Arid Land Research and Management, 28:383–394, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 1532-4982 print/1532-4990 online
DOI: 10.1080/15324982.2013.871599

Feature Selection Using Parallel Genetic Algorithm for the Prediction of Geometric Mean Diameter of Soil Aggregates by Machine Learning Methods

A. A. Besalatpour[1], S. Ayoubi[2], M. A. Hajabbasi[2],
A. Yousefian Jazi[3], and A. Gharipour[4]

[1]Department of Soil Science, Vali-e-Asr University of Rafsanjan,
Rafsanjan, Iran
[2]Department of Soil Science, Isfahan University of Technology,
Isfahan, Iran
[3]Department of Biomedical Engineering, Seoul National University,
Jongno-gu, Seoul, Korea
[4]School of Information and Communication Technology,
Gold Coast Campus, Griffith University, Australia

Aggregate stability is a useful soil physical dynamic index of soil resistivity to surface wind and water erosion in all ecosystems, especially, in arid and semi-arid regions. Two machine learning techniques including support vector machines (SVMs) and artificial neural networks (ANNs) were used to develop predictive models for the estimation of geometric mean diameter (GMD) of soil aggregates. An empirical multiple linear regression (MLR) model was also constructed as the benchmark to compare their performances. Furthermore, the influence of feature space dimension reduction using parallel genetic algorithm (PGA) on the prediction accuracy of all investigated techniques was evaluated. The ANN model achieved greater accuracy in GMD prediction as compared to the MLR and SVM models. The obtained ERROR% value in GMD prediction using the ANN model was 6.9%, while it was 15.7 and 10.6% for the MLR and SVM models, respectively. Feature selection using PGA improved the prediction accuracy of all investigated techniques. The coefficient of determination (R^2) values between the measured and the predicted GMD values using PGA-based MLR, SVM, and ANN models increased by 20.0, 12.2, and 8.8% in comparison with the proposed MLR, SVM, and ANN models. In conclusion, it appears that the PGA-based ANN model could be considered as an alternative to conventional regression models for the GMD prediction.

Keywords artificial neural networks (ANNs), geometric mean diameter (GMD), parallel genetic algorithms (PGAs), support vector machines (SVMs)

Aggregate stability (AS) is a useful physical dynamic index of soil resistivity to wind and water erosion that can be evaluated by various laboratory-based techniques and indices (such as geometric mean diameter, GMD; Calero et al., 2008). Nevertheless,

Received 20 May 2013; accepted 29 November 2013.

Address correspondence to A. A. Besalatpour, Department of Soil Science, Vali-e-Asr University of Rafsanjan, 7718897111, Rafsanjan, Iran. E-mail: a.besalatpour@vru.ac.ir; a_besalatpour@yahoo.com

most of these techniques are generally time-consuming and/or rather cumbersome, particularly, when a large number of samples are required to be characterized for application on a large scale. Therefore, it would be advantageous if AS could be estimated indirectly from more easily available data.

Thus far, in many researches the emphasis has been placed on conventional linear regression methods (such as multiple linear regression, MLR) to predict AS while they can only fit a linear function to predictor-AS data pairs. However, the effect of the predictors on AS is not usually linear in nature. Recently, soil scientists have shown a keen interest in developing nonlinear indirect approaches to overcome this problem. Among the evaluated techniques, machine learning (ML) approaches have attracted greater interest (Muttil & Chau, 2006; Twarakavi et al., 2009). The ML techniques such as artificial neural networks (ANNs) and support vector machines (SVMs) can be used to provide a low-cost approach with a tolerance of imprecision, uncertainty and approximation, and to avoid over-fitting problems. This makes ML capable of analyzing large-scale data and thus solving the problems which conventional linear methods have not yet been able to solve in a satisfactory cost-effective manner (Chau et al., 2005; Wang et al., 2009; Huang et al., 2010).

Various ML techniques have been studied and applied in the last decades for scientific research and have been found to be useful in agricultural sciences (Muttil & Chau, 2006; Qiao et al., 2010; Besalatpour et al., 2012). Owing to the large number of them now available, finding an appropriate one to use for a site-specific problem is becoming increasingly difficult for novice modelers. Therefore, it is distinctly desirable to introduce expertise in the system with a view to helping novice users to choose an appropriate ML technique. Hence, this study was conducted to evaluate the effectiveness of SVM and ANN techniques in developing prediction functions for estimating soil aggregate stability by considering this hypothesis that: they may provide much lower variance and smaller magnitude of errors for the GMD prediction in comparison with the commonly used linear regression prediction techniques. We also expected that SVM technique will show a greater potential in the GMD prediction in comparison with ANN technique.

On the other hand, in many scientific research and engineering applications, researchers are interested in identifying the most important factors influencing a certain outcome of interest. In addition, original datasets often contain features some of which are either redundant or irrelevant to the target concept (Fayyad et al., 1996). Another objective of this study was to evaluate the effects of feature space dimension reduction using parallel genetic algorithm (PGA) on prediction accuracy of the investigated techniques.

Materials and Methods

Brief Description of the Modeling Approaches

Feature Selection Using Parallel Genetic Algorithm

Real-world datasets often contain a large number of features some of which are redundant and/or irrelevant to the target variable(s). This usually happens when it is unknown which features are related to a target concept and particularly when domain knowledge is unavailable or incomplete. Many features are then introduced to represent an unknown domain. The existence of irrelevant and/or redundant features may make vague the distribution of really relevant features for a target concept and thus cause damage to the model. Data feature selection and reduction techniques may be useful in solving this problem (Fayyad et al., 1996).

Different data feature selection techniques can be adopted for feature space dimension reduction. Among them, the parallel genetic algorithm (PGA) outperforms all algorithms in complex real problems (Zhu and Chipman, 2006). PGAs are an extension of the traditional genetic algorithm (GA) sequential models which represent a new class of algorithms in that they search the space of solutions differently. A PGA basically consists of various GAs, each processing a part of the population or independent populations, with or without communication between them. The main advantage of a PGA over GA is to reduce the processing time required-in order to achieve an acceptable solution-to explore a solution space.

Support Vector Machines (SVMs)

SVMs are extended based on statistical learning to solve a regression problem with a given set of training data

$$D = \{(x_i, y_i)\}_{i=1}^{n}$$

where x_i is the sample vector and $x_i \in X$, y_i is the corresponding response, and $y_i \in R$, and n is the total number of samples (Vapnik, 1995, 1998). The regression function of SVM is represented as:

$$y = f(x) = w_i \phi_i(x) + b \tag{1}$$

ϕ_i is the input sample and w_i and b are coefficients estimating by minimizing the risk function:

$$r(C) = C\frac{1}{N}\sum_{i=1}^{N} l_\varepsilon(d_i, y_i) + \frac{1}{2}\|\omega\|^2 \tag{2}$$

where

$$L_\varepsilon(d, y) = \left\{ \begin{matrix} |d - y| - \varepsilon & if|d - y| \geq \varepsilon \\ 0 & otherwise \end{matrix} \right\} \tag{3}$$

Then, Eq. (2) is transformed into the following constrained form:

$$\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

subject to:

$$\omega_i \phi(x_i) + b_i - d_i \leq \varepsilon + \xi_i^*,$$

$$d_i - \omega_i \phi(x_i) - b_i \leq \varepsilon + \xi_i, \;\; \xi_i, \xi_i^*, \qquad i = 1, 2, \ldots, N$$

where ξ_i and ξ_i^* are the corresponding positive and negative errors at the *i-th* point, respectively. The constrained optimization problem then solved using the

Lagrangian theory by:

$$L = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N} a_i(\omega_i\phi(x_i) + b - d_i + \varepsilon + \xi_i)$$
$$- \sum_{i=1}^{N} a_i^*(d_i - \omega_i\phi(x_i) - b + \varepsilon + \xi_i^*) - \sum_{i=1}^{N}(\beta_i\xi_i + \beta_i^*\xi_i^i) \quad (5)$$

Equation (5) is minimized with respect to primal variables ω_i, b, ξ_i and ξ_i^*, and maximized with respect to the non-negative Lagrangian multipliers α_i^* and β_i^*. Finally, the Karush-Kuhn-Tucker conditions are applied to the regression to have a dual Lagrangian form of:

$$\nu(\alpha_i, \alpha_i^i) = \sum_{i=1}^{N} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - a_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \quad (6)$$

The Lagrange multipliers α_i and α_i^* are calculated and satisfying to the equality $\alpha_i, \alpha_i^* = 0$ by using Eq. (6). The optimal desired weight vector of the regression hyper plane is expressed as:

$$\omega^* = \sum_{i=1}^{N}(\alpha_i - \alpha_i^j)k(x, x_i) \quad (7)$$

Thus, the regression function can be explained as:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x, x_i) + b \quad (8)$$

The $K(x_i, x)$ is named the kernel function. The radial basis function (RBF) was used as the kernel functions in this study:

$$K(x_i, x) = \exp\left(\frac{\|x_i - x\|}{2\sigma^2}\right) \quad (9)$$

where σ is kernel parameter (Cristianini and Taylor, 2000; Li et al., 2009).

Artificial Neural Networks (ANNs)

ANNs are an extension of ML methods made up of a number of interconnected processing layers. Generally, an ANN is consisted of three parts: an input layer, one (or several) hidden layer(s), and an output layer of neurons. Thus, the layers between the input and output layers are called hidden layers which may contain a large number of hidden processing elements. The weights are used to fully interconnect each neighboring layer. The received information from the outside by the input layer neurons are transmitted to the neurons of the hidden layer. Finally, the output layer neurons produce the network estimations to the outside world (Qiao et al., 2010).

Compilation of the Data

The study area was a part of the Bazoft watershed (31° 37′ to 32° 39′ N and 49° 34′ to 50° 32′ E) located in northern part of Karun river basin in central Iran. Many parts of the watershed are severely susceptible to wind and water erosion where surface soil aggregate stability could be a proper index of their resistivity to erosion and thus its estimation can be a valuable source of information for other modelers. A total of 160 soil samples were collected (October 2010) from the top 5 cm soil layer to produce a measurement of the diversity of soil properties in those parts. The soil samples were then air-dried and ground to pass a 2-mm sieve for the determination of intrinsic soil properties. Soil organic matter (SOM) content was determined by the Walkley-Black method (Nelson and Sommers, 1986). Particle size distribution in the soil samples (clay, silt, and sand) were measured using the procedure described by Gee and Bauder (1986) and calcium carbonate equivalent (CCE) content was determined by the back-titration method (Nelson, 1982).

The Kemper and Rosenau (1986) method was used to determine wet-aggregate stability and geometric mean diameter (GMD) of the aggregates was determined as an indicator of AS. Briefly, 50 g of the <4.75 mm aggregates were placed on the topmost of a nest of sieves of mesh size 2, 1, 0.5, and 0.25 mm. The samples were first immersed in the water and then sieved by moving the sieve set vertically. The soil mass on each sieve was dried at 105°C for 24 h, weighted and corrected for the sand/gravel particles to obtain the proportion of the water-stable aggregates. The GMD (mm) of water-stable aggregates was calculated using the following equation:

$$GMD = \exp\left[\sum_{i=1}^{n} W_i Log X_i\right] \tag{10}$$

where X_i is the arithmetic mean diameter of each size fraction (mm), and W_i is the proportion of the total water-stable aggregates in the corresponding size fraction after deducting the weight of sand/gravel particles as previously indicated.

The topographic attributes of the representative points including elevation, slope, and aspect were characterized using a 20-m by 20-m digital elevation model (DEM). For quantifying the vegetation in each representative point, the normalized difference vegetation index (NDVI) was derived using Indian Remote Sensing (IRS) satellite photo of April 2008 at a spatial resolution of 24-m by 24-m (Indian Space Applications Centre, Ahmedabad, India).

Two different sets of the available properties were then prepared as inputs to each investigated model. The first set consisted of all measured parameters and the second set included selected features resulted from the PGA analysis. Each data set was divided into two sets consisted of training and testing. The training set of 113 samples was obtained out of total 160 and the remained 47 soil samples were used as the testing set. The Clementine software (International Business Machines Corporation, Chicago, USA) was used to build the models. For the SVM analysis, the RBF was used as the kernel function and a feed forward neural network with back propagation training algorithm was employed for the development of the ANN models. The number of hidden neurons and epochs in the ANN analysis as well as number of generations and parallel paths in the PGA analyses were determined by a trial and error procedure. The mean square error (MSE), mean absolute error (MAE), and error percentage (ERROR%) between the measured and the predicted GMD values were used to evaluate the performance of the models.

Results

Based on the PGA analysis, the clay, NDVI, and aspect parameters were accounted as the redundant features among the input parameters (i.e., clay, sand, silt, SOM,

Table 1. Goodness-of-fit of the proposed MLR, SVM, and ANN models for the prediction of GMD

	Evaluation criterion		
Model type	ERROR (%)	MAE (%)	MSE (%)
MLR	15.7	9.0	1.2
PGA-MLR	15.5	8.9	1.1
SVM	10.6	6.0	0.5
PGA-SVM	6.9	4.0	0.2
ANN	6.9	4.0	0.2
PGA-ANN	5.6	3.2	0.1

Note: MAE: mean absolute error, MSE: mean square error, GMD: geometric mean diameter, MLR: multiple linear regression, SVM: support vector machine, ANN: artificial neural network, and PGA: parallel genetic algorithm.

CCE, slope, aspect, elevation, and NDVI) for the GMD prediction and thus they were not included in the databases used to build the models combined with PGA (PGA-based models).

Data redundancy reduction using PGA improved the model performances in the GMD prediction (Table 1). For instance, the obtained MSE value in GMD estimation by PGA-MLR model was 9% lower than that obtained by the constructed MLR model using the original data set (all of the features). The MAE and ERROR% values for the PGA-based MLR model were 8.9% and 15.5%, respectively. A similar trend in GMD prediction using both constructed MLR models was also observed for most of the samples (Figure 1). According to the evaluation indices, it appears that the conventional MLR model did not perform well in estimating the GMD. The high observed error values in GMD prediction using both

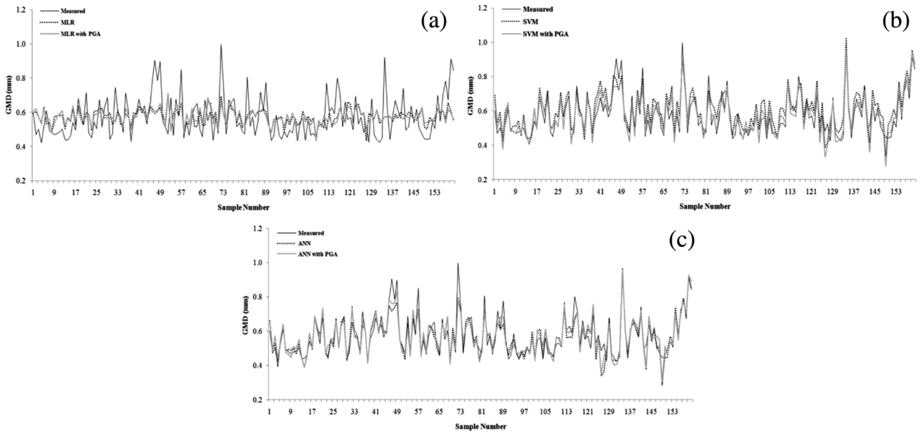

Figure 1. Comparison of the measured and predicted GMD values using the MLR (a), SVM (b), and ANN (c) models with and without PGA (GMD: geometric mean diameter, MLR: multiple linear regression, SVM: support vector machine, ANN: artificial neural network, PGA: parallel genetic algorithm).

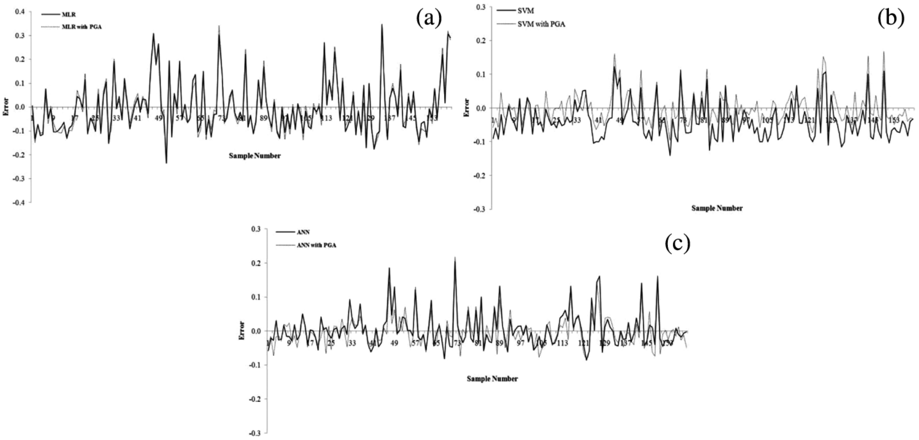

Figure 2. Comparison of the observed errors in the GMD prediction using the MLR (a), SVM (b), and ANN (c) models with and without PGA (GMD: geometric mean diameter, MLR: multiple linear regression, SVM: support vector machine, ANN: artificial neural network, PGA: parallel genetic algorithm).

MLR models (Figure 2) also confirm this finding that MLR models seem to be unreliable for the GMD prediction in the study area.

Table 2 shows the values of SVM parameters for the two proposed SVM models (i.e., the SVM and PGA-SVM models). Whilst the SVM parameters resulting from both SVM models may be satisfactory in terms of C and σ parameter values, the PGA-based SVM model seems to be better. For accuracy measures, the SVM

Table 2. SVM parameter values for the prediction of GMD

	SVM parameter		
Model type	Kernel parameter (σ)	Insensitive parameter (ϵ)	Punishment coefficient (C)
SVM	0.5	0.1	10
SVM with PGA	0.4	0.05	30

Note: GMD: geometric mean diameter, SVM: support vector machine, and PGA: parallel genetic algorithm.

Table 3. ANN parameter values for the prediction of GMD

	ANN parameter		
Model type	No. of hidden neurons	No. of epochs	Learning function
ANN	15	32	TRILMN
ANN with PGA	7	21	TRILMN

Note: GMD: geometric mean diameter, ANN: artificial neural network, and PGA: parallel genetic algorithm.

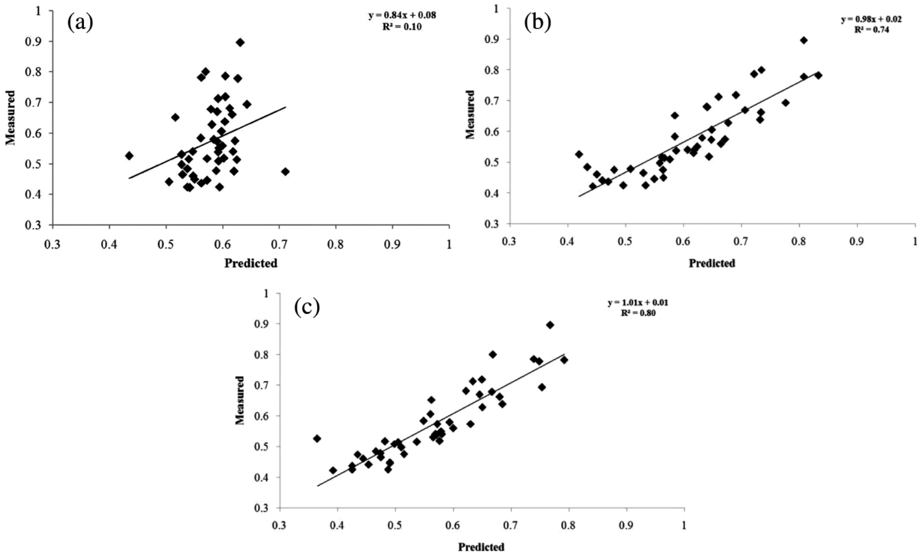

Figure 3. Scatter plots displaying relationships between the measured and the predicted GMD values for the test sample sets of the MLR (a), SVM (b), and ANN (c) models (GMD: geometric mean diameter, MLR: multiple linear regression, SVM: support vector machine, ANN: artificial neural network).

technique provided lower variance and smaller magnitude of errors in the GMD prediction than the MLR models (see Table 1 and Figures 1 and 2). The MAE, MSE, and ERROR% values for the GMD prediction using the SVM model were 6.0, 0.5,

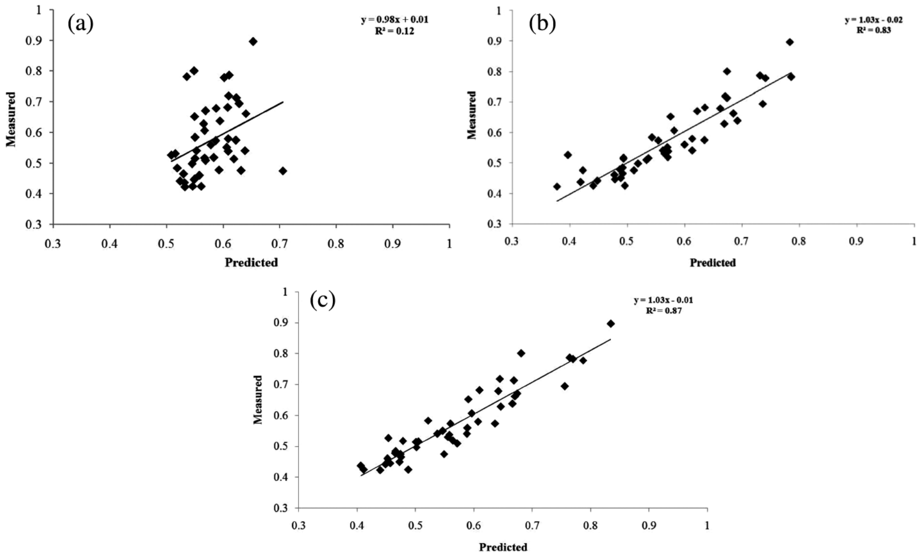

Figure 4. Scatter plots displaying relationships between the measured and the predicted GMD values for the test sample sets of the PGA-based MLR (a), SVM (b), and ANN (c) models (GMD: geometric mean diameter, MLR: multiple linear regression, ANN: artificial neural network, PGA: parallel genetic algorithm).

and 10.6%, respectively (Table 1). Application of the PGA-based SVM model also resulted in the lower ERROR% value of 6.9% compared with the SVM model.

In neural network analysis, 15 and 7 hidden neurons and epoch set numbers of 32 and 21 were generated as the satisfactory results (evaluated by the network performances) for the PGA-ANN and ANN models, respectively (Table 3). The advantage of ANN was more pronounced in the GMD prediction, where lower relative errors have been encountered and the predicted GMD values were in close proximity of the observed values (Figures 1 and 2). The MAE and MSE values for the constructed ANN model with all of the features were 4.0 and 0.2%, respectively (Table 1). The coupling of ANN and PGA improved the prediction accuracy where the ERROR% in the GMD prediction decreased by around 23% using the PGA-based ANN model as compared to the ANN model without PGA.

Comparing the obtained results from the proposed MLR, SVM, and ANN models also revealed that the PGA-based ANN model was better in predicting the soil aggregate stability among the evaluated models (see Table 1). The obtained coefficients of determination (R^2) values between the measured and the predicted GMD values using all investigated approaches also confirm this finding (Figures 3 and 4).

Discussion

In the PGA analysis, the sand, silt, SOM, CCE, slope, and elevation parameters were accounted as important features affecting the GMD prediction. In many researches, the direct and indirect effects of these soil properties and topographic parameters on AS are reported (Amezketa, 1999; Canton et al., 2009). Soil particles influence aggregation because of their specific surface area, cation exchange capacity, and other their physical and chemical properties. SOM content is also considered as a cementing agent for aggregation and can affect GMD value by stabilizing the aggregates. CCE content also influences soil aggregation through its cementing effects and preventing aggregate dispersion. Topography characteristics also affect AS, in particular, through their influences on the dynamics of soil structure and soil properties such as soil water content, SOM, soil texture, carbonate concentration, mineralogy, and plant establishment and development. In addition, they may influence the rate of weathering and erodibility of soils and thus geometric mean diameter of soil aggregates.

From the obtained results in the GMD prediction, it appears that MLR and SVM approaches may be poorer in predicting the GMD than ANN technique. However, the predictive accuracy of the proposed SVM models (especially, the PGA-based SVM model) was not considerably lower than that of ANN technique. The main reason for these findings is that the effect of the predictors on the GMD may not be linear in nature. While the investigated linear models can only fit a linear function to input-GMD data pairs, the nonlinear ML models were probably capable of contorting themselves into a complex form to accommodate the spatial and temporal changes of the input-GMD data pairs. Another reason may be attributed to less data availability for developing reasonable MLR models for the GMD estimation. ANNs, in contrast, can recognize the relationships between input-GMD data pairs with relatively less data because of their distributed and parallel computing nature. Also, in constructed SVM models with the kernel function, the original inputs are first nonlinearly mapped into the feature space and the resulted ϵ-SVM becomes so flexible that can be used to model complicated nonlinear relationships (Li et al., 2009). Therefore, it appears that in the case of insufficient data for reliable regression models to predict GMD, advanced models such as ANNs show a better performance.

Stated plainly, the optimal feature set obtained from the PGA analysis resulted in maximizing the performance of investigated methods for the GMD prediction. The coupling of investigated methods and PGA improved the model performances since it takes advantages of the local optimization of the ML techniques and the global optimization of PGA. It is reasonable that a ML-PGA model with no irrelevant and redundant features is more flexible than a ML model with real-world dataset that often contain a large number of features some of which are redundant and/ or irrelevant to the GMD (Chau et al., 2005). In addition, existence of irrelevant and redundant features in the input data sets will increase the dimensionality of the feature space which can lead to increasing the complexity of interactions among the features and thus the degree of noise in the GMD prediction.

There are limited published studies dealing with the use of nonlinear approaches for the prediction of AS and the emphasis has been placed on the use of conventional linear regression models in many researches. Bazzoffi et al. (1995), for instance, evaluated the efficacy of different linear models for predicting AS from intrinsic soil components and reported that the model developed from soil chemical properties was the most reliable for estimating the AS; however, its construction is very time-consuming. Scheyer (1998) used a linear model to quantify the contributions of different parameters to the composition and size distribution of water-stable aggregates and found that the chemical binding of water-stable aggregates smaller than fine sand size was a function of organic carbon content, iron oxide content, and clay activity. Skidmore and Layton (1992) developed an empirical linear model to predict dry soil aggregate stability from non-easily available soil properties (specific surface area, water content at -1500 J/Kg, clay content, and geometric mean diameter of primary particles). They reported that the relationship between AS and specific surface area was better than that with geometric mean diameter of primary particles, but neither predicted AS as well as water content and clay fraction.

Conclusion

The presented method in the current study provides a robust statistical framework to compare models developed under distinct learning techniques and feature sets for the geometric mean diameter of soil aggregates. The optimal feature set obtained from the PGA analysis resulted in maximizing the performance of all investigated methods. A vast difference appears to be between the investigated techniques by considering their prediction accuracies: ANN and SVM techniques provided much lower variance and smaller magnitude of errors as compared to MLR technique. Whereas the results from both PGA-based ANN and SVM models may be satisfactory, the performance comparison between them shows that ANN technique may provide smaller magnitudes of errors in the GMD estimation. Nevertheless, as other researchers have reported a better prediction performance of SVM as compared to ANN (Twarakavi et al., 2009, for instance), we hypothesize that it might be worthwhile to use a combined SVM-ANN model (complemented by PGA) for the prediction of aggregate stability. However, further researches in this area should be conducted and need to be validated in the future, especially, for soils in different management systems. Finally, we believe that the introduced methods here will provide a novel tool for quantitative estimation of soil aggregate stability as an alternative to existing conventional linear models for soil scientists who look for an aggregate stability prediction tool to achieve smallest error and highest efficiency.

References

Amezketa, E. 1999. Soil aggregate stability: A review. *Journal of Sustainable Agriculture* 14: 83–151.

Bazzoffi, P., J. S. C. Mbagwu, and W. I. E. Chukwu. 1995. Statistical models for predicting aggregate stability from intrinsic soil components. *International Agrophysics* 9: 1–9.

Besalatpour, A., M. A. Hajabbasi, S. Ayoubi, M. Afyuni, A. Jalalian, and R. Schulin. 2012. Soil shear strength prediction using intelligent systems: artificial neural networks and adaptive neuro-fuzzy inference system. *Soil Science and Plant Nutrition* 58: 149–160.

Calero, N., V. Barron, and J. Torrent. 2008. Water dispersible clay in calcareous soils of southwestern Spain. *Catena* 74: 22–30.

Canton, Y., A. Sole-Benet, C. Asensio, S. Chamizo, and J. Puigdefabregas. 2009. Aggregate stability in range sandy loam soils: relationships with runoff and erosion. *Catena* 77: 192–199.

Chau, K. W., C. L. Wu, and Y. S. Li. 2005. Comparison of several flood forecasting models in Yangtze River. *Journal of Hydrologic Engineering ASCE* 10: 485–491.

Cristianini, N., and J. S. Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, New York.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery in databases. *Artificial Intelligence* 17(4): 37–54.

Gee, G. W., and J. W. Bauder. 1986. Particle size analysis, pp. 383–411, in A. Klute, ed., *Methods of Soil Analysis: Part 1: Agronomy Handbook No 9*. American Society of Agronomy and Soil Science Society of America, Madison, WI.

Huang, Y., Y. Lan, S. J. Thomson, A. Fang, W. C. Hoffmann, and R. E. Lacey. 2010. Development of soft computing and applications in agricultural and biological engineering. *Computer and Electronic in Agriculture* 71: 107–127.

Kemper, W. D., and K. Rosenau. 1986. Size distribution of aggregates, pp. 425–442, in A. Klute, ed., *Methods of Soil Analysis: Part 1: Physical and Mineralogical Methods*. American Society of Agronomy, Madison, WI.

Li, H., Y. Liang, and Q. Xu. 2009. Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems* 95: 188–198.

Muttil, N., and K. W. Chau. 2006. Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution* 28: 223–238.

Nelson, D. W., and L. P. Sommers. 1986. Total carbon, organic carbon and organic matter, pp. 539–579, in A. L. Page, ed., *Methods of Soil Analysis: Part 2: Chemical Methods*. American Society of Agronomy and Soil Science Society of America, Madison, WI.

Nelson, R. E. 1982. Carbonate and gypsum, pp. 181–197, in A. L. Page, ed., *Methods of Soil Analysis: Part 2: Chemical Methods*. American Society of Agronomy and Soil Science Society of America, Madison, WI.

Qiao, D. M., H. B. Shi, H. B. Pang, X. B. Qi, and F. Plauborg. 2010. Estimating plant root water uptake using a neural network approach. *Agricultural Water Management* 98: 251–260.

Scheyer, J. C. M. 1998. Modeling soil aggregate size-distribution and water-stability in eroded sediment. Ph.D. Dissertation, University of Nebraska, Lincoln.

Skidmore, E. L., and J. B. Layton. 1992. Dry-soil aggregate stability as influenced by selected soil properties. *Soil Science Society of America Journal* 56: 557–561.

Twarakavi, N. K. C., J. Simunek, and M. G. Schaap. 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. *Soil Science Society American Journal* 73: 1443–1452.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York.

——— 1998. *Statistical Learning Theory*. Wiley, New York.

Wang, W. C., K. W. Chau, C. T. Cheng, and L. Qiu. 2009. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology* 374: 294–306.

Zhu, M., and H. A. Chipman. 2006. Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. *Technometrics* 48: 491–502.