Computers and Electronics in Agriculture 135 (2017) 242–251

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Modeling soil cation exchange capacity using soil parameters: Assessing the heuristic models

Jalal Shiri [a,*], Ali Keshavarzi [b], Ozgur Kisi [c], Ursula Iturraran-Viveros [d], Ali Bagherzadeh [e], Rouhollah Mousavi [b], Sepideh Karimi [a]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran
[c] Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia
[d] Departamento de Matematicas, Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Circuito Escolar, Cd. Universitaria, Coyoacán 04510, Ciudad de México, Mexico
[e] Department of Agriculture, Islamic Azad University, Mashhad Branch, Emamyeh Boulevard, P.O. Box: 91735-413, Mashhad, Iran

A R T I C L E I N F O

Article history:
Received 22 November 2016
Received in revised form 10 January 2017
Accepted 16 February 2017
Available online 24 February 2017

Keywords:
Cation exchange capacity
Heuristic models
k-fold testing

A B S T R A C T

Accurate knowledge about soil cation exchange capacity (CEC) is very important in land drainage and reclamation, groundwater pollution studies and modeling chemical characteristics of the agricultural lands. The present study aims at developing heuristic models, e.g. gene expression programming (GEP), neuro-fuzzy (NF), neural network (NN), and support vector machine (SVM) for modeling soil CEC using soil parameters. Soil characteristic data including soil physical parameters (e.g. silt, clay and sand content), organic carbon, and pH from two different sites in Iran were utilized to feed the applied heuristic models. The models were assessed through a k-fold test data set scanning procedures, so a complete scan of the possible train and test patterns was carried out at each site. Comparison of the models showed that the NF outperforms the other applied models in both studied sites. The obtained results revealed that the performance of the applied models fluctuated throughout the test stages and between two sites, so a reliable assessment of the model should consider a complete scan of the utilized data set, which will be a good option for preventing partially valid conclusions obtained from assessing the models based on a simple data set assignment.

1. Introduction

The soil cation exchange capacity (CEC) is the total exchangeable cations which may be hold in soil by electrostatic forces at a specific pH level (Bauer and Velde, 2014). The knowledge about CEC values is important in land drainage and reclamation as well as groundwater pollution studies (van Hoorn and van Alphen, 1994). It is one of the most important chemical characteristics of agricultural lands (Ghaemi et al., 2013). CEC influences the stability of soil structure, nutrient availability, soil pH and the soil's reaction to fertilizers and other ameliorants, as well as it and provides a buffer against soil acidification (Hazelton and Murphy, 2007). It is also used as a measure of soil fertility, nutrient retention capacity, and the capacity to protect groundwater from cation contamination (Robertson et al., 1999). Usually, heavy clay soils present higher magnitudes of CEC, expressing the higher availability of nutrients in these soils.

CEC is usually measured on the fine earth fraction (soil particles lower than 2 mm in size). In gravelly soils, the effective soil CEC as a whole is diluted, and if only the clay fraction is analyzed, the obtained CEC values will be higher than the actual field values. Measuring CEC includes washing the soil for removing excess salts and using an 'index ion' for determining the total positive charge in relation to original soil mass. This includes bringing the soil to a predetermined pH level before analysis. Further details about CEC measurement techniques might be found in e.g. Rengasamy and Churchman (1999) and Rayment and Higginson (1992). However, these methods are time consuming, laborious and expensive, especially in remote areas, e.g. Aridisols in Iran. Alternatively, heuristic data driven models [e.g. gene expression programming (GEP), neuro-fuzzy (NF) technique, neural networks (NN) and support vector machine (SVM)] which can relate the CEC to its influential parameters might be applied for simulating CEC. Were et al. (2015) compared different heuristic models for predicting soil organic carbon stocks across an Afromontane landscape and found the SVM as the superior model in this issue. Keshavarzi et al. (2015) applied neural-network for defining pedotransfer functions

* Corresponding author.
E-mail addresses: j_shiri2005@yahoo.com, jalalshiri@tabrizu.ac.ir (J. Shiri).

http://dx.doi.org/10.1016/j.compag.2017.02.016
0168-1699/

in estimating soil phosphorous. Keshavarzi et al. (2017) developed ANFIS-based subtractive clustering algorithm in estimating soil CEC through using soil and remotely sensed data in a semi-arid region of Iran. Emamgolizadeh et al. (2016) compared different heuristic models for predicting CEC and found that the multivariate adaptive regression splines (MARS) and GEP models are superior in this issue. Zolfaghari et al. (2016) applied k-nearest neighbor technique for predicting soil cation exchange capacity. All the reported literature have used single data set assignment for developing and testing the applied models, where the models are trained by using a portion of the whole data and tested using the remain data patterns. The present study, however, aims at assessing heuristic data driven approaches, namely GEP, NF, NN and SVM, in modeling soil CEC through k-fold testing. The necessary input variables of the models were identified by utilizing Gamma-test. This is the first attempt that compares the GEP, NF, NN and SVM methods accuracy in modeling soil CEC.

2. Materials and methods

2.1. Gene expression programming (GEP)

In contrast to common applications of classical regression models to estimate CEC indirectly based on other data (Bishop and McBratney, 2001; Park and Vlek, 2002; Triantafilis et al., 2011), GP (genetic programming) has not been exploited for this purpose, although it has shown much potential in similar applications (Johari et al., 2006; Makkeasorn et al., 2006; Parasuraman et al., 2007; Padarian et al., 2012).

GP-based models (Koza, 1992), utilize a "parse tree" structure for the search of their solutions. GP has the ability for creating an explicit formulations set that govern the studied phenomenon, to map the relationship(s) between the input-target parameters using different operators. Gene expression programming (GEP) is similar to GP, in a way that selects the best governing formulations based on fitness values and introduces genetic variation using a unique or various genetic operators (Ferreira, 2006). One of the advantages of GP (i.e. GEP) over other heuristic techniques (e.g. NF, NN and SVM) is in giving explicit expression of the input-target relationship. Gene Xpro program was used in the present study for GEP-based modeling. Different fitness functions and function sets were tried in the applied models and the best ones were selected. Details for model development will be given in the next sections. Further details about modeling process with GEP can be read in e.g. Ferreira (2006).

2.2. Neuro-fuzzy systems (NF)

Neuro-fuzzy technique (NF) is a combination of adaptive artificial neural network and fuzzy inference systems, where the parameters of the fuzzy system are computed by the neural networks training algorithms. NF calculates a set of parameters via a hybrid learning rule composed of back-propagation gradient descent error (BPGDE) and a least squared error (LSE). The Sugeno's fuzzy approach (Takagi and Sugeno, 1985) was used here to relate the target variable (*CEC*) to input variables. Different membership functions were evaluated here to find the optimal one. The hybrid optimization method (the combination of LSE and BPGDE) was used for obtaining the membership functions parameters to emulate the training data. For a given input-output matrix, various fuzzy-Sugeno models can be employed using different identification methods (i.e. grid partitioning and subtractive clustering, etc.). The commonly used grid partitioning (GP) identification method was utilized here for modeling soil CEC. The GP method proposes independent partitions of each antecedent variable by defining the membership functions (MFs) of all antecedent variables. Fuzzy MFs might take different forms and the optimal numbers of MFs is computed by trial and error. In selecting the number of MFs, large numbers of MFs or parameters should be avoided to save time and computational costs (Kisi and Shiri, 2012). For this reason, 2 or 3 numbers of MFs were used in the applied ANFIS models. Details for NF model structures used in the applications will be provided in the next sections.

2.3. Neural networks (NNs)

Neural networks are parallel information-processing systems which have been originally designed for the modeling of the performance of a biological neural system. The most common architecture of NNs is composed of the input, hidden, and output layers, which is called multilayer perceptron (MLP) (Fausset, 1994). Here, three-layer feed-forward networks were utilized with different transfer functions in the hidden and output layers. The hidden-layer-node numbers of each model were determined iteratively. At each training process, 100 networks were evaluated and the optimum architecture for each case (transfer functions) was selected. Also minimum and maximum values of 0.0001 and 0.001 were found to be optimum values of weight decay in hidden layer.

With modeling CEC through NN technique, the input and output values were normalized using the following equation:

$$h_{ni} = a\frac{CEC_i - CEC_{\min}}{CEC_{\max} - CEC_{\min}} + b \qquad (1)$$

where CEC_{ni} is the normalized data at time i, CEC_{min} and CEC_{max} denote the minimum and maximum of the data set and CEC_i stands for the observed *CEC* value at time *i*. Different values can be assigned for the scaling factors *a* and *b*. The *a* and *b* were taken as 0.8 and 0.2 herein, respectively according to Cigizoglu (2003) and Kisi et al. (2013). Thus, the training data were scaled in the range [0.2, 0.8].

Detailed descriptions of NN techniques can be read in e.g. Bishop (1995) or Haykin (1999).

2.4. Support vector machine (SVM)

SVMs are regression procedures, with a structural risk minimization (SRM) principle formulation, which is superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks. Traditional ERM minimizes the error on the training data, while SRM minimizes an upper bound on the expected risk, providing SVM a greater ability to generalize, which is the goal in statistical learning (Vapnik et al., 1997; Gunn, 1998). Further details on the application of SVM can be found e.g. in Vapnik et al. (1997).

2.5. Gamma test for input selection

The Gamma test is a non-linear analysis and modeling method which allows examining the nature of a hypothetical input/output relationship in a numerical data-set. The Gamma statistic Γ is calculated utilizing the Gamma test and least Γ indicates the best input combination. First reported in Stefánsson et al. (1997) with the conjecture that a very simple method (the Gamma test) could be utilized to directly estimate from a given input/output data set the extent to which the data identified from an underlying smooth model, even though the model was unknown.

The set of input vectors in this study are values of Clay, Silt, Sand, OC and pH. The corresponding output is the soil *CEC*. We assume that the input vectors contain factors which are useful for influencing the output *CEC*. A second assumption is that the