TẠP CHÍ KHOA HỌC & CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 05(42) (2020) 20-25
TRƯỜNG ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

A sequential piecewise linear regression model for data analysis developed with Visual C# .NET

Mô hình hồi quy tuyến tính từng phần sử dụng cho phân tích dữ liệu được phát triển với ngôn ngữ C# .NET

Xuan Linh Tran[a,b], Nhat Duc Hoang[a,b]
Trần Xuân Linh[a,b], Hoàng Nhật Đức[a,b]

[a]*Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam*
[a]*Viện Nghiên cứu và Phát triển Công nghệ Cao, Trường Đại học Duy Tân, Đà Nẵng, Việt Nam*
[b]*Faculty of Civil Engineering, Duy Tan University, Da Nang, 550000, Vietnam*
[b]*Khoa Xây dựng, Trường Đại học Duy Tân, Đà Nẵng, Việt Nam*

(Ngày nhận bài: 09/9/2020, ngày phản biện xong: 24/9/2020, ngày chấp nhận đăng: 30/9/2020)

Abstract

This study develops a software program used for nonlinear data analysis based on the Sequential Piecewise Linear Regression (SPLR) [1]. The SPLR is a regression analysis method relying on the concept of hinge function to identify locally linear relationship in datasets. Thus, this method can effectively used to capture nonlinear functional mapping. In this study, the SPRL software program has been developed with Visual C# .NET framework 4.6.1. The usefulness of the newly developed program is verified via several data modeling tasks.

Keywords: Piecewise Linear; Regression Analysis; Visual C# .NET; Machine Learning; Mathematical Modeling.

Tóm tắt

Nghiên cứu này phát triển một chương trình phần mềm được sử dụng để phân tích dữ liệu phi tuyến dựa trên thuật toán Hồi quy tuyến tính từng phần tuần tự (SPLR) [1]. SPLR là một phương pháp phân tích hồi quy dựa trên khái niệm hàm bản lề để xác định mối quan hệ tuyến tính cục bộ trong tập dữ liệu. Do đó, phương pháp này có thể được sử dụng hiệu quả để mô phỏng các mối liên hệ phi tuyến giữa các biến số. Trong nghiên cứu này, một chương trình phần mềm SPRL đã được phát triển với ngôn ngữ Visual C # nền tảng .NET 4.6.1. Chương trình phần mềm mới phát triển được kiểm chứng thông qua một số vấn đề mô hình hóa dữ liệu.

Từ khóa: Tuyến tính từng phần; Phân tích hồi quy; Ngôn ngữ C # .NET; Học máy; Mô hình toán học.

1. Introduction

In civil engineering field, regression analysis is a common method used for analyzing functional mapping between a dependent variable of interest and a set of predicting variables [2-4]. Mathematical models, expressed in the form of mathematical equations, can be utilized to aid civil engineers

* *Corresponding Author:* Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam; Faculty of Civil Engineering, Duy Tan University, Da Nang, 550000, Vietnam
Email: tranxuanlinh@dtu.edu.vn (Xuan Linh Tran); hoangnhatduc@duytan.edu.vn (Nhat Duc Hoang)

in various tasks, e.g. structural design [5, 6], project management [7, 8], concrete mix component design [9-11], etc.

Traditional regression analysis approach basically relied on multiple linear regression models for constructing various mathematical models based on collected datasets [12]. This conventional method has the desired properties of simplicity and transparency. However, the critical assumption of linearity significantly hinders the capability of this method in modeling complex and nonlinear problems. To avoid such hinderance, scholars have resorted to sophisticated models for regression analysis such as artificial neural network, support vector regression, and least squares support vector regression. The advanced methods are highly powerful in data fitting [13-17]. Nevertheless, those sophisticated models have a black-box structure which creates certain difficulties for civil engineers to comprehend and interpret those models' structure.

Another direction to improve the predictive accuracy of the conventional multiple linear regression models is to employ piecewise linear regression models [18-20]. These models assume that the functional mapping of interest can be satisfactorily approximated via locally linear models [21]. By employing the concept of hinge function, piecewise linear regression models can be construct to accurately estimate complex and nonlinear mathematic relationships.

A piecewise linear regression model trained with a sequential algorithm, named as Sequential Piecewise Linear Regression Model (SPLRM), has been put forward in [1] and programmed in MATLAB environment [22]. This study further enhances the applicability of this model via a software program developed with Visual C# .NET framework 4.6.1. The rest of the paper is organized as follows: the second section briefly mentions the formulation of SPMR; two application cases of the newly developed program are demonstrated in the third section; concluding remarks of this paper are stated in the final section.

2. The Used Method for Constructing Piecewise Linear Regression Model

The SPLRM utilizes various linear models to fit subsets of the input data X. Herein, the overall space of X is divided into disjoint regions within which a linear model can be used to describe the relationship between X and a dependent variable Y. The disjoint regions are separated via identification of various knots or break points [23]. Values of knots are sequentially identified and included in a SPLRM structure.

The model structure with one knot is shown as follows [24]:

$$Y(X_i) = \begin{cases} \sum_{d=1}^{D+1} \beta_d X_{i,d} \ \text{if} \ X_{i,d} \le b \\ \sum_{d=1}^{D+1} \beta_d X_{i,d} \ \text{if} \ X_{i,d} > b \end{cases} \quad (1)$$

where X_i denotes the vector of the i^{th} explanatory variable consisting of D elements. b denotes the breaking point value. Y denotes the response variable.

It is noted that the least square method is employed to compute the parameter β of the linear regression model shown in Eq. (1) as follows:

$$\beta^* = (X^T X)^{-1} X^T Y \quad (2)$$

The model with multiple variables and knots can be generally expressed as follows:

$$Y = \sum_{d=1}^{D} \sum_{v1}^{V_d} LF_{d,v}(X_d) \quad (3)$$

where D and V_d denotes the number of predicting variables and the number of hinge function of the d^{th} predicting variable.

In addition to accept or reject a knot candidate, the root mean squared error (RMSE) index is used as follows:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Y_{A,i} - Y_{P,i})}{N}} \quad (4)$$

where $Y_{A,i}$ and $Y_{P,i}$ are the actual and predicted values of Y. N denotes the total number of data samples.

3. Program Applications

To verify the developed SPLRM program, a simple regression analysis problem (Dataset 1), which has 1 break point, is used. The functions used to generate the first dataset are described as follows:

$$Y = -1.5X + 17 + r/10 \text{ if } X \le 1.5 \quad (5)$$

$$Y = 1.5X + 12 + r/10 \text{ if } X > 1.5 \quad (6)$$

Herein, X is of [0, 3] and generated with an interval of 0.1. The symbol r denotes a Gaussian random variable with mean = 0 and standard deviation = 1.

The second problem (Dataset 2) involves a simple regression analysis problem with two break points as follows:

$$Y = -1.5X + 14 + r/10 \text{ if } X \le 1 \quad (7)$$

$$Y = 1.5X + 11 + r/10 \text{ if } X > 1.5 \text{ and } X < 2 \quad (8)$$

$$Y = -1.5X + 17 + r/10 \text{ if } X \ge 2 \quad (9)$$

Table 1. Prediction Performance

Phase	Indices	Dataset 1	Dataset 2	Interface yield stress	Plastic viscosity
Training Phase	RMSE	0.09	0.11	6.42	66.60
	MAPE	0.43	0.69	12.05	8.72
	MAE	0.07	0.09	4.43	41.25
	R^2	0.98	0.94	0.89	0.90
Testing Phase	RMSE	0.13	0.16	9.74	89.81
	MAPE	0.69	0.98	14.66	13.64
	MAE	0.11	0.13	7.02	64.20
	R^2	0.97	0.90	0.90	0.78

Besides the aforementioned simulation cases, real-world datasets regarding the prediction of the interface yield stress and plastic viscosity of fresh concrete [25, 26] are used. The SPLRM is employed to capture the mapping relationships between the interface yield stress and plastic viscosity and their influencing factors. The content of cement, water, sand, small coarse gravel, medium coarse gravel, superplasticizer, and time after mixing are used as influencing variables. This dataset includes including 142 experimental tests. To evaluate the model performances, the indices of RMSE, mean absolute percentage error (MAPE), mean absolute error (MAE), and coefficient of determination (R^2) are employed. The model prediction outcomes are reported in Table 1 as well as Fig. 1. The exemplary SPLRM program used for predicting the variable of plastic viscosit is provided in Fig. 2.

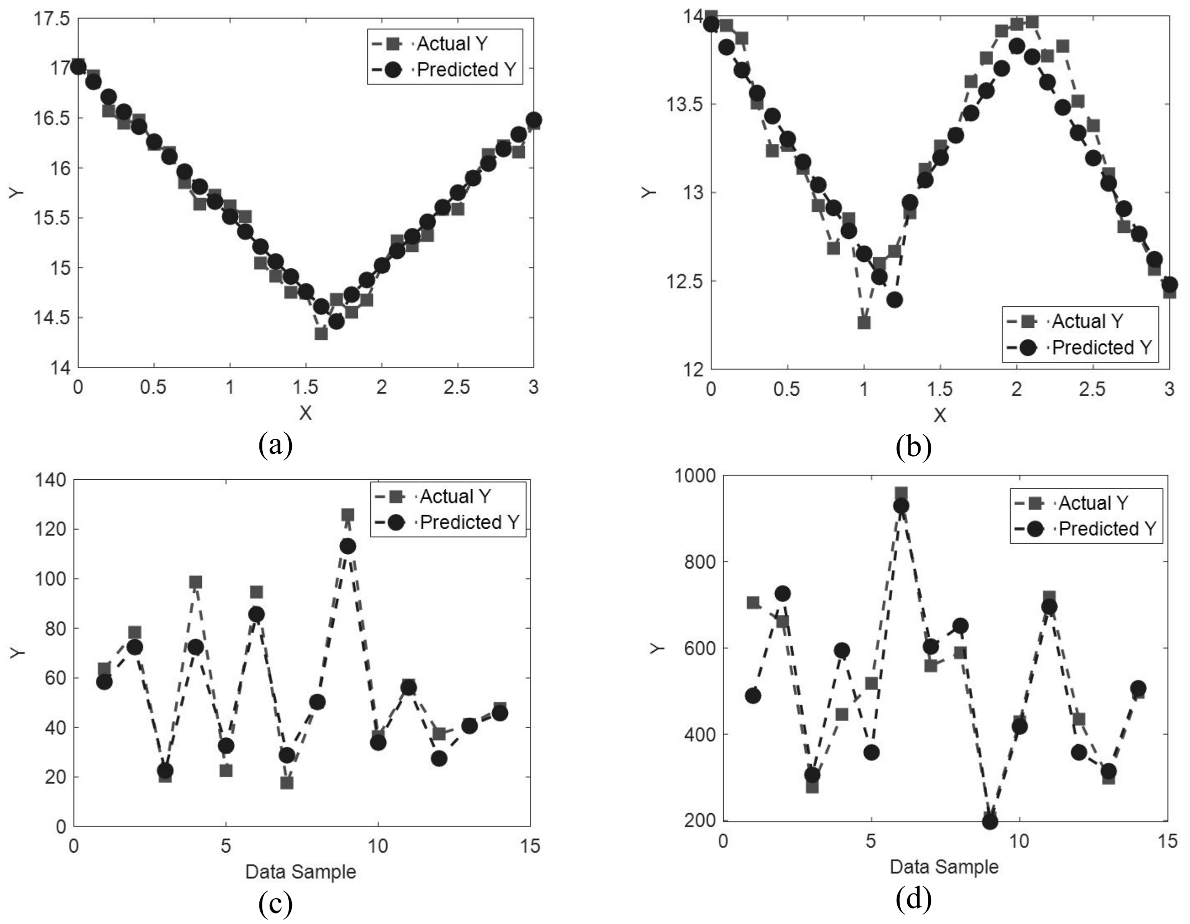

Fig. 1. Demonstrations of testing performances: (a) Dataset 1, (b) Dataset 2, (c) Interface yield stress, and (d) Plastic viscosity

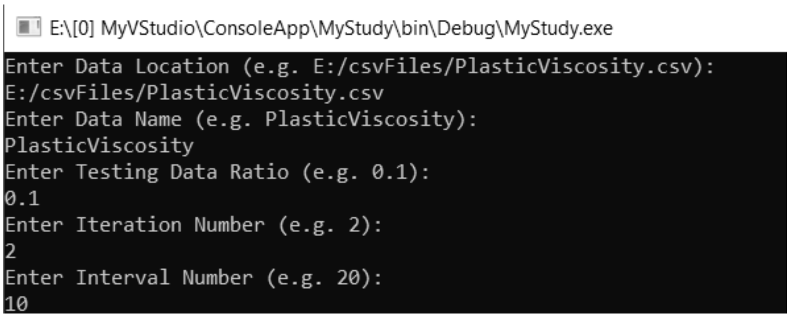

Fig. 2. The complied program SPLRM developd with Visual C# .NET

4. Conclusion

In civil engineering, data fitting via regression analysis is an important task. This study develops a computer program based on the established SPRLM used for approximating nonlinear data relationships. The software program has been developed in Visual C# .NET and its performances have been demonstrated via 4 applications. Good predictive performances show that the newly developed tool can be helpful to assist civil engineers in various data modeling tasks.

Supplementary materials

The compiled SPLRM program and the experimental datasets can be accessed via: https://github.com/NDHoangDTU/SPLRM-Program_VC

References

[1] N.-D. Hoang, "Estimating Punching Shear Capacity of Steel Fibre Reinforced Concrete Slabs Using Sequential Piecewise Multiple Linear Regression and Artificial Neural Network," *Measurement,* vol. 137, pp. 58-70, 2019/01/18/ 2019.

[2] M. D. Nguyen, B. T. Pham, L. S. Ho, H.-B. Ly, T.-T. Le, C. Qi*, et al.*, "Soft-computing techniques for prediction of soils consolidation coefficient," *CATENA,* vol. 195, p. 104802, 2020/12/01/ 2020.

[3] R. J. Freund , W. J. Wilson , and P. Sa, *Regression Analysis: Statistical Modeling of a Response Variable*: Academic Press, 2006.

[4] S. Weisberg, *Applied Linear Regression, Third Edition*: John Wiley & Sons, Printed in the United States of America, 2005.

[5] M. A. Mashrei and A. M. Mahdi, "An Adaptive Neuro-Fuzzy Inference Model to Predict Punching Shear Strength of Flat Concrete Slabs," *Applied Sciences,* vol. 9, p. 809, 2019.

[6] V.-H. Nhu, N.-D. Hoang, V.-B. Duong, H.-D. Vu, and D. Tien Bui, "A hybrid computational intelligence approach for predicting soil shear strength for urban housing construction: a case study at Vinhomes Imperia project, Hai Phong city (Vietnam)," *Engineering with Computers,* February 06 2019.

[7] M.-Y. Cheng and N.-D. Hoang, "Interval estimation of construction cost at completion using least squares support vector machine," *Journal of Civil Engineering and Management,* vol. 20, pp. 223-236, 2014/03/04 2014.

[8] T. Yi, H. Zheng, Y. Tian, and J.-p. Liu, "Intelligent Prediction of Transmission Line Project Cost Based on Least Squares Support Vector Machine Optimized by Particle Swarm Optimization," *Mathematical Problems in Engineering,* vol. 2018, p. 11, 2018.

[9] A. Hocine, "Compressive strength prediction of limestone filler concrete using artificial neural networks," *Advances in Computational Design, Vol. 3, No. 3 (2018) 289-302,* 2018.

[10] A.-D. Pham, N.-D. Hoang, and Q.-T. Nguyen, "Predicting Compressive Strength of High-Performance Concrete Using Metaheuristic-Optimized Least Squares Support Vector Regression," *Journal of Computing in Civil Engineering,* vol. 30, p. 06015002, 2016.

[11] N.-D. Hoang, A.-D. Pham, Q.-L. Nguyen, and Q.-N. Pham, "Estimating Compressive Strength of High Performance Concrete with Gaussian Process Regression Model," *Advances in Civil Engineering,* p. 8, 2016.

[12] F. Khademi, S. M. Jamal, N. Deshpande, and S. Londhe, "Predicting strength of recycled aggregate concrete using Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System and Multiple Linear Regression," *International Journal of Sustainable Built Environment,* vol. 5, pp. 355-369, 2016/12/01/ 2016.

[13] X. Ding, M. Hasanipanah, H. Nikafshan Rad, and W. Zhou, "Predicting the blast-induced vibration velocity using a bagged support vector regression optimized with firefly algorithm," *Engineering with Computers,* 2020/01/23 2020.

[14] N.-D. Hoang, X.-L. Tran, and H. Nguyen, "Predicting ultimate bond strength of corroded reinforcement and surrounding concrete using a metaheuristic optimized least squares support vector regression model," *Neural Computing and Applications,* May 16 2019.

[15] H. Han, X. Cui, Y. Fan, and H. Qing, "Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features," *Applied Thermal Engineering,* vol. 154, pp. 540-547, 2019/05/25/ 2019.

[16] S. Heddam and O. Kisi, "Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree," *Journal of Hydrology,* vol. 559, pp. 499-509, 2018/04/01/ 2018.

[17] N.-D. Hoang, K.-W. Liao, and X.-L. Tran, "Estimation of scour depth at bridges with complex pier foundations using support vector regression integrated with feature selection," *Journal of Civil Structural Health Monitoring,* June 02 2018.

[18] S. Shi, Y. Li, and C. Wan, "Robust continuous piecewise linear regression model with multiple change points," *The Journal of Supercomputing,* pp. 1-23, September 07 2018.

[19] N. Martinez, H. Anahideh, J. M. Rosenberger, D. Martinez, V. C. P. Chen, and B. P. Wang, "Global optimization of non-convex piecewise linear regression splines," *Journal of Global Optimization,* vol. 68, pp. 563-586, July 01 2017.

[20] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "Mathematical programming for piecewise linear regression analysis," *Expert Systems with Applications,* vol. 44, pp. 156-167, 2016/02/01/ 2016.

[21] N.-D. Hoang, Q.-L. Nguyen, and X.-L. Tran, "Automatic Detection of Concrete Spalling Using Piecewise Linear Stochastic Gradient Descent Logistic Regression and Image Texture Analysis," *Complexity,* vol. 2019, p. 14, 2019.

[22] N. D. Hoang and C. H. Le, "Sequential Piecewise Linear Regression software program for nonlinear regression analysis in structural engineering," *DTU Journal of Science and Technology,* vol. 05, pp. 03-09, 2019.

[23] S. E. Ryan and L. S. Porth, "A tutorial on the piecewise regression approach applied to bedload transport data," *Gen. Tech. Rep. RMRS-GTR-189.*

Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 41 p., 2007.

[24] M. E. Greene, O. Rolfson, G. Garellick, M. Gordon, and S. Nemes, "Improved statistical analysis of pre- and post-treatment patient-reported outcome measures (PROMs): the applicability of piecewise linear regression splines," *Quality of Life Research,* vol. 24, pp. 567-573, March 01 2015.

[25] T.-D. Nguyen, T.-H. Tran, and N.-D. Hoang, "Prediction of interface yield stress and plastic viscosity of fresh concrete using a hybrid machine learning approach," *Advanced Engineering Informatics,* vol. 44, p. 101057, 2020/04/01/ 2020.

[26] T.-D. Nguyen, T.-H. Tran, H. Nguyen, and H. Nhat-Duc, "A success history-based adaptive differential evolution optimized support vector regression for estimating plastic viscosity of fresh concrete," *Engineering with Computers,* December 18 2019.