TẠP CHÍ KHOA HỌC & CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 5(54) (2022) 110-115

TRƯỜNG ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

Logistic regression for data classification developed in Excel VBA

Mô hình hồi quy logistic cho phân loại dữ liệu được phát triển trong Excel VBA

Hoang Nhat Duc[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, 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: 16/6/2022, ngày phản biện xong: 20/6/2022, ngày chấp nhận đăng: 25/8/2022)

Abstract

This research work aims at developing a logistic regression based data classification model. This model method is developed in Excel VBA to ease its practical implementations. The newly developed program has been tested with two basic data classification tasks.

Keywords: Logistic regression; Data classification; Excel VBA; Civil engineering.

Tóm tắt

Bài báo xây dựng một công cụ hồi quy logistic được sử dụng cho phân loại dữ liệu. Mô hình này được phát triển trên nền tảng Excel VBA để nâng cao tính ứng dụng của nó trong thực tiễn. Chương trình hồi quy logistic đã được kiểm chứng với hai ứng dụng phân loại cơ bản.

Từ khóa: Mô hình hồi quy logistic*;* phân loại dữ liệu; Excel VBA; Xây dựng dân dụng.

1. Introduction

Data classification is an important task in civil engineering. This task significantly helps enhance the effectiveness of the decision-making process and productivity during various phases of a project. A variety of data-driven methods have been successfully proposed and verified to deal with complex problems such as concrete strength estimation [1-3], pile bearing capacity estimation [4, 5], prediction of behavior of reinforced concrete elements [6, 7], damage recognition [8, 9], estimation of concrete's mechanical properties [10, 11], etc.

Among these data-driven methods, logistic regression model still plays an important role due to its simple structure and ease of implementation. This method is a statistical model that helps derive the probability of one event (out of two alternatives) [12]. For instance, Kim et al. [13] develops a logistic regression model for estimating sinkhole susceptibility due to damaged sewer pipes.

[*]*Corresponding Author:* Hoang Nhat Duc, Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam; Faculty of Civil Engineering, Duy Tan University, Da Nang, 550000, Vietnam
Email: hoangnhatduc@duytan.edu.vn

Hoang [14] relied on this statistical model for detecting asphalt pavement raveling. A Visual C# based stochastic gradient descent logistic regression software program for classifying data has been developed in [15]. Computer vision based approaches for concrete spall detection that utilizes the method of interest has been put forward in [16] and [17].

Although various models based on the logistic regression have been proposed, few studies have dedicated to developing a logistic regression model for data classification in Excel Visual Basic for Applications (VBA). Microsoft Excel is a popular tool for performing calculations in civil engineering. Therefore, the ability of constructing logistic regression models in Excel can be helpful for practicing engineers who wish to apply this statistical method to deal with various data classification problems.

2. Logistic regression developed in Excel VBA

A logistic regression model relies on the logistic function to compute the probability of an event (out of two alternatives). The formula of the logistic function is given by:

$$Logistic(z) = \frac{1}{1+\exp(-z)} \quad (1)$$

Given an input variable $x = [x_0, x_1, x_2, ..., x_D]$ (x_0 is always 1) that provides the features of a data sample, the corresponding probability of the positive class label (or event) can be compute as follows:

$$Logistic(z) = \frac{1}{1+\exp(-w.x)} = \frac{1}{1+\exp(-1\times z)} \quad (2)$$

where

$z = w_0 \times x_0 + w_1 \times x_1 + w_2 \times x_2 + ... + w_D \times x_D$; the vector $w = [w_0, w_1, w_2, ..., w_D]$ denotes the model weight. The $w_1, w_2, ..., w_D$ correspond to D features of a data sample. The quantity w_0 is the bias. $w.x$ denotes the dot product of these two vectors.

Let t denote the target class label (t is either 0 or 1), the L_2 loss function used to train a logistic regression model is given by:

$$L(t, y) = \frac{1}{2}(t-y)^2 \quad (2)$$

Herein, we rely on the gradient descent computation [18] to adapt the vector w of a logistic regression model. To do so, it is required to compute the partial derivative of the loss function L with respect to each element w_i of the vector w. Based on the chain rule, this partial derivative can be stated as follows:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial y} \times \frac{\partial y}{\partial z} \times \frac{\partial z}{\partial w_i} \quad (3)$$

Due to the fact that $\frac{\partial L}{\partial y} = -(t-y)$, $\frac{\partial y}{\partial z} = y \times (1-y)$, and $\frac{\partial z}{\partial x_i} = x_i$, the above equation is equivalent to the following one:

$$\frac{\partial L}{\partial w_i} = -(t-y) \times y \times (1-y) \times x_i \quad (4)$$

Accordingly, the equation used to adapt the model weight is given by:

$$w_i = w_i - \frac{\partial L}{\partial w_i} = w_i - \eta \times (-1) \times (t-y) \times y \times (1-y) \times x_{ii} \quad (5)$$

where η denotes the learning rate parameter.

In general, the steps used to construct a logistic regression model are as follows: (i) Specifying a dataset including a matrix of feature x and a matrix of target t, (ii) Generate an initial value of w randomly, and (iii) Adaptation of w using the equation 6. The Unified Modeling Language (UML) diagram of the Logistic Regression class coded in Excel VBA is presented in **Fig. 2.1**. The function used to train a logistic regression model in Excel VBA has been demonstrated by **Fig. 2.2**. Herein, the function named "*Predict_Probability* ()" is used to compute the quantity *Logistic(z)* stated in equation 2.

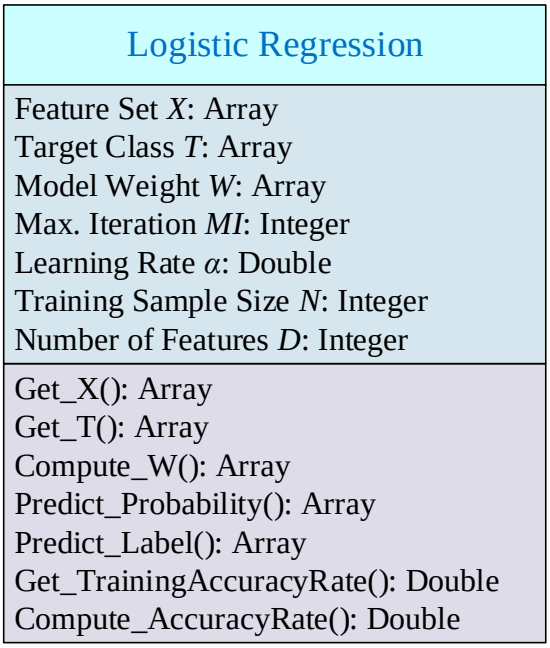

Fig. 2.1 The UML diagram of the Logistic Regression class

```
Public Function Compute_W()
    ReDim W(D, 1)
    Dim i, j, k As Integer
    For k = 1 To D
        W(k, 1) = -0.5 + 1 * Rnd()
    Next k
    Dim X_i As Variant
    ReDim X_i(1, D)
    Dim mM As myMatrix
    Set mM = New myMatrix
    Dim Y As Variant
    For j = 1 To MaxIter ' Loop through each iteration
        Y = Predict_Probability(X)
        For i = 1 To N ' Loop through each data sample
            X_i = mM.ExtractMatrixRow(X, i)
            'Debug.Print "X_i = "
            'mM.Print_Matrix (X_i)

            For k = 1 To D ' Loop through each feature including bias
                W(k, 1) = W(k, 1) + alpha * (T(i, 1) - Y(i, 1)) * Y(i, 1) * _
                (1 - Y(i, 1)) * X(i, k)
            Next k
        Next i
    Next j
    Compute_W = W
End Function
```

Fig. 2.2 The function used to train a logistic regression model in Excel VBA

3. Model applications

In this section, the logistic regression model developed in Excel VBA is used to classify datasets in two applications. The first application involves categorizing a dataset including two variables X_1 and X_2 as shown in Table 1 (for training data) and Table 2 (for testing). Herein, X_0 is for computing the bias element w_0 in w.

Table 1. The training dataset in the first application

X_0	X_1	X_2	T	$P(Y=1)$	Y
1	1.00	5.00	0	0.14	0
1	1.50	6.00	0	0.12	0
1	2.50	5.50	0	0.28	0
1	3.50	7.00	0	0.26	0
1	5.00	3.00	1	0.92	1
1	3.50	2.00	1	0.87	1
1	3.30	3.20	1	0.74	1
1	4.20	2.50	1	0.90	1
1	3.80	3.30	1	0.80	1

Table 2. The testing dataset in the first application

X_0	X_1	X_2	T	$P(Y=1)$	Y
1	1.2	8.8	0	0.02	0
1	4.3	4.9	1	0.69	1
1	1.4	6.8	0	0.07	0
1	5.3	3.2	1	0.93	1

In Tables 1 and 2, T denotes the target output class; T = 1 denotes the event of the positive class. $P(Y=1)$ is the probability of the event Y = 1. Y denotes the predicted class label. The logistic regression model has been trained during 10 iterations and with the learning rate = 0.5. The optimized w = [0.3995, 0.7720, -0.5961]. With this value of w, the logistic regression model achieves a classification accuracy rate = 100% for both training and testing data samples.

The second applications involves the prediction of groutability of granular soil [19]. Ten data samples extracted from [19] has been used to train the logistic regression model. Herein, the $D10_{soil}$ (X_1) and $d90_{grout}$ (X_2) are used to determine the state of groutability. The training and testing datasets are shown in Table 3 and Table 4. The optimized w = [-1.59885, 259334, -0.20492]. With this value of w, the logistic regression model achieves a classification accuracy rate = 90% for the training dataset and 70% for the testing dataset.

Table 3. The training dataset in the second application

X_0	X_1	X_2	T	$P(Y=1)$	Y
1	2.50	0.04	1	0.99	1
1	1.28	0.04	1	0.85	1
1	0.60	0.04	1	0.49	0
1	0.76	0.04	1	0.59	1
1	0.76	0.04	1	0.59	1
1	0.32	0.04	0	0.31	0
1	0.32	0.04	0	0.31	0
1	0.32	0.04	0	0.31	0
1	0.32	0.04	0	0.31	0
1	0.32	0.04	0	0.31	0

Table 4. The Testing dataset in the second application

X_0	X_1	X_2	T	$P(Y=1)$	Y
1	0.54	0.04	0	0.45	0
1	0.60	0.04	0	0.49	0
1	0.76	0.04	1	0.59	1
1	0.12	0.00	1	0.22	0
1	0.11	0.00	0	0.21	0
1	0.12	0.00	1	0.22	0
1	0.54	0.04	0	0.45	0
1	0.60	0.04	0	0.49	0
1	0.76	0.04	1	0.59	1
1	0.76	0.04	0	0.59	1

4. Conclusion

Data classification is a crucial task in civil engineering. This paper has developed and verified a logistic regression model developed in Excel VBA to assist the decision-making process involving pattern classification. The capability of this a logistic regression model has been demonstrated via two simple pattern classification problems. Future extensions of the current work may consider the applications of the logistic regression model in multi-class pattern classification problems and other advanced algorithms for training the model.

Supplementary material

The Excel VBA code of the program can be accessed at:

https://github.com/NDHoangDTU/LR_ExcelVBA

References

[1] H. Nguyen, T. Vu, T.P. Vo, H.-T. Thai (2021) *Efficient machine learning models for prediction of concrete strengths*, Construction and Building Materials, 266 120950.

[2] J.-S. Chou, C.-F. Tsai, A.-D. Pham, Y.-H. Lu (2014) *Machine learning in concrete strength simulations: Multi-nation data analytics*, Construction and Building Materials, 73 771-780.

[3] M. Castelli, L. Vanneschi, S. Silva (2013) *Prediction of high performance concrete strength using Genetic Programming with geometric semantic genetic operators*, Expert Systems with Applications, 40 6856-6862.

[4] N.-D. Hoang, X.-L. Tran, T.-C. Huynh (2022) *Prediction of Pile Bearing Capacity Using Opposition-Based Differential Flower Pollination-Optimized Least Squares Support Vector Regression (ODFP-LSSVR)*, Advances in Civil Engineering, 2022 7183700.

[5] H. Harandizadeh, D. Jahed Armaghani, M. Khari (2021) *A new development of ANFIS–GMDH optimized by PSO to predict pile bearing capacity based on experimental datasets*, Engineering with Computers, 37 685-700.

[6] N.-D. Hoang (2019) *Estimating Punching Shear Capacity of Steel Fibre Reinforced Concrete Slabs Using Sequential Piecewise Multiple Linear Regression and Artificial Neural Network*, Measurement, 137 58-70.

[7] M.N. Uddin, K. Yu, L.-z. Li, J. Ye, T. Tafsirojjaman, W. Alhaddad (2022) *Developing machine learning model to estimate the shear capacity for RC beams with stirrups using standard building codes*, Innovative Infrastructure Solutions, 7 227.

[8] N.D. Hoang (2021) *Image processing based concrete crack classification using Logistic Regression model* DTU Journal of Science and Technology, 2 3-9.

[9] N.-D. Hoang (2019) *Image processing based automatic recognition of asphalt pavement patch using a metaheuristic optimized machine learning approach*, Advanced Engineering Informatics, 40 110-120.

[10] T.-D. Nguyen, T.-H. Tran, N.-D. Hoang (2020) *Prediction of interface yield stress and plastic viscosity of fresh concrete using a hybrid machine learning approach*, Advanced Engineering Informatics, 44 101057.

[11] D.-K. Bui, T. Nguyen, J.-S. Chou, H. Nguyen-Xuan, T.D. Ngo (2018) *A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete*, Construction and Building Materials, 180 320-333.

[12] D.W. Hosmer, S. Lemeshow (2000) *Applied Logistic Regression*, Wiley. ISBN 978-0-471-35632-5.

[13] K. Kim, J. Kim, T.-Y. Kwak, C.-K. Chung (2018) *Logistic regression model for sinkhole susceptibility due to damaged sewer pipes*, Natural Hazards, 93 765-785.

[14] N.-D. Hoang (2019) *Automatic detection of asphalt pavement raveling using image texture based feature extraction and stochastic gradient descent logistic regression*, Automation in Construction, 105 102843.

[15] N.D. Hoang, H.T. Nguyen (2019) *A stochastic gradient descent logistic regression software program for civil engineering data classification developed in .NET framework*, DTU Journal of Science and Technology, 4.

[16] N.-D. Hoang (2020) *Image Processing-Based Spall Object Detection Using Gabor Filter, Texture Analysis, and Adaptive Moment Estimation (Adam) Optimized Logistic Regression Models*, Advances in Civil Engineering, 2020 8829715.

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

[18] S.J. Russell , P. Norvig (2010), *Artificial Intelligence - A Modern Approach*, Pearson Education, Upper Saddle River, New Jersey.

[19] E. Tekin, S.O. Akbas (2017) *Predicting groutability of granular soils using adaptive neuro-fuzzy inference system*, Neural Computing and Applications.