EFFECT OF FEATURE SELECTION ON CLOUD METAHEURISTIC MACHINE LEARNING

[1]JUI-SHENG CHOU, [2]LUCIA DEWI SANTOSO

[1,2]National Taiwan University of Science and Technology
E-mail: [1]jschou@mail.ntust.edu.tw, [2]lucia.94ds@gmail.com

Abstract - Data mining approaches and their associated optimization algorithms have made big data more attractive because they can be used to extract global information that supports the construction of reliable prediction models. We developed a novel and interactive cloud-computing tool, called the Cloud Metaheuristic Machine Learning System, which enables universal access for mining data through the internet without the need of any computer software. The web-based computing system can be trained to solve many kinds of regression problem, such as the prediction of the compressive strength of concretes. However, to increase predictive accuracy and computing efficiency for engineering purposes, the most relevant information should be selected. Irrelevant variables or features must be eliminated to yield the most useful result. Therefore, this work seeks to determine significant features that can be used to predict the outputs and to improve the accuracy of the previous full-mode prediction model without feature selection. In the preprocessing stage, stepwise regression is used to delete redundant or insignificant input variables. The result of the case study herein concerning the compressive strength of selected multi-country concretes after a standard curing of 28-days demonstrates that the current model outperforms the full-mode model with respect to accuracy and computing time.

Index Terms - Civil Engineering Application, Construction Material, Feature Selection, Machine Learning, Web-based Analytics.

I. INTRODUCTION

Data mining is used to uncover hidden information and to construct non-trivial knowledge from seemly useless data by revealing important relationships, patterns, and trends.

Most concrete specimens are tested after a standard curing of 28 days. This conventional method wastes time, money, and the natural sources that are used to generate the trial mixes. Reference [1] discovered that conventional methods of forecasting the 28-day compressive strength of concrete are based upon statistical analysis in which many linear and nonlinear regression equations are developed to model the problem. Owing to the shortcomings of compressive tests in the laboratory, a means of accurately predicting concrete compressive strength is favored.

Researchers and scholars have carried out many investigations of the prediction of concrete compressive strength. Reference [2] used a combined classification and regression technique to predict the compressive strength of high-performance concrete. An ensemble model of artificial intelligence that was presented by [3] yields performance 4.2- 69.7% better error rates than earlier prediction models in determining the compressive strength of high-performance concrete compared to the previous prediction models. Reference [4] also examined the prediction of 28-day compressive strength; they generated 150 different concrete mixes in the laboratory and developed an initial real model for ANN prediction in MATLAB Platform.

In particular, we developed a novel and interactive online-computing tool, called the Cloud Metaheuristic Machine Learning System (CMMLS), which enables universal access to data mining through the internet. The CMMLS does not have the computational complexity of standalone applications. CMMLS uses a machine learning (ML) technique, called Least Squares Support Vector Regression (LSSVR), to solve non-linear regression problems. An optimization algorithm (OA)-the Smart Firefly Algorithm-is utilized to improve the hyperparameters in LSSVR, yielding more accurate predictions.

The multi-dimensional input variables not only result in the high computational cost of training, but also reduce the accuracy of the results because of the "curse of dimensionality"; a higher input dimensionality corresponds to a need for more data, therefore, slower its computation [5]. A combination of SFA and LSSVR, in fact, leads to a longer computing time. The LSSVR model in CMMLS does not have this issue, but its accuracy is questioned. This study introduces a dimensionality reduction method, feature selection using stepwise regression to improve the quality of CMMLS predictions in terms of computational performance.

II. LITERATURE REVIEW

A. Reduction of Dimensionality

Redundant features in high dimensional data significantly increase the computational complexity and the memory storage requirements of the input variables [6]. Dimensionality can be reduced using two approaches, which are feature extraction and feature selection. Feature extraction transforms the

components of original features while feature selection cuts out only unnecessary features without any feature conversion.

Methods of feature selection include embedding, wrapping, and filtering. Wrapper methods select features based on the predictive performance of a predefined classifier. However, the wrapper approach is largely avoided because it consumes extensive computation resources and has a high risk of overfitting [7]. Embedded methods use ML models for classification, and a classifier to optimize the subset of features. The embedded approaches give more complexity to the classifier. Indeed, to improve the performance of the computation is not easy [8].

Notably, filter methods, which are done before the classifier runs, are the most efficient [6]. Since the filter model applies independent evaluation criteria without any mining algorithm, it does not inherit any of the bias of a mining algorithm, thus, computes efficiently [9].

B. Least Squared Vector Regression

Suykens et al. [10] proposed a well-developed machine learning (ML) technique with many advanced features that support a high generalization capacity and fast computation, called Least Squares Support Vector Regression (LSSVR). LSSVR uses a least square formulation to solve regression problems. LSSVR is more effective than Support Vector Regression [11]. However, the predictive accuracy of LSSVR depends strongly on the determination of its hyperparameters, the regularization parameter (C) and the sigma of the RBF kernel (σ).

C. Optimization Algorithm

The CMMLS includes machine learning, LSSVR, and an optimization algorithm, the Smart Firefly Algorithm (SFA), to analyze model parameters and prevent over-fitting and local minima problems. Reference [12] extended the Firefly Algorithm [13] by adding metaheuristic components, which were the logistic map, the chaotic Gauss or mouse map, adaptive inertia weight, and Lévy flight, to develop the SFA. The purpose of this refined OA is to find the optimal hyperparameters, C and σ. It is suggested to read on [11] to gain an advance understanding of this optimization algorithm.

III. METHODOLOGY

A. Cloud Metaheuristic Machine Learning

The CMMLS comprises the Single Machine Learning Model (SML) and the Optimized Machine Learning Model (OML). An input form on a Java Server Page (JSP) will allow users to import their datasets which must be in text format (.TXT) and to set the input parameters for analysis. Then, the website sends it to Apache Tomcat and performs the analysis in the core system (JAR). Ultimately, after the predictions are made and the performance measures have been calculated, the Servlet collects the outcome and transfers it to the website output page. The output page will display numerical results, which are the prediction values and the value of objective function (R, RMSE, MAE, MAPE), and graphical results.

For the OML model, graphs of performance trajectory and tracing paths are presented. The CMML solve a diverse regression problem over the internet and has many favorable functions, including (a) default setting, (b) export-import, (c) saving detailed outputs, (d) saving the model and its optimal hyperparameters, and (e) providing feedback to the users.

Fig. 1 shows the flowchart of CMMLS, consisting of two different models, SML Model that performs LSSVR and the OML Model, with the SFA as the OA and the LSSVR as the computing engine. Fig. 2 displays the sitemap of CMMLS.

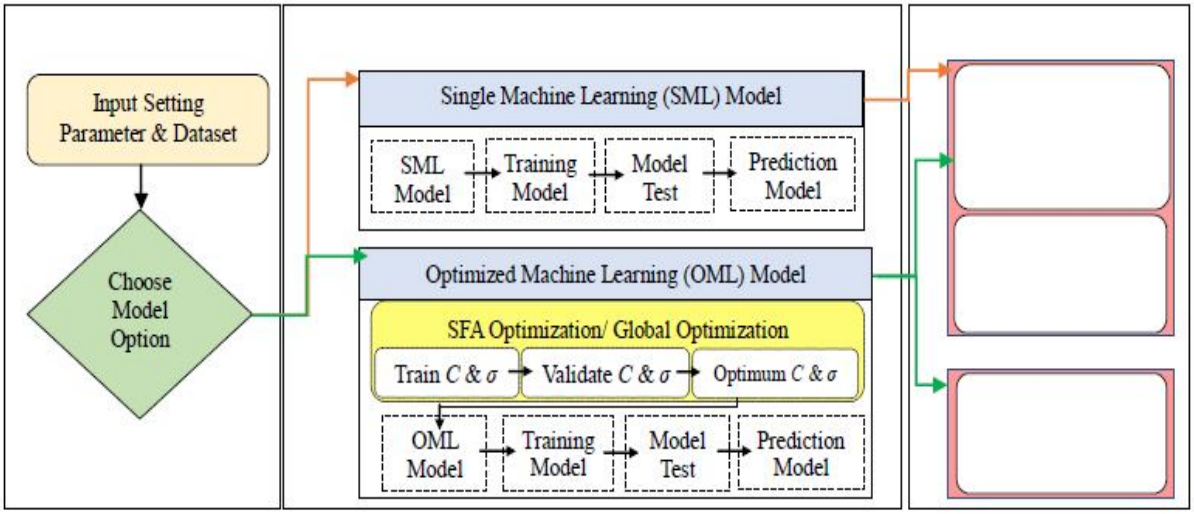

Fig. 1. Cloud environment of CMMLS.

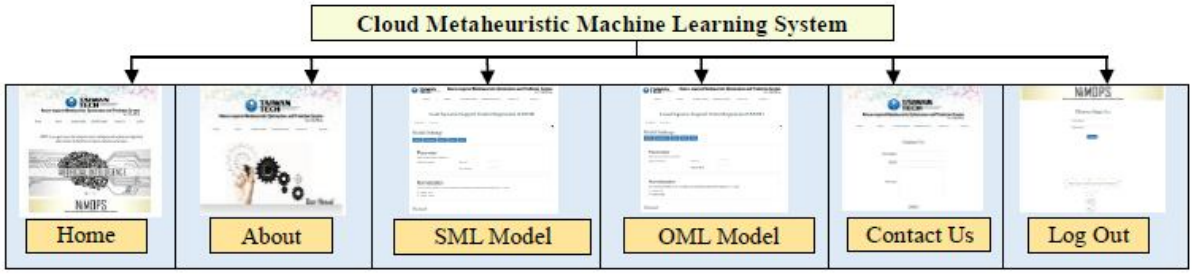

Fig. 2. Sitemap of CMMLS

B. Feature Selection Workflow in CMMLS

The filter method, which involves stepwise regression, is used to select significant predictor attributes and to remove insignificant or redundant input variables. The stepwise regression technique is implemented in IBM SPSS.

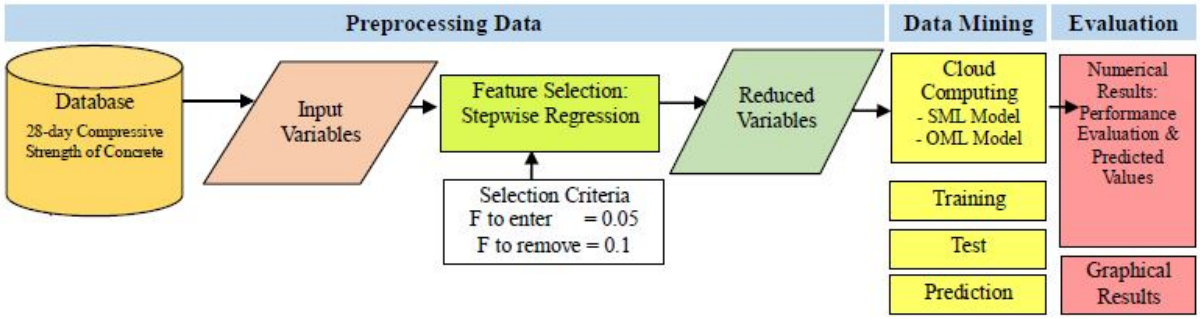

Fig. 3 presents the workflow of feature selection in the first operation of data mining using CMMLS.

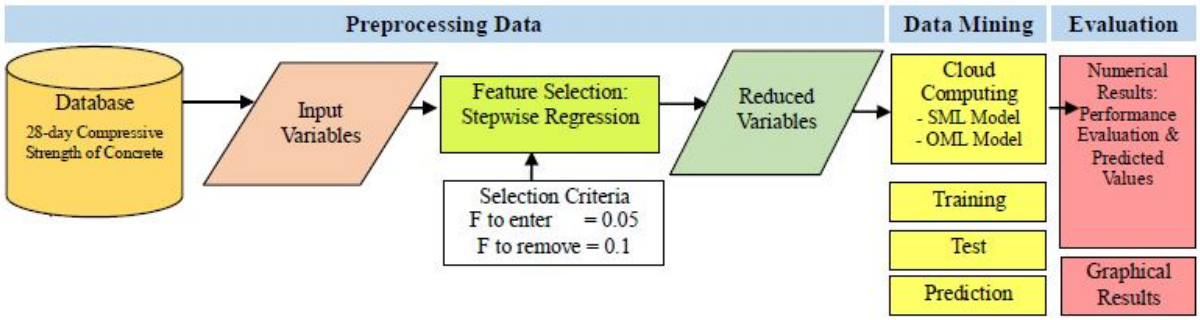

Fig. 3. Proposed workflow.

IV. DATA COLLECTION

Multi-country 28-day compressive strength of concrete data are used in this investigation. The dataset comprises 437 data points from UCI [14], ten data points from Toronto, Canada [15], 103 data points from Taiwan [16], six data points from Australia [17], 18 data points from Mexico [18], 42 data points from Turkey [19], 12 data points from China [20], 13 data points from USA [21], 15 data points from South Australia [22]. Table 1 presents the data attributes and statistical parameters of the dataset.

The objective functions that specify accuracy are the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the correlation coefficient (R), and the mean square error (MSE). These indices measure whether the predicted values are close to the actual values.

Table 1. Statistical parameters of dataset.

Parameter	Unit	Max. Value	Min. Value	Mean	Std. Dev.
Cement	kg/m^3 mixture	540	35	255.9	Input
Blast Furnace Slag (BFS)	kg/m^3 mixture	870	0	76.4	Input
Fly Ash (FA)	kg/m^3 mixture	319	0	85.0	Input
Water	kg/m^3 mixture	247	109	183.7	Input
Superplasticizer	kg/m^3 mixture	32.2	0	6.1	Input
Coarse Aggregate	kg/m^3 mixture	5600	609	1145.6	Input
Fine Aggregate	kg/m^3 mixture	3200	0	854.6	Input
28-day Compressive Strength	MPa	81.8	5	37.4	Output

V. RESULT AND DISCUSSION

A. Stepwise Regression Result

Table 2 shows the result of feature selection by stepwise regression, which reveals that cement, water, blast furnace slag, fine aggregate, and fly ash are the most important variables accounting for 79.5% variance of the 28-day compressive strength of concrete.

Table 2. Stepwise model summary.

Model Summary					Variable Combination in Model
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.625[a]	.391	.390	11.161	a. Predictors: Cement
2	.685[b]	.469	.467	10.434	b. Predictors: Cement, Water
3	.729[c]	.532	.530	9.804	c. Predictors: Cement, Water, BFS
4	.762[d]	.580	.578	9.288	d. Predictors: Cement, Water, BFS, Fine Aggregate.
5	.795[e]	.632	.629	8.708	e. Predictors: Cement, Water, BFS, Fine Aggregate., FA

B. Analytical Results

Table 3 shows ten trials that were carried out using baseline model. Table 3 displays the summary of the results of OML model. The effect of feature selection on SML model is clearly observed as most of the performance measurements of baseline model is improved. The average improvements range from 4% - 21.35 %. Small values of the error performance parameters demonstrate that the values predicted using the model are very close to the actual values. A higher number of R implies a stronger linear correlation relationship in the model. For the OML model, the effect of feature selection only decreased RMSE, R, and the computing time by 24%, 19.4%, and 28%, respectively. Due to this limitation, future investigation can address another advanced feature selection method that can enhance all the performance measurements of OML model.

Table 3. Analytical results using the proposed workflow for the Single Machine Learning Model.

Trial	Before Feature Selection					After Feature Selection					Improvement				
	Performance Measurement					Performance Measurement					Performance Measurement (%)				
	R	RMSE	MAE	MAPE	Time (sec)	R	RMSE	MAE	MAPE	Time (sec)	R	RMSE	MAE	MAPE	Time
1	0.88	8.12	5.12	16.46	2.84	0.93	4.48	4.70	15.24	1.48	5.70	44.85	8.21	7.41	47.96
2	0.86	7.93	5.31	17.23	1.47	0.92	5.83	4.77	15.24	1.35	6.72	26.48	10.21	11.52	7.91
3	0.91	6.08	5.29	17.26	1.84	0.91	6.66	4.67	0.86	1.42	-0.33	-9.46	11.76	95.04	22.83
4	0.88	7.37	5.37	17.83	1.44	0.93	5.37	4.87	16.41	1.56	5.70	27.13	9.31	7.96	-8.64
5	0.88	7.60	5.18	16.69	1.70	0.92	4.66	4.70	15.25	1.35	4.12	38.68	9.30	8.63	20.63
6	0.84	6.11	5.08	15.98	1.40	0.92	6.29	4.91	16.10	1.36	8.65	-2.95	3.41	-0.75	3.35
7	0.91	6.00	5.06	16.24	1.41	0.92	5.98	4.66	15.75	1.37	1.20	0.32	7.85	3.03	2.56
8	0.87	8.59	5.33	18.06	1.73	0.91	5.45	4.75	15.51	1.45	5.04	36.55	10.88	14.12	16.38
9	0.86	8.59	5.34	18.06	1.73	0.90	6.16	4.70	15.61	1.32	4.44	28.29	11.97	13.57	23.88
10	0.91	6.00	5.23	17.45	1.74	0.90	6.06	4.72	15.31	1.42	-1.44	-1.05	9.68	12.26	18.39
Avg.	0.88	7.24	5.23	17.13	1.73	0.92	5.69	4.74	14.13	1.41	**4.00**	**21.35**	**9.29**	**17.51**	**18.65**

Table 4. Analytical results using the proposed workflow for the CMMLS.

Before Feature Selection					After Feature Selection					Percentage of Improvement (%)				
RMSE Test	MAE Test	MAPE Test	R Test	Time (s)	RMSE Test	MAE Test	MAPE Test	R Test	Time (s)	RMSE	MAE	MAPE	R	Time
13.66	5.79	0.18	0.63	6468	11.02	8.77	0.27	0.75	4666	↑ 24.0	↓ 51.4	↓ 55.4	↑ 19.4	↑ 28

CONCLUSION

To avoid the misinterpretation of predictions of 28-day compressive strength owing to redundant variables in the dataset, feature selection was implemented to select the most significant variables and to remove the less important ones. Stepwise regression method was applied herein to a customized web-based system optimized by a metaheuristic machine learning technique. The results obtained in the preliminary analysis indicate feature selection can improve the performance of cloud computing without losing adequate prediction accuracy.

REFERENCES

[1] L. Snell, V. R. J., and W. ND, "Predicting Early Concrete Strength," presented at the Concrete Int., 1 December, 1989.

[2] J.-S. Chou and C.-F. Tsai, "Concrete Compressive Strength Analysis Using a Combined Classification and Regression Technique," Automation in Construction, vol. 24, pp. 52-60, 2012. Elsevier

[3] J.-S. Chou and A.-D. Pham, "Enhanced Artificial Intelligence

for Ensemble Approach to Predicting High Performance Concrete Compressive Strength," Construction and Building Materials, vol. 49, pp. 554-563, December 2013.
[4] F. Khademi and J. Sayed Mohammadmehdi, "Predicting the 28 Days Compressive Strength of Concrete Using Artificial Neural Network," (in English), i-Manager's Journal on Civil Engineering, vol. 6, no. 2, pp. 1-7, 2016.
[5] E. Keogh and A. Mueen, "Curse of Dimensionality," in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010.
[6] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature Selection Considering Two Types of Feature Relevancy and Feature Interdependency," Expert Systems with Applications, vol. 93, pp. 423-434, 1 March 2018.
[7] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent Advances and Emerging Challenges of Feature Selection in the Context of Big Data," Knowledge-Based Systems, vol. 86, pp. 33-45, 2015.
[8] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay, "Pareto Front Feature Selection Based on Artificial Bee Colony Optimization," Information Sciences, vol. 422, pp. 462-479, January 2018.
[9] L. Huan and Y. Lei, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, pp. 491-502, 7 March 2005. IEEE
[10] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, Least squares support vector machines. Singapore: World Scientific Publishing Co. Pte. Ltd, Singapore, 2002.
[11] J.-S. Chou, W. K. Chong, and D.-K. Bui, "Nature-Inspired Metaheuristic Regression System: Programming and Implementation for Civil Engineering Applications," Journal of Computing in Civil Engineering, vol. 30, no. 5 September 2016.
[12] J.-S. Chou and A.-D. Pham, "Smart Artificial Firefly Colony Algorithm-Based Support Vector Regression for Enhanced Forecasting in Civil Engineering," Computer-Aided Civil and Infrastructure Engineering, vol. 30, no. 9, pp. 715–732, September 2015.
[13] H. Gandomi, X. S. Yang, S. Talatahari, and A. H. Alavi, "Firefly Algorithm with Chaos," Communications in Nonlinear Science and Numerical Simulation, vol. 18, no. 1, pp. 89-98, January 2013.
[14] I.-C. Yeh, "Modeling of Strength of High Performance Concrete Using Artificial Neural Networks," Cement and Concrete Research, vol. 28, pp. 1797-1808, December 1998.
[15] N. Bouzoubaâ and M. Lachemi, "Self-Compacting Concrete Incorporating High Volumes of Class F Fly Ash: Preliminary Results," Cement and Concrete Research, vol. 31, no. 3, pp. 413-420, 1 March 2001.
[16] I.-C. Yeh, "Modeling Slump Flow of Concrete Using Second-Order Regressions and Artificial Neural Networks," Cement and Concrete Composites, vol. 29, no. 6, pp. 474-480, July 2007.
[17] K. Komnitsas, D. Zaharaki, and V. Perdikatsis, "Effect of Synthesis Parameters on the Compressive Strength of Low-Calcium Ferronickel Slag Inorganic Polymers," Journal of Hazardous Materials, vol. 161, no. 2-3, pp. 760-768, 30 January 2009.
[18] Durán-Herrera, C. A. Juárez, P. Valdez, and D. P. Bentz, "Evaluation of Sustainable High-Volume Fly Ash Concretes," Cement and Concrete Composites, vol. 33, no. 1, pp. 39-45, 1 January 2011.
[19] M. Sumer, "Compressive Strength and Sulfate Resistance Properties of Concretes Containing Class F and Class C Fly Ashes," Construction and Building Materials, vol. 34, no. Supplement C, pp. 531-536, 1 September 2012.
[20] W. Wang, C. Lu, Y. Li, G. Yuan, and Q. Li, "Effects of Stress and High Temperature on the Carbonation Resistance of Fly Ash Concrete," Construction and Building Materials, vol. 138, pp. 486-495, 1 May 2017.
[21] E. G. Moffatt, M. D. A. Thomas, and A. Fahim, "Performance of High-Volume Fly Ash Concrete in Marine Environment," Cement and Concrete Research, vol. 102, no. Supplement C, pp. 127-135, 1 December 2017.
[22] Gholampour and T. Ozbakkaloglu, "Performance of Sustainable Concretes Containing Very High Volume Class-F Fly Ash and Ground Granulated Blast Furnace Slag," Journal of Cleaner Production, vol. 62, pp. 1407-1417, 20 September 2017.

★★★