Scientific Journal Impact Factor (SJIF) : 3.518

International Journal of Modern Trends in Engineering and Research

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS) HEURISTIC MODEL: APPLICATION OF HEAVY METAL PREDICTION

Iman Hussein AL-Qinani [1]

[1] *Department of Computer Science, College of Education, Al -Mustansiriyah University*

Abstract—In the last two decades, soft computing modeling such as Artificial Intelligence (AI) approaches have gained a massive attention by the information technology researchers. Nowadays, AI models are improving human abilities in several areas of engineering and science problems. In this paper, we investigate the proficiency of modern heuristic approach called Multivariate adaptive regression splines (MARS) in prediction regression problem. The experimental data set of heavy metal is selected as a case study. The predictive model is conducted based on several physiochemical inputs variables. In order to inspect the MARS model accuracy, Multiple Linear Regression (MLR) model is chosen to compare the results. Here, the root mean square error (RMSE), mean absolute percentage error, and coefficient of determination are examined the accuracy of the models. The finding of the investigated model (i.e., MARS) exhibited a noticeable improving in the prediction accuracy in comparison with MLR.

Keywords— Multivariate adaptive regression splines; Multiple Linear Regression; heavy metal; predictive metal.

I. INTRODUCTION

Artificial intelligence is a modern approach that is inherited a fixable mathematical structure in which capable to identify the high non-linearity, non-stationarity, and randomness relationship between input and output in comparison with the classical modeling approaches[5,6,17,26].Within the scope of the AI society, the phrase of machine learning is generally understood as a several data-driven models including; fuzzy set [9], support vector machine [25], artificial neural network [21], and evolutionary computing [17]. All the forgoing heuristic models have been used in the literature to solve both classification and regression problems. Those data-driven models are capable to solve complex problems in different categories of science and technology fields. Zadeh (1994) has coined the term of soft computing and defined it as “collection of methodologies that aim to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost” [16]. Therefore, soft computing should be seen as a partnership of distinct methods, rather than as a homogeneous body of concepts and techniques.

Speaking within the scope of the application, water quality is the backbone and the significant concern of the environmental engineering [2]. Heavy metals have serious negative effects on the water quality indexing due to the presence of the toxicity [3]. Studying the behavior of heavy metals in the aquatic system is very important for numerous prospective of environmental engineers. For instance, set up an alarm system to measure the toxicity in water, reduce the threats to the human life as well as the water creatures. So far, heavy metals prediction accomplished with minority researches in the literature using soft computing. Rooki et al. (2011) [20] predicted heavy metals parameters using three types of intelligent model algorithm namely, back propagation neural network, generalized regression neural network and multiple linear regression. The selected case study was in acid mine drainage. The finding was encouraging the back propagation neural network over the other algorithms in modeling heavy metal. Aryafar et al. (2012) [1] investigated the application of support vector machine (SVM) in predicting four heavy metals parameters including Manganese, Copper, Lead and Iron. The accuracies were compared with generalized regression neural network for evaluation. The results obtained in this research indicated that SVM model can be

DOI:10.21884/IJMTER.2016.3027.7NUQV

applied in monitoring ground and surface water quality. Most recently, Elzwayie et al. (2016) applied the radial basis function neural network (RBF) in predicting heavy metals concentration in two different climate zones tropical and arid [3]. Authors concluded that RBF can sufficiently use in this scope of research.

Multivariate adaptive regression splines (MARS) is a relatively modern artificial intelligence approach that proposed by (Friedman 1991) [7]. This approach inherits several features such as the capacity to handle the natural complication of the data mapping in high-dimensional data patterns, fast and bendable model, and perform the prediction of continuous and binary output variables accurately. Furthermore, it distinguishes itself by flexible procedure in which organizing the relationship between the inputs/output parameters with reducing the variable interactions [12]. In the light of the literature, several studies proved the successfulness of MARS algorithm in different types of applications [4, 10, 11, 12, 13]. We present in this study, the use of MARS algorithm in heavy metal prediction as our originality.

This research investigates the capability of MARS model in regression problem. Heavy metal laboratory samples collected in monthly time scale measurement to be implemented as a case study. Several inputs parameters including physiochemical and atmospheric are used in this modeling. In order to diagnostic and evaluate the accuracy of the investigated model, MLR model is conducted for comparison purposes. Root mean square error (RMSE), mean absolute error (MAE), and determination coefficient (R) are used to assess the outcome accuracies. The rest of the article organized in this order. Section 2 presents the methodology including models, data base, and performance indicators. In section 3, application and analysis are discussed. Finally, the conclusion and remarks are drawn.

II. THEORETICAL BACKGROUND AND CASE STUDY

2.1 Multivariate Adaptive Regression Splines (MARS)

In 1991, Friedman proposed MARS model as a nonparametric heuristic model that used in solving regression problems (i.e., forecasting) [7]. MARS model main advantage is the forward and backward stepwise procedure that can controls and explains the complex nonlinear mapping between the inputs and output variables [29]. The feature of the backward stepwise procedure is to remove the unnecessary input candidates from the previous selected data set in order to enhance the forecasting accuracy. The new output variable Y is forecasted in accordance to the input variables via either of the two basis functions, using a knot or value of variable that defines the inflection point along the inputs range [10, 23, 24]:

$$Y = \max\ (0, X - c) \qquad (1)$$

$$Y = \max\ (0, c - X) \qquad (2)$$

Where the c parameter donates the threshold value. There are two adjacent splines intersect at a knot, in order to maintain the continuity of the basis functions. The function is used in the forward and backward stepwise procedure to each input parameter is to identify the precise location of knots where the function value changes. Great to mention, MARS model is a data-driven process that gained popularity in time series analysis, most recently. In addition, it is even better to explore its capability to improve heavy metal prediction models. Authors recommend the following references for the reader to refer for more comprehensive details of MARS model [7, 24, 27].

2.2 Multiple Linear Regression (MLR)

For the purpose of comparison, MLR model was selected. In a MLR model with *N*-data points, the relationship is examined between the input variables (x_1, x_2. . . x_N) and the output variable (y_1, y_2…, y_N) as [15]:

$$\hat{y}_t = \omega_0 + \omega_1 x_{t,1} + \omega_2 x_{t,2} + \ldots + \omega_N x_{N,d} + e_t \qquad (3)$$

Here, $\hat{y}_t$ = predicted variable, e_t = error (noise) term, $X_{N,d}$ = input value of d^{th} predictor variable in month t, ω_o = regression constant, ω_d = coefficient of d^{th} predictor variable and d = number of predictor variables.

2.3 The Collected Data Set: Case Study

Heavy metal data set “targeted parameter” is modeled based on the physiochemical parameters (i.e., total suspended solids (TSS), chloride (CL), Phosphate (PO_4), and sulfate (SO_4)) and atmospheric parameters (i.e., rainfall (R), weather temperature (T), and humidity (H)) “predictors parameters”. Those physiochemical parameters were obtained from the laboratory test of Darbandikhan Lake, which is located in Kurdistan, Northern of Iraq [14]. This lake fed by Tanjero and Sirwan Rivers, that are flowing from north/northwest of Iraq and Iran, respectively. There are several benefits of this lake including source of water for irrigation and agriculture purposes, drinking water, producing electricity via Darbandikhan Dam, as well as beautiful site fused by many for recreation. As a matter of fact, this lake is currently facing a high risk due to the sources of sewage system pollutions and municipal wastes. Thus, undertaken this research can give a reliable mathematical model for prediction the environmental health of this lake.

2.4 Data Set Division and Model Configuration

Both the physiochemical, heavy metal, and atmospheric parameters were obtained in the form of monthly time scale for the period of (2002-2009). 75 percentage of the data set was used for training the algorithm, 10 percentage of the data set was used to validate the algorithm. Whereas, 15 percentages were used to test the algorithm modeling. MARS algorithm use the piecewise linear spline functions to fit the examined data pattern, Figure 1 illustrates the example of MARS algorithm.

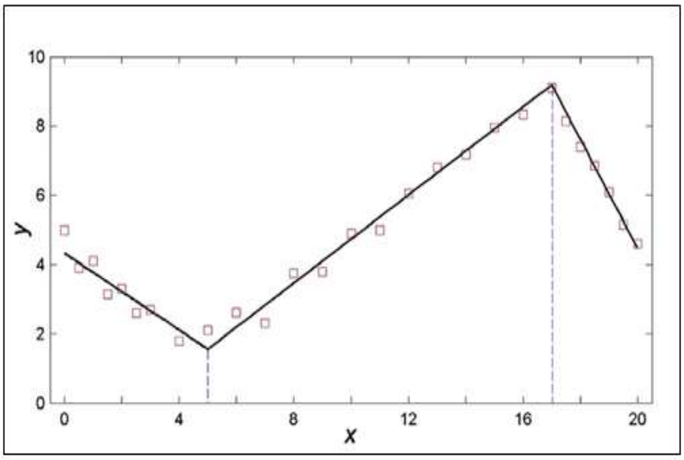

Figure1. Knots and linear splines of MARS algorithm [27].

2.5 Diagnostic Indicators

Two different types of performance indicators are inspected the modeling results including absolute error and best-fit-goodness criteria. Those quantitative metrics are root mean square error (RMSE), mean absolute error (MAE), and determination coefficient (R^2). The statistic measures are [19] formulated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(E_a - E_p)^2} \qquad (4)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n} |E_a - E_p| \qquad (5)$$

$$R = \frac{\sum_{t=1}^{n}\left[(E_a - \overline{E}_a)(E_p - \overline{E}_p)\right]}{\sqrt{\sum_{t=1}^{n}(E_a - \overline{E}_a)^2 \sum_{t=1}^{n}(E_p - \overline{E}_p)^2}} \qquad (6)$$

Where n is the number of the actual data. $E_a, \overline{E_a}, E_p$ are the observed, average observed and predicted values of heavy metal variable. *R or r : is* Correlation Coefficient

III. APPLICATION AND ANALYSIS

In this section, a detailed discussion has been demonstrated the application of multivariate adaptive regression splines model in prediction heavy metal concentration. Laboratory chemical parameters tests of Darbandikhan Lake was selected as the case study to demonstrate the prediction model. In order to present the performance of the proposed model, the results of MARS model have been compared with MLR model. Table 1 demonstrates the statistical performance criteria in order to judge the effectiveness of the predictive modeling in addition to reveal the level of the accuracies improvement. This accuracy improvement was computed through:

$$\text{accuracy improvement} = MARS - MLR/MARS$$

MARS model absolute error indicators accuracies including (i.e., RMSE and MAE) were increased by 47.9 and 33.5 % comparing with MLR model, as tabulated in table 1. Whereas, the best goodness indicator (i.e., correlation coefficient (r)) was increased using MARS model up to 35.4 % with comparison to MLR model.

Table 1: Performance criteria including correlation coefficient (r), Root mean square error (RMSE), and mean absolute error (MAE) for MARS and MLR models evaluated for the testing period.

Models	**Correlation** **Coefficient,** r	**Root mean square error** RMSE	**Mean Absolute Error** MAE
MLR	0.51	46.15	38.85
MARS	0.79	23.75	25.83

Another effective inspection carried out which is the scatter plot. It is important to investigated the diversion from the ideal fit line of the linear regression examination between the actual and the predicted values for the testing samples. Scatter plots for MARS and MLR models are displayed in figure 2 and 3, respectively. MARS model indicates a good performance of regression coefficient R^2 in comparison with MLR.

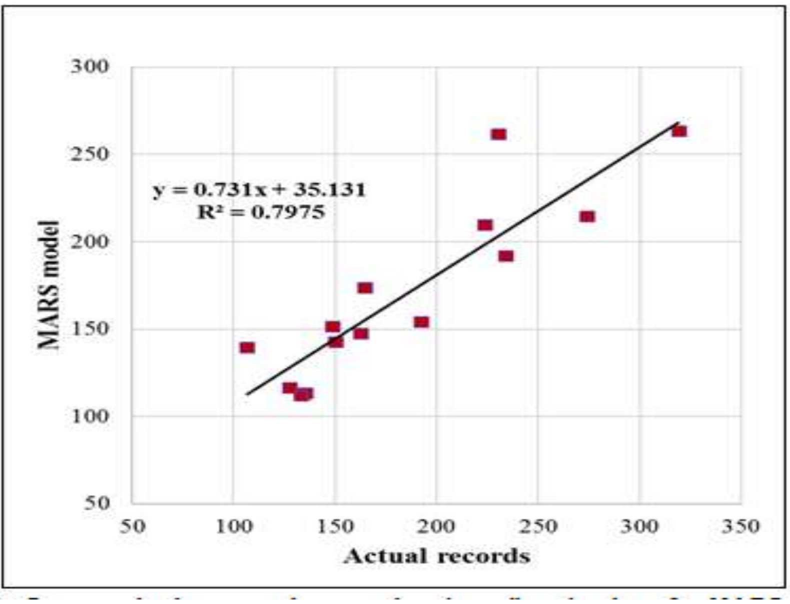

Figure2. Scatter plot between the actual and predicted values for MARS model.

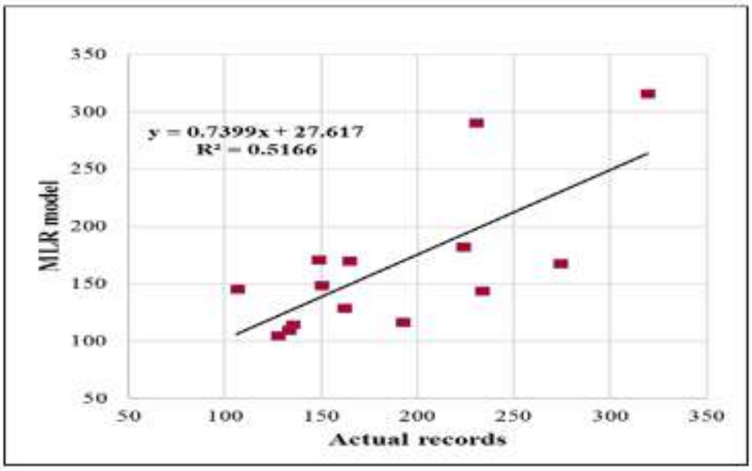

Figure3. Scatter plot between the actual and predicted values for MLR model.

Further assessment worth to be visualized, the general behavior of the prediction of the models in accordance with the actual records of the testing samples. Figure 4 illustrates the phenomena of the predictive model and actual records in order to present the consistency of the modeling.

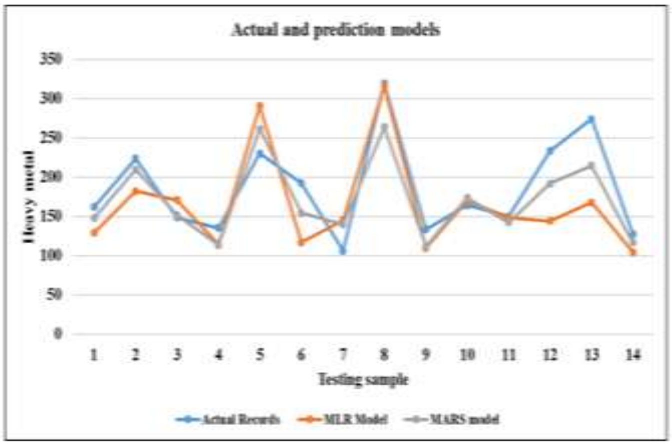

Figure4. Actual and predicted heavy metal concentration using MARS and MLR models.

Several state-of-the-art research are stated that, there is no regression predictive model can be applied for all application under all the circumstances [8, 18, 22, 28]. Proceeding from this prospective, investigating new model that can be applied in several applications and give high level of accuracies, is still ongoing mission for Computer Science. Thus, this research conducted to explore the MARS model in new domain of application and concluded with high level of satisfactory.

IV. CONCLUSION

In the last two decades, soft computing approaches have shown a reliable and robust modeling in term of prediction. In the current research, we studied the applicability of multivariate adaptive regression splines model in prediction problem, precisely in the field of environment and water science. Heavy metal prediction model was accomplished based on two types of inputs variables including physiochemical and atmospheric variables. The results accuracies of MARS model were compared with MLR model for evaluation and assessment. MARS model displayed a reliable accuracies based on the statistical measures for example R, RMSE and MAE. In addition, this modern model exhibited a trustful and robust alternative model that can be applied in this kind of regression problem in the field of environmental engineering. In general, (i) MARS model presents more flexibility than MLR, (ii) it shows the ability to module the prediction data set for heavy metal with limited data, and finally (iii) MARS algorithm tend to has a very good bias-variance trade-off in which is quite flexible to model the high non-linearity and variables interactions.

REFERENCES

[1] A. Aryafar, R. Gholami, R. Rooki, and F. Doulati Ardejani, “Heavy metal pollution assessment using support vector machine in the Shur River, Sarcheshmeh copper mine, Iran,” Environmental Earth Sciences, vol. 67, Issue 4, pp. 1191–1199, October 2012.

[2] A. A. Najah, A. El-Shafie, O. A. Karim, and O. Jaafar, “Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation,” Neural Computing and Applications, vol. 21(5), pp. 833–841, 2012.

[3] A. Elzwayie, A. El-shafie, Z. M. Yaseen, H. A. Afan, and M. F. Allawi, “RBFNN-based model for heavy metal prediction for different climatic and pollution conditions,” Neural Computing and Applications, 2016.

[4] A. Sarangi, and A.K. Bhattacharya, “Comparison of Artificial Neural Network and regression models for sediment loss prediction from Banha watershed in India,” Agricultural Water Management, vol. 78 (3), pp. 195–208, December 2005.

[5] B. Coppin, Artificial Intelligence Illuminated, 1st ed, Sudbury, MA :Jones and Bartlett Publishers, 2014.

[6] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, January 1996.

[7] F. J. H. Friedman,“Multivariate Adaptive Regression Splines,” The Annals of Statistics, vol. 19(1), pp. 1–67, 1991.

[8] G. Huang, G.B. Huang, S. Song, and K. You,” Trends in extreme learning machines: A review”, Neural Networks, vol. 61, pp. 32–48, January 2015.

[9] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, 1st ed, Publisher: Prentice Hall, Upper Saddle River,1995.

[10] J. Adamowski, H. F. Chan, S. O. Prasher, and V. N. Sharda, “Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data,” Journal of Hydroinformatics, vol. 14, Issue 3, pp. 731-744, July 2012.

[11] J. E. Shortridge , S. D. Guikema , and B. F. Zaitchik, “Empirical streamflow simulation for water resource management in data-scarce seasonal watersheds,” Hydrology and Earth System Sciences Discussions, vol. 12 (10), pp. 11083-11127, 2015.

[12] J.R. Leathwick , J. Elith, and T. Hastie,” Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions”, Ecological Modelling, vol. 199(2), pp.188–196, 2006.

[13] K. A. L. Sotomayor, “Comparison of adaptive methods using multivariate regression splines (MARS) and artificial neural networks backpropagation (ANNB) for the forecast of rain and temperatures in the Mantaro river basin,” IGP – Instituto Geofísico del Perú, pp. 58–68, 2010.

[14] K. Ararat, R. A. Mehdi, H. A. Falih, and A. M. Maher, “Darbandikhan Lake Poisoning Event,” Nature Iraq Preliminary Field & Lab Report, Kurdistan, Iraq, September 2008.

[15] K. J. Preacher, P. J. Curran, and D. J. Bauer, “Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis,” Journal of Educational and Behavioral Statistics, vol. 31(4), pp. 437-448, 2006.

[16] L. A. Zadeh, “Soft computing and fuzzy logic,” IEEE Software, vol. 11 (6), pp. 48–56, Nov. 1994.

[17] L. J. Fogel, A. J. Owens, and M. J. Walsh, Artificial Intelligence Through Simulated Evolution, John Wiley & Sons , 1966.

[18] M. S. Hossain and A. El-shafie,” Intelligent Systems in Optimizing Reservoir Operation Policy: A Review”, Water Resources Management, vol. 27 (9), pp. 3387–3407, 2013.

[19] R..l. David, and J. M. Jr. Gregory, "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation," Water Resources Research, vol. 35 (1), pp. 233-241, January 1999.

[20] R. Rooki, F. D. Ardejani, A. Aryafar, and A. B. Asadi, "Prediction of heavy metals in acid mine drainage using artificial neural network from the Shur River of the Sarcheshmeh porphyry copper mine, Southeast Iran," Environmental Earth Sciences, vol. 64(5), pp. 1303–1316, 2011.

[21] S. Haykin, Neural networks: a comprehensive foundation, 2nd ed, Publication: Pearson, Prentice Hall, Pearson Education,1999.

[22] S. Raghavendra, and P. C. Deka, "Support vector machine applications in the field of hydrology: A review," Applied Soft Computing, vol. 19, pp. 372–386, June 2014.

[23] V.N. Sharda, R.M. Patel, S.O. Prasher, P.R. Ojasvi, and C. Prakash, "Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques," Agricultural Water Management, vol. 83 (3), pp. 233-242, 2006.

[24] V.N. Sharda, S.O. Prasher, R.M. Patel, P.R. Ojasvi, and C. Prakash, "Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data," Hydrological Sciences Journal, vol. 53(6), pp. 1165-1175, 2008.

[25] V.N. Vapnik, The Nature of statistical Learning Theory, 2nd ed, Publication Springer-Verlag: Information Science and Statistics, New York, p. 188, 1995.

[26] W. B. Gevarter," Introduction to Artificial Intelligence", Chemical Engineering Progress, vol. 83(9), pp. 21–37, 1987.

[27] W. Zhang,and A.T.C . Goh ,"Multivariate adaptive regression splines and neural network models for prediction of pile drivability," Geoscience Frontiers, vol. 7 (1), pp. 45–52, 2014.

[28] Z.M. Yaseen, A. El-shafie, O . Jaafar , H. A. Afan,and K. N.Say, "Artificial intelligence based models for stream-flow forecasting: 2000–2015," Journal of Hydrology, vol. 530, pp. 829–844, 2015.

[29] Z.M. Yaseen, O . Kisi, and V. Demir, "Enhancing Long-Term Streamflow Forecasting and Predicting using Periodicity Data Component: Application of Artificial Intelligence," Published: Springer in Water Resources Management, vol. 30 (12), pp. 4125–4151, 2016.