Contents lists available at ScienceDirect

Geoderma

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

Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm

Yao Zhang[a,b], Minzan Li[a,*], Lihua Zheng[a], Qiming Qin[b], Won Suk Lee[c]

[a] Key Laboratory of Modern Precision Agriculture System Integration Research, Ministry of Education, China Agricultural University, Beijing 100083, China
[b] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing 100871, China
[c] Department of Agricultural & Biological Engineering, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Editor: Dr. A.B. McBratney

Keywords:
Ant colony optimization
Feature selection
Mutual information
Near infrared
Spectroscopy
Soil total nitrogen

ABSTRACT

Nondestructive and rapid estimation of soil total nitrogen (TN) content by using near-infrared spectroscopy plays a crucial role in agriculture. The obtained original spectrum, however, presents several disadvantages, such as high redundancy, large computation, and complex model, because it generally processes a large amount of data. This study aimed to determine soil TN content-sensitive wavebands with high information quality, considerable predictive ability, and low redundancy. This paper proposes an evaluation criterion in selecting sensitive wavebands based on three factors, namely, degree of relevance with target variables, representative ability of the entire spectral information, and redundancy of the selected wavebands. Based on these three factors, two methods, namely, mutual information (MI) algorithm and the combination of ant colony optimization (ACO) and MI, were innovatively developed to identify soil TN content-sensitive wavebands. After the analysis and comparison, a set of wavelengths, including 943, 1004, 1097, 1351, 1550, 1710, 2123, and 2254 nm, using the ACO–MI combined method was selected as the soil TN content-sensitive wavebands to estimate the TN content of soil samples, under four soil types, collected from different regions. The partial least squares (PLS) models based on full-spectral information, multiple linear regression (MLR) models and support vector machine (SVM) regression models based on the eight selected wavelengths for soil TN content were established separately. After the comparison, the MLR and SVM models achieved higher accuracies than the PLS models based on the full spectral information. In addition, the SVM models got the best results. In the calibration group, the coefficients of determination (R^2) was 0.989, and the root mean square errors (RMSE) of calibration was 0.078 g/kg. In the validation group, the R^2 was 0.96, and the RMSE of prediction was 0.219 g/kg. The residual predictive deviation (RPD) was 5.426. For the soil samples with TN content in the range of 0–1 g/kg, the detection precision also reached a high level. Therefore, the eight sensitive wavebands selected through the ACO–MI method performed good mechanism, universality and predictive ability in soil TN content estimation. The ACO–MI method would be valuable for soil sensing in precision agriculture.

1. Introduction

Soil is the primary support for soil-grown crops. It is an important medium for plant root extension and the main nutrient source for crop growing. The main soil nutrients include TN, OM, available potassium, and available phosphorus (Chacón Iznaga et al., 2014; Sinfield et al., 2010). Among those soil nutrients, soil nitrogen (TN and available nitrogen) plays the most important role in promoting the growth of leaf, root, and stem and is a decisive factor to the crop yield (Bansod and Thakre, 2014). Excessive nitrogenous fertilization however will cause environmental pollution and crop distortion of growth and quality. The amount of nitrogen fertilizer applied therefore needs to be precisely controlled to ensure the crop yield and environmental protection. The fundament of precision fertilizing is effectively acquiring the soil information in the field. Rapid and precise acquisition of soil nitrogenous information in farmlands hence becomes increasingly important. The conventional method of detecting soil nitrogen content usually takes several days and consumes toxic chemicals. The conventional method

Abbreviations: TN, total nitrogen; OM, organic matter; NIR, near-infrared; MI, mutual information; ACO, ant colony optimization; PLS, partial least squares; MLR, multiple linear regression; SVM, support vector machine; R^2, coefficients of determination; SSR, sum of squares for regression; SST, sum of squares for total; SSE, sum of squares for error; RMSE, root mean square errors; RPD, residual predictive deviation

[*] Corresponding author.
E-mail address: limz@cau.edu.cn (M. Li).

https://doi.org/10.1016/j.geoderma.2018.07.004
Received 19 December 2017; Received in revised form 25 June 2018; Accepted 3 July 2018

also has several disadvantages, such as high requirements for detection personnel, expensive testing equipment, low efficiency, and environmental pollution (Debaene et al., 2014; Florinsky et al., 2002; Kuang and Mouazen, 2013; Moore et al., 1993; Nocita et al., 2014). By contrast, spectral analysis techniques are based on the internal relations between radiation energy and the composition and structure of matters. According to the characteristic spectra of matter, the target concentration or properties can be determined rapidly without chemicals (Igne et al., 2010; Vohland et al., 2011). For soil, the spectral information related to most of the organic radical groups containing hydrogen was in the NIR region (Li, 2006). NIR spectroscopy is a rapid, non-destructive, and non-pollutant testing method that plays an growing important role in soil nutrition measurement and exhibits extraordinary development potential in applications of soil TN content detection (Chang et al., 2001; Lucà et al., 2017; Morellos et al., 2016).

In the detection of soil TN content with NIR spectroscopy, the spectra of soil samples are first measured, and the NIR spectral data are then used as the input variables to establish the prediction models. Modern spectrometers possess high spectral resolution, and spectral data measurement generally involves hundreds or thousands of wavelength variables. Three kinds of information variables are involved in measurement of such superlarge-scale data. One is the effective informative variable, which can improve the model predictive ability because it reflects the characteristics of the target substance in the NIR region. The second is redundant or interfering variable, which is related with other targets. The last one is uninformative variable, which is irrelevant to the target material and usually caused by the measurement environment, such as noise. If the prediction model was established by the entire-spectrum information, then the latter two kinds of variables would increase the computation complexity and reduce the target prediction accuracy of the model. The multivariate calibration model is therefore a better choice when the informative variables could be selected appropriately, which can help in simplifying the calibration model and improving the model's predictive ability in terms of accuracy, speed and robustness (Petropoulos et al., 2012; Sorol et al., 2010). Several studies confirmed that the models adopting limited characteristic wavebands only are better than the ones with entire-spectrum information. Cai et al. (2008) employed Monte Carlo uninformative variable elimination method to extract the characteristic bands and then established a PLS model to predict the sugar content of tomato. The results showed that the prediction accuracy and robustness of the proposed model were all better than the model on the basis of the entire band. Gao et al. (2009) proposed a method combining screened contribution and SPA to select soil TN content-sensitive wavebands. The result of the established multivariable model was more precise than the result derived from PLS for the entire spectra. Several scientists took advantage of the specific wavelength of the light source to develop a soil spectral detection device and obtained good detection accuracy (An et al., 2014; Li et al., 2010). The results revealed that the use of limited wavelength of light source in detecting soil nutrients had reached a practical level. The existing characteristic wavelength selection algorithm however revealed that several drawbacks in soil TN prediction remained. The selected wavebands had a weak mechanism interpretation about soil TN, and the accuracies of the established models were relatively low. The methods for selecting the characteristic wavebands therefore need to be further explored for the study of NIR spectral technology on soil TN content detection.

ACO is an evolutionary algorithm used to simulate the natural foraging behavior of ant colonies and was introduced in the early 1990s (Colorni et al., 1991). The technique is based on updating the coordination mechanism of seeking the shortest path to achieve intelligent search and parametric optimization. The ACO algorithm has been extensively applied in feature selection because of its prominent advantages, such as information positive feedback, distributed computation, heuristic search, robustness, and easy to combine with other algorithms (Dorigo et al., 1996). Aghdam and Kabiri (2016) proposed an intrusion detection system with features optimally selected using ACO to improve the performance. Varma et al. (2016) also used fuzzy entropy-based heuristic for ACO to search for the global best smallest set of network traffic features for real-time intrusion detection data set. Selection of the optimal spectral characteristic variable is also a combinatorial optimization issue. The features, such as global, discrete, and probability selection of self-adaptive ACO, are applicable to spectroscopy analysis. Hou et al. (2016) applied ant colony clustering algorithm to detect Grapevine leafroll disease (GLD) spectral anomalies on four GLD-infected vineyards from multi-spectral images for precision disease management. The classification accuracies of Non-, GLD1-, GLD2-, and GLD3-infected grapevines were 94.4%, 75%, 84.6%, and 83.3%, respectively. Guo et al. (2014) selected sensitive wavebands from apple NIR spectroscopy using ACO–PLS optimized algorithm based on the features of heuristic global search and the random selection mechanism of Monte Carlo roulette to predict soluble solid content. The prediction model obtained a good prediction performance for SSC with correlation coefficients of 0.970, and RMSE of prediction of Brix of 0.514. Allegrini and Olivieri (2011) employed the concept of cooperative pheromone accumulation, which is typical of ACO selection methods, and optimized PLS models using a pre-defined number of variables and employing a Monte Carlo approach to discard irrelevant sensors. Ke et al. (2008) proposed an ACO-based algorithm to deal with feature selection in rough set theory and compared its performance with the simulated annealing-, genetic algorithm-, and Tabu search-based algorithms. The results showed the proposed algorithm achieved better performance according to both the classification results and the number of features. Santana et al. (2010) found ACO performed better than genetic algorithm-based feature selection method for ensemble classifiers when the number of individual classifiers was small. Agrawal and Kaur (2018) compared ACO and Hybrid Particle Swarm Optimization in test case selection. The results indicated ACO outperforms Hybrid Particle Swarm Optimization in the calculating efficiency.

The studies discussed above used ACO to carry out feature selection in numerous areas. After the comparison between ACO and other heuristic algorithms, ACO performs more flexible and efficient. It is especially suitable for the relatively small-scale problems (Xue et al., 2016). In order to make further improvement on the performance of ACO, the information theory, as a supplementary, could explore more and deeper information from the variables themselves. When using ACO, few researchers however adopt information theory to improve the performance of ACO in feature selection. Several studies adopted the Monte Carlo roulette principle to select features, which is random and lack connection with the variable information (Allegrini and Olivieri, 2011; Guo et al., 2014). In this way, the selected features were mostly dependent only on the performance of ACO. Merging ACO and information theory could significantly benefit feature selection, especially when the training set is not big enough to represent the whole application space.

In the early 1990s, Battiti (1994) and Lewis (1992)introduced MI theory into variable selection research. The feature selection method based on MI has gained considerable attention after 20 years of development. MI is an excellent tool for quantitatively calculating the common information between two random variables. The MI theory has therefore been extensively applied as an effective indicator for investigating correlation. MI can also be used to measure the arbitrary dependencies between random variables, which makes it suitable for assessing the "information content" of features in complex tasks.

Filter methods are defined by a criterion *J* based on MI, also referred to as a "relevance index" or "scoring" criterion, which is intended to imply the potentially predictive ability of the feature (Duch, 2006). Moreover, there is another widely accepted point that an useful and parsimonious set of features is considered individually relevant and should not be redundant with respect to each other, which means that the selected features should not be highly correlated (Brown et al., 2012). This heuristic has been adopted numerous times. Battiti (1994)