

Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction

**A. A. Jafarzadeh, M. Pal, M. Servati,
M. H. FazeliFard & M. A. Ghorbani**

**International Journal of
Environmental Science and
Technology**

ISSN 1735-1472

Int. J. Environ. Sci. Technol.
DOI 10.1007/s13762-015-0856-4



Your article is protected by copyright and all rights are held exclusively by Islamic Azad University (IAU). This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

Comparative analysis of support vector machine and artificial neural network models for soil cation exchange capacity prediction

A. A. Jafarzadeh¹ · M. Pal² · M. Servati³ · M. H. FazeliFard⁴ · M. A. Ghorbani⁴

Received: 13 December 2014 / Revised: 27 April 2015 / Accepted: 6 July 2015
© Islamic Azad University (IAU) 2015

Abstract The aim of this study was to compare the performance of support vector machine and artificial neural network techniques to predict the soil cation exchange capacity of an agricultural research station in terms of soil characteristics (clay, silt, sand, gypsum, organic matter). The data consist of 380 soil samples collected from different horizons of 80 soil profiles located in the Khoja (Khajeh) region of Azerbaijani provinces, Iran. The support vector machine and artificial neural network models predict the cation exchange capacity from the above soil characteristics of the samples. The models' results are compared using three criteria, i.e., root-mean-square errors, Nash–Sutcliffe and the correlation coefficient. A comparison of support vector machine results with artificial neural network method indicates that artificial neural network is better than the support vector machine method in prediction of the cation exchange capacity.

Keywords Clay · Khajeh · Modeling · Pedo-transfer function · Sand

Introduction

Soil cation exchange capacity or CEC is a very important characteristic, essential for measuring fertility and nutrient retention capacity, and is a commonly applied indicator of soil condition or vulnerability. Soil management of a region has invoked this research to develop a predictive model using measurements of a sample of CEC values of a region. The predictive models are used to study the CEC behavior in terms of contributory factors of soil composition of clay, silt, sand, gypsum and organic matter. These soil constituents carry intrinsically negative charges on their surface and therefore adsorb exchangeable cations (Tang et al. 2009).

The two properties that most account for the reactivity of soils are surface area and surface charge. Surface area is a direct result of particle size and shape. Most of the total surface area of mineral soil is due to clay-sized particle and soil organic matter. Charge development in soils is intimately associated with these same two fractions, although the sand and silt size fractions may contribute some CEC if coarse-grained vermiculite is present. Charge development in soils occurs as a result of both isomorphic substitution and ionization of functional groups on the surface of solids that make up the soil matrix. These two mechanisms give rise to the permanent and the pH-dependent charges of soils, respectively (Yola et al. 2014). Plant roots are capable of cation exchanges and root zones are where this capability is exploited. The quantity of negative charges in soil, known as CEC, is a measurable quantity, but their direct measurements are expensive, difficult and time-consuming (Evans 1989). Therefore, this paper aims to gain an insight into modeling the CEC in terms of soil composition of clay, silt, sand, gypsum and organic matter. The significance of the CEC of the soil particles stems from

✉ M. A. Ghorbani
ghorbani@tabrizu.ac.ir

¹ Department of Soil Science, University of Tabriz, Tabriz, Iran

² Department of Civil Engineering, NIT, Kurukshetra 136119, India

³ Shahid Bakeri Higher Education Center of Miandoab, Urmia University, Urmia, Iran

⁴ Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran



the 'ease' of exchange of cations with one another to the extent that they are readily available for plants. The CEC is among the most important soil properties required in soil databases (Manrique et al. 1991) and serves as an input to soil and environmental models (Keller et al. 2001). Different study topics on the CEC include: fluoride sorption and desorption on soils (Gago et al. 2014), monitoring general variability of soil attributes to different land use types in calcareous soils (Rezapour 2014), influence of pasture degradation on soil quality indicators that included physical, chemical, biological and micromorphological attributes (Ayoubi et al. 2014), isolation of different soil organic carbon fractions and coal C in a reclaimed minesoil chronosequence (Chaudhuri et al. 2013), determination of adsorption efficiency based on cation exchange capacity (Gatima et al. 2006), study on adsorption of aqueous phenol solution in soil (Subramanyam and Das 2009), impact of sewage and mining activities on distribution of heavy metals in the water–soil–vegetation system (Semhi et al. 2013), potential impact of fluorine-rich fertilizers on the aquifer system (Marimon et al. 2013), prediction of subsurface heterogeneity of contaminated soil management (Moon et al. 2013), identification of heavy metal sources in agricultural soil (Huang et al. 2013), effects of cations and anions on iron and manganese sorption and desorption capacity in calcareous soils (Moharami and Jalali 2013), effect of the addition of granitic powder to an acidic soil (Silva et al. 2013), effects of agricultural practice and land use on the distribution and origin of some potentially toxic metals in the soils (Moghaddas et al. 2013). In the laboratory procedure, CEC of the soils is measured by using the ammonium acetate (NH_4OAc) method through the replacement of sodium (Na^+) ions with ammonium (NH_4^{++}) ions, but this is difficult, time-consuming and expensive (Carpena et al. 1972). An alternative CEC estimation approach is through the more easily measurable and readily available soil properties such as particle size distribution (clay, sand and silt content), gypsum and organic matter, which are referred to as pedo-transfer functions (PTFs) as coined in soil science (Bouma 1989). In recent years, data mining techniques such as artificial neural networks (ANN) and support vector machines (SVM) have been widely used in the modeling of various complex environmental problems (Yilmaz and Kaynar 2011). A neural network is an adaptable system that learns relationships from the input and output datasets and is able to predict a previously unseen dataset of similar characteristics to the input set (Haykin 1999; ASCE 2000). Previous PTF studies for formulating CEC models using ANNs include (Van Bladel et al. 1975; Baker and Ellison 2008; Tang et al. 2009; Minasny and McBratney 2002; Minasny et al. 1999;

Schaap et al. 1998; Silveira et al. 2013). Gruszczyński (2009) applied multiple regression, polynomial neural network and fuzzy neural network (ANFIS) to prediction of CEC in southern Poland. The results showed that ANFIS is better than the multiple regression and polynomial neural network models. Tang et al. (2009) investigated four radial basis function neural networks (RBFN) into the estimation of CEC from 457 soil physicochemical properties. Soil horizon; pH; organic carbon content; and clay, silt and sand contents were taken as the input variables. The results show high correlation coefficients between predicted and measured CEC values.

As an alternative to ANN, the SVM proposed by Vapnik (1998) is a powerful tool for nonlinear classification, regression and time series prediction (Wang et al. 2008). SVM belong to kernel-based learning approaches and have gained wide popularity. SVM are a kind of supervised machine learning system that use a linear high-dimensional hypothesis space called feature space. The basic idea of working principle of SVM is provided by the use of kernel functions that implicitly map the data to a higher-dimensional space. This makes SVM a powerful tool for modeling the nonlinear complex environmental problems (Bhagwat and Maity 2012). Several studies reported the use of SVM in forecasting the soil water (Wu et al. 2008), soil moisture prediction (Gill et al. 2006), estimation of soil hydraulic parameters (Twarakavi et al. 2009), modeling soil diffuse reflectance spectra (Rossel and Behrens 2010) and soil type classification (Kovačević et al. 2010). Liao et al. (2014) also used SVM, multiple stepwise regression (MSR) and ANN models to estimate the CEC from easily measured physicochemical properties (e.g., texture, soil organic matter, pH) based on 208 soil samples in Qingdao City, China. They indicated that the SVM model has better results than the ANN and MSR models to estimate CEC values. Sensitivity analysis was also conducted to explore the influence of each input parameter on the CEC predictions. The clay and sand content are the most and weakest parameters, respectively.

Study area

The area under study covers approximately 8000 ha (between the longitude: $46^{\circ}35'$ – $46^{\circ}40'E$ and latitude: $35^{\circ}08'$ – $35^{\circ}12'N$) in the Khoja region, Tabriz, Azerbaijan, Iran. The Khoja Soil and Water Conservation Research Station is located 30 km at the east of Tabriz and 60 km from Eher (also mentioned as Ahar). The prevailing semiarid climate in the region is cold continental with a mean annual rainfall of 220–270 mm. The altitude in the vicinity of the region is approx. 1550 m above the mean sea level. The region is

surrounded by mountains and hills through which the River Aji flows to Lake Urmu. The underlying rock formations are often granitic intrusion related to the Oligo-Miocene age, intruding through a sequence of calcareous rocks. The sampling procedure was started by carrying out a preliminary soil investigation and then designing a random sampling procedure, but flexibility is exercised during the setting out of the sampling locations to reflect the inherent variability of soil (e.g., the sampling points are set out capturing the locations with tree lines, variability due to soil texture or locations with notable land use changes). The underlying principle in selected sample location is that the results are representative of the soil, such that the fitted models are not a construct of the particular measurement but are capable of being a representative of the whole area by capturing its inherent variability. The study area is located on floodplain with well-established flooding patterns, and therefore, soil variability depends more on local weathering than regional variability or land use. There have been past attempts to assess land suitability of the site for agricultural purposes by using various methods (Malékian and Jafarzadeh 2011), which were used as preliminary basis before taking the necessary steps to gather data. On this basis, the decisions on designing the sampling procedure were made, and subsequently, sampling at regular grid points was found to be more appropriate for this study area, but the randomness was allowed within the number of samples at each soil horizon.

Materials and Methods

Artificial neural networks (ANNs)

ANNs, often traced back to McCulloch and Pitts (1943), are inspired by the working of the brain and nerve systems in biological organisms with a capability for self-learning and automatic abstracting and with a possible benefit of reducing modeling times. One application of ANNs is an alternative modeling strategy to traditional methods of data and time series analysis. Fundamental processing element

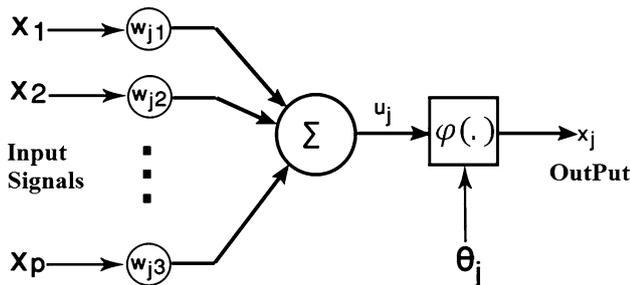


Fig. 1 Nonlinear model of a neuron (Haykin 1999)

of ANNs is a neuron. At the hidden layers, each neuron computes w_{ij} , a weighted sum of its p input signals, x_i , for $i = 0, 1, 2, \dots, n$ and then applies a nonlinear activation function to produce an output signal, u_j . The model of a neuron is shown in Fig. 1. A neuron j is described mathematically by the following pair of equations:

$$u_j = \sum_{i=1}^p w_{ij} \cdot x_i \tag{1}$$

and

$$x_j = \varphi(u_j - \theta_j) \tag{2}$$

where θ is a threshold function, and its use has the effect of applying an affine transformation to the output of the linear combiner in the model of Fig. 1 (Haykin 1999; Melesse and Hanley 2005), and in this study, the logistic sigmoid nonlinear function (Bilgili et al. 2007) is used for this purpose, expressed as:

$$\varphi_x = \frac{1}{1 + e^{-x}} \tag{3}$$

The type of ANN used in this study is a feed-forward multilayer perceptron (MLP) with back propagation (BP) learning algorithm, as commonly used in various complex environmental problems such as soil science applications of MLP. The structure of a three-layer MLP is shown in Fig. 2. MLP with back propagation (BP) is a popular form of training multilayer neural networks learning algorithm, and it is widely used in solving various classification and prediction problems. Back propagation convergence is slow, but it has the advantages of accuracy and adaptability (Kisi 2005).

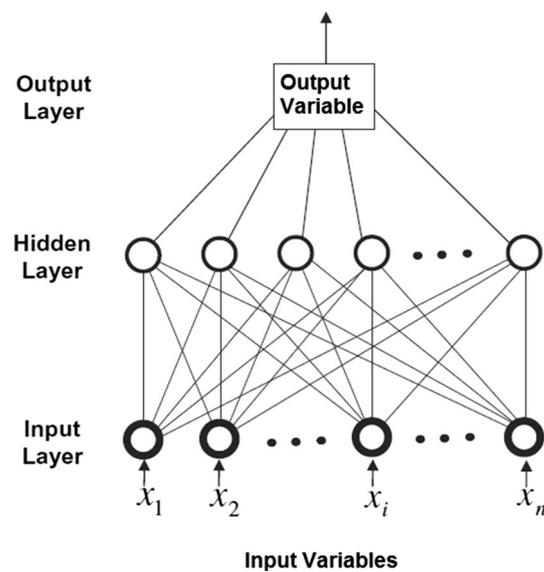


Fig. 2 Simple configuration of multilayer perceptron neural network

It consists of three layers: an input layer, a hidden layer and an output layer. A set of neurons or nodes are arranged in each layer. The number of neurons in the input and output layers is defined depending on the number of input and output variables of the system under investigation, respectively. However, the number of neurons in the hidden layer(s) is usually determined via a trial-and-error procedure. As seen from the figure, the neurons of each layer are connected to the neurons of the next layer by weights. The typical performance function used for training feed-forward neural networks is the mean sum of squares (MSE) of the network errors:

$$MSE = \frac{1}{N} \sum_{j=1}^N (T_j - O_j)^2 \tag{4}$$

where T_j is the target output, O_j is the actual output at output unit j , and N is the number of data patterns.

Support vector machines (SVM)

The SVM is based on statistical learning theory (Vapnik 1998) Because of good generalization performance, SVM is receiving increasing attention in pattern classification and nonlinear regression estimation (Cao and Tay Francis 2003). For a given training data with N number of samples, represented by $(x_1, y_1), \dots, (x_N, y_N)$, where x is an input vector and y is a corresponding output value, SVM estimator (f) on regression can be represented by:

$$f(x) = w \cdot \phi(x) + b \tag{5}$$

where w is a weight vector, b is a bias, ‘ \cdot ’ denotes the dot product and ϕ a nonlinear mapping function. A smaller value of w indicates the flatness of Eq. (5), which can be obtained using minimizing the Euclidean norm as defined by $\|w\|^2$. Vapnik (1995) introduced the following convex optimization problem with an ϵ -insensitive loss function for regression using SVM:

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} w^2 + C \sum_{k=1}^N (\xi_k^+ + \xi_k^-) \\ &\text{subject to} \quad \begin{cases} y_k - (w \cdot \phi(x_k) + b) \leq \epsilon + \xi_k^+ \\ -y_k + (w \cdot \phi(x_k) + b) \leq \epsilon + \xi_k^- \\ \xi_k^+, \xi_k^- \geq 0 \end{cases} \quad k = 1, 2, \dots, N \end{aligned} \tag{6}$$

where C is a positive trade-off parameter or capacity parameter that determines the degree of the empirical error in the optimization problem and determines the trade-off between the flatness of the function and the amount to which deviations larger than ϵ are tolerated. Also ξ_k^-, ξ_k^+ are slack variables representing upper and lower constraints on the output system over the error tolerance ϵ

(Misra et al. 2009). Lagrangian multipliers and the Karush–Kuhn–Tucker (KKT) condition were used to solve the optimization of Eq. (6) in a dual form. Support vectors are the input vectors that have nonzero Lagrangian multipliers under the KKT condition (Yoon et al. 2011). Figure 3 shows a schematic diagram of the SVM used in this study.

In natural processes, almost all the predictor variables (input space) are nonlinearly related to the predicted variable. This limits a linear formulation of the problem as shown in Eq. (6). This limitation is solved by mapping the input space on to some higher-dimensional space (feature space) using a kernel function. The kernel function enables us to implicitly work in a higher-dimensional feature space. Eq. (5) can then be modified by using Lagrangian multipliers (α_i and α_i^*) as following:

$$f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) \cdot K(x, x_i) + b \tag{7}$$

where $K(x, x_i)$ is the kernel function. Commonly used kernel functions include the linear, polynomial, radial basis and sigmoid kernel functions. In this study, the widely used radial basis function (RBF) kernel function was used. The RBF kernel function is defined by Eq. (8):

$$K(x, x_i) = \exp\left(-\frac{x - x_i^2}{\sigma^2}\right) \tag{8}$$

while using SVM with RBF kernel function, one has to optimize three parameters during training, which includes kernel parameter (σ), capacity parameter (C) and insensitive loss function (ϵ). In this study, an internal cross-validation (Wang and Hu 2005) during creation of SVM model

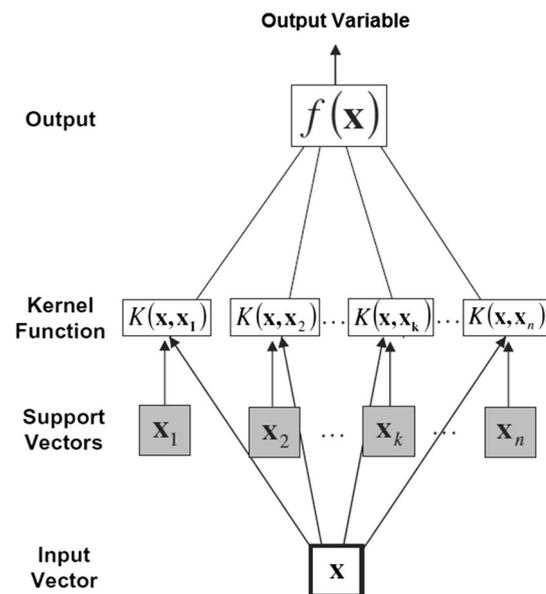


Fig. 3 A schematic structure of SVM model (Yoon et al. 2011)

is adopted to have an optimal combination of the three parameters.

Data specification

Present study includes 380 soil samples collected from different horizons of 80 soil profiles. The approach adopted in this study consists of marking 80 sites spaced at approximately 1000 m on the map of the area at a scale of 1:25,000. These sites were set out using the GPS, and bores were dug of the size of 1 m × 2 m × 1.5 m (depth). Samples were taken from each site and were analyzed in laboratory to measure their content of CEC, clay, silt, sand, gypsum and OM. For this study, texture represented by clay, silt and sand as well as gypsum and organic matter was measured by Bouyoucos hydrometer, Stone's and Walkley-Black methods, respectively (Sayegh et al. 1978; Nelson and Sommers 1982) and the CEC values (in Cmol Kg⁻¹) were determined using the Bower method (Sparks et al. 1996). Total data of 380 samples were divided randomly in a way to use 323 for model training, whereas remaining 57 samples (15 % of total samples) were used for testing the models. The size of sand particles ranges between 2.0 and 0.05 mm; silt, 0.05 and 0.002 mm; and clay, less than 0.002 mm. Gypsum, known to chemists as calcium sulfate dihydrate (CaSO₄·2H₂O), is a soft naturally occurring mineral and primarily used in agriculture as a 'clean green' soil conditioner and fertilizer (Anonymus 1992). Organic matter in soil is composed of carbon, oxygen, hydrogen, nitrogen, phosphorus and sulfur, which is a complex mixture of substances that range from freshly deposited plant and animal parts to the residual humus-stable organic compounds. Clay minerals usually range from 10 to 150 meq/100 g in CEC values. Organic matter ranges from 200 to 400 meq/100 g. So, the kind and amount of clay and organic matter content greatly influence the CEC of soils (Parker 2010). The key statistic values of

the dataset used are presented in Table 1, and Table 2 presents the estimated values of correlation between the variables, according to which: (1) The correlation between CEC, silt and gypsum is poor, and (2) CEC is directly correlated with clay, silt and OM but inversely with sand and gypsum. Based on Table 1, the training and testing data are broadly agreeable to each other, whereas Table 2 shows that the correlation of the CEC with the input variables is ranked in the order of clay, OM, sand (inverse correlation), silt and gypsum, but it is rather concerning that the cross-correlation between silt and sand is quite strong.

Performance measures

Three performance criteria, correlation coefficient (CC), root-mean-square error (RMSE) and the Nash–Sutcliffe efficiency coefficient (E), are used in this study to assess the goodness of fit of the models. The CC, which ranges from -1 to 1, is a statistical measure of how well the regression line is close to the observed data, and a coefficient of ±1 indicates that the regression line perfectly fits the observed data. The RMSE can provide a balanced evaluation of the goodness of fit of the model as it is more sensitive to the larger relative errors caused by the low value, and the perfect model will have a value of zero. The E-values range between -∞ and 1.0, with = 1 obtained by perfect fits. These performance measures and information criteria are calculated by (Ghorbani et al. 2013):

$$CC = \frac{\sum_{i=1}^n [(CEC_o)_i - (\overline{CEC_o})] [(CEC_e)_i - (\overline{CEC_e})]}{\sqrt{\sum_{i=1}^n [(CEC_o)_i - (\overline{CEC_o})]^2 \sum_{i=1}^n [(CEC_e)_i - (\overline{CEC_e})]^2}} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n ((CEC_o)_i - (CEC_e)_i)^2}{n}} \tag{10}$$

Table 1 Summary of statistics characteristics of input and output parameters

	Training						Testing					
	Data points	Mean	Max	Min	Std	CV	Data points	Mean	Max	Min	Std	CV
Input												
Clay (%)	323	31.43	55.72	3.86	9.69	0.31	57	30.24	53.26	10.02	10.40	0.34
Silt (%)	323	28.92	54.63	3.03	9.35	0.32	57	26.22	49.96	6.95	10.71	0.41
Sand (%)	323	39.58	89.96	3.61	15.70	0.40	57	43.59	79.60	13.33	17.98	0.41
Gypsum (%)	323	7.65	49.30	0.10	8.56	1.12	57	5.90	21.50	0.70	5.73	0.97
OM (%)	323	0.44	1.45	0.00	0.28	0.64	57	0.39	1.11	0.00	0.27	0.69
Output												
CEC (Cmol K g ⁻¹)	323		26.21	5.02	4.61	0.31	57	13.48	25.19	50.3	5.24	0.39

Data points = number of data; max = maximum value; min = minimum value; std = standard deviation; CV = coefficient of variation

Table 2 Correlation matrix of the data used

	Clay (%)	Silt (%)	Sand (%)	Gypsum (%)	OM (%)	CEC (Cmol Kg ⁻¹)
Clay (%)	1	0.37 ^a	-0.82 ^a	0.03	0.06	0.61 ^a
Silt (%)	0.37 ^a	1	-0.83 ^a	-0.03	0.18 ^a	0.37 ^a
Sand (%)	-0.82 ^a	-0.83 ^a	1	-0.002	-0.14 ^a	-0.59 ^a
Gypsum (%)	0.03	-0.03	-0.002	1	-0.27 ^a	-0.17 ^a
OM (%)	0.06	0.18 ^a	-0.14 ^a	-0.27 ^a	1	0.59 ^a
CEC (Cmol Kg ⁻¹)	0.62 ^a	0.37 ^a	-0.59 ^a	-0.17 ^a	0.59 ^a	1

^a At 0.05 level of significance

$$E = 1 - \frac{\sum_{i=1}^n ((CEC_o)_i - (CEC_e)_i)}{\sum_{i=1}^n [(CEC_o)_i + (CEC_e)_i]} \quad (11)$$

where the subscripts ‘o’ and ‘e’ represent the observed and estimated values; the average value of the associated variable is represented with a bar above it and *n* is the total number of records.

Results and Discussion

In the application of ANN and SVM in this study for the predication model, five soil characteristics including clay, silt, sand, gypsum and organic matter are considered as the input variables, and CEC is considered as the output variable. In order to develop PTFs using proposed models to predict CEC, both training and test data were normalized using following equation:

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (12)$$

where *x_n* is the normalized value (lying between 0 and 1), *x* is the original data point, *x_{min}*, and *x_{max}* are the minimum and maximum values in the dataset, respectively. In this study, an MLP model has been constructed for predicting the CEC using the MATLAB software, which also randomizes the dataset, as the data points are not time series. Neurons in the input layer have no transfer function. Logistic sigmoid transfer function was used in the hidden layer, and linear transfer function (purelin) was employed from the hidden to the output layers as an activation function. Logistic sigmoid activation function is known as the most used nonlinear activation function. Most important features of the sigmoid functions are continuity and differentiability. Purelin activation function is a pure linear function which outputs any inputs without any change. This is to enable the network to be able to take care of any nonlinearities in the input data and at the output, to be able to give a wide range of values (Zhang et al. 1998).

The neural network used in this study was trained using 1000 epochs and the Levenberg–Marquardt learning algorithm. The optimal value of learning rate and momentum used in present study was 0.001 and 0.9, respectively. Mean square error (MSE) is used as the performance measure of MLP. The optimal number of neuron in the hidden layer was identified using a trial-and-error procedure by varying the number of hidden neurons from 1 to 20. Results from Fig. 4 suggest that a neural network with one hidden layer having 13 nodes performs well with this dataset and achieves CC value = 0.920, RMSE = 1.803 Cmol Kg⁻¹ and *E* = 0.846 with training and CC = 0.860, RMSE = 2.731 Cmol Kg⁻¹ and *E* = 0.723 with testing dataset. The ANN modeling was implemented using MATLAB software. Figure 5 shows the scatter plots obtained from the optimum ANN model for training and testing dataset. Results in terms of various performance measures and Fig. 4 suggest that used neural network model achieves close approximations of the actual observations, suggesting effectiveness of this modeling approach in predicting CEC values.

To compare the performance of neural network modeling approach, SVM, another modeling approach is used for predicting CEC values. Many studies on the use of SVM in soil science suggested the improved performance by RBF kernel-based SVM (Twarakavi et al. 2009; Zhu and Xu 2011; Cisty et al. 2011; Lamorski et al. 2013; Shi et al. 2013). Therefore, in present study, the RBF kernel with parameter *γ* is used as the kernel function. Therefore, in this study, the RBF kernel with parameters (*C*, *ε*, *σ*) is used as the kernel function for CEC modeling and the accuracy of a SVM model is dependent to identify the parameters. To obtain a suitable value of these parameters (*C*, *ε*, *σ*), the RMSE was used to optimize parameters. The performance criteria with optimal parameters (*C*, *ε*, *σ*) = (4.68, 0.79, 2.36) suggest that SVM perform well in predicting CEC values (CC = 0.896, RMSE = 2.040 Cmol Kg⁻¹ and *E* = 0.803 with training set and CC = 0.849, RMSE = 2.796 Cmol Kg⁻¹ and *E* = 0.709 with testing

Fig. 4 **a** The RMSE and **b** the CC of different neural networks

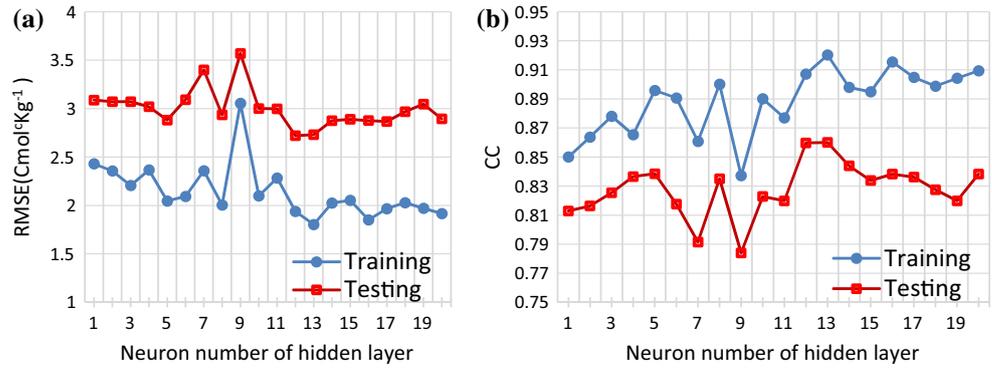


Fig. 5 Scatter plots of measured and predicted CEC by ANN model in **a** training set **b** testing set

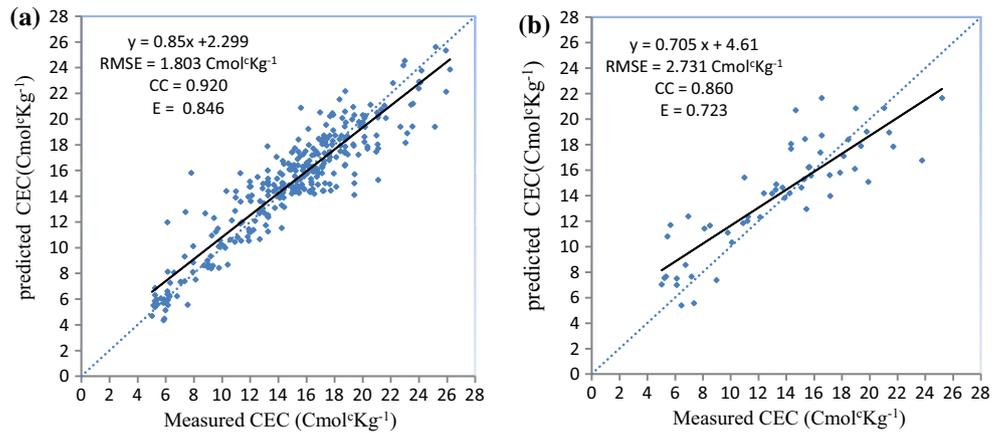


Fig. 6 Scatter plots of measured and predicted CEC by SVM model in **a** training set **b** testing set

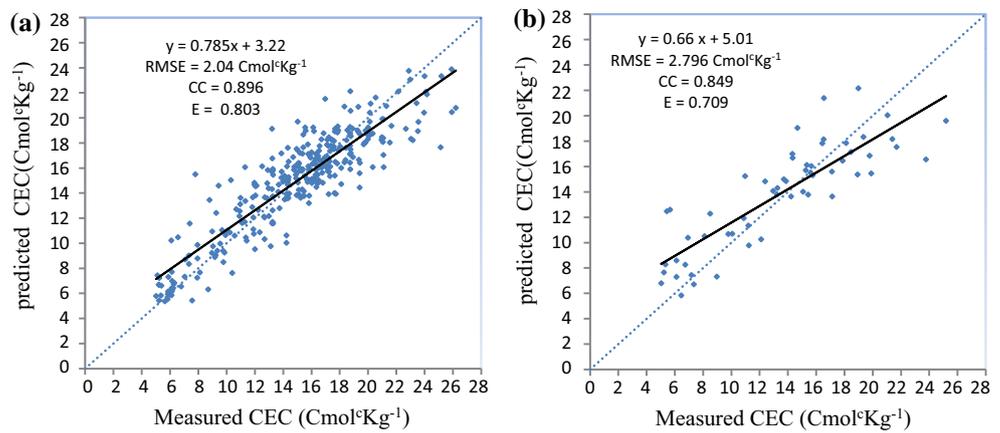
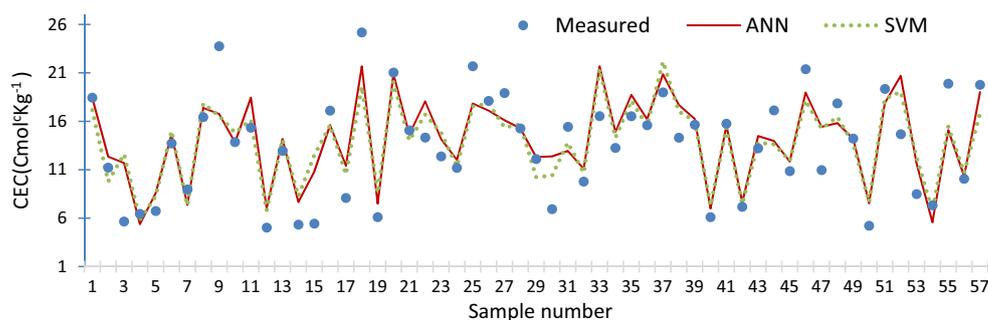


Table 3 Performance evaluation criteria of SVM and ANN models

Model	Training			Testing		
	CC	RMSE (Cmol Kg ⁻¹)	E	CC	RMSE (Cmol Kg ⁻¹)	E
ANN	0.92	1.803	0.846	0.86	2.731	0.723
SVM	0.896	2.04	0.803	0.849	2.796	0.709

Fig. 7 a Comparison of the SVM and ANN models predicted and measured CEC values in testing set



set). Figure 6 shows the scatter plots of the results obtained using the SVM model for training and testing dataset. The results show almost perfect agreement between predicted and measured values.

The performance of the ANN and SVM techniques is compared in Table 3. Given the obtained results in Table 3, results indicate an appropriate performance by both models for CEC prediction. ANN model with high CC (0.920), lowest RMSE equal to (1.803 Cmol Kg^{-1}) and highest Nash–Sutcliffe coefficient (0.846) with training data and high CC (0.860), lowest RMSE equal to (2.731 Cmol Kg^{-1}) and highest Nash–Sutcliffe coefficient (0.723) with test dataset found to be performing well in comparison with SVM models with this dataset. Comparison of ANN and SVM models as illustrated in Fig. 7 suggests that both ANN and SVM methods perform poorly in extrapolating the maximum and minimum values of CEC data.

Conclusion

In this study, ANN and SVM were used to predict the CEC for an agricultural site using five input variables. From these results, following conclusion can be drawn:

The results from this study suggest that both ANN and SVM models had the ability to predict CEC within acceptable limits. ANN and SVM methods perform poorly in extrapolating maximum and minimum values of CEC data. ANN model provided better estimation in the testing period in comparison with SVM for CEC prediction. Before using both ANN and SVM modeling approaches for CEC prediction, it is suggested that these techniques may be used with the datasets from different regions as all machine learning approaches are data-dependent in nature.

Acknowledgments We are grateful to the University of Tabriz for their kind encouragement and their laboratory facilities.

References

Anonymus (1992) Soil survey laboratory methods and procedures for collection soil sample. In: Soil Conservation Service, Investment Report Government Printing Office, Washington, DC

- ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000) Artificial neural networks in hydrology I: preliminary concepts. *J Hydrol Eng* 5:115–123
- Ayoubi S, Emami N, Ghaffari N, Honarjoo N, Sahrawat KL (2014) Pasture degradation effects on soil quality indicators at different hillslope positions in a semiarid region of western Iran. *Environ Earth Sci* 71(1):375–381
- Baker L, Ellison D (2008) Optimisation of pedotransfer functions using an artificial neural network ensemble method. *Geoderma* 144:212–224
- Bhagwat PP, Maity R (2012) Multistep-ahead river flow prediction using LS–SVR at daily scale. *J Water Resource Prot* 4:528–539
- Bilgili M, Sahin B, Yasar A (2007) Application of artificial neural networks for the wind speed prediction of target station using reference stations data. *Renew Energy* 32:2350–2360
- Bouma J (1989) Using soil survey data for quantitative land evaluation. *Adv Soil Sci* 9:177–213
- Cao LJ, Tay Francis EH (2003) Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans Neural Netw* 14:1506–1518
- Carpena O, Lux A, Vahtras K (1972) Determination of exchangeable calcareous soils. *Soil Sci* 33:194–199
- Chaudhuri S, McDonald LM, Pena-Yewtukhiw EM, Skousen J, Roy M (2013) Chemically stabilized soil organic carbon fractions in a reclaimed minesoil chronosequence: implications for soil carbon sequestration. *Environ Earth Sci* 70(4):1689–1698
- Cisty M, Bajtek Z, Bezak J (2011) Support vector machine based model for water content in soil interpolation. *Geophys Res Abstr* 13:1–2
- Evans LJ (1989) Chemistry of metal retention by soils. *Environ Sci Technol* 23:1046–1056
- Gago C, Romar A, Fernandez-Marcos ML, Alvarez E (2014) Fluoride sorption and desorption on soils located in the surroundings of an aluminium smelter in Galicia (NW Spain). *Environ Earth Sci* 72(10):4105–4114
- Gatima E, Mwinyihija M, Killham K (2006) Determination of adsorption efficiency based on cation exchange capacity related to red earth, bone meal and pulverised fly ash as ameliorants to lead contaminated soils. *Int J Environ Sci Technol* 3(3):269–280
- Ghorbani MA, Khatibi R, Hosseini B, Bilgili M (2013) Relative importance of parameters affecting wind speed prediction using artificial neural networks. *Theor Appl Climatol* 114:107–114
- Gill MK, Tirusew A, Mariush WK, Mac M (2006) Soil moisture prediction using support vector machines. *J Am Water Resour Assoc* 42:1033–1046
- Gruszczyński S (2009) Assessment of suitability of various models for estimating cation exchange capacity (CEC). *Pol J Soil Sci* 42(1):16–29
- Haykin S (1999) *Neural networks: a comprehensive foundation*. Macmillan Publishing, New York
- Huang LM, Deng CB, Huang N, Huang XJ (2013) Multivariate statistical approach to identify heavy metal sources in

- agricultural soil around an abandoned Pb–Zn mine in Guangxi Zhuang Autonomous Region, China. *Environ Earth Sci* 68(5):1331–1348
- Keller A, Von Steiger B, Vander Zee ST, Schulin R (2001) A stochastic empirical model for regional heavy metal balances in agroecosystems. *J Environ Qual* 30:1976–1989
- Kisi O (2005) Daily river flow forecasting using artificial neural networks and auto-regressive models. *Turk J Eng Environ Sci* 29:9–20
- Kovačević M, Bajat B, Gajić B (2010) Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154(3–4):340–347
- Lamorski K, Pastuszka T, Krzyszczyk J, Sławiński C, Witkowska-Walczak B (2013) Soil water dynamic modeling using the physical and support vector machine methods. *Vadose Zone J* 12(4). doi:10.2136/VZJ2013.05.0085
- Liao K, Xu S, Wu J, Zhu Q, An L (2014) Using support vector machines to predict cation exchange capacity of different soil horizons in Qingdao City, China. *J Plant Nutr Soil Sci* 177(5):775–782
- Malekian A, Jafarzadeh AA (2011) Qualitative land suitability evaluation of the Khajeh research station for wheat, barley, alfalfa, maize and safflower. *Res Plant Biol* 1(5):33–40
- Manrique LA, Jones CA, Dyke PT (1991) Predicting cation exchange capacity from soil physical and chemical properties. *Soil Sci Soc Am J* 55:787–794
- Marimon MPC, Roisenberg A, Viero AP, Camargo FAD, Suhogsoff AV (2013) Evaluation of the potential impact of fluorine-rich fertilizers on the Guarani Aquifer System, Rio Grande do Sul, Southern Brazil. *Environ Earth Sci* 69(1):77–84
- McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in neurons activity. *Bull Math Biophys* 5:115–133
- Melesse AM, Hanley RS (2005) Artificial neural network application for multi ecosystem carbon flux simulation. *Ecol Model* 189:305–314
- Minasny B, McBratney AB (2002) The neuro-m methods for fitting neural network parametric pedotransfer functions. *Soil Sci Soc Am J* 66:352–361
- Minasny B, McBratney AB, Bristow KL (1999) Comparison of different approaches to the development of pedotransfer functions for water retention curves. *Geoderma* 93:225–253
- Misra D, Oommen T, Agarwal A, Mishra SK, Thompson AM (2009) Application and analysis of support vector machine based simulation for runoff and sediment yield. *Biosyst Eng* 103:527–535
- Moghaddas NH, Namaghi HH, Ghorbani H, Dahrazma B (2013) The effects of agricultural practice and land-use on the distribution and origin of some potentially toxic metals in the soils of Golestan province, Iran. *Environ Earth Sci* 68(2):487–497
- Moharami S, Jalali M (2013) Effects of cations and anions on iron and manganese sorption and desorption capacity in calcareous soils from Iran. *Environ Earth Sci* 68(3):847–858
- Moon Y, Zhang YS, Song Y, Park E, Moon HS (2013) Multivariate statistical analysis and 3D-coupled Markov chain modeling approach for the prediction of subsurface heterogeneity of contaminated soil management in abandoned Guryong Mine Tailings, Korea. *Environ Earth Sci* 68(6):1527–1538
- Nelson DW, Sommers LE (1982) Total carbon, organic carbon and organic matter. In: Page AL, Miller RH, Keeney DR (eds) *Methods of soil analysis, Part II: Chemical and microbiological properties*. American Society of Agronomy, Madison, pp 539–579
- Parker R (2010) *Plant and soil science: Fundamentals & applications*. Clifton Park, NY: Delmar Cengage Learning
- Rezapour S (2014) Response of some soil attributes to different land use types in calcareous soils with Mediterranean type climate in north-west of Iran. *Environ Earth Sci* 71(5):2199–2210
- Rossel RAV, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158(1–2):46–54
- Sayegh AH, Khan P, Ryan J (1978) Factors affecting gypsum and cation exchange capacity determination in gypsiferous soils. *SSJ* 125:294–300
- Schaap MG, Leij FJ, van Genuchten MT (1998) Neural network analysis for hierarchical prediction of soil hydraulic properties. *Soil Sci Soc Am J* 62:847–855
- Semhi K, Al Abri R, Al Khanbashi S (2013) Impact of sewage and mining activities on distribution of heavy metals in the water–soil–vegetation system. *Int J Environ Sci Tech* 11(5):1285–1296
- Shi T, Cui L, Wang J, Fei T, Chen Y, Wu G (2013) Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant Soil* 366(1–2):363–375
- Silva B, Paradelo R, Vazquez N, Garcia-Rodeja E, Barral MT (2013) Effect of the addition of granitic powder to an acidic soil from Galicia (NW Spain) in comparison with lime. *Environ Earth Sci* 68(2):429–437
- Silveira CT, Oka-Fiori C, Santos LJC, Sirtoli AE, Silva CR, Botelho MF (2013) Soil prediction using artificial neural networks and topographic attributes. *Geoderma* 195–196:165–172
- Sparks DL, Page AL, Helmke PA, Leoppert RH, Soltanpour PN, Tabatabai MA, Johnston GT, Summer ME (1996) *Methods of soil analysis*. Soil Science Society of America, Madison
- Subramanyam B, Das A (2009) Linearized and non-linearized isotherm models comparative study on adsorption of aqueous phenol solution in soil. *Int J Environ Sci Technol* 6(4):633–640
- Tang L, Zeng G, Nourbakhsh F, Guoli L, Shen GL (2009) Artificial neural network approach for predicting cation exchange capacity in based on physico-chemical chemical properties. *Environ Eng Sci* 26(1):137–146
- Twarakavi NKC, Šimůnek J, Schaap MG (2009) Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. *Soil Sci Soc Am J* 73:1443–1452
- Van Bladel R, Frankart R, Gheyi HR (1975) A comparison of three methods of determining the cation exchange capacity of calcareous soils. *Geoderma* 13:289–298
- Vapnik VN (1995) *The nature of statistical learning theory*. Springer, New York
- Vapnik VN (1998) *Statistical learning theory*. Wiley, New York
- Wang H, Hu D (2005) Comparison of SVM and LS–SVM for regression. In: *Proceedings of the international conference on neural networks and brain proceedings (ICNNB '05)*, pp 279–283
- Wang W, Men C, Lu W (2008) Online prediction model based on support vector machine. *Neurocomputing* 71:550–558
- Wu W, Wang X, Xie D, Liu H (2008) Soil water content forecasting by support vector machine in purple hilly region. *Int Fed Inf Proc* 258:223–230
- Yilmaz I, Kaynar O (2011) Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert Syst Appl* 38:5958–5966



- Yola ML, Eren T, Atar N (2014) A novel efficient photocatalyst based on TiO₂ nanoparticles involved boron enrichment waste for photocatalytic degradation of atrazine. *Chem Eng J* 250:288–294
- Yoon H, Jun SC, Hyun Y, Bae GO, Lee KK (2011) A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J Hydrol* 396:128–138
- Zhang G, Patuwo EB, Hu M (1998) Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 14:35–62
- Zhu P, Xu B (2011) Fusion of ECa data using SVM and rough sets augmented by PSO. *J Comput Inf Syst* 7–1:295–302

