Computers and Electronics in Agriculture 147 (2018) 158–165

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Soil microbial dynamics prediction using machine learning regression methods

Sunil Kr. Jha[a,b,*], Zulfiqar Ahmad[c,d]

[a] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[b] Faculty of Applied Sciences, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
[c] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, Hubei 430072, China
[d] Department of Environmental Sciences, University of California, Riverside, CA 92521, USA

ARTICLE INFO

Keywords:
ANN
SVR
SC-FIS
WM-FIS
Soil microbial dynamics prediction

ABSTRACT

Soil microbial dynamics is significant for the soil productivity. The present study explores the application of machine learning based regression methods in the prediction of selected soil microbial dynamics, including bacterial population (BP), phosphate solubilization (PS), and enzyme activities. An experiment was designed in a salt medium with rock phosphate inoculated with the solubilizing microorganism to measure the PS, BP, and 1-Aminocyclopropane-1-carboxylate (ACC) deaminase activity at a different temperature, pH, and incubation period. The artificial neural network (ANN), support vector regression (SVR), Wang and Mendel's (WM) - fuzzy inference systems (FIS), and subtractive clustering (SC)-FIS methods have been applied in the estimation of PS, BP, and ACC deaminase activity using the experimental conditions. The performance of four regression methods has been evaluated in the terms of the coefficient of determination (R^2), root mean square error (RMSE), and correlation coefficient (ρ). The SC-FIS method has better performance than the rest three methods in the prediction of each of the soil microbial dynamics (R^2 of 0.99 in the prediction of PS).

1. Introduction

The soil is a vital system of life. It contains water, nutrients, microorganism, and inorganic and organic materials, etc. and provides the living ecosystem of all organisms and plants. The condition of soil changes over time due to continuous degradation and variations in its composition (Doran et al., 1999; Karlen et al., 2003; Ghosh et al., 2017). A proper balance of physical, chemical and biological constituents is essential for maintaining the productivity of soil for agriculture (Schoenholtz et al., 2000; Paz et al., 2016). The estimation of the constituents of soil is necessary for monitoring of strength of soil. Microorganisms are a major part of biological constituents of soil and have a greater contribution to the strength of soil than physical and chemical constituents (Winding et al., 2005; Hermans et al., 2017). Microorganisms exhibit high activity towards any change in soil composition and get used to environmental conditions (Lange et al., 2015; Žifčáková et al., 2016). This adaptation hypothetically results in variations in microbial populations and their activities. Consequently, it is an indication of changes in soil condition that can be used in discrimination in soil fitness (Van Bruggen and Semenov, 2000; Doran and Zeiss, 2000; Hermans et al., 2017). Microorganisms have multifunctional support to the soil ecosystem by increasing the crop production by controlling the weed, insects, and other infections, increasing water and nutrients holding capacity, reutilizing nutrients, and nitrogen-fixing, etc. (Finlay, 2004; Barrios, 2007; Adnan et al., 2017).

The accurate evaluation of soil microbial conditions and fitness of soil is one of the significant research objectives at the present (Doran et al., 1999; Liu et al., 2006; Armenise et al., 2013). The phosphate solubilization (PS), bacterial population (BP), and 1-Aminocyclopropane-1-carboxylate (ACC) deaminase activity are substantial soil microbial condition indicators. Consequently, their estimation is significant in soil fitness monitoring. The prediction of PS, BP, and ACC deaminase activity using mathematical models results in lower estimation accuracy. The artificial intelligence (AI) methods have advanced computing capability and better efficiency, therefore widely adopted in modeling applications

Since last few years, some statistical and AI based methods have been implemented successfully in soil parameter modeling and classification (Coopersmith et al., 2014; Shiri et al., 2017a, 2017b; Sirsat et al., 2017; Wu et al., 2018) and modeling of soil microbial dynamics (Hughes et al., 2001; Haider et al., 2008; Tajik et al., 2012; Mukhlisin et al., 2012; Taghavifar and Mardani, 2014; Ebrahimi et al., 2017). K-

* Corresponding author.
E-mail address: drsuniljha@tdt.edu.vn (S.K. Jha).

https://doi.org/10.1016/j.compag.2018.02.024
Received 23 October 2017; Received in revised form 18 February 2018; Accepted 19 February 2018
0168-1699/

nearest-neighbor and boosted perceptron algorithm (Coopersmith et al., 2014), gene expression programming (GEP), and artificial neural network (ANN) (Shiri et al., 2017a, 2017b; Wu et al., 2018; Haider et al., 2008; Tajik et al., 2012; Mukhlisin et al., 2012; Taghavifar and Mardani, 2014; Ebrahimi et al. 2017), support vector machine (SVM) (Shiri et al., 2017a, 2017b; Sirsat et al., 2017; Wu et al., 2018), neuro-fuzzy (Shiri et al., 2017a, 2017b), random forest (RF) (Sirsat et al., 2017; Shiri et al., 2017a, 2017b), regression analysis (Tajik et al., 2012; Ebrahimi et al. 2017) are the commonly used methods in soil parameter modeling. Amongst earlier stated methods, ANN and fuzzy methods result in better accuracy in the modeling of nonlinear problems, like modeling soil cation exchange (Shiri et al., 2017a, 2017b) and water capacity (Shiri et al., 2017a, 2017b), and estimation of biosurfactant production (Ahmad et al., 2016). Both methods have some pros and cons like for ANN method, it is difficult to understand the rules of the model due to weights and error adjustment in each iteration but easy in learning from the data. Fuzzy logic methods use *if-then* rules and it is easy to interpret them. Meanwhile, the combination of ANN and fuzzy method based fuzzy inference system (FIS) results in better decision rules in modeling of a system (Shiri et al., 2017a, 2017b). FIS methods are more efficient in several research applications. Specifically, two categories of FIS methods have been implemented in modeling, including the Wang and Mendel's (WM) rule-based FIS, and subtractive clustering (SC)-FIS (Eftekhari and Katebi, 2008; Yang et al., 2010; Lohani et al., 2014).

Though we have not noticed the application of FIS methods in the prediction of soil microbial dynamics. With this motivation, we have implemented the WM-FIS and SC-FIS methods in the prediction of soil microbial dynamics. Also, the performance of FIS methods has been compared with the ANN and support vector regression (SVR) methods in terms of root mean square error (RMSE), correlation coefficient in between the actual and predicted values of the soil microbial dynamics and the coefficient of determination (R^2). The FIS methods create a linguistic regression relation in between soil microbial dynamics and experimental condition and results in better performance than the ANN, and SVR methods.

2. Analysis methods

The regression analysis method is used to establish a relationship between the dependent (soil microbial dynamics) and independent variables (experimental conditions). Regression methods are categorized into linear and nonlinear methods, including linear regression method, logistic regression method, polynomial regression method, etc. (Kleinbaum et al., 2013). In the present study, four AI based non-linear regression methods, including ANN, SVR, WM-FIS, and SC-FIS have been implemented. The non-linear regression models have been formulated for PS, BP, and ACC deaminase activity, independently using the earlier mentioned methods and their performance is compared in order to obtain optimal prediction accuracy. A short description and implementation details of four regression methods are as follows.

2.1. Artificial neural network (ANN)

ANN is an organization of artificial neurons in different layers (input, hidden, and output) (Theodoridis and Koutroumbas, 2006). The basic parameters of artificial neurons, including weight, threshold, and activation functions were optimized during the training of ANN method when used for regression analysis. Back-propagation feed forward artificial neural network (BPNN) is selected in the prediction of soil microbial dynamics and experimental conditions. The Levenberg-Marquardt algorithm is used in training of BPNN method. A $3 \times 10 \times 1$ structure of BPNN method with sigmoid activation function in the hidden layer is implemented in R (R Core Team, 2017) using the *nnet* package (Venables and Ripley, 2002) in the prediction of PS, BP, and ACC deaminase activity using the experimental parameters, incubation period, temperature and pH as inputs. The minimum error is achieved at 29, 23 and 20 iterations for PS, BP, and ACC deaminase activity, respectively. The details of the method can be seen in Ref. (Theodoridis and Koutroumbas, 2006; Venables and Ripley, 2002).

2.2. Support vector regression (SVR)

Support vector machine (SVM) is termed as support vector regression (SVR) when used for the regression analysis (Theodoridis and Koutroumbas, 2006; Bishop, 2006). In the SVR method the actual values of PS, BP, and ACC deaminase activity are assumed as output and experimental condition, incubation period, temperature and pH as inputs. The kernel function (linear, Gaussian, sigmoid, radial basis, and polynomial, etc.) computes the inner product of input feature vectors in high dimensional space and reduces the time and space complexity of the SVR method. The Radial basis kernel function is used in SVR method in the prediction of PS, BP, and ACC deaminase activity independently using the *e1071* package in R (Meyer et al., 2017). The kernel function has similar values of parameters such as $\cos t = 1$, $\gamma = 0.33$ and $\xi = 0.1$ for SVR models in the prediction of PS, BP, and ACC deaminase activity. Though, the number of support vectors varies (34, 19, and 29 for three SVR models, respectively).

2.3. Fuzzy inference system (FIS)

FIS creates a map in between the dependent and independent variables using the fuzzy logic. In which, initially, a partial membership is assigned to each input, thereafter complete membership is formulated by composing the partial truths using certain *if-then* rules (Klir and Yuan, 1995). Several clustering methods (subtractive, mountain, c-means, etc.) have been used in the implementation of FIS. Amongst them, subtractive clustering based FIS and Wang and Mendel's rule-based general FIS have been implemented in the prediction of soil microbial dynamics. A short description of both methods is as follows.

2.3.1. Wang and Mendel's fuzzy inference system (WM-FIS)

WM-FIS is the most basic type of fuzzy inference system. The method is described in (Wang and Mendel, 1992). WM-FIS is implemented using the *frbs* package in R (Riza et al., 2015). The method divides the independent and dependent variable spaces into several fuzzy regions and assign a partial membership, generates *if-then* rules and assign a degree to each rule, then creates a joint rule for final mapping of a test input. The basic steps and implementation details of the method are shown in Fig. 1(a). The details of the method can be seen in Ref. (Wang and Mendel, 1992).

2.3.2. Subtractive clustering fuzzy inference system (SC-FIS)

In SC-FIS, initially, SC finds fuzzy clusters according to a Gaussian measure. Data points close to a point of maximum potential (center) are assigned to the first cluster. The rest of the data points are assigned to the second cluster and the process is repeated till all data points are clustered. After that fuzzy *if-then* rules are generated using previous clusters and the dependent variable is predicted using rules (Chiu, 1994). A schematic representation of basic steps and implementation details of SC-FIS method is shown in Fig. 1(b). The method is implemented using the *frbs* package in R (Riza et al., 2015).

3. Materials and methods

3.1. Sampling of soil

The wheat plants were uprooted at tillering stage and stored in polythene bags. The non-rhizospheric soil was detached by agitating the roots and thereafter the soil strictly adhering to the roots was collected as a rhizospheric soil sample. A composite sample was prepared by using seven rhizospheric samples of wheat plants. For the simplicity,