Optimal feature selection for predicting soil CEC: Comparing the hybrid of ant colony organization algorithm and adaptive network-based fuzzy system with multiple linear regression

Hosein Shekofteh [a,*], Fatemeh Ramazani [b], Hossein Shirani [c]

[a] Department of Soil Science, College of Agriculture, Vali-e-Asr University of Rafsanjan, P.O. Box 518, Rafsanjan, Iran
[b] Soil Science Department, Vali-e-Asr Rafsanjan University, Rafsanjan, Iran
[c] Department of Soil Science, College of Agriculture, Vali-e-Asr University of Rafsanjan, P.O. Box 518, Rafsanjan, Iran

ABSTRACT

Soil CEC is a very important property that represents soil fertility status. Though difficult to measure, it can be predicted by soil physicochemical properties that can be easily measured. Researchers have used different input soil properties to derive pedo-transfer functions (PTFs) and predict soil CEC. To select properties that influence soil CEC, we have introduced a hybrid algorithm: an advance ant colony organization (ACO) in combination with an adaptive network-based fuzzy inference system (ANFIS). The potential power of the advance ACO-ANFIS algorithm in setting up a framework for identifying the most determinant parameters of agricultural soils CEC in an Iranian semiarid region (Rabor region, 29° 27′ N to 38° 54′ N and 56° 45′ E to 57° 16′ E) was also investigated. To make sure that ACO-AN FIS algorithm reaches its global minimum, features were selected by ANFIS. A multiple linear regression (MLR) model was constructed as benchmark for the comparison of performances. The results from ACO-ANFIS and ANFIS for feature selection and their RMSEs were the same. Results of ACO-ANFIS and ANFIS for selecting best dataset showed that five properties including soil organic matter (SOM), clay, silt, pH and bulk density (BD) had the lowest error. The ANFIS method resulted in higher model efficiency and coefficient of determination ($R^2 = 0.91$) than MLR approach ($R^2 = 0.74$). This study provides a strong basis for predicting soil CEC and identifying the most determinant parameters influencing soil CEC in the Iranian semiarid regions; however, its general analytical framework could be applied to other parts of the world with similar challenges.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The soil cation exchange capacity (CEC) is defined as the number of adsorbed cation charge moles that are desorbed from a unit mass of soil under such specific conditions as temperature, pressure, and soil solution composition (Sposito, 2008). It is commonly referred to as the quantity of negative charges in soil. The negative charge may be pH-dependent (soil organic matter) or permanent (some clay minerals) (Evans, 1989). CEC is a good indicator of soil fertility, crop growth and pollutant transport which determines a soil buffering capacity for holding cationic nutrients and organic pollutants (Arias et al., 2005; Tang et al., 2009; Visconti et al., 2012); it is therefore an important parameter for predicting crop yield. Precise knowledge of CEC data helps determine the accuracy of crop yield simulation. Overall, a good understanding of soil CEC is necessary for crop, soil and environmental research studies. Direct and accurate measurement of soil CEC is expensive and

time consuming, especially for Iranian soils with high calcium carbonate contents. Thus, it is worth to apply indirect methods for an accurate prediction of soil CEC, which is possible through measuring basic soil properties (Krogh et al., 2000; Seybold et al., 2005). Researchers have used different soil properties to predict soil CEC. To do this, soil organic matter (SOM) and clay content (Helling et al., 1964), as well as pH, SOM and clay content (Seybold et al., 2005) were used as input variables for creating PTFs. Properties like soil structure, water content at permanent wilting point, hydraulic conductivity and soil horizons have also been applied (Hartmann et al., 1998; Madeira et al., 2003; Tang et al., 2009). Numerous studies have been conducted on prediction of soil CEC in Iran using soft computing techniques. For example, Kashi et al. (2014) employed artificial neural networks (multilayer perceptron: MLP and Radial basis function: RBF), ANFIS and multiple regression (MLR) techniques. The input data for models included electrical conductivity (EC), soil texture, lime percentage, sodium adsorption ratio (SAR) and bulk density. Ghorbani et al. (2015) estimated soil cation exchange capacity through MLR, RBF and ANFIS models and compared them to MLR. The input data consisted of clay, silt, sand, soil organic carbon

and pH. Emamgolizadeh et al. (2015) applied genetic expression programming (GEP) and multivariate adaptive regression splines (MARS) and developed functional relations to estimate CEC from more readily measurable soil physical and chemical variables (e.g. OM, clay, and pH). Zolfaghari et al. (2016) took the nonparametric k-nearest neighbor approach, using two data sets to construct a PTF and predict soil CEC. The first data set consisted of clay, silt, sand and organic carbon (OC) contents. The second data set included OC and clay contents, found in both k-NN and ANN models too; the larger the number of input variables, the more correct the estimation of CEC was. However, this improvement was not statistically significant at the 0.05 level.

As indicated by a variety of literature, various inputs have been used for estimation and prediction of soil CEC. Input parameters are so different that there have been no methods so far for selecting input properties that affect soil CEC. Given the progress made in feature selection algorithm in other areas, such algorithms can be applied to soil CEC. The best approach for selecting properties is the feature selection (FS) approach. It is generally used in machine learning, especially when the learning task involves datasets with further dimensions. It is mainly aimed at selecting a subset of features or a set of properties, by removing those with little or no predictive information as well as redundant features that are strongly correlated (Vieira et al., 2010). In general, the availability of large amounts of input data represents a challenge to regression and classification analysis methods. For example, the use of numerous input properties for deriving a PTF for prediction of soil CEC may require estimation of a considerable number of equation parameters during regression process and therefore more data to be measured. Ideally, for each property used in the regression process an independent set of information should be added. However, the properties may be highly correlated and this can suggest a degree of redundancy of the available information on the input variables which may have a negative impact on the accuracy of regression (Pal and Foody, 2010). Recently, nature-inspired meta-heuristic algorithms have been used for feature selection in soil science, e.g. particle swarm optimization (PSO) for selecting influential properties in soil quality indices (Shirani et al., 2015). One of the nature inspired meta-heuristic algorithms is ant colony optimization (ACO), an extensively-used promising approach for solving combinational optimization problems and feature selection (Aghdam et al., 2009; Ahmed, 2005). So far, the ACO algorithm has been applied to soil and water sciences (Abbaspour et al., 2001; Jalali et al., 2006; Pour

and Zeynali, 2015) but not to feature selection in soil science. In this study, a hybrid algorithm advance ACO-ANFIS was used for selecting the best set of properties affecting soil CEC. The aims of this study were: 1. to assess advance ACO combined with ANFIS to select the best set of input properties influencing soil CEC, 2. to run the ANFIS method for all of the input properties with all states and calculate their error, 3. to compare the results of ACO-ANFIS algorithm with those of ANFIS in feature selection, 4. to do modeling after selecting the properties that influence soil CEC by ANFIS and multiple linear regression (MLR) approaches, and 5. to compare ANFIS and MLR results.

2. Materials and methods

2.1. Study area

The data required for selecting the set of properties affecting soil CEC were collected from some parts of Rabor region (29° 27′ N to 38° 54′ N and 56° 45′ E to 57° 16′ E) located in the south-west part of Kerman Province, south east Iran (Fig. 1). Rabor is a typical semiarid land farming area with a cold temperate climate. The annual mean temperature is 15 °C with an average annual precipitation of 250 mm. Based on Soil Survey Staff (2014), the study soils mainly belong to Typic Calcixerepts subgroup. Different parent materials including limestone, dolomite, and conglomerate can be seen in this area. The main textural classes are loam and sandy loam.

2.2. Soil sampling and measurement

A total of 104 soil samples were collected from the soil surface (0–15 cm depth) of four land uses which included gardens with 20 year-old walnut trees, pasture, agriculture and a mountain almond forest. A grid sampling strategy was designed by ILWIS 3.4 software (ITC, University of Twente, Netherlands) to properly select soil sampling locations and consider spatial variations of parameters influencing the soil CEC in the area. At each point, disturbed and undisturbed samples were taken. For the former, large plant materials (i.e., roots and shoots) and pebbles in each sample were separated by hand and discarded. The positions of the sampling points were identified in the field using GPS (model 76 CSx, Gramin Co., Taiwan). These samples were air-dried and ground to pass a 2 mm sieve. Soil organic matter (SOM) content
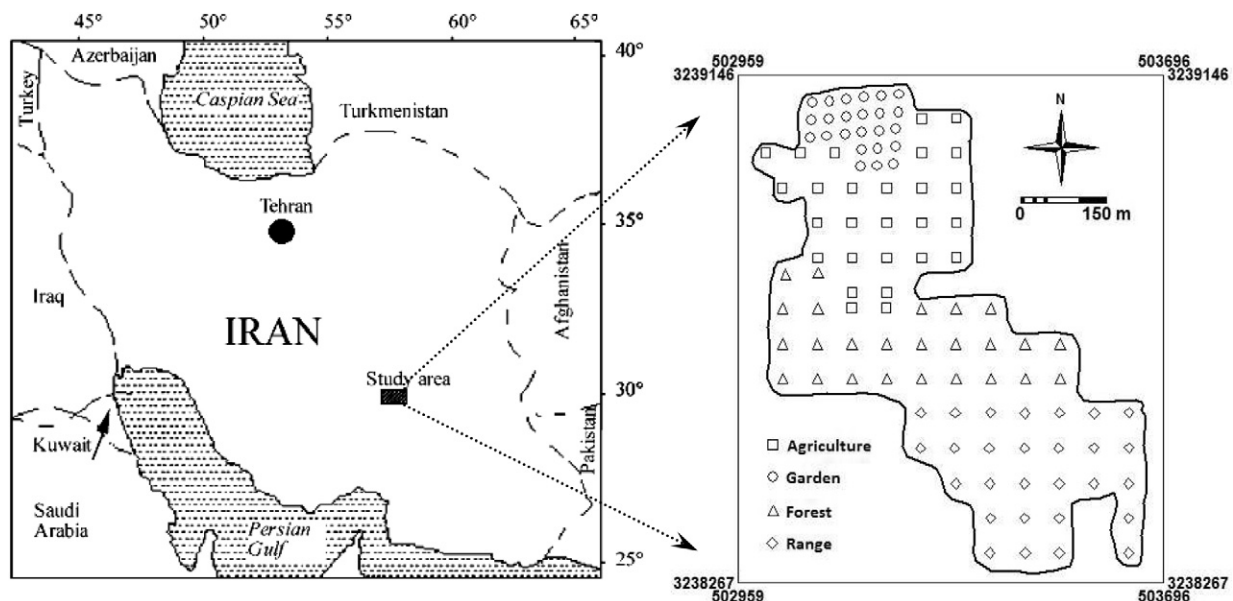

Fig. 1. Location of the study area along with sampling points in different land uses.