Optimizing the prediction accuracy of load-settlement behavior of single pile using a self-learning data mining approach

Doddy Prayogo[1,*], *Yudas Tadeus Teddy* Susanto[1,2]

[1]Petra Christian University, Dept. of Civil Engineering, Jalan Siwalankerto 121-131, Surabaya 60236, Indonesia
[2]PT. Sarana Data Persada, Margorejo Indah XIX/35-Blok D-507, Surabaya, Indonesia

Abstract. Pile foundations usually are used when the upper soil layers are soft clay and, hence, unable to support the structures' loads. Piles are needed to carry these loads deep into the hard soil layer. Therefore, the safety and stability of pile-supported structures depends on the behavior of the piles. Additionally, an accurate prediction of the piles' behavior is very important to ensure satisfactory performance of the structures. Although many methods in the literature estimate the settlement of the piles both theoretically and experimentally, methods for comprehensively predicting the load-settlement of piles are very limited. This study develops a new data mining approach called self-learning support vector machine (SL-SVM) to predict the load-settlement behavior of single piles. SL-SVM performance is investigated using 446 training data points and 53 test data points of cone penetration test (CPT) data obtained from the previous literature. The actual prediction accuracy is then compared to other prediction methods using three statistical measurements, including mean absolute error (MAE), coefficient of correlation (R), and root mean square error (RMSE). The obtained results show that SL-SVM achieves better accuracy than does LS-SVM and BPNN. This confirms the capability of the proposed data mining method to model the accurate load-settlement behavior of single piles through CPT data. The paper proposes beneficial insights for geotechnical engineers involved in estimating pile behavior.

1 Introduction

Pile foundations are usually used to transmit the axial load from upper structures to the hard soil layer. At times, a pile foundation can be more advantageous than a shallow foundation due to the cost-effectiveness of its construction [1-3]. One important aspect in the design of the pile foundation is the evaluation of its load-settlement. Poulos and Davis [4] showed that the elastic settlement of the pile makes a major contribution to the total settlement. Especially in pile on sand, the elastic settlement is almost as much as the total settlement. Usually, the elastic settlement is analyzed using the semi-empirical method.

Although many methods in geotechnical engineering predict the pile's settlement, both theoretical and experimental methods of thoroughly predicting the load-settlement of the pile are very limited. In the civil engineering world, data mining techniques have become an important research area. Several studies have shown the advantages of data mining technique in producing better prediction models than traditional methods [5,6]. Shahin [7] developed an artificial neural network (ANN) model to predict the load-settlement of a steel pile using a recurrent neural network (RNN). This RNN model had been calibrated using 23 in situ, full-scale load tests, as well as cone penetration test (CPT) data. Even though the RNN model from Shahin [7] showed good results, this model was derived from limited data, i.e. 23 full-scale load tests. In addition, the Shahin model is focused on steel driven piles and has only one input parameter to calculate the variation of the soil strength along the pile shaft, i.e. the mean value of cone resistance of CPT, q_c.

Lately, the least squares support vector machine (LS-SVM) has become one of the most prominent data mining techniques used to solve a complex problem in the world [8,9]. Although LS-SVM has produced more accurate prediction results, an incorrect tuning parameter can reduce the accuracy of LS-SVM. The objective of this study is to improve the accuracy of the prediction model using parameter optimization. Identifying the most optimal parameters is an optimization problem. Therefore, the latest studies integrate a machine learning technique with a metaheuristic-based optimization tool

* Corresponding author: prayogo@petra.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

instead of using only a machine learning technique [10-13]. This study introduces a new hybrid data mining model called the self-learning support vector machine (SL-SVM) to accurately predict the individual pile behavior in test records. Tests were conducted directly in the field and took into account various types of soil, several types of pile, and various geotechnical problems commonly encountered in the field. The hybrid approach used by SL-SVM combines techniques from SOS and LS-SVM. SOS is used to optimize the γ and σ parameters of LS-SVM; then LS-SVM creates an improved input-output relationship from a dataset by performing a supervised-learning-based predictor.

In this study, 499 test records were obtained from the previous literature. The proposed SL-SVM model can fully predict the load-settlement behavior of concrete, steel, and composite piles, as well as bored or driven piles. To accurately model the non-uniformity of the soil along the pile shaft, the length of the embedded pile is divided into 5 segments of equal length. In each segment, the mean value of q_c and shaft friction of CPT (f_s) are calculated.

2 Methodology

2.1 Regression model: LS-SVM

LS-SVM was first developed by [8] as an improved version of the support vector machine (SVM). As a data mining technique, LS-SVM has been successfully applied in many civil engineering-related problems [14-17]. LS-SVM utilizes a cost function based on the least squares principle as opposed to the quadratic loss function that had been used in the original SVM [18]. The objective function and constraints for minimizing the cost function of LS-SVM are shown as follows:

$$\text{Minimize } J_p(w,e) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 \quad (1)$$

$$\text{Subjected to } y_k = w^T\phi(x_k) + b + e_k,\ k = 1,...,N \quad (2)$$

where γ is a regularization constant, e_k denotes the error variable, and x_k and y_k are the input and output data points of the given training dataset of N data points.

For function estimation, the following equation expressed the LS-SVM model:

$$y(x) = \sum_{k=1}^{N}\alpha_k K(x_k, x_l) + b \quad (3)$$

where α_k and b represent the solutions to the linear system.

This study employed the radial basis function (RBF) kernel with the following formula:

$$K(x_k, x_l) = \exp\left(-\frac{\|x_k - x_l\|^2}{2\sigma^2}\right) \quad (4)$$

where σ denotes the kernel function parameter.

2.2 Optimization algorithm: SOS

Initially developed by Cheng and Prayogo [10], the SOS algorithm took its inspiration from the symbiotic interactions among a group of organisms. Its initial application was to solve continuous optimization problems [10] and it has been used to solve various problems in multiple disciplines [19-27]. SOS utilized nature-inspired operators – the mutualism phase, commensalism phase, and parasitism phase – to guide the organisms (solutions) to the global optima region (best solution).

In the "mutualism phase," each organism is modified as follows:

$$new\_O_i = O_i + U(0,1) \times [O_{best} - (1 + round(rand(0,1)) \times (O_i + O_j)/2] \quad (5)$$

$$new\_O_j = O_j + U(0,1) \times [O_{best} - (1 + round(rand(0,1)) \times (O_i + O_j)/2] \quad (6)$$

where O_i and O_j denote the i-th and j-th organism vectors, respectively, such that $i \neq j$; $U(0,1)$ denotes the uniform random numbers between 0 and 1; O_{best} represents the best organism; and new_O_i and new_O_j are the generated candidate solutions after O_i and O_j perform the interaction.

In the "commensalism phase," each organism is modified as follows:

$$new\_O_i = O_i + U(-1,1) \times (O_{best} - O_j) \quad (7)$$

where $U(-1,1)$ denotes the uniform random numbers between –1 and 1.

In the "parasitism phase," each organism is modified as follows:

$$O_{par} = F \times O_i + (1 - F) \times (U(0,1) \times (ub - lb) + lb) \quad (8)$$

where O_{par} denotes the parasite that attempts to eliminate the host O_j; ub and lb represent the lower and upper bounds of the given problem, respectively; and F and $(1 - F)$ are the binary random matrix and its inverse, respectively.

2.3 SL-SVM system integration

In this study, two different forms of artificial intelligence (AI), which are SOS and LS-SVM, are combined to form a new hybrid data-mining technique called SL-SVM. The relationship between the input and output variables of a given set of data is accurately mapped out through the LS-SVM that has a key role as a predictor. The SOS is utilized to find the most suitable LS-SVM parameters γ and σ.

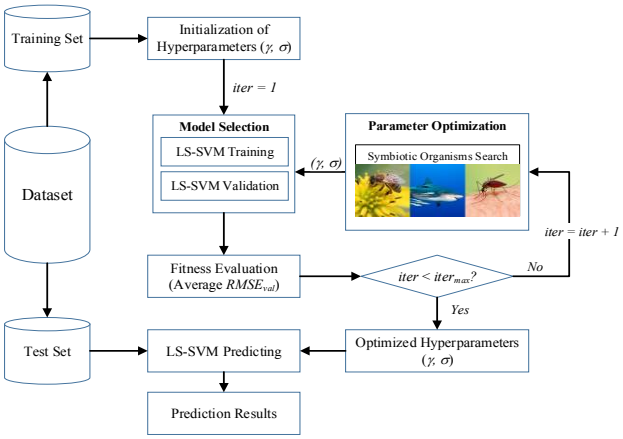

Fig. 1. Flow-chart of SL-SVM.

The architecture of SL-SVM is shown in Fig. 1. Throughout these test phases and training, the six main steps of the SL-SVM are conducted and are as delineated below:

(1) Dataset: The dataset is usually grouped into a test set and a training set. Furthermore, the datasets were scaled into a (0,1) range [28] to curb circumstances in which one or some of the input variables are dominant over others.

(2) Hyperparameters' initialization: Using the formula written below in the first iteration, the parameters are randomly initialized within the boundary range.

$$x = U(0,1) \times (ub - lb) + lb \tag{9}$$

where x represents the candidate solution (hyperparameters of LS-SVM).

(3) Model selection: This is a critical step in building an accurate learning model. Utilizing the initial hyperparameters and the training set, the LS-SVM model is trained with a key focus on determining the true nature of the relationship between the input and output variables. The training process is conducted in an iterative manner and the tuning parameters from LS-SVM are gradually optimized by utilizing the SOS algorithm. A fitness function that correlates with the accuracy of the prediction model is now developed in the bid to evaluate the accuracy of the learning system. k-fold cross validation, a well-known sampling technique, is incorporated in the fitness function. The dataset is now grouped into k-folds in which the $(k - 1)/k$ part of the given dataset is assigned to training and the remaining part is assigned to validating the trained model.

Thus, a k sets of training and validation subset are formed and carried out for model selection. For measuring the model accuracy, the root mean square error (RMSE) is selected as the fitness function, as shown in the following equation:

$$fit\_val = \frac{\sum_{k=1}^{S} RMSE_{\text{val}}}{S} \tag{10}$$

where fit_val is a fitness value calculated from RMSE between the predicted output and actual output from the validation subset and S is the total number of folds.

(4) SOS for parameter search: To identify the best set of these hyperparameters, the hybrid AI system utilizes SOS to explore various simulations of γ and σ. Through the generation of the initial population, the search process commences. The initial population, however, serves as the initial candidate for the hyperparameters searched. SOS uses the parasitism, commensalism, and mutualism phases for each iteration to gradually bring about improvement in the fitness value of every candidate solution present in the population.

(5) Optimal hyperparameters: When the stopping criterion is met, the loop stops. This implies that the prediction model has identified the input–output mapping relationship with optimal γ and σ parameters.

(6) LS-SVM predicting: To predict the test set, the prediction model must be established. Thus, the given training phase brought about the optimal LS-SVM γ and σ parameters that were utilized to establish the prediction model.

3 Data preprocessing

Four fundamental parameters are used in many established methods to predict the load-settlement behavior of single pile. These main parameters are: the geometry of the pile, material properties of the pile, soil properties, and load applied to the pile. In addition to the main parameters are several extra parameters, such as: the pile installation method and load test type, as well as whether the pile tip is open or closed. The geometry of the pile, material properties of the pile, and load applied to the pile are easy to quantify and identify. However, soil properties are tricky to quantify and identify.

Table 1. Statistical description of the dataset.

Attributes	Unit	Min	Max	Avg	Std
X_1: Type of load test	1: Maintained load, 2: Constant rate of penetration				
X_2: Material properties of the pile	1: Concrete, 2: Steel, 3: Composite				
X_3: Pile installation method	1: Bored, 2: Driven				
X_4: End of pile	1: Closed, 2: Open				
X_5: Axial rigidity of the pile	MN	796.74	33106.3	11459.8	11680.0
X_6: Cross-sectional area of the pile	cm^2	100.00	7854.00	3411.78	2638.14
X_7: Perimeter of the pile	cm^2	58.50	957.56	320.31	280.47
X_8: Pile length	m	5.50	56.39	21.84	13.63
X_9: Embedded length of the pile	m	5.50	45.00	18.03	10.39
X_{10}: q_{c1}	MPa	0.00	10.38	3.57	2.64
X_{11}: f_{s1}	KPa	0.00	273.91	59.11	52.29
X_{12}: q_{c2}	MPa	0.05	17.16	4.73	3.51
X_{13}: f_{s2}	KPa	1.83	275.50	75.59	63.24
X_{14}: q_{c3}	MPa	0.30	31.54	6.18	6.55
X_{15}: f_{s3}	KPa	1.62	618.67	90.95	97.18
X_{16}: q_{c4}	MPa	0.25	33.37	8.52	7.72
X_{17}: f_{s4}	KPa	4.42	1292.67	200.31	215.31
X_{18}: q_{c5}	MPa	0.25	53.82	10.54	10.30
X_{19}: f_{s5}	KPa	7.99	559.00	139.86	144.50
X_{20}: q_c at the end of the pile	MPa	0.25	70.29	13.40	13.02
X_{21}: load applied to the pile	KN	0.00	30000.0	2585.01	3652.62
Y: Pile settlement	mm	0.00	137.88	10.57	16.14

In this study, the dataset is derived from load tests which comprised 499 data points, obtained from Pooya Nejad and Jaksa [29]. In the literature, CPT is used to quantify and identify soil properties. The 499 data points are divided into 446 training data points and 53 test data points. To accurately model the non-uniformity of the soil along the pile shaft, the length of the embedded pile is divided into 5 segments of equal length. In each segment, the mean value of q_c and f_s are calculated. Finally, the attributes of the dataset are shown in Table 1 alongside the statistical description of the dataset.

4 SL-SVM application

4.1 Model selection and training results

This study implements the parameter setting of SOS as follows: ecosystem size = 50 and total iterations = 30. The searching range for the tuning parameters, γ and σ^2, was between 10^{-5} and 10^5. To have a balance between training and validation data points, cross-validation was used. To have a splitting ratio of 2:1 between training and validation, 3-fold cross validation is used. SOS is then performed on the model selection using the 3 sets of training and validation data subsets. The fitness value was determined as the average validation errors in the model selection. The model performance in the training process is shown in Fig. 2. The optimal hyperparameters found by SOS were as follows: final γ = 28.9507 and final σ^2 = 0.0547 with the fitness value of 10.5514 mm.

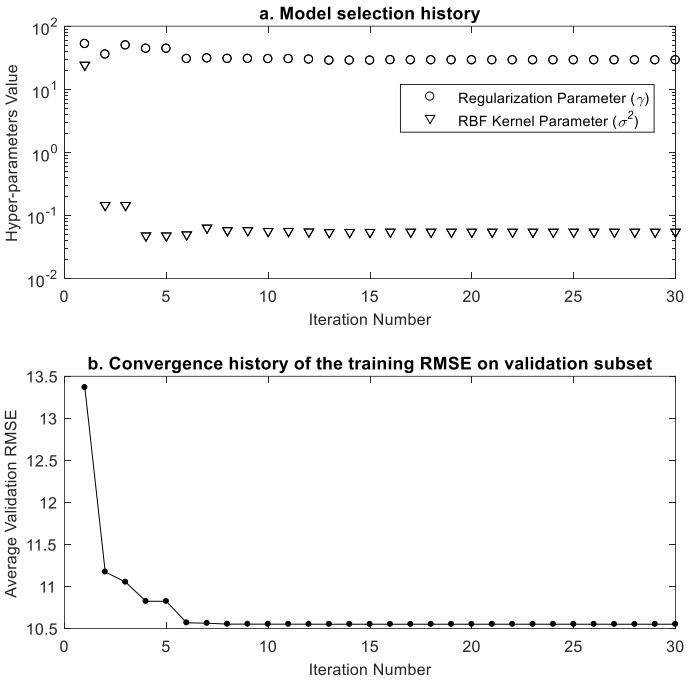

Fig. 2. Model selection process and convergence history of the training RMSE.

4.2 Prediction results

The accuracy of the training and test results between the predicted output (y') and actual output (y) of n data points can be compared using three metrics: correlation coefficient (R), root mean square error (RMSE), and mean absolute error (MAE). Each metric can be expressed as shown in Table 2.

Table 2. Performance metrics for measuring prediction results.

Performance Metrics	Formula
R	$\frac{n\sum y \times y' - (\sum y)(\sum y')}{\sqrt{n(\sum y^2) - (\sum y)^2}\sqrt{n(\sum y'^2) - (\sum y')^2}}$
RMSE	$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - y_j')^2}$
MAE	$\frac{1}{n}\sum_{j=1}^{n}\lvert y_j - y_j'\rvert$

The developed SL-SVM was validated and compared to other predictive models, including the original LS-SVM and back-propagation neural network (BPNN). The comparison between SL-SVM and other predictive algorithms may indicate the advantages of using the optimization method to tune the optimal parameters. BPNN settings included: learning rate = 1, maximum hidden layers = 1, and number of neurons in the hidden layer = 21 (following the total input variables). Finally, the LS-SVM parameters for γ and σ^2 were set to 1 as suggested in [8].

The experimental results between the proposed method and other prediction method are shown in Table 3. It is shown that the SL-SVM model outperformed LS-SVM and BPNN in all performance metrics. The SL-SVM produces the best value in R, RMSE, and MAE. Meanwhile, Fig. 3 further illustrates the actual and predicted settlement of the developed model in both the training and test datasets.

Table 3. Training and test performance of SL-SVM and other methods.

AI methods	Training		
	R	RMSE (mm)	MAE (mm)
BPNN	0.7264	10.4378	6.602
LS-SVM	0.7354	12.0865	6.9043
SL-SVM	0.9513	5.2468	2.5634
AI methods	Test		
	R	RMSE (mm)	MAE (mm)
BPNN	0.7876	8.5783	6.0188
LS-SVM	0.5501	7.4134	5.2907
SL-SVM	0.7523	7.0517	4.7118

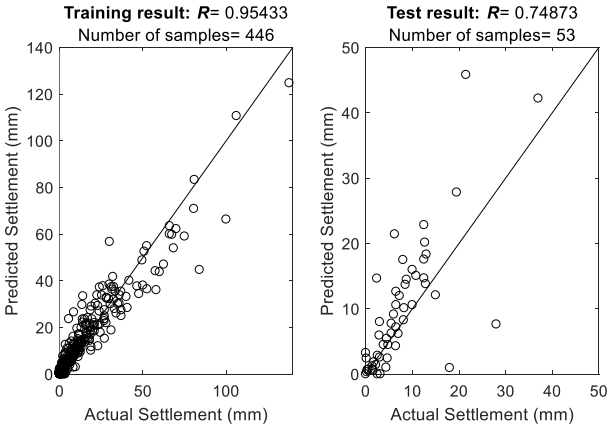

Fig. 3. Actual and predicted settlement of SL-SVM in training and test datasets.

5 Conclusions

In this study, we propose an automatic-tuning data mining technique called the self-learning least squares support vector machine (SL-SVM) to predict the settlement of a single pile. The experimental dataset was acquired from previous literature that contained 499 samples of load tests. Three performance metrics were utilized to assess the proposed data mining technique and a comparison with various predictive techniques was conducted. The result indicates that the most accurate prediction model is the proposed SL-SVM. The SL-SVM is able to outperform the original LS-SVM due to the SOS's success in searching for the most suitable LS-SVM parameters. This established data mining technique, SL-SVM, can potentially help geotechnical engineers model pile behavior for pile design. The trained model can model pile settlement with higher accuracy in comparison to other predictive techniques.

This work was supported by Petra Christian University under the internal research grant scheme no. 03/HB-PENELITIAN/LPPM-UKP/I/2018.

References

1. U. Smoltczyk, Geotechnical Engineering Handbook, Procedures, John Wiley & Sons,. ISBN:3433014507 (2003)
2. D. Tjandra, Indarto, R.A.A. Soemitro, Behavior of expansive soil under water content variation and its impact to adhesion factor on friction capacity of pile foundation, International Journal of Applied Engineering Research **10**, 38913-38917 (2015) [in Indonesian]
3. D. Tjandra, Indarto, R.A.A. Soemitro, Effect of drying-wetting process on friction capacity and adhesion factor of pile foundation in clayey soil, Jurnal Teknologi **77**, 145-150 (2015) [in Indonesian]
4. H.G. Poulos, E.H. Davis, Pile foundation analysis and design, ISBN:0471020842 (1980)
5. J.-S. Chou, C.-K. Chiu, M. Farfoura, I. Al-Taharwa, Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques, J. Comput. Civ. Eng. **25** , 242-253 (2011)
6. S.-H. Liao, P.-H. Chu, P.-Y. Hsiao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications **39** , 11303-11311 (2012)
7. M.A. Shahin, Load–settlement modeling of axially loaded steel driven piles using CPT-based recurrent neural networks, Soils and Foundations **54**, 515-522 (2014)
8. J.A.K. Suykens, J. Vandewalle, Least Squares Support Vector Machine Classifiers, Neural Process. Lett. **9**, 293-300 (1999)
9. P. Samui, Least square support vector machine and relevance vector machine for evaluating seismic liquefaction potential using SPT, Natural Hazards **59**, 811-822 (2011)

10. M.-Y. Cheng, P.M. Firdausi, D. Prayogo, High-performance concrete compressive strength prediction using Genetic Weighted Pyramid Operation Tree (GWPOT), Engineering Applications of Artificial Intelligence **29**, 104-113 (2014)
11. M.-Y. Cheng, D. Prayogo, Y.-W. Wu, Novel Genetic Algorithm-Based Evolutionary Support Vector Machine for Optimizing High-Performance Concrete Mixture, Journal of Computing in Civil Engineering **28**, 06014003 (2014)
12. M.-Y. Cheng, D.K. Wibowo, D. Prayogo, A.F.V. Roy, Predicting productivity loss caused by change orders using the evolutionary fuzzy support vector machine inference model, Journal of Civil Engineering and Management **21**, 881-892 (2015)
13. M.-Y. Cheng, D. Prayogo, Modeling the permanent deformation behavior of asphalt mixtures using a novel hybrid computational intelligence, ISARC 2016 - 33rd International Symposium on Automation and Robotics in Construction, International Association for Automation and Robotics in Construction, Auburn, USA, 1009-1015 (2016)
14. D. Prayogo, M.Y. Cheng, J. Widjaja, H. Ongkowijoyo, H. Prayogo, *Prediction of concrete compressive strength from early age test result using an advanced metaheuristic-based machine learning technique* ISARC 2017 - Proceedings of the 34th International Symposium on Automation and Robotics in Construction (2017)
15. M.-Y. Cheng, D. Prayogo, Y.-W. Wu, Prediction of permanent deformation in asphalt pavements using a novel symbiotic organisms search–least squares support vector regression, Neural Comput. Appl. (2018)
16. D. Prayogo, Metaheuristic-Based Machine Learning System for Prediction of Compressive Strength based on Concrete Mixture Properties and Early-Age Strength Test Results, Civil Engineering Dimension **20**, 21-29 (2018)
17. D. Prayogo, Y.T.T. Susanto, Optimizing the Prediction Accuracy of Friction Capacity of Driven Piles in Cohesive Soil Using a Novel Self-Tuning Least Squares Support Vector Machine, Adv. Civ. Eng. **2018** (2018)
18. K.S. Kulkrni, D.-K. Kim, S.K. Sekar, P. Samui, Model of Least Square Support Vector Machine (LSSVM) for Prediction of Fracture Parameters of Concrete, International Journal of Concrete Structures and Materials **5**, 29-33 (2011)
19. D.-H. Tran, M.-Y. Cheng, D. Prayogo, A novel Multiple Objective Symbiotic Organisms Search (MOSOS) for time–cost–labor utilization tradeoff problem, Knowledge-Based Systems **94**, 132-145 (2016)
20. M.-Y. Cheng, D. Prayogo, D.-H. Tran, Optimizing Multiple-Resources Leveling in Multiple Projects Using Discrete Symbiotic Organisms Search, J. Comput. Civ. Eng. **30**, 04015036 (2016)
21. D. Prayogo, M.-Y. Cheng, H. Prayogo, A Novel Implementation of Nature-inspired Optimization for Civil Engineering: A Comparative Study of Symbiotic Organisms Search, Civil Engineering Dimension **19**, 36-43 (2017)
22. V.F. Yu, A.A.N.P. Redi, C.-L. Yang, E. Ruskartina, B. Santosa, Symbiotic organisms search and two solution representations for solving the capacitated vehicle routing problem, Applied Soft Computing **52**, 657-672 (2017)
23. G.G. Tejani, V.J. Savsani, V.K. Patel, Adaptive symbiotic organisms search (SOS) algorithm for structural design optimization, J. Comput. Des. Eng. **3**, 226-249 (2016)
24. G.G. Tejani, V.J. Savsani, S. Bureerat, V.K. Patel, Topology and Size Optimization of Trusses with Static and Dynamic Bounds by Modified Symbiotic Organisms Search, J. Comput. Civ. Eng. **32**, 04017085 (2018)
25. G.G. Tejani, V.J. Savsani, V.K. Patel, S. Mirjalili, Truss optimization with natural frequency bounds using improved symbiotic organisms search, Knowl.-Based Syst. **143**, 162-178 (2018)
26. D. Prayogo, M.-Y. Cheng, F.T. Wong, D. Tjandra, D.-H. Tran, Optimization model for construction project resource leveling using a novel modified symbiotic organisms search, Asian Journal of Civil Engineering (2018)
27. D. Prayogo, R.A. Gosno, R. Evander, S. Limanto, Implementasi Metode Metaheuristik Symbiotic Organisms Search Dalam Penentuan Tata Letak Fasilitas Proyek Konstruksi Berdasarkan Jarak Tempuh Pekerja, Jurnal Teknik Industri **19**, 103-114 (2018) [in Indonesian]
28. C.W. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector classification (2003)
29. F. Pooya Nejad, M.B. Jaksa, Load-settlement behavior modeling of single piles using artificial neural networks and CPT data, Computers and Geotechnics **89**, 9-21 (2017)