Soft Comput (2012) 16:1347–1351
DOI 10.1007/s00500-012-0815-7

ORIGINAL PAPER

Multivariate adaptive regression spline (MARS) and least squares support vector machine (LSSVM) for OCR prediction

Pijush Samui · Pradeep Kurup

Published online: 18 January 2012

Abstract This article investigates the feasibility of multivariate adaptive regression spline (MARS) and least squares support vector machine (LSSVM) for the prediction of over consolidation ratio (OCR) of clay deposits based on Piezocone Penetration Tests (PCPT) data. MARS uses piece-wise linear segments to describe the non-linear relationships between input and output variables. LSSVM is firmly based on the theory of statistical learning, and uses regression technique. The input parameters of the models are corrected cone resistance (q_t), vertical total stress (σ_v), hydrostatic pore pressure (u_0), pore pressure at the cone tip (u_1), and the pore pressure just above the cone base (u_2). The developed LSSVM model gives error bar of predicted OCR. Equations have also been developed for prediction of OCR. The performance of MARS and LSSVM models has been compared with the traditional methods for OCR prediction. As the results reveal, the proposed MARS and LSSVM models are robust models for determination of OCR.

Keywords OCR · Clay · Piezocone · Multivariate adaptive regression spline · Least squares support vector machine · Error bar

P. Samui (✉)
Centre for Disaster Mitigation and Management,
VIT University, Vellore 632014, India
e-mail: pijush.phd@gmail.com

P. Kurup
Department of Civil and Environmental Engineering,
University of Massachusetts Lowell, 1 University Ave.,
Lowell, MA 01854, USA
e-mail: Pradeep_Kurup@uml.edu

1 Introduction

Due to the complex physical process associated with the formation of soil, the behavior and engineering properties of soil are greatly varied and exhibit heterogeneity and anisotropic behavior. Data-driven models have been shown to have some degree of success to cope with the complexity of geotechnical behavior. In data-driven modeling, the model tries to find a relationship between a training set of input vectors and corresponding outputs. Once training has been accomplished, the goal is to make predictions for new input data. Artificial neural networks (ANNs) are examples of such models that have been used in geotechnical engineering (Goh 1994; Chan et al. 1995; Lee and Lee 1996; Nawari et al. 1999; Rahman et al. 2001; Kurup and Dudani 2002; Celik and Tan 2005; Kumar and Samui 2008). However, ANNs have limitations such as being called a black box approach, arriving at local minima, slow convergence speed, over fitting problems and absence of probabilistic output (Park and Rilett 1999; Kecman 2001). Geotechnical engineers successfully employed support vector machine (SVM) to overcome the problems of ANN partly (Pal 2006; Goh and Goh 2007; Samui et al. 2008; Das et al. 2010). However, the limitations of SVM are given below:

- SVM has high computational complexity due to quadratic programing (Vapnik 1998).
- It has three tuning parameters [capacity factor (C), error insensitive zone (ε), and kernel parameter (Samui 2008)]. The determination of design value of these tuning parameters is a difficult task.

This study investigates the capability of multivariate adaptive regression splines (MARS) and least squares support vector machine (LSSVM) for determination of

OCR based on PCPT data. MARS is a flexible, more accurate, and faster simulation method for both regression and classification problems (Friedman 1991; Salford Systems 2001). LSSVM is a statistical learning theory which adopts least squares as a loss function (Suykens et al. 1999; Suykens and Vandewalle 1999).

This study uses the database from the work of Kurup and Dudani (2002). The database contains information about corrected cone resistance (q_t), vertical total stress (σ_v), hydrostatic pore pressure (u_0), pore pressure at the cone tip (u_1), the pore pressure just above the cone base (u_2), and OCR. The paper has the following aims:

- To investigate the feasibility of MARS and LSSVM for determination of OCR from PCPT data.
- To develop equations based on the MARS and LSSVM models for prediction of OCR.
- To determine the error bar of the predicted OCR based on the developed LSSVM model.
- To make a comparative study between the developed MARS, LSSVM and other available methods for prediction of OCR.

2 Details of MARS

MARS is a method that is used for fitting the relationship between a set of input and output variables (Friedman 1991). It is built by taking the form of an expansion in product spline basis functions, where the number of basis functions as well as the parameters associated with each one is automatically determined by the data. The approach is analogous to the use of splines. The general form of a MARS predictor is as follows:

$$y = \beta_0 + \sum_{j=1}^{P}\sum_{b=1}^{B}\left[\beta_{jb}(+)\max\left(0, x_j - H_{bj}\right) + \beta_{jb}(-)\max\left(0, H_{bj} - x_j\right)\right] \quad (1)$$

For P predictor variables and B basis function. The basis functions $\max(0, x - H)$ and $\max(0, H - x)$ are univariate and do not have to each be present if their β coefficients are 0. The H values are called "hinges" or "knots". The MARS algorithm consists of (i) a forward stepwise algorithm to select certain spline basis functions, (ii) a backward stepwise algorithm to delete basis functions until the "best" set is found, and (iii) a smoothing method which gives the final MARS approximation a certain degree of continuity.

This article adopts the above methodology for prediction of OCR based on PCPT data. The dataset consist the magnitude of q_t, σ_v, u_0, u_1, u_2 and OCR from 202 test sites around the world. The input parameters of MARS are q_t, σ_v, u_0, u_1, and u_2. OCR is considered as output of MARS. The data are normalized between 0 and 1. The following formula has been adopted for normalization.

$$d_{\text{normalized}} = \frac{(d - d_{\min})}{(d_{\max} - d_{\min})} \quad (2)$$

where d is any data (input or output), $d_{\min}$ is minimum value of the entire dataset, $d_{\max}$ is maximum value of the entire dataset, and $d_{\text{normalized}}$ is normalized value of the data. The data have been divided into two sub-sets; a training dataset, to construct the model, and a testing dataset to estimate the model performance. In this study 137 data points were considered for training and the remaining 65 data points were considered for testing. The programing of MARS was carried using MATLAB.

3 Details of LSSVM

Least squares support vector machine models are an alternate formulation of SVM regression (Vapnik 1998) proposed by Suykens et al. (2002). Consider a given training set of N data points $\{x_k, y_k\}_{k=1}^{N}$ with input data $x_k \in R^N$ and output $y_k \in r$ where R^N is the N-dimensional vector space and r is the one-dimensional vector space. The five input variables of the LSSVM model are q_t, σ_v, u_0, u_1, and u_2. The output of the LSSVM model is OCR. So, in this study, $x = [q_t, \sigma_v, u_0, u_1, u_2]$ and $y = \text{OCR}$. In feature space, LSSVM models take the form

$$y(x) = w^{\mathrm{T}}\varphi(x) + b \quad (3)$$

where the nonlinear mapping $\varphi(.)$ maps the input data into a higher dimensional feature space; $w \in R^n$; $b \in r$; w an adjustable weight vector, b the scalar threshold, R^n the n dimensional vector space and r is one dimensional vector space. In LSSVM for function estimation, the following optimization problem is formulated:

$$\text{Minimize: } \frac{1}{2}w^{\mathrm{T}}w + \gamma\frac{1}{2}\sum_{k=1}^{N}e_k^2$$
$$\text{Subject to: } y(x) = w^{\mathrm{T}}\varphi(x_k) + b + e_k, \quad k = 1, \ldots, N. \quad (4)$$

where e_k is error variable and γ is regularization parameter.

The following equation for OCR prediction has been obtained by solving the above optimization problem (Vapnik 1998; Smola and Scholkopf 1998).

$$\text{OCR} = y(x) = \sum_{k=1}^{N}\alpha_k K(x, x_k) + b \quad (5)$$

The radial basis function has been used in this analysis. The radial basis function is given by:

$$K(x_k, x_l) = \exp\left\{-\frac{(x_k - x_l)(x_k - x_l)^{\mathrm{T}}}{2\sigma^2}\right\}k, \quad l = 1, \ldots, N \tag{6}$$

where σ is the width of radial basis function.

The above LSSVM model has been employed for determination of OCR. The same input, output, training dataset, testing dataset and normalization technique have been used in LSSVM as used in MARS model. The program of LSSVM has been constructed using MATLAB.

4 Results and discussion

Coefficient of correlation (R) has been adapted to asses the performance of the developed MARS and LSSVM models. The value of R has been determined using the following equation:

$$R = \frac{\sum_{i=1}^{n}\left(\mathrm{OCR_{ai}} - \overline{\mathrm{OCR}}_{\mathrm{a}}\right)\left(\mathrm{OCR_{pi}} - \overline{\mathrm{OCR}}_{\mathrm{p}}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathrm{OCR_{ai}} - \overline{\mathrm{OCR}}_{\mathrm{a}}\right)}\sqrt{\sum_{i=1}^{n}\left(\mathrm{OCR_{pi}} - \overline{\mathrm{OCR}}_{\mathrm{p}}\right)}} \tag{7}$$

where $\mathrm{OCR_{ai}}$ and $\mathrm{OCR_{pi}}$ are the actual and predicted OCR values, respectively, $\overline{\mathrm{OCR}}_{\mathrm{a}}$ and $\overline{\mathrm{OCR}}_{\mathrm{p}}$ are mean of actual and predicted OCR values corresponding to n patterns. For a predictive model of high accuracy, the value of R should be close to one. First, the forward stepwise procedure was carried out to select 15 basis functions (BF) to build the MARS model. This was followed by the backward stepwise procedure to remove redundant basis functions. The final model includes 11 basis functions, which are listed in Table 1 together with their corresponding equations. The final equation for the prediction of OCR based on MARS model is given below:

Table 1 The basis functions (BF) and the corresponding equation in the MARS model

Basis function	Equation
BF1	$\max(0,\ 0.359 - u_2)$
BF2	$\mathrm{BF1} \times \max(0,\ u_1 - 0.0214)$
BF3	$\mathrm{BF1} \times \max(0,\ 0.0214 - u_1)$
BF4	$\mathrm{BF1} \times \max(0,\ q_t - 0.0320)$
BF5	$\mathrm{BF1} \times \max(0,\ 0.0320 - q_t)$
BF6	$\mathrm{BF1} \times \max(0,\ 0.177 - u_1)$
BF7	$\mathrm{BF6} \times \max(0,\ 0.035 - u_2)$
BF8	$\mathrm{BF1} \times \max(0,\ u_0 - 0.0390)$
BF9	$\mathrm{BF1} \times \max(0,\ 0.0390 - u_0)$
BF10	$\max(0,\ \sigma_{\mathrm{v}} - 0.086)$
BF11	$\max(0,\ 0.086 - \sigma_{\mathrm{v}})$

$$\begin{aligned}\mathrm{OCR} = {} & 0.215 + 25.213 \times \mathrm{BF1} - 32.0380 \times \mathrm{BF2} \\ & + 189.809 \times \mathrm{BF3} - 32.793 \times \mathrm{BF4} + 153.679 \\ & \times \mathrm{BF5} - 1.025 \times \mathrm{BF6} - 140.294 \times \mathrm{BF7} - 7.117 \\ & \times \mathrm{BF8} + 24.269 \times \mathrm{BF9} - 0.349 \times \mathrm{BF10} \\ & - 0.614 \times \mathrm{BF11}\end{aligned} \tag{8}$$

The above Eq. (8) has been used to determine the performance of training and testing data points. Figure 1 presents scatter plot of the actual OCR and predicted OCR by the MARS model for training dataset. Figure 2 depicts the performance of testing dataset for MARS model. The R values are quite close to one for training as well as testing data points. Therefore, the developed MARS model has the capability to predict OCR.

For LSSVM model, the design values of γ and σ have been determined by trial and error approach. The design values of γ and σ are 100 and 3, respectively. The performance of training and testing dataset has been computed using the design values of γ and σ. Figures 1 and 2 illustrate the performance of training and testing dataset, respectively. It is observed that the value of R is close to one for LSSVM model. So, the developed LSSVM model can predict OCR reasonably well. By substituting $K(x, x_k) = \exp\left\{-\frac{(x_k - x)(x_k - x)^{\mathrm{T}}}{2\sigma^2}\right\}$, $\sigma = 3$, $b = 1.5713$, and $N = 137$ in Eq. 5, the following equation has also been developed for the prediction of OCR based on the developed LSSVM model.

$$\mathrm{OCR} = \sum_{k=1}^{137}\alpha_k \exp\left\{-\frac{(x_k - x)(x_k - x)^{\mathrm{T}}}{18}\right\} + 1.5713 \tag{9}$$

The values of α have been given in Fig. 3. The developed LSSVM model has also been used to compute the error bar of training and testing data points. Figures 4 and 5 illustrate the 95% error bar of training and testing dataset, respectively. The obtained error bar can be used to determine uncertainty.

A comparative study has been carried out between MARS, LSSVM and other traditional methods (Sully et al. 1988; Mayne 1991; Chen and Mayne 1994; Tumay et al. 1995; Kurup and Dudani 2002; Samui et al. 2008) for prediction of OCR. Comparison has been done in terms of root mean square error (RMSE) and mean absolute error (MAE). RMSE and MAE have been computed using the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\mathrm{OCR_{ai}} - \mathrm{OCR_{pi}}\right)^2}{n}} \tag{10}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|\mathrm{OCR_{ai}} - \mathrm{OCR_{pi}}\right|}{n} \tag{11}$$

Figures 6 and 7 present the bar chart of RMSE and MAE for the different models, respectively. In terms of

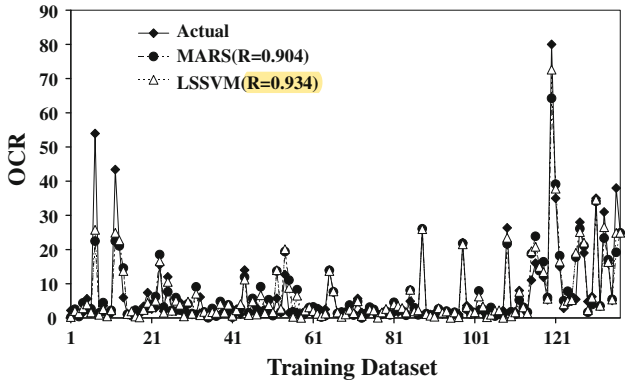

Fig. 1 Performance of MARS and LSSVM for training dataset

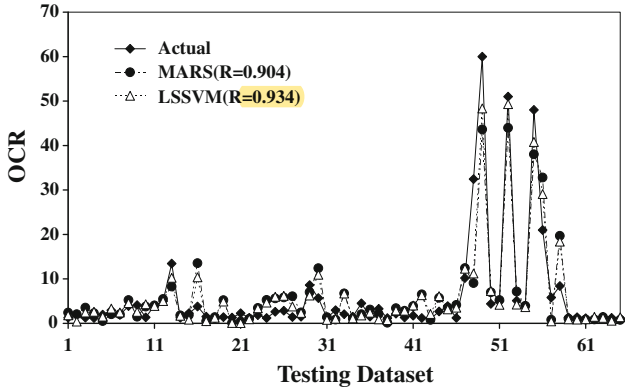

Fig. 2 Performance of MARS and LSSVM models for testing dataset

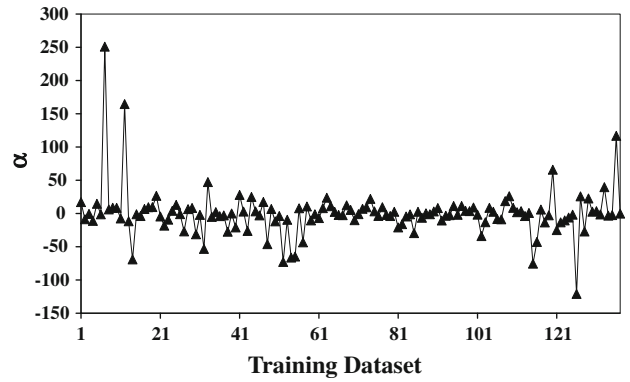

Fig. 3 Values of α for LSSVM model

RMSE, the performance of MARS and LSSVM models is better than the SVM model. Whereas, in terms of MAE, the performance of the SVM model is slightly better than the developed LSSVM and MARS models. The RMSE gives more attention on large errors than small error (Hecht-Nielsen 1990). In contrast, the MAE eliminates the importance given to large errors. The developed LSSVM

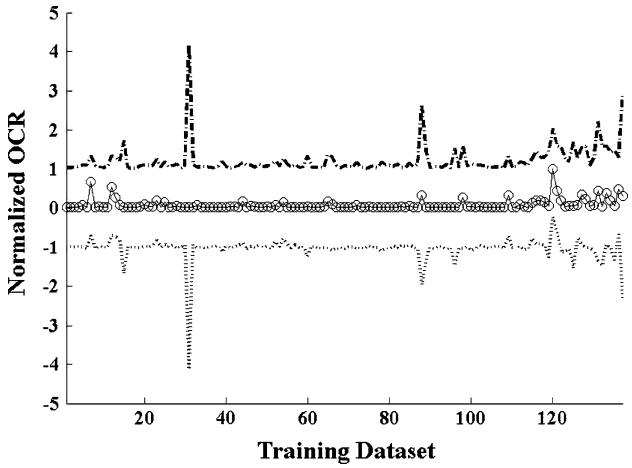

Fig. 4 95% error bar for training dataset

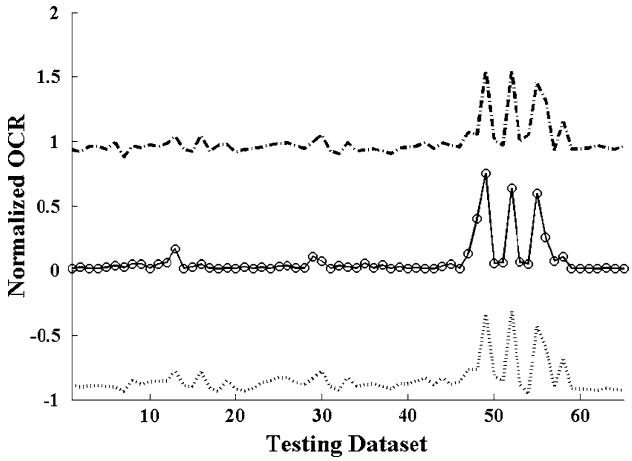

Fig. 5 95% error bar for testing dataset

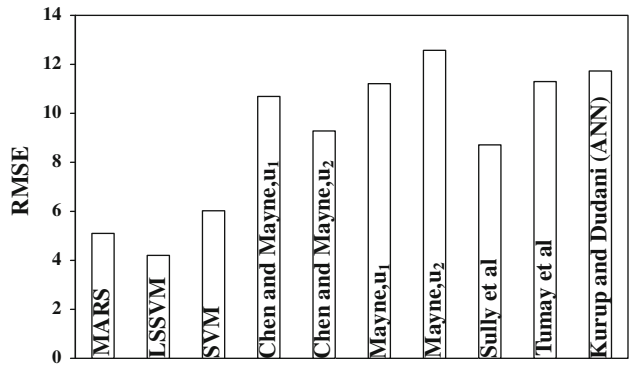

Fig. 6 Values of RMSE for different models

uses only two parameters (γ and σ). Whereas, SVM uses three parameters (C, ε, and σ). The overall performance of LSSVM is better than the MARS model.

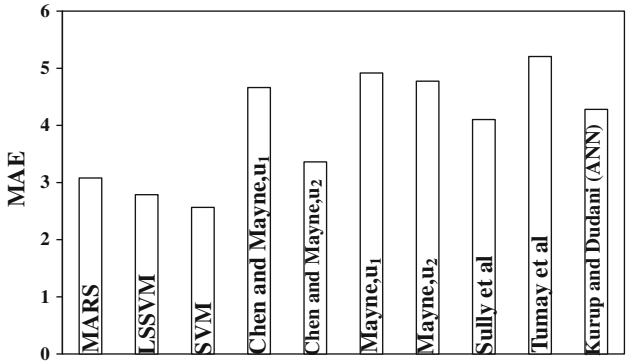

Fig. 7 Values of MAE for different models

5 Conclusion

The potential of the MARS and LSSVM models in predicting OCR from PCPT data was critically examined. Both the models produce encouraging result. The developed MARS and LSSVM models outperform the traditional methods (except SVM) for OCR determination. The overall performance of LSSVM is better than the MARS model. Practitioners can use the developed equations for the determination of OCR. The predicted error bar can be used to determine the corresponding uncertainty and risk. It is concluded that the developed MARS and LSSVM techniques are effective tools for prediction of OCR from PCPT data.

References

Celik S, Tan O (2005) Determination of pre-consolidation pressure with artificial neural network. Civ Eng Environ Syst 22(4):217–231

Chan WT, Chow YK, Liu LF (1995) Neural network: an alternative to pile driving formulas. Comput Geotech 17:135–156

Chen BSY, Mayne PW (1994) Profiling the OCR of clays by piezocone tests, Rep. No. CEEGEO-94-1, Georgia Institute of Technology, Atlanta: pp 280

Das S, Samui P, Sabat AK, Sitharam TG (2010) Prediction of swelling pressure of soil using artificial intelligence techniques. Environ Earth Sci 61(2):393–403

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–141

Goh ATC (1994) Nonlinear modelling in geotechnical engineering using neural networks. Aust Civ Eng Trans CE36(4):293–297

Goh ATC, Goh SH (2007) Support vector machines: their use in geotechnical engineering as illustrated using seismic liquefaction data. Comput Geotech 34(5):410–421

Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading

Kecman V (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. The MIT press, Cambridge

Kumar B, Samui P (2008) Application of ANN for predicting pore water pressure response in a shake table test. Int J Geotech Eng 2(2):153–160

Kurup PU, Dudani NK (2002) Neural network for profiling stress history of clays from PCPT data. J Geotech Geoenviron Eng 128(7):569–579

Lee IM, Lee JH (1996) Prediction of pile bearing capacity using artificial neural networks. Comput Geotech 18(3):189–200

Mayne PW (1991) Determination of OCR in clays by PCPT using cavity expansion and critical state concepts. Soils Found 31(2):65–76

Nawari NO, Liang R, Nusairat J (1999) Artificial intelligence techniques for the design and analysis of deep foundations. Electronic J Geotech Eng. http://geotech.civeng.okstate.edu/ejge/ppr9909/index.html

Pal M (2006) Support vector machines-based modelling of seismic liquefaction potential. Int J Numer Anal Methods Geomech 30(10):983–996

Park D, Rilett LR (1999) Forecasting freeway link travel times with a multi-layer feed forward neural network. Comp Aided Civ Infra Struct Eng 14:358–367

Rahman MS, Wang J, Deng W, Carter JP (2001) A neural network model for the uplift capacity of suction cassions. Comput Geotech 28(4):269–287

Samui P (2008) Support vector machine applied to settlement of shallow foundations on cohesionless soils. Comput Geotech 35(3):419–427

Samui P, Kurup P, Sitharam TG (2008) OCR prediction using support vector machine based on piezocone data. J Geotech Geoenviron Eng 134(6):894–898

Smola A, Scholkopf B (1998) On a kernel based method for pattern recognition, regression, approximation and operator inversion. Algorithmica 22:211–231

Sully JP, Campanella RG, Robertson PK (1988) Overconsolidation ratio of clays from penetration pore pressures. J Geotech Eng 114(2):209–216

Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300

Suykens JAK, Lukas L, Van Dooren P, De Moor B, Vandewalle J (1999) Least squares support vector machine classifiers: a large scale algorithm. In: Proceedings of the European Conference on Circuit Theory and Design (ECCTD'99), Stresa, pp 839–842

Suykens JAK, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation. Neurocomputing 48(1–4):85–105

Systems Salford (2001) MARSTM User Guide. Salford Systems, San Diego

Tumay MT, Kurup PU, Voyiadjis GZ (1995) Profiling OCR and Ko from piezocone penetration tests. In: Proceedings of International Symposium on Cone Penetration Testing, vol 95, SGF Report No. 3, Swedish Geotechnical Society, Linkoping, pp 337–342

Vapnik VN (1998) Statistical learning theory. Wiley, New York