a local model according to the local dataset might be more suitable for specific test data. We applied

* Corresponding author.
E-mail address: funatsu@chemsys.t.u-tokto.ac.jp (K. Funatsu).

http://dx.doi.org/10.1016/j.chemolab.2017.04.001
Received 24 August 2016; Accepted 3 April 2017
Available online 1 May 2017

global and local strategies for linear method PLS and nonlinear method SVR to predict the aforementioned five soil properties. For model superiority comparison, conventional root mean square of prediction (RMSEp) and a model comparison statistic called randomization *t*-test [11] were applied to this study.

Wavelength selection is a graphical technique used to investigate the (non)linearity between NIR wavelength and objective variables. H.Arakawa and K.Funatsu developed a GA-based wavelength selection (GAWLS), which uses regional variable selection to make it more suitable for NIR spectra calibration [5]. Nonlinear regression method could also be used as the fitness function in GA, one application is nonlinear regional variable optimization method (GAWLS-SVR) [12]. If nonlinearity exists, optimized variables found using each method should be different.

For regression analysis, applicability domain (AD) evaluation is essential. By using specific criteria to estimate prediction errors such as data similarity and distance, we can know how a model will perform for new data [13,14]. AD evaluation based on ensemble learning was proposed and reported to be appropriate for regression methods [15–18]. Small variance usually suggests small error; therefore, using standard deviation obtained from predicted values to estimate prediction error is an interesting approach. In this study, GAWLS and GAWLS-SVR were run multiple times for predicting each sample; thus, they are applicable to AD comparison.

By comparing various results, one can represent only the trend or indication of nonlinearity. A nonlinearity detection method based on PLS model was already proposed, which has become a universal tool [19], where plotting residuals against latent variables is used to inspect which latent variable contributes to PLS fitness. Moreover, the Durbin–Watson (D–W) test evaluates the residual prediction errors and reports on linearity, nonlinearity, or inconclusiveness [19–21].

2. Review on chemometric methods

Before introducing the data to be used and regression modeling results, some basic concepts of chemometric methods are briefly explained in this section. PLS and SVR methods were the two basic algorithms used in this study. These two methods were combined with various chemometric tools.

2.1. PLS

A Swedish statistician Herman Wold introduced the PLS method; then his son, Svante Wold, developed it [22]. PLS is a regression method that, different from multiple linear regression (MLR), determines multiple hyperplanes of maximum variance between the explainable variables (**X**) and objective variables (**Y**). PLS projects both **X** and **y** to latent space, where the amount of information between **X** and **y** is defined by their covariance. Generally, **y** is a column vector, PLS1 is a widely used algorithm for the vector **y** case [23],

$$\mathrm{X} = \sum_{i=1}^{A} \mathbf{t}_a \mathbf{p}_a^{\mathrm{T}} + \mathbf{E} = \mathrm{TPT} + \mathrm{E} \tag{1}$$

$$\mathrm{y} = \sum_{i=1}^{A} \mathbf{t}_a \mathbf{q}_a + \mathbf{f} = \mathrm{Tq} + \mathrm{F} \tag{2}$$

where A is the number of latent variables, **T** is the projection matrix of **X**, **P** is an orthogonal loading matrix, q is a coefficient, and matrices **E** and **F** are the error terms. Latent variable **t** plays a very important role in PLS algorithm because it is determined by the linear combination **t**=**Xw**, where **w** is a d×1 weight vector, and its norm is 1.

When the sum of squares of error terms **E** and **F** is minimal, the loading matrix **P** and coefficient q are given as follows

$$P = \mathbf{X}^{\mathrm{T}}\mathbf{t}/\mathbf{t}^{\mathrm{T}}\mathbf{t} \tag{3}$$

$$q = \mathbf{y}^{\mathrm{T}}\mathbf{t}/\mathbf{t}^{\mathrm{T}}\mathbf{t} \tag{4}$$

If the next latent variables, $\mathbf{X}_{\mathrm{new}}$ and $\mathbf{y}_{\mathrm{new}}$, are necessary, then they are calculated using the following equations:

$$\mathbf{X}_{\mathrm{new}} = \mathbf{X} - \mathbf{tp} \tag{5}$$

$$\mathbf{y}_{\mathrm{new}} = \mathbf{y} - \mathbf{tq} \tag{6}$$

One of the most important advantages of PLS is that it can treat data with more variables than observations, which cannot be accomplished by MLR. The more the latent variables are used, the more accurate is the model for training samples. However, its generalization ability for other test samples would be decreased considerably, which is called overfitting. In general, the number of latent variables is decided by cross-validation. By randomly separating training samples into several folds, the properties of one fold are predicted by a PLS constructed by the remaining folds. The number of latent variables is decided according to the minimum prediction error obtained. Although the original applications of PLS were in social science, today PLS regression is more widely used in chemometrics.

2.2. SVR

In machine learning, support vector machine (SVM) is a well-known supervised learning method that is already widely applied in classification and regression analysis. The original motivation of SVM was initiated from binary classifier. The application of SVM to regression problem was proposed by Vladimir N. Vapnik [24]. In linear regression problem, a regularized error function could be minimized by

$$\frac{1}{2}\sum_{n=1}^{N}\{\mathrm{y_n} - \mathrm{t_n}\}^2 + \frac{\lambda}{2}||w||^2 \tag{7}$$

In SVR, this quadratic error function is replaced by a ε-insensitive error function for a sparse solution; if the absolute difference between the prediction y and the target t is less than ε, then the error is considered zero. Thus, we can minimize a regularized error function as follows:

$$C\sum\nolimits_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{1}{2}|\mathbf{w}^2| \tag{8}$$

where $\mathrm{y(x_n)}=\mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x})+\mathrm{b}$, φ(x) denotes the fixed feature space transformation, and C is the regularization parameter. Then, by introducing the Lagrange multipliers and minimizing the error function, we set the derivative of the Lagrangian with respect to ***w***,

$$w = \sum_{n=1}^{N}(a_n - \widehat{a}_n)\phi(x_n) \tag{9}$$

Predictions for new inputs can be made by substituting Eq. (9) in $\mathrm{y(x_n)}=\mathrm{w}^{\mathrm{T}}\varphi(\mathrm{x})+\mathrm{b}$

$$y(\mathbf{x}_n) = \sum\nolimits_{n=1}^{N}\left(a_n - \widehat{a}_n\right)k(\mathbf{x},\mathbf{x}_n) + b \tag{10}$$

which is expressed in terms of a kernel function, $\mathrm{k(x,x_n)}=\varphi(\mathrm{x'})^{\mathrm{T}}\varphi(\mathrm{x})$. The parameter b can be determined by considering a data point that lies on the error tube, which has $\xi_n = 0$; therefore, it satisfies $\varepsilon + y_n - t_n = 0$. Using (10) hen, b can be determined using the following equation:

$$\mathrm{b} = t_n - \varepsilon - w^T\phi(x_n) = t_n - \varepsilon - \sum_{m=1}^{N}\left(a_m - \hat{a}_m\right)k(x_n,x_m) \tag{11}$$

According to the corresponding Karush–Kuhn–Tucker (KKT)